Superlinear and Quadratic Convergence of Riemannian Interior Point Methods

ZHJI\-JAN LAI∗ AKIKO YOSHISE†

April 17, 2022

Abstract

We extend the classical primal-dual interior point algorithms from the Euclidean setting to the Riemannian one. Our method, named the Riemannian interior point (RIP) method, is for solving Riemannian constrained optimization problems. Under the standard assumptions in the Riemannian setting, we establish locally superlinear, quadratic convergence for the Newton version of RIP and locally linear, superlinear convergence for the quasi-Newton version. These are generalizations of the classical local convergence theory of primal-dual interior point algorithms for nonlinear programming proposed by El-Bakry et al. and Yamashita et al. in 1996.

Keywords: Riemannian manifolds, Riemannian constrained optimization problem, Interior point methods, Local convergence, Riemannian Newton methods, Quasi-Newton.

1 Introduction

We consider the following problem,

\[
\min_{x \in M} \quad f(x) \\
\text{s.t.} \quad h(x) = 0, \quad \text{and} \quad g(x) \geq 0,
\]

(RCOP)

where \(M\) is a \(d\)-dimensional Riemannian manifold and \(f : M \to \mathbb{R}, h : M \to \mathbb{R}^l, l < d,\) and \(g : M \to \mathbb{R}^m\) are \(C^2\) (twice continuously differentiable) on the manifold. This problem is called the nonlinear programming problem (NLP) on a Riemannian manifold, or the Riemannian constrained optimization problem (RCOP). It appears in many applications, for instance, matrix factorization with nonnegative constraints on a fixed-rank manifold [34] and \(k\)-means via low-rank SDP as a constrained optimization problem on the Stiefel manifold [6]; for more applications, see [21, 18].

The body of knowledge on the Riemannian unconstrained optimization problem (i.e., \(h = g = 0\)), often called simply Riemannian optimization, has grown considerably in the last 20 years. In particular, well-known methods in the Euclidean setting, such as steepest descent, Newton, conjugate gradient and trust region, have been extended to the Riemannian setting [1, 16, 4, 27]. By contrast, research on the Riemannian constrained optimization problem is still in its infancy. The earliest studies go back to the optimality conditions in the Riemannian case. Yang et al. [33] extended the KKT conditions and the second-order

∗Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan. E-mail:s2130117@tsukuba.ac.jp
†Corresponding author. Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan. E-mail:yoshise@sk.tsukuba.ac.jp
necessary and sufficient conditions to (RCOP). Bergmann et al. [2] considered more constraint qualifications (CQs) on manifolds. Yamakawa et al. [31] proposed sequential optimality conditions, called approximate KKT conditions in the Riemannian case. Moreover, Liu et al. [21] were the first to develop practical algorithms. They extended the augmented Lagrangian method and exact penalty method to (RCOP). Yamakawa et al. [31] improved the augmented Lagrangian method to obtain a solution without CQs. Schiela et al. [28] and Obara et al. [25] proposed the Riemannian sequential quadratic programming method. However, to our knowledge, interior point methods are yet to be considered for (RCOP).

The advent of interior point methods in the 1980s greatly advanced the field of optimization [30, 35, 15]. By the early 1990s, the success of these methods in linear and quadratic programming ignited interest in using them on nonlinear cases [10, 32]. From the 1990s to the first decade of the 21st century, a large number of interior point methods for nonlinear programming emerged. They proved to be as successful in the nonlinear cases as in the linear ones [24, Chapter 19]. A subclass known as primal-dual interior point method is the most efficient practical approach. As described in [22], the primal-dual approach to linear programming was introduced in [23]: it was first developed as an algorithm in [20] and eventually became standard for the nonlinear case as well [10, 32]. Since it seems to be an application of the Newton method for solving the KKT conditions, it is called the Newton interior point method in some of the literature.

In this paper, we extend the classical primal-dual interior point algorithms from the Euclidean setting; i.e., $\mathbb{M} = \mathbb{R}^d$ in (RCOP), to the Riemannian setting. We call this extension the Riemannian interior point (RIP) method. Under the meaningful standard assumptions in the Riemannian setting, we establish locally superlinear, quadratic convergence of RIP with an exact Newton update and locally linear, superlinear convergence of RIP with quasi-Newton updates. These are generalizations of the classical local convergence theory of interior point methods for nonlinear programming first proposed by El-Bakry et al. [10] and Yamashita et al. [32]. To our knowledge, this paper is the first study to apply the primal-dual interior point method to optimization on Riemannian manifolds.

In this paper, we extend the classical primal-dual interior point algorithms from the Euclidean setting; i.e., $\mathbb{M} = \mathbb{R}^d$ in (RCOP), to the Riemannian setting. We call this extension the Riemannian interior point (RIP) method. Under the meaningful standard assumptions in the Riemannian setting, we establish locally superlinear, quadratic convergence of RIP with an exact Newton update and locally linear, superlinear convergence of RIP with quasi-Newton updates. These are generalizations of the classical local convergence theory of interior point methods for nonlinear programming first proposed by El-Bakry et al. [10] and Yamashita et al. [32]. To our knowledge, this paper is the first study to apply the primal-dual interior point method to optimization on Riemannian manifolds.

In Section 2, we give an interpretation of the Riemannian interior point methods; in particular, we define the concept of a KKT vector field and give the formulation of its covariant derivative. The implication of the standard assumptions motivates RIP. We end the section by describing a prototype algorithm 2. In Section 3, we describe the notation, preliminaries, and auxiliary results that we will need later. In Section 4, we prove locally superlinear and quadratic convergence of RIP with an exact Newton update. In Section 5, we prove locally linear convergence of RIP with quasi-Newton updates and in Section 6 give the conditions for its superlinear convergence. Section 7 concludes the paper.

2 Interpretation of Riemannian Interior Point Methods

We will use three symbols for the various manifolds that appear in this paper: $\mathbb{M}$ denotes the Riemannian manifold appearing in (RCOP); $\mathbb{M} := \mathbb{M} \times \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m$ is the Riemannian product manifold consisting of $\mathbb{M}$ and three Euclidean spaces, and $\mathbb{M}$ refers to a general Riemannian manifold. Following common usage in the interior-point literature, big letters denote the associated diagonal matrix by $Z = \text{diag}(z_1, \ldots, z_n)$ for $z \in \mathbb{R}^n$, and $e$ the vector of all ones whose dimension is clear in context. In the following overview of the Riemannian interior point method, many of the symbols appearing in Riemannian optimization will be used without giving their definitions. The readers may refer to Section 3.1 for the detailed definitions.
2.1 KKT Vector Field

The Lagrangian function of (RCOP) is
\[
L(x, y, z) = f(x) - y^T h(x) - z^T g(x),
\]
where \( y \in \mathbb{R}^l \) and \( z \in \mathbb{R}^m \) are Lagrange multipliers. With respect to the variable \( x \), \( L(\cdot, y, z) \) is a real function on \( M \), and its Riemannian gradient is
\[
\nabla_x L(x, y, z) = \nabla f(x) - \sum_{i=1}^l y_i \nabla h_i(x) - \sum_{i=1}^m z_i \nabla g_i(x),
\]
where \( \nabla f(x), \{\nabla h_i(x)\}, \{\nabla g_i(x)\} \) are the Riemannian gradients for the respective component functions of \( f, h, g \). The active set at \( x \in M \), \( A(x) = \{i : g_i(x) = 0, i = 1, \ldots, m\} \) consists of indices of the active constraints at \( x \). The Riemannian versions of the KKT conditions [21, Definition 2.3] for (RCOP) are given by
\[
\begin{cases}
\nabla_x L(x, y, z) = 0_x, \\
h(x) = 0, \\
g(x) \geq 0, \\
Z g(x) = 0, \\
z \geq 0.
\end{cases}
\]

With slack variables \( s := g(x) \), the above KKT conditions can be written as\(^1\)
\[
F(w) := \begin{pmatrix}
\nabla_x L(x, y, z) \\
h(x) \\
g(x) - s \\
Z Se
\end{pmatrix} = 0 := \begin{pmatrix}
0_x \\
0 \\
0 \\
0
\end{pmatrix},
\]
and \( (s, z) \geq 0 \), where \( w := (x, y, s, z) \in M \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m \). Note that for the Riemannian KKT conditions, we generate a vector field \( F \) on the Riemannian product manifold \( M \), i.e.,
\[
F : M \to T_M \equiv TM \times T\mathbb{R}^l \times T\mathbb{R}^m \times T\mathbb{R}^m,
\]
where \( T_M := \bigsqcup_{w \in M} T_w \) denotes the tangent bundle of \( M \). At a point \( w = (x, y, s, z) \in M \), the tangent space appears as
\[
T_w M = T_x M \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m
\]
under the identifications \( T_v E = E \) for any vector space \( E \) and any \( v \in E \). In particular, for any \( w = (x, y, s, z) \) with \( x \in M \), the first component of \( F(w) \) is in \( T_x M \); the Lagrange multipliers \( y, z \) and slack variables \( s \), in turn, are treated as they usually are.

**Definition 2.1.** Let \( M = \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m \). The vector field \( F \) on \( M \) defined in (2) is called the KKT vector field of (RCOP).

Thus, the KKT conditions for (RCOP) can be interpreted as ones for finding a singularity of a vector field on a product Riemannian manifold with partial nonnegative requirements, namely,
\[
F(w) = 0 \quad \text{and} \quad (s, z) \geq 0.
\]
\(^1\)The brackets denote vertically aligned elements. For example, \( \begin{pmatrix} x \\ y \end{pmatrix} = (x, y) \) in some product space \( X \times Y \).
2.2 Implication of Standard Assumptions

Let \( \mathcal{X}(\mathcal{M}) \) denote the set of all \( C^1 \) (continuously differentiable) vector fields on \( \mathcal{M} \). The Newton method is a powerful tool for finding the zeros of nonlinear functions in the Euclidean setting. The generalized Newton method has been studied in the Riemannian setting; it aims to find the singularity of a vector field \( F \in \mathcal{X}(\mathcal{M}) \), specifically, a point \( p \in \mathcal{M} \) such that,

\[
F(p) = 0.
\]  

(4)

Just as vector fields correspond to nonlinear functions from and to \( \mathbb{R}^n \), covariant derivatives of vector fields correspond to the usual Jacobian matrices. Let \( \nabla \) be the Riemannian connection on \( \mathcal{M} \). The covariant derivative of \( F \in \mathcal{X}(\mathcal{M}) \) assigns each point \( p \in \mathcal{M} \) a linear operator \( \nabla F(p) \) on \( T_p\mathcal{M} \), defined as

\[
\nabla F(p) : T_p\mathcal{M} \to T_p\mathcal{M}, \nabla F(p) \xi := \nabla_Y F(p),
\]

where \( Y \in \mathcal{X}(\mathcal{M}) \) such that \( Y(p) = \xi \). Then, (standard) Riemannian Newton iteration for (4) can be performed as follows.

**Algorithm 1 (Standard Riemannian Newton Method).**

(Step 1) For \( p_k \in \mathcal{M} \), compute \( v_k \in T_{p_k}\mathcal{M} \) as a solution of the linear system on \( T_{p_k}\mathcal{M} \) by

\[
\nabla F(p_k)v_k = -F(p_k).
\]

(Step 2) Let \( p_{k+1} := R_{p_k}(v_k) \), where \( R \) denotes a retraction on \( \mathcal{M} \). Return to Step 1.

We know that, if \( p^* \) is a solution of (4) and the linear operator \( \nabla F(p^*) \) is nonsingular, then local superlinear and quadratic convergence holds under certain mild conditions on the map \( p \to \nabla F(p) \) [12]. Therefore, the requirement of nonsingularity of the covariant derivative at the solution is essential if the Newton method is to be applied to (2).

For our purpose, we must formulate the covariant derivative of KKT vector field \( F \) at arbitrary \( w \in \mathcal{M} \). Let \( \text{Hess}_x \mathcal{L}(w) \) be the Riemannian Hessian of the real function \( \mathcal{L}(\cdot, y, z) \). It is a linear operator on \( T_x\mathcal{M} \) such that, for any \( \Delta x \in T_x\mathcal{M} \),

\[
\text{Hess}_x \mathcal{L}(w)\Delta x = \text{Hess} f(x)\Delta x - \sum_{i=1}^l y_i \text{Hess} h_i(x)\Delta x - \sum_{i=1}^m z_i \text{Hess} g_i(x)\Delta x,
\]

where \( \text{Hess} f(x) \), \{\( \text{Hess} h_i(x) \)\}, \{\( \text{Hess} g_i(x) \)\} are Riemannian Hessians of the component functions. By using the notation \( \text{Hess}_x \mathcal{L}(w) \), we have the following.

**Lemma 2.2** (covariant derivative of KKT vector field). Given any \( w \in \mathcal{M} \), for the KKT vector field \( F \) defined in (2), the linear operator \( \nabla F(w) : T_w\mathcal{M} \to T_{w}\mathcal{M} \) is given by

\[
\nabla F(w)\Delta w = \begin{pmatrix}
\text{Hess}_x \mathcal{L}(w)\Delta x - \sum_{i=1}^l y_i \text{grad} h_i(x) - \sum_{i=1}^m z_i \text{grad} g_i(x) \\
\langle \text{grad} h_i(x), \Delta x \rangle_x, \text{ for } i = 1, \ldots, l \\
\langle \text{grad} g_i(x), \Delta x \rangle_x - \Delta s_i, \text{ for } i = 1, \ldots, m \\
Z\Delta z + S\Delta \zeta
\end{pmatrix}
\]

(5)

where \( \Delta w = (\Delta x, \Delta y, \Delta s, \Delta z) \in T_x\mathcal{M} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m \). Moreover, its adjoint is given by

\[
\nabla F(w)^*\Delta w = \begin{pmatrix}
\text{Hess}_x \mathcal{L}(w)^*\Delta x + \sum_{i=1}^l y_i \text{grad} h_i(x) + \sum_{i=1}^m z_i \text{grad} g_i(x) \\
-\langle \text{grad} h_i(x), \Delta x \rangle_x, \text{ for } i = 1, \ldots, l \\
Z\Delta z - \Delta s \\
s_i \Delta z_i - \langle \text{grad} g_i(x), \Delta x \rangle_x, \text{ for } i = 1, \ldots, m
\end{pmatrix}
\]
Proof. See Appendix A for a rigorous proof. \[\square\]

**Remark 2.3.** If \(\mathcal{M}\) is a Riemannian submanifold of \(\mathbb{R}^n\) equipped with the inherited metric \(\langle \cdot, \cdot \rangle\), we can express the above more easily. For example, for \(h_i\) with \(i = 1, \ldots, l\),
\[
\langle \text{grad } h_i(x), \Delta x \rangle_x = \langle \text{Proj}_x \text{egrad } h_i(x), \Delta x \rangle = \langle \text{egrad } h_i(x), \text{Proj}_x \Delta x \rangle = \langle \text{egrad } h_i(x), \Delta x \rangle,
\]
where \(\text{Proj}_x\) denotes the orthogonal projector onto \(T_x \mathcal{M} \subseteq \mathbb{R}^n\). Hence, the second and third rows of (5) only need the Euclidean gradients \(\text{egrad } h_i(x)\), \(\text{egrad } g_i(x)\).

A prior study on Riemannian optimal conditions [33, 2] showed that the following assumptions for (RCOP) are meaningful. The corresponding Euclidean versions can be found in [10, Section 4].

**Assumption 1** (standard Riemannian assumptions of (RCOP)).

(A1) **Existence.** There exists \((x^*, y^*, z^*)\), solution to (RCOP) and associated multipliers, satisfying the KKT conditions (1).

(A2) **Smoothness of \(C^2\).** The functions \(f, g, h\) are \(C^2\) on \(\mathcal{M}\).

(A3) **Regularity.** The set \(\{\text{grad } h_i(x^*) : i = 1, \ldots, l\} \cup \{\text{grad } g_i(x^*) : i \in \mathcal{A}(x^*)\}\) is linearly independent in \(T_{x^*} \mathcal{M}\).

(A4) **Strict complementarity.** \((z^*)_i > 0\) if \(g_i(x^*) = 0\) for all \(i = 1, \ldots, m\).

(A5) **Second order sufficiency.** \(\langle \text{Hess}_x \mathcal{L}(w^*) \xi, \xi \rangle > 0\) for all nonzero \(\xi \in T_{x^*} \mathcal{M}\) satisfying \(\langle \xi, \text{grad } h_i(x^*) \rangle = 0\) for \(i = 1, \ldots, l\), and \(\langle \xi, \text{grad } g_i(x^*) \rangle = 0\) for \(i \in \mathcal{A}(x^*)\).

The following result motivates the use of the Newton method for solving (2). Proposition 2.4 is a generalization of [10, Proposition 4.1].

**Proposition 2.4.** Let assumptions (A1)-(A5) hold at \(w^*\). Then the operator \(\nabla F(w^*)\) in (5) is nonsingular.

Proof. This proof omits all the asterisks of \(w^*\). Let \(\mathcal{E} = \{1, \ldots, l\}, \mathcal{I} = \{1, \ldots, m\}\) and \(\mathcal{A} = \mathcal{A}(x) \subseteq \mathcal{J}\). Take a triple \((x, y, z)\) satisfying (A1)-(A5). Set \(s_i := g_i(x)\) for \(i \in \mathcal{I}\), and define \(w = (x, y, s, z)\). Suppose that \(\nabla F(w) \Delta w = 0\) for some \(\Delta w \in T_x \mathcal{M}\). We will show that \(\Delta w = (\Delta x, \Delta y, \Delta s, \Delta z) = 0\) to prove its nonsingularity.

Expanding \(\nabla F(w) \Delta w = 0\) gives
\[
\begin{aligned}
0 &= \text{Hess}_x \mathcal{L}(w) \Delta x - \sum_{i \in \mathcal{E}} \Delta y_i \text{grad } h_i(x) - \sum_{i \in \mathcal{I}} \Delta z_i \text{grad } g_i(x), \\
0 &= \langle \text{grad } h_i(x), \Delta x \rangle, \text{ for all } i \in \mathcal{E}, \\
0 &= \langle \text{grad } g_i(x), \Delta x \rangle - s_i \Delta s_i, \text{ for all } i \in \mathcal{I}, \\
0 &= z_i \Delta s_i + s_i \Delta z_i, \text{ for all } i \in \mathcal{I}.
\end{aligned}
\]

Strict complementarity (A4) and the last equalities above imply that \(\Delta s_i = 0\) for \(i \in \mathcal{A}\) and \(\Delta z_i = 0\) for \(i \in \mathcal{J} \setminus \mathcal{A}\). Substituting those values into the system (6) reduces it to
\[
\begin{aligned}
0 &= \text{Hess}_x \mathcal{L}(w) \Delta x - \sum_{i \in \mathcal{E}} \Delta y_i \text{grad } h_i(x) - \sum_{i \in \mathcal{A}} \Delta z_i \text{grad } g_i(x), \\
0 &= \langle \text{grad } h_i(x), \Delta x \rangle, \text{ for all } i \in \mathcal{E}, \\
0 &= \langle \text{grad } g_i(x), \Delta x \rangle, \text{ for all } i \in \mathcal{A}.
\end{aligned}
\]

\[\text{Here, we do not assume Lipschitz continuity of their Hessians for the time being. Lipschitz continuity becomes more complicated in the Riemannian case, so we will leave its description for later.}\]
and
\[ \Delta s_i = \langle \text{grad} g_i(x), \Delta x \rangle \text{ for all } i \in \mathcal{J}\backslash \mathcal{A}. \]  
(8)

If \( \Delta x \neq 0 \), it follows from system (7) that
\[
0 = \langle \text{Hess}_x \mathcal{L}(w) \Delta x - \sum_{i \in \mathcal{E}} \Delta y_i \text{grad} h_i(x) - \sum_{i \in \mathcal{A}} \Delta z_i \text{grad} g_i(x), \Delta x \rangle
\]
\[
= \langle \text{Hess}_x \mathcal{L}(w) \Delta x, \Delta x \rangle - \sum_{i \in \mathcal{E}} \Delta y_i \langle \text{grad} h_i(x), \Delta x \rangle - \sum_{i \in \mathcal{A}} \Delta z_i \langle \text{grad} g_i(x), \Delta x \rangle
\]
\[
= \langle \text{Hess}_x \mathcal{L}(w) \Delta x, \Delta x \rangle,
\]
which is a contradiction by second-order sufficiency (A5). Thus, \( \Delta x \) must be zero, and by (8), \( \Delta s_i = 0 \) for all \( i \in \mathcal{J}\backslash \mathcal{A} \).

Next, substituting \( \Delta x = 0 \) into the first equation in (7) yields
\[
0 = \sum_{i \in \mathcal{E}} \Delta y_i \text{grad} h_i(x) + \sum_{i \in \mathcal{A}} \Delta z_i \text{grad} g_i(x). \]  
(9)

The linear independence of the gradients in \( T_x \mathcal{M} \) of (A3) implies that \( \Delta y \) and \( \Delta z_i \) for \( i \in \mathcal{A} \) must be zero. This completes the proof. \( \square \)

### 2.3 Prototype Algorithm

If we directly apply the Newton method to the KKT vector field, the Newton equation for (2) is
\[
\nabla F(w) \Delta w + F(w) = 0. \]  
(10)

As in the usual Euclidean setting, once the iterates reach the boundary of the feasible region, they are forced to stick to it. To keep the iterates sufficiently far from the boundary, we introduce a perturbed complementary equation for some positive number \( \mu > 0 \) and define
\[
F(w; \mu) := F(w) - \mu \hat{e}, \text{ and } \hat{e} := \hat{e}(w) := (0_x, 0, 0, e),
\]  
(11)

with the zero element \( 0_x \) of \( T_x \mathcal{M} \). Notice that the perturbation term \( \hat{e} \), indeed, is a special vector field on \( \mathcal{M} \), not a constant, because \( 0_x \) is essentially dependent on \( w \) or \( x \).

**Definition 2.5.** For some parameter \( \mu > 0 \), the vector field \( F(w; \mu) \) defined in (11) is called the **perturbed KKT vector field** of (RCOP).

Note that the covariant derivative of the perturbed KKT vector field is the same as that of the original. At any point \( w \in \mathcal{M} \), we have
\[
\nabla F(w; \mu) = \nabla F(w) - \mu \nabla \hat{e}(w) = \nabla F(w),
\]  
(12)

since \( \nabla \hat{e}(w) \Delta w = (0_x, 0, 0, 0) \) for any \( \Delta w \in T_w \mathcal{M} \). Applying the Newton method to
\[
F(w; \mu) = 0
\]
yields the perturbed Newton equation, \( \nabla F(w; \mu) \Delta w + F(w; \mu) = 0 \). From (11) and (12), this equation is equivalent to
\[
\nabla F(w) \Delta w + F(w) = \mu \hat{e},
\]
which reduces to the ordinary Newton equation (10) when \( \mu = 0 \).

On the other hand, the next lemma indicates a homotopy (or, continuation) derivation as in the case of the Euclidean interior point method [24, Chapter 19]. Very recently, Séguin et al. [29] developed continuation methods for Riemannian optimization, which are closely related to our Riemannian interior point method.
Lemma 2.6. Under the standard assumptions (A1)-(A5) at $w^*$, there exist a sufficiently small $\bar{\mu} > 0$ and a smooth curve $w: [0, \bar{\mu}) \to \mathcal{M}$ such that $w(0) = w^*$ and

$$F(w(\mu); \mu) = 0, \quad \forall \mu \in [0, \bar{\mu}).$$  

(13)

Proof. By Proposition 2.4, we have $F(w^*; 0) = 0$ and $\nabla F(w^*; 0)$ is nonsingular. The proof of this lemma uses the same technique as in [29, Theorem 3.1]. Roughly speaking, it applies the implicit function theorem to the local coordinate representations of the vector field $F$ and its full rank Jacobian matrix at the solution. \hfill \Box

This smooth curve $\mu \mapsto w(\mu)$ is called the central path, whose endpoint $w(0) = w^*$ is a solution satisfying the KKT conditions. $\mu$ is customarily called the barrier parameter because (13) can be interpreted as the Riemannian KKT conditions of the following barrier problem:

$$\min_{(x, s) \in \mathbb{M} \times \mathbb{R}^m} f(x) - \mu \sum_{i=1}^{m} \log s_i$$

s.t. $h(x) = 0$, and $g(x) - s = 0.$

Now, we describe the prototype algorithm of the Riemannian interior point (RIP) method and its quasi-Newton version.

Algorithm 2 (Prototype algorithm of RIP).

(Step 0) Let $R$ be a retraction on $\mathbb{M}$. Given an initial point $w_0 = (x_0, y_0, z_0, s_0) \in \mathcal{M}$ with $(s_0, z_0) > 0$, for $k = 0, 1, 2, \ldots$, do:

(Step 1) Choose the barrier parameter $\mu_k > 0$.

(Step 2) Solve the following linear system for $\Delta w_k = (\Delta x_k, \Delta y_k, \Delta s_k, \Delta z_k)$,

$$\nabla F(w_k)\Delta w_k = -F(w_k) + \mu_k \hat{e}. \quad (14)$$

(Step 3) Choose $\gamma_k$ with $0 < \gamma_k \leq 1$ for some constant $\gamma$ and compute the step size,

$$\alpha_k = \min \left\{1, \gamma_k \min_i \left\{ -\frac{(s_k)_i}{(\Delta s_k)_i} \mid (\Delta s_k)_i < 0 \right\}, \gamma_k \min_i \left\{ -\frac{(z_k)_i}{(\Delta z_k)_i} \mid (\Delta z_k)_i < 0 \right\} \right\}. \quad (15)$$

(Step 4) Update: $w_{k+1} = R_{w_k}(\alpha_k \Delta w_k)$, or, $(x_{k+1}, y_{k+1}, s_{k+1}, z_{k+1}) = (R_{x_k}(\alpha_k \Delta x_k), y_k + \alpha_k \Delta y_k, s_k + \alpha_k \Delta s_k, z_k + \alpha_k \Delta z_k)$. Return to Step 1.

When $\nabla F(w_k)$ is ill-conditioned, we may want to modify it. Consider an iterate $w_k$ and a linear operator $G_k : T_{x_k} \mathbb{M} \to T_{x_k} \mathcal{M}$. We can define an approximate operator of $\nabla F(w_k)$, denoted by $B_k$, by replacing $\text{Hess}_x \mathcal{L}(w_k)$ in the first row of (5) with $G_k$ and keeping everything else the same:

$$B_k \Delta w = \begin{pmatrix}
G_k \Delta x - \sum_{i=1}^{l} \Delta y_i \text{grad} h_i(x_k) - \sum_{i=1}^{m} \Delta z_i \text{grad} g_i(x_k) \\
\langle \text{grad} h_i(x_k), \Delta x \rangle, \text{ for } i = 1, \ldots, l \\
\langle \text{grad} g_i(x_k), \Delta x \rangle - \Delta s_i, \text{ for } i = 1, \ldots, m \\
Z_k \Delta s + S_k \Delta z
\end{pmatrix} \quad (16)$$

for any $\Delta w = (\Delta x, \Delta y, \Delta s, \Delta z) \in T_{w_k} \mathbb{M}$. $G_k$ should of course be chosen appropriately to approximate to $\text{Hess}_x \mathcal{L}(w_k)$. We obtain quasi-Newton RIP by replacing (14) in (Step 2) with

$$B_k \Delta w_k = -F(w_k) + \mu_k \hat{e}. \quad (17)$$

We conclude this section with the following lemma, which shows the relationship between the parameter $\gamma_k$ and step size $\alpha_k$ in (Step 3). To simplify the problem, we will consider only a simple step-size selection (15).
Lemma 2.7. Let the assumptions (A1) and strict complementarity (A4) hold. Suppose that the step size $\alpha_k$ is as in Algorithm 2, i.e., (15). Define a constant,

$$\Omega := 2 \max \left\{ \max_i \left\{ \frac{1}{(s^*)_i} \right| (s^*)_i > 0 \right\}, \max_i \left\{ \frac{1}{(z^*)_i} \right| (z^*)_i > 0 \right\}.$$

For $\gamma_k \in (0, 1)$, if

$$\Omega \|\Delta w_k\| \leq \gamma_k,$$

then

$$0 \leq 1 - \alpha_k \leq (1 - \gamma_k) + \Omega \|\Delta w_k\|.$$

Proof. Note that the Riemannian Newton iteration in (14) for the modified complementarity condition yields

$$S_k^{-1} \Delta s_k + Z_k^{-1} \Delta z_k = \mu_k (S_k Z_k)^{-1} e - e,$$

which is exactly the same as in the usual interior point method in the Euclidean setting. Thus, the proof entails directly applying the lemmas [32, Lemma 3 and 4] for the Euclidean case to the Riemannian case. \qed

Remark 2.8. Rule (15) uses a single step size for the variables. Another popular rule, also mentioned in [32, (3.15)], uses different step sizes as follows: let $w_{k+1} = (R_{x_k} (\alpha_{x_k} \Delta x_k), y_k + \alpha_{y_k} \Delta y_k, s_k + \alpha_{s_k} \Delta s_k, z_k + \alpha_{z_k} \Delta z_k)$, where

$$\alpha_{s_k} = \min \left\{ 1, \gamma_k \min_i \left\{ \frac{(s_k)_i}{(\Delta s_k)_i} \right| (\Delta s_k)_i < 0 \right\} \right\},$$

$$\alpha_{z_k} = \min \left\{ 1, \gamma_k \min_i \left\{ \frac{(z_k)_i}{(\Delta z_k)_i} \right| (\Delta z_k)_i < 0 \right\} \right\},$$

and $\alpha_{y_k} = 1$ or $\alpha_{s_k}$ or $\alpha_{z_k}$. For this step-size rule, we can obtain a similar result as in Lemma 2.7. Refer to [32, Lemma 5] for details.

3 Preliminaries and Auxiliary Results

In this section, we describe some of the preliminary and auxiliary results necessary for understanding Riemannian optimization on the basis of the literature [19, 17]. All of the symbols are defined in Section 3.1.

3.1 Notations

Let $E$ and $E'$ be two $d$-dimensional vector spaces with inner products $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle'$. Let $A : E \to E'$ be a linear operator. The adjoint of $A$ is the linear operator $A^* : E' \to E$ defined by the property: for all $v \in E, u \in E'$, $\langle A(v), u \rangle = \langle v, A^*(u) \rangle$. The norm of $A$ is defined by $\|A\| := \sup \{ \|Av\| : v \in E, \|v\| = 1 \}$. Given two nonnegative infinite sequences of reals $\{u_k\}$ and $\{v_k\}$, we write $u_k = O(v_k)$ if there is a positive constant $M$ such that $u_k \leq Mv_k$ for all sufficiently large $k$. We write $u_k = o(v_k)$ if $v_k > 0$ and the sequence of ratios $\{u_k/v_k\}$ approaches zero.
Riemannian metrics and gradients  Let $\mathcal{M}$ be a finite-dimensional manifold. The tangent space $T_x\mathcal{M}$ at a point $x$ of a manifold $\mathcal{M}$ is the set of tangent vectors of all the curves at $x$. The tangent bundle is $T\mathcal{M} := \{(x,v) : x \in \mathcal{M}, v \in T_x\mathcal{M}\}$. A vector field on a manifold $\mathcal{M}$ is a map $V : \mathcal{M} \to T\mathcal{M}$ such that $V(x)$ is in $T_x\mathcal{M}$ for all $x \in \mathcal{M}$. The set of all $C^1$ vector fields on $\mathcal{M}$ is denoted by $\mathfrak{X}(\mathcal{M})$. The inner product on $T_x\mathcal{M}$ is a bilinear, symmetric, positive definite function $\langle \cdot , \cdot \rangle_x$. It induces the norm $\|v\|_x := \sqrt{\langle v,v \rangle_x}$. A Riemannian metric on $\mathcal{M}$ is a choice of inner product $\langle \cdot , \cdot \rangle_x$ for each $x \in \mathcal{M}$ and it varies smoothly with $x$. We will often omit the subscript if it is clear. A manifold with a Riemannian metric is called a Riemannian manifold. Let $\mathfrak{F}(\mathcal{M})$ denote the set of all smooth real functions $f : \mathcal{M} \to \mathbb{R}$. The Riemannian gradient of $f$ is the vector field, $\text{grad} f$, uniquely defined by the identities: $\nabla f(x)[v] = \langle v, \text{grad} f(x) \rangle$ for all $(x,v) \in T\mathcal{M}$, where $\nabla f : T_x\mathcal{M} \to \mathbb{R}$ is the differential of $f$ at $x$.

Connections, covariant derivatives and Hessians  An (affine) connection on $\mathcal{M}$ is an operator

$$\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M}) : (U,V) \mapsto \nabla_U V$$

which satisfies the following three properties for arbitrary $U,V,W \in \mathfrak{X}(\mathcal{M})$, $f,g \in \mathfrak{F}(\mathcal{M})$ and $a,b \in \mathbb{R}$:

1. $(\mathfrak{F}(\mathcal{M})$-linearity in $U$): $\nabla_U fV = f\nabla_U V + g\nabla_W V$;
2. $\mathbb{R}$-linearity in $V$: $\nabla_U (aV + bW) = a\nabla_U V + b\nabla_W V$; and
3. The Leibniz rule: $\nabla_U (fV) = (Uf)V + f\nabla_U V$.

The vector field $\nabla_U V$ at $x$ depends on $U$ only through $U(x)$. Thus, we can write $\nabla_v$ to mean $\nabla_U (V(x))$ for arbitrary $U \in \mathfrak{X}(\mathcal{M})$ such that $U(x) = v$. On a Riemannian manifold $\mathcal{M}$, there exists a unique connection $\nabla$ which satisfies two additional properties for all $U,V,W \in \mathfrak{X}(\mathcal{M})$: symmetry: $[U,V] = \nabla_U V - \nabla_V U$; and compatibility with the metric: $U\langle V,W \rangle = \langle \nabla_U V,W \rangle + \langle V,\nabla_U W \rangle$.

This connection is called the Levi-Civita or Riemannian connection. Throughout the paper, $\nabla$ stands for the Riemannian connection.

The covariant derivative of $F \in \mathfrak{X}(\mathcal{M})$ determined by $\nabla$ defines a linear operator at each $p \in \mathcal{M}$: $\nabla F(p) : T_p\mathcal{M} \to T_p\mathcal{M}$, by $\nabla F(p)v := \nabla_v F$. Particularly, the Riemannian Hessian of $f \in \mathfrak{F}(\mathcal{M})$ at $x \in \mathcal{M}$ is a linear operator, $\text{Hess} f(x) : T_x\mathcal{M} \to T_x\mathcal{M}$, defined as $\text{Hess} f(x)u := \nabla_u \text{grad} f$.

Retraction and vector transport  A retraction $R$ on a manifold $\mathcal{M}$ is a smooth map from the tangent bundle $T\mathcal{M}$ to $\mathcal{M}$ with the following properties, where $R_x$ denotes the restriction of $R$ to $T_x\mathcal{M}$:

1. $R_x(0_x) = x$, where $0_x$ denotes the zero element of $T_x\mathcal{M}$.
2. With the canonical identification $T_0 T_x\mathcal{M} = T_x\mathcal{M}$, $R_x$ satisfies $\mathcal{D} R_x (0_x) = \text{id}_{T_x\mathcal{M}}$, where $\mathcal{D} R_x (0_x)$ denotes the differential of $R_x$ at $0_x$ and $\text{id}_{T_x\mathcal{M}}$ denotes the identity map on $T_x\mathcal{M}$.

A vector transport $\mathcal{T}$ on a manifold $\mathcal{M}$ is a smooth map

$$T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M} : (\eta, \xi) \mapsto \mathcal{T}_{\eta}(\xi) \in T\mathcal{M}$$

satisfying the following three properties for all $x \in \mathcal{M}$, where $\oplus$ denotes the Whitney sum $T\mathcal{M} \oplus T\mathcal{M} := \{(\xi_x, \eta_x) : \xi_x, \eta_x \in T_x\mathcal{M}, x \in \mathcal{M}\}$:

1. (Associated retraction) There exists an associated retraction $R$ such that $\mathcal{T}_{\eta_x}(\xi_x) \in T_{R_x(\eta_x)}\mathcal{M}$ for all $\eta_x, \xi_x \in T_x\mathcal{M}$.
2. (Consistency) $\mathcal{T}_{0_x}(\xi_x) = \xi_x$ for all $\xi_x \in T_x\mathcal{M}$.
3. (Linearity) $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x}(\xi_x) + b\mathcal{T}_{\eta_x}(\zeta_x)$ for all $a, b \in \mathbb{R}$ and $\eta_x, \xi_x, \zeta_x \in T_x\mathcal{M}$. 


Additionally, \( T \) is isometric if the following equation holds, for all \( x \in \mathcal{M} \) and all \( \xi_x, \zeta_x, \eta_x \in T_x\mathcal{M} \):

\[
\langle T_{\eta_x} \xi_x, T_{\eta_x} \zeta_x \rangle = \langle \xi_x, \zeta_x \rangle.
\]

In other words, the adjoint and the inverse coincide. For the submanifolds \( \mathcal{M} \) of a linear space \( \mathcal{E} \), which are mostly what we consider in practice, there are many ways to construct an isometric vector transport \cite[Section 2.3]{19}. \textit{Throughout this paper, we will suppose an isometric vector transport} \( T \). Isometry plays an irreplaceable role in the proofs about quasi-Newton RIP.

**Distance, geodesics and completeness** Given a piecewise smooth curve segment \( \gamma : [a, b] \to \mathcal{M} \), we define the length of \( \gamma \) as \( L(\gamma) := \int_a^b \| \gamma'(t) \|_{\gamma(t)} \, dt \). The Riemannian distance between \( x \) and \( y \), denoted by \( d(x, y) \), is defined to be the infimum of \( L(\gamma) \) over all piecewise smooth curve segments \( \gamma \) that connect \( x \) to \( y \). Riemannian distance induces the original topology on \( \mathcal{M} \); namely, \( (\mathcal{M}, d) \) is a metric space. The open ball of radius \( r > 0 \) centered at \( p \) is defined as \( B_r(p) := \{ q \in \mathcal{M} : d(p, q) < r \} \).

A vector field \( V \) along a smooth curve \( \gamma \) on \( \mathcal{M} \) is said to be parallel, if and only if \( \nabla_\gamma V = 0 \). If \( \gamma' \) itself is parallel we say that \( \gamma \) is a geodesic. A geodesic segment joining \( p \) to \( q \) in \( \mathcal{M} \) is said to be minimal if its length is equal to \( d(p, q) \).

A Riemannian manifold is complete if the geodesics are defined for all values of \( t \in \mathbb{R} \). Hopf-Rinow’s theorem asserts that every pair of points in a complete Riemannian manifold \( \mathcal{M} \) can be joined by a (not necessarily unique) minimal geodesic segment. If a unique geodesic segment joining \( p \) to \( q \) exists, we denote it by \( \gamma_{pq} \). We will often omit the subscript if it is clear. Throughout the paper, \( \mathcal{M} \) stands for a complete Riemannian manifold.

**Exponential mapping and parallel transport** The exponential map and parallel transport are special examples of retraction and vector transport, respectively. They are perfect in theory, but often computationally intractable.

For every \( \xi \in T_x\mathcal{M} \), there exists an interval \( I \) around zero and a unique geodesic \( \gamma(t) : I \to \mathcal{M} \) such that \( \gamma(0) = x \) and \( \gamma'(0) = \xi \). The map \( \exp_x : T_x\mathcal{M} \to \mathcal{M} : \xi \mapsto \exp_x \xi = \gamma(1) \) is called the exponential map at \( x \). The exponential map is a retraction. The domain of \( \exp_x \) is the whole \( T_x\mathcal{M} \) for all \( x \in \mathcal{M} \) if and only if \( \mathcal{M} \) is complete.

Given a smooth curve \( \gamma \) on \( \mathcal{M} \), parallel transport of the tangent space at \( \gamma(t_0) \) to the tangent space at \( \gamma(t_1) \) along \( \gamma \), \( \mathcal{P}_\gamma^{t_1 \to t_0} : T_{\gamma(t_0)}\mathcal{M} \to T_{\gamma(t_1)}\mathcal{M} \) is defined by \( \mathcal{P}_\gamma^{t_1 \to t_0} (u) = Z(t_1) \), where \( Z \) is a unique parallel vector field such that \( Z(t_0) = u \). Parallel transport from \( x \) to \( y \) depends on the curve connecting \( x \) and \( y \). The parallel transport operator \( \mathcal{P}_\gamma^{t_1 \to t_0} \) is linear; \( \mathcal{P}_\gamma^{t_2 \to t_1} \circ \mathcal{P}_\gamma^{t_1 \to t_0} = \mathcal{P}_\gamma^{t_2 \to t_0} \); \( \mathcal{P}_\gamma^{t \to t} \) is the identity. In particular, the inverse of \( \mathcal{P}_\gamma^{t_1 \to t_0} \) is \( \mathcal{P}_\gamma^{t_0 \to t_1} \). Moreover, parallel transport is an isometry.

Let \( R \) be a retraction on a manifold \( \mathcal{M} \); then \( T_{\eta_x}(\xi_x) := \mathcal{P}_\gamma^{1 \to 0} \xi_x \) is a vector transport with an associated retraction \( R_x \), where \( \mathcal{P}_\gamma \) denotes parallel transport along the curve \( t \mapsto \gamma(t) = R_x \ (t\eta_x) \).

**Riemannian product manifold** \( \mathcal{M} \) For the Riemannian product manifold \( \mathcal{M} = \mathcal{M} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m \), at a point \( w = (x, y, s, z) \), the tangent space \( T_w\mathcal{M} = T_x\mathcal{M} \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^m \). For any \( w \) and \( \xi, \zeta \in T_w\mathcal{M} \), the Riemannian product metric is defined as \( \langle \xi, \zeta \rangle_w := \langle \xi_x, \zeta_x \rangle + \xi_y^T \zeta_y + \xi_s^T \zeta_s + \xi_z^T \zeta_z \), where \( \xi = (\xi_x, \xi_y, \xi_s, \xi_z) \) and \( \zeta = (\zeta_x, \zeta_y, \zeta_s, \zeta_z) \). Accordingly, the induced norm \( \| \xi \|_w := \sqrt{\langle \xi, \xi \rangle_w} \) satisfies

\[
\| \xi \|_w^2 = \| \xi_x \|^2 + \| \xi_y \|^2 + \| \xi_s \|^2 + \| \xi_z \|^2.
\]
Here, \( \| \cdot \| \) denotes the usual \( l_2 \) norm. The Riemannian distance on \( \mathcal{M} \) is defined as
\[
d(w_1, w_2) := \sqrt{d^2(x_1, x_2) + \|y_1 - y_2\|^2 + \|s_1 - s_2\|^2 + \|z_1 - z_2\|^2}.
\]
For any \( w \in \mathcal{M} \) and \( \xi \in T_w \mathcal{M} \),
\[
\bar{R}_w(\xi) := (R_x(\xi_x), y + \xi_y, s + \xi_s, z + \xi_z)
\]
defines a retraction on \( \mathcal{M} \). \( \bar{R} \) is the exponential map on \( \mathcal{M} \) if \( R \) is the exponential map on \( \mathbb{M} \). For any \( w \in \mathcal{M} \) and \( \xi, \zeta \in T_w \mathcal{M} \),
\[
T_\zeta \xi = T_{(\xi_x, \xi_y, \xi_s, \xi_z)}(\xi_x, \xi_y, \xi_s, \xi_z) := (T_{\zeta_x}(\xi_x), \xi_y, \xi_s, \xi_z)
\]
is a vector transport with an associated retraction \( \bar{T} \); equivalently, the linear operator \( \bar{T}_\zeta : T_w \mathcal{M} \to T_{\bar{R}_w(\xi)} \mathcal{M} \) is defined as \( \bar{T}_\zeta = (T_{\zeta_x}, \text{id}, \text{id}, \text{id}) \). If it exists, the inverse \( \bar{T}_\zeta^{-1} = (T_{\zeta_x}^{-1}, \text{id}, \text{id}, \text{id}) \).

Note that \( \bar{T} \) is isometric if \( T \) is isometric. Consider \( \mathbb{R}^d \) and \( \mathbb{R}^m \) equipped with the canonical Euclidean connection; then, the connection \( \nabla \) on \( \mathbb{M} \) essentially determines a connection on \( \mathcal{M} \) (see Lemma A.1 for a simple version), which is Riemannian provided that \( \nabla \) is Riemannian. In this paper, we conflate the notations of \( \nabla \) and \( d \) for \( \mathcal{M} \) and \( \mathbb{M} \) since they are clear from context.

### 3.2 Preliminaries

Here, we introduce two important topological concepts in Riemannian geometry: retractive neighborhood and totally retractive neighborhood. These concepts were first proposed in [19] and formally stated in [36].

#### 3.2.1 Totally retractive neighborhood

From \( DR_x(0_x) = \text{id}_{T_x \mathcal{M}} \) and the inverse function theorem, there exists a neighborhood \( V \) of \( 0_x \) in \( T_x \mathcal{M} \) such that \( R_x \) is a diffeomorphism on \( V \). We call \( U = R_x(V) \) a **retractive neighborhood** of \( x \). The well-known concept of the **normal neighborhood** of \( x \) is defined in the same way, where \( R \) is the exponential map.

The following existence theorem gives the definition of a **totally retractive neighborhood**. The proof is omitted from [19] and [36], since it can be arrived at along the lines of [9, Theorem 3.7], which is the same statement as Theorem 3.1 but restricted to the exponential map \( R = \exp \).

**Theorem 3.1 ([36, Theorem 2]).** *Let \( R \) be a retraction on \( \mathcal{M} \). For any \( \bar{x} \in \mathcal{M} \), there exists a so-called **totally retractive neighborhood** \( W \) of \( \bar{x} \) and a number \( \delta > 0 \) such that, for every \( x \in W \), \( R_x \) is a diffeomorphism on the open ball \( B_{\delta}(0_x) \subset T_x \mathcal{M} \) centered at \( 0_x \) with radius \( \delta \). \( W \subset R_x(B_{\delta}(0_x)) \), and \( R : (x, \xi_x) \mapsto (x, R_x(\xi_x)) \) is a diffeomorphism on \( \{(x, \xi_x) : x \in W, \xi_x \in B_{\delta}(0_x)\} \subset TM \).*

**Remark 3.2.** The existence of a totally retractive neighborhood shows that there is a neighborhood \( W \) of \( \bar{x} \) such that \( R_x^{-1}(y) \) is well defined for any \( x, y \in W \). In what follows, we choose such a neighborhood of \( \bar{x} \) by default if necessary.
3.2.2 Lipschitz continuity with respect to a vector transport

Multiple Riemannian versions of Lipschitz continuity have been defined. Traditionally, Lipschitz continuity in the Riemannian setting, e.g., [4, Section 10.4], is defined using parallel transport along the minimal geodesic connecting \(x\) and \(y\). When applying the standard Newton update, parallel transport is satisfactory for theoretical purposes, but it is of limited help in practice. For example, the quasi-Newton update in the Riemannian case requires tangent vectors in different spaces to be compared; thus, those algorithms require a tractable vector transport. Here, we consider Lipschitz continuity with respect to general vector transport and its associated retraction.

First, let us consider the Lipschitz-continuous gradient of a scale function \(f\).

**Definition 3.3** ([17, Definition 5.2.1]). Let \(M\) be a Riemannian manifold endowed with a vector transport \(T\) and an associated retraction \(R\). A function \(f: M \rightarrow \mathbb{R}\) is Lipschitz continuously differentiable with respect to \(T\) in \(U \subset M\) if it is differentiable and if there exists a number \(\kappa > 0\) such that, for all \(x, y \in U\),

\[
\| \text{grad} f(y) - T_\xi \text{grad} f(x) \| \leq \kappa \| \xi \|,
\]

where \(\xi = R_x^{-1} y\).

**Lemma 3.4** ([17, Lemma 5.2.1]). If a function \(f: M \rightarrow \mathbb{R}\) is \(C^2\), then, for any \(\bar{x} \in M\) and for any given vector transport \(T\), there exists a neighborhood \(U\) of \(\bar{x}\) such that \(f\) is Lipschitz continuously differentiable with respect to \(T\) in \(U\).

We get the following result when the gradient is replaced by a general vector field.

**Definition 3.5.** Let \(M\) be a Riemannian manifold endowed with a vector transport \(T\) and an associated retraction \(R\). A vector field \(F\) is Lipschitz continuous with respect to \(T\) in \(U \subset M\) if there exists a number \(\kappa > 0\) such that, for all \(x, y \in U\),

\[
\| F(y) - T_\xi F(x) \| \leq \kappa \| \xi \|,
\]

where \(\xi = R_x^{-1} y\).

**Lemma 3.6.** If \(F\) is a \(C^1\) vector field, then, for any \(\bar{x} \in M\) and any given vector transport \(T\), there exists a neighborhood \(U\) of \(\bar{x}\) such that \(F\) is Lipschitz continuously differentiable with respect to \(T\) in \(U\).

**Proof.** The proof of this lemma is similar to that of [17, Lemma 5.2.1]. \(\square\)

Going one degree higher, let us now discuss the Lipschitz-continuous Hessian of a scale function \(f\). Recall that the Hessian of \(f\) associates to each \(x\) a linear operator \(\text{Hess} f(x)\) from \(T_x M\) to \(T_x M\).

**Definition 3.7** ([19, Assumption 3]). Let \(M\) be a Riemannian manifold endowed with a vector transport \(T\) and an associated retraction \(R\). A function \(f: M \rightarrow \mathbb{R}\) is twice Lipschitz continuously differentiable with respect to \(T\) in \(U \subset M\) if it is twice differentiable and if a number \(\kappa > 0\) exists such that, for all \(x, y \in U\),

\[
\| \text{Hess} f(y) - T_\xi \text{Hess} f(x)T_\xi^{-1} \| \leq \kappa d(x, y),
\]

where \(\xi = R_x^{-1} y\).
Lemma 3.8 ([19, Lemma 4]). If \( f : \mathcal{M} \to \mathbb{R} \) is \( C^3 \), then, for any \( \bar{x} \in \mathcal{M} \) and for any given vector transport \( \mathcal{T} \), there exists a neighborhood \( U \) of \( \bar{x} \), such that \( f \) is twice Lipschitz continuously differentiable with respect to \( \mathcal{T} \) in \( U \).

Similarly, when the Hessian is replaced by a general covariant derivative, i.e., Hess \( f \) by \( \nabla F \), we get the following result.

**Definition 3.9.** Let \( \mathcal{M} \) be a Riemannian manifold endowed with a vector transport \( \mathcal{T} \) and an associated retraction \( R \). The covariant derivative \( \nabla F \) is **Lipschitz continuous with respect to** \( \mathcal{T} \) in \( U \subset \mathcal{M} \) if there exists a number \( \kappa > 0 \) such that, for all \( x, y \in U \),

\[
\| \nabla F(y) - T_\xi \nabla F(x) T_\xi^{-1} \| \leq \kappa d(x, y),
\]

where \( \xi = R_x^{-1} y \).

**Lemma 3.10.** If \( f \) is a \( C^2 \) vector field, then, for any \( \bar{x} \in \mathcal{M} \) and any given vector transport \( \mathcal{T} \), there exists a neighborhood \( U \) of \( x \) such that covariant derivative \( \nabla F \) is Lipschitz continuous with respect to \( \mathcal{T} \) in \( U \).

*Proof.* This lemma can be proven in a similar way as [19, Lemma 4]. \( \square \)

### 3.3 Auxiliary Results

#### 3.3.1 Local errors of retraction, vector transport

**Lemma 3.11.** Let \( \mathcal{M} \) be a Riemannian manifold endowed with a (smooth) retraction \( R \) and let \( \bar{x} \in \mathcal{M} \). Then,

(i) for any \( \epsilon > 0 \) there is an \( \delta_\epsilon > 0 \) such that for all \( x \) in a sufficiently small neighborhood of \( \bar{x} \) and all \( \xi, \eta \in T_x \mathcal{M} \) with \( \| \xi \| \leq \delta_\epsilon \) and \( \| \eta \| \leq \delta_\epsilon \),

\[
(1 - \epsilon) \| \xi - \eta \| \leq d(R_x(\eta), R_x(\xi)) \leq (1 + \epsilon) \| \xi - \eta \|;
\]

(ii) there exist \( a_0 > 0, a_1 > 0 \), and \( \delta_{a_0,a_1} > 0 \) such that for all \( x \) in a sufficiently small neighborhood of \( \bar{x} \) and all \( \xi, \eta \in T_x \mathcal{M} \) with \( \| \xi \| \leq \delta_{a_0,a_1} \) and \( \| \eta \| \leq \delta_{a_0,a_1} \),

\[
a_0 \| \xi - \eta \| \leq d(R_x(\eta), R_x(\xi)) \leq a_1 \| \xi - \eta \|.
\]

(iii) there exist \( a_0 > 0, a_1 > 0 \), and \( \delta_{a_0,a_1} > 0 \) such that for all \( x \) in a sufficiently small neighborhood of \( \bar{x} \) and all \( \xi \in T_x \mathcal{M} \) with \( \| \xi \| \leq \delta_{a_0,a_1} \),

\[
a_0 \| \xi \| \leq d(x, R_x(\xi)) \leq a_1 \| \xi \|.
\]

(iv) there exist \( a_0 > 0, a_1 > 0 \) such that for all \( x \) in a sufficiently small neighborhood of \( \bar{x} \), \( \xi = R^{-1}_x (x) \) is well defined and

\[
a_0 \| \xi \| \leq d(x, \bar{x}) \leq a_1 \| \xi \|.
\]

*Proof.* (i) and (ii) come directly from [26, Lemma 6] and [19, Lemma 2]. Note that if we select any \( \epsilon > 0 \) and let \( a_0 := 1 - \epsilon, a_1 := 1 + \epsilon \), then (i) implies (ii). (iii) follows from (ii) when we take \( \eta = 0 \). Next, we show (iv).

Taking \( x = \bar{x} \) in (iii), we have \( a_0 \| \xi \| \leq d(\bar{x}, R_\bar{x} (\xi)) \leq a_1 \| \xi \| \) for all \( \xi \in T_\bar{x} \mathcal{M} \) with \( \| \xi \| \leq \delta_{a_0,a_1} \). Since \( R_\bar{x} \) is a local diffeomorphism at \( 0_\bar{x} \in T_\mathcal{M} \), for all \( x \) in a sufficiently small neighborhood of \( \bar{x} \), \( \xi = R^{-1}_\bar{x} (x) \) is well defined and \( \| \xi \| \leq \delta_{a_0,a_1} \). Therefore, by substituting \( R_\bar{x} (\xi) = x \), we have \( a_0 \| \xi \| \leq d(x, \bar{x}) \leq a_1 \| \xi \| \). \( \square \)
Lemma 3.12 ([14, Lemma 14.5]). Let $F$ be a $C^1$ vector field on a Riemannian manifold $\mathcal{M}$ and let $x^* \in \mathcal{M}$ be a nonsingular zero of $F$, i.e., $F(x^*) = 0$ and $\nabla F(x^*)$ be nonsingular. Then, there exists a neighborhood $U$ of $x^*$ with $a_2 > 0$, $a_3 > 0$ such that, for all $x \in U$,

$$a_2 d(x, x^*) \leq \| F(x) \| \leq a_3 d(x, x^*).$$

The following corollary combines Lemma 3.12 with (iv) of Lemma 3.11.

Corollary 3.13. Let $F$ be a $C^1$ vector field on a Riemannian manifold $\mathcal{M}$ and let $x^* \in \mathcal{M}$ be a nonsingular zero of $F$. Then, there exists a neighborhood $U$ of $x^*$ with $a_4 > 0$, $a_5 > 0$ such that, for all $x \in U$,

$$a_4 \| \xi \| \leq \| F(x) \| \leq a_5 \| \xi \|.$$

where $\xi = R^{-1}_{x^*}(x)$.

The next lemma compares any two vector transports. In particular, we often compare general vector transport with parallel transport.

Lemma 3.14 ([19, Lemma 6]). Let $\mathcal{M}$ be a Riemannian manifold endowed with two vector transports $\bar{T}_1$ and $\bar{T}_2$, and let $\bar{x} \in \mathcal{M}$. Then, there exist a constant $a_6$ and a neighborhood $U$ of $\bar{x}$ such that, for all $x, y \in U$ and all $\xi \in T_{x} \mathcal{M}$,

$$\| \bar{T}^{-1}_{1 \eta} \xi - \bar{T}^{-1}_{2 \eta} \xi \| \leq a_6 \| \xi \| \| \eta \|,$$

where $\eta = R^{-1}_x(y)$.

3.3.2 Fundamental theorem of calculus in the Riemannian case

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be $C^1$ in an open convex set $D \subseteq \mathbb{R}^n$ containing $x$. We know that the fundamental theorem of calculus, e.g., [8, Lemma 4.1.9], is stated as: for $x + p =: y \in D$,

$$F(y) - F(x) = \int_{0}^{1} J(x + tp)pd, t,$$

where $J(x) \in \mathbb{R}^{n \times n}$ denotes the Jacobian matrix of $F$ at $x$.

In the Euclidean case, the next lemma holds with $c_1 = 0$ and reduces to the above.

Lemma 3.15 ([19, Lemma 8]). Let $F$ be a $C^1$ vector field on a Riemannian manifold $\mathcal{M}$, let $R$ be a retraction on $\mathcal{M}$, and let $\bar{x} \in \mathcal{M}$. Then, there exist a neighborhood $U$ of $\bar{x}$ and a constant $c_1$ such that, for all $x, y \in U$,

$$\| \mathcal{P}^{0}_{\gamma} - F(x) - \int_{0}^{1} \mathcal{P}^{0}_{\gamma} \nabla F(\gamma(t)) \mathcal{P}^{t - 0}_{\gamma} \eta dt \| \leq c_1 \| \eta \|^2,$$

where $\eta = R^{-1}_x(\eta)$ and $\mathcal{P}_\gamma$ is a parallel transport along the curve $t \to \gamma(t) = R_x(t\eta)$.

In particular, $c_1 = 0$ also holds if $\gamma(t) = \exp_x(t\eta)$. Note that this result is also shown in Lemma 3.16.

Lemma 3.16 ([13, equality (2.4)]). Let $F$ be a $C^1$ vector field on a Riemannian manifold $\mathcal{M}$ and let $\bar{x} \in \mathcal{M}$. Then, there exists a neighborhood $U$ of $\bar{x}$ such that, for all $x, y \in U$,

$$\mathcal{P}^{0}_{\gamma} - F(x) = \int_{0}^{1} \mathcal{P}^{0}_{\gamma} \nabla F(\gamma(t)) \mathcal{P}^{t - 0}_{\gamma} \eta dt,$$

where $\eta = R^{-1}_x(\eta)$ and $\mathcal{P}_\gamma$ is a parallel transport along the curve $t \to \gamma(t) = \exp_x(t\eta)$. 

14
### 3.3.3 Some estimations in Riemannian case

In the Euclidean setting, there are two important estimations in the analysis of the Newton and quasi-Newton methods (See [8, Lemma 4.1.12 & 4.1.15]). Here, let $F : \mathbb{R}^n \to \mathbb{R}^n$ be $C^1$ in the open convex set $D \subset \mathbb{R}^n$ containing $x$, and let Jacobian $x \mapsto J(x)$ be Lipschitz continuous with constant $\gamma$. Then, for any $x + p = y \in D$,

$$\|F(y) - F(x) - J(x)p\| \leq \frac{\gamma}{2}\|p\|^2.$$ 

Moreover, for any $v, u \in D$,

$$\|F(v) - F(u) - J(x)(v - u)\| \leq \gamma \max\{\|v - x\|, \|u - x\|\}\|v - u\|.$$

The first estimation in the Riemannian case can also be used for analysis of the standard Riemannian Newton update. Since no tangent vectors need to be transported in the standard Riemannian Newton algorithm, we can use parallel transport to simplify the theoretical analysis.

**Lemma 3.17.** Let $F$ be a $C^2$ vector field and $\bar{x} \in \mathcal{M}$. Then, there exist a neighborhood $\mathcal{U}$ of $\bar{x}$ and a constant $c_2$ such that, for all $x \in \mathcal{U}$,

$$\|\mathcal{P}_\gamma^{0-t}F(x) - F(\bar{x}) - \nabla F(\bar{x})\eta\| \leq c_2 d^2(\bar{x}, x).$$

where $\eta = R_{\bar{x}}^{-1}x$ and $\mathcal{P}_\gamma$ is a parallel transport along the curve $t \mapsto \gamma(t) = R_{\bar{x}}(t\eta)$.

**Proof.** It follows from

$$\begin{align*}
\|\mathcal{P}_\gamma^{0-t}F(x) - F(\bar{x}) - \nabla F(\bar{x})\eta\| & \leq \|\mathcal{P}_\gamma^{0-t}F(x) - F(\bar{x}) - \int_0^1 \mathcal{P}_\gamma^{0-t}\nabla F(\gamma(t))\mathcal{P}_\gamma^{t}\eta dt\| \\
& \quad + \|\int_0^1 \mathcal{P}_\gamma^{0-t}\nabla F(\gamma(t))\mathcal{P}_\gamma^{t}\eta dt - \nabla F(\bar{x})\eta\| \\
& \leq c_1 \|\eta\|^2 + \left\|\int_0^1 \left[\mathcal{P}_\gamma^{0-t}\nabla F(\gamma(t))\mathcal{P}_\gamma^{t} - \nabla F(\bar{x})\right] \eta dt\right\|. \quad \text{(by Lemma 3.15)}
\end{align*}$$

Let $\theta$ be the last term of the above inequality, i.e., $\theta := \int_0^1 \left[\mathcal{P}_\gamma^{0-t}\nabla F(\gamma(t))\mathcal{P}_\gamma^{t} - \nabla F(\bar{x})\right] \eta dt$. Note that

$$\|\theta\| \leq \int_0^1 \|\mathcal{P}_\gamma^{0-t}\nabla F(\gamma(t))\mathcal{P}_\gamma^{t} - \nabla F(\bar{x})\| \|\eta\| dt$$

$$\leq \int_0^1 c_0d(\bar{x}, R_{\bar{x}}(t\eta)) \|\eta\| dt \quad \text{(by Lemma 3.10 and Lipschitz continuity at $\bar{x}$)}$$

$$\leq \int_0^1 c_0 a_1 \|\eta\| \|\eta\| dt = \frac{1}{2} c_0 a_1 \|\eta\|^2. \quad \text{(by (iv) of Lemma 3.11)}$$

Again, by (iv) of Lemma 3.11, we have

$$\|\mathcal{P}_\gamma^{0-t}F(x) - F(\bar{x}) - \nabla F(\bar{x})\eta\| \leq (c_1 + \frac{1}{2} c_0 a_1) \|\eta\|^2 \leq (c_1 + \frac{1}{2} c_0 a_1) a_0^2 d^2(\bar{x}, x).$$

Letting $c_2 := (c_1 + \frac{1}{2} c_0 a_1)/a_0^2$ completes the proof. \qed
Likewise, the second estimation in the Riemannian case is for the analysis of the Riemannian quasi-Newton updates.

**Lemma 3.18.** Let $F$ be a $C^2$ vector field and $\bar{x} \in \mathcal{M}$. Then, there exist a neighborhood $\mathcal{U}$ of $\bar{x}$ and a constant $c_3$ such that, for all $x, y \in \mathcal{U}$,

\[
(i) \quad \|T_s^{-1} F(y) - F(x) - T_s \nabla F(\bar{x}) T_s^{-1} s\| \leq c_3 \|s\| \max\{d(x, \bar{x}), d(y, \bar{x})\},
\]

where $\zeta = R_{\bar{x}}^{-1}(x), s = R_{\bar{x}}^{-1}(y)$.

\[
(ii) \quad \|T_{\eta}^{-1} F(\bar{x}) - F(x) - T_{\eta} \nabla F(\bar{x}) T_{\eta}^{-1} \eta\| \leq c_3 \|\eta\| d(x, \bar{x}),
\]

where $\eta = R_{\bar{x}}^{-1}(x), \eta' = R_{\bar{x}}^{-1}(x)$.

**Proof.** (i) can be proven in the same way as [19, Lemma 12], where $F$ is specified as the Riemannian gradient and $\nabla F$ as the Riemannian Hessian. (ii) is a special case of (i): set $y = \bar{x}, s = \eta$, and $\zeta = \eta'$.

We end this section with the following lemmas.

**Lemma 3.19 (Banach’s Lemma).** Let $A, B$ be linear operators on $T_x \mathcal{M}$. If $A$ is nonsingular and $\|A^{-1}\| B - A \| < 1$, then $B$ is nonsingular and

\[
\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| B - A \|.}
\]

**Lemma 3.20 ([11, Lemma 3.2]).** Suppose that $\nabla F$ is continuous at $p^*$. If $\nabla F(p^*)$ is nonsingular; then, there exist a neighborhood $\mathcal{U}$ of $p^*$ and a positive constant $\Theta$ such that, for all $p \in \mathcal{U}$, $\nabla F(p)$ is nonsingular and $\|\nabla F(p)^{-1}\| \leq \Theta$.

### 4 Local Convergence of Newton RIP

Along the lines of the Euclidean local analysis in [10, Section 5], we begin with a perturbed damped Newton method and then derive its local convergence theory. Note that the convergence of Newton RIP results from an application of the perturbed damped Newton method and will be described later. Here, consider again the problem,

\[ F(p) = 0, \tag{24} \]

where $F$ is a $C^1$ vector field on $\mathcal{M}$. The standard assumptions for (24) are as follows:

**Assumption 2 (Riemannian Newton assumptions [3, Theorem 2]).**

(B1) There exists $p^* \in \mathcal{M}$ such that $F(p^*) = 0$.

(B2) The operator $\nabla F(p^*)$ is nonsingular.

(B3) The covariant derivative $\nabla F$ is locally Lipschitz continuous at $p^*$.

Then, the iteration sequence of the perturbed damped Newton method for (24) is as follows: let $0 < \alpha_k \leq 1, \mu_k > 0$ and $k = 0, 1, 2, \ldots$

**Algorithm 3 (Perturbed Damped Newton Method).**

(Step 1) Compute $v_k \in T_{p_k} \mathcal{M}$ by solving the perturbed equation $\nabla F(p_k) v_k + F(p_k) = \mu_k \hat{e}$. 

16
(Step 2) Compute \( p_{k+1} := R_{p_k}(\alpha_k v_k) \) by using a damped step size \( \alpha_k \). Return to Step 1.

**Proposition 4.1.** Consider the perturbed damped Newton method for problem \((24)\). Let the standard assumptions \((B1)-(B3)\) hold at \( p^* \). Choose parameters \( \mu_k, \alpha_k \) as below; then there exists a positive constant \( \delta \) such that for all \( d(p_k, p^*) < \delta \), the sequence \( \{ p_k \} \) is well defined. Furthermore,

(i) if \( \mu_k = o(\| F(p_k) \|) \) and \( \alpha_k \to 1 \), then \( \{ p_k \} \) converges to \( p^* \) superlinearly;

(ii) if \( \mu_k = O(\| F(p_k) \|^2) \) and \( 1 - \alpha_k = O(\| F(p_k) \|) \), then \( \{ p_k \} \) converges to \( p^* \) quadratically.

**Proof.** By Lemma 3.20, we can let \( p_k \) be sufficiently close to \( p^* \) such that \( \nabla F(p_k) \) is nonsingular, and \( \| \nabla F(p_k)^{-1} \| \leq \Theta \). Then, the next iterate,

\[
p_{k+1} := R_{p_k}[\alpha_k \nabla F(p_k)^{-1}(-F(p_k) + \mu_k \hat{e})],
\]

is well defined, and hence, it follows from \( p^* = R_{p_k}(\eta) \) with \( \eta := R_{p_k}^{-1} p^* \) and Lemma 3.11 that

\[
d(p_{k+1}, p^*) \leq a_1 \| \alpha_k \nabla F(p_k)^{-1}(-F(p_k) + \mu_k \hat{e}) - \eta \|
= a_1 \| \eta + \alpha_k \nabla F(p_k)^{-1}(F(p_k) - \mu_k \hat{e}) \|.
\]

Let \( r_k := \eta + \alpha_k \nabla F(p_k)^{-1}(F(p_k) - \mu_k \hat{e}) \). Algebraic manipulations show that

\[
r_k = (1 - \alpha_k) \eta + \alpha_k \nabla F(p_k)^{-1} \nabla F(p_k) \eta + \alpha_k \nabla F(p_k)^{-1}(F(p_k) - \mu_k \hat{e})
= (1 - \alpha_k) \eta + \alpha_k \nabla F(p_k)^{-1}[\nabla F(p_k) \eta + F(p_k) - \mu_k \hat{e}]
= (1 - \alpha_k) \eta + \alpha_k \nabla F(p_k)^{-1}[\nabla F(p_k) \eta + F(p_k) - T^\gamma \nabla F(p^*) - \mu_k \hat{e}],
\]

where \( T^\gamma \) is a parallel transport along the curve \( \gamma \) given by \( \gamma(t) = R_{p_k}(t \eta) \).

Thus, using \( \| \eta \| \leq \frac{1}{a_0} d(p_k, p^*) \) from (iii) of Lemma 3.11 and the first estimation of Lemma 3.17, we have

\[
\| r_k \| \leq (1 - \alpha_k) \| \eta \|
+ \alpha_k \| \nabla F(p_k)^{-1} \| \| T^\gamma \nabla F(p^*) - F(p_k) - \nabla F(p_k) \eta \|
+ \alpha_k \| \nabla F(p_k)^{-1} \| \| \hat{e} \| \mu_k
\leq \frac{1}{a_0} (1 - \alpha_k) d(p_k, p^*) + \alpha_k \| \nabla F(p_k)^{-1} \| c_2 d^2(p_k, p^*) + \alpha_k \| \nabla F(p_k)^{-1} \| \| \hat{e} \| \mu_k
\leq \frac{1}{a_0} (1 - \alpha_k) d(p_k, p^*) + \Theta c_2 d^2(p_k, p^*) + \Theta \| \hat{e} \| \mu_k. \quad \text{(by Lemma 3.20 and } 0 < \alpha_k \leq 1\)
\]

Therefore, by combining the above with \((25)\), we conclude that

\[
d(p_{k+1}, p^*) \leq \kappa_1 (1 - \alpha_k) d(p_k, p^*) + \kappa_2 d^2(p_k, p^*) + \kappa_3 \mu_k
\]

for some positive constants \( \kappa_1, \kappa_2, \kappa_3 \). On the other hand, by Lemma 3.12, we have

\[
\| F(p_k) \| = O(d(p_k, p^*)), \quad \text{(27)}
\]

In what follows, we prove assertions (i) and (ii).

(i) Suppose that \( \alpha_k \to 1 \) and \( \mu_k = o(\| F(p_k) \| \) , which together imply \( \mu_k = o(d(p_k, p^*)) \).

By (26), we have

\[
\frac{d(p_{k+1}, p^*)}{d(p_k, p^*)} \leq \kappa_1 (1 - \alpha_k) + \kappa_2 d(p_k, p^*) + \kappa_3 \frac{\mu_k}{d(p_k, p^*)}, \quad \text{(28)}
\]
and we can take $\delta$ sufficiently small and $k$ sufficiently large, if necessary, to conclude that
\[ d(p_{k+1}, p^*) < \frac{1}{2}d(p_{k}, p^*) < \delta. \]

Thus, $p_{k+1} \in B_{\delta}(p^*)$. By mathematical induction, it is easy to show that the sequence $\{p_k\}$ is well defined and converges to $p^*$. Taking the limit of both sides of (28) proves superlinear convergence.

(ii) Again, we start from (26):
\[ d(p_{k+1}, p^*) = (1 - \alpha_k)O(d(p_{k}, p^*)) + O(d^2(p_{k}, p^*)) + O(\mu_k). \]  

(29)

Suppose that $1 - \alpha_k = O(\|F(p_k)\|)$ and $\mu_k = O(\|F(p_k)\|^2)$. Using (27), the above equality reduces to
\[ d(p_{k+1}, p^*) = O(d^2(p_{k}, p^*)). \]

This implies that there exists a positive constant $\nu$ such that $d(p_{k+1}, p^*) \leq \nu d^2(p_{k}, p^*)$, and hence,
\[ d(p_{k+1}, p^*) \leq \nu d^2(p_{k}, p^*) \leq \nu \delta^2 < \delta, \]

if $\delta$ is sufficiently small. Again, by mathematical induction, it is easy to show that the sequence $\{p_k\}$ converges to $p^*$ quadratically. 

Now, let us establish local convergence of Newton RIP in a way that almost replicates the perturbed damped Newton method except for taking care of $\gamma_k$. We consider the next stronger assumption instead of (A2):

(A2') Smoothness of $C^3$. The functions $f, g, h$ are $C^3$ on $\mathbb{M}$.

From Lemma 3.8, (A2') implies that the covariant derivative $\nabla F$ is Lipschitz continuous with respect to any vector transport, where $F$ is the associated KKT vector field. Thus, from Proposition 2.4, assumptions (A1)-(A5) together with (A2') show that $F$ satisfies the standard Riemannian Newton assumptions (B1)-(B3).

**Theorem 4.2** (Locally quadratic convergence of Newton RIP). Consider Algorithm 2 for solving problem (RCOP). Let the standard assumptions (A1)-(A5) and (A2') hold at $w^*$. Choose the parameters such that
\[ \mu_k = O(\|F(w_k)\|^2) \quad \text{and} \quad 1 - \gamma_k = O(\|F(w_k)\|). \]  

(30)

Then, there exists a positive constant $\delta$ such that, for all $d(w_0, w^*) < \delta$, $w_0 \in \mathcal{M}$, the sequence $\{w_k\}$ is well defined and converges quadratically to $w^*$.

**Proof.** Suppose that $d(w_k, w^*) < \delta$ for sufficiently small $\delta$. Since $F$ satisfies assumptions (B1)-(B3), form the proof of Proposition 4.1 and equation (29), we also have
\[ d(w_{k+1}, w^*) = (1 - \alpha_k)O(d(w_k, w^*)) + O(d^2(w_k, w^*)) + O(\mu_k). \]

Since $\mu_k = O(\|F(w_k)\|^2)$, and $\|F(w_k)\| = O(d(w_k, w^*))$ by equation (27), we obtain
\[ \mu_k = O(d^2(w_k, w^*)). \]  

(31)
Thus, we have

\[
\|\Delta w_k\| = \|\nabla F(w_k)^{-1}(-F(w_k) + \mu_k \hat{e})\|
\leq \Theta(\|F(w_k)\| + \mu_k \|\hat{e}\|) \quad (\text{by Lemma 3.20})
\leq O(F(w_k)) + O(\mu_k)
= O(d(w_k, w^*)) + O(d^2(w_k, w^*)) = O(d(w_k, w^*)) \quad (\text{by equation (31)}).
\]

Since \(\delta\) is sufficiently small, from equation (31) and the above inequalities, the conditions of Lemmas 2.7 are satisfied. Hence, we have

\[
0 \leq 1 - \alpha_k \leq (1 - \gamma_k) + \Omega \|\Delta w_k\| = (1 - \gamma_k) + O(d(w_k, w^*)),
\]

and

\[
d(w_{k+1}, w^*) = (1 - \alpha_k)O(d(w_k, w^*)) + O(d^2(w_k, w^*)) + O(\mu_k)
\leq [(1 - \gamma_k) + O(d(w_k, w^*))]O(d(w_k, w^*))
+ O(d^2(w_k, w^*)) + O(d^2(w_k, w^*)) \quad (\text{by equations (31) and (32)})
= O(d^2(w_k, w^*)�)
\]

The next theorem can be proven similarly.

**Theorem 4.3** (Locally superlinear convergence of Newton RIP). *Let the assumptions of Theorem 4.2 hold. Choose the parameters such that

\[
\mu_k = o(\|F(w_k)\|) \quad \text{and} \quad \gamma_k \to 1.
\]

Then, there exists a positive constant \(\delta\) such that, for all \(d(w_0, w^*) < \delta, w_0 \in \mathcal{M}\), the sequence \(\{w_k\}\) is well defined and converges superlinearly to \(w^*\).*

### 5 Local and Linear Convergence of quasi-Newton RIP

In this section, we will show local and linear convergence of quasi-Newton RIP. Quasi-Newton RIP is obtained by replacing only (step 2) of Algorithm 2, as follows: Instead of equation (14) in (step 2), solve equation (17), where \(\text{Hess}_x \mathcal{L}(w_k)\) in the first line of (5) is approximated by \(G_k\), and \(\nabla F(w_k)\) is approximated by \(B_k\); see (16).

Recall that \(w_k = (x_k, y_k, s_k, z_k), \ w^* = (x^*, y^*, s^*, z^*)\). Let \(\zeta_k^x := R_{x^*}^{-1}(x_k), \ \zeta_k^y := y_k - y^*, \ \zeta_k^t := z_k - s^*; \) then, by (22) we have

\[
(R_{x^*}(\zeta_k^x), y^* + \zeta_k^y, s^* + \zeta_k^t, z^* + \zeta_k^z) = R_{w^*}(\zeta_k) = w_k,
\]

where \(\zeta_k = (\zeta_k^x, \zeta_k^y, \zeta_k^t, \zeta_k^z) \in T_{w^*}\mathcal{M}\). In what follows, we assume that the parameter selection satisfies

\[
\mu_k = O(\|F(w_k)\|) \quad \text{and} \quad 0 < \hat{\gamma} \leq \gamma_k \leq 1
\]

for a given constant \(\hat{\gamma} \in (0, 1)\).

Lemma 5.1 is a generalization of [32, Lemma 7].
Lemma 5.1. Let $\mathbb{M}$ be a Riemannian manifold endowed with an isometric vector transport $T$ and an associated retraction $R$. Suppose that (A1)-(A5) and (A2') hold and that Algorithm 2 with equation (17) generates the sequence $\{w_k\}$. If there exist constants $\varepsilon > 0$ and $\delta > 0$ for each $w_k$ and linear operator $G_k$ satisfying

$$
d(w_k, w^*) \leq \varepsilon, \quad \left\| G_k - \mathcal{T}_{\xi_k} \text{Hess}_x \mathcal{L}(w^*) \mathcal{T}_{\xi_k}^{-1} \right\| \leq \delta, \quad (34)
$$

where $\xi_k \in T_{w_k}\mathbb{M}$ is defined by $\xi_k = \mathcal{R}_{x_k}^{-1}(x_k)$, then the following hold:

(i) let $\xi_k \in T_{w^*}\mathbb{M}$ be given by $\xi_k = \mathcal{R}_{w^*}^{-1}(w_k)$; then, for each $k$,

$$
\left\| B_k - \mathcal{T}_{\xi_k} \nabla F (w^*) \mathcal{T}_{\xi_k}^{-1} \right\| \leq \sqrt{\delta^2 + \delta c_1 \varepsilon + c_2 \varepsilon},
$$

for some positive constant $c_1, c_2$;

(ii) $B_k$ is nonsingular and $\left\| B_k^{-1} \right\| \leq \Phi$ with a positive constant $\Phi$ for each $k$;

(iii) furthermore, if $\mu_k = O(\|F(w_k)\|)$,

$$
1 - \alpha_k \leq (1 - \gamma_k) + O(F(w_k)) + O(\mu_k). \quad (35)
$$

Proof. Take any $\Delta w = (\Delta x, \Delta y, \Delta s, \Delta z) \in T_{u_k}\mathbb{M}$. First, we compute the expression of $\left( B_k - \mathcal{T}_{\xi_k} \nabla F (w^*) \mathcal{T}_{\xi_k}^{-1} \right) \Delta w$. By vector transport on the product manifold $\mathcal{T}_{\xi_k} = (\mathcal{T}_{\xi_k}, \text{id}, \text{id}, \text{id})$ in (23), its inverse $\mathcal{T}_{\xi_k}^{-1} = (\mathcal{T}_{\xi_k}^{-1}, \text{id}, \text{id}, \text{id})$, and (5) of Lemma 2.2, we obtain

$$
\mathcal{T}_{\xi_k} \nabla F (w^*) \mathcal{T}_{\xi_k}^{-1} \Delta w = \left( \frac{\mathcal{T}_{\xi_k} \text{Hess}_x \mathcal{L}(w^*) \mathcal{T}_{\xi_k}^{-1} \Delta x - \sum_{i=1}^{m} (\Delta y)_i \mathcal{T}_{\xi_k}^{-1} \text{grad} h_i(x^*) \text{grad}(x^*)}{\sum_{i=1}^{m} (\Delta z)_i \mathcal{T}_{\xi_k}^{-1} \text{grad} g_i(x^*)} \right). \quad (36)
$$

Then, subtracting (36) from $B_k \Delta w$ in (16) yields

$$
\left( B_k - \mathcal{T}_{\xi_k} \nabla F (w^*) \mathcal{T}_{\xi_k}^{-1} \right) \Delta w = \left( \frac{G_k - \mathcal{T}_{\xi_k} \text{Hess}_x \mathcal{L}(w^*) \mathcal{T}_{\xi_k}^{-1} \Delta x}{\sum_{i=1}^{m} (\Delta z)_i \mathcal{T}_{\xi_k}^{-1} \text{grad} g_i(x^*)} \right).
$$

For $h_i, i = 1, \ldots, l$, by the isometry of vector transport $T$, we get

$$
\langle \text{grad} h_i(x^*), \mathcal{T}_{\xi_k}^{-1} \Delta x \rangle = \langle \mathcal{T}_{\xi_k}^{-1} \text{grad} h_i(x^*), \Delta x \rangle = \langle \mathcal{T}_{\xi_k} \text{grad} h_i(x^*), \Delta x \rangle.
$$

Hence, the second row of (37) can be simplified as

$$
\langle \text{grad} h_i(x_k), \Delta x \rangle - \langle \text{grad} h_i(x^*), \mathcal{T}_{\xi_k}^{-1} \Delta x \rangle = \langle \text{grad} h_i(x_k), \Delta x \rangle - \langle \mathcal{T}_{\xi_k} \text{grad} h_i(x^*), \Delta x \rangle = \langle \text{grad} h_i(x_k) - \mathcal{T}_{\xi_k} \text{grad} h_i(x^*), \Delta x \rangle.
$$
In the same way, for \( g_i, i = 1, \ldots, m \), we have
\[
\langle \nabla g_i (x_k^*), \Delta x \rangle - \langle \nabla g_i (x^*), \mathcal{T}_{\xi^*_k}^{-1} \Delta x \rangle = \langle \nabla g_i (x_k^*) - \mathcal{T}_{\xi^*_k} \nabla g_i (x^*), \Delta x \rangle.
\]
Define
\[
\begin{align*}
\theta_i &:= \nabla h_i (x_k) - \mathcal{T}_{\xi^*_k} \nabla h_i (x^*), \quad \text{for } i = 1, \ldots, l, \\
\lambda_i &:= \nabla g_i (x_k) - \mathcal{T}_{\xi^*_k} \nabla g_i (x^*), \quad \text{for } i = 1, \ldots, m, \\
\alpha_1 &:= \left( G_k - \mathcal{T}_{\xi^*_k} \text{Hess}_x \mathcal{L}(w^*) \mathcal{T}_{\xi^*_k}^{-1} \right) \Delta x.
\end{align*}
\]
Then, (37) reduces to
\[
\left( B_k - \mathcal{T}_{\xi^*_k} \nabla F (w^*) \mathcal{T}_{\xi^*_k}^{-1} \right) \Delta w = \begin{pmatrix}
\alpha_1 - \sum_{i=1}^{l} (\Delta y)_i \theta_i - \sum_{i=1}^{m} (\Delta z)_i \lambda_i \\
\langle \theta_i, \Delta x \rangle, \quad \text{for } i = 1, \ldots, l \\
\langle \lambda_i, \Delta x \rangle, \quad \text{for } i = 1, \ldots, m \\
(Z_k - Z^*) \Delta s + (S_k - S^*) \Delta z
\end{pmatrix}.
\]
(38)

Note that, assumption (A2') and Lipschitz continuity of \( \{ h_i \}, \{ g_i \} \) at \( x^* \) (see Definition 3.3) imply that
\[
\| \theta_i \| \leq \kappa \| \xi^*_k \| \quad \text{for } i = 1, \ldots, l, \quad \text{and} \quad \| \lambda_i \| \leq \kappa \| \xi^*_k \| \quad \text{for } i = 1, \ldots, m.
\]
(39)

Without loss of generality, we can take a single constant \( \kappa \) for each \( h_i, g_i \) above. Again, define
\[
\begin{align*}
\alpha_2 &:= \sum_{i=1}^{l} (\Delta y)_i \theta_i + \sum_{i=1}^{m} (\Delta z)_i \lambda_i \quad \text{and} \quad \alpha := \alpha_1 - \alpha_2, \\
\beta_i &:= \langle \theta_i, \Delta x \rangle \quad \text{for } \beta \in \mathbb{R}^l, \\
\gamma_i &:= \langle \lambda_i, \Delta x \rangle \quad \text{for } \gamma \in \mathbb{R}^m, \\
\delta &:= \delta_1 + \delta_2 \in \mathbb{R}^m, \quad \text{where } \delta_1 := (Z_k - Z^*) \Delta s, \quad \text{and} \quad \delta_2 := (S_k - S^*) \Delta z.
\end{align*}
\]
These definitions and (38) imply that
\[
\left\| \left( B_k - \mathcal{T}_{\xi^*_k} \nabla F (w^*) \mathcal{T}_{\xi^*_k}^{-1} \right) \Delta w \right\|_{w_k}^2 = \| \alpha \|_{\xi^*_k}^2 + \| \beta \|^2 + \| \gamma \|^2 + \| \delta \|^2.
\]
(40)

Now, let us examine each term on the right-hand side of the equation above. First, we assert that
\[
\| \beta \|^2 \leq \sum_{i=1}^{l} \| \langle \theta_i, \Delta x \rangle \|^2 \leq \sum_{i=1}^{l} \left( \| \theta_i \| \| \Delta x \| \right)^2 \quad \text{(by the Cauchy–Schwarz inequality)}
\leq \kappa^2 \| \xi^*_k \|^2 \| \Delta x \|^2 \quad \text{(by (39))}
\leq \kappa^2 \frac{1}{a_0^2} d^2 (w_k, w^*) \| \Delta w \|^2 \quad \text{(by (iv) of Lemma 3.11 and (20))}
= O (d^2 (w_k, w^*)) \| \Delta w \|^2.
\]
(41)

In a similar manner, we obtain \( \| \gamma \|^2 = \sum_{i=1}^{m} \| \langle \lambda_i, \Delta x \rangle \|^2 = O (d^2 (w_k, w^*)) \| \Delta w \|^2 \). Next, for \( \| \delta \|^2 \) we see that
\[
\| \delta_1 \| = \| (Z_k - Z^*) \Delta s \|
\]
21
To deal with δ2, we note that
\[ |p\|^2 = \|\alpha_1 - \alpha_2\|^2 \leq (\|\alpha_1\|^2 + \|\alpha_2\|^2) = \|\alpha_1\|^2 + \|\alpha_2\|^2 + 2\|\alpha_1\|\|\alpha_2\|. \] (43)

Next, consider
\[ \|\alpha_2\| = \|\sum_{i=1}^l (\Delta y)_i \theta_i + \sum_{i=1}^m (\Delta z)_i \lambda_i\| \leq \sum_{i=1}^l |(\Delta y)_i| \|\theta_i\| + \sum_{i=1}^m |(\Delta z)_i| \|\lambda_i\| \leq \sum_{i=1}^l \|\Delta y\| \|\theta\| + \sum_{i=1}^m \|\Delta z\| \|\lambda\| \leq \kappa \max \{\sqrt{l}, \sqrt{m}\} \|\Delta w\| \leq \kappa \max \{\sqrt{l}, \sqrt{m}\} \frac{1}{a_0} d(\Delta w). \] (44)

Finally, combining the results from (40) to (44), we conclude that
\[ \left\| \left( B_k - \tilde{T}_{\zeta_k} \nabla F (w^*) \tilde{T}_{\zeta_k}^{-1} \right) \Delta w \right\|^2 \leq \left( \left( G_k - \tilde{T}_{\zeta_k} \text{Hess}_x \mathcal{L}(w^*) \tilde{T}_{\zeta_k}^{-1} \right) \Delta x \right)^2 + O(d^2(w_k, w^*)) \|\Delta w\|^2 + O(d(w_k, w^*)) \|\Delta w\| \left\| \left( G_k - \tilde{T}_{\zeta_k} \text{Hess}_x \mathcal{L}(w^*) \tilde{T}_{\zeta_k}^{-1} \right) \Delta x \right\|. \] (45)

Take any \( \Delta w = (\Delta x, \Delta y, \Delta s, \Delta z) \in T_{w_k} \mathcal{M} \) such that \( \|\Delta w\| = 1 \). Define
\[ S := B_k - \tilde{T}_{\zeta_k} \nabla F (w^*) \tilde{T}_{\zeta_k}^{-1}, \] (46)
\[ T := G_k - \tilde{T}_{\zeta_k} \text{Hess}_x \mathcal{L}(w^*) \tilde{T}_{\zeta_k}^{-1}. \] (47)

From (34), we have \( d(w_k, w^*) \leq \varepsilon \); thus, inequality (45) and (46) and (47) imply
\[ \|S\Delta w\|^2 \leq \|T\Delta x\|^2 + \|T\Delta x\|c_1\varepsilon + c_2\varepsilon \] (48)
for some constant \( c_1, c_2 > 0 \). Consider \( \Delta x \), which is a component of \( \Delta w \) of the unit norm; (20) implies \( \|\Delta x\| \leq \|\Delta w\| = 1 \). Thus, we have
\[ \|T\Delta x\| \leq \sup\{\|T\Delta x\| : \|\Delta x\| \leq 1, \Delta x \in T_{x_k} \mathcal{M} \} \text{ (since } \|\Delta x\| \leq 1) \]
Thus,

\[ \|S\| = \sup \{ \| S \Delta w \| : \| \Delta w \| = 1, \Delta w \in T_{x_k} \mathcal{M} \} \] (by the defintion of the operator norm)

\[ \leq \delta \] (by assumption (34)). \hspace{1cm} (49)

Thus,

\[ \|S\| \leq \sqrt{\delta^2 + \delta c_1\varepsilon + c_2\varepsilon} \] (by (48) and (49)). \hspace{1cm} (50)

Thus, by (46)

\[ \| B_k - \bar{T}_{x_k} \nabla F(w^*) \bar{T}_{x_k}^{-1} \| \leq \sqrt{\delta^2 + \delta c_1\varepsilon + c_2\varepsilon}. \]

This proves assertion (i).

Since \( \bar{T} \) is isometric, \( \| (\bar{T}_{x_k} \nabla F(w^*) \bar{T}_{x_k}^{-1})^{-1} \| = \| \bar{T}_{x_k} \nabla F(w^*)^{-1} \bar{T}_{x_k}^{-1} \| = \| \nabla F(w^*)^{-1} \|. \) By choosing \( \varepsilon \) and \( \delta \) such that

\[ \sqrt{\delta^2 + \delta c_1\varepsilon + c_2\varepsilon} \| \nabla F(w^*)^{-1} \| < 1, \]

we have

\[ \| (\bar{T}_{x_k} \nabla F(w^*) \bar{T}_{x_k}^{-1})^{-1} \| B_k - \bar{T}_{x_k} \nabla F(w^*) \bar{T}_{x_k}^{-1} \| \leq \sqrt{\delta^2 + \delta c_1\varepsilon + c_2\varepsilon} \| \nabla F(w^*)^{-1} \| < 1, \]

and it follows from the Banach Lemma 3.19 that \( B_k \) is nonsingular and

\[ \| B_k^{-1} \| \leq \Phi := \frac{\| \nabla F(w^*)^{-1} \|}{1 - \sqrt{\delta^2 + \delta c_1\varepsilon + c_2\varepsilon} \| \nabla F(w^*)^{-1} \|} \] \hspace{1cm} (52)

for some constant \( \Phi \). This proves assertion (ii).

Since \( B_k \) is nonsingular, \( \Delta w_k \) is well defined and we have

\[ \| \Delta w_k \| = \| B_k^{-1} (-F(w_k) + \mu_k \hat{e}) \| \] (by (17))

\[ \leq \| B_k^{-1} \| (\| F(w_k) \| + \mu_k \| \hat{e} \|) \]

\[ = O(F(w_k)) + O(\mu_k) \] (by (52)). \hspace{1cm} (53)

To prove (35), we note that if \( \varepsilon \) and \( \delta \) are sufficiently small, then from the parameter conditions (33) and inequality (53), the assumptions (18) of Lemmas 2.7 are satisfied. By (19), we have

\[ 0 \leq 1 - \alpha_k \leq (1 - \gamma_k) + \Omega \| \Delta w_k \| = (1 - \gamma_k) + O(F(w_k)) + O(\mu_k). \]

This proves assertion (iii). \hspace{1cm} \( \Box \)

The bounded deterioration property \([5, \text{Theorem 3.2}]\) is a sufficient condition for the general local convergence of Euclidean quasi-Newton methods. It covers most of the quasi-Newton update formulas, such as BFGS. The next theorem shows the local convergence property of our quasi-Newton RIP. We require the sequence of linear operators \( \{ G_k \} \) on \( T_{x_k} \mathcal{M} \) to satisfy an analogous bounded deterioration property. Theorem \( 5.2 \) is a generalization of \([32, \text{Theorem 2}]\) to a Riemannian manifold.

**Theorem 5.2** (Local and linear convergence of quasi-Newton RIP). Suppose that \( (A1)-(A5) \) and \( (A\bar{2}') \) hold. Choose the parameter such that

\[ \mu_k = O(\| F(w_k) \|^{1+\tau}) \]
for some positive constant $\tau$. Suppose further that the sequence of linear operators $\{G_k\}$ satisfies the bounded deterioration property:

$$\left\| G_{k+1} - \mathcal{T}_{\zeta_{k+1}} \right\| \leq \| \nabla^2 L(x^*) \mathcal{T}_{\zeta_{k+1}} \| \leq (1 + \beta_1 \sigma_k) \left\| G_k - \mathcal{T}_{\zeta_k} \nabla^2 L(x^*) \mathcal{T}_{\zeta_k}^{-1} \right\| + \beta_2 \sigma_k, \tag{54}$$

where $\zeta_k = R_{x^*}(x_k)$, $\beta_1$ and $\beta_2$ are positive constants, and $\sigma_k := \max\{d(w_k, w^*), d(w_k, w^*)\}$.

Then for Algorithm 2 with equation (17) and for each $\nu \in (1 - \gamma, 1)$, there exist positive constants $\delta_1 = \delta_2(\nu)$ such that if

$$d(w_0, w^*) \leq \delta, \quad \left\| G_0 - \mathcal{T}_{\zeta_0} \nabla^2 L(x^*) \mathcal{T}_{\zeta_0}^{-1} \right\| \leq \frac{1}{2} \delta,$$

the sequence $\{w_k\}$ is well defined and converges to $w^*$. Furthermore,

$$d(w_{k+1}, w^*) \leq \nu d(w_k, w^*), \quad \left\| G_k - \mathcal{T}_{\zeta_k} \nabla^2 L(x^*) \mathcal{T}_{\zeta_k}^{-1} \right\| \leq \delta, \tag{55}$$

for each $k \geq 0$.

Proof. By mathematical induction, we will prove that if for $i = 0, 1, \ldots, k$,

$$d(w_i, w^*) < \delta, \quad \left\| G_i - \mathcal{T}_{\zeta_i} \nabla^2 L(x^*) \mathcal{T}_{\zeta_i}^{-1} \right\| \leq \delta. \tag{56}$$

Then,

$$d(w_{k+1}, w^*) \leq \nu d(w_k, w^*) < \delta, \quad \left\| G_{k+1} - \mathcal{T}_{\zeta_{k+1}} \nabla^2 L(x^*) \mathcal{T}_{\zeta_{k+1}}^{-1} \right\| \leq \delta. \tag{57}$$

Condition (56) clearly holds when $k = 0$. If $\varepsilon$ and $\delta$ are sufficiently small, it follows from (ii) of Lemma 5.1 that $B_k$ is nonsingular and $\|B_k^{-1}\| \leq \Phi$ with a positive constant. From the linear system (17), we have

$$\Delta w_k = B_k^{-1}(-F(w_k) + \alpha k \epsilon). \tag{58}$$

Noting that $w^* = \tilde{R}_{w_k}(\xi_k)$ where $\xi_k := \tilde{R}_{w_k}^{-1}(w^*), w_{k+1} = \tilde{R}_{w_k}(\alpha k \Delta w_k)$, and (ii) of Lemma 3.11, we obtain

$$d(w_{k+1}, w^*) \leq a_1 \| \alpha k \Delta w_k - \xi_k \| = a_1 \| \alpha k B_k^{-1}(-F(w_k) + \alpha k \epsilon) - \xi_k \|. \tag{59}$$

Let $r := \alpha k B_k^{-1}(-F(w_k) + \alpha k \epsilon)$ then $\xi_k := \tilde{R}_{w_k}(\xi_k - B_k^{-1}(-F(w_k) + \alpha k \epsilon)) - \tilde{R}_{w_k}(\alpha k \Delta w_k)$, we have

$$r = (\alpha k - 1) \xi_k + \alpha k \mu_k B_k^{-1} \epsilon + \alpha k B_k^{-1} \left( B_k - \mathcal{T}_{\zeta_k} \nabla F(w^*) \mathcal{T}_{\zeta_k}^{-1} \right) \mathcal{T}_{\zeta_k} \nabla F(w^*)^{-1} \mathcal{T}_{\zeta_k}^{-1} F(w_k) + \alpha k \mathcal{T}_{\zeta_k} \nabla F(w^*)^{-1} \mathcal{T}_{\zeta_k}^{-1} \left( \mathcal{T}_{\zeta_k}^{-1} F(w^*) - F(w_k) \right) \mathcal{T}_{\zeta_k} \nabla F(w^*) \mathcal{T}_{\zeta_k}^{-1} \xi_k. \tag{60}$$

Let $\theta_1$ and $\theta_2$ respectively denote the last two terms of the above equality, i.e.,

$$\theta_1 := \alpha k B_{k}^{-1} \left( B_k - \mathcal{T}_{\zeta_k} \nabla F(w^*) \mathcal{T}_{\zeta_k}^{-1} \right) \mathcal{T}_{\zeta_k} \nabla F(w^*)^{-1} \mathcal{T}_{\zeta_k}^{-1} F(w_k),$$

$$\theta_2 := \alpha k \mathcal{T}_{\zeta_k} \nabla F(w^*)^{-1} \mathcal{T}_{\zeta_k}^{-1} \left( \mathcal{T}_{\zeta_k}^{-1} F(w^*) - F(w_k) \right) \mathcal{T}_{\zeta_k} \nabla F(w^*) \mathcal{T}_{\zeta_k}^{-1} \xi_k.$$
\[
\|\theta_2\| \leq \alpha_k \left\| \tilde{T}_{\xi_k} \nabla F \left( w^* \right)^{-1} \tilde{T}_{\xi_k}^{-1} \right\| \left\| \tilde{T}_{\xi_k}^{-1} F \left( w^* \right) - F(w_k) - \tilde{T}_{\xi_k} \nabla F \left( w^* \right) \tilde{T}_{\xi_k}^{-1} \xi_k \right\| \\
\leq c_3 \left\| \nabla F \left( w^* \right)^{-1} \right\| \xi_k \| d(w_k, w^*) \\
= O(d^2(w_k, w^*)) \text{ (by (iv) of Lemma 3.11).} \tag{62}
\]

Since \(d(w_{k+1}, w^*) \leq a_1 \|r\|\) by (59), it follows from (60) to (62) and (iii) of Lemma 5.1 that
\[
d(w_{k+1}, w^*) \leq a_1 (1 - \alpha_k) \left\| \xi_k \right\| + O(\mu_k) + O(d^2(w_k, w^*)) + \sqrt{\delta^2 + \delta c_1 \varepsilon + c_2 \varepsilon d(w_k, w^*)})
\]
\[
\leq \left\{ \frac{1}{1 - \epsilon} (1 - \gamma_k) + M_1 \sqrt{\delta^2 + \delta c_1 \varepsilon + c_2 \varepsilon} + M_2 \varepsilon \right\} d(w_k, w^*)
\]
where \(a_0 = 1 - \epsilon, a_1 = 1 + \epsilon\) for any \(\epsilon > 0\) by (i) of Lemma 3.11, \(M_1\) and \(M_2\) are constants. Let
\[
\psi(\bar{\epsilon}, \varepsilon, \delta) := \left\{ \frac{1 + \bar{\epsilon}}{1 - \epsilon} (1 - \gamma) + M_1 \sqrt{\delta^2 + \delta c_1 \varepsilon + c_2 \varepsilon} + M_2 \varepsilon \right\}
\]
\[
\lim_{\epsilon \to 0^+, \bar{\epsilon} \to 0^+, \delta \to 0^+} \psi(\bar{\epsilon}, \varepsilon, \delta) = 1 - \gamma < \nu.
\]
By choosing \(\bar{\epsilon}, \varepsilon,\) and \(\delta\) such that \(\psi(\bar{\epsilon}, \varepsilon, \delta) < \nu\), we obtain \(d(w_{k+1}, w^*) \leq \nu d(w_k, w^*) < \varepsilon\).

The second part of the induction uses the same technique as in [5, Theorem 3.2]. For \(i = 0, 1, \ldots, k,\)
\[
\left\| G_{i+1} - T_{c_{i+1}} \right\| \nabla \mathcal{L}(w^*) \left\| T_{c_{i+1}}^{-1} \right\| \\
\leq \beta_1 \left\| G_i - T_{c_i} \right\| \nabla \mathcal{L}(w^*) \left\| T_{c_i}^{-1} \right\| \sigma_i + \beta_2 \sigma_i \text{ (by (54))}
\]
\[
\leq (\beta_1 \delta + \beta_2) \sigma_i \text{ (by (56))}
\]
\[
\leq (\beta_1 \delta + \beta_2) \nu^i \varepsilon.
\]
By summing both sides from \(i = 0\) to \(k,\) we obtain
\[
\left\| G_{k+1} - T_{c_{k+1}} \right\| \nabla \mathcal{L}(w^*) \left\| T_{c_{k+1}}^{-1} \right\| \leq \left\| G_0 - T_{c_0} \right\| \nabla \mathcal{L}(w^*) \left\| T_{c_0}^{-1} \right\| + \frac{(\beta_1 \delta + \beta_2) \varepsilon}{1 - \nu}.
\]
By choosing \(\varepsilon\) and \(\delta\) such that \(\frac{(\beta_1 \delta + \beta_2) \varepsilon}{1 - \nu} < \frac{1}{2} \delta,\) we obtain
\[
\left\| G_{k+1} - T_{c_{k+1}} \right\| \nabla \mathcal{L}(w^*) \left\| T_{c_{k+1}}^{-1} \right\| \leq \delta.
\]
The proof is complete.
6 Local and Superlinear Convergence of quasi-Newton RIP

The well-known Dennis Moré condition [7, Theorem 2.2] has a very important place in the analyses of quasi-Newton methods in the Euclidean setting. It is a necessary and sufficient condition for superlinear convergence of the quasi-Newton method. This condition encompasses a large category of well-known update formulas, such as BFGS, DFP. Gallivan et al. [14] have generalized it to cover Riemannian quasi-Newton methods. In this section, we give an analogous Dennis Moré condition for our quasi-Newton RIP (Theorem 6.1) and establish its superlinear convergence (Theorem 6.2). On the basis of the previous section, we assume that a well-defined sequence \( \{w_k\} \) converges linearly to \( w^* \).

**Theorem 6.1** (Dennis Moré Condition for quasi-Newton RIP). Let \( \mathbb{M} \) be a Riemannian manifold endowed with an isometric vector transport \( T \) and an associated retraction \( R \). Suppose that (A1)-(A5) and (A2') hold, \( \{B_k^{-1}\} \) is a bounded sequence, and the sequence \( \{w_k\} \) generated by Algorithm 2 with equation (17) converges linearly to \( w^* \). Choose the parameters such that

\[
\mu_k = o(\|F(w_k)\|) \quad \text{and} \quad \gamma_k \to 1.
\]

Then, the following are equivalent:

(a) The sequence of linear operators \( \{B_k\} \) satisfies

\[
\lim_{k \to \infty} \frac{\|B_k - \bar{T}_{\zeta_k} \nabla F(w^*) \bar{T}_{\zeta_k}^{-1}\| w_k \| \Delta w_k \|}{\| \Delta w_k \|} = 0,
\]

where \( \zeta_k \in T_{w^*} \mathbb{M} \) is defined by \( \zeta_k = R_{w^*}^{-1}(w_k) \), i.e. \( R_{w^*}(\zeta_k) = w_k \);

(b) The sequence \( \{F(w_k)\} \) satisfies

\[
\lim_{k \to \infty} \frac{\|T_{\alpha k w_k}^{-1} F(w_{k+1})\|}{\| \alpha_k \Delta w_k \|} = 0;
\]

(c) The sequence \( \{F(w_k)\} \) satisfies

\[
\lim_{k \to \infty} \frac{\|F(w_{k+1})\|}{\| \Delta w_k \|} = 0;
\]

(d) The sequence \( \{w_k\} \) converges superlinearly to \( w^* \), i.e., \( d(w_{k+1}, w^*) / d(w_k, w^*) \to 0 \).

**Proof.** Let us show an auxiliary result:

\[
\|F(w_k)\| = O(\|\Delta w_k\|).
\]

This comes from that the sequence \( \{w_k\} \) converges linearly to the solution \( w^* \). Note that linear convergence implies, for some \( \nu \in (0,1) \),

\[
d(w_k, w^*) \leq d(w_k, w_{k+1}) + d(w_{k+1}, w^*) \leq d(w_k, w_{k+1}) + \nu d(w_k, w^*).
\]

Thus, we have

\[
d(w_k, w^*) / d(w_k, w_{k+1}) \leq 1 / (1 - \nu).
\]

Since \( w_{k+1} = \bar{R}_{w_k} (\alpha_k \Delta w_k) \), by (iii) of Lemma 3.11 we have \( d(w_k, w_{k+1}) \leq a_1 \|\alpha_k \Delta w_k\|. \) By Lemma 3.12, \( \|F(w_k)\| \leq a_3 d(w_k, w^*) \). Finally, we see that

\[
\|F(w_k)\| / \|\Delta w_k\| = \frac{\alpha_k a_1 \|F(w_k)\|}{a_1 \|\alpha_k \Delta w_k\|} \leq \frac{\alpha_k a_1 a_3 d(w_k, w^*)}{d(w_k, w_{k+1})} \leq \frac{a_1 a_3}{1 - \nu},
\]

\( 26 \)
and that \( \|F(w_k)\| = O(\|\Delta w_k\|) \).

Next, we show that
\[
\alpha_k \to 1. \quad (68)
\]

Since we have assumed that \( \{\|B_k^{-1}\|\} \) is bounded above, we have
\[
\|\Delta w_k\| = \|B_k^{-1}(-F(w_k) + \mu_k \hat{e})\| \quad (\text{by (17)})
\leq \|B_k^{-1}\| \|F(w_k)\| + \mu_k \|\hat{e}\|
= O(F(w_k)) + O(\mu_k),
= O(F(w_k)) \quad (\text{by (64) for } \mu_k),
\]
and (69) together with \( w_k \to w^* \) implies that the conditions of Lemmas 2.7 are satisfied,
\[
0 \leq 1 - \alpha_k \leq (1 - \gamma_k) + \Omega \|\Delta w_k\| = (1 - \gamma_k) + O(F(w_k)).
\]

Thus, \( \gamma_k \to 1 \) implies \( \alpha_k \to 1 \).

(a \iff b). Let (a) hold. By \( B_k \Delta w_k = \mu_k \hat{e} - F(w_k) \), we have
\[
\tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) = \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) - F(w_k) - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \alpha_k \Delta w_k
- \left( B_k - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \right) \alpha_k \Delta w_k
+ (1 - \alpha_k) F(w_k) + \alpha_k \mu_k \hat{e},
\]
where \( \zeta_k = R_{w^*}^{-1}(w_k) \). Thus, by Lemma 3.18, we see that
\[
\left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\| \leq \left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) - F(w_k) - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \alpha_k \Delta w_k \right\|
+ \left\| \left( B_k - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \right) \alpha_k \Delta w_k \right\|
+ (1 - \alpha_k) \|F(w_k)\| + \alpha_k \mu_k \|\hat{e}\|
\leq c_3 \|\alpha_k \Delta w_k\| \max\{d(w_k, w^*), d(w_k, w^*)\}
+ \left\| \left( B_k - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \right) \alpha_k \Delta w_k \right\|
+ (1 - \alpha_k) \|F(w_k)\| + \alpha_k \mu_k \|\hat{e}\|,
\]
and by dividing both sides by \( \|\alpha_k \Delta w_k\| \), we get
\[
\left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\| \|\alpha_k \Delta w_k\| \leq c_3 \max\{d(w_k, w^*), d(w_k, w^*)\}
+ \left\| \left( B_k - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \right) \alpha_k \Delta w_k \right\|
+ \left( 1 \alpha_k - 1 \right) \|F(w_k)\| \|\Delta w_k\| + \frac{\mu_k \|F(w_k)\| \|\Delta w_k\|}{\|\Delta w_k\|} \|\hat{e}\|.
\]

Note that \( \frac{\|F(w_k)\|}{\|\Delta w_k\|} \) is bounded by (67) and \( \mu_k = o(\|F(w_k)\|) \). Taking the limit of the above gives (b).

Conversely, let (b) hold. From (70), we have
\[
\left( B_k - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \right) \alpha_k \Delta w_k = \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) - F(w_k) - \tilde{T}_{\zeta_k} \nabla F(w^*) \tilde{T}_{\zeta_k}^{-1} \alpha_k \Delta w_k
- \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1})
+ (1 - \alpha_k) F(w_k) + \alpha_k \mu_k \hat{e}.
\]
Again, taking the norm and using Lemma 3.18, we get
\[
\left\| \left( B_k - \tilde{T}_{q_k} \nabla F (w^*) \tilde{T}_{q_k}^{-1} \right) \alpha_k \Delta w_k \right\| \leq \left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) - F(w_k) - \tilde{T}_{\alpha_k} \nabla F (w^*) \tilde{T}_{\alpha_k}^{-1} \alpha_k \Delta w_k \right\|
\]
\[
+ \left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\|
\]
\[
+ (1 - \alpha_k) \| F(w_k) \| + \alpha_k \mu_k \| \epsilon \|
\]
\[
\leq c_3 \| \alpha_k \Delta w_k \| \max \{d(w_k, w^*), d(w_{k+1}, w^*)\}
\]
\[
+ \left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\|
\]
\[
+ (1 - \alpha_k) \| F(w_k) \| + \alpha_k \mu_k \| \epsilon \|.
\]

Dividing both sides by \( \| \alpha_k \Delta w_k \| \) and taking the norm gives (a) for the same reason described above.

(b \( \iff \) c). This part mainly uses Lemma 3.14. Let (b) hold. Observe that
\[
\tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) = \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) - P_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) + P_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}).
\]
Thus, Lemma 3.14 and the isometry of parallel transport \( P_{\alpha_k \Delta w_k}^{-1} \) show that
\[
\left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\| \geq \left\| P_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\| \left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\|
\]
\[
\geq \left\| F(w_{k+1}) \right\| - a_4 \left\| F(w_{k+1}) \right\| \left\| \alpha_k \Delta w_k \right\|
\]
for some constant \( a_4 \), and hence,
\[
\frac{\left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\|}{\left\| \alpha_k \Delta w_k \right\|} + a_4 \left\| F(w_{k+1}) \right\| \geq \frac{\left\| F(w_{k+1}) \right\|}{\left\| \alpha_k \Delta w_k \right\|} \geq \frac{\left\| F(w_{k+1}) \right\|}{\left\| \Delta w_k \right\|}.
\]
Since \( \| F(w_k) \| \to 0 \), taking the limit of the above gives (c).

Conversely, let (c) hold. Again, (71), Lemma 3.14, and the isometry of parallel transport yield
\[
\left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\| \leq \left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) - P_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\| + \left\| P_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\|
\]
\[
\leq a_4 \left\| F(w_{k+1}) \right\| \left\| \alpha_k \Delta w_k \right\| + \left\| F(w_{k+1}) \right\|
\]
and hence,
\[
\frac{\left\| \tilde{T}_{\alpha_k \Delta w_k}^{-1} F(w_{k+1}) \right\|}{\left\| \alpha_k \Delta w_k \right\|} \leq a_4 \left\| F(w_{k+1}) \right\| + \frac{\left\| F(w_{k+1}) \right\|}{\left\| \Delta w_k \right\|} \frac{1}{\alpha_k}.
\]
Since \( \| F(w_k) \| \to 0 \) and \( \alpha_k \to 1 \), taking the limit of the above gives (b).

(c \( \iff \) d). Note that the statement that \( \{ w_k \} \) converges superlinearly to \( w^* \) can be rewritten equivalently as
\[
\lim_{k \to \infty} \frac{\| \zeta_{k+1} \|}{\| \zeta_k \|} = 0,
\]
where \( \zeta_k \in T_{w^*} \mathcal{M} \) is defined by \( \zeta_k = R_w^{-1}(w_k) \). We will show that the above equation holds if and only if (c) holds.

This part mainly uses Corollary 3.13 and (iv) of Lemma 3.11. Let (c) hold. By (iv) of Lemma 3.11, it follows that
\[
a_0 \| \alpha_k \Delta w_k \| \leq d(w_k, w_{k+1}) \leq d(w_k, w^*) + d(w_{k+1}, w^*) \leq a_1 (\| \zeta_k \| + \| \zeta_{k+1} \|),
\]
thus,
\[ \| \alpha_k \Delta w_k \| \leq \frac{a_1}{a_0} (\| \xi_k \| + \| \xi_{k+1} \|). \]  
(72)

On the other hand, by Corollary 3.13,
\[ a_4 \| \xi_{k+1} \| \leq \| F(w_{k+1}) \| \leq a_5 \| \xi_{k+1} \|. \]  
(73)

Therefore, by (72) and (73), we obtain
\[
\frac{\| F(w_{k+1}) \|}{\| \Delta w_k \|} = \alpha_k \frac{\| F(w_{k+1}) \|}{\| \alpha_k \Delta w_k \|} \geq \alpha_k \frac{a_0 a_4 \| \xi_{k+1} \|}{a_1 (\| \xi_k \| + \| \xi_{k+1} \|)} = \alpha_k \frac{a_0 a_4}{a_1} \cdot \frac{\| \xi_{k+1} \|}{(1 + \| \xi_{k+1} \| / \| \xi_k \|)}.
\]

By (68), taking the limit of the above gives (d).

Conversely, let (d) hold. Again by (iv) of Lemma 3.11, similarly to the proof of [14, Theorem 14.1 (Page 292)], we find that
\[
\| \alpha_k \Delta w_k \| \geq \frac{1}{a_1} d(w_k, w_{k+1}) \geq \frac{1}{a_1} (d(w_k, w^*) - d(w_{k+1}, w^*)) \geq \frac{a_0}{a_1} (\| \xi_k \| - \| \xi_{k+1} \|).
\]  
(74)

It follows that
\[
\frac{\| F(w_{k+1}) \|}{\| \Delta w_k \|} = \alpha_k \frac{\| F(w_{k+1}) \|}{\| \alpha_k \Delta w_k \|} \leq \alpha_k \frac{a_1^2 \| \xi_{k+1} \|}{a_0 (\| \xi_k \| - \| \xi_{k+1} \|)} = \alpha_k \frac{a_1^2}{a_0} \cdot \frac{\| \xi_{k+1} \|}{(1 - \| \xi_{k+1} \| / \| \xi_k \|)}.
\]

Taking the limit of the above gives (c). This completes the proof.

Finally, Theorem 6.2 guarantees superlinear convergence of quasi-Newton RIP and is a direct requirement for \( \{G_k\} \).

\textbf{Theorem 6.2 (Superlinear Convergence of the quasi-Newton RIP).} Suppose that (A1)-(A5) and (A2′) hold. Choose the parameter such that
\[ \mu_k = o(\| F(w_k) \|) \quad \text{and} \quad \gamma_k \to 1. \]  
(75)

Suppose that the sequence of linear operators \( \{G_k\} \) satisfies the bounded deterioration property (54). In addition, if the sequence of linear operators \( \{G_k\} \) satisfies
\[ \lim_{k \to \infty} \frac{\left\| (G_k - T_{\xi_{k}^*} \text{Hess}_x \mathcal{L}(w^*) T_{\xi_{k}^*}^{-1}) \Delta x_k \right\|}{\| \Delta x_k \|} = 0, \]  
(76)

where \( \xi_{k}^* = R_{w^*}^{-1} (x_k) \), then for Algorithm 2 with equation (17), the sequence \( \{w_k\} \) converges locally and superlinearly to \( w^* \).

\textbf{Proof.} Recall that \( \Delta w_k = (\Delta x_k, \Delta y_k, \Delta s_k, \Delta z_k) \in T_{w_k} \mathcal{M} \). By Theorem 5.2 and (54), the sequence \( \{w_k\} \) locally converges to \( w^* \) and \( \| G_k - T_{\xi_{k}^*} \text{Hess}_x \mathcal{L}(w^*) T_{\xi_{k}^*}^{-1} \| \) is bounded for all \( k \geq 0 \). From inequality (45) in the proof of Lemma 5.1, we conclude that
\[
\left\| (B_k - T_{\xi_{k}^*} \nabla F (w^*) T_{\xi_{k}^*}^{-1}) \Delta w_k \right\|^2 \leq \left\| (G_k - T_{\xi_{k}^*} \text{Hess}_x \mathcal{L}(w^*) T_{\xi_{k}^*}^{-1}) \Delta x_k \right\|^2 + \| G_k - T_{\xi_{k}^*} \text{Hess}_x \mathcal{L}(w^*) T_{\xi_{k}^*}^{-1} \| \| \Delta x_k \| O (d (w_k, w^*)) \| \Delta w_k \| + O (d (w_k, w^*)) \| \Delta w_k \|^2
\]
\[ \leq \left\| (G_k - T_{\xi_k}^z \text{Hess}_x L (w^*) T_{\xi_k}^{-1}) \Delta x_k \right\|^2 + O (d (w_k, w^*)) \left\| \Delta w_k \right\|^2. \quad (77) \]

The last inequality comes from
\[ \left\| G_k - T_{\xi_k}^z \text{Hess}_x L (w^*) T_{\xi_k}^{-1} \right\| \left\| \Delta x_k \right\| O (d(w_k, w^*)) \left\| \Delta w_k \right\| \]
\[ = \left\| \Delta x_k \right\| O (d(w_k, w^*)) \left\| \Delta w_k \right\| \text{ (by boundedness of } \left\| G_k - T_{\xi_k}^z \text{Hess}_x L (w^*) T_{\xi_k}^{-1} \right\|) \]
\[ \leq O (d(w_k, w^*)) \left\| \Delta w_k \right\|^2 \text{ (by } \left\| \Delta x_k \right\| \leq \left\| \Delta w_k \right\|). \]

Thus, by dividing both sides of (77) by \( \left\| \Delta w_k \right\|^2 \), we have
\[ \frac{\left\| (B_k - T_{\xi_k} \nabla F (w^*) T_{\xi_k}^{-1}) \Delta w_k \right\|^2}{\left\| \Delta w_k \right\|^2} = \frac{\left\| (G_k - T_{\xi_k}^z \text{Hess}_x L (w^*) T_{\xi_k}^{-1}) \Delta x_k \right\|^2}{\left\| \Delta w_k \right\|^2} + O(d(w_k, w^*)) \]
\[ \leq \frac{\left\| (G_k - T_{\xi_k}^z \text{Hess}_x L (w^*) T_{\xi_k}^{-1}) \Delta x_k \right\|^2}{\left\| \Delta x_k \right\|^2} + O(d(w_k, w^*)). \]

Taking the limit above with (76) and invoking Theorem 6.1 complete the proof. \( \square \)

7 Conclusion

We proposed a Riemannian version of the classical interior point method and established some local convergence results. To our knowledge, this is the first study to apply the primal-dual interior point method to constrained optimization on manifolds.

Acknowledgments

We are deeply grateful to Hiroyuki Sato of Kyoto University for his valuable comments and suggestions. This work was supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center, at Kyoto University, by a JSPS KAKENHI Grant, Number (B)19H02373, and by a JST SPRING Grant, Number JPMJSP2124.

A Proof of Lemma 2.2

For simplicity, we will prove the result for the case in which (RCOP) contains only inequality constraints. The same technique can be used for (RCOP). Let us consider
\[ \min_{x \in \mathbb{M}} f(x) \quad \text{s.t.} \quad c(x) \geq 0, \quad \text{(RCOP-I)} \]
where \( c : \mathbb{M} \to \mathbb{R}^m \). The KKT conditions with Lagrange multiplier \( \lambda \in \mathbb{R}^m \) are
\[ \begin{cases} \text{grad} f(x) - \sum_{i=1}^{m} \lambda_i \text{grad} c_i(x) = 0, \\ c_i(x) \lambda_i = 0, \quad i = 1, \ldots, m, \end{cases} \]
and \( c(x) \geq 0, \lambda \geq 0 \). Let \( C(x) \) be a diagonal matrix of vectors \( c(x) \); the KKT vector field \( F : \mathbb{M} \times \mathbb{R}^m \to T(\mathbb{M} \times \mathbb{R}^m) \equiv T\mathbb{M} \times T\mathbb{R}^m \) is defined by
\[ F(x, \lambda) := \left( \text{grad} f(x) - \sum_{i=1}^{m} \lambda_i \text{grad} c_i(x), C(x)\lambda \right). \quad (78) \]
The next lemma establishes a way of defining a connection on product manifolds. It can be found in [4, Exercise 5.4 & 5.13]. It can easily be extended to the product of more than two manifolds.

**Lemma A.1** (product connection [4, Exercise 5.4 & 5.13]). Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two Riemannian manifolds, respectively equipped with Riemannian connections $\nabla^1$ and $\nabla^2$. Consider the product manifold $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$. Let $(u_1, u_2)$ be tangent to $\mathcal{M}$ at $(x_1, x_2)$. Then the map $\nabla : T\mathcal{M} \times \mathfrak{X}(\mathcal{M}) \to T\mathcal{M}$ defined by

$$
\nabla_{(u_1, u_2)} (F_1, F_2) = \left( \nabla^1_{u_1} F_1(\cdot, x_2) + D F_1(x_1, \cdot)(x_2)[u_2], \right.
\nabla^2_{u_2} F_2(x_1, \cdot) + D F_2(\cdot, x_2)(x_1)[u_1] \big)
$$

is a Riemannian connection on $\mathcal{M}$; we call it the product connection. The notation $F_1(\cdot, x_2)$ represents the map obtained from $F_1 : \mathcal{M}_1 \times \mathcal{M}_2 \to T\mathcal{M}_1$ by fixing the second input to $x_2$. In particular, $F_1(\cdot, x_2)$ is a vector field on $\mathcal{M}_1$, while $F_1(x_1, \cdot)$ is a map from $\mathcal{M}_2$ to $T_{x_1} \mathcal{M}_1$.

**Proof of Lemma 2.2.** We apply the Lemma A.1 to (78) directly. $\mathbb{R}^m$ is equipped with the canonical Euclidean connection, and we will not distinguish those connections with superscripts, as they should be clear from context. Let

$$F(x, \lambda) = (F_1(x, \lambda), F_2(x, \lambda)),$$

where

$$F_1 : \mathbb{M} \times \mathbb{R}^m \to T\mathbb{M}, F_1(x, \lambda) = \text{grad } f(x) - \sum_{i=1}^{m} \lambda_i \text{grad } c_i(x),$$

$$F_2 : \mathbb{M} \times \mathbb{R}^m \to T\mathbb{R}^m, F_2(x, \lambda) = C(x)\lambda.$$

We will compute the covariant derivative of $F$ at $(x, \lambda) \in \mathbb{M} \times \mathbb{R}^m$ step by step in accordance with Lemma A.1. Let $(u_x, u_{\lambda}) \in T_{x} \mathbb{M} \times T_{x, \lambda} \mathbb{M}$ be the tangent vector at $(x, \lambda)$.

1. $F_1(\cdot, \lambda) : \mathbb{M} \to T\mathbb{M}$ (a vector field on $\mathcal{M}$). From the linearity of the connection $\nabla$ on $\mathbb{M}$, we have

$$\nabla_{u_x} F_1(\cdot, \lambda) = \nabla_{u_x} \left( \text{grad } f(x) - \sum_{i=1}^{m} \lambda_i \text{grad } c_i(x) \right) = \nabla_{u_x} \text{grad } f(x) - \sum_{i=1}^{m} \lambda_i \nabla_{u_x} \text{grad } c_i(x)$$

$$= \text{Hess } f(x)[u_x] - \sum_{i=1}^{m} \lambda_i \text{Hess } c_i(x)[u_x] = \text{Hess}_x L(x, \lambda)[u_x].$$

2. $F_1(x, \cdot) : \mathbb{R}^m \to T_{x} \mathbb{M}$ (a map between two vector spaces).

$$D F_1(x, \cdot)(\lambda)[u_{\lambda}] = \lim_{t \downarrow 0} \{ F_1(x, \lambda + tu_{\lambda}) - F_1(x, \lambda) \} / t$$

$$= \lim_{t \downarrow 0} \left\{ \sum_{i=1}^{m} \lambda_i \text{grad } c_i(x) - \sum_{i=1}^{m} [\lambda_i + t(u_{\lambda})_i] \text{grad } c_i(x) \right\} / t$$

$$= - \sum_{i=1}^{m} (u_{\lambda})_i \text{grad } c_i(x).$$

3. $F_2(x, \cdot) : \mathbb{R}^m \to T\mathbb{R}^m \equiv \mathbb{R}^m$ (a trivial function).

$$\nabla_{u_{\lambda}} F_2(x, \cdot) = D F_2(x, \cdot)[u_{\lambda}] = C(x)u_{\lambda} = \begin{pmatrix} c_1(x)(u_{\lambda})_1 \\ \vdots \\ c_m(x)(u_{\lambda})_m \end{pmatrix}.$$
4. $F_2(\cdot, \lambda) : \mathbb{M} \to T_\lambda \mathbb{R}^m \equiv \mathbb{R}^m$ (a map from $\mathbb{M}$ to $\mathbb{R}^m$). Let $F_2^i(\cdot, \lambda) = c_i(x)\lambda_i$ be the $i$th component function for $i = 1, \ldots, m$. Since

$$DF_2^i(\cdot, \lambda)(x)[u_x] = \langle \text{grad}_x F_2^i(\cdot, \lambda), u_x \rangle_x = \langle \lambda_i \text{grad} c_i(x), u_x \rangle_x,$$

we have

$$DF_2(\cdot, \lambda)(x)[u_x] = \begin{pmatrix} DF_2^1(\cdot, \lambda)(x)[u_x] \\ \vdots \\ DF_2^m(\cdot, \lambda)(x)[u_x] \end{pmatrix} = \begin{pmatrix} \lambda_1 \langle \text{grad} c_1(x), u_x \rangle_x \\ \vdots \\ \lambda_m \langle \text{grad} c_m(x), u_x \rangle_x \end{pmatrix}.$$ 

Finally, by combining 1-4, we obtain

$$\nabla F(x, \lambda) [(u_x, u_\lambda)] = \left( \text{Hess}_x L(x, \lambda)[u_x] - \sum_{i=1}^m (u_\lambda)_i \text{grad} c_i(x), \right) \begin{pmatrix} \lambda_1 \langle \text{grad} c_1(x), u_x \rangle_x + c_1(x) (u_\lambda)_1 \\ \vdots \\ \lambda_m \langle \text{grad} c_m(x), u_x \rangle_x + c_m(x) (u_\lambda)_m \end{pmatrix}.$$ 

The expression for the adjoint operator follows from the definition.

References


