A Cubic Regularized Newton’s Method over Riemannian Manifolds

Junyu Zhang∗ Shuzhong Zhang†

May 16, 2018

Abstract

In this paper we present a cubic regularized Newton’s method to minimize a smooth function over a Riemannian manifold. The proposed algorithm is shown to reach a second-order $\epsilon$-stationary point within $O(1/\epsilon^2)$ iterations, under the condition that the pullbacks are locally Lipschitz continuous, a condition that is shown to be satisfied if the manifold is compact. Furthermore, we present a local superlinear convergence result under some additional conditions.

Keywords: Riemannian optimization, Stiefel Manifold, second-order $\epsilon$-stationary solution, cubic regularization, iteration complexity.

1 Introduction

Optimization over a Riemannian manifold (e.g. Stiefel manifolds) is an important model with numerous applications, including: the PCA, the sparse PCA, eigenvalue and combinatorial optimization; see [6,15,22,23,25]. Specifically, this paper considers the following model

$$\min f(x), \text{ subject to } x \in \mathcal{M},$$

(1)

where $\mathcal{M}$ is a Riemannian manifold. The most commonly encountered Riemannian manifolds include: the Stiefel manifold $\text{St}_{n,k} := \{X \in \mathbb{R}^{n \times k} : X^\top X = I_{k \times k}\}$, the Grassmann manifold $\text{Gr}_{n,k} := \text{St}_{n,k}/\text{St}_{k,k}$, the generalized orthogonal matrices $\{X \in \mathbb{R}^{n \times k} : X^\top MX = I_{k \times k}\}$ with $M > 0$, the sphere $S^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, the low-rank elliptope $\{X \in \mathbb{R}^{n \times n} : \text{diag}(X) = 1_n, X \succeq 0, \text{rank}(X) \leq r \leq n\}$, the low-rank spectrahedron $\{X \in \mathbb{R}^{n \times n} : \text{tr}(X) = 1, X \succeq 0, \text{rank}(X) \leq r \leq n\}$, and a Cartesian product of the above.

∗Department of Industrial & System Engineering, University of Minnesota (zhan4393@umn.edu).
†Department of Industrial & System Engineering, University of Minnesota (zhangs@umn.edu).
Model (1) is not only non-convex from a Euclidean standpoint, but also difficult to preserve feasibility once an iterate steps out of $\mathcal{M}$. An effective way to resolve such difficulty under the framework of Riemannian optimization is to incorporate a so-called retraction operation, which gets the iterate back to the manifold in close proximity. Hence, gradient-type approach with retraction becomes a natural choice as solution method for Riemannian optimization. Numerous studies have been conducted along this line; cf. [17, 18, 29]. Note that such first-order methods typically assure an iteration complexity of $O(1/\epsilon^2)$ to reach a first-order $\epsilon$-stationary point, without guaranteeing any second-order optimality condition. Speaking of which, on the positive side, recent results (such as the ones in [19, 20, 27, 28]) show that many first-order methods, including the Riemannian gradient descent method, actually converge to a strict saddle point with probability 0 if initialized at random. Yet, there is no guarantee of iteration complexity of convergence to a second-order stationary point. Counter examples do exist, showing that the gradient descent algorithm may take exponential number of steps to converge to a second-order stationary point [14]. In other words, these methods do not guarantee to reach a point satisfying the second-order optimality condition in a reasonable amount of time in the worst case. On the other hand, the second-order optimality condition actually turns out to be a key quality to possess in many applications. For example, in [9] the authors proposed a non-convex low-rank approximation

$$\min \langle C, VV^\top \rangle \quad \text{s.t.} \quad V \in \mathbb{R}^{n \times k}, A(VV^\top) = b$$

for the standard SDP

$$\min \langle C, X \rangle \quad \text{s.t.} \quad X \in S^{n \times n}, X \succeq 0, A(X) = b. \quad (3)$$

It was shown in [8] that if $V^*$ is a second-order stationary point of (2) and is rank-deficient then $V^*(V^*)^\top$ is optimal to (3). Under a similar setting, [24] presented a Grothendieck-type inequality

$$\langle C, X^* \rangle \leq \langle C, V^*(V^*)^\top \rangle \leq \langle C, X^* \rangle + \frac{8n}{\sqrt{k}}\|C\|_2,$$

for the SDP relaxation of the max-cut problem, where $X^*$ is optimal to (3) and $V^*$ is an arbitrary second-order stationary point of (2). This bound was later improved in [23] for the low-rank max-cut-SDP (MC-SDP) problem and the max-orthogonal-cut-SDP (MOC-SDP) problem, whose feasible regions are a product of spheres and a product of Stiefel manifolds, respectively. The authors also extend the result to the second-order $\epsilon$-stationary points. Such results reveal the importance of the second-order $\epsilon$-stationary solutions, thus promoting the use of the Hessian information. In the literature, globally convergent algorithms guaranteeing second-order optimality conditions for Riemannian optimization are based on the trust-region method [1, 2, 7]. In general, the Riemannian trust-region (RTR) algorithms return a first-order $\epsilon$-stationary point within $O(1/\epsilon^2)$ iterations, and returns a second-order $\epsilon$-stationary point (to be defined in later in the paper) in $O(1/\epsilon^{2.5})$ iterations. In this paper, we propose a cubic regularized Riemannian Newton’s (CRRN) method to solve the smooth optimization problems on Riemannian manifolds. This method follows the line of research originated from Nesterov’s pioneering work [26], which has drawn significant research attention
in the classical Euclidean optimization context; see e.g. [5, 10–13]. Such methods typically find a second-order \(\varepsilon\)-stationary point within \(O(1/\varepsilon^{3/2})\) iterations in the Euclidean case. In this paper, we prove that this iteration complexity result carries over to general Riemannian optimization. By assuming a certain local Lojasiewicz inequality property or the non-degeneracy of the Riemannian Hessian, local superlinear convergence can be further guaranteed.

**Organization.** In Section 2 we introduce some basic properties of Riemannian manifolds, as well as the notions that are essential for Riemannian optimization. In Section 3, we present our algorithms and their iteration complexity bounds. In Section 4, we discuss the application of our algorithm specialized to the Stiefel manifold, and report results of numerical experiments. All the relevant constants required by the algorithm are explicitly computed in this special case. In Section 5, we conclude the paper.

**Notations.** The Jacobian of a vector function \(g(x) : \mathbb{R}^n \to \mathbb{R}^m\) is denoted as \(Jg(x)\). When dealing with a composite function \(g(f(\xi))\) with \(f(\xi) : \mathbb{R}^k \to \mathbb{R}^n\), we use \(Jg(f(\xi))\) to denote the Jacobian of \(g\) at point \(f(\xi)\) and use \(J_{\xi}g(f(\xi))\) to denote the Jacobian of the function \(g \circ f\) at point \(\xi\). We also differentiate \(\nabla\) and \(\nabla^2\) with \(\nabla_{\xi}\) and \(\nabla^2_{\xi}\) if \(g\) is a scaler function. For a Hessian \(\nabla^2 f(x)\) operating along a direction \(z\), we shall write \(\nabla^2 f(x)[z]\) instead of \(\nabla^2 f(x)z\), as the former is less confusing especially when \(z\) itself is a matrix.

## 2 Riemannian Optimization

This section provides some preliminaries regarding Riemannian optimization, which aims to minimize a smooth function over a Riemannian submanifold \(\mathcal{M}\) of a Euclidean space \(\mathcal{E}\). For an in-depth discussion of Riemannian manifolds, we refer the interested readers to [3,21]. Our brief introduction is to be followed by a discussion about the optimality conditions under the Riemannian optimization setting. We then introduce the notion of the extended retraction. Finally, we introduce the pullback operation and its properties, for the benefit of analyzing the performance of our algorithms in later sections.

### 2.1 Riemannian Manifolds

Consider a differentiable submanifold in a Euclidean space (we follow the notations of [4]).

**Definition 2.1 (Differentiable Submanifold [4])** We call \(\mathcal{M}\) to be a \(d\)-dimensional \(C^k\) differentiable submanifold of \(\mathbb{R}^n\), \(k \geq 2\), if for any \(x \in \mathcal{M}\) there exists a neighbourhood \(B_x\) of \(x\) in \(\mathbb{R}^n\) and a \(C^k\) diffeomorphism \(\psi\) on \(B_x\) into \(\mathbb{R}^n\) such that \(\forall y \in B_x, y \in \mathcal{M}\) if and only if

\[
\psi_{d+1}(y) = \cdots = \psi_n(y) = 0.
\]

A useful insight of this definition is to recognize that a submanifold can locally be induced by a set
of equations. By applying the implicit function theorem, the following is readily seen:

**Corollary 2.2** Let \( \mathcal{M} := \{ x \in \mathbb{R}^n : \phi(x) = 0 \} \) where \( \phi : \mathbb{R}^n \to \mathbb{R}^m \) is a \( C^k \) smooth mapping. Then \( \mathcal{M} \) is an \( n - m \) dimensional \( C^k \) submanifold of \( \mathbb{R}^n \) if for any \( x \in \mathcal{M} \), the Jacobian matrix \( J\phi(x) \) has full rank.

For example, the sphere \( S^{n-1} := \{ x \in \mathbb{R}^n : \|x\|^2 = 1 \} \) is an \( n - 1 \) dimensional \( C^\infty \) submanifold and the Stiefel manifold \( \text{St}_{n,r} = \{ X \in \mathbb{R}^{n \times r} : X^\top X = I_{r \times r} \} \) is an \( nr - \frac{r(r+1)}{2} \) dimensional \( C^\infty \) submanifold. For a submanifold \( \mathcal{M} \) embedded in a Euclidean space \( \mathcal{E} \), the tangent space of \( \mathcal{M} \) at point \( x \in \mathcal{M} \) is denoted by \( T_x \mathcal{M} \), which can be characterized by the following subspace of \( \mathcal{E} \):

**Definition 2.3 (Tangent Space)** Suppose \( \mathcal{M} \) is a submanifold of \( \mathcal{E} \). The tangent space of \( \mathcal{M} \) at \( x \) is defined as

\[
T_x \mathcal{M} = \left\{ \gamma'(0) : \gamma \text{ is a smooth curve with } \gamma(0) = x, \gamma([-\delta, \delta]) \subset \mathcal{M}, \delta > 0 \right\}.
\]

Then the tangent bundle is defined as \( T \mathcal{M} = \{(x, \xi) : x \in \mathcal{M}, \xi \in T_x \mathcal{M}\} \).

For a submanifold induced by \( \mathcal{M} = \{ x \in \mathbb{R}^n : \phi(x) = 0 \} \), an effective way to characterize the tangent space is

\[
T_x \mathcal{M} = \text{Ran}(J\phi(x)^\top)^\perp, \tag{5}
\]

namely the orthogonal complement of the range space of \( J\phi(x)^\top \). If the tangent spaces is equipped with an inner product (hence inducing a metric), then this submanifold is known as Riemannian.

**Definition 2.4 (Riemannian Submanifold)** Suppose \( \mathcal{M} \) is a differentiable submanifold of \( \mathcal{E} \). We call \( \mathcal{M} \) to be a Riemannian submanifold of \( \mathcal{E} \), if for any \( x \in \mathcal{M} \) the tangent space \( T_x \mathcal{M} \) is endowed with the Euclidean inner product; that is, for any \( \eta, \xi \in T_x \mathcal{M} \), if we let \( T_x \mathcal{M} \) be embedded in \( \mathcal{E} \) as a subspace, then the inner product on \( T_x \mathcal{M} \) is defined as \( \langle \xi, \eta \rangle_x := \langle \xi, \eta \rangle \), where the latter is the standard Euclidean inner product. Hence the norm \( \| \cdot \|_x \) induced by \( \langle \cdot, \cdot \rangle_x \) is also the same as the standard \( L_2 \)-norm (or the Frobenius norm in the matrix case).

Roughly speaking, a Riemannian manifold is a differentiable manifold \( \mathcal{M} \) with an inner product \( \langle \cdot, \cdot \rangle_x \) on its tangent spaces, which will be our subject of study henceforth. Let \( f \) be a smooth function defined on \( \mathcal{E} (= \mathbb{R}^n) \). Then \( f \) is also a smooth function on \( \mathcal{M} \). If we denote \( \text{Proj}_{T_x \mathcal{M}} \) to be the orthogonal projection onto the tangent space \( T_x \mathcal{M} \), then one can define the *Riemannian gradient* and the *Riemannian Hessian* as follows:

**Definition 2.5 (Riemannian Gradient and Hessian)** Let \( f \) be a smooth function on \( \mathcal{E} \). The Riemannian gradient \( \text{grad} f(x) \) of \( f \) with respect to a submanifold \( \mathcal{M} \) is a tangent vector in \( T_x \mathcal{M} \) defined by

\[
\text{grad} f(x) = \text{Proj}_{T_x \mathcal{M}}(\nabla f(x)). \tag{6}
\]
For any \( z \in \mathcal{T}_x \mathcal{M} \), the operation of the Riemannian Hessian \( \text{Hess} f(z) \) of \( f \) on \( z \) is defined as
\[
\text{Hess} f(x)[z] = \text{Proj}_{\mathcal{T}_x \mathcal{M}}(D\text{grad} f(x)[z]),
\]
where \( D\text{grad} f(x) \) is the differential of \( \text{grad} f(x) \) in the usual sense. In terms of Jacobian matrix \( J\text{grad} f(x) \), we have \( D\text{grad} f(x)[z] = (J\text{grad} f(x))[z] \).

A core ingredient in Riemannian optimization is the retraction defined as follows (see [4] for more details).

**Definition 2.6 (Retraction)** Let \( \text{Retr}(x, \xi) : \mathcal{T} \mathcal{M} \rightarrow \mathcal{M} \) be a mapping from the tangent bundle \( \mathcal{T} \mathcal{M} \) to the manifold \( \mathcal{M} \). Then we call \( \text{Retr}(\cdot, \cdot) \) a retraction if
\[
\text{Retr}(x, 0) = x, \quad \frac{d}{dt} \text{Retr}(x, t\xi) \bigg|_{t=0} = \xi, \quad \forall x \in \mathcal{M}, \forall \xi \in \mathcal{T}_x \mathcal{M}.
\]
We call \( \text{Retr}(\cdot, \cdot) \) a second-order retraction if it further satisfies
\[
\frac{d^2}{dt^2} \text{Retr}(x, t\xi) \bigg|_{t=0} \in \mathcal{T}_x \mathcal{M}^\perp, \quad \forall x \in \mathcal{M}, \forall \xi \in \mathcal{T}_x \mathcal{M},
\]
where \( \mathcal{T}_x \mathcal{M}^\perp \) denotes the orthogonal complement of \( \mathcal{T}_x \mathcal{M} \) in \( \mathcal{E} \).

Below is a useful property of retraction.

**Proposition 2.7** For a retraction \( \text{Retr}(\cdot, \cdot) \) on a compact submanifold \( \mathcal{M} \subset \mathcal{E} \), there exist constants \( L_1, L_2 > 0 \) such that the following inequalities hold
\[
\|\text{Retr}(x, \xi) - x\| \leq L_1\|\xi\|, \quad \|\text{Retr}(x, \xi) - x - \xi\| \leq L_2\|\xi\|^2,
\]
for all \( x \in \mathcal{M} \) and all \( \xi \in \mathcal{T}_x \mathcal{M} \).

The constants \( L_1 \) and \( L_2 \) may depend on the manifold and the dimensions. As an example, the polar retraction and the QR retraction for the Stiefel manifold \( \text{St}_{n,r} \) all satisfy this regularity condition with some universal constants \( L_1 \) and \( L_2 \) independent of \( n \) and \( r \); see [17]. This Proposition was initially shown as a by-product of Lemma 3 in [7].

### 2.2 The 1st and 2nd Order Optimality Conditions for Riemannian Optimization

Consider the unconstrained problem (1). If \( x^* \) is the optimal solution of the above problem, then
\[
\text{grad} f(x^*) = 0.
\]
Furthermore, if $f$ is second-order continuously differentiable, then

$$\langle \text{Hess} f(x^\ast)[\xi], \xi \rangle \geq 0, \forall \xi \in \mathcal{T}_{x^\ast} \mathcal{M}$$

is also satisfied. We refer the interested readers to [3, 30] for more information on these optimality conditions. Consequently, we call a point to be a first-order $\epsilon$-stationary point if

$$\|\text{grad} f(x^\ast)\| \leq \epsilon;$$

we call $x^\ast$ a second-order $\epsilon$-stationary point if, in addition, it also satisfies

$$\langle \text{Hess} f(x^\ast)[\xi], \xi \rangle \geq -\sqrt{\epsilon} \|\xi\|^2, \forall \xi \in \mathcal{T}_{x^\ast} \mathcal{M}.$$

### 2.3 The Extended Retraction

To begin with, we first propose the extended retraction and discuss its properties. As is defined in Definition 2.6, a retraction is a mapping from the tangent bundle $\mathcal{T} \mathcal{M}$ to the manifold $\mathcal{M}$ and is not defined outside of $\mathcal{T} \mathcal{M}$. Note that both $\mathcal{M}$ and the tangent spaces $\mathcal{T}_x \mathcal{M}$ are embedded in $\mathcal{E}$ and are parameterized with Euclidean coordinates, it will be convenient if we can use the usual differential operators in the Euclidean space. This requires the differentiability of the mapping in an open set of $\mathcal{E}$ rather than restricted to $\mathcal{M}$ or $\mathcal{T}_x \mathcal{M}$ with no interior. Hence we propose to work with the following extended retraction, which extends $\text{Retr}(x, \cdot)$ to the whole $\mathcal{E}$ for all $x \in \mathcal{M}$.

**Definition 2.8 (Extended Retraction)** For a given retraction $\text{Retr}(\cdot, \cdot) : \mathcal{T} \mathcal{M} \to \mathcal{M}$, and a given $x \in \mathcal{M}$, we can continuously and smoothly extend $\text{Retr}(x, \cdot)$ on $\mathcal{T}_x \mathcal{M}$ to the whole space $\mathcal{E}$ by defining the following extended retraction

$$\text{Retr}(x, z) := \text{Retr}(x, \text{Proj}_{\mathcal{T}_x \mathcal{M}}(z)), \forall z \in \mathcal{E}, \forall x \in \mathcal{M},$$

which naturally extends $\text{Retr}(x, \cdot)$ on $\mathcal{T}_x \mathcal{M}$ to the whole space $\mathcal{E}$.

Note that for any $z \in \mathcal{T}_x \mathcal{M}$, the extended retraction remains the original retraction. Without loss of generality, we make the following assumption.

**Assumption 2.9** For the retraction $\text{Retr}(\cdot, \cdot)$ under consideration, we assume that they are already extended smoothly to the whole space $\mathcal{E}$ by incorporating (16). Consequently, the following relationship holds

$$\text{Retr}(x, \cdot) = \text{Retr}(x, \cdot) \circ \text{Proj}_{\mathcal{T}_x \mathcal{M}}.$$

**Proposition 2.10** Suppose that for a submanifold $\mathcal{M} \subset \mathcal{E}$ the retraction $\text{Retr}(\cdot, \cdot)$ satisfies Assumption 2.9. Then,

$$J_\xi \text{Retr}(x, 0) = \text{Proj}_{\mathcal{T}_x \mathcal{M}}, \forall x \in \mathcal{M}.$$
Proof. For any \( x \in \mathcal{M}, \forall \eta \in \mathcal{E} \), Assumption \([2.9]\) and \([8]\) guarantee that
\[
J_\xi \text{Retr}(x, 0)[\eta] = \frac{d}{dt} \text{Retr}(x, t\eta) \bigg|_{t=0} = \frac{d}{dt} \text{Retr}(x, t\text{Proj}_{T_x\mathcal{M}}(\eta)) \bigg|_{t=0} = \text{Proj}_{T_x\mathcal{M}}(\eta).
\]
The result thus follows. \( \square \)

2.4 The Pullback and Its Properties

For any smooth function \( f \) on \( \mathcal{M} \) and a retraction \( \text{Retr}(\cdot, \cdot) \), the pullback of \( f \) at point \( x \), denoted by \( \hat{f}_x \), is defined as
\[
\hat{f}_x(\xi) = f(\text{Retr}(x, \xi)), \forall \xi \in T_x\mathcal{M}.
\] (18)
It locally reparametrizes a function with the points on a subspace \( T_x\mathcal{M} \subset \mathcal{E} \) instead of the points on the manifold \( \mathcal{M} \). When differentiation is performed, the pullback is automatically extended to the whole space \( \mathcal{E} \) through the extended retractions. The gradient and Hessian of the pullbacks connect to the Riemannian gradient and the Riemannian Hessian through the relationships shown in the next three propositions.

**Proposition 2.11 (Pullback Gradient)** Under Assumption \([2.9]\) the pullback \( \hat{f}_x \) satisfies
\[
\nabla_\xi \hat{f}_x(0) = \text{grad} f(x), \forall x \in \mathcal{M}.
\] (19)

Proof. By Proposition \([2.10]\) \( \nabla_\xi \hat{f}_x(0) = \nabla_\xi f(\text{Retr}(x, 0)) = J_\xi \text{Retr}(x, 0) [\nabla f(x)] = \text{grad} f(x). \) \( \square \)

The gradient of the pullback is equal to the Riemannian gradient. However, the Hessian of the pullback is not necessarily equal to the Riemannian Hessian.

**Proposition 2.12 (Pullback Hessian \([7]\))** Under Assumption \([2.9]\) and \([18]\), it holds that
\[
\langle \nabla_\xi^2 \hat{f}_x(0)(x)[\xi], \xi \rangle = \langle \text{Hess} f(x)[\xi], \xi \rangle + \langle \text{grad} f(x), \frac{d^2}{dt^2} \text{Retr}(x, t\xi) \bigg|_{t=0} \rangle, \forall \xi \in T_x\mathcal{M}.
\] (20)

The proof of this proposition can be found in Appendix C of \([7]\).

**Corollary 2.13** If \( \text{Retr}(\cdot, \cdot) \) is a second-order retraction, then the pullback Hessian at 0 coincides with the Riemannian Hessian on the tangent space, i.e.,
\[
\langle \nabla_\xi^2 \hat{f}_x(0)(x)[\xi], \xi \rangle = \langle \text{Hess} f(x)[\xi], \xi \rangle, \forall \xi \in T_x\mathcal{M}.
\]

Proof. By \([9]\), we have \( \frac{d^2}{dt^2} \text{Retr}(x, t\xi) \bigg|_{t=0} \in T_x\mathcal{M}^\perp \). Therefore, we have
\[
\langle \text{grad} f(x), \frac{d^2}{dt^2} \text{Retr}(x, t\xi) \bigg|_{t=0} \rangle = 0.
\]
In combination with Proposition \([2.12]\) this proves the corollary. \( \square \)
such that \( \forall U \) neighbourhood as the unit sphere and Stiefel manifold, this local equation is actually global. Therefore, \( L^2 \) with parameter \( Y \) and \( \eta \).

Proof. For any fixed \( x \in \mathcal{M} \) and \( \eta \in T_x \mathcal{M} \), let us denote \( Y(t) = \text{Retr}(x, t\eta) \). Consequently \( Y(0) = x \) and \( Y'(0) = \eta \). By the continuity of the norm,

\[
\|Y''(0)\| = \left\| \lim_{t \to 0} \frac{Y(t) - Y(0) - Y'(0)t}{t^2} \right\| = \lim_{t \to 0} \frac{\|Y(t) - Y(0) - Y'(0)t\|}{t^2} \leq \lim_{t \to 0} \frac{L_2\|\eta\|^2}{t^2} = 2L_2\|\eta\|^2.
\]

On the other hand, by Proposition 2.12

\[
\left\langle (\nabla_x^2 f_x(0) - \text{Hess}(x))[\eta], \eta \right\rangle \leq \left\| \nabla f(x) \right\| \cdot \left\| Y''(0) \right\|.
\]

Combining this with the bound on \( \|Y''(0)\| \) yields the desired result. \( \square \)

For the iteration complexity of the Riemannian gradient descent or the Riemannian trust-region methods, it is sufficient to know the gradient of the pullback at the origin, i.e., \( \nabla_x \hat{f}_x(0) \). However, to derive faster local convergence we will also need to analyze the pullback gradient in a neighbourhood of 0.

Proposition 2.15 (Pullback Gradient in a Neighbourhood) Under Assumption 2.9 and \( \|\xi\| \) sufficiently small, we have

\[
\left\| \nabla \hat{f}_x(\xi) \right\| = \frac{1}{1 - \left\| \text{Proj}_{T_y \mathcal{M}} - J_\xi \text{Retr}(x, \xi) \right\|_2} \left\| \nabla f(x) \right\|,
\]

where \( y = \text{Retr}(x, \xi) \), and \( \cdot \) denotes the matrix spectral norm.

Before proving this proposition, let us consider an example. Consider the retraction to the unit sphere. In that case, the retraction is \( \text{Retr}(x, \xi) := \frac{x + \xi}{\|x + \xi\|} \). By direct calculation, we have \( \nabla_x \hat{f}_x(\xi) = \frac{1}{\|x + \xi\|} (I - yy^\top) \nabla f(y) = \frac{1}{\|x + \xi\|} \text{grad} f(y) \). Therefore, if \( \|\xi\| \) is small enough, then the difference between \( \nabla_x \hat{f}_x(\xi) \) and \( \text{grad} f(y) \) can be controlled.

Proof. Suppose \( \mathcal{E} = \mathbb{R}^n \), \( \dim(\mathcal{M}) = d < n \). For any \( x \in \mathcal{M} \) and any \( \xi \in T_x \mathcal{M} \), let \( \phi : \mathcal{E} \to \mathbb{R}^{n-d} \) be a smooth local equation of \( \mathcal{M} \) around point \( y := \text{Retr}(x, \xi) \) (see e.g. \([2]\))

That is, there exists a local neighbourhood \( U_y \) of \( y \) in \( \mathcal{E} \), satisfying \( z \in \mathcal{M} \cap U_y \iff \phi(z) = 0, z \in U_y \). For special examples, such as the unit sphere and Stiefel manifold, this local equation is actually global. Therefore, \( \exists \delta > 0 \) such that \( \forall \eta \in \mathcal{E}, \|\eta\|^2 = 1 \), we have \( \phi(\text{Retr}(x, \xi + t\eta)) = 0, \forall t \in (-\delta, \delta) \).

Hence

\[
0 = \frac{d}{dt} \phi(\text{Retr}(x, \xi + t\eta)) \bigg|_{t=0} = J\phi(y)J_\xi \text{Retr}(x, \xi)\eta, \forall \|\eta\|^2 = 1.
\]
Consequently, $J\phi(y)J_x\text{Retr}(x, \xi) = 0$, which means that

$$\text{Ran}(J_x\text{Retr}(x, \xi)) \subset \text{Ran}(J\phi(y)^\top) = \text{Ran}(\text{Proj}_{\tau_xM}).$$

Therefore $\text{Proj}_{\tau_xM}J_x\text{Retr}(x, \xi) = J_x\text{Retr}(x, \xi)$. By direct calculation we have

$$\text{grad}f(y) - \nabla_{\hat{f}_x}(\xi) = \text{Proj}_{\tau_xM}\nabla f(y) - J_x\text{Retr}(x, \xi)^\top\nabla f(y)$$

$$= \text{Proj}_{\tau_xM}^2\nabla f(y) - J_x\text{Retr}(x, \xi)^\top\text{Proj}_{\tau_xM}\nabla f(y)$$

$$= \left(\text{Proj}_{\tau_xM} - J_x\text{Retr}(x, \xi)^\top\right)\text{grad} f(y).$$

Since $y = x$ when $\xi = 0$ and $J_x\text{Retr}(x, \xi)^\top = \text{Proj}_{\tau_xM}$, it follows that

$$\|\text{Proj}_{\tau_xM} - J_x\text{Retr}(x, \xi)^\top\|_2 < 1$$

when $\xi$ is sufficiently small. Therefore,

$$\|\text{grad} f(y)\| - \|\nabla_{\hat{f}_x}(\xi)\| \leq \|\text{Proj}_{\tau_xM} - J_x\text{Retr}(x, \xi)^\top\|_2\|\text{grad} f(y)\|.$$ 

The proposition follows by rearranging the terms in the above inequality. \qed

If $M$ is compact, then we have the following:

**Corollary 2.16 (Proposition 2.15 Refined)** Under Assumption 2.9 and (18), supposing that $M$ is compact, then there exists a constant $C_g > 0$ such that for all $x \in M$ and for all $\xi \in T_xM$ with $\|\xi\|_F \leq C_g$, we have

$$\|\text{grad} f(\text{Retr}(x, \xi))\|_F \leq 2\|\nabla_{\hat{f}_x}(\xi)\|. \quad (23)$$

The constant $C_g$ depends on the manifold $M$ and the retraction. For instance, $C_g = 1/8.62$ if $M$ is the Stiefel manifold with polar retraction.

Next we consider the Lipschitz continuity of the pullback Hessian at the origin. That is, we wish to establish for any $x \in M$ an inequality in the form of

$$|\langle (\nabla^2_{\xi}\hat{f}_x(\eta) - \nabla^2_{\xi}\hat{f}_x(0))[\nu], \nu \rangle| \leq L_H\|\eta\|, \forall \eta \in T_xM, \forall \nu \in T_xM, \|\nu\| = 1, \quad (24)$$

where $L_H$ is independent of $x$. Such property is non-trivial because the tangent bundle is noncompact (unbounded). However, it is true in the case of the Stiefel manifold with polar retraction, and we shall explicitly compute this constant in later sections. Fortunately, we only need a weaker form of local Lipschitz continuity, which is true and is sufficient for our analysis of the general case:

**Lemma 2.17 (Local Lipschitz Continuity of Pullback Hessian)** For any $R > 0$, there exists a constant $L_H^R > 0$ such that condition (24) holds for $\|\eta\| \leq R$:

$$|\langle (\nabla^2_{\xi}\hat{f}_x(\eta) - \nabla^2_{\xi}\hat{f}_x(0))[\nu], \nu \rangle| \leq L_H^R\|\eta\|, \forall \eta \in T_xM, \|\eta\| \leq R, \forall \nu \in T_xM, \|\nu\| = 1. \quad (25)$$
This lemma is a direct consequence of the Lipschitz continuity of $\nabla^2 f$, the smoothness of $\text{Retr}(\cdot, \cdot)$ and the compactness of $\{\eta : \eta \in T_xM, x \in M, \|\eta\| \leq R\}$. Finally, we present a local Lipschitz property on the Riemannian Hessian in the next lemma, whose proof is in Appendix A.

**Lemma 2.18 (Local Lipschitz Continuity of Riemannian Hessian)** Let $M \subset E$ be a compact submanifold and let $f$ be a smooth function with Lipschitz continuous Hessian. Then there exist constants $d_M, D > 0$ such that for any $x, y \in M$, if $\|x - y\| \leq d_M$ then

$$|\lambda_{\min}^M(\text{Hess}(x)) - \lambda_{\min}^M(\text{Hess}(y))| \leq D\|x - y\|$$

where $\lambda_{\min}^M(\text{Hess}(x)) := \inf_{\xi \in T_xM} \left\{ \frac{\langle \text{Hess}(x)[\xi], \xi \rangle}{\|\xi\|^2} \right\}$ and $\|\cdot\|$ is the Euclidean norm.

### 3 Cubic Regularized Newton’s Method

#### 3.1 The Basic Algorithm and Its Convergence

We shall now expand on the local Lipschitz property (25). The following are some estimations of the residuals.

**Proposition 3.1** Let $\text{Retr}(\cdot, \cdot)$ be a second-order retraction on $M$. Suppose condition (25) holds for all $x \in M$ with a uniform constant $L_R^H$ for the pullbacks $\hat{f}_x$. Then

$$\|\nabla_\xi \hat{f}_x(\xi) - \nabla_\xi \hat{f}_x(0) - \nabla^2_\xi \hat{f}_x(0)[\xi]\| \leq \frac{L_R^H}{2}\|\xi\|^2,$$

(27)

$$|\hat{f}_x(\xi) - f(x) - \langle \text{grad}(f)(x), \xi \rangle - \frac{1}{2} \langle \text{Hess}(f)(x)[\xi], \xi \rangle| \leq \frac{L_R^H}{6}\|\xi\|^3$$

(28)

for all $\xi \in T_M, \|\xi\| \leq R$, where the constant $L_R^H$ is independent of $x$.

We remark here that estimate (28) first appeared in [7] without condition (25) and the constraint that $\|\xi\| \leq R$. As a consequence, it now follows from (24). This proposition immediately suggests that for any $\sigma > L_R^H$,

$$m_{x,\sigma}(\xi) := \hat{f}_x(0) + \langle \text{grad}(f)(x), \xi \rangle + \frac{1}{2} \langle \text{Hess}(f)(x)[\xi], \xi \rangle + \frac{\sigma}{6}\|\xi\|^3$$

(29)

is an upper bound of $\hat{f}_x(\xi)$ in the subspace $T_xM$ if $\|\xi\| \leq R$. Therefore, whenever $R$ is large enough so as to ensure the global minimum of $m_{x,\sigma}(\xi)$ is in the interior of the disk $\xi \in T_M, \|\xi\| \leq R$, then following the principle of majorization-minimization it makes sense to minimize $m_{x,\sigma}(\xi)$ over $T_xM$ as an iterative subroutine, instead of minimizing $\hat{f}_x(\xi)$ itself. More specifically, the iterates run as:

$$\begin{cases}
\xi_k := \arg \min m_{x,\sigma}(\xi), \text{ subject to } \xi \in T_xM, \\
x_{k+1} := \text{Retr}(x_k, \xi_k).
\end{cases}$$

(30)
Note that if we denote the projection onto $\mathcal{T}_x\mathcal{M}$ to be operator $P_x$, then the constrained cubic subproblem can be equivalently rewritten as

$$
\xi_k = \arg \min_{\xi} \hat{m}_{x_k, \sigma}(\xi) \quad (31)
$$

$$
:= \arg \min_{\xi} \hat{f}_x(0) + \langle \nabla f(x), \xi \rangle + \frac{1}{2}(P_x \circ \text{Hess} f(x) \circ P_x)[\xi] + \sigma \|\xi\|^3. \quad (32)
$$

In principle, we should also replace grad $f(x)$ by $P_x[\text{grad} f(x)]$. Since grad $f(x)$ is in $\mathcal{T}_x\mathcal{M}$, the projection becomes redundant. This means that we essentially end up with an unconstrained problem over $\mathcal{E}$, which can be solved to global optimality; see e.g. [10][20].

Below we present a number of constants to be used later. Let $\nabla f$ be Lipschitz continuous on the convex hull $\text{Conv}(\mathcal{M})$ with Lipschitz constant $\ell_f$. Define $\ell_f := \max_{x \in \text{Conv}(\mathcal{M})} \|\nabla^2 f(x)\|_2$ as the Lipschitz constant for $\nabla f$ over $\text{Conv}(\mathcal{M})$, denote $G := \max_{x \in \mathcal{M}} \|\nabla f(x)\|_F$, and $k_B := \max_{x \in \mathcal{M}} \max_{\xi \in \mathcal{T}_x\mathcal{M}, \|\xi\|=1} \|\text{Hess} f(x)[\xi]\|$.  

**Algorithm 1:** Cubic Regularized Newton’s Method over Riemannian Manifold

1. **Input:** an initial point $x_0 \in \mathcal{M}$, a retraction $\text{Retr}(\cdot, \cdot)$, a parameter
   $$
   \sigma > \max \left\{ \left( \sqrt{10L_2k_B + \frac{2}{3}L^R_H + 9L^3_2G + 3L_2\sqrt{G}} \right)^2, 1 \right\}
   $$
   where $L^R_H$ is defined in Lemma [2,17] with $R = 3k_B + 3\sqrt{G}$ and an iteration number $T$.

2. **for** $k = 0, 2, ..., T - 1$ **do**
   3. **Solve** $\xi_k = \arg \min_{\xi} \hat{m}_{x_k, \sigma}(\xi)$.
   4. **Update** $x_{k+1} = \text{Retr}(x_k, \xi_k)$.

5. **Output:** Let $k^* := \arg \min_{0 \leq k \leq T - 1} \|\xi_k\|^3$, and return $x_{k^* + 1}$.

**Theorem 3.2** Let the sequence $(X_k, \xi_k)$ be generated by Algorithm 1, with the parameters chosen to satisfy $\sigma > \max \left\{ \left( \sqrt{10L_2k_B + \frac{2}{3}L^R_H + 9L^3_2G + 3L_2\sqrt{G}} \right)^2, 1 \right\}$ and $R = 3k_B + 3\sqrt{G}$, if we choose to set $k^* := \arg \min_{0 \leq k \leq T - 1} \|\xi_k\|^3$ and

$$
T \geq \frac{4(f(x_0) - f)}{\tau_1} \cdot \max \left\{ \frac{1}{C_g^2}, \left( \frac{L_1}{d_{\mathcal{M}}} \right)^3, \left( \frac{DL_1 + \sigma/2}{\epsilon^2} \right)^3, \frac{\tau_2^3}{\epsilon^2} \right\},
$$

where $C_g$ is defined in Corollary [2,10] $d_{\mathcal{M}}$ is defined in Lemma [2,18] $f$ is any lower bound of the optimal value, $\tau_1 = \sigma - 10L_2k_B - 6L_2\sqrt{\sigma G} - \frac{2}{3}L^R_H$ and $\tau_2 = \sigma + L^R_H + 10L_2k_B + 6L_2\sqrt{\sigma G}$. Then, the returned point $x_{k^* + 1}$ is a second-order $\epsilon$-stationary point satisfying [14] and [15].

**Proof.** For the ease of notation, denote $g_k := \text{grad} f(x_k)$ and $B_k = P_{x_k} \text{Hess} f(x_k) P_{x_k}$, where $P_{x_k}$ is the orthogonal projection onto $\mathcal{T}_{x_k}\mathcal{M}$. Then the subproblem can be represented as $\xi_k = \arg \min_{\xi} g_k^T \xi + \frac{1}{2} \xi^T B_k \xi + \frac{\sigma}{6} \|\xi\|^3$. According to Lemma 2.2 in [14], $\|\xi_k\| \leq \frac{\epsilon}{\sigma} \max \left\{ k_B, \sqrt{\sigma \|g_k\|} \right\}$. 

11
Since \( \|g_k\| \leq \|\nabla f(x_k)\| \leq G, \sigma \geq 1 \), we have
\[
\|\xi_k\| \leq \frac{3}{\sigma} \max\{k_B, \sqrt{\sigma G}\} \leq \frac{3k_B}{\sigma} + 3 \sqrt{\frac{G}{\sigma}} \leq R. \tag{33}
\]

This validate our choice of \( R \). By the global optimality conditions provided in \cite{26}
\[
(B_k + \lambda^*_k I)\xi_k + g_k = 0, \quad \lambda^*_k = \frac{\sigma}{2}\|\xi_k\|, \quad B_k + \lambda^*_k I \succeq 0. \tag{34}
\]

The first two conditions of \( \ref{41} \) further result in
\[
g_k^\top \xi_k + \xi^\top_k B_k \xi_k + \frac{\sigma}{2}\|\xi_k\|^3 = 0. \tag{35}
\]

In the absence of second-order retraction, Hess \( f(x_k) \) is no longer equal to the Hessian of the pullback \( \nabla_x^2 \hat{f}_{x_k}(0) \) and the majorization property \( \hat{m}_{x_k,\sigma}(\xi) \geq \hat{f}_{x_k}(\xi) \) no longer holds. In this case, let us denote the matrix \( H_k := P_{x_k} \nabla^2_{\xi} \hat{f}_{x_k}(0) P_{x_k} \). We start from Lemma \ref{2.17} with constant \( L_H^R \) and get
\[
\hat{f}_{x_k}(\xi_k) \leq \hat{f}_{x_k}(0) + g_k^\top \xi_k + \frac{1}{2} \xi^\top_k H_k \xi_k + \frac{L_H^R}{6}\|\xi_k\|^3
\]
\[
= \hat{f}_{x_k}(0) - \xi^\top_k B_k \xi_k + \frac{1}{2} \xi^\top_k H_k \xi_k + \left( \frac{L_H^R}{6} - \frac{\sigma}{2} \right)\|\xi_k\|^3
\]
\[
= \hat{f}_{x_k}(0) - \frac{1}{2} \xi^\top_k \left( B_k + \frac{\sigma}{2}\|\xi_k\|I \right) \xi_k + \frac{1}{2} \xi^\top_k (H_k - B_k) \xi_k - \left( \frac{\sigma}{4} - \frac{L_H^R}{6} \right)\|\xi_k\|^3.
\]

Note that by condition \( \ref{41} \) we have \( \frac{1}{2} \xi^\top_k (B_k + \frac{\sigma}{2}\|\xi_k\|I) \xi_k \geq 0 \), and by Corollary \ref{2.13} we have
\( \frac{1}{2} \xi^\top_k (H_k - B_k) \xi_k \leq L_2\|g_k\|\|\xi_k\|^2 \). Combining these inequalities leads to
\[
f(x_k) - f(x_{k+1}) \geq \left( \frac{\sigma}{4} - \frac{L_H^R}{6} - L_2 \cdot \frac{\|g_k\|}{\|\xi_k\|} \right)\|\xi_k\|^3. \tag{36}
\]

Recall that the optimality condition in \( \ref{41} \) gives \( -g_k = (B_k + \frac{\sigma}{2}\|\xi_k\|I)\xi_k \). By combining this equality with \( \ref{33} \), we have
\[
\|g_k\| \leq \left( \|B_k\|_2 + \frac{\sigma}{2}\|\xi_k\| \right)\|\xi_k\| \leq \left( \frac{5}{2}k_B + \frac{3}{2}\sqrt{\sigma G} \right)\|\xi_k\|. \tag{37}
\]

Putting \( \ref{36} \) and \( \ref{37} \) together yields
\[
f(x_k) - f(x_{k+1}) \geq \frac{1}{4} \left( \sigma - 10L_2k_B - 6L_2\sqrt{\sigma G} - \frac{2}{3}L_H^R \right)\|\xi_k\|^3 = \frac{\tau_1}{4}\|\xi_k\|^3. \tag{38}
\]

When \( \sigma > \left( \sqrt{10L_2k_B + \frac{2}{3}L_H^R + 9L_2^2G + 3L_2\sqrt{G}} \right)^2 \), the decrease is positive. Summing the inequalities up yields
\[
\sum_{k \in I} \|\xi_k\|^3 \leq \frac{4(f(x_0) - f)}{\tau_1}.
\]
where the existence of $f$ is guaranteed by the compactness of $M$. Following the way that $T$ and $k^*$ are set, we have
\[
\|\xi_{k^*}\| \leq \min \left\{ C_g, d_M / L_1, \tau_2^{-\frac{1}{2}} \epsilon \frac{1}{2}, (DL_1 + \sigma / 2)^{-\frac{1}{2}} \epsilon \frac{1}{2} \right\}. \tag{39}
\]

It remains to prove that it is an $\epsilon$-solution. To this end, note that
\[
\begin{cases}
\| \nabla \hat{f}_{x_{k^*}}(\xi_{k^*}) - g_{k^*} - H_{k^*} \xi_{k^*} \| \leq \frac{L R_H}{2} \| \xi_{k^*} \|^2, \\
\| g_{k^*} + B_{k^*} \xi_{k^*} \| = \frac{\sigma}{2} \| \xi_{k^*} \|^2, \\
\| (H_{k^*} - B_{k^*}) \xi_{k^*} \| \leq 2L_2 \| g_{k^*} \| \| \xi_{k^*} \| \leq L_2 (5k_B + 3 \sqrt{\sigma G}) \| \xi_{k^*} \|^2,
\end{cases}
\]
where the first inequality is due to Lemma 2.17 and Proposition 3.1, the equality is due to condition (34), and the last inequality is due to Corollary 2.14 and (37). Combining these, we have
\[
\| \nabla \hat{f}_{x_{k^*}}(\xi_{k^*}) \| \leq \frac{T_2}{2} \| \xi_{k^*} \|^2.
\]
Since $\| \xi_{k^*} \| \leq C_g$, we further obtain
\[
\| \text{grad} f(x_{k^*+1}) \| \leq 2 \| \nabla \hat{f}_{x_{k^*}}(\xi_{k^*}) \| \leq T_2 \| \xi_{k^*} \|^2 \leq \epsilon. \tag{40}
\]

By condition (34) and Lemma 2.18
\[
\lambda_{\text{min}}(B_{k^*+1}) \geq \lambda_{\text{min}}(B_{k^*}) - DL_1 \| \xi_{k^*} \| \geq - \left( DL_1 + \frac{\sigma}{2} \right) \| \xi_{k^*} \| \geq - \sqrt{\epsilon}.
\]
Hence $\langle \text{Hess} f(x_{k^*+1})[\eta], \eta \rangle \geq - \sqrt{\epsilon} \| \eta \|^2, \ \forall \eta \in T_{x_{k^*+1}} M$. The proof is complete. \qed

### 3.2 Speeding Up Local Convergence

In this subsection, we investigate the possibility of speeding up the theoretical convergence rate under some additional conditions. One such condition is the so-called gradient-dominant property.

#### 3.2.1 Gradient-Dominant Functions

**Definition 3.3 (Locally Gradient-Dominant Function)** For a smooth function $f$ defined on a manifold $M$, if for any local minimum point $\bar{x}$, there exists a neighbourhood $U_{\bar{x}}$ of $\bar{x}$ such that for all $x \in U_{\bar{x}}$, we have
\[
f(x) - f(\bar{x}) \leq \tau_f \| \text{grad} f(x) \|^p,
\]
where $\tau_f > 0$ is some universal constant independent of $\bar{x}$, then we call $f$ to be a locally gradient-dominant function of degree $p$.\[13\]
This definition stipulates that in a neighbourhood of a local minimum point, the function value is dominated by the size of Riemannian gradient. As an example, consider the principal component analysis (PCA), which can be posed as
\[
\min_X \langle A, XX^\top \rangle, \text{ subject to } X \in \text{St}_{n,r}.
\]
The objective function of the problem actually satisfies the gradient domination property with degree \( p = 2 \). One can further prove that every second-order \( \varepsilon \)-stationary point is close to the global optimum and every exact second-order stationary point is a global minimum.

**Theorem 3.4** Let the sequence \( \{ x_k, \xi_k \} \) be generated by Algorithm 1. Assume that \( \{ x_k \} \) is converging to a local minimum \( \bar{x} \) and the whole sequence \( \{ x_k \} \) lies within the neighbourhood \( U_{\bar{x}} \) where the objective function \( f \) is locally gradient-dominant with degree \( p \). Define
\[
z_k = \tau_f^{\frac{3}{2p-3}} \left( \frac{4}{\tau_1} \right)^{\frac{2p}{2p-3}} \tau_2^{\frac{3p}{2p-3}} \cdot (f(x_k) - f(\bar{x})),
\]
where \( \tau_1, \tau_2 \) are defined in Theorem 3.2. We have:

- For \( p = \frac{3}{2} \), it holds that \( z_k \leq (\frac{1}{2})^k z_0 \).
- For \( 1 \leq p < \frac{3}{2} \), it holds that:
  - if \( z_0 \geq 2^{\frac{3}{3-2p}} \), then \( z_k \leq z_{k-1} \);
  - if \( z_0 \leq 2^{\frac{3}{3-2p}} \), then after at most \( t = 1 + \left\lceil \frac{3}{3-2p} \right\rceil \) steps, we have \( z_t < 1 \);
  - if \( z_0 < 1 \), then by letting \( \beta = \frac{2p}{3-2p} \) we have \( z_k \leq \frac{1}{((1-\beta^{-1})^{k+1})^p} = \mathcal{O}(k^{-\beta}) \). In particular, when \( p = 1 \) then \( z_k \leq \mathcal{O}(k^{-2}) \).
- For \( p > \frac{3}{2} \), it holds that:
  - if \( z_0 \geq 1 \), then \( \frac{z_k}{z_{k-1}} \leq \left( 1 + z_0^{\frac{3-2p}{2p}} \right)^{-1} < 1 \);
  - if \( z_0 < 1 \), then \( z_k \leq z_{k-1}^{\frac{2p}{3}} \).

We remark that the rate in Theorem 3.2 yields \( \|\nabla f(x_k)\| \leq \mathcal{O}(k^{-\frac{4}{3}}) \). Combined with the gradient-dominant condition of degree \( p \), this results in a sub-linear convergence rate of \( \mathcal{O}(k^{-\frac{2p}{3-2p}}) \), which is slower than the \( \mathcal{O}(k^{-\frac{2p}{3-2p}}) \) rate here with \( 1 \leq p < \frac{3}{2} \).

**Proof.** Recall (38) and (40) state that
\[
f(x_k) - f(x_{k+1}) \geq \frac{T_1}{4} \|\xi_k\|^2 \text{ and } \|\nabla f(x_{k+1})\| \leq \tau_2 \|\xi_k\|^2.
\]
These relationships indicate
\[
f(x_k) - f(x_{k+1}) \geq \frac{T_1}{4} \tau_2^{-\frac{3}{2}} \|\nabla f(x_{k+1})\|^\frac{3}{2}.
\]
Together with the gradient-dominant condition \((41)\), we get
\[
f(x_k) - f(x_{k+1}) \geq \frac{T_1}{4r^2}(f(x_{k+1}) - f(\bar{x}))^{\frac{3}{2p}}.
\]

If we define \(z_k\) according to the equation \((42)\), then the above inequality can be simplified to
\[
z_k \geq z_{k+1} + z_{k+1}^{\frac{3}{2p}}.
\]

Next let us discuss various values of \(p\).

First, when \(p = \frac{3}{2}\), \((43)\) becomes \(z_k \geq 2z_{k+1}\) and the result follows.

Second, consider \(1 \leq p < \frac{3}{2}\). Suppose \(z_k \geq 1\). If \(z_{k+1} < 1\), this stage is over; otherwise, \(z_{k+1} \geq 1\), and since \(\frac{3}{2p} > 1\) we have \(z_{k+1} \geq z_{k+1}^{\frac{3}{2p}}\). Then \((43)\) implies that \(z_k \geq \max\{2z_{k+1}, z_{k+1}^{\frac{3}{2p}}\}\). In case \(z_{k+1} \geq 2^{\frac{3}{2p}}\), we have \(2z_{k+1} \leq z_{k+1}^{\frac{3}{2p}}\) and so the inequality \(z_k \geq z_{k+1}^{\frac{3}{2p}}\) dominates. Consequently, in that scenario after \(t\) steps, we have \(\ln z_t \leq (\frac{2p}{3})^t \ln z_0\).

Whenever \(z_{k_0} \leq 2^{\frac{3}{2p}}\) for some \(k_0\), then \(z_k \geq 2z_{k+1}\) starts to dominate henceforth. Therefore, \(z_{k_0+t}\) is guaranteed to be less than 1 when \(t \geq 1 + \lceil \frac{3}{3-2p} \rceil\).

Now, for simplicity suppose \(z_0 < 1\). Letting \(\beta = 2p - \frac{3}{2p}\), \((43)\) leads to
\[
\left(\frac{1}{z_k}\right)^{\frac{1}{\beta}} \leq \left(\frac{1}{z_{k+1}}\right)^{\frac{1}{\beta}} \left(1 + z_{k+1}^{\frac{1}{\beta}}\right)^{-\frac{1}{\beta}}.
\]

Note that function \(r(s) := (1 + s)^{-1/\beta}\) is strictly convex in \(s\). Hence for any \(0 < s < 1\), we have \(r(s) < r(0) + \frac{r(1) - r(0)}{1-s} \cdot s = 1 - (1 - 2^{-1/\beta})s\). Substituting this inequality into \((44)\) with \(s = z_{k+1}^{1/\beta}\) we obtain
\[
\left(\frac{1}{z_k}\right)^{\frac{1}{\beta}} \leq \left(\frac{1}{z_{k+1}}\right)^{\frac{1}{\beta}} - (1 - 2^{-1/\beta}),
\]
which further implies that
\[
z_k \leq \frac{1}{((1 - 2^{-1/\beta})k + z_0^{-1/\beta})^{\beta}} \leq \frac{1}{((1 - 2^{-1/\beta})k + 1)^{\beta}} = O(k^{-\beta}).
\]

This completes our analysis for the case \(p < \frac{3}{2}\).

For the case \(p > \frac{3}{2}\), when \(z_0 \geq 1\), \((43)\) immediately leads to
\[
\frac{z_{k+1}}{z_k} \leq (1 + z_{k+1}^{\frac{1}{2p}})^{-1} \leq \left(1 + z_0^{\frac{2p}{3-2p}}\right)^{-1} < 1,
\]
which is a linear rate of convergence to the interval \((0, 1)\). Whenever \(z_0 < 1\), then \((43)\) yields
\[
z_{k+1} \leq z_k^{\frac{3}{2p}}
\]
where \(\frac{3}{2p} > 1\). The theorem is proven. \(\square\)
3.2.2 Nondegenerate Riemannian Hessian

A second condition under which a faster local convergence holds is when a local minimum point has positive definite Riemannian Hessian. Under this condition, it maintains a local quadratic rate of convergence, which is typical for the Newton type methods in the usual Euclidean case. It is interesting to note that this property carries over to Riemannian optimization as well.

Formally, let us call a second-order stationary point \( \bar{x} \) to be non-degenerate if there exists a constant \( \delta > 0 \) such that for any \( \xi \in T_{\bar{x}}M \), we have

\[
\langle \text{Hess} f (\bar{x})[\xi], \xi \rangle \geq \delta \|\xi\|^2.
\] (46)

**Theorem 3.5 (Local Quadratic Convergence)** Let the sequence \( \{x_k, \xi_k\} \) be generated by Algorithm 1. Suppose \( x_k \to \bar{x} \) where \( \bar{x} \) is a nondegenerate local minimum satisfying (46). Suppose that \( x_0 \) satisfies (46) with some constant \( \delta_0 > 0 \) and \( \|\xi_0\| \leq \min \left\{ \frac{3d_M}{5L_1}, \frac{3\delta_0}{10DL_1}, \frac{\delta_0}{\tau_2} \right\} \), where \( D, d_M \) are defined in Lemma 2.18, and \( L_1 \) is defined in Proposition 2.7. Then, a quadratic rate of convergence holds:

\[
\frac{\tau_2}{\delta_0/2} \|\xi_{k+1}\| \leq \left( \frac{\tau_2}{\delta_0/2} \|\xi_k\| \right)^2,
\] where \( \tau_2 \) is defined in Theorem 3.2. As a result,

\[
\|\text{grad} f (x_k)\| \leq O(2^{-2k}), \text{ and } \langle \text{Hess} f (x_k)[\xi], \xi \rangle \geq \frac{\delta_0}{2} \|\xi\|^2, \forall \xi \in T_{x_k}M.
\]

**Proof.** Since \( \bar{x} \) satisfies the non-degeneracy condition (46) and \( \text{Hess} f \) is Lipschitz continuous in a neighbourhood of \( \bar{x} \), there exists a neighbourhood \( U_{\bar{x}} \subset M \) of \( \bar{x} \) and a positive constant \( 0 < \delta \leq \delta_0 \), such that for any \( x \in U_{\bar{x}} \), (46) is satisfied for this \( \delta \). Since \( \{x_k\} \) converges to \( \bar{x} \) and \( \|\xi_k\| \) converges to 0, the condition regarding \( \xi_0 \) will be satisfied for some \( \xi_k \). One may redefine this \( x_k \) to be \( x_0 \), and the condition on \( \xi_0 \) is then satisfied.

Let us proceed to the proof of the theorem. Now the assumption implies that \( \frac{2\tau_2}{\delta_0} \|\xi_0\| \leq \frac{1}{2} < 1 \) and \( \langle \text{Hess} f (x_0)[\xi], \xi \rangle \geq \delta_0 \|\xi\|^2 \geq \frac{1}{2} \delta_0 \|\xi\|^2, \forall \xi \in T_{x_0}M \). We claim that (will prove this claim in one moment) for any nonnegative integer \( k \),

\[
\langle \text{Hess} f (x_k)[\xi], \xi \rangle \geq \frac{1}{2} \delta_0 \|\xi\|^2, \forall \xi \in T_{x_k}M.
\] (47)

Notice that (34) and (40) give \( \langle \text{Hess} f (x_k) + \frac{\sigma^2}{\delta_0} \|\xi_k\| I \rangle[\xi_k] = g_k \) and \( \|g_k\| \leq \tau_2 \|\xi_{k-1}\|^2 \). Therefore,

\[
\|\xi_k\| \leq \frac{\|g_k\|}{\frac{\sigma^2}{\delta_0} + \frac{\tau_2}{\delta_0} \|\xi_k\|} \leq \frac{\tau_2 \|\xi_{k-1}\|^2}{\frac{\sigma^2}{\delta_0} + \frac{\tau_2}{\delta_0} \|\xi_k\|} \leq \frac{\tau_2 \|\xi_{k-1}\|^2}{\delta_0/2}.
\]
Defining \( z_k := \frac{2^{n-1}}{2^n} \|\xi_k\| \), the above inequality is equivalent to \( z_k \leq z_{k-1}^2 \). Since \( z_0 \leq \frac{1}{2} \), the whole sequence converge quadratically to 0. Specifically, this leads to

\[
\|\xi_k\| = \frac{\delta_0}{2^{2k}} z_k \leq \frac{\delta_0}{2^{2k}} z_0^2 \leq \frac{\delta_0}{2^{2k}} \cdot \frac{1}{2^{2k}}.
\]

By (10), we have \( \|\text{grad}(f(x_k))\| \leq \tau_2 \|\xi_k-1\|^2 = \mathcal{O}(2^{-2k}) \), as required. Now it remains only to show (17). We shall prove by induction. The base case holds trivially. Suppose we already have \( \|x_k - x_0\| \leq \sum_{k=1}^{k_0} \|x_k - x_{k-1}\| \leq L \sum_{k=0}^{k_0-1} \|\xi_k\| \leq L \frac{\delta_0}{2^{2k}} \left(z_0 + \sum_{k=1}^{\infty} (z_0)^{2k}\right) \leq L \frac{\delta_0}{2^{2k}} \left(z_0 + \frac{z_0^2}{1-z_0}\right) \).

Since \( z_0 \leq \frac{1}{2} \), we have \( \frac{z_2}{1-z_0} = \frac{1-z_0}{1-(\frac{1}{2})^2} = \frac{3}{2} z_0 \), consequently \( \|x_k - x_0\| \leq L \frac{\delta_0}{2^{2k}} \cdot \frac{3}{2} z_0 \). Suppose \( \|\xi_0\| \) satisfies the condition of the theorem, \( \|x_k - x_0\| \leq d_M \). Then by Lemma 2.18

\[
\langle \text{Hess}f(x_k)\xi, \xi \rangle \geq (\delta_0 - D \|x_k - x_0\|) \|\xi\|^2 \geq \frac{1}{2} \delta_0 \|\xi\|^2, \forall \xi \in T_{x_0}M,
\]

completing the proof.

\[\square\]

4 The Case of Stiefel Manifolds

When equipped with the standard Euclidean inner product, the so-called Stiefel manifold \(\text{St}_{n,r}\) is an \(nr - \frac{r(r+1)}{2}\) dimensional Riemannian submanifold. Its tangent space is given by \(T_X \text{St}_{n,r} = \{Z \in \mathbb{R}^{n \times r} : X^\top Z + Z^\top X = 0\}\). The orthogonal projection onto the tangent space is given by

\[
P_X(G) = G - \frac{1}{2}(XX^\top G + XG^\top X), \quad \forall G \in \mathbb{R}^{n \times r}.
\]  

Consequently, the Riemannian gradient of \(f\) at point \(X \in \text{St}_{n,r}\) equals \(\text{grad}(f)(X) = P_X(\nabla f(X))\). Consider the polar retraction defined as

\[
\text{Retr}(X, Z) = (X + Z)(I_r + Z^\top Z)^{-\frac{1}{2}}, \quad \forall Z \in T_X \text{St}_{n,r}.
\]

For \(Z\) outside of \(T_X \text{St}_{n,r}\), the extended polar retraction is defined by replacing \(Z\) with \(P_X(Z)\) in the above formula. It is worth noting that the polar retraction is a second-order retraction according to [4]. Moreover, for polar retraction, Proposition 2.7 is satisfied with \(L_1 = 1\) and \(L_2 = \frac{1}{2}\) according to [17]. For the algorithmic setup, the constants \(L^R_R\) and \(C_R\) are characterized in the following lemmas, whose proofs are lengthy and technical; they are in Appendices B and C.
Proposition 4.1 (Global Pullback Hessian Lipschitz Continuity) If $\text{Retr}(\cdot, \cdot)$ is chosen to be the polar retraction on the Stiefel manifold $\text{St}_{n,r}$ (or a products of Stiefel manifolds), then the condition \[(24)\] holds with \[L_H = 13.66G + 12.55\ell_f + 4\ell_H, \] which is independent of $x \in M$ as well as the dimensions $n, r$.

Though we only considered the local Lipschitz continuity of the pullback Hessian, we actually have a stronger global Lipschitz continuity property on the Stiefel manifold.

Proposition 4.2 For the polar retraction on the Stiefel manifold $\text{St}_{n,r}$ (or a product of Stiefel manifolds), Corollary \[2.16\] holds with constant $C_g = \frac{1}{8.02}$.

4.1 Numerical experiments

As a numerical illustration, we consider the following problem (cf. \[23\]):

\[
\max \langle A, UU^\top \rangle \text{ subject to } u_i^\top u_i = I_{d \times d}, \ i = 1, ..., n, \tag{51}
\]

where $A \in \mathbb{R}^{dn \times dn}$, $u_i \in \mathbb{R}^{k \times d}$, $U = [u_1, \ldots, u_n]^\top \in \mathbb{R}^{dn \times k}$. When $d = 1$, $u_i^\top u_i = I_{d \times d} = 1$, the problem is constrained on a product of $n$ spheres, and the problem is called the low-rank max-cut-SDP problem. When $d \geq 2$, the problem is constrained on a product of $n$ Stiefel manifolds, and the problem is called the low-rank max-orthogonal-cut-SDP problem. In the experiments, we test our algorithm in the cases where $d = 1$ and $d = 3$ respectively.

We sample the matrices $A \sim \text{GOE}(300)$, where $A \sim \text{GOE}(n)$ stands for a matrix with $A_{ij} \sim N(0, 1/n)$ when $i \neq j$ and $A_{ii} \sim N(0, 2/n)$ (see \[23\]). In the implementation of Algorithm 1, we follow \[10\] in that the subproblems are solved approximately with a fixed stepsize gradient method. However, we change the stopping criterion of the subproblem to be $\|\nabla m_{x_k, \sigma}(\xi_t)\| \leq c\|\nabla f(x_k)\|$, where $c$ is some constant. The performance of the algorithm is plotted in Figure 1 and Figure 2, where Algorithm 1 (CRRN), the Riemannian trust-region (RTR) algorithm \[7\] and the Riemannian gradient descent (RGD) algorithm are compared. For each of $d = 1$ and $d = 3$ problems, we randomly generate 3 initial solutions and show the performance of the three algorithms on these cases.

Finally, we show two examples where CRRN converges superlinearly to a local minimum point for the case $d = 1$ in Figure 3. The data is from the Caltech students Facebook social network dataset. The network consists of 597 nodes. In these two cases, we set $k = 6$ and $k = 8$ respectively.

As is observed in \[23\], the gradient descent algorithm with fixed stepsize actually works very well for this problem and finally converges to a second-order stationary point. Similar observations can be made on the general behaviors of the second-order methods. In terms of Riemannian gradient size, the CRRN runs slight slower than the RGD and RTR at first, and then quickly catches up and takes over both algorithms as the Riemannian gradient gets smaller.
Figure 1: Experiments on low-rank Max-Cut-SDP \((d = 1)\) problem

Figure 2: Experiments on low-rank Max-Orthogonal-Cut-SDP \((d = 3)\) problem

Figure 3: Superlinear convergence example on low-rank Max-Cut-SDP \((d = 1)\) problem
5 Conclusion

In this paper, we extend Nesterov’s cubic regularized Newton’s method to Riemannian optimization. It is shown that under mild conditions on the objective function and the Riemannian manifold, an $O(1/\epsilon^3)$ complexity bound can be guaranteed. This establishes that Riemannian optimization essentially bears the same degree of iteration complexity as the classical unconstrained optimization over Euclidean space. Specifically, we consider optimization over the Stiefel manifold (or a product of Stiefel manifolds). In this case, all the constants established in the general setting can be exactly computed. Our numerical experiments show that our method is competitive compared against the Riemannian gradient descent and the Riemannian trust-region method in terms of iteration complexity, although we notice that solving the subproblems using Carmon’s gradient descent method is indeed more expensive than solving the Riemannian trust-region subproblem. This motivates the study of effective schemes to solve the subproblems. To compute the required constants/parameters maybe a non-trivial task. However, once it is done, then it will be valid for the entire class of manifolds. In this paper, we carried out this computations for the case of Stiefel manifolds. It will be interesting to design an adaptive and parameter-free computational scheme, which is a topic for the future research.

References

[1] P. A. Absil, C. G. Baker, and K. A. Gallivan, *Convergence analysis of Riemannian trust-region methods*, Technical Report, (2006).

[2] ———, *Trust-region methods on Riemannian manifolds*, Foundations of Computational Mathematics, 7 (2007), pp. 303–330.

[3] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[4] P. A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM Journal on Optimization, 22 (2012), pp. 135–158.

[5] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, *Finding approximate local minima faster than gradient descent*, in Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2017, pp. 1195–1199.

[6] N. Boumal, *A Riemannian low-rank method for optimization over semidefinite matrices with block-diagonal constraints*, arXiv preprint arXiv:1506.00575, (2015).

[7] N. Boumal, P. A. Absil, and C. Cartis, *Global rates of convergence for nonconvex optimization on manifolds*, arXiv preprint arXiv:1605.08101, (2016).
[8] N. Boumal, V. Voroninski, and A. Bandeira, *The non-convex Burer-Monteiro approach works on smooth semidefinite programs*, in Advances in Neural Information Processing Systems, 2016, pp. 2757–2765.

[9] S. Burer and R. D. C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming, 95 (2003), pp. 329–357.

[10] Y. Carmon and J. C. Duchi, *Gradient descent efficiently finds the cubic-regularized non-convex Newton step*, arXiv preprint arXiv:1612.00547, (2016).

[11] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results*, Mathematical Programming, 127 (2011), pp. 245–295.

[12] ———, *Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity*, Mathematical Programming, 130 (2011), pp. 295–319.

[13] ———, *An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity*, IMA Journal of Numerical Analysis, 32 (2012), pp. 1662–1695.

[14] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, *Gradient descent can take exponential time to escape saddle points*, in Advances in Neural Information Processing Systems, 2017, pp. 1067–1077.

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

[16] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 2012.

[17] B. Jiang, S. Ma, A. M.-C. So, and S. Zhang, *Vector transport-free SVRG with general retraction for Riemannian optimization: Complexity analysis and practical implementation*, Preprint available at https://arxiv.org/abs/1705.09059, (2017).

[18] H. Kasai, H. Sato, and B. Mishra, *Riemannian stochastic variance reduced gradient on Grassmann manifold*, arXiv preprint arXiv:1605.07367, (2016).

[19] J. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. Jordan, and B. Recht, *First-order methods almost always avoid saddle points*, arXiv preprint arXiv:1710.07406, (2017).

[20] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, *Gradient descent only converges to minimizers*, in Conference on Learning Theory, 2016, pp. 1246–1257.

[21] J. M. Lee, *Introduction to smooth manifolds*, 41 (2008), p. 573.
A Proof of Lemma 2.18

For ease of notation, we denote $B_x = P_x \circ \text{Hess} f(x) \circ P_x$, where $P_x$ is the orthogonal projection onto the tangent space $T_x \mathcal{M}$. Then $\lambda_{\text{min}}^M (\text{Hess} f(x))$ corresponds to smallest eigenvalue of $B_x$ among the eigenvalues whose eigenvectors lie within $T_x \mathcal{M}$, which we denote as $\lambda_{\text{min}}^M (B_x)$ for simplicity. We now prove [26] by the following steps.

[Step 1.] To show: There exists a finite closed covering for $\mathcal{M}$ such that $\mathcal{M} \subset \bigcup_{i=1}^{n_0} U_i$. Each $U_i$ is a closed ball in $\mathcal{E}$ and there exists a local equation characterization of $\mathcal{M} \cap U_i$, i.e., $x \in \mathcal{M} \cap U_i$ if and only if $\phi^{(i)}(x) = 0, x \in U_i$.

Proof. First, for any $x \in \mathcal{M}$, there exists an open ball within which a local equation characterization is valid. Hence the union of these balls provide an open covering for $\mathcal{M}$. By the compactness of $\mathcal{M}$, there exists a finite open covering $\bigcup_{i=1}^{n_0} \bar{U}_i \supset \mathcal{M}$ with $\bar{U}_i = B(x_i, r_i)$, where $B(x, r)$ denotes an
open ball centered at \( x \) with radius \( r \). Define \( S = \partial (\cup_{i=1}^{n} \bar{U}_i) \) be the boundary of \( \cup_{i=1}^{n} \bar{U}_i \), then we have \( \varepsilon := \inf_{z \in \mathcal{M}, y \in S} \| z - y \| > 0 \). This is because both \( \mathcal{M} \) and \( S \) are compact sets and \( \| \cdot \| \) is continuous, the infimum is attained at some \( z_0 \) and \( y_0 \). However, since \( \cup_{i=1}^{n} \bar{U}_i \) is an open covering for \( \mathcal{M} \), \( z_0 \neq y_0 \), hence \( \varepsilon = \| z_0 - y_0 \| > 0 \). Therefore, if we let \( U_i = \text{cl}(B(x_i, r_i - \frac{1}{2} \varepsilon)) \) where \( \text{cl}(\cdot) \) denotes the closure of a set, then we have that \( \cup_{i=1}^{n} U_i \) is a finite closed covering for \( \mathcal{M} \). Within each \( U_i \), a local equation characterization \( \phi_i \) exists as given in the description of \( \bar{U}_i \).

**Step 2.** To show: The inequality \( (26) \) holds within each \( U_i \) with constant \( D_i \).

*Proof.* With the local equation \( \phi_i(x) = 0 \), according to the results in \( [30] \), one can write \( B_x = P_x(\nabla^2 f(x) - \sum_{j=1}^{n-d} \mu_j(x) \nabla^2 \phi_j(x))P_x \), where we have \( P_x = I - J \phi_i(x)^\top(J \phi_i(x)(J \phi_i(x)^\top)^{-1})J \phi_i(x) \), \( \mu(x) = [J \phi_i(x)J \phi_i(x)^\top]^{-1}J \phi_i(x)\nabla f(x) \). By the nonsingularity of \( J \phi_i(x) \) and compactness of \( U_i \cap \mathcal{M} \), the smoothness of \( \phi_i \) indicates that \( P_x \) and \( \mu(x) \) are all smooth function in \( x \), by the Lipschitz continuity of \( \nabla f(x) \) and \( \nabla^2 f(x) \) and the compactness of \( U_i \), we conclude that \( B_x \) is Lipschitz continuous on \( U_i \). Then there exists a \( c_i \) such that

\[
\|B_x - B_y\|_F \leq c_i \|x - y\| \quad \text{for } \forall x, y \in U_i \cap \mathcal{M}.
\]

Now we note the following well-known result on the continuity of the eigenvalues of symmetric matrices.

**Lemma A.1** (Corollary 6.3.8, in \( [10] \) on page 407) Let \( A, E \in \mathbb{R}^{n \times n} \). Assume that \( A, E \) are symmetric. Let \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \) be the eigenvalues of \( A \), and let \( \hat{\lambda}_1 \leq \hat{\lambda}_2 \leq \cdots \leq \hat{\lambda}_n \) be the eigenvalues of \( A + E \). Then we have

\[
\sum_{j=1}^{n} \| \lambda_j - \hat{\lambda}_j \|^2 \leq \| E \|_F^2.
\]

Note that the eigenvalues of \( B_x \) are \( n - d \) zeros and \( d \) eigenvalues with corresponding eigenvectors lying within \( \mathcal{T}_x \mathcal{M} \). Now consider the matrices \( B_x \) and \( B_y \) in the following three cases. First, when both \( \lambda_{\min}^M(B_x), \lambda_{\min}^M(B_y) \leq 0 \), we have \( \lambda_{\min}^M(B_x) = \lambda_1(B_x), \lambda_{\min}^M(B_y) = \lambda_1(B_y) \). Then applying Lemma A.1 we have

\[
|\lambda_{\min}^M(B_x) - \lambda_{\min}^M(B_y)| \leq \sqrt{\sum_{j=1}^{n} (\lambda_j(B_x) - \lambda_j(B_y))^2} \leq \|B_x - B_y\|_F \leq c_i \|x - y\|.
\]

Second, when both \( \lambda_{\min}^M(B_x) \geq 0 \) and \( \lambda_{\min}^M(B_y) \geq 0 \), we have \( \lambda_{\min}^M(B_x) = \lambda_{n-d+1}(B_x) \) and \( \lambda_{\min}^M(B_y) = \lambda_{n-d+1}(B_y) \). Then the same argument of the first case goes through similarly. Third, when \( \lambda_{\min}^M(B_x) \) and \( \lambda_{\min}^M(B_y) \) have different signs, e.g., \( \lambda_{\min}^M(B_x) < 0 \) while \( \lambda_{\min}^M(B_y) > 0 \), we have the following arguments,

\[
\lambda_1(B_x) = \lambda_{\min}^M(B_x) < 0, \quad \lambda_{n-d+1}(B_x) \leq 0,
\]

\[
\lambda_1(B_y) = 0, \quad \lambda_{n-d+1}(B_y) = \lambda_{\min}^M(B_y) > 0.
\]

Applying Lemma A.1 in a similar way as before we have

\[
-c_i \|x - y\| \leq \lambda_{\min}^M(B_x) < 0, \quad \text{and } 0 < \lambda_{\min}^M(B_y) \leq c_i \|x - y\| + \lambda_{n-d+1}(B_x) \leq c_1 \|x - y\|.
\]
Here we consider the extended polar retraction on the Stiefel manifold $St$. Proof of Proposition 4.1

Then this indicates that $z$.

Since $|\text{Hence } \lambda^M_{\text{min}}(B_x) - \lambda^M_{\text{min}}(B_y)| \leq 2c_1||x - y||$. Therefore if we take $D_i = 2c_i$, then the statement of Step 2 is proved.

[Step 3.] To show: There exists a constant $d_M > 0$ such that for $x, y \in M$, if $||x - y|| \leq d_M$ then there exists a $U_i$ from the finite closed cover of $M$ such that $x, y \in \text{int}(U_i)$, where $\text{int}(\cdot)$ indicates the interior of a set.

Proof. First, by our construction of the $U_i$’s, we known that $\cup_{i=1}^{n_0} \text{int}(U_i)$ is also a finite open cover of $M$. Suppose the statement is not true, then there exists a sequence $\{x_k, y_k\} \subset M$ such that the pair $x_k, y_k$ does not belong to the interior of a same $U_i$ for $\forall k, \forall i$, but $||x_k, y_k|| \to 0$. By the compactness of $M$, we have convergent subsequence $\{x_{k_i}, y_{k_i}\}$ such that both $x_{k_i} \to z_0, y_{k_i} \to z_0$. Then this indicates that $z_0 \notin \cup_{i=1}^{n_0} \text{int}(U_i)$, otherwise for sufficiently large $r$, $x_{k_i}, y_{k_i}$ shall lie in some same $\text{int}(U_i)$ and hence yields a contradiction. Therefore $z_0 \notin M$, but this also contradicts the compactness of $M$ which infers that $z_0 \in M$.

[Step 4.] Combining all the previous results and letting $D = \max_{1 \leq i \leq n_0} D_i$, the conclusion of the theorem follows. In the special case of Stiefel Manifolds, which can be characterized by a smooth and uniform equation $X^T X = I$, the finite covering arguments in Step 1 is unnecessary; $M$ itself is a valid closed covering. In that case, $d_M$ can be set to $+\infty$.

B Proof of Proposition 4.1

Here we consider the extended polar retraction on the Stiefel manifold $St_{n,r}$, written as $\text{Retr}(X, Z) = (X + Z_P)(I + Z_P^T Z_P)^{-\frac{1}{2}}$, where $Z_P := P_z[Z]$ is the orthogonal projection of $Z$ onto $T_X St_{n,r}$. However, since $Z$ and $V$ are already in the tangent space $T_X St_{n,r}$, we drop the subscript $P$ in the subsequent discussion. This extended version is applied to enable the usage of Euclidean calculus tools. For the ease of notation, we define $Y_{X,Z,V}(t) = \text{Retr}(X, Z + tV)$ and $g_{X,Z,V}(t) = f(Y_{X,Z,V}(t))$. Therefore,

$$\langle (\nabla^2 f(Z) - \nabla^2 f(0))[V], V \rangle = g''_{X,Z,V}(0) - g''_{X,0,V}(0).$$

This suggests that to prove (24) it suffices to prove

$$|g''_{X,Z,V}(0) - g''_{X,0,V}(0)| \leq L_H \|Z\|_F, \forall V \in T_X St_{n,r}, \|V\|_F = 1,$$

with an estimation of $L_H$. By direct computation,

$$g''_{X,Z,V}(0) = \langle \nabla f(Y_{X,Z,V}(0)), Y''_{X,Z,V}(0) \rangle + \langle \nabla^2 f(Y_{X,Z,V}(0))[Y'_{X,Z,V}(0)], Y_{X,Z,V}(0) \rangle.$$ 

Applying the triangular inequality yields

$$|g''_{X,Z,V}(0) - g''_{X,0,V}(0)| \leq \langle \nabla f(Y_{X,Z,V}(0)), Y''_{X,Z,V}(0) - Y''_{X,0,V}(0) \rangle$$

$$+ \langle \nabla f(Y_{X,Z,V}(0)) - \nabla f(Y_{X,0,V}(0)), Y''_{X,0,V}(0) \rangle$$

$$+ \|Y'_{X,Z,V}(0)\|_F \|\nabla^2 f(Y_{X,Z,V}(0))\|_F \|Y_{X,Z,V}(0) - Y_{X,0,V}(0)\|_F$$

$$+ \|Y'_{X,Z,V}(0)\|_F \|\nabla^2 f(Y_{X,Z,V}(0)) - \nabla^2 f(Y_{X,0,V}(0))\|_F \|Y_{X,0,V}(0)\|_F$$

$$+ \|Y'_{X,Z,V}(0) - Y'_{X,0,V}(0)\|_F \|\nabla^2 f(Y_{X,0,V}(0))\|_F \|Y_{X,0,V}(0)\|_F.$$
Recalling the nature of the parameters $G, \ell_f, \ell_H$, and Proposition 2.7, the above inequality can be simplified to

\[
|g''_{X,Z,V}(0) - g''_{X,0,V}(0)| \leq G\|Y''_{X,Z,V}(0) - Y''_{X,0,V}(0)\|_F + \ell_f L_1\|Y''_{X,0,V}(0)\|_F \|Z\|_F
\]

\[
+ \ell_f (\|Y'_{X,Z,V}(0)\|_F + \|Y'_{X,0,V}(0)\|_F)\|Y'_{X,Z,V}(0) - Y'_{X,0,V}(0)\|_F
\]

\[
+ \ell_H L_1\|Y'_{X,Z,V}(0)\|_F \|Y'_{X,0,V}(0)\|_F \|Z\|_F.
\]

Thus we need only to bound the following terms by

\[
\|Y''_{X,Z,V}(0) - Y''_{X,0,V}(0)\|_F \text{ and } \|Y'_{X,Z,V}(0) - Y'_{X,0,V}(0)\|_F \leq O(\|Z\|_F),
\]

\[
\|Y'_{X,0,V}(0)\|_F, \|Y''_{X,0,V}(0)\|_F \text{ and } \|Y'_{X,Z,V}(0)\|_F = O(1).
\]

Now we bound these terms in the following steps.

[Step 1.] First we characterize the derivatives of $Y_{X,Z,V}(t)$. Define $S_Z := I_r + Z^T Z$ and define

\[
F_{Z,V}(t) := (I_r + (Z + tV)^T (Z + tV))^{-\frac{1}{2}} = (S_Z + t(Z^T V + V^T Z + tV^T V))^{-\frac{1}{2}}.
\]

Then we have $Y_{X,Z,V}(t) = (X + Z + tV)F_{Z,V}(t)$ and $F_{Z,V}(t) = (I_r + tS^{-1}_Z(V^T Z + Z^T V + tV^T V))^{-1}S^{-1}_Z$. Let the expansion of $F_{Z,V}(t)$ be

\[
F_{Z,V}(t) = F_{Z,V}(0) + tC_{Z,V} + t^2 D_{Z,V} + O(t^3)
\]

and let us calculate the expansion of $F_{Z,V}(t)$ by

\[
F_{Z,V}(t) = [I_r - tS^{-1}_Z(V^T Z + Z^T V + tV^T V) + (tS^{-1}_Z(V^T Z + Z^T V + tV^T V))^2]S^{-1}_Z + O(t^3)
\]

where we use the formula $(I + A)^{-1} = I + \sum_{i=1}^{\infty} (-1)^i A^i$ when $\|A\| < 1$. Then, by comparing the coefficients of $t$ and $t^2$ terms between the above expansion of $F_{Z,V}(t)$ and (53), $F_{Z,V}(t)^2 = (F_{Z,V}(0) + tC_{Z,V} + t^2 D_{Z,V})^2 + O(t^3)$, we have

\[
S^{-\frac{1}{2}}_Z C_{Z,V} + C_{Z,V} S^{-\frac{1}{2}}_Z = -S^{-1}_Z(V^T Z + Z^T V)S^{-1}_Z,
\]

and

\[
S^{-\frac{1}{2}}_Z D_{Z,V} + D_{Z,V} S^{-\frac{1}{2}}_Z + C^2_{Z,V}
\]

\[
= -S^{-1}_Z(V^T V)S^{-1}_Z + S^{-1}_Z(V^T Z + Z^T V)S^{-1}_Z(V^T Z + Z^T V)S^{-1}_Z.
\]

With $C_{Z,V}$ and $D_{Z,V}$ in place, we can write the derivatives of $Y_{X,Z,V}(t)$ explicitly as

\[
\begin{cases}
Y'_{X,Z,V}(0) = (X + Z)C_{Z,V} + VS^{-\frac{1}{2}}_Z, \\
Y''_{X,Z,V}(0) = 2VC_{Z,V} + 2(X + Z)D_{Z,V}.
\end{cases}
\]

Note that when $Z = 0$ and $S_Z = I_r$, we can solve (57) and (58) to yield $S_0 = I, C_{0,V} = 0, D_{0,V} = -\frac{1}{2} V^T V$. Consequently,

\[
Y'_{X,0,V}(0) = V \text{ and } Y''_{X,0,V}(0) = -XV^T V.
\]
[Step 2.] Bound the term \( \|Y''_{X,Z,V}(0) - Y''_{X,0,V}(0)\|_F \leq \mathcal{O}(\|Z\|_F) \) for \( \forall V \in T_X\text{St}_{n,r}, \|V\|_F \leq 1 \). By (59) and (60), we have

\[
\|Y''_{X,Z,V}(0) - Y''_{X,0,V}(0)\|_F = \|2VC_{Z,V} + 2(X + Z)D_{Z,V} + XV^TV\|_F \\
\leq 2\|C_{Z,V}\|_F + 2\|DX_{Z,V} + XV^TV\|_F + 2\|D_{Z,V}\|_F\|Z\|_F.
\]

(61)

First, let us consider the term \( T_1 \). Since \( S_Z \succeq I_r \), we have \( \|C_{Z,V}\|_F \leq \|S_{Z}^{\frac{1}{2}}C_{Z,V}\|_F \). Thus we choose to bound \( \|S_{Z}^{\frac{1}{2}}C_{Z,V}\|_F \) which will be useful later. If we denote by \( \text{vec}(X) \) the vectorization operator for a matrix \( X \), then a handy formula gives \( \text{vec}(AXB) = (B^\top \otimes A)\text{vec}(X) \). Note that by using the \( \text{vec}(\cdot) \) operator, equation (67) has an explicit solution

\[
\text{vec}(C_{Z,V}) = (I_r \otimes S_Z^{\frac{1}{2}} + S_Z^{\frac{1}{2}} \otimes I_r)^{-1}\text{vec}(S_Z^{-1}(V^TZ + Z^TV)S_Z^{-1}) \equiv (I_r \otimes S_Z^{\frac{1}{2}} + S_Z^{\frac{1}{2}} \otimes I_r)^{-1}\text{vec}(V^TZ + Z^TV).
\]

Therefore

\[
\text{vec}(S_Z^{\frac{1}{2}}C_{Z,V}) = (I_r \otimes S_Z^{\frac{1}{2}})\text{vec}(C_{Z,V}) \equiv (I_r \otimes S_Z^{\frac{1}{2}} + S_Z^{\frac{1}{2}} \otimes S_Z^{\frac{1}{2}})^{-1}\left[ \text{vec}(V^TZ) + \text{vec}(Z^TV) \right].
\]

Since \( S_Z \succeq I_r, S_Z^{\frac{1}{2}} \succeq I_r \), we have \( I_r \otimes S_Z^{\frac{1}{2}} + S_Z^{\frac{1}{2}} \otimes S_Z^{\frac{1}{2}} \succeq 2I_r \), and therefore

\[
\|(I_r \otimes S_Z^{\frac{1}{2}} + S_Z^{\frac{1}{2}} \otimes S_Z^{\frac{1}{2}})^{-1}\|_2 \leq \frac{1}{2}
\]

(63)

\[
\|S_Z^{\frac{1}{2}}C_{Z,V}\|_F = \|\text{vec}(S_Z^{\frac{1}{2}}C_{Z,V})\|_F \leq \frac{1}{2}\|\text{vec}(V^TZ + Z^TV)\|_F
\]

\[
\leq \frac{1}{2}(\|V^TZ\|_F + \|Z^TV\|_F) \leq \|Z\|_F
\]

(64)

where the last inequality is due to \( \|V\|_F \leq 1 \). Hence

\[
T_1 \leq 2\|Z\|_F.
\]

(65)

Now for the benefit of discussion later, let us further bound \( \|S_Z^{\frac{1}{2}}C_{Z,V}\|_F \) by a constant. Let the SVD of \( Z \) be \( Z = QA^\frac{1}{2}U^\top \). Then \( Z^TZ = U\Lambda U^\top \), \( S_Z = U(I_r + \Lambda)U^\top \) and \( S_Z^{\frac{1}{2}} = U(I_r + \Lambda)^{\frac{1}{2}}U^\top \). Hence we have

\[
(I_r \otimes S_Z + S_Z^{\frac{1}{2}} \otimes S_Z^{\frac{1}{2}})^{-1} = (U \otimes U)(I_r \otimes (\Lambda + I_r) + (\Lambda + I_r)^{\frac{1}{2}} \otimes (\Lambda + I_r)^{\frac{1}{2}})(U \otimes U)^\top.
\]

26
Therefore, instead of bounding \( \| (I_r \otimes S_Z + S_Z^\frac{1}{2} \otimes S_Z^\frac{1}{2})^{-1}\text{vec}(Z^TV) \|_F \) in (63) by means of (63), we now take a different approach:

\[
\| (I_r \otimes S_Z + S_Z^\frac{1}{2} \otimes S_Z^\frac{1}{2})^{-1}\text{vec}(Z^TV) \|_F
\]
\[
= \| (I_r \otimes S_Z + S_Z^\frac{1}{2} \otimes S_Z^\frac{1}{2})^{-1}\text{vec}(U\Lambda^\frac{1}{2}Q^TV)\|_F
\]
\[
= \| (U \otimes U)(I_r \otimes (\Lambda + I_r) + (\Lambda + I_r)^\frac{1}{2} \otimes (\Lambda + I_r)^\frac{1}{2})^{-1}(U^T \otimes U^T)(I_r \otimes U\Lambda^\frac{1}{2})\text{vec}(Q^TV)\|_F
\]
\[
\leq \| U \otimes U \|_2 \| (I_r \otimes (\Lambda + I_r) + (\Lambda + I_r)^\frac{1}{2} \otimes (\Lambda + I_r)^\frac{1}{2})^{-1}(I_r \otimes \Lambda^\frac{1}{2}) \|_2 \| U^T \otimes I_r \|_2 \| Q \|_2 \| V \|_F.
\]

Note that \( \| U \otimes U \|_2 = \| U^T \otimes I_r \|_2 = \| Q \|_2 = 1 \), \( \| V \|_F \leq 1 \). If we denote \( \lambda_i \) as the \( i \)th diagonal element of \( \Lambda \), then all the eigenvalues of the diagonal matrix \((I_r \otimes (\Lambda + I_r) + (\Lambda + I_r)^\frac{1}{2} \otimes (\Lambda + I_r)^\frac{1}{2})^{-1}(I_r \otimes \Lambda^\frac{1}{2})\) can be written as

\[
\sigma_{ij} = \frac{\lambda_j^\frac{1}{2}}{1 + \lambda_j + (1 + \lambda_i)^\frac{1}{2}(1 + \lambda_j)^\frac{1}{2}} \leq \frac{1}{2}.
\]

Hence we end up with \( \| (I_r \otimes S_Z + S_Z^\frac{1}{2} \otimes S_Z^\frac{1}{2})^{-1}\text{vec}(Z^TV) \|_F \leq \frac{1}{2} \). Similarly, we have \( \| (I_r \otimes S_Z + S_Z^\frac{1}{2} \otimes S_Z^\frac{1}{2})^{-1}\text{vec}(V^TZ) \|_F \leq \frac{1}{2} \). Together with (62), they give an alternative bound which states \( \| S_Z^\frac{1}{2}C_{Z,V} \|_F \leq 1 \). In total,

\[
\| C_{Z,V} \|_F \leq \| S_Z^\frac{1}{2}C_{Z,V} \|_F \leq \min\{ \| Z \|_F, 1 \}.
\]

Second, we now come to bound the term \( T_2 \). Reformulating (58) slightly, we have

\[
D_{Z,V}S_Z^\frac{1}{2} + S_Z^\frac{1}{2}D_{Z,V} = -S_Z^\frac{1}{2}C_{Z,V}S_Z^\frac{1}{2} - S^{-\frac{1}{4}}_ZV^TVS^{-\frac{1}{4}}_Z + \left(S^{-\frac{1}{4}}_Z(Z^TV + V^TZ)S^{-\frac{1}{4}}_Z\right)^2.
\]

Let us define \( H_{Z,V} \) be the matrix that satisfies

\[
H_{Z,V}S_Z^\frac{1}{2} + S_Z^\frac{1}{2}H_{Z,V} = S^{-\frac{1}{2}}_ZV^TVS^{-\frac{1}{2}}_Z,
\]

and define \( J_{Z,V} = D_{Z,V} + H_{Z,V} \). Then,

\[
T_2 = \| 2D_{Z,V} + V^TV \|_F \leq 2\| D_{Z,V} + H_{Z,V} \|_F + \| 2H_{Z,V} - V^TV \|_F = 2\| J_{Z,V} \|_F + \| 2H_{Z,V} - V^TV \|_F.
\]

Note that by the definition of \( J_{Z,V} \) and \( H_{Z,V} \),

\[
J_{Z,V}S_Z^\frac{1}{2} + S_Z^\frac{1}{2}J_{Z,V} = -S_Z^\frac{1}{2}C_{Z,V}S_Z^\frac{1}{2} + \left(S^{-\frac{1}{4}}_Z(Z^TV + V^TZ)S^{-\frac{1}{4}}_Z\right)^2.
\]

Similar to the bound for \( T_1 \), we have

\[
\| J_{Z,V} \|_F = \| \text{vec}(J_{Z,V}) \|_F \leq \left\| \left(I_r \otimes S_Z^\frac{1}{2} + S_Z^\frac{1}{2} \otimes I_r\right)^{-1}\left(\| S_Z^\frac{1}{2}C_{Z,V} \|_F^2 + \| S^{-\frac{1}{4}}_Z(Z^TV + V^TZ)S^{-\frac{1}{4}}_Z \|_F^2\right)\right\|_2.
\]

(69)
Note that
\[
\| S_Z^{-\frac{1}{2}}(Z^\top V + V^\top Z)S_Z^{-\frac{1}{2}} \|_F \leq 2\| S_Z^{-\frac{1}{2}}V^\top ZS_Z^{-\frac{1}{2}} \|_F \leq 2\| S_Z^{-\frac{1}{2}} \|_2 \| ZS_Z^{-\frac{1}{2}} \|_2 \| V \|_F \leq 2\| ZS_Z^{-\frac{1}{2}} \|_2,
\]
where the last inequality is due to \(\| S_Z^{-\frac{1}{2}} \|_2, \| V \|_F \leq 1\). By the SVD of \(Z\),
\[
\| ZS_Z^{-\frac{1}{2}} \|_2 = \| Q\Lambda^{\frac{1}{2}}U(I_r + \Lambda)^{-\frac{1}{2}}U^\top \|_2 = \| \Lambda^{\frac{1}{2}}(I_r + \Lambda)^{-\frac{1}{2}} \|_2 \leq 1,
\]
and so
\[
\| S_Z^{-\frac{1}{2}}(Z^\top V + V^\top Z)S_Z^{-\frac{1}{2}} \|_F \leq 2.
\]
On the other hand,
\[
\| S_Z^{-\frac{1}{2}}(Z^\top V + V^\top Z)S_Z^{-\frac{1}{2}} \|_F \leq 2\| S_Z^{-\frac{1}{2}} \|_2 \| V \|_F \| Z \|_F \leq 2\| Z \|_F.
\]
Therefore,
\[
\| S_Z^{-\frac{1}{2}}(Z^\top V + V^\top Z)S_Z^{-\frac{1}{2}} \|_F^2 \leq \min\{4, 4\| Z \|_F\}. \tag{70}
\]
Similarly, \(\| S_Z^{-\frac{1}{2}}C_{Z,V} \|_F^2 \leq \min\{1, \| Z \|_F\}\). \tag{66}
Together with \(\| S_Z^{-\frac{1}{2}} \|_2\), the above bounds and \(\| S_Z^{-\frac{1}{2}} \|_2\), we have
\[
\| J_{Z,V} \|_F \leq \min\left\{ \frac{5}{2}, \frac{5}{2}\| Z \|_F \right\}. \tag{71}
\]
For \(\| 2H_{Z,V} - V^\top V \|_F\), let us start with the explicit solution of equation \(\| 67\), which is
\[
2\text{vec}(H_{Z,V}) = 2(S_Z^{\frac{1}{2}} \otimes S_Z + S_Z \otimes S_Z^{\frac{1}{2}})^{-1}\text{vec}(V^\top V)
\]
leading to
\[
\| H_{Z,V} \|_F \leq \frac{1}{2}. \tag{72}
\]
and
\[
\| 2\text{vec}(H_{Z,V}) - \text{vec}(V^\top V) \|_F \leq \| I_{r^2} - 2(S_Z^{\frac{1}{2}} \otimes S_Z + S_Z \otimes S_Z^{\frac{1}{2}})^{-1}\|_2\| \text{vec}(V^\top V) \|_F
\]
\[
\leq 1 - \frac{2}{\lambda_{\max}(S_Z^{\frac{1}{2}} \otimes S_Z + S_Z \otimes S_Z^{\frac{1}{2}})}
\]
\[
\leq 1 - \frac{1}{\lambda_{\max}(S_Z \otimes S_Z^{\frac{1}{2}})}
\]
\[
= 1 - \frac{1}{\lambda_{\max}^3(S_Z)}
\]
\[
\leq 1 - \frac{1}{(1 + \| Z \|_F^2)^{\frac{3}{2}}}.
\]
Let $w = \|Z\|_F$, and define the function $q(w) = 1 - (1 + w^2)^{-\frac{3}{2}}$, $w \geq 0$. Then it is easy to prove that $q(w) \leq 0.66w$ for all $w \geq 0$. Consequently, $$\|H_{Z,V} - V^TV\|_F = \|2\text{vec}(H_{Z,V}) - \text{vec}(V^TV)\|_F \leq 0.66\|Z\|_F. \tag{73}$$ Therefore, combining (69), (71) and (73) yields $$T_2 \leq 2 \cdot \frac{5}{2} \|Z\|_F + 0.66\|Z\|_F = 5.66\|Z\|_F. \tag{74}$$

Third, we bound the term $T_3$ by $$T_3 = 2\|D_{Z,V}\|_F\|Z\|_F \leq 2(\|J_{Z,V}\|_F + \|H_{Z,V}\|_F) \leq 6\|Z\|_F, \tag{75}$$ where the last inequality is due to (71) and (72). Now combining the bounds on $T_1, T_2$ and $T_3$, we finally finish the Step 2 with $$\|Y''_{X,Z,V}(0) - Y''_{X,0,V}(0)\|_F \leq T_1 + T_2 + T_3 \leq 13.66\|Z\|_F. \tag{76}$$

**Step 3.** Bound the term $\|Y'_{X,Z,V}(0) - Y'_{X,0,V}(0)\|_F$ by

$$\|Y'_{X,Z,V}(0) - Y'_{X,0,V}(0)\|_F = \|(X + Z)C_{Z,V} + VS_Z^{-\frac{1}{2}} - V\|_F$$

$$\leq \|X\|_2\|C_{Z,V}\|_F + \|Z\|_F\|C_{Z,V}\|_F + \|I_r - S_Z^{-\frac{1}{2}}\|_2\|V\|_F$$

$$\leq \|Z\|_F + \|Z\|_F + (1 - \lambda_{\text{min}}^{-\frac{1}{2}}(S_Z))$$

$$\leq 2\|Z\|_F + (1 - (1 + \|Z\|_F^2)^{-\frac{1}{2}})$$

$$\leq 2.31\|Z\|_F.$$

The second inequality is due to $\|X\|_2 = 1, \|V\|_F \leq 1, \|C_{Z,V}\|_F \leq \min\{1, \|Z\|_F\}$, the last inequality is due to $(1 - (1 + w^2)^{-\frac{3}{2}}) \leq 0.31w, \forall w \geq 0$.

**Step 4.** Bound the terms $\|Y'_{X,Z,V}(0)\|_F, \|Y'_{X,0,V}(0)\|_F$ and $\|Y''_{X,0,V}(0)\|_F$. By previous results, directly, we have $$\|Y'_{X,0,V}(0)\|_F = \|V\|_F \leq 1, \quad \|Y''_{X,0,V}(0)\|_F = \|XV^TV\|_F \leq 1,$$

and $$\|Y'_{X,Z,V}(0)\|_F = \|(X + Z)C_{Z,V} + VS_Z^{-\frac{1}{2}}\|_F \leq 2 + \|ZC_{Z,V}\|_F.$$

Again, by the SVD of $Z$ we obtain$$\|\text{vec}(ZC_{Z,V}I)\|_F = \|(I_r \otimes Z)\text{vec}(C_{Z,V})\|_F$$

$$\leq \|(I_r \otimes Z)(S_Z^{\frac{1}{2}} \otimes S_Z + S_Z \otimes S_Z^{\frac{1}{2}})^{-1}\text{vec}(V^TZ + Z^TV)\|_F.$$
Note that
\[
\|(I_r \otimes Z)(S_Z^\frac{1}{2} \otimes S_Z + S_Z \otimes S_Z^\frac{1}{2})^{-1}\text{vec}(Z^TV)\|_F
= \|(I_r \otimes QA^\frac{1}{2}U^T)(U \otimes U)(\Lambda^\frac{1}{2} \otimes \Lambda + \Lambda \otimes \Lambda^\frac{1}{2})^{-1}(U^T \otimes U^T)(I_r \otimes UA^\frac{1}{2}Q^T)\text{vec}(V)\|_F
\leq \|(U \otimes Q)\|_2\|(I_r \otimes \Lambda^\frac{1}{2})(\Lambda^\frac{1}{2} \otimes \Lambda + \Lambda \otimes \Lambda^\frac{1}{2})^{-1}(I_r \otimes \Lambda^\frac{1}{2})\|_2\|U^T \otimes Q^T\|_2
\leq \max_{i,j} \frac{\lambda_j}{1 + \lambda_j + (1 + \lambda_j)\frac{1}{2}(1 + \lambda_i)^\frac{1}{2}} \leq 1.
\]

Similarly, \||(I_r \otimes Z)(S_Z^\frac{1}{2} \otimes S_Z + S_Z \otimes S_Z^\frac{1}{2})^{-1}\text{vec}(V^TZ)\|_F \leq 1. Consequently
\[
\|Y'_{X,Z,V}(0)|_F \leq 2 + \|\text{vec}(ZC_{Z,V}I)\|_F \leq 4.
\]

Now with all the bounds for (54) and (55), and \(L_1 = 1\), we derive a bound for (53) as
\[
|g''_{X,Z,V}(0) - g''_{X,0,V}(0)| \leq L_H\|Z\|_F,
\]
where \(L_H = 13.66G + 12.55\ell_f + 4\ell_H\).

\section*{C Proof of Proposition 4.2}

Due to Proposition 2.15 we only need to bound the difference \(\|P_W - J_x\text{Retr}(X,Z)\|_2\), where \(W = \text{Retr}(X,Z)\). It suffices to show for any \(V \in \mathbb{R}^{n \times r}\) and \(\|V\|_F = 1\) that
\[
\|P_W[V] - J_x\text{Retr}(X,Z)[V]\|_F = \|P_W[V] - Y'_{X,Z,V}(0)\|_F \leq \mathcal{O}(\|Z\|_F), \forall Z \in T_{XSt_{n,r}},
\]
where \(Y'_{X,Z,V}(t)\) is used in the proof of Proposition 4.1. By (59) we have
\[
\begin{cases}
Y'_{X,0,V}(0) = V_P := P_X[V], \\
Y'_{X,Z,V}(0) = (X + Z)C_{Z,V_P} + V_PS_Z^{-\frac{1}{2}}.
\end{cases}
\]

Note that in this case \(V\) is not required to lie within \(T_{XSt_{n,r}}\), so ‘\(V\)’ in (59) should be changed to ‘\(V_P\)’. Observe,
\[
\|P_W[V] - Y'_{X,Z,V}(0)\|_F \leq \|Y'_{X,Z,V}(0) - Y'_{X,0,V}(0)\|_F + \|P_W[V] - Y'_{X,0,V}(0)\|_F
\leq 2.31\|Z\|_F + \|P_W[V] - P_X[V]\|_F,
\]
where the second inequality is due to Step 3 in the proof of Proposition 4.1. We have
\[
\|P_W[V] - P_X[V]\|_F = \frac{1}{2}\|WW^TV + W^TV - XX^TV - XV^TX\|_F
\leq \frac{1}{2}\|WW^TV - WX^TV\|_F + \frac{1}{2}\|WX^TV - XX^TV\|_F
+ \frac{1}{2}\|W^TV - WV^TX\|_F + \frac{1}{2}\|W^TV - XV^TX\|_F
\leq 2\|W - X\|_F \leq 2L_1\|Z\|_F = 2\|Z\|_F.
\]
Hence $\|P_W[V] - Y_{X,Z,V}'(0)\|_F \leq 4.31\|Z\|_F$. As long as $\|Z\|_F \leq \frac{1}{8.62}$, we have $\|P_W[V] - Y_{X,Z,V}'(0)\|_F \leq \frac{1}{2}$, $\forall X \in \text{St}_{n,r}$, $\forall Z \in T_X \text{St}_{n,r}$, $\forall \|V\| \leq 1$, and consequently

$$\|P_W - J_\xi \text{Retr}(X, Z)\|_2 \leq \frac{1}{2}.$$ 

One last remark is that with some additional effort, this threshold on $\|Z\|_F$ can be slightly improved from $1/8.62$ to $1/8$. However, we shall leave it out here for simplicity.