A PRIMAL-DUAL INTERIOR POINT TRUST REGION METHOD FOR INEQUALITY-CONSTRAINED OPTIMIZATION PROBLEMS ON RIEMANNIAN MANIFOLDS[*]

MITSUAKI OBARA[†], TAKAYUKI OKUNO[‡], AND AKIKO TAKEDA[†§]

Abstract. We consider Riemannian inequality-constrained optimization problems and propose a Riemannian primal-dual interior point trust region method (RIPTRM) for solving them. We prove its global convergence to an approximate Karush-Kuhn-Tucker point and a second-order stationary point. We also establish the local near-quadratic convergence. To the best of our knowledge, this is the first algorithm that incorporates the trust region strategy and has the second-order convergence property for optimization problems on Riemannian manifolds with nonlinear inequality constraints. It is also the first Riemannian interior point method that possesses both global and local convergence properties. We conduct numerical experiments in which we introduce a truncated conjugate gradient method and an eigenvalue-based subsolver for RIPTRM to approximately and exactly solve the trust region subproblems, respectively. Empirical results show that RIPTRMs find solutions with higher accuracy compared to an existing Riemannian interior point method and other algorithms. Additionally, we observe that RIPTRM with the exact search direction shows significantly promising performance in an instance where the Hessian of the Lagrangian has a large negative eigenvalue.

Key words. Riemannian optimization, Inequality-constrained optimization, Interior point trust region method, Eigenvalue-based solver.

AMS subject classifications. 65K05, 90C30

1. Introduction. In this paper, we consider the following optimization problem:

$$
\begin{aligned}
&\underset{x \in \mathcal{M}}{\text{minimize}} && f(x) \\
&\text{subject to} && g_i(x) \geq 0, \quad i \in \mathcal{I} := \{1, \ldots, m\},
\end{aligned}
$$

(1.1)

where $\mathcal{M}$ is a d-dimensional, connected, complete Riemannian manifold, and $f \colon \mathcal{M} \to \mathbb{R}$ and $\{g_i\}_{i \in \mathcal{I}} \colon \mathcal{M} \to \mathbb{R}$ are twice continuously differentiable functions. If $\mathcal{M}$ is not connected, we can consider our result on the connected components of $\mathcal{M}$ separately. We call problem (1.1) the Riemannian inequality-constrained optimization problem and abbreviate it as the RICO problem. RICO is a natural extension of the standard inequality-constrained optimization problem from a Euclidean space to a Riemannian manifold. Indeed, when $\mathcal{M} = \mathbb{R}^d$, RICO (1.1) reduces to the standard problem on $\mathbb{R}^d$. Due to its versatility, RICO (1.1) has applications across various fields; for example, stable linear system identification [50], nonnegative principal component analysis [75], and robot posture computation [13].

Considering optimization on Riemannian manifolds, or Riemannian optimization, has several advantages compared with the classical Euclidean optimization. First, the Riemannian modeling ensures the feasibility for every iterate being in $\mathcal{M}$. In contrast, their feasibility of iterates and the final solution in Euclidean optimization depends on the behavior of numerical algorithm, which may lead to an infeasible output [5].

1

Second, the Riemannian approach serves as a means to handle specific structures in applications. For example, the stability and other properties of linear systems can be modeled using the set of positive-definite matrices and the geometry of quotient manifolds [50, 58, 56, 57]. Posture computation in robotics [13] and synchronization of rotation [11] reduce to optimization problems on the special orthogonal group. The flag manifold, a sequence of nested subspaces, appears in numerical PDE and statistics [74]. Third, leveraging the inherent geometry of optimization problems provides better understanding and improves computational performance. For example, the formulations of geodesically convex optimization problems, which arise in various applications [62, 61, 2, 11], guarantee that any local optimum is also a global optimum [11]. Levin et al. [43] developed lift, a tool that theoretically and practically relates Riemannian optimization to other problems. Furthermore, the relationships between Riemannian and other modelings for specific applications are also studied [46, 45, 69].

Riemannian optimization has been extensively investigated over the last two decades especially for the unconstrained case. Absil et al. [1] established the modern theory of Riemannian optimization, where they proposed the geometric Newton method and the Riemannian trust region methods. Based on their work, classical algorithms for the Euclidean optimization have been extended to the Riemannian settings with the guarantees of the global convergence to a first-order stationary point; for example, the trust region method with symmetric rank-one update [36], the quasi-Newton methods [37, 47, 53], and the nonlinear conjugate gradient methods [55]. Several algorithms, such as the cubic-regularized Newton method [4, 76], the trust region methods [12, 33, 38], and the perturbed gradient descent [20, 63], are designed to achieve the global convergence to a second-order stationary point (SOSP) and efficiently escape the saddle points. We also refer readers to the recent books [11, 54] for introduction to Riemannian optimization.

On the other hand, Riemannian optimization with nonlinear constraints is developing. Yang et al. [73] derived optimality conditions for Riemannian optimization with inequality and equality constraints. Constraint qualifications have been examined for the Riemannian cases [9, 71, 5, 6]. Liu and Boumal [44] developed an exact penalty method (EPM) combined with smoothing and an augmented Lagrangian method (ALM). They proved the global convergence of EPM to a Karush-Kuhn-Tucker (KKT) point and that of ALM to a KKT point, an SOSP in the equality-constrained case, or the global optimum depending on the subsolver's quality. Note that, however, the concrete analyses of the sequence by the subsolver are beyond the scope of their work, and computing the global optimum is generally hard due to nonconvexity. ALM has been also developed to achieve the global convergence to an approximate-KKT (AKKT) point or a positive-AKKT (PAKKT) point, both of which are first-order stationary points without constraint qualifications and coincide with a KKT point under mild constraint qualifications [71, 5, 6]. In particular, Andreani et al. [6] derived new weaker constraint qualifications and clarified the superiority of the Riemannian approach in that their Riemannian ALM has the global convergence property under these constraint qualifications. Several studies have also developed Riemannian ALM for nonsmooth [77, 25, 24] and stochastic settings [30]. Recently, EPM has been refined to guarantee global convergence to AKKT points and, under weak constraint qualifications, to KKT points [10]. Obara et al. [49] proposed Riemannian sequential quadratic optimization (SQO) for optimization problems with inequality and equality constraints, proving its global convergence to a KKT point and the local quadratic convergence to a solution. Schiela and Ortiz [59] proposed SQO for optimization problems on Hilbert manifolds with equality constraints and proved the local quadratic

convergence.

Interior point method, abbreviated as IPM, is known as one of the most prominent algorithms for optimization problems with nonlinear constraints in the Euclidean setting; we refer readers to [48, 28] for comprehensive reviews of IPMs in the Euclidean setting. Comparisons of IPM with other algorithms, including SQO, in the Euclidean setting are discussed in the survey paper by Gould et al. [32]. Interior point trust region method, or IPTRM, is a variant of IPM that uses the trust region approach as the globalization mechanism; that is, it employs the trust region scheme when a trial iterate fails to make sufficient progress toward the solution set and is rejected [19]. The strengths of IPTRM include strong convergence properties and the practicality. For instance, Wächter and Biegler [68] provided an example where several IPMs using line search fail to achieve feasibility when starting at reasonable infeasible points, while the IPTRM by Byrd et al. [14] avoids this issue. Additionally, the trust region framework allows for the use of the exact Hessian of the Lagrangian when computing the search directions by solving subproblems, whereas the line search methods cannot since the quadratic term in the subproblems need to be positive semidefinite. In the Euclidean setting, Conn et al. [18] proposed a primal-dual IPTRM for optimization problems with nonlinear inequality constraints and linear equality constraints, proving its global convergence to an SOSP. For the same problem, Gould et al. [31] provided the local superlinear convergence of IPM using an extrapolation step. Tseng [65] independently proposed an infeasible primal-dual IPTRM and proved its global convergence to an SOSP. Primal IPTRMs have been shown to enjoy the global convergence to a KKT point and local superlinear convergence for optimization problems with nonlinear inequality and equality constraints [14, 15]. Yamashita et al. [72] proposed a primal-dual IPTRM with both global and local superlinear convergence to a KKT point. Filter IPTRMs have been developed to ensure the global convergence to a KKT point [66, 60].

In the Riemannian setting, Lai and Yoshise [40] proposed two types of IPMs that have global convergence to a KKT point and the local quadratic convergence properties, respectively. Hirai et al. [35] analyzed IPM for convex optimization problems on Riemannian manifolds. Note that these IPMs are based on the line search for the globalization mechanism. To the best of our knowledge, no IPTRMs have been developed in the Riemannian setting. Building on prior studies in the Euclidean setting, the trust region approach is expected to exhibit strong convergence properties and high practicality for constrained optimization on Riemannian manifolds.

1.1. Our contribution. In this paper, we propose a primal-dual Riemannian IPTRM, abbreviated as RIPTRM, for RICO (1.1). RIPTRM consists of outer and inner iterations. In outer iteration, starting from an interior point of RICO (1.1), RIPTRM generates a sequence by adjusting the barrier parameter and the tolerance level for residuals used in the inner iterations. In the inner iteration, the algorithm computes a search direction by approximately or exactly solving a trust region subproblem defined on a tangent space of $\mathcal{M}$. The subproblem uses the full Hessian of f and $\{g_i\}_{i \in \mathcal{I}}$, which may be indefinite. Then, the search direction is evaluated to decide if the algorithm updates the iterate and adjusts the trust region radius. Here, we make use of the retraction, a smooth map for computing the next iterate in $\mathcal{M}$ according to the search direction, and the log barrier function as a merit function.

Our contributions are summarized as follows:

1. We propose RIPTRM for solving RICO (1.1). To the best of our knowledge, this is the first algorithm incorporating the trust region strategy for

TABLE 1
Comparison of algorithms for Riemannian optimization with nonlinear constraints. The symbols $\mathcal{I}$ and $\mathcal{E}$ denote the capability of handling inequality and equality constraints, respectively. The symbols ✔ and ✔ represent the global convergence to an AKKT point and a PAKKT point, respectively.

Reference	Method	Constraints	KKT	SOSP	Local
[44]	ALM	$\mathcal{I}, \mathcal{E}$	✓	✓[†]	
[44]	EPM	$\mathcal{I}, \mathcal{E}$	✓		
[59]	SQO	$\mathcal{E}$			quadratic
[49]	SQO	$\mathcal{I}, \mathcal{E}$	✓		quadratic
[71]	ALM	$\mathcal{I}, \mathcal{E}$	✔		
[40]	IPM	$\mathcal{I}, \mathcal{E}$			quadratic
[40]	IPM	$\mathcal{I}, \mathcal{E}$	✓		
[5, 6]	ALM	$\mathcal{I}, \mathcal{E}$	✔		
[10]	EPM	$\mathcal{I}, \mathcal{E}$	✔		
Ours	IPTRM	$\mathcal{I}$	✔	✓	near-quadratic

† ALM in [44] has the second-order convergence property under only equality-constrained setting.

    optimization problems on Riemannian manifolds with nonlinear constraints.
2. We prove that our RIPTRM achieves the global convergence to an AKKT point and an SOSP, depending on the chosen quality of the search direction. Additionally, we prove the local near-quadratic convergence of RIPTRM under standard assumptions. To the best of our knowledge, this is the first algorithm achieving the second-order global convergence property for Riemannian optimization with inequality constraints and also the first Riemannian IPM with both global and local convergence. We summarize the comparison of our algorithm with the existing ones for Riemannian optimization with nonlinear constraints in Table 1.
3. We conduct numerical experiments in which we introduce an truncated conjugate gradient (tCG) method and eigenvalue-based subsolver [3] for RIPTRM to approximately and exactly solve the trust region subproblems, respectively. Empirical results show that RIPTRMs find solutions with higher accuracy compared to an existing Riemannian IPM and other algorithms. Additionally, we observe that RIPTRM with the exact search direction shows significantly promising performance in an instance where the Hessian of the Lagrangian has a large negative eigenvalue.

1.2. Paper organization. The rest of this paper is organized as follows. In Section 2, we review fundamental concepts from Riemannian geometry and Riemannian optimization. In Section 3, we describe RIPTRM consisting of outer and inner iterations. We prove its global and local convergence properties in Sections 4 and 5, respectively. In Section 6, we provide numerical experiments on the stable linear system identification and the Rosenbrock function minimization and compare our algorithm with the existing methods. In Section 7, we summarize our research and state future work.

2. Preliminaries and auxiliary results. Define $\mathbb{N}_0 \coloneqq \{0, 1, 2, \dots\}$ and let $\mathbb{R}^d, \mathbb{R}^d_+$, and $\mathbb{R}^d_{++}$ be d-dimensional Euclidean space, its nonnegative orthant, and its positive orthant, respectively. We denote by $\mathbf{1}_m$ the m-dimensional vector of ones. We omit the subscript m when the context is clear. A continuous function $\varepsilon \colon \mathbb{R}_+ \to \mathbb{R}_+$

is said to be a forcing function if $\varepsilon(\mu) = 0$ holds if and only if $\mu = 0$. For related positive quantities α and β, we write $\alpha = \mathcal{O}(\beta)$ if there exists a constant $c > 0$ such that $\alpha \leq c\beta$ for all β sufficiently small. We write $\alpha = o(\beta)$ if $\lim_{\beta \to 0} \frac{\alpha}{\beta} = 0$ holds. We also write $\alpha = \Omega(\beta)$ if $\beta = \mathcal{O}(\alpha)$, and $\alpha = \Theta(\beta)$ if $\alpha = \mathcal{O}(\beta)$ and $\beta = \mathcal{O}(\alpha)$. Given two normed vector spaces $\mathcal{E}, \mathcal{V}$ and a linear operator $\mathcal{A} \colon \mathcal{E} \to \mathcal{V}$, we define the operator norm as $\|\mathcal{A}\|_{\mathrm{op}} := \sup\{\|\mathcal{A}v\|_{\mathcal{V}} \colon v \in \mathcal{E} \text{ and } \|v\|_{\mathcal{E}} \leq 1\}$, where $\|\cdot\|_{\mathcal{E}}$ and $\|\cdot\|_{\mathcal{V}}$ are the norms on $\mathcal{E}$ and $\mathcal{V}$, respectively.

2.1. Notation and terminology from Riemannian geometry. We briefly review some concepts from Riemannian geometry by following the notation of [1, 11]. Let $x \in \mathcal{M}$, and let $T_x\mathcal{M}$ be the tangent space to $\mathcal{M}$ at x. We denote by $(\mathcal{U}, \varphi)$ a chart of $\mathcal{M}$. Here, $\mathcal{U} \subseteq \mathcal{M}$ is an open set, and $\varphi \colon \mathcal{U} \to \varphi(\mathcal{U}) \subseteq \mathbb{R}^d$ is a homeomorphism. Given a chart $(\mathcal{U}, \varphi)$ with $x \in \mathcal{U}$, we write the coordinate expressions as

$$\overline{x} := \varphi(x), \overline{\xi} := \mathrm{D}\varphi(x)[\xi], \text{ and } \overline{\theta} := \theta \circ \varphi^{-1}$$

for any $\xi \in T\mathcal{M}$ and $\theta \colon \mathcal{M} \to \mathbb{R}$. A vector field on $\mathcal{M}$ is a map $V \colon \mathcal{M} \to T\mathcal{M}$ with $V(x) \in T_x\mathcal{M}$, where $T\mathcal{M}$ is the tangent bundle. Let $I \subseteq \mathbb{R}$ be an open interval, and let $c \colon I \to \mathcal{M}$ be a smooth curve. A Riemannian metric on $\mathcal{M}$ is a choice of inner product $\langle \cdot, \cdot \rangle_x \colon T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ for every $x \in \mathcal{M}$ satisfying that, for all smooth vector fields V, W on $\mathcal{M}$, the function $x \mapsto \langle V(x), W(x) \rangle_x$ is smooth from $\mathcal{M}$ to $\mathbb{R}$. A Riemannian manifold is a smooth manifold endowed with a Riemannian metric. The Riemannian metric induces the norm $\|\xi_x\|_x := \sqrt{\langle \xi_x, \xi_x \rangle_x}$ for $\xi_x \in T_x\mathcal{M}$ and the Riemannian distance $\mathrm{dist}(\cdot, \cdot) \colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. It follows from [41, Theorem 13.29] that $\mathcal{M}$ is a metric space under the Riemannian distance. According to the Hopf-Rinow theorem, every closed bounded subset of $\mathcal{M}$ is compact for a finite-dimensional, connected, complete Riemannian manifold by regarding $\mathcal{M}$ as a metric space [26, Chapter 7, Theorem 2.8]. The following lemma shows the local equivalence of the Riemannian distance and the Euclidean distance in $\varphi(\mathcal{U})$.

LEMMA 2.1 ([29, Lemma 14.1]). *Let $\mathcal{M}$ be a Riemannian manifold, $(\mathcal{U}, \varphi)$ be a chart, and $\mathcal{Q} \subseteq \mathcal{U}$ be a compact subset, respectively. Then, there exist $a_1, a_2 \in \mathbb{R}_{++}$ such that, for all $x_1, x_2 \in \mathcal{Q}$,*

$$a_1 \|\overline{x_1} - \overline{x_2}\| \leq \mathrm{dist}(x_1, x_2) \leq a_2 \|\overline{x_1} - \overline{x_2}\|,$$

where $\|\cdot\|$ denotes the Euclidean norm.

For two smooth manifolds $\mathcal{M}_1, \mathcal{M}_2$ and a differentiable map $F \colon \mathcal{M}_1 \to \mathcal{M}_2$, we denote by $\mathrm{D}F(x) \colon T_x\mathcal{M}_1 \to T_{F(x)}\mathcal{M}_2$ the differential of F at $x \in \mathcal{M}_1$. We use the canonical identification $T_x\mathcal{E} \simeq \mathcal{E}$ for a vector space $\mathcal{E}$ and $x \in \mathcal{E}$. Let $\mathfrak{F}(\mathcal{M})$ denote the set of sufficiently differentiable real-valued functions. Given $\theta \in \mathfrak{F}(\mathcal{M})$, $\mathrm{D}\theta(x)[\xi_x] \in T_{\theta(x)}\mathbb{R} \simeq \mathbb{R}$ is the differential of θ at $x \in \mathcal{M}$ along $\xi_x \in T_x\mathcal{M}$. The Riemannian gradient of θ at x, denoted by $\mathrm{grad}\theta(x)$, is defined as a unique element of $T_x\mathcal{M}$ that satisfies

$$(2.1) \qquad \langle \mathrm{grad}\theta(x), \xi_x \rangle_x = \mathrm{D}\theta(x)[\xi_x]$$

for any $\xi_x \in T_x\mathcal{M}$. Here, $\mathrm{grad}\theta \colon \mathcal{M} \to T\mathcal{M} \colon x \mapsto \mathrm{grad}\theta(x)$ is the gradient vector field. Let ∇ be the Levi-Civita connection and $\mathfrak{X}(\mathcal{M})$ be the set of sufficiently smooth vector fields. For any $V \in \mathfrak{X}(\mathcal{M})$, we define the Jacobian of V at x as $\mathrm{J}_V(x) \colon T_x\mathcal{M} \to T_x\mathcal{M} \colon \xi_x \mapsto \nabla_{\xi_x} V$. The nonsingularity and the boundedness of the Jacobian around a nonsingular point is shown in the following lemma:

LEMMA 2.2 ([27, Lemma 3.2], [40, Lemma 3.13]). *Let $V \in \mathfrak{X}(\mathcal{M})$. If J_V is continuous at $x^* \in \mathcal{M}$ and $\mathrm{J}_V(x^*)$ is nonsingular, then there exist $r > 0$ and a neighborhood $\mathcal{U} \subseteq \mathcal{M}$ of x^* such that, for all $x \in \mathcal{U}$, $\mathrm{J}_V(x)$ is nonsingular and $\|\mathrm{J}_V(x)\|_{\mathrm{op}} \leq r$.*

In particular, for the case $V = \mathrm{grad}\theta$, the operator $\mathrm{Hess}\theta(x) \colon T_x\mathcal{M} \to T_x\mathcal{M}$ denotes the Riemannian Hessian of θ at x; that is, $\mathrm{Hess}\theta(x)[\xi_x] := \nabla_{\xi_x}\mathrm{grad}\theta$ for all $\xi_x \in T_x\mathcal{M}$. When $\mathcal{M}$ is a Euclidean space, we have $\mathrm{Hess}\theta(x)[\xi_x] = \mathrm{D}(\mathrm{grad}\theta)(x)[\xi_x]$ and $\langle \mathrm{Hess}\theta(x)[\xi_x^1], \xi_x^2 \rangle_x = \mathrm{D}^2\theta(x)[\xi_x^1, \xi_x^2]$ for all $\xi_x, \xi_x^1, \xi_x^2 \in T_x\mathcal{M}$.

For each $x \in \mathcal{M}$, let $\mathrm{Exp}_x \colon T_x\mathcal{M} \to \mathcal{M}$ denote the exponential map at x, the map satisfying that $t \mapsto \mathrm{Exp}_x(t\xi_x)$ is the unique geodesic passing through x with velocity $\xi_x \in T_x\mathcal{M}$ at $t = 0$. The injectivity radius at x is defined as

$$\mathrm{inj}(x) := \sup\left\{ r > 0 \colon \mathrm{Exp}_x\big|_{\{\xi_x \in T_x\mathcal{M} \colon \|\xi_x\|_x < r\}} \text{ is a diffeomorphism} \right\}.$$

Note that $\mathrm{inj}(x) > 0$ for any $x \in \mathcal{M}$. For any $x_1, x_2 \in \mathcal{M}$ with $\mathrm{dist}(x_1, x_2) < \mathrm{inj}(x_1)$, there is a unique minimizing geodesic connecting x_1 and x_2. Given a smooth curve c connecting $x_1, x_2 \in \mathcal{M}$, we denote by $\mathrm{PT}^c_{x_2 \leftarrow x_1} \colon T_{x_1}\mathcal{M} \to T_{x_2}\mathcal{M}$ the parallel transport of $T_{x_1}\mathcal{M}$ to $T_{x_2}\mathcal{M}$. We omit the superscript c when $\mathrm{dist}(x_1, x_2) < \mathrm{inj}(x_1)$ holds, and we use the unique minimizing geodesic as the curve c. Note that the parallel transport is isometric, i.e., $\|\mathrm{PT}_{x_2 \leftarrow x_1}[\xi_{x_1}]\|_{x_2} = \|\xi_{x_1}\|_{x_1}$ for any $\xi_{x_1} \in T_{x_1}\mathcal{M}$ and that $\mathrm{PT}_{x \leftarrow x} \colon T_x\mathcal{M} \to T_x\mathcal{M}$ is the identity map. The adjoint of the parallel transport corresponds with its inverse, that is, $\langle \mathrm{PT}_{x_2 \leftarrow x_1}[\xi_{x_1}], \zeta_{x_2} \rangle_{x_2} = \langle \xi_{x_1}, \mathrm{PT}_{x_1 \leftarrow x_2}[\zeta_{x_2}] \rangle_{x_1}$ for any $\xi_{x_1} \in T_{x_1}\mathcal{M}$ and any $\zeta_{x_2} \in T_{x_2}\mathcal{M}$.

A retraction $\mathrm{R} \colon T\mathcal{M} \to \mathcal{M}$ is a smooth map with the following properties: by letting $\mathrm{R}_x \colon T_x\mathcal{M} \to \mathcal{M}$ be the restriction of R to $T_x\mathcal{M}$, it satisfies

$$(2.2a) \qquad\qquad\qquad \mathrm{R}_x(0_x) = x,$$
$$(2.2b) \qquad\qquad\qquad \mathrm{D}\,\mathrm{R}_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}$$

under $T_{0_x}(T_x\mathcal{M}) \simeq T_x\mathcal{M}$, where 0_x is the zero vector of $T_x\mathcal{M}$ and $\mathrm{id}_{T_x\mathcal{M}}$ denotes the identity map on $T_x\mathcal{M}$. Let

$$(2.3) \qquad\qquad\qquad \hat{\theta} := \theta \circ \mathrm{R} \ \text{ and } \ \hat{\theta}_x := \theta \circ \mathrm{R}_x$$

denote the pullback of the function $\theta \colon \mathcal{M} \to \mathbb{R}$ and the restriction of $\hat{\theta}$ to $T_x\mathcal{M}$, respectively. Note that, it follows from (2.2) that

$$(2.4) \qquad\qquad\qquad \mathrm{grad}\hat{\theta}(0_x) = \mathrm{grad}\theta(x).$$

The equivalence of the Riemannian norm between two tangent vectors and the Riemannian distance between their retracted points is shown as follows:

LEMMA 2.3 ([36, Lemma 2]). *Let $\mathcal{M}$ be a Riemannian manifold endowed with a retraction R and let $x_1 \in \mathcal{M}$. There exist positive scalars $a_1, a_2, \delta_{a_1,a_2} \in \mathbb{R}_{++}$ such that, for all x_2 in a sufficiently small neighborhood of x_1 and all $\xi_{x_2}, \zeta_{x_2} \in T_{x_2}\mathcal{M}$ with $\|\xi_{x_2}\|_{x_2} \leq \delta_{a_1,a_2}$ and $\|\zeta_{x_2}\|_{x_2} \leq \delta_{a_1,a_2}$,*

$$a_1\|\xi_{x_2} - \zeta_{x_2}\|_{x_2} \leq \mathrm{dist}(\mathrm{R}_{x_2}(\xi_{x_2}), \mathrm{R}_{x_2}(\zeta_{x_2})) \leq a_2\|\xi_{x_2} - \zeta_{x_2}\|_{x_2}.$$

A second-order retraction is a retraction satisfying that, for all $x \in \mathcal{M}$ and all $\xi_x \in T_x\mathcal{M}$, the curve $c(t) = \mathrm{R}_x(t\xi_x)$ has zero acceleration at $t = 0$. Second-order

retractions are not a restrictive class. For example, metric projection retractions and the exponential maps meet the condition; see [11, Section 5.12] and [1, Section 5.5] for details. Second-order retractions have the following property:

PROPOSITION 2.4 ([1, Proposition 5.5.5]). *If the retraction* R *is second order, then, for any* $\theta \in \mathfrak{F}(\mathcal{M})$,

$$\text{(2.5)} \qquad \qquad \text{Hess}\hat{\theta}_x(0_x) = \text{Hess}\theta(x),$$

where the left-hand side is the Hessian of $\hat{\theta}_x \colon T_x\mathcal{M} \to \mathbb{R}$ *at* $0_x \in T_x\mathcal{M}$.

We say that $\hat{\theta}$ is radially Lipschitz continuously differentiable (radially L-C^1) on $\mathcal{U} \subseteq \mathcal{M}$ if there exist positive scalars $\beta_{RL}^\theta, \delta_{RL}^\theta \in \mathbb{R}_{++}$ such that, for any $x \in \mathcal{U}$ and all $t \geq 0, \xi_x \in T_x\mathcal{M}$ with $t\|\xi_x\|_x \leq \delta_{RL}^\theta$, it holds that

$$\text{(2.6)} \qquad \left| \left\langle \text{grad}\hat{\theta}_x(t\xi_x) - \text{grad}\hat{\theta}_x(0_x), \xi_x \right\rangle_x \right| \leq \beta_{RL}^\theta t\|\xi_x\|_x^2.$$

Note that the definition above is equivalent to that in [1, Definition 7.4.1] with the trivial case of $\xi_x = 0_x$. Similarly, we say that $\hat{\theta}$ is radially Lipschitz twice continuously differentiable (radially L-C^2) on $\mathcal{U} \subseteq \mathcal{M}$ if there exist positive scalars $\beta_{RL2}^\theta, \delta_{RL2}^\theta \in \mathbb{R}_{++}$ such that, for any $x \in \mathcal{U}$ and all $t \geq 0, \xi_x \in T_x\mathcal{M}$ with $t\|\xi_x\|_x \leq \delta_{RL2}^\theta$, it holds that

$$\text{(2.7)} \qquad \left| \left\langle \left(\text{Hess}\hat{\theta}_x(t\xi_x) - \text{Hess}\hat{\theta}_x(0_x) \right)[\xi_x], \xi_x \right\rangle_x \right| \leq \beta_{RL2}^\theta t\|\xi_x\|_x^3.$$

under $T_{t\xi_x}(T_x\mathcal{M}) \simeq T_x\mathcal{M}$ and $T_{0_x}(T_x\mathcal{M}) \simeq T_x\mathcal{M}$. The following lemma provides sufficient conditions for the radially L-C^1 and radially L-C^2 properties:

LEMMA 2.5. *Let* $\mathcal{U}$ *be a compact subset of* $\mathcal{M}$ *and* $\theta \colon \mathcal{M} \to \mathbb{R}$. *If* θ *is of class* C^2, *then* θ *is radially L-C^1 on* $\mathcal{U}$. *Additionally, if* θ *is of class* C^3, *then* θ *is radially L-C^2 on* $\mathcal{U}$.

Proof. See Appendix C.1. □

Let us end the subsection by introducing the continuity of the Riemannian Hessian combined with the parallel transport in the following lemma:

LEMMA 2.6. *Let* $\theta \colon \mathcal{M} \to \mathbb{R}$ *be of class* C^3. *Given* $x^* \in \mathcal{M}$, *there exist* $s_\theta > 0$ *and a closed neighborhood* $\mathcal{P}_\theta \subseteq \mathcal{M}$ *of* x^* *such that, for any* $x \in \mathcal{P}_\theta$,

$$\left\| \text{Hess}\theta(x^*) - \text{PT}_{x^* \leftarrow x} \circ \text{Hess}\theta(x) \circ \text{PT}_{x \leftarrow x^*} \right\|_{\text{op}} \leq s_\theta \, \text{dist}(x, x^*).$$

Proof. See Appendix C.1. □

2.2. Optimality conditions for RICO. Define the Lagrangian of RICO (1.1) as

$$\mathcal{L}(\omega) \coloneqq f(x) - \sum_{i \in \mathcal{I}} y_i g_i(x),$$

where $\omega \coloneqq (x, y) \in \mathcal{M} \times \mathbb{R}^m$ and $y \in \mathbb{R}^m$ is the vector of Lagrange multipliers for the inequality constraints. The Riemannian gradient and the Riemannian Hessian of the Lagrangian with respect to $x \in \mathcal{M}$ are represented as

$$\text{(2.8)} \qquad \text{grad}_x\mathcal{L}(\omega) = \text{grad}f(x) - \sum_{i \in \mathcal{I}} y_i\text{grad}g_i(x),$$

$$\text{Hess}_x\mathcal{L}(\omega) = \text{Hess}f(x) - \sum_{i \in \mathcal{I}} y_i\text{Hess}g_i(x),$$

respectively. For each $(x, y) \in \mathcal{M} \times \mathbb{R}^m$, any $v \in \mathbb{R}^m$, and all $\xi_x \in T_x \mathcal{M}$, we define the following maps:

$$G(x) := \operatorname{diag}(g(x)) \in \mathbb{R}^{m \times m}, \quad Y := \operatorname{diag}(y) \in \mathbb{R}^{m \times m},$$

$$\mathcal{G}_x[v] := \sum_{i \in \mathcal{I}} v_i \operatorname{grad} g_i(x) \in T_x \mathcal{M},$$

$$\mathcal{G}_x^*[\xi_x] := \left(\langle \operatorname{grad} g_1(x), \xi_x \rangle_x, \dots, \langle \operatorname{grad} g_m(x), \xi_x \rangle_x \right)^\top \in \mathbb{R}^m,$$

where $\operatorname{diag} : \mathbb{R}^m \to \mathbb{R}^{m \times m}$ is the diagonal operator. Let

$$\mathcal{F} := \{x \in \mathcal{M} | g_i(x) \geq 0 \text{ for all } i \in \mathcal{I}\} \text{ and } \operatorname{str} \mathcal{F} := \{x \in \mathcal{M} | g_i(x) > 0 \text{ for all } i \in \mathcal{I}\}$$

be the feasible region and the strictly feasible region of RICO (1.1), respectively. We define the index set of active inequalities at $x \in \mathcal{M}$ as

$$(2.9) \qquad \mathcal{A}(x) := \{i \in \mathcal{I} | g_i(x) = 0\}.$$

We now introduce the optimality conditions and related concepts.

DEFINITION 2.7 ([73, Equation (4.3)]). *The linear independence constraint qualification (LICQ) holds at* $x^* \in \mathcal{M}$ *if* $\{\operatorname{grad} g_i(x^*)\}_{i \in \mathcal{A}(x^*)}$ *is linearly independent on* $T_{x^*} \mathcal{M}$.

THEOREM 2.8 ([73, Theorem 4.1]). *Suppose that* $x^* \in \mathcal{M}$ *is a local minimum of RICO* (1.1) *and the LICQ holds at* x^*. *Then, there exist a vector of Lagrange multipliers for the inequality constraints* $y^* \in \mathbb{R}^m$ *such that the following hold:*

$$(2.10) \qquad \operatorname{grad}_x \mathcal{L}(\omega^*) = 0_{x^*}, y^* \geq 0, g(x^*) \geq 0, y_i^* g_i(x^*) = 0 \text{ for all } i \in \mathcal{I}.$$

We call (2.10) *the KKT conditions of RICO* (1.1) *and* x^* *a KKT point of RICO* (1.1).

We also introduce a sequential optimality condition that works without any constraint qualification:

THEOREM 2.9 ([71, Definition 5, Theorem 1]). *Suppose that* $x^* \in \mathcal{M}$ *is a local minimum of RICO* (1.1). *Then, there exist sequences* $\{x_k\}_k \subseteq \mathcal{M}$ *and* $\{y_k\}_k \subseteq \mathbb{R}_+^m$ *such that*

$$(2.11) \qquad \lim_{k \to \infty} x_k = x^*, \lim_{k \to \infty} \operatorname{grad}_x \mathcal{L}(\omega_k) = 0, \lim_{k \to \infty} \sum_{i=1}^m \max\{y_{ki} g_i(x_k), 0\} = 0,$$

where $\omega_k := (x_k, y_k)$. *We call* (2.11) *the AKKT conditions of RICO* (1.1) *and* x^* *satisfying the AKKT conditions an AKKT point of RICO* (1.1).

PROPOSITION 2.10 ([71, Theorem 2]). *Suppose that* $x^* \in \mathcal{M}$ *is an arbitrary point satisfying the LICQ. Then, the following two statements are equivalent:*
1. x^* *is a KKT point of RICO* (1.1).
2. x^* *is an AKKT point of RICO* (1.1).

DEFINITION 2.11 ([49, Definition 2.5]). *Given* $x^* \in \mathcal{F}$ *satisfying the KKT conditions with associated Lagrange multipliers* $y^* \in \mathbb{R}_+^m$, *we say that the strict complementarity condition (SC) holds if exactly one of* y_i^* *and* $g_i(x^*)$ *is zero for each index* $i \in \mathcal{I}$. *Hence, we have* $y_i^* > 0$ *for every* $i \in \mathcal{A}(x^*)$.

We define the critical cone associated with $\omega^* := (x^*, y^*) \in \mathcal{M} \times \mathbb{R}_+^m$ as

(2.12)
$$\mathcal{C}(\omega^*) := \left\{ \xi_{x^*} \in T_{x^*}\mathcal{M} \left| \begin{array}{l} \langle \mathrm{grad}\, g_i(x^*), \xi_{x^*} \rangle_{x^*} = 0 \text{ for all } i \in \mathcal{A}(x^*) \text{ with } y_i^* > 0, \\ \langle \mathrm{grad}\, g_i(x^*), \xi_{x^*} \rangle_{x^*} \geq 0 \text{ for all } i \in \mathcal{A}(x^*) \text{ with } y_i^* = 0 \end{array} \right. \right\}.$$

THEOREM 2.12 ([73, Theorem 4.3]). *Let $x^* \in \mathcal{F}$ be a KKT point with associated Lagrange multipliers $y^* \in \mathbb{R}^m$. Suppose that*

(2.13)
$$\langle \mathrm{Hess}_x \mathcal{L}(\omega^*)[\xi_{x^*}], \xi_{x^*} \rangle_{x^*} > 0 \text{ for all } \xi_{x^*} \in \mathcal{C}(\omega^*) \backslash \{0_{x^*}\}$$

holds. Then, x^* is a strict local minimum of RICO (1.1). We call (2.13) the second-order sufficient condition (SOSC).

THEOREM 2.13 ([73, Theorem 4.2]). *Suppose that $x^* \in \mathcal{M}$ is a local minimum of RICO (1.1) and the LICQ holds at x^*. Let $y^* \in \mathbb{R}^m$ be the vector of Lagrange multipliers for the inequality constraints. Then,*

(2.14)
$$\langle \mathrm{Hess}_x \mathcal{L}(\omega^*)[\xi_{x^*}], \xi_{x^*} \rangle_{x^*} \geq 0 \text{ for all } \xi_{x^*} \in \mathcal{C}(\omega^*).$$

We call (2.14) the second-order necessary condition and such x^* an SOSP.

We next define the weak second-order necessary condition as

(2.15)
$$\langle \mathrm{Hess}_x \mathcal{L}(\omega^*)[\xi_{x^*}], \xi_{x^*} \rangle_{x^*} \geq 0 \text{ for all } \xi_{x^*} \in \mathcal{C}_{\mathrm{w}}(x^*),$$

where

(2.16)
$$\mathcal{C}_{\mathrm{w}}(x^*) := \left\{ \xi_{x^*} \in T_{x^*}\mathcal{M} | \langle \mathrm{grad}\, g_i(x^*), \xi_{x^*} \rangle_{x^*} = 0 \text{ for all } i \in \mathcal{A}(x^*) \right\}.$$

We call such a point $x^* \in \mathcal{F}$ a weak second-order stationary point (w-SOSP) [44, Definition 3.3]. By definition, an SOSP is also a w-SOSP. The converse is not necessarily true. Under the strict complementarity condition, however, these two points are identical.

PROPOSITION 2.14. *Let $x^* \in \mathcal{M}$ be a w-SOSP of RICO (1.1). Suppose that the SC holds at x^*. Then, x^* is an SOSP of RICO (1.1).*

Proof. It follows from the SC that the i-th Lagrange multiplier y_i^* is positive for every index $i \in \mathcal{A}(x^*)$. Therefore, $\mathcal{C} = \mathcal{C}_{\mathrm{w}}$ holds, which completes the proof. □

3. Proposed method. In this section, we propose RIPTRM for solving RICO (1.1). First, we introduce the KKT vector field, which was originally proposed by [40, Section 2.2]. The vector field will be the basis of the subproblems to be solved in our algorithm. Then, we propose RIPTRM, which consists of outer and inner iterations, and characterize three search directions obtained by approximately or exactly solving the subproblems in the inner iterations.

3.1. Overview of RIPTRM. Let $\mu > 0$ be a barrier parameter. We define the barrier KKT vector field as

(3.1)
$$\Psi(\omega; \mu) := \begin{bmatrix} \mathrm{grad}_x \mathcal{L}(\omega) \\ Yg(x) - \mu\mathbf{1} \end{bmatrix}.$$

This vector field originates from the KKT conditions (2.10); it consists of the Riemannian gradient of the Lagrangian and the complementarity condition relaxed by

the barrier parameter. With μ replaced by zero, equation (3.1) coincides with the KKT conditions (2.10) for $\omega \in \mathcal{F} \times \mathbb{R}_+^m$. For brevity, we often write $\Psi(\omega)$ for $\Psi(\omega; 0)$.

RIPTRM aims to find the solution of $\Psi(\omega) = 0$ with $\omega \in \mathcal{F} \times \mathbb{R}_+^m$ by approximately solving

$$(3.2) \qquad \Psi(\omega; \mu_k) = 0, \quad \omega \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$$

for a sequence of barrier parameters $\{\mu_k\}_k \subset \mathbb{R}_{++}$ that converges to zero. Hereafter, for $\omega = (x, y)$, we refer to x as the primal variable and y as the dual variable, respectively. Let $\omega_k := (x_k, y_k) \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$ denote an approximate solution of (3.2). We call the sequence $\{\omega_k\}_k$ the outer iterates. The associated adjustment of the barrier parameter and tolerances for residuals defines the outer iteration, whose index is the subscript $k \in \mathbb{N}_0$. Each ω_k is an output of the inner iterations, where a corresponding sequence of inner iterates is indexed by the superscript $\ell \in \mathbb{N}_0$. At each inner iteration, the algorithm approximately solves (3.2) with $\mu_k > 0$ fixed while keeping $\omega^\ell := (x^\ell, y^\ell) \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$. In the following, we explain the details of the outer and inner iterations in turn.

3.2. Outer iteration. In this subsection, we present the specific outer process of the proposed algorithm. Let $\varepsilon_{\mathcal{L}}, \varepsilon_C, \varepsilon_S \colon \mathbb{R}_+ \to \mathbb{R}_+$ be forcing functions. At the k-th outer iteration, we find a point $\omega_{k+1} \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$ that satisfies the following stopping conditions with $\mu_k > 0$:

$$(3.3a) \qquad \left\| \operatorname{grad}_x \mathcal{L}(\omega_{k+1}) \right\|_{x_{k+1}} \leq \varepsilon_{\mathcal{L}}(\mu_k),$$

$$(3.3b) \qquad \left\| Y_{k+1} g(x_{k+1}) - \mu_k \mathbf{1} \right\| \leq \varepsilon_C(\mu_k),$$

$$(3.3c) \qquad \lambda^{\min}\!\left[H_{k+1} \right] \geq -\varepsilon_S(\mu_k),$$

$$(3.3d) \qquad g(x_{k+1}) > 0, \text{ and } y_{k+1} > 0,$$

where we write Y_{k+1} and H_{k+1} for $\operatorname{diag}(y_{k+1})$ and $H(\omega_{k+1})$, as defined later in (3.8), respectively. If our goal is to compute not an SOSP but an AKKT point, we can remove (3.3c) from the stopping conditions. To gain such ω_{k+1}, for example, we use Algorithm 3.2 with an initial point $\omega_k \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$, the k-th barrier parameter $\mu_k > 0$, and the k-th initial trust region radius $\Delta_k^{\mathrm{init}} \in (0, \Delta_{\max}]$. We also store the final trust region radius $\Delta_k^{\mathrm{final}}$ used in the inner iteration. Then, the algorithm defines the next initial trust region radius $\Delta_{k+1}^{\mathrm{init}} := \max\!\left(\Delta_k^{\mathrm{final}}, \Delta_{\min}^{\mathrm{init}} \right)$, where $\Delta_{\min}^{\mathrm{init}} > 0$ is a predefined parameter, and the next barrier parameter $\mu_{k+1} > 0$ so that $\lim_{k \to \infty} \mu_k = 0$. The outer iteration is formally presented as Algorithm 3.1.

3.3. Inner iteration. To approximately solve (3.2) with $\mu = \mu_k > 0$ fixed, we apply the Riemannian Newton method to (3.2): for $\omega = (x, y) \in \operatorname{str} \mathcal{F} \times \mathbb{R}^m$, the Newton equation is

$$(3.4) \qquad \mathrm{J}_\Psi(\omega)[d_\omega] = -\Psi(\omega; \mu),$$

where, under $T_y \mathbb{R}^m \simeq \mathbb{R}^m$,

$$(3.5) \qquad \begin{aligned} \mathrm{J}_\Psi(\omega) \colon &T_x \mathcal{M} \times \mathbb{R}^m \to T_x \mathcal{M} \times \mathbb{R}^m \\ d_\omega := (d_x, d_y) \mapsto &\begin{bmatrix} \operatorname{Hess}_x \mathcal{L}(\omega)[d_x] - \mathcal{G}_x[d_y] \\ Y \mathcal{G}_x^*[d_x] + G(x) d_y \end{bmatrix} \end{aligned}$$

Algorithm 3.1 Outer iteration of RIPTRM

Require: Riemannian manifold $\mathcal{M}$, twice continuously differentiable functions $f \colon$
$\mathcal{M} \to \mathbb{R}$ and $\{g_i\}_{i \in \mathcal{I}} \colon \mathcal{M} \to \mathbb{R}$, maximal trust region radius $\Delta_{\max} > 0$,
initial trust region radius $\Delta_0^{\mathrm{init}} \in (0, \Delta_{\max}]$, minimum initial trust re-
gion radius $\Delta_{\min}^{\mathrm{init}} \in (0, \Delta_{\max}]$, forcing functions $\varepsilon_{\mathcal{L}}, \varepsilon_C, \varepsilon_S \colon \mathbb{R}_+ \to \mathbb{R}_+$,
2ND_ORDER $\in \{\mathrm{True}, \mathrm{False}\}$.

1 **Input**   : Initial point $\omega_0 = (x_0, y_0) \in \mathrm{str}\,\mathcal{F} \times \mathbb{R}_{++}^m$, initial barrier parameter $\mu_0 > 0$.
2 **for** $k = 0, 1, \ldots$ **do**
3   Set COND $\leftarrow$ (3.3) if 2ND_ORDER is True. Else, set COND $\leftarrow$ (3.3a), (3.3b),
      and (3.3d).
4   Compute $\omega_{k+1} \in \mathrm{str}\,\mathcal{F} \times \mathbb{R}_{++}^m$ satisfying COND and $\Delta_k^{\mathrm{final}} > 0$ using, e.g., Algo-
      rithm 3.2 $\left(\omega_k, \mu_k, \Delta_k^{\mathrm{init}}, \mathrm{COND}\right)$.
5   Set $\Delta_{k+1}^{\mathrm{init}} \leftarrow \max\left(\Delta_k^{\mathrm{final}}, \Delta_{\min}^{\mathrm{init}}\right)$ and $\mu_{k+1} > 0$ so that $\lim_{k \to \infty} \mu_k = 0$.
6 **end for**

---

is the Jacobian of $\Psi(\cdot; \mu)$ at ω. Note that the Jacobian is independent of the value
of the barrier parameter μ. Since $G(x) \in \mathbb{R}^{m \times m}$ is nonsingular due to $x \in \mathrm{str}\,\mathcal{F}$,
equation (3.4) is equivalent to

$$(3.6) \qquad\qquad H(\omega)[d_x] = -c_\mu(x),$$

$$(3.7) \qquad\qquad d_y = -y + \mu G(x)^{-1}\mathbf{1} - YG(x)^{-1}\mathcal{G}_x^*[d_x],$$

where

$$(3.8) \qquad\qquad H(\omega) \coloneqq \mathrm{Hess}_x \mathcal{L}(\omega) + \mathcal{G}_x YG(x)^{-1}\mathcal{G}_x^*,$$

$$(3.9) \qquad\qquad c_\mu(x) \coloneqq \mathrm{grad}\,f(x) - \mu \mathcal{G}_x \left[G(x)^{-1}\mathbf{1}\right].$$

Based on (3.6), for each ℓ-th iteration, RIPTRM computes a search direction
for the primal variable using the trust region strategy; we introduce the trust region
subproblem at $x \in \mathrm{str}\,\mathcal{F}$ with the dual variable $y \in \mathbb{R}_{++}^m$, the trust region radius
$\Delta > 0$, and the barrier parameter $\mu > 0$ as follows:

$$(3.10) \qquad \begin{aligned} &\underset{d \in T_x \mathcal{M}}{\mathrm{minimize}} \quad m_{\omega,\mu}(d) \coloneqq \frac{1}{2}\langle H(\omega)[d], d\rangle_x + \langle c_\mu(x), d\rangle_x \\ &\text{subject to} \quad \|d\|_x \le \Delta. \end{aligned}$$

Despite the fact the subproblem is generally a nonconvex quadratic optimization
problem, we can solve the problem exactly; we provide a detailed discussion in Sec-
tion 6.2. RIPTRM computes the search direction for the primal variable, denoted by
$d_{x^\ell} \in T_{x^\ell}\mathcal{M}$, by approximately or exactly solving the trust region subproblem (3.10)
at $x^\ell \in \mathrm{str}\,\mathcal{F}$ with $y^\ell \in \mathbb{R}_{++}^m$ and $\Delta^\ell, \mu \in \mathbb{R}_{++}$.

By using $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ that satisfies $\left\|d_{x^\ell}\right\|_{x^\ell} \le \Delta^\ell$, RIPTRM computes the search
direction for the dual variable $y^\ell \in \mathbb{R}^m$, denoted by $d_{y^\ell} \in T_{y^\ell}\mathbb{R}^m \simeq \mathbb{R}^m$, according to
(3.7). If the point $\left(\mathrm{R}_{x^\ell}\left(d_{x^\ell}\right), y^\ell + d_{y^\ell}\right)$ satisfies the stopping conditions with $\mu > 0$,
the algorithm outputs this point. Otherwise, several tests are performed to determine
whether to update the trust region radius and the variables. First, RIPTRM tests
whether $\mathrm{R}_{x^\ell}\left(d_{x^\ell}\right)$ belongs to $\mathrm{str}\,\mathcal{F}$. If not, it shrinks the trust region radius $\Delta^{\ell+1} \leftarrow$

$\gamma\|d_{x^\ell}\|_{x^\ell}$ with $0 < \gamma < 1$ and starts over from the beginning of the next iteration, computing another search direction due to the smaller trust region radius. Otherwise, the algorithm proceeds to the other tests: for these tests, we introduce the log barrier function

$$(3.11) \qquad P_\mu(x) := f(x) - \mu \sum_{i \in \mathcal{I}} \log g_i(x)$$

for $x \in \operatorname{str}\mathcal{F}$, which serves as a merit function. We also define two reductions as

$$(3.12\text{a}) \qquad \operatorname{ared}_\mu(d_x) := \hat{P}_{\mu_x}(0_x) - \hat{P}_{\mu_x}(d_x),$$

$$(3.12\text{b}) \qquad \operatorname{pred}_{\omega,\mu}(d_x) := m_{\omega,\mu}(0_x) - m_{\omega,\mu}(d_x),$$

which we call the actual and predicted reductions, respectively. The actual reduction is the change in (3.11) produced by the step, and the predicted reduction is that of the objective function in (3.10). RIPTRM sets the trust region radius at the next iteration according to the ratio of the two reductions:

$$(3.13) \qquad \Delta^{\ell+1} \leftarrow \begin{cases} \frac{1}{4}\Delta^\ell & \operatorname{ared}^\ell < \frac{1}{4}\operatorname{pred}^\ell, \\ \min\left(2\Delta^\ell, \Delta_{\max}\right) & \operatorname{ared}^\ell \geq \frac{3}{4}\operatorname{pred}^\ell \text{ and } \|d_{x^\ell}\|_{x^\ell} = \Delta^\ell, \\ \Delta^\ell & \text{otherwise}, \end{cases}$$

where $\Delta_{\max} > 0$ is a predefined parameter called the maximal trust region radius, and we write ared^ℓ and pred^ℓ for $\operatorname{ared}_\mu\left(d_{x^\ell}\right)$ and $\operatorname{pred}_{\omega^\ell,\mu}\left(d_{x^\ell}\right)$, respectively. Here, we omit the subscript μ for brevity. Subsequently, the algorithm updates the iterate as

$$(3.14) \qquad x^{\ell+1} \leftarrow \begin{cases} \mathrm{R}_{x^\ell}\left(d_{x^\ell}\right) & \operatorname{ared}^\ell > \rho'\operatorname{pred}^\ell, \\ x^\ell & \text{otherwise}, \end{cases}$$

where $\rho' \in \left(0, \frac{1}{4}\right)$ is a predefined parameter working as the threshold of the ratio. We say that the ℓ-th iteration is *successful* if $\mathrm{R}_{x^\ell}\left(d_{x^\ell}\right) \in \operatorname{str}\mathcal{F}$ and $\operatorname{ared}^\ell > \rho'\operatorname{pred}^\ell$ hold and the iterate is redefined. Otherwise, we say that the iteration is *unsuccessful*. RIPTRM also updates $y^{\ell+1} \in \mathbb{R}_{++}^m$ using y^ℓ and d_{y^ℓ} when the iteration is successful. In this paper, we introduce a clipping for this update in Section 4.4. Algorithm 3.2 formally states the procedure of the inner iteration.

Now, we present three search directions obtained from the subproblem (3.10), namely, the Cauchy step, the eigenstep, and the exact step. We first define the Cauchy step as follows:

DEFINITION 3.1 ([1, Equation (7.8)]). *The Cauchy step $d_x^{\mathrm{C}} \in T_x\mathcal{M}$ of the subproblem (3.10) is defined as*

$$d_x^{\mathrm{C}} := -\tau^* c_\mu(x),$$

where we define $\tau^* = 0$ if $c_\mu(x) = 0_x$, otherwise,

$$\tau^* := \underset{\tau \geq 0}{\operatorname{argmin}}\left\{m_{\omega,\mu}(-\tau c_\mu(x)) \text{ subject to } \|\tau c_\mu(x)\|_x \leq \Delta\right\}$$

$$= \begin{cases} \frac{\Delta}{\|c_\mu(x)\|_x} & \text{if } \langle H(\omega)[c_\mu(x)], c_\mu(x)\rangle_x \leq 0, \\ \min\left(\frac{\|c_\mu(x)\|_x^2}{\langle H(\omega)[c_\mu(x)], c_\mu(x)\rangle_x}, \frac{\Delta}{\|c_\mu(x)\|_x}\right) & \text{otherwise}. \end{cases}$$

Algorithm 3.2 Inner iteration of RIPTRM

Require: Riemannian manifold $\mathcal{M}$, twice continuously differentiable functions f : $\mathcal{M} \to \mathbb{R}$ and $\{g_i\}_{i \in \mathcal{I}} \colon \mathcal{M} \to \mathbb{R}$, maximal trust region radius $\Delta_{\max} > 0$, threshold $\rho' \in \left(0, \frac{1}{4}\right)$, coefficient $\gamma \in (0, 1)$.

7 **Input** : Initial point $\omega^0 = \left(x^0, y^0\right) \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$, barrier parameter $\mu > 0$, initial trust region radius $\Delta^0 \in (0, \Delta_{\max}]$, stopping conditions COND

8 . **Output :** Final point $\omega = (x, y) \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$ and $\Delta^{\text{final}} > 0$.

9 **for** $\ell = 0, 1, \dots$ **do**

10   Compute $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ by approximately or exactly solving the subproblem (3.10).

11   Compute $d_{y^\ell} \in T_{y^\ell}\mathbb{R}^m$ according to (3.7).

12   **if** $\left(\mathrm{R}_{x^\ell}\left(d_{x^\ell}\right), y^\ell + d_{y^\ell}\right)$ *satisfies* COND **then**

13   **return** $\omega = \left(\mathrm{R}_{x^\ell}\left(d_{x^\ell}\right), y^\ell + d_{y^\ell}\right)$ and $\Delta^{\text{final}} = \Delta^\ell$.

14   **end if**

15   **if** $\mathrm{R}_{x^\ell}\left(d_{x^\ell}\right) \notin \operatorname{str}\mathcal{F}$ **then**

16   Set $\Delta^{\ell+1} \leftarrow \gamma\left\|d_{x^\ell}\right\|_{x^\ell}$ and go to line 9.

17   **end if**

18   Compute ared^ℓ and pred^ℓ by (3.12a) and (3.12b), respectively.

19   Update $\Delta^{\ell+1}$ according to (3.13).

20   Define $x^{\ell+1}$ by (3.14) and update $y^{\ell+1} \in \mathbb{R}_{++}^m$ by y^ℓ and d_{y^ℓ}; see (4.34) in Section 4.4, for example.

21 **end for**

The bound on the predicted reduction from the Cauchy step is shown below.

LEMMA 3.2 ([1, Equation (7.14)]). *The following holds:*

$$m_{\omega, \mu}(0_x) - m_{\omega, \mu}\left(d_x^{\mathrm{C}}\right) \geq \frac{1}{2}\|c_\mu(x)\|_x \min\left(\Delta, \frac{\|c_\mu(x)\|_x}{\|H(\omega)\|_{\text{op}}}\right).$$

If $\|H(\omega)\|_{\text{op}} = 0$*, then we regard* $\min\left(\Delta, \frac{\|c_\mu(x)\|_x}{\|H(\omega)\|_{\text{op}}}\right) = \Delta$*.*

We will prove that, if a search direction is adopted such that the decrease in the objective function of (3.10) is not less than that achieved by the Cauchy step, RIPTRM has a global convergence property to an AKKT point under certain assumptions. In the experiments described in Section 6, we use the search direction computed by the tCG method, which satisfies the aforementioned condition.

Next, we define the eigenstep. Let $\lambda^{\min}[H(\omega)] \in \mathbb{R}$ be the minimum eigenvalue of $H(\omega)$.

DEFINITION 3.3 ([11, Lemma 6.16]). *Given* $y \in \mathbb{R}_{++}^m$ *and* $\Delta, \mu \in \mathbb{R}_{++}$*, let* $s_x \in T_x\mathcal{M}$ *satisfy*

$$\|s_x\|_x = 1, \ \langle c_\mu(x), s_x \rangle_x \leq 0, \ \text{and} \ \langle H(\omega)[s_x], s_x \rangle_x < -\varepsilon_H,$$

where $\varepsilon_H \in \mathbb{R}$ *is a predefined parameter satisfying* $\lambda^{\min}[H(\omega)] < -\varepsilon_H$*. We call* $d_x^{\mathrm{E}} := \Delta s_x$ *an eigenstep.*

Note that an eigenvector corresponding to $\lambda^{\min}[H(\omega)] \in \mathbb{R}$ satisfies the conditions on the eigenstep with the appropriate choice of its sign and the scaling. The bound on the predicted reduction is also known when using the eigenstep:

LEMMA 3.4 ([12, Lemma 3.3]). *For any $\varepsilon_H \in \mathbb{R}$, any point $\omega = (x,y) \in \operatorname{str} \mathcal{F} \times \mathbb{R}^m_{++}$ with $\lambda^{\min}[H(\omega)] < -\varepsilon_H$, and $\mu > 0$, the corresponding eigenstep $d_x^{\mathrm{E}} \in T_x\mathcal{M}$ satisfies that*

$$(3.15) \qquad m_{\omega,\mu}(0_x) - m_{\omega,\mu}(d_x^{\mathrm{E}}) \geq \frac{1}{2}(\Delta)^2 \varepsilon_H.$$

We will prove that, if we adopt a search direction whose decrease in the objective function of (3.10) is not less than those of the Cauchy step and the eigenstep, RIPTRM has a global convergence property to an SOSP under assumptions. The eigenstep, however, is rarely computed in practice as mentioned in [11, p.130]. In this paper, the eigenstep mainly serves to show that the exact step of the subproblem (3.10) satisfies (3.15).

Let us end the subsection by considering the exact step, that is, a global optimum of the subproblem (3.10). Note that the subproblem (3.10) has a global optimum since the feasible region is bounded and closed, and the objective function is continuous. We provide the necessary and sufficient condition for the global optimality of (3.10). We will discuss the computation of $d_{x^\ell}^*$ in Section 6.2.

PROPOSITION 3.5 ([1, Proposition 7.3.1]). *The vector $d_x^* \in T_x\mathcal{M}$ is a global optimum of (3.10) if and only if there exists a scalar $\nu \geq 0$ such that the following hold:*

$$(3.16\mathrm{a}) \qquad \big(H(\omega) + \nu \operatorname{id}_{T_x\mathcal{M}}\big)d_x^* = -c_\mu(x),$$

$$(3.16\mathrm{b}) \qquad \nu\Big(\Delta - \big\|d_x^*\big\|_x\Big) = 0,$$

$$\big\|d_x^*\big\|_x \leq \Delta,$$

$$H(\omega) + \nu \operatorname{id}_{T_x\mathcal{M}} \succeq 0.$$

We will prove that, if we adopt the exact solution of the subproblem (3.10) as the search direction, RIPTRM also possesses a local near-quadratic convergence property to a solution of RICO (1.1) under assumptions.

4. Global convergence analysis. In this section, we establish the global convergence properties of RIPTRM. We first prove the global convergence of Algorithm 3.1 in Section 4.1. Then, we analyze Algorithm 3.2; we prove its consistency and the global convergence in Sections 4.2 and 4.3, respectively. We also provide clipping for computing the dual variables in Section 4.4.

4.1. Global convergence of Algorithm 3.1. In this subsection, we first establish the global convergence of Algorithm 3.1. In particular, we show the convergence to an AKKT point and an SOSP when the generated sequence accumulates at some point.

We first assume the following:

ASSUMPTION A.1. *For any $k \in \mathbb{N}_0$, the point $x_k \in \mathcal{M}$ satisfies COND in Algorithm 3.1.*

ASSUMPTION A.2. *The sequence $\Big\{\|\operatorname{grad}g_i(x_k)\|_{x_k}\Big\}_k$ is bounded for every $i \in \mathcal{I}$.*

Note that, as we will analyze in Section 4.3, Assumption A.1 is fulfilled if Assumptions B.1-B.11 hold for every inner iteration. Moreover, Assumption A.2 holds if $\{x_k\}_k$ is bounded. This boundedness of the generated sequence is assumed in the literature, for example, [40, Assumption 3.(C2)].

For the global convergence, we define a subset of the subsequences. This concept was originally introduced in the Euclidean setting by Conn et al. [18, Section 4.4].

DEFINITION 4.1. *Let* $\{x_{k_n}\}_n$ *be a subsequence of Algorithm 3.1. We say that* $\{x_{k_n}\}_n$ *is asymptotically consistent if, for each* $i \in \mathcal{I}$, *either*

$$\lim_{n\to\infty} g_i(x_{k_n}) = 0 \ \text{or} \ \liminf_{n\to\infty} g_i(x_{k_n}) > 0$$

holds. The former and the latter constraints are said to be asymptotically active and inactive, respectively. For such $\{x_{k_n}\}_n$, *we define* $\mathcal{W} \coloneqq \{i \in \mathcal{I}: \lim_{n\to\infty} g_i(x_{k_n}) = 0\}$.

Now, we prove the global convergence of Algorithm 3.1 by examining the limiting behavior of $\{\omega_k\}_k$.

THEOREM 4.2. *Suppose Assumption A.1. Then, the following hold:*

(4.1) $$\lim_{k\to\infty} \|\mathrm{grad}\mathcal{L}(\omega_k)\|_{x_k} = 0,$$

(4.2) $$\lim_{k\to\infty} \|Y_k g(x_k)\| = 0,$$

(4.3) $$\liminf_{k\to\infty} g_i(x_k) \geq 0 \ \text{and} \ \liminf_{k\to\infty} y_{ki} \geq 0 \ \text{for all} \ i \in \mathcal{I}.$$

Let $\{x_{k_n}\}_n$ *be an asymptotically consistent sequence, and let* $\{\xi_{x_{k_n}}\}_n$ *be any sequence of tangent vectors satisfying* $\xi_{x_{k_n}} \in T_{x_{k_n}}\mathcal{M}$ *for all* $n \in \mathbb{N}_0$,

(4.4a) $$\left\{\|\xi_{x_{k_n}}\|_{x_{k_n}}\right\}_n \ \text{is bounded},$$

(4.4b) $$\left\langle \mathrm{grad}g_i(x_{k_n}), \xi_{x_{k_n}} \right\rangle_{x_{k_n}} = 0 \ \text{for all} \ i \in \mathcal{W} \ \text{and all} \ n \in \mathbb{N}_0.$$

Additionally, suppose that Assumption A.2 holds and that 2ND_ORDER is True in Algorithm 3.1. Then, it also follows that

(4.5) $$\liminf_{n\to\infty}\left\langle \mathrm{Hess}\mathcal{L}(\omega_{k_n})[\xi_{x_{k_n}}], \xi_{x_{k_n}} \right\rangle_{x_{k_n}} \geq 0.$$

Proof. We can easily obtain (4.1) and (4.3) from (3.3a) and (3.3d), respectively. As for (4.2), it follows from (3.3b) that

$$\|Y_k g(x_k)\| \leq \|Y_k g(x_k) - \mu_{k-1}\mathbf{1}\| + \mu_{k-1}\|\mathbf{1}\| \leq \varepsilon_C(\mu_{k-1}) + m\mu_{k-1},$$

where the right-hand side converges to zero as $k \to \infty$. Thus, equation (4.2) holds.

Next, additionally supposing that Assumption A.2 holds and 2ND_ORDER is True in Algorithm 3.1, we show (4.5) by contradiction. Assume that there exist an asymptotically consistent subsequence $\{x_{k_n}\}_n$ and a sequence of tangent vectors $\{\xi_{x_{k_n}}\}_n$ satisfying (4.4) such that

(4.6) $$\liminf_{n\to\infty}\left\langle \mathrm{Hess}\mathcal{L}(\omega_{k_n})[\xi_{x_{k_n}}], \xi_{x_{k_n}} \right\rangle_{x_{k_n}} < 0.$$

Define $v \coloneqq \min_{i\in\mathcal{I}\setminus\mathcal{W}} \liminf_{n\to\infty} g_i(x_{k_n}) > 0$. Note that $g_i(x_{k_n}) \geq \frac{1}{2}v > 0$ holds for every $i \in \mathcal{I}\setminus\mathcal{W}$ and all $n \in \mathbb{N}_0$ sufficiently large. Therefore, together with (4.2), it follows that $\lim_{n\to\infty}[y_{k_n}]_i = 0$ for every $i \in \mathcal{I}\setminus\mathcal{W}$. For all $n \in \mathbb{N}_0$ sufficiently large,

we write Y_{k_n} for $\mathrm{diag}\big(y_{k_n}\big)$ and have

$$
\big\langle \mathrm{Hess}\,\mathcal{L}(\omega_{k_n})\big[\xi_{x_{k_n}}\big],\xi_{x_{k_n}}\big\rangle_{x_{k_n}}
$$

$$
= \big\langle H(\omega_{k_n})\big[\xi_{x_{k_n}}\big],\xi_{x_{k_n}}\big\rangle_{x_{k_n}} - \big\langle \mathcal{G}_{x_{k_n}} Y_{k_n} G(x_{k_n})^{-1}\mathcal{G}^*_{x_{k_n}}\big[\xi_{x_{k_n}}\big],\xi_{x_{k_n}}\big\rangle_{x_{k_n}}
$$

$$
\geq -\varepsilon_S\big(\mu_{k_n-1}\big) - \sum_{i\in\mathcal{I}} \frac{[y_{k_n}]_i}{g_i(x_{k_n})}\big\langle \mathrm{grad}\,g_i(x_{k_n}),\xi_{x_{k_n}}\big\rangle^2_{x_{k_n}}
$$

$$
\geq -\varepsilon_S\big(\mu_{k_n-1}\big) - \sum_{i\in\mathcal{I}\setminus\mathcal{W}} \frac{2[y_{k_n}]_i}{v}\big\|\mathrm{grad}\,g_i(x_{k_n})\big\|^2_{x_{k_n}}\big\|\xi_{x_{k_n}}\big\|^2_{x_{k_n}},
$$

where the equality follows from (3.8), the first inequality follows from (3.3c), and the second one from $g_i(x_{k_n}) \geq \frac{1}{2}v$ and (4.4b). Since $\left\{\big\|\mathrm{grad}\,g_i(x_{k_n})\big\|_{x_{k_n}}\right\}_n$ and $\left\{\big\|\xi_{x_{k_n}}\big\|_{x_{k_n}}\right\}_n$ are bounded by Assumption A.2 and (4.4a), $\lim_{n\to\infty}\varepsilon_S\big(\mu_{k_n-1}\big) = 0$ holds, and we have $\lim_{n\to\infty}[y_{k_n}]_i = 0$ for any $i \in \mathcal{I}\setminus\mathcal{W}$, the right-hand side converges to zero, which contradicts (4.6). The proof is complete. $\qquad\square$

We proceed to consider the global convergence when $\{x_k\}_k$ has an accumulation point. Let $x^* \in \mathcal{F}$ be any accumulation point, and let $\{x_{k_n}\}_n$ be a subsequence of the outer iterates that realizes $x_{k_n} \to x^*$ as $n \to \infty$. Here, $\{x_{k_n}\}_n$ is asymptotically consistent since the set of asymptotically active constraints is determined by the values of the inequality constraints at x^*, i.e., $\mathcal{W} = \mathcal{A}(x^*)$. In this case, we additionally assume the following conditions:

ASSUMPTION A.3. *The point $x^* \in \mathcal{M}$ satisfies the LICQ.*

ASSUMPTION A.4. *There exist $s_f > 0$ and a closed neighborhood $\mathcal{P}_f \subseteq \mathcal{M}$ of x^* such that, for all $x \in \mathcal{P}_f$,*

$$
(4.7) \qquad \big\|\mathrm{Hess}\,f(x^*) - \mathrm{PT}_{x^*\leftarrow x}\circ\mathrm{Hess}\,f(x)\circ\mathrm{PT}_{x\leftarrow x^*}\big\|_{\mathrm{op}} \leq s_f\,\mathrm{dist}(x,x^*).
$$

For each $i \in \mathcal{I}$, there exist $s_{g_i} > 0$ and a closed neighborhood $\mathcal{P}_{g_i} \subseteq \mathcal{M}$ of x^* such that, for all $x \in \mathcal{P}_{g_i}$,

$$
(4.8) \qquad \big\|\mathrm{Hess}\,g_i(x^*) - \mathrm{PT}_{x^*\leftarrow x}\circ\mathrm{Hess}\,g_i(x)\circ\mathrm{PT}_{x\leftarrow x^*}\big\|_{\mathrm{op}} \leq s_{g_i}\,\mathrm{dist}(x,x^*).
$$

Note that Assumption A.3 is standard in the literature; for example, [40, Assumption 3.(C1)] and [44, Propositions 3.2, 3.4, 4.2]. Assumption A.4 holds if f and $\{g_i\}_{i\in\mathcal{I}}$ are of class C^3 as in Lemma 2.6.

To prove the global convergence to x^*, we first derive the following lemma:

LEMMA 4.3. *Under Assumption A.3, there exists a neighborhood $\mathcal{N} \subseteq \mathcal{M}$ of $x^* \in \mathcal{M}$ such that $\{\mathrm{grad}\,g_i(x)\}_{i\in\mathcal{A}(x^*)}$ are linearly independent for any $x \in \mathcal{N}$.*

Proof. See Appendix C.2. $\qquad\square$

Now, we prove the global convergence to x^* in the following theorem. In the proof, we use the parallel transport to prove the weak second-order stationarity at x^*, which is peculiar to the Riemannian setting.

THEOREM 4.4. *Under Assumption A.1, $x^* \in \mathrm{str}\,\mathcal{F}$ is an AKKT point of RICO (1.1). Suppose additionally that Assumptions A.2-A.4 hold and 2ND_ORDER is True in Algorithm 3.1. Then, x^* is a w-SOSP of RICO (1.1) with associated Lagrange multipliers y^*.*

Proof. Theorem 4.2 directly implies that the sequence $\{x_{k_n}\}_n$ and the associated sequence $\{y_{k_n}\}_n$ satisfy the AKKT conditions (2.11). Hereafter, we consider the weak second-order stationarity when 2ND_ORDER is True in Algorithm 3.1. To this end, we first analyze the limiting behaviors of the orthogonal projection and the parallel transport. Let $\{e_i(x)\}_{i \in \mathcal{A}(x^*)}$ be the orthonormal basis for $\text{span}\{\text{grad}\,g_i(x)\}_{i \in \mathcal{A}(x^*)} \subseteq T_x\mathcal{M}$. Under Assumption A.3, we can construct the basis around x^* using the Gram-Schmidt process; from Lemma 4.3, there exists a neighborhood $\mathcal{N} \subseteq \mathcal{M}$ of x^* such that $\{\text{grad}\,g_i(x)\}_{i \in \mathcal{A}(x^*)}$ are linearly independent for any $x \in \mathcal{N}$. Without loss of generality, we let the indices of $\mathcal{A}(x^*) = \{1, 2, 3, \ldots, |\mathcal{A}(x^*)|\}$; see (2.9) for the definition of $\mathcal{A}(x^*)$. We define the orthonormal basis function $e_i \colon \mathcal{N} \to T\mathcal{M}$ as

$$e_i(x) := \frac{u_i(x)}{\|u_i(x)\|_x}, \quad \text{where } u_i(x) := \text{grad}\,g_i(x) - \sum_{t=1}^{i-1} \langle \text{grad}\,g_i(x), e_t(x) \rangle_x e_t(x)$$

for $i = 1, 2, 3, \ldots, |\mathcal{A}(x^*)|$. Note that, due to the continuity of the Riemannian gradients and the Riemannian metric, e_i is a continuous vector field. Using the basis, we define the orthogonal projection operator as

$$\mathrm{P}^*_{\mathcal{C}_\mathrm{w}}(x)[\xi_x] := \xi_x - \sum_{i \in \mathcal{A}(x^*)} \langle e_i(x), \xi_x \rangle_x e_i(x)$$

for $\xi_x \in T_x\mathcal{M}$. By replacing $\mathcal{N}$ with a sufficiently small neighborhood of x^* if necessary, it holds that, for any $\xi_{x^*} \in \mathcal{C}_\mathrm{w}(x^*)$, the map $\mathcal{N} \ni x \mapsto \mathrm{P}^*_{\mathcal{C}_\mathrm{w}}(x)\big[\mathrm{PT}_{x \leftarrow x^*}[\xi_{x^*}]\big] \in T_x\mathcal{M}$ is continuous from the continuities of e_i and the parallel transport [44, Lemma A.1]. Therefore,

(4.9)
$$\lim_{x \to x^*} \big\| \mathrm{PT}_{x \leftarrow x^*}[\xi_{x^*}] - \mathrm{P}^*_{\mathcal{C}_\mathrm{w}}(x)\big[\mathrm{PT}_{x \leftarrow x^*}[\xi_{x^*}]\big] \big\|_x$$
$$= \left\| \sum_{i \in \mathcal{A}(x^*)} \langle e_i(x^*), \xi_{x^*} \rangle_{x^*} e_i(x^*) \right\|_{x^*} = 0,$$

where the first equality follows from $\lim_{x \to x^*} \mathrm{PT}_{x \leftarrow x^*}[\xi_{x^*}] = \mathrm{PT}_{x^* \leftarrow x^*}[\xi_{x^*}] = \xi_{x^*}$ and the second one from ξ_{x^*} and each $e_i(x^*)$ being orthogonal. We also note that, by the continuity again, $\big\| \mathrm{PT}_{x \leftarrow x^*}[\xi_{x^*}] \big\|_x$ and $\big\| \mathrm{P}^*_{\mathcal{C}_\mathrm{w}}(x)\big[\mathrm{PT}_{x \leftarrow x^*}[\xi_{x^*}]\big] \big\|_x$ are bounded around x^*.

Notice that, from Proposition 2.10 and Assumption A.3, the point x^* is a KKT point, and hence there exist associated Lagrange multipliers $y^* \in \mathbb{R}^m_+$. Define $\omega := (x^*, y^*) \in \mathcal{F} \times \mathbb{R}^m_+$. We next analyze the behavior of the Hessian of the Lagrangian around ω^*. It follows that, for all $n \in \mathbb{N}_0$ sufficiently large,

$$\left\| \mathrm{Hess}\,\mathcal{L}(\omega^*) - \mathrm{PT}_{x^* \leftarrow x_{k_n}} \circ \mathrm{Hess}\,\mathcal{L}(\omega_{k_n}) \circ \mathrm{PT}_{x_{k_n} \leftarrow x^*} \right\|_{\mathrm{op}}$$
$$\leq \left\| \mathrm{Hess}\,f(x^*) - \mathrm{PT}_{x^* \leftarrow x_{k_n}} \circ \mathrm{Hess}\,f(x_{k_n}) \circ \mathrm{PT}_{x_{k_n} \leftarrow x^*} \right\|_{\mathrm{op}}$$
$$+ \sum_{i \in \mathcal{I}} \big|y_i^* - [y_{k_n}]_i\big| \|\mathrm{Hess}\,g_i(x^*)\|_{\mathrm{op}}$$
$$+ \sum_{i \in \mathcal{I}} \big|[y_{k_n}]_i\big| \left\| \mathrm{Hess}\,g_i(x^*) - \mathrm{PT}_{x^* \leftarrow x_{k_n}} \circ \mathrm{Hess}\,g_i(x_{k_n}) \circ \mathrm{PT}_{x_{k_n} \leftarrow x^*} \right\|_{\mathrm{op}}$$

$$\leq s_f \,\mathrm{dist}(x_{k_n}, x^*) + \sum_{i \in \mathcal{I}} \big|y_i^* - [y_{k_n}]_i\big| \|\mathrm{Hess}\,g_i(x^*)\|_{\mathrm{op}} + \sum_{i \in \mathcal{I}} \big|[y_{k_n}]_i\big| s_{g_i} \,\mathrm{dist}(x_{k_n}, x^*),$$

where the second inequality follows from (4.7) and (4.8). Since the right-hand side converges to zero as $n \to \infty$ due to the boundedness of $\left\{ \left\| y_{k_n} \right\| \right\}_n$ and $y_{k_n} \to y^*$, we have

$$(4.10) \qquad \lim_{n \to \infty} \left\| \operatorname{Hess} \mathcal{L}(\omega^*) - \operatorname{PT}_{x^* \leftarrow x_{k_n}} \circ \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \circ \operatorname{PT}_{x_{k_n} \leftarrow x^*} \right\|_{\operatorname{op}} = 0.$$

Using the results above, we now establish the weak second-order stationarity. For any $\xi_{x^*} \in \mathcal{C}_{\mathrm{w}}(x^*)$ and all $n \in \mathbb{N}_0$ sufficiently large, it follows that

$$\langle \operatorname{Hess} \mathcal{L}(\omega^*)[\xi_{x^*}], \xi_{x^*} \rangle_{x^*}$$
$$= \langle \operatorname{Hess} \mathcal{L}(\omega^*)[\xi_{x^*}], \xi_{x^*} \rangle_{x^*} - \left\langle \operatorname{PT}_{x^* \leftarrow x_{k_n}} \circ \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \circ \operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}], \xi_{x^*} \right\rangle_{x^*}$$
$$+ \left\langle \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right], \operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right\rangle_{x_{k_n}}$$
$$- \left\langle \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \left[\operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right], \operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right\rangle_{x_{k_n}}$$
$$+ \left\langle \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \left[\operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right], \operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right\rangle_{x_{k_n}}$$
$$(4.11) - \left\langle \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \left[\operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right], \operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right\rangle_{x_{k_n}}$$
$$+ \left\langle \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \left[\operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right], \operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right\rangle_{x_{k_n}}$$
$$\geq - \left\| \operatorname{Hess} \mathcal{L}(\omega^*) - \operatorname{PT}_{x^* \leftarrow x_{k_n}} \circ \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \circ \operatorname{PT}_{x_{k_n} \leftarrow x^*} \right\|_{\operatorname{op}} \| \xi_{x^*} \|^2_{x^*}$$
$$- \| \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \|_{\operatorname{op}} \left\| \operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] - \operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right\|_{x_{k_n}}$$
$$\left(\left\| \operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right\|_{x_{k_n}} + \left\| \operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right\|_{x_{k_n}} \right)$$
$$+ \left\langle \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \left[\operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right], \operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right\rangle_{x_{k_n}}.$$

By (4.9) and (4.10) and the boundedness of $\left\{ \left\| \operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right\|_{x_{k_n}} \right\}_n$, $\left\{ \left\| \operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right\|_{x_{k_n}} \right\}_n$, and $\left\{ \| \operatorname{Hess} \mathcal{L}(\omega_{k_n}) \|_{\operatorname{op}} \right\}_n$ around ω^*, the first two terms on the right-hand side of (4.11) converge to zero as $n \to \infty$. In addition, since the sequence $\left\{ \operatorname{P}^*_{\mathcal{C}_{\mathrm{w}}}(x_{k_n}) \left[\operatorname{PT}_{x_{k_n} \leftarrow x^*}[\xi_{x^*}] \right] \right\}_n$ satisfies (4.4) by definition, the last term on the right-hand side accumulates at a nonnegative value by taking a subsequence that realizes the limit inferior in (4.5). Therefore, equation (4.11) implies that ω^* is a w-SOSP, which completes the proof.   □

The following corollary immediately follows from Proposition 2.14:

COROLLARY 4.5. *Suppose that the SC holds at x^* and 2ND_ORDER is True in Algorithm 3.1. Under Assumptions A.1-A.4, (x^*, y^*) is an SOSP of RICO (1.1).*

4.2. Consistency of Algorithm 3.2. In this subsection, we prove that Algorithm 3.2 is consistent in the sense that, if the current point is not a solution of $\Psi(\cdot; \mu) = 0$ and the trust region radius is sufficiently small, the iteration becomes successful; that is, $\operatorname{R}_{x^\ell}(d_{x^\ell}) \in \operatorname{str} \mathcal{F}$ and $\operatorname{ared}^\ell > \rho' \operatorname{pred}^\ell$ hold. Hereafter, for brevity,

we write m^ℓ, H^ℓ, and c^ℓ for $m_{\omega^\ell,\mu}$, $H(\omega^\ell)$, and $c_\mu(x^\ell)$, respectively. We omit the subscript μ for brevity.

For the consistency, we assume the following:

ASSUMPTION B.1. *There exists $\kappa_C > 0$ such that, for all $\ell \in \mathbb{N}_0$, the search direction $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ satisfies*

$$m^\ell(0_{x^\ell}) - m^\ell(d_{x^\ell}) \geq \kappa_C \|c^\ell\|_{x^\ell} \min\left(\Delta^\ell, \frac{\|c^\ell\|_{x^\ell}}{\|H^\ell\|_{\mathrm{op}}}\right).$$

Note that, under Assumption B.1, $d_{x^\ell} \neq 0_{x^\ell}$ holds if $\|c^\ell\|_{x^\ell} \neq 0$. From Lemma 3.2, the Cauchy step satisfies Assumption B.1 with $\kappa_C = \frac{1}{2}$, and so does the exact step (3.16). Assumption B.1 is standard in the literature; it is made in [11, A6.3] and [1, Section 7.4.1], for example.

We first derive a bound on the values of the inequality constraints around each feasible point: recall the definitions of the pullback and its restriction (2.3).

LEMMA 4.6. *Choose $x \in \mathrm{str}\,\mathcal{F}$ arbitrarily. Then, there exists $\delta_x > 0$ such that $\hat{g}_{i_x}(\xi_x) \geq \frac{1}{2}g_i(x) > 0$ holds for all $i \in \mathcal{I}$ and any $\xi_x \in T_x\mathcal{M}$ with $\|\xi_x\|_x \leq \delta_x$.*

Proof. See Appendix C.3. □

Using Lemma 4.6, we conclude that the search direction obtained from (3.10) passes the test in line 15 of Algorithm 3.2 when the trust region radius is sufficiently small. We formally state this at the end of this section.

Next, we consider the tests on the ratio of the actual and predicted reductions. To this end, we derive the first and second directional derivatives of the merit function:

LEMMA 4.7. *Choose $x \in \mathrm{str}\,\mathcal{F}$ and $\mu > 0$ arbitrarily. Then, for all $\xi_x \in T_x\mathcal{M}$, the following holds:*

$$\mathrm{D}\hat{P}_{\mu_x}(0_x)[\xi_x] = \langle c_\mu(x), \xi_x \rangle_x. \tag{4.12}$$

Additionally, choose $y \in \mathbb{R}^m_{++}$ and $\Delta > 0$ arbitrarily. Let $d_x^* \in T_x\mathcal{M}$ and $\nu \geq 0$ be the global optimum of (3.10) at x with y, Δ, μ and its associated scalar, respectively. Then,

$$\mathrm{D}\hat{P}_{\mu_x}(0_x)[d_x^*] = -\left\langle \left(H(\omega) + \nu\,\mathrm{id}_{T_x\mathcal{M}}\right)[d_x^*], d_x^* \right\rangle_x. \tag{4.13}$$

Proof. See Appendix C.3. □

LEMMA 4.8. *Choose $x \in \mathrm{str}\,\mathcal{F}$ and $\mu > 0$ arbitrarily. Let $\xi_x \in T_x\mathcal{M}$ be any tangent vector satisfying $\hat{g}_{i_x}(\xi_x) \neq 0$ for all $i \in \mathcal{I}$. Then, for all $\zeta_x, \eta_x \in T_{\xi_x}T_x\mathcal{M} \simeq T_x\mathcal{M}$,*

$$\mathrm{D}^2\hat{P}_{\mu_x}(\xi_x)[\zeta_x, \eta_x] = \left\langle \left(\mathrm{Hess}\,\hat{f}_x(\xi_x) - \sum_{i \in \mathcal{I}} \frac{\mu}{\hat{g}_{i_x}(\xi_x)}\mathrm{Hess}\,\hat{g}_{i_x}(\xi_x)\right)[\zeta_x], \eta_x \right\rangle_x$$
$$+ \sum_{i \in \mathcal{I}} \frac{\mu}{\hat{g}_{i_x}(\xi_x)^2}\langle \mathrm{grad}\,\hat{g}_{i_x}(\xi_x), \zeta_x \rangle_x \langle \mathrm{grad}\,\hat{g}_{i_x}(\xi_x), \eta_x \rangle_x. \tag{4.14}$$

Proof. See Appendix C.3. □

Using Lemmas 4.7 and 4.8, we derive a bound on the gap between the predicted and actual reductions.

LEMMA 4.9. *Choose $\omega = (x, y) \in \operatorname{str} \mathcal{F} \times \mathbb{R}_{++}^m$ and $\mu > 0$ arbitrarily. Then, there exists $\alpha > 0$ such that*

$$\left| \operatorname{pred}_{\omega,\mu}(d_x) - \operatorname{ared}_\mu(d_x) \right| \leq \alpha \| d_x \|_x^2$$

for any $d_x \in T_x \mathcal{M}$ sufficiently small.

Proof. See Appendix A. □

Now, we prove that the iteration is successful if the current iterate is not a solution of $\Psi(\,\cdot\,;\mu) = 0$ for a given $\mu > 0$ and the trust region radius is sufficiently small. Let $\mathcal{S} \subseteq \mathbb{N}_0$ be the set of successful iterations.

PROPOSITION 4.10. *Suppose Assumption B.1. Let $\ell^\times \in \mathbb{N}_0 \backslash \mathcal{S}$ be any unsuccessful iteration of Algorithm 3.2 with $\Psi\left(\omega^{\ell^\times}; \mu\right) \neq 0$ for a given $\mu > 0$. Then, there exists $\ell \in \mathcal{S}$ such that $\ell \geq \ell^\times$.*

Proof. We first consider the test in line 15 of Algorithm 3.2 on the feasibility of the primal variable. Recall that, when the iterations are unsuccessful, the iterates remain the same and only the trust region radius continues to shrink, meaning that the norm of the search directions is made sufficiently small since $\left\| d_{x^\ell} \right\|_{x^\ell} \leq \Delta^\ell$ holds. Therefore, it follows from Lemma 4.6 that these sufficiently small directions pass the test in line 15 of Algorithm 3.2.

In the following, we focus on such iterations and additionally prove the existence of $\ell \in \mathbb{N}$ satisfying that $\ell \geq \ell^\times$ and $\operatorname{ared}^\ell > \rho' \operatorname{pred}^\ell$ hold. We argue by contradiction. Suppose that $\operatorname{ared}^\ell \leq \rho' \operatorname{pred}^\ell$ holds for every $\ell \geq \ell^\times$. Since $\operatorname{ared}^\ell \leq \rho' \operatorname{pred}^\ell < \frac{1}{4} \operatorname{pred}^\ell$ holds for every $\ell \geq \ell^\times$, the sequence $\left\{ \Delta^\ell \right\}_\ell$ converges to zero as ℓ tends to infinity according to (3.13). Combining this limit with $\left\| c^\ell \right\|_{x^\ell} = \left\| c^{\ell^\times} \right\|_{x^{\ell^\times}}$ and $\left\| H^\ell \right\|_{\operatorname{op}} = \left\| H^{\ell^\times} \right\|_{\operatorname{op}}$ for all $\ell \geq \ell^\times$ yields

$$(4.15) \qquad \min\left(\Delta^\ell, \frac{\left\| c^\ell \right\|_{x^\ell}}{\left\| H^\ell \right\|_{\operatorname{op}}} \right) = \min\left(\Delta^\ell, \frac{\left\| c^{\ell^\times} \right\|_{x^{\ell^\times}}}{\left\| H^{\ell^\times} \right\|_{\operatorname{op}}} \right) = \Delta^\ell$$

for any ℓ sufficiently large. We have that, for every $\ell \geq \ell^\times$ sufficiently large,

$$\operatorname{pred}^\ell - \operatorname{ared}^\ell \geq (1 - \rho') \operatorname{pred}^\ell$$
$$(4.16) \qquad \geq (1 - \rho') \kappa_C \left\| c^\ell \right\|_{x^\ell} \min\left(\Delta^\ell, \frac{\left\| c^\ell \right\|_{x^\ell}}{\left\| H^\ell \right\|_{\operatorname{op}}} \right) = (1 - \rho') \kappa_C \left\| c^\ell \right\|_{x^\ell} \Delta^\ell,$$

where the first inequality follows from $\operatorname{ared}^\ell \leq \rho' \operatorname{pred}^\ell$, the second one from Assumption B.1, and the equality from (4.15). Hence, it follows from (4.16) and Lemma 4.9 and $\left\| d_{x^\ell} \right\|_{x^\ell} \leq \Delta^\ell$ that

$$(1 - \rho') \kappa_C \left\| c^{\ell^\times} \right\|_{x^{\ell^\times}} \Delta^\ell \leq \operatorname{pred}^\ell - \operatorname{ared}^\ell \leq \left| \operatorname{pred}^\ell - \operatorname{ared}^\ell \right| \leq \alpha \| d_{x^\ell} \|_{x^\ell}^2 \leq \alpha \left(\Delta^\ell \right)^2$$

for all $\ell \geq \ell^\times$ sufficiently large. This is, however, a contradiction since the left-hand side is $\mathcal{O}\left(\Delta^\ell \right)$, whereas the right-hand side is $\mathcal{O}\left(\left(\Delta^\ell \right)^2 \right)$. The proof is complete. □

4.3. Global convergence of Algorithm 3.2. In this subsection, we prove the global convergence of Algorithm 3.2 to a solution of (3.2) with the second-order condition, which ensures that Algorithm 3.2 terminates in a finite number of iterations to satisfy the stopping conditions (3.3). For the global convergence, we additionally assume the following:

ASSUMPTION B.2. *The retraction* R *is second order.*

ASSUMPTION B.3. *The functions* $\{g_i\}_{i\in\mathcal{I}}$ *are radially L-C^1 on some* $\mathcal{U} \subseteq \mathcal{M}$ *with* $\{x^\ell\}_\ell \subseteq \mathcal{U}$.

ASSUMPTION B.4. *The functions* f *and* $\{g_i\}_{i\in\mathcal{I}}$ *are radially L-C^2 on some* $\mathcal{U} \subseteq \mathcal{M}$ *with* $\{x^\ell\}_\ell \subseteq \mathcal{U}$.

ASSUMPTION B.5. *The following hold:*
1. *The sequence* $\{g_i(x^\ell)\}_\ell$ *is bounded above for every* $i \in \mathcal{I}$.
2. *The sequences* $\{\|\mathrm{grad}\, g_i(x)\|_{x^\ell}\}_\ell$ *and* $\{\|\mathrm{Hess}\, g_i(x)\|_{\mathrm{op}}\}_\ell$ *are bounded above for every* $i \in \mathcal{I}$.

ASSUMPTION B.6. *The following hold:*
1. *The sequence* $\{f(x^\ell)\}_\ell$ *is bounded below.*
2. *The sequence* $\{\|\mathrm{Hess}\, f(x^\ell)\|_{\mathrm{op}}\}_\ell$ *is bounded above.*

ASSUMPTION B.7. *For each* $i \in \mathcal{I}$ *and any* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that, for any* $\ell \in \mathbb{N}_0$ *and all* $\xi_{x^\ell} \in T_{x^\ell}\mathcal{M}$ *with* $\|\xi_{x^\ell}\|_{x^\ell} \leq \delta$,

$$\left|\hat{g}_{i_{x^\ell}}(\xi_{x^\ell}) - \hat{g}_{i_{x^\ell}}(0_{x^\ell})\right| \leq \varepsilon. \tag{4.17}$$

ASSUMPTION B.8. *The sequence* $\{\|y^\ell\|\}_\ell$ *is bounded above.*

ASSUMPTION B.9. $\lim_{\ell\to\infty}\|y^\ell - \mu G(x^\ell)^{-1}\mathbf{1}\| = 0$.

ASSUMPTION B.10. *There exist positive scalars* $\delta_R, L_R \in \mathbb{R}_{++}$ *such that, for all* $\ell \in \mathbb{N}_0$ *and all* $\xi_{x^\ell} \in T_{x^\ell}\mathcal{M}$ *with* $\|\xi_{x^\ell}\|_{x^\ell} \leq \delta_R$,

$$\mathrm{dist}\left(x^\ell, \mathrm{R}_{x^\ell}(\xi_{x^\ell})\right) \leq L_R\|\xi_{x^\ell}\|_{x^\ell}. \tag{4.18}$$

ASSUMPTION B.11. *There exists* $\kappa_E > 0$ *such that, for any* $\varepsilon_H > 0$ *and any* $\ell \in \mathbb{N}_0$ *with* $\lambda^{\min}[H^\ell] < -\varepsilon_H$, *the search direction* $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ *satisfies*

$$m^\ell(0_{x^\ell}) - m^\ell(d_{x^\ell}^{\mathrm{E}}) \geq \kappa_E \varepsilon_H (\Delta^\ell)^2. \tag{4.19}$$

In Section 4.4, we will discuss Assumptions B.8 and B.9 when using a clipping to update the dual variable. We defer the discussion for the other assumptions to Appendix B. In short, Assumption B.2 is not restrictive, the eigenstep and the exact step satisfy Assumption B.11 with $\kappa_E = \frac{1}{2}$, and Assumptions B.3-B.7 and B.10 are fulfilled if the generated sequence $\{x^\ell\}_\ell$ is bounded and all functions $f, \{g_i\}_{i\in\mathcal{I}}$ are of class C^3, for example. We also note that all the assumptions are standard in the literature, as detailed in Appendix B.

For the global convergence analysis, we first investigate the properties of the sequence of the merit function's values. By assigning indices to the elements of $\mathcal{S}$ in order, we define the sequence of successful iterations as $\{\ell_j\}_j \subseteq \mathbb{N}_0$. Note that any iterate between ℓ_j and ℓ_{j+1}, if it exists, is unsuccessful by definition; that is, the $(1+\ell_j)$-, $(2+\ell_j)$-, ..., $(\ell_{j+1}-1)$-th iterations are all unsuccessful.

LEMMA 4.11. *Suppose Assumption* B.1. *The following hold:*
1. *The sequence $\left\{P_\mu\left(x^\ell\right)\right\}_\ell$ is monotonically non-increasing for a given $\mu > 0$.*
2. *Under Assumption* B.5.1 *and Assumption* B.6.1, $\left\{P_\mu\left(x^\ell\right)\right\}_\ell$ *is convergent for a given $\mu > 0$.*

Proof. We first consider Item 1. When $|\mathcal{S}| < \infty$, the entire sequence $\left\{x^\ell\right\}_\ell$ reduces to the finite sequence of the successful iterates $\left\{x^{\ell_j}\right\}_j$, and we can easily obtain the result in this case. In the following, we consider the case where $|\mathcal{S}| = \infty$. Note that, for any $j \in \mathbb{N}_0$, $x^{1+\ell_j} = x^{2+\ell_j} = \cdots = x^{\ell_{j+1}}$ holds: once the iterate is updated and set to $x^{1+\ell_j}$ at the ℓ_j-th iteration, it remains unchanged until it is updated again and set to $x^{1+\ell_{j+1}}$ at the ℓ_{j+1}-th iteration. Thus, we have

$$P_\mu\left(x^{\ell_j}\right) - P_\mu\left(x^{\ell_{j+1}}\right) = P_\mu\left(x^{\ell_j}\right) - P_\mu\left(x^{1+\ell_j}\right)$$
$$\geq \rho'\left(m^{\ell_j}\left(0_{x^{\ell_j}}\right) - m^{\ell_j}\left(d_{x^{\ell_j}}\right)\right) \geq \rho'\kappa_C\left\|c^{\ell_j}\right\|_{x^{\ell_j}} \min\left(\Delta^{\ell_j}, \frac{\left\|c^{\ell_j}\right\|_{x^{\ell_j}}}{\left\|H^{\ell_j}\right\|_{\mathrm{op}}}\right) \geq 0,$$

where the first inequality holds since the ℓ_j-th iteration is successful and the second one follows from Assumption B.1. Therefore, $\left\{P_\mu\left(x^{\ell_j}\right)\right\}_j$ is monotonically non-increasing. Since, again, all iterates between ℓ_j and ℓ_{j+1} are unsuccessful, the lower bound extends to all the iterates, and the proof is complete.

Next, we consider Item 2. By Assumption B.5.1 and Assumption B.6.1, the sequence $\left\{P_\mu\left(x^{\ell_j}\right)\right\}_j$ is bounded below, Item 2 follows by the monotone convergence theorem. □

We next derive positive lower bounds on $\left\{g_i\left(x^\ell\right)\right\}_\ell$ and its sufficiently small neighborhoods.

LEMMA 4.12. *Suppose Assumption* B.1 *and Assumption* B.6.1. *The following hold:*
1. *There exists $\underline{\varepsilon} > 0$ such that $g_i\left(x^\ell\right) \geq \underline{\varepsilon}$ holds for any $i \in \mathcal{I}$ and all $\ell \in \mathbb{N}_0$.*
2. *Under Assumption* B.7, *there exist $\varepsilon' > 0$ and $\delta' > 0$ such that, for any $\ell \in \mathbb{N}_0$ and all $\xi_{x^\ell} \in T_{x^\ell}\mathcal{M}$ with $\|\xi_{x^\ell}\|_{x^\ell} \leq \delta'$, $\hat{g}_{i_{x^\ell}}(\xi_{x^\ell}) \geq \varepsilon'$ holds.*

Proof. We first consider Item 1. When $|\mathcal{S}| < \infty$, the entire sequence $\left\{x^\ell\right\}_\ell$ reduces to the finite one of the successful iterates $\left\{x^{\ell_j}\right\}_j$. In this case, we can easily obtain the conclusion by letting $\underline{\varepsilon} := \min_{i,j} \frac{1}{2}g_i\left(x^{\ell_j}\right)$ and $\delta' := \min_j \delta_{x^{\ell_j}}$ in Lemma 4.6. In the following, we consider the case where $|\mathcal{S}| = \infty$. Recall that, from Item 1 of Lemma 4.11, the sequence $\left\{P_\mu\left(x^\ell\right)\right\}_\ell$ is bounded above. Combining this with (3.11) and Assumption B.6.1, we obtain that $\sum_{i \in \mathcal{I}} \log g_i\left(x^\ell\right) = \mu^{-1}\left(f\left(x^\ell\right) - P_\mu\left(x^\ell\right)\right)$ is bounded below. This implies that there exists $\underline{\varepsilon} > 0$ such that $g\left(x^\ell\right) \geq \underline{\varepsilon}$ for all $\ell \in \mathbb{N}_0$. Thus, Item 1 is established.

Next, we consider Item 2. Let $\varepsilon' := \frac{3}{4}\underline{\varepsilon} > 0$. By Assumption B.7, for each $i \in \mathcal{I}$, there exists $\delta_i > 0$ such that, for any $\ell \in \mathbb{N}_0$ and all $\xi_{x^\ell} \in T_{x^\ell}\mathcal{M}$ with $\|\xi_{x^\ell}\|_{x^\ell} \leq \delta_i$, it follows that $\left|\hat{g}_{i_{x^\ell}}(\xi_{x^\ell}) - g_i\left(x^\ell\right)\right| \leq \frac{1}{3}\varepsilon'$, which implies $\hat{g}_{i_{x^\ell}}(\xi_{x^\ell}) \geq g_i\left(x^\ell\right) - \frac{1}{3}\varepsilon' \geq \frac{4}{3}\varepsilon' - \frac{1}{3}\varepsilon' = \varepsilon'$. By letting $\delta' := \min_{i \in \mathcal{I}} \delta_i > 0$, we complete the proof of Item 2. □

Note that Lemma 4.12 provides a positive lower bound on the values of the inequality constraints during the inner iterations, while Lemma 4.6 ensures the positivity of the inequality constraints around a fixed point. We use Lemma 4.12 to prove the boundedness of several sequences, one of which is $\left\{\left\|H^\ell\right\|_{\mathrm{op}}\right\}_\ell$ in the following lemma.

LEMMA 4.13. *Under Assumptions B.1, B.5, B.6, and B.8, there exists $\kappa^H > 0$ such that $\left\|H^\ell\right\|_{\mathrm{op}} \leq \kappa^H$ for all $\ell \in \mathbb{N}_0$.*

Proof. See Appendix C.4. □

The following lemma is a crucial ingredient for proving the global convergence property.

LEMMA 4.14. *Suppose that Assumptions B.1-B.4, B.5.2, B.6.1, B.7 and B.8 hold. The following hold:*

1. *There exist $\Delta' > 0$ and $\beta > 0$ such that, for all $\ell \in \mathbb{N}_0$, if the search direction $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ satisfies $\left\|d_{x^\ell}\right\|_{x^\ell} \leq \Delta'$, then*

$$(4.20) \qquad \left|\mathrm{pred}^\ell - \mathrm{ared}^\ell\right| \leq \beta\left\|d_{x^\ell}\right\|_{x^\ell}^2.$$

2. *Suppose Assumption B.9 in addition. Choose $\gamma > 0$ arbitrarily. Then, there exist $\Delta'' > 0$ and $K' \in \mathbb{N}_0$ such that, for all $\ell \geq K'$, if the search direction $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ satisfies $\left\|d_{x^\ell}\right\|_{x^\ell} \leq \Delta''$, then*

$$\left|\mathrm{pred}^\ell - \mathrm{ared}^\ell\right| \leq \gamma\left\|d_{x^\ell}\right\|_{x^\ell}^2.$$

Proof. See Appendix A. □

In contrast to Euclidean optimization, we make use of the retraction in the Riemannian setting, which may cause the curve $t \mapsto \mathrm{R}_x(td_x)$ to reach the boundary of the feasible region and potentially leads to a division by zero. Additionally, since the pullbacks of the Riemannian gradient and Hessian are defined on $T_x\mathcal{M}$, which varies depending on ℓ, their norms may diverge as ℓ tends to infinity. Our analysis in the proof appropriately addresses these concerns using Lemma 4.12 and Assumptions B.2-B.4.

Compared with Lemma 4.9, Lemma 4.14 estimates the gap between the predicted and actual reductions during the inner iteration. In particular, by additionally supposing Assumption B.9, we obtain a bound whose coefficient can be made arbitrarily small as proved in Item 2 of Lemma 4.14. We will use Item 1 of Lemma 4.14 and Item 2 of Lemma 4.14 to prove the global convergence to an AKKT point and an SOSP, respectively.

In the following theorem, we analyze the limit inferior of $\left\{\left\|c^\ell\right\|_{x^\ell}\right\}_\ell$ that will be used for analyzing the entire sequence.

THEOREM 4.15. *Under Assumptions B.1-B.8, $\liminf_{\ell \to \infty}\left\|c^\ell\right\|_{x^\ell} = 0$ holds.*

Proof. We argue by contradiction. Suppose that there exists $\varepsilon_g > 0$ and $K \in \mathbb{N}_0$ such that $\left\|c^\ell\right\|_{x^\ell} \geq \varepsilon_g$ holds for all $\ell \geq K$. It follows from Assumption B.1 and Lemma 4.13 that, for all $\ell \geq K$,

$$(4.21) \qquad \mathrm{pred}^\ell \geq \kappa_C\left\|c^\ell\right\|_{x^\ell} \min\left(\Delta^\ell, \frac{\left\|c^\ell\right\|_{x^\ell}}{\left\|H^\ell\right\|_{\mathrm{op}}}\right) \geq \kappa_C\varepsilon_g \min\left(\Delta^\ell, \frac{\varepsilon_g}{\kappa^H}\right) > 0.$$

Therefore, letting $\Delta' > 0$ be the threshold in Item 1 of Lemma 4.14, we have

$$\left|\frac{\mathrm{ared}^\ell}{\mathrm{pred}^\ell} - 1\right| = \frac{\left|\mathrm{pred}^\ell - \mathrm{ared}^\ell\right|}{\left|\mathrm{pred}^\ell\right|} \leq \frac{\beta\left\|d_{x^\ell}\right\|_{x^\ell}^2}{\kappa_C\varepsilon_g \min\left(\Delta^\ell, \frac{\varepsilon_g}{\kappa^H}\right)} \leq \frac{\beta\left(\Delta^\ell\right)^2}{\kappa_C\varepsilon_g \min\left(\Delta^\ell, \frac{\varepsilon_g}{\kappa^H}\right)}$$

whenever $\ell \geq K$ and $\Delta^\ell \leq \Delta'$, where the first inequality follows from (4.20) and (4.21) and the second one from $\left\|d_{x^\ell}\right\|_{x^\ell} \leq \Delta^\ell$. Define $\Delta^\sharp := \min\left(\Delta', \frac{\kappa_C \varepsilon_g}{2\beta}, \frac{\varepsilon_g}{\kappa^H}\right)$. If $\Delta^\ell \leq \Delta^\sharp$, we have $\min\left(\Delta^\ell, \frac{\varepsilon_g}{\kappa^H}\right) = \Delta^\ell$ and hence $\left|\frac{\mathrm{ared}^\ell}{\mathrm{pred}^\ell} - 1\right| \leq \frac{\beta\Delta^\ell}{\kappa_C \varepsilon_g} \leq \frac{1}{2}$. This implies $\frac{\mathrm{ared}^\ell}{\mathrm{pred}^\ell} \geq \frac{1}{2} > \frac{1}{4} > \rho'$, and thus the update $\Delta^{\ell+1} = \frac{1}{4}\Delta^\ell$ can occur only if $\Delta^\ell > \Delta^\sharp$. Therefore, for all $\ell \geq K$, we have

$$(4.22) \qquad \Delta^\ell \geq \min\left(\Delta^K, \frac{\Delta^\sharp}{4}\right).$$

On the other hand, recall that $\mathcal{S}$ is the set of the successful iterations, and $\left\{x^{\ell_j}\right\}_j$ is the ordered sequence of $\mathcal{S}$. When $|\mathcal{S}|$ is finite, all the sufficiently large iterations become unsuccessful. Thus, it follows that $\lim_{\ell\to\infty} \Delta^\ell = 0$, which contradicts (4.22). In the following, we consider the case where $|\mathcal{S}|$ is infinite. Recall that $x^{1+\ell_j} = x^{\ell_{j+1}}$ holds since the $(1 + \ell_j)$-, $(2 + \ell_j)$-, ..., $(\ell_{j+1} - 1)$-th iterations are all unsuccessful. For all $j \in \mathbb{N}_0$ satisfying $\ell_j \geq K$, we obtain

$$(4.23)$$
$$P_\mu\left(x^{\ell_j}\right) - P_\mu\left(x^{\ell_{j+1}}\right) = P_\mu\left(x^{\ell_j}\right) - P_\mu\left(x^{1+\ell_j}\right) \geq \rho' \mathrm{pred}_{\ell_j} \geq \rho' \varepsilon_g \min\left(\Delta^{\ell_j}, \frac{\varepsilon_g}{\kappa^H}\right) \geq 0,$$

where the equality follows from $x^{\ell_{j+1}} = x^{1+\ell_j}$, the first inequality from the fact that the ℓ_j-th iterate is successful, and the second one from (4.21). From Item 2 of Lemma 4.11, $\left\{P_\mu\left(x^\ell\right)\right\}_\ell$ is a Cauchy sequence. Thus, (4.23) implies $\liminf_{\ell_j\to\infty} \Delta^{\ell_j} = 0$, which contradicts (4.22).

Now, we have that, for any $\varepsilon_g > 0$ and any index $K \in \mathbb{N}_0$, there exists $\ell_K \geq K$ such that $\left\|c^{\ell_K}\right\|_{x^{\ell_K}} < \varepsilon_g$ holds. This implies $\liminf_{\ell\to\infty}\left\|c^\ell\right\|_{x^\ell} = 0$. $\qquad \square$

Now, we present the following theorem on the limit of $\left\{\left\|c^\ell\right\|_{x^\ell}\right\}_\ell$. Recall that $\mathcal{S}$ is the set of the successful iterations.

THEOREM 4.16. *Under Assumptions B.1-B.8 and B.10, $\lim_{\ell\to\infty}\left\|c^\ell\right\|_{x^\ell} = 0$ holds.*

Proof. If $|\mathcal{S}|$ is finite, the point $x^\ell \in \mathrm{str}\,\mathcal{F}$ remains the same for all $\ell \in \mathbb{N}_0$ sufficiently large. Hence, from Theorem 4.15, it follows that $\lim_{\ell\to\infty}\left\|c^\ell\right\|_{x^\ell} = 0$. In the following, we consider the case where $|\mathcal{S}|$ is infinite. Suppose that there exists an infinite subsequence of $\mathcal{S}$, denoted by $\left\{x^{\ell_{j_r}}\right\}_r$, and a constant $\varepsilon > 0$ such that $\left\|c^{\ell_{j_r}}\right\|_{x^{\ell_{j_r}}} \geq 3\varepsilon$ holds for all $r \in \mathbb{N}_0$. From Theorem 4.15, for any $r \in \mathbb{N}_0$, there exists the first index $p_r \in \mathbb{N}_0$ satisfying $p_r > \ell_{j_r}$ and $\left\|c^{p_r}\right\|_{x^{p_r}} < \varepsilon$. Define $\mathcal{R}_r := \{q \in \mathcal{S} : \ell_{j_r} \leq q < p_r\}$ for each $r \in \mathbb{N}_0$ and $\mathcal{R} := \bigcup_{r\in\mathbb{N}_0} \mathcal{R}_r$. Notice $\mathcal{R}_r$ is nonempty since $\ell_{j_r} \in \mathcal{R}_r \subseteq \mathcal{S}$ holds. Note also that, for any $q \in \mathcal{R}$, $\left\|c^q\right\|_{x^q} \geq \varepsilon$ holds by definition. For any $q \in \mathcal{R}$, since the q-th iterate is successful, we have

$$P_\mu(x^q) - P_\mu\left(x^{q+1}\right) \geq \rho'(m^q(0_{x^q}) - m^q(d_{x^q}))$$
$$(4.24) \qquad \geq \rho' \kappa_C \left\|c^q\right\|_{x^q} \min\left(\Delta^q, \frac{\left\|c^q\right\|_{x^q}}{\left\|H^q\right\|_{\mathrm{op}}}\right) \geq \rho' \kappa_C \varepsilon \min\left(\Delta^q, \frac{\varepsilon}{\kappa^H}\right) \geq 0,$$

where the second inequality follows from Assumption B.1 and the third one from Lemma 4.13 and the definition of the set $\mathcal{R}$. Thus, it holds by Item 2 of Lemma 4.11 that the left-hand side converges to zero as $q \in \mathcal{R} \to \infty$ and hence $\lim_{q\in\mathcal{R}\to\infty} \Delta^q = 0$.

For all $q \in \mathcal{R}$ sufficiently large, since $\min\left(\Delta^q, \frac{\varepsilon}{\kappa_H}\right) = \Delta^q$ holds, it follows from (4.24) that

$$(4.25) \qquad \Delta^q \leq \frac{1}{\rho' \kappa_C \varepsilon}\left(P_\mu(x^q) - P_\mu\left(x^{q+1}\right)\right).$$

Additionally, note that (4.18) holds for all $q \in \mathcal{R}$ sufficiently large from Assumption B.10. We have

$$\mathrm{dist}\left(x^{\ell_{j_r}}, x^{p_r}\right) \leq \sum_{n=\ell_{j_r}}^{p_r-1} \mathrm{dist}\left(x^n, x^{n+1}\right) = \sum_{q \in \mathcal{R}_r} \mathrm{dist}\left(x^q, \mathrm{R}_{x^q}(d_{x^q})\right)$$

$$\leq \sum_{q \in \mathcal{R}_r} L_R \Delta^q \leq \sum_{q \in \mathcal{R}_r} \frac{L_R}{\rho' \kappa_C \varepsilon}\left(P_\mu(x^q) - P_\mu\left(x^{q+1}\right)\right)$$

$$= \sum_{n=\ell_{j_r}}^{p_r-1} \frac{L_R}{\rho' \kappa_C \varepsilon}\left(P_\mu(x^n) - P_\mu\left(x^{n+1}\right)\right) = \frac{L_R}{\rho' \kappa_C \varepsilon}\left(P_\mu\left(x^{\ell_{j_r}}\right) - P_\mu(x^{p_r})\right),$$

for any $r \in \mathbb{N}_0$ sufficiently large, where the first and the second equalities follow from the definition of $\mathcal{R}_r$, the second inequality from (4.18), and the third one from (4.25). From Item 2 of Lemma 4.11, the right-hand side converges to zero as $r \to \infty$, which means $\lim_{r \to \infty} \mathrm{dist}\left(x^{\ell_{j_r}}, x^{p_r}\right) = 0$. Therefore, from the continuities of c_μ defined in (3.9) and the norm, we obtain

$$\left|\left\|c^{\ell_{j_r}}\right\|_{x^{\ell_{j_r}}} - \left\|c^{p_r}\right\|_{x^{p_r}}\right| \leq \varepsilon$$

for all $r \in \mathbb{N}_0$ sufficiently large, which, together with the definitions of ℓ_{j_r} and p_r, implies that

$$2\varepsilon = 3\varepsilon - \varepsilon < \left\|c^{\ell_{j_r}}\right\|_{x^{\ell_{j_r}}} - \left\|c^{p_r}\right\|_{x^{p_r}} \leq \left|\left\|c^{\ell_{j_r}}\right\|_{x^{\ell_{j_r}}} - \left\|c^{p_r}\right\|_{x^{p_r}}\right| \leq \varepsilon$$

holds for some $r \in \mathbb{N}_0$. This is, however, a contradiction.

Now, we verify that, for any infinite subsequence of $\mathcal{S}$, denoted by $\left\{x^{\ell_{j_r}}\right\}_r$, and any $\varepsilon > 0$, there exists $r \in \mathbb{N}_0$ such that $\left\|c^{\ell_{j_r}}\right\|_{x^{\ell_{j_r}}} < \varepsilon$. Consider a subsequence of $\mathcal{S}$ that realizes $\limsup_{\ell \to \infty}\left\|c^\ell\right\|_{x^\ell}$. For any $r \in \mathbb{N}_0$, there exists an index $\ell_{j_r} \in \mathcal{S}$ of the subsequence that satisfies $\left\|c^{\ell_{j_r}}\right\|_{x^{\ell_{j_r}}} < r^{-1}$. Since the sequence $\left\{\left\|c^{\ell_{j_r}}\right\|_{x^{\ell_{j_r}}}\right\}_r$ converges to zero as $r \to \infty$, we see that the original subsequence also converges to zero, meaning that $\limsup_{\ell \to \infty}\left\|c^\ell\right\|_{x^\ell} = 0$. Thus, combining this with Theorem 4.15 yields $\lim_{\ell \to \infty}\left\|c^\ell\right\|_{x^\ell} = 0$. The proof is complete. $\square$

Using Theorem 4.16, we derive the bound on $\left\|\mathrm{grad}\mathcal{L}\left(\omega^\ell\right)\right\|_{x^\ell}$: for any $x^\ell \in \mathbb{N}_0$,

$$\left\|\mathrm{grad}\mathcal{L}\left(\omega^\ell\right)\right\|_{x^\ell} \leq \left\|c^\ell\right\|_{x^\ell} + \sum_{i \in \mathcal{I}}\left|\frac{\mu}{g_i\left(x^\ell\right)} - y_i^\ell\right|\left\|\mathrm{grad}g_i\left(x^\ell\right)\right\|_{x^\ell}.$$

By Theorem 4.16 and additionally supposing Assumption B.9, the right-hand side converges to zero as $\ell \to \infty$. We formally state the result in the following corollary.

COROLLARY 4.17. *Under Assumptions B.1-B.10, it holds that*

$$\lim_{\ell \to \infty}\left\|\mathrm{grad}\mathcal{L}\left(\omega^\ell\right)\right\|_{x^\ell} = 0.$$

Next, we provide the second-order analysis. Recall that $\lambda^{\min}[H^\ell] \in \mathbb{R}$ is the minimum eigenvalue of H^ℓ.

THEOREM 4.18. *Under Assumptions B.1-B.5 and B.7-B.9, B.6.1, and B.11,*

$$(4.26) \qquad \limsup_{\ell \to \infty} \lambda^{\min}[H^\ell] \geq 0.$$

Proof. We argue by contradiction. Suppose that there exist $\kappa_E > 0$ and an index $K \in \mathbb{N}_0$ such that $\lambda^{\min}[H^\ell] < -\kappa_E$ for all $\ell \geq K$. Let $\gamma := \frac{1}{2}\kappa_E\varepsilon_H > 0$ and let $\Delta'' > 0$ be the associated threshold value in Item 2 of Lemma 4.14. Then, since $\mathrm{pred}^\ell > 0$ holds by (4.19), we have

$$\left| \frac{\mathrm{ared}^\ell}{\mathrm{pred}^\ell} - 1 \right| = \frac{\left| \mathrm{pred}^\ell - \mathrm{ared}^\ell \right|}{\left| \mathrm{pred}^\ell \right|} \leq \frac{\gamma \|d_{x^\ell}\|_{x^\ell}^2}{\kappa_E \varepsilon_H (\Delta^\ell)^2} \leq \frac{1}{2}$$

whenever $\Delta^\ell \leq \Delta''$, where the first inequality follows from Item 2 of Lemma 4.14 and (4.19) again and the second one holds by the definition of γ and $\|d_{x^\ell}\|_{x^\ell} \leq \Delta^\ell$. Thus, if $\Delta^\ell \leq \Delta''$, we have $\frac{\mathrm{ared}^\ell}{\mathrm{pred}^\ell} \geq \frac{1}{2} > \frac{1}{4} > \rho'$, which implies that the update $\Delta^{\ell+1} = \frac{1}{4}\Delta^\ell$ can occur only if $\Delta^\ell > \Delta''$. Hence,

$$(4.27) \qquad \Delta^\ell \geq \min\left(\Delta^K, \frac{\Delta''}{4} \right)$$

holds for all $\ell \geq K$.

On the other hand, recall that $\mathcal{S}$ is the set of the successful iterations and $\{x^{\ell_j}\}_j$ is the ordered sequence of $\mathcal{S}$. When $|\mathcal{S}|$ is finite, for all $\ell \in \mathbb{N}_0$, the iteration is unsuccessful. Thus, it follows that $\lim_{\ell \to \infty} \Delta^\ell = 0$, which contradicts (4.27). In the following, we consider the case where $|\mathcal{S}|$ is infinite. For all $j \in \mathbb{N}_0$, we obtain

$$(4.28) \qquad \begin{aligned} P_\mu(x^{\ell_j}) - P_\mu(x^{\ell_{j+1}}) &= P_\mu(x^{\ell_j}) - P_\mu(x^{1+\ell_j}) \\ &\geq \rho'\left(m^{\ell_j}(0_{x^{\ell_j}}) - m^{\ell_j}(d_{x^{\ell_j}}) \right) \geq \rho'\kappa_E(\Delta^{\ell_j})^2 \varepsilon_H \geq 0, \end{aligned}$$

where the equality follows since $x^{\ell_{j+1}} = x^{1+\ell_j}$ holds, the first inequality does since the ℓ_j-th iterate is successful, and the second one follows from Assumption B.11. From Item 2 of Lemma 4.11, $\{P_\mu(x^\ell)\}_\ell$ is a Cauchy sequence. Therefore, it follows from (4.28) that $\lim_{j \to \infty} \Delta^{\ell_j} = 0$, which contradicts (4.27).

Now, we have that, for any $\kappa'_E > 0$ and any $K \in \mathbb{N}_0$, there exists $\ell_K \geq K$ such that $\lambda^{\min}[H^{\ell_K}] \geq -\kappa'_E$ holds. This implies the limit superior of the sequence $\lambda^{\min}[H^\ell]$ is not less than zero. Indeed, letting $\kappa'_E = K^{-1}$ for all $K \in \mathbb{N}$, we obtain $\sup_{\ell \geq K} \lambda^{\min}[H^\ell] \geq \lambda^{\min}[H^{\ell_K}] \geq -K^{-1}$ for all $K \in \mathbb{N}$. Taking the limit as $K \to \infty$ leads to $\limsup_{\ell \to \infty} \lambda^{\min}[H^\ell] \geq 0$. □

Based on the analyses above, we prove the global convergence of Algorithm 3.2 in the following theorem.

THEOREM 4.19. *Let $\varepsilon_L, \varepsilon_C, \varepsilon_S \in \mathbb{R}_{++}$ be any positive scalars. Under Assumptions B.1-B.10, the following hold after a finite number of iterations of Algorithm 3.2:*

$$(4.29) \qquad \left\| \mathrm{grad}\mathcal{L}(\omega^\ell) \right\|_{x^\ell} \leq \varepsilon_L,$$

$$(4.30) \qquad \left\| G(x^\ell)y^\ell - \mu\mathbf{1} \right\| \leq \varepsilon_C,$$

$$(4.31) \qquad y_i^\ell > 0, g(x^\ell) > 0.$$

Additionally, suppose Assumption B.11. *Then, equations* (4.29)–(4.31), *and*

$$\lambda^{\min}[H^\ell] \geq -\varepsilon_S \tag{4.32}$$

hold after a finite number of iterations of Algorithm 3.2.

Proof. Equation (4.29) follows from Corollary 4.17. As for (4.30), we have

$$\left\|G(x^\ell)y^\ell - \mu\mathbf{1}\right\| \leq \left\|G(x^\ell)\right\|\left\|y^\ell - \mu G(x^\ell)^{-1}\mathbf{1}\right\|.$$

Under Assumption B.5.1 and Assumption B.9, the right-hand side converges to zero as $\ell \to \infty$, which implies that (4.30) holds for all $\ell \in \mathbb{N}_0$ sufficiently large. Notice that (4.31) holds for every $\ell \in \mathbb{N}_0$ by lines 15 and 20 of Algorithm 3.2. Therefore, we conclude that (4.29)–(4.31) hold under Assumptions B.1-B.10.

Next, suppose Assumption B.11 additionally. Considering a subsequence that realizes (4.26), we guarantee (4.32) after finitely many iterations. □

4.4. Computation of dual variables. Since the simple update for the dual variables $y^\ell + d_{y^\ell}$ does not always preserve positivity, Assumption B.8, and Assumption B.9, we introduce a clipping to address the issue. Letting $\underline{c}, \tilde{c} \in \mathbb{R}$ be constants satisfying $0 < \underline{c} < 1 < \tilde{c}$, we define the interval for the i-th element as $I_i^\ell := [I_{\min i}^\ell, I_{\max i}^\ell]$, where

$$I_{\min i}^\ell := \underline{c}\min\left\{1, y_i^\ell, \frac{\mu}{\hat{g}_{i_{x^\ell}}(d_{x^\ell})}\right\}, \quad I_{\max i}^\ell := \max\left\{\tilde{c}, y_i^\ell, \frac{\tilde{c}}{\mu}, \frac{\tilde{c}}{\hat{g}_{i_{x^\ell}}(d_{x^\ell})}\right\} \tag{4.33}$$

for each $i \in \mathcal{I}$. We update the dual variable as follows:

$$y^{\ell+1} \leftarrow \begin{cases} \mathrm{clip}_{I^\ell}(y^\ell + d_{y^\ell}) & \text{if } x^{\ell+1} = \mathrm{R}_{x^\ell}(d_{x^\ell}), \\ y^\ell & \text{if } x^{\ell+1} = x^\ell, \end{cases} \tag{4.34}$$

where $\mathrm{clip}_{I^\ell}\colon \mathbb{R}^m \to \mathbb{R}^m$ is the operator whose i-th component is defined by

$$[\mathrm{clip}_{I^\ell}(v)]_i = \max\left\{I_{\min i}^\ell, \min\left\{I_{\max i}^\ell, v_i\right\}\right\}.$$

In Section 6, we observe that the choices $\underline{c} = 0.5$ and $\tilde{c} = 10^{20}$ work satisfactorily in practice.

Now, we provide theoretical analyses for (4.34). Note that the update (4.34) satisfies the positivity of the dual variable by definition. We prove that the strategy satisfies Assumptions B.8 and B.9 in the following theorem: recall that $\{\ell_j\}_j \subseteq \mathbb{N}_0$ is the sequence of successful iterations.

THEOREM 4.20. *Let $\left\{(x^\ell, y^\ell)\right\}_\ell$ be the sequence generated by Algorithm 3.2 with $\{y^\ell\}_\ell$ updated by (4.34). Suppose Assumption B.1 and Assumption B.6.1. Then, the following hold:*

1. *Assumption B.8 holds.*
2. *Under Assumptions B.5 and B.7, if*

$$\lim_{j\to\infty}\left\|d_{x^{\ell_j}}\right\|_{x^{\ell_j}} = 0, \tag{4.35}$$

   then Assumption B.9 holds.

Proof. See Appendix C.5 □

We provide a sufficient condition for (4.35) in the following.

Remark 4.21. If the maximal trust region radius $\Delta_{\max} > 0$ is sufficiently small and the sequence $\{x^{\ell_j}\}_j$ converges to some $x^* \in \mathcal{M}$, then (4.35) holds. Indeed, by Lemma 2.3, there exists $a > 0$ such that, for all $\ell_j \in \mathbb{N}_0$ sufficiently large,

$$(4.36) \qquad a\left\|d_{x^{\ell_j}}\right\|_{x^{\ell_j}} \leq \mathrm{dist}\left(\mathrm{R}_{x^{\ell_j}}\left(d_{x^{\ell_j}}\right), x^{\ell_j}\right) = \mathrm{dist}\left(x^{1+\ell_j}, x^{\ell_j}\right)$$

since the point $x^{\ell_j} \in \mathrm{str}\,\mathcal{F}$ is sufficiently close to x^* and the search direction $d_{x^{\ell_j}} \in T_{x^{\ell_j}}\mathcal{M}$ satisfies $\left\|d_{x^{\ell_j}}\right\|_{x^{\ell_j}} \leq \Delta_{\max}$. By the assumption, $\{x^{\ell_j}\}_j$ is a Cauchy sequence. Hence, equation (4.36) implies (4.35) by taking the limit as $\ell_j \to \infty$.

5. Local convergence analysis. In this section, we establish the local convergence property of RIPTRM. In Section 5.1, we prove that the exact step reaches a point that satisfies the stopping condition (3.3) in a small neighborhood of ω^* under assumptions, meaning that Algorithm 3.2 terminates in one iteration. In Section 5.2, we prove the locally near-quadratic convergence of Algorithm 3.1 using the exact step and a specific update for the sequence of the barrier parameter.

5.1. Analysis of the exact step satisfying the stopping conditions in Algorithm 3.2. In this subsection, we show that the exact step, that is, the global optimum of (3.10), is identical to the Newton step of (3.2) if the current iterate lies in a sufficiently small neighborhood of $\omega^* = (x^*, y^*) \in \mathcal{F} \times \mathbb{R}_+^m$ and the barrier parameter is sufficiently small. Then, we show that the first iterate of Algorithm 3.2 satisfies the stopping condition in Algorithm 3.2, implying that Algorithm 3.2 terminates in one iteration. This result will be the basis of the local convergence analysis in Section 5.2.

We assume the following:

ASSUMPTION C.1. *The point ω^* satisfies the SC, the LICQ, and the SOSC.*

ASSUMPTION C.2. *There exist $\varepsilon_{\mathrm{c}}, \tilde{\varepsilon}_{\mathrm{c}}, \varepsilon_{\mathcal{L}}, \tilde{\varepsilon}_{\mathcal{L}} \in \mathbb{R}$ with $0 < \varepsilon_{\mathrm{c}} < 1 < \tilde{\varepsilon}_{\mathrm{c}}$ and $0 < \varepsilon_{\mathcal{L}} < 1 < \tilde{\varepsilon}_{\mathcal{L}}$ such that, for all $k \in \mathbb{N}_0$,*

$$(5.1) \qquad \varepsilon_{\mathrm{c}}\mu_k \leq \varepsilon_C(\mu_k) \leq \tilde{\varepsilon}_{\mathrm{c}}\mu_k \ \text{and} \ \varepsilon_{\mathcal{L}}\mu_k \leq \varepsilon_{\mathcal{L}}(\mu_k) \leq \tilde{\varepsilon}_{\mathcal{L}}\mu_k.$$

ASSUMPTION C.3. *$\mu_k^2 = o(\mu_{k+1})$ holds.*

ASSUMPTION C.4. *$\mu_{k+1} \leq \mu_k$ holds for all $k \in \mathbb{N}_0$ sufficiently large.*

ASSUMPTION C.5. *There exist positive scalars $\hat{L}_{\mathrm{H}}^f, \{\hat{L}_{\mathrm{H}}^{g_i}\}_{i \in \mathcal{I}} \in \mathbb{R}_{++}$ such that, for all $x \in \mathcal{M}$ sufficiently close to $x^* \in \mathcal{F}$ and all $\xi_x \in T_x\mathcal{M}$,*

$$(5.2) \qquad \left\|\mathrm{Hess}\,\hat{f}_x(\xi_x) - \mathrm{Hess}\,\hat{f}_x(0_x)\right\|_{\mathrm{op}} \leq \hat{L}_{\mathrm{H}}^f\|\xi_x\|_x,$$

$$(5.3) \qquad \left\|\mathrm{Hess}\,\hat{g}_{i_x}(\xi_x) - \mathrm{Hess}\,\hat{g}_{i_x}(0_x)\right\|_{\mathrm{op}} \leq \hat{L}_{\mathrm{H}}^{g_i}\|\xi_x\|_x \ \text{for all} \ i \in \mathcal{I}.$$

ASSUMPTION C.6. *There exist positive scalars $L_{\mathrm{H}}^f, \{L_{\mathrm{H}}^{g_i}\}_{i \in \mathcal{I}} \in \mathbb{R}_{++}$ such that, for all $x \in \mathcal{M}$ satisfying $\mathrm{dist}(x, x^*) < \mathrm{inj}(x^*)$,*

$$(5.4)\left\|\mathrm{Hess}\,f(x) - \mathrm{PT}_{x \leftarrow x^*} \circ \mathrm{Hess}\,f(x^*) \circ \mathrm{PT}_{x^* \leftarrow x}\right\|_{\mathrm{op}} \leq L_{\mathrm{H}}^f\,\mathrm{dist}(x, x^*),$$

$$(5.5)$$
$$\left\|\mathrm{Hess}\,g_i(x) - \mathrm{PT}_{x \leftarrow x^*} \circ \mathrm{Hess}\,g_i(x^*) \circ \mathrm{PT}_{x^* \leftarrow x}\right\|_{\mathrm{op}} \leq L_{\mathrm{H}}^{g_i}\,\mathrm{dist}(x, x^*) \ \text{for all} \ i \in \mathcal{I}.$$

Note that Assumption C.1 is standard in the literature [40, Assumptions 1.(A3)–(A5)] and [49, Assumption B1]. We will discuss the relaxation of Assumption C.1 in Section 7. Assumption C.2 is fulfilled if we employ updates $\varepsilon_C(\mu_k) \leftarrow c_C \mu_k$ and $\varepsilon_{\mathcal{L}}(\mu_k) \leftarrow c_{\mathcal{L}} \mu_k$ for some $c_C, c_{\mathcal{L}} \in \mathbb{R}_{++}$. We will provide a specific update for $\{\mu_k\}_k$ that satisfies Assumptions C.3 and C.4 in Section 5.2. Assumption C.5 holds if the functions f and $\{g_i\}_{i \in \mathcal{I}}$ are of class C^3 [11, Lemma 10.57]. Assumption C.6 is essentially the same as Assumption A.4; we state it again simply for the notational convenience.

We first prove that the Newton step is well-defined around ω^*. For this, we show the nonsingularity of J_Ψ around ω^* in a similar manner to [40, Proposition 2.3]. Recall the definitions of the barrier KKT vector field (3.1), its Jacobian (3.5), and the critical cone (2.12).

LEMMA 5.1. *Under* Assumption C.1, $\mathrm{J}_\Psi(\omega^*)$ *is nonsingular.*

Proof. See Appendix C.6. $\qquad\qquad\square$

For $\omega \in \mathcal{M} \times \mathbb{R}^m_+$ with $\mathrm{J}_\Psi(\omega)$ being nonsingular, we define the Newton step $d^{\mathrm{N}}_{\omega,\mu} = \left(d^{\mathrm{N}}_{x,\mu}, d^{\mathrm{N}}_{y,\mu} \right)$ as

$$(5.6) \qquad d^{\mathrm{N}}_{\omega,\mu} := -\mathrm{J}_\Psi(\omega)^{-1} \Psi(\omega; \mu) \in T_\omega \mathcal{M} \times T_y \mathbb{R}^m \simeq T_\omega \mathcal{M} \times \mathbb{R}^m.$$

LEMMA 5.2. *Choose* $\Delta > 0$ *arbitrarily. Under* Assumption C.1, $\mathrm{J}_\Psi(\omega)$ *is nonsingular, and the Newton step* $d^{\mathrm{N}}_{\omega,\mu} \in T_x \mathcal{M} \times \mathbb{R}^m_+$ *is well-defined for any* $\omega \in \mathcal{F} \times \mathbb{R}^m_+$ *sufficiently close to* ω^* *under* $T_y \mathbb{R}^m \simeq \mathbb{R}^m$. *Moreover,* $\left\| d^{\mathrm{N}}_{\omega,\mu} \right\|_\omega \leq \Delta$ *holds for any* $\omega \in \mathcal{F} \times \mathbb{R}^m_+$ *sufficiently close to* ω^* *and any* $\mu > 0$ *sufficiently small.*

Proof. See Appendix C.6. $\qquad\qquad\square$

In the following proposition, we prove that the exact step is identical to the Newton step under the given assumptions.

PROPOSITION 5.3. *Let* $\Delta, \mu \in \mathbb{R}_{++}$ *be any positive scalars,* $\omega \in \mathcal{F} \times \mathbb{R}^m_+$ *be any point satisfying that* $\mathrm{J}_\Psi(\omega)$ *is nonsingular,* $d^*_x \in T_x \mathcal{M}$ *be the global optimum of* (3.10) *at* $x \in \mathcal{M}$ *with* Δ *and* μ, *and* $d^*_y \in \mathbb{R}^m$ *be the vector* (3.7) *with* d_x *replaced by* d^*_x. *Under* Assumption C.1, *if the search direction* $d^*_x \in T_x \mathcal{M}$ *satisfies* $\left\| d^*_x \right\|_x < \Delta$, *then* $d^*_\omega := \left(d^*_x, d^*_y \right)$ *is equivalent to* $d^{\mathrm{N}}_{\omega,\mu}$.

Proof. See Appendix C.6. $\qquad\qquad\square$

Recall that the iterate $x_k \in \mathcal{M}$ and the radius $\Delta^{\mathrm{init}}_{k+1} \geq \Delta^{\mathrm{init}}_{\min} > 0$ are used as the initial point and the initial trust region radius for the $(k+1)$-th iteration of Algorithm 3.1. Under Assumption C.1, it follows from Lemma 5.2 that the exact step is equivalent to the Newton step if the initial point ω_k is sufficiently close to $\omega^* \in \mathcal{F} \times \mathbb{R}^m_+$ and $\mu_k > 0$ is sufficiently small. In the following, we will prove that the exact step at the first iteration of Algorithm 3.2 satisfies the stopping condition of the k-th outer iteration in such cases; that is, the point $\left(\mathrm{R}_{x_k} \left(d^{\mathrm{N}}_{x_k,\mu_k} \right), y_k + d^{\mathrm{N}}_{y_k,\mu_k} \right)$, or equivalently $\left(\mathrm{R}_{x_k} \left(d^*_{x_k} \right), y_k + d^*_{y_k} \right)$, satisfies (3.3) with $\mu_k > 0$. To this end, we first derive the bounds on the inequality constraints, the Lagrange multipliers, and the norm of the Newton step associated with the barrier parameter, step by step.

LEMMA 5.4. *Suppose the SC and Assumption C.2. For any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^*, the following hold:*

1. *There exist $\underline{v}_y, \tilde{v}_y \in \mathbb{R}_{++}$ such that $\tilde{v}_y \geq \underline{v}_y > 0$ and, for any $i \in \mathcal{A}(x^*)$,*

$$(5.7) \qquad \frac{1}{\tilde{v}_y}(1 - \tilde{\varepsilon}_c)\mu_{k-1} \leq g_i(x_k) \leq \frac{1}{\underline{v}_y}(1 + \tilde{\varepsilon}_c)\mu_{k-1} \text{ and } \underline{v}_y \leq y_{ki} \leq \tilde{v}_y.$$

2. *There exist $\underline{v}_c, \tilde{v}_c \in \mathbb{R}_{++}$ such that $\tilde{v}_c \geq \underline{v}_c > 0$ and, for any $i \notin \mathcal{A}(x^*)$,*

$$(5.8) \qquad \frac{1}{\tilde{v}_c}(1 - \tilde{\varepsilon}_c)\mu_{k-1} \leq y_{ki} \leq \frac{1}{\underline{v}_c}(1 + \tilde{\varepsilon}_c)\mu_{k-1} \text{ and } \underline{v}_c \leq g_i(x_k) \leq \tilde{v}_c.$$

Proof. We first prove Item 1. Let $\underline{v}_y := \min_{i \in \mathcal{A}(x^*)} \frac{1}{2}y_i^*$ and $\tilde{v}_y := \max_{i \in \mathcal{A}(x^*)} 2y_i^*$. Under the SC, for each $i \in \mathcal{A}(x^*)$ and any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^*, it holds that $0 < \underline{v}_y \leq y_{ki} \leq \tilde{v}_y$ by definition. From (5.1), we have $\left|y_{ki}g_i(x_k) - \mu_{k-1}\right| \leq \tilde{\varepsilon}_c\mu_{k-1}$, which implies

$$\frac{1}{\tilde{v}_y}(1 - \tilde{\varepsilon}_c)\mu_{k-1} \leq \frac{1}{y_{ki}}(1 - \tilde{\varepsilon}_c)\mu_{k-1} \leq g_i(x_k) \leq \frac{1}{y_{ki}}(1 + \tilde{\varepsilon}_c)\mu_{k-1} \leq \frac{1}{\underline{v}_y}(1 + \tilde{\varepsilon}_c)\mu_{k-1}.$$

The proof of Item 1 is complete.

Similarly, we prove Item 2. Define $\underline{v}_c := \min_{i \notin \mathcal{A}(x^*)} \frac{1}{2}g_i(x_k)$ and let $\tilde{v}_c := \max_{i \notin \mathcal{A}(x^*)} 2g_i(x_k)$. Under the SC, for each $i \notin \mathcal{A}(x^*)$ and any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^*, it holds that $0 < \underline{v}_c \leq g_i(x_k) \leq \tilde{v}_c$. By (5.1), we have $\left|y_{ki}g_i(x_k) - \mu_{k-1}\right| \leq \tilde{\varepsilon}_c\mu_{k-1}$, which implies

$$\frac{1}{\tilde{v}_c}(1 - \tilde{\varepsilon}_c)\mu_{k-1} \leq \frac{1}{g_i(x_k)}(1 - \tilde{\varepsilon}_c)\mu_{k-1} \leq y_{ki} \text{ and}$$

$$y_{ki} \leq \frac{1}{g_i(x_k)}(1 + \tilde{\varepsilon}_c)\mu_{k-1} \leq \frac{1}{\underline{v}_c}(1 + \tilde{\varepsilon}_c)\mu_{k-1}.$$

The proof of Item 2 is complete. $\qquad\square$

LEMMA 5.5. *Under Assumptions C.1, C.2, and C.4, there exists $c_N > 0$ such that $\left\|d_{\omega_k, \mu_k}^N\right\|_{\omega_k} \leq c_N\mu_{k-1}$ holds for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_k is sufficiently small.*

Proof. Recall that, from Lemma 5.2, the Newton step is well-defined in a sufficiently small neighborhood of ω^*. From the continuity of J_Ψ and the nonsingularity of $J_\Psi(\omega^*)$, there exists $r > 0$ such that, for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_k is sufficiently small, $\left\|J_\Psi(\omega_k)^{-1}\right\|_{\text{op}} \leq r$ holds and hence

$$\left\|d_{\omega_k, \mu_k}^N\right\|_{\omega_k} \leq \left\|J_\Psi(\omega_k)^{-1}\right\|_{\text{op}} \|\Psi(\omega_k; \mu_k)\|_{\omega_k}$$

$$\leq r\left(\left\|\text{grad}_x\mathcal{L}(\omega_k)\right\|_{x_k} + \|Y_k g(x_k) - \mu_{k-1}\mathbf{1}_m\| + (\mu_{k-1} - \mu_k)\left\|\begin{bmatrix} 0_{x_k} \\ \mathbf{1}_m \end{bmatrix}\right\|_{\omega_k}\right)$$

$$\leq r\left(\varepsilon_\mathcal{L}(\mu_{k-1}) + \varepsilon_C(\mu_{k-1}) + \sqrt{m}(\mu_{k-1} + \mu_k)\right) \leq r\left(\tilde{\varepsilon}_\mathcal{L} + \tilde{\varepsilon}_c + 2\sqrt{m}\right)\mu_{k-1},$$

where the third inequality follows from (3.3a) and (3.3b) and the fourth one from (5.1) and Assumption C.4. Letting $c_N := \tilde{\varepsilon}_\mathcal{L} + \tilde{\varepsilon}_c + 2\sqrt{m} > 0$, we complete the proof.$\square$

Using Lemmas 5.4 and 5.5, we prove that the Newton step satisfies the stopping conditions (3.3) under assumptions one by one. The following lemmas ensure the feasibility of the iterate. Note that, since the Newton step $d^{\mathrm{N}}_{\omega_k,\mu_k}$ is equivalent to $d^*_{\omega_k}$ for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small, we have

$$\tag{5.9} \left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i g_i(x_k) = -y_{ki}g_i(x_k) + \mu_k - y_{ki}\mathrm{D}g_i(x_k)\left[d^{\mathrm{N}}_{x_k,\mu_k}\right]$$

for all $i \in \mathcal{I}$ by (3.7) with d_x replaced by $d^{\mathrm{N}}_{x_k,\mu_k}$.

LEMMA 5.6. *Under Assumptions* C.1-C.4, $g_i\left(\mathrm{R}_{x_k}\left(d^{\mathrm{N}}_{x_k,\mu_k}\right)\right) > 0$ *holds for any* $i \in \mathcal{I}$ *and any k-th iteration of Algorithm* 3.1 *satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small.*

Proof. For each $i \notin \mathcal{A}(x^*)$, it follows from the continuities of g_i and R that

$$g_i(x_k) \geq \frac{1}{2}g_i(x^*) > 0 \text{ and } \left|g_i \circ \mathrm{R}\left(x_k, d^{\mathrm{N}}_{x_k,\mu_k}\right) - g_i \circ \mathrm{R}(x_k, 0_{x_k})\right| \leq \frac{1}{3}g_i(x^*)$$

for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small, which, together with (2.2a), implies that

$$g_i\left(\mathrm{R}_{x_k}\left(d^{\mathrm{N}}_{x_k,\mu_k}\right)\right) \geq g_i(x_k) - \frac{1}{3}g_i(x^*) \geq \frac{1}{6}g_i(x^*) > 0.$$

Next, we consider the case $i \in \mathcal{A}(x^*)$. We have

$$\tag{5.10}\begin{aligned}\hat{g}_{ix_k}\left(d^{\mathrm{N}}_{x_k,\mu_k}\right) &= \hat{g}_{ix_k}(0_{x_k}) + \mathrm{D}\hat{g}_{ix_k}(0_{x_k})\left[d^{\mathrm{N}}_{x_k,\mu_k}\right] \\ &+ \int_0^1 \mathrm{D}\left(\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \hat{g}_{ix_k}(0_{x_k})\right)\left[d^{\mathrm{N}}_{x_k,\mu_k}\right]\mathrm{d}t\end{aligned}$$

In the following, we provide the bounds on the first two terms and the last one. As for the first two terms, for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small, we have

$$\tag{5.11}\begin{aligned}\hat{g}_{ix_k}(0_{x_k}) + \mathrm{D}\hat{g}_{ix_k}(0_{x_k})\left[d^{\mathrm{N}}_{x_k,\mu_k}\right] &= g_i(x_k) + \mathrm{D}g_i(x_k)\left[d^{\mathrm{N}}_{x_k,\mu_k}\right] \\ &= -\frac{\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i g_i(x_k)}{y_{ki}} + \frac{\mu_k}{y_{ki}} \geq -\frac{c_{\mathrm{N}}(1+\tilde{\varepsilon}_{\mathrm{c}})}{\underline{v}_y}\mu^2_{k-1} + \frac{\mu_k}{y_{ki}},\end{aligned}$$

where the first equality follows from (2.2), the second one from (5.9), and the inequality from (5.7) and Lemma 5.5. As for the third term, for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small, we have

$$\tag{5.12}\begin{aligned}&\left|\int_0^1 \mathrm{D}\left(\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \hat{g}_{ix_k}(0_{x_k})\right)\left[d^{\mathrm{N}}_{x_k,\mu_k}\right]\mathrm{d}t\right| \\ &\leq \int_0^1 \left|\left\langle\mathrm{grad}\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{grad}\hat{g}_{ix_k}(0_{x_k}), d^{\mathrm{N}}_{x_k,\mu_k}\right\rangle_{x_k}\right|\mathrm{d}t \\ &\leq \int_0^1 \beta^{g_i}_{RL}t\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|^2_{x_k}\mathrm{d}t = \frac{\beta^{g_i}_{RL}}{2}\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|^2_{x_k} \leq \frac{c^2_{\mathrm{N}}\beta^{g_i}_{RL}}{2}\mu^2_{k-1},\end{aligned}$$

where the second inequality follows from Lemma 2.5 and the last one from Lemma 5.5. Combining (5.10) with (5.11) and (5.12) yields

$$\hat{g}_{i x_k}\left(d_{x_k, \mu_k}^{\mathrm{N}}\right) \geq \mu_k\left(\frac{1}{y_{ki}} - \left(\frac{c_{\mathrm{N}}(1 + \tilde{\varepsilon}_{\mathrm{c}})}{\underline{v}_y} + \frac{c_{\mathrm{N}}^2 \beta_{RL}^{g_i}}{2}\right) \frac{\mu_{k-1}^2}{\mu_k}\right).$$

From Assumption C.3, the right-hand side is positive for $\mu_k > 0$ sufficiently small, which implies $\hat{g}_{i x_k}\left(d_{x_k, \mu_k}^{\mathrm{N}}\right) > 0$ for any $\mu_k > 0$ sufficiently small. We complete the proof. □

LEMMA 5.7. *Under Assumptions* C.1-C.4, $y_k + d_{y_k, \mu_k}^{\mathrm{N}} > 0$ *holds for any k-th iteration of Algorithm* 3.1 *satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small.*

Proof. Recall the definition (2.9) of $\mathcal{A}(x)$ with $x \in \mathcal{M}$. For each $i \in \mathcal{A}(x^*)$, since the point $y_{ki} + \left[d_{y_k, \mu_k}^{\mathrm{N}}\right]_i$ can be made arbitrarily close to y_i^* by considering the k-th iterations of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small, it holds that $y_{ki} + \left[d_{y_k, \mu_k}^{\mathrm{N}}\right]_i > 0$ for such ω_k and μ_k under the SC.

Next, we consider the case $i \notin \mathcal{A}(x^*)$. Note that the function $x \mapsto \|\mathrm{grad}g_i(x)\|_x$ is bounded around $x^* \in \mathcal{M}$ for each $i \in \mathcal{I}$ by the continuity of the gradient. Letting $\kappa > 0$ be a sufficiently large scalar, we obtain

$$\begin{aligned}
y_{ki} + \left[d_{y_k, \mu_k}^{\mathrm{N}}\right]_i &= g_i(x_k)^{-1}\left(\mu_k - y_{ki}\left\langle\mathrm{grad}g_i(x_k), d_{x_k, \mu_k}^{\mathrm{N}}\right\rangle_{x_k}\right) \\
&\geq g_i(x_k)^{-1}\left(\mu_k - y_{ki}\|\mathrm{grad}g_i(x_k)\|_{x_k}\left\|d_{x_k, \mu_k}^{\mathrm{N}}\right\|_{x_k}\right) \\
&\geq g_i(x_k)^{-1}\left(\mu_k - \frac{c_{\mathrm{N}}(1 + \tilde{\varepsilon}_{\mathrm{c}})}{\underline{v}_c}\|\mathrm{grad}g_i(x_k)\|_{x_k}\mu_{k-1}^2\right) \\
&\geq \frac{\mu_k}{g_i(x_k)}\left(1 - \kappa\frac{\mu_{k-1}^2}{\mu_k}\right) \geq \frac{\mu_k}{\bar{v}_c}\left(1 - \kappa\frac{\mu_{k-1}^2}{\mu_k}\right) > 0,
\end{aligned}$$

where the first equality follows from (5.9), the second inequality follows from (5.8) and Lemma 5.5, the third one from the boundedness of $\|\mathrm{grad}g_i(\cdot)\|_{(\cdot)}$ around ω^*, and the fourth one from (5.8) again and the positivity of $\left(1 - \kappa\frac{\mu_{k-1}^2}{\mu_k}\right)$ by Assumption C.3. The proof is complete. □

Before proving that the iterate obtained from the Newton step satisfies (3.3a) and (3.3b), we prove one more lemma, which is an extension of [1, Lemma 7.4.9].

LEMMA 5.8. *Let* R *be a retraction on $\mathcal{M}$, and let $\theta^1, \ldots, \theta^n \in \mathfrak{F}(\mathcal{M})$ be continuously differentiable functions. Given $x^* \in \mathcal{M}$ and $c > 1$, there exists a closed neighborhood $\mathcal{P} \subseteq \mathcal{M}$ of x^* and $\delta > 0$ such that, for all $x \in \mathcal{P}$, any $a \in \mathbb{R}^n$, and all $\xi_x \in T_x\mathcal{M}$ with $\|\xi_x\|_x \leq \delta$,*

$$(5.13) \qquad \left\|\sum_{t=1}^n a_t \mathrm{grad}\theta^t(\mathrm{R}_x(\xi_x))\right\|_x \leq c\left\|\sum_{t=1}^n a_t \mathrm{grad}\hat{\theta}_x^t(\xi_x)\right\|_x.$$

Proof. See Appendix C.6. □

Now, we prove the lemma that the iterate obtained from the Newton step satisfies (3.3a) and (3.3b) using Lemma 5.8. Note that the coefficients of the following bound

can be made arbitrarily small by taking ω_k sufficiently close to ω^* and μ_k sufficiently small.

LEMMA 5.9. *Choose* $c_{\mathcal{L}}, c_{\mathrm{c}} \in \mathbb{R}_{++}$ *arbitrarily. Under* Assumptions *C.1-C.5,*

$$(5.14) \quad \left\| \mathrm{grad} f\left(\mathrm{R}_x\left(d^{\mathrm{N}}_{x_k,\mu_k} \right) \right) - \sum_{i\in\mathcal{I}} \left(y_{ki} + d^{\mathrm{N}}_{y_k,\mu_k} \right) \mathrm{grad} g_i\left(\mathrm{R}_x\left(d^{\mathrm{N}}_{x_k,\mu_k} \right) \right) \right\|_{x_k} \le c_{\mathcal{L}}\mu_k,$$

$$(5.15) \quad \left\| G\left(\mathrm{R}_{x_k}\left(d^{\mathrm{N}}_{x_k,\mu_k} \right) \right)\left(y_{ki} + d^{\mathrm{N}}_{y_k,\mu_k} \right) - \mu_k \mathbf{1} \right\| \le c_{\mathrm{c}}\mu_k$$

hold for any k-th iteration of Algorithm 3.1 *satisfying that* ω_k *is sufficiently close to* ω^* *and* μ_{k-1} *is sufficiently small.*

Proof. Given $\omega \in \mathcal{M} \times \mathbb{R}^m$ and $\mu > 0$, we define the operator $\widetilde{\Psi}^\mu_\omega : T_\omega\mathcal{M} \to T_x\mathcal{M} \times \mathbb{R}^m$ as

$$(5.16) \qquad \widetilde{\Psi}^\mu_\omega(\xi_\omega) := \begin{bmatrix} \mathrm{grad}\,\hat{f}_x(\xi_x) - \sum_{i\in\mathcal{I}}(y_i+\xi_y)\mathrm{grad}\,\hat{g}_{i_x}(\xi_x) \\ G(\mathrm{R}_x(\xi_x))(y+\xi_y) - \mu\mathbf{1} \end{bmatrix}$$

for $\xi_\omega \in T_\omega\mathcal{M}$. Note that, for any $\xi_\omega, \zeta_\omega \in T_\omega\mathcal{M}$, its directional derivative is

$$\mathrm{D}\widetilde{\Psi}^\mu_\omega(\xi_\omega)[\zeta_\omega]$$
$$(5.17) \begin{bmatrix} \mathrm{Hess}\,\hat{f}_x(\xi_x)[\zeta_x] - \sum_{i\in\mathcal{I}}(y_i+\xi_{y_i})\mathrm{Hess}\,\hat{g}_{i_x}(\xi_x)[\zeta_x] - \sum_{i\in\mathcal{I}}\zeta_{y_i}\mathrm{grad}\,\hat{g}_{i_x}(\xi_x) \\ \left[(y_i+\xi_{y_i})\cdot\mathrm{D}\hat{g}_{i_x}(\xi_x)[\zeta_x] + \zeta_{y_i}\cdot\hat{g}_{i_x}(\xi_x)\right]_{i=1,\ldots,m} \end{bmatrix}.$$

We have

$$(5.18) \begin{aligned} \widetilde{\Psi}^{\mu_k}_{\omega_k}\left(d^{\mathrm{N}}_{\omega_k,\mu_k} \right) &= \widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k}) + \mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k})\left[d^{\mathrm{N}}_{\omega_k,\mu_k} \right] \\ &\quad + \int_0^1 \left(\mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}\left(t d^{\mathrm{N}}_{\omega_k,\mu_k} \right) - \mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k}) \right)\left[d^{\mathrm{N}}_{\omega_k,\mu_k} \right] \mathrm{d}t. \end{aligned}$$

In the following, we analyze each term of (5.18). For the first term, it follows from (2.4), (3.1), and (5.16) that

$$(5.19) \qquad \widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k}) = \Psi(\omega_k;\mu_k).$$

As for the second term, substituting $(\omega,\mu,\xi_\omega,\zeta_\omega) = \left(\omega_k, \mu_k, 0_{\omega_k}, d^{\mathrm{N}}_{\omega_k,\mu_k} \right)$ into (5.17) yields

$$\mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k})\left[d^{\mathrm{N}}_{\omega_k,\mu_k} \right]$$

$$(5.20) \begin{aligned} &= \begin{bmatrix} \mathrm{Hess}\,\hat{f}_{x_k}(0_{x_k})\left[d^{\mathrm{N}}_{x_k,\mu_k} \right] - \sum_{i\in\mathcal{I}} y_{ki}\mathrm{Hess}\,\hat{g}_{i_{x_k}}(0_{x_k})\left[d^{\mathrm{N}}_{x_k,\mu_k} \right] \\ - \sum_{i\in\mathcal{I}}\left[d^{\mathrm{N}}_{y_k,\mu_k} \right]_i \mathrm{grad}\,\hat{g}_{i_{x_k}}(0_{x_k}) \\ \left[y_{ki}\mathrm{D}\hat{g}_{i_{x_k}}(0_{x_k})\left[d^{\mathrm{N}}_{x_k,\mu_k} \right] + \left[d^{\mathrm{N}}_{y_k,\mu_k} \right]_i \hat{g}_{i_{x_k}}(0_{x_k}) \right]_{i=1,\ldots,m} \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{Hess}_x\mathcal{L}(\omega_k)\left[d^{\mathrm{N}}_{x_k,\mu_k} \right] - \sum_{i\in\mathcal{I}}\left[d^{\mathrm{N}}_{y_k,\mu_k} \right]_i \mathrm{grad}\,g_i(x_k) \\ \left[y_{ki}\left\langle \mathrm{grad}\,g_i(x_k), d^{\mathrm{N}}_{x_k,\mu_k} \right\rangle_{x_k} + \left[d^{\mathrm{N}}_{y_k,\mu_k} \right]_i g_i(x_k) \right]_{i=1,\ldots,m} \end{bmatrix} \\ &= \mathrm{J}_\Psi(\omega_k)\left[d^{\mathrm{N}}_{\omega_k,\mu_k} \right], \end{aligned}$$

where the second equality follows from (2.1), (2.2), and (2.5) and the third one from (3.5). Next, we consider the third term. Note that, by the twice continuous differentiablity of $\{g_i\}_{i\in\mathcal{I}}$ and [11, Lemma 10.57], there exist positive scalars $\{L^{g_i}\}_{i\in\mathcal{I}}$ and $\left\{\hat{L}^{g_i}_{\mathrm{g}}\right\}_{i\in\mathcal{I}}$ such that, for all $x \in \mathcal{M}$ sufficiently close to x^* and any $\xi_x \in T_x\mathcal{M}$ sufficiently small,

$$(5.21) \qquad \left|\hat{g}_{ix}(\xi_x) - \hat{g}_{ix}(0_x)\right| \leq L^{g_i}\,\mathrm{dist}\left(\mathrm{R}_x(\xi_x), x\right) \text{ for all } i \in \mathcal{I},$$

$$(5.22) \qquad \left\|\mathrm{grad}\hat{g}_{ix_k}(\xi_x) - \mathrm{grad}\hat{g}_{ix}(0_x)\right\| \leq \hat{L}^{g_i}_{\mathrm{g}}\|\xi_x\|_x \text{ for all } i \in \mathcal{I}.$$

It also follows from (5.17) with $(\omega, \mu, \xi_\omega, \zeta_\omega)$ replaced by $\left(\omega_k, \mu_k, td^{\mathrm{N}}_{\omega_k,\mu_k}, d^{\mathrm{N}}_{\omega_k,\mu_k}\right)$ and (5.20) that

$$\left(\mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}\left(td^{\mathrm{N}}_{\omega_k,\mu_k}\right) - \mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k})\right)\left[d^{\mathrm{N}}_{\omega_k,\mu_k}\right] = \begin{bmatrix} U \\ L \end{bmatrix},$$

where

$$U := \left(\mathrm{Hess}\hat{f}_{x_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{Hess}\hat{f}_{x_k}(0_{x_k})\right)\left[d^{\mathrm{N}}_{x_k,\mu_k}\right]$$
$$\quad - \sum_{i\in\mathcal{I}}\left(y_{ki} + t\left[d^{\mathrm{N}}_{y_k,\mu_k i}\right]_i\right)\left(\mathrm{Hess}\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{Hess}\hat{g}_{ix_k}(0_{x_k})\right)\left[d^{\mathrm{N}}_{x_k,\mu_k}\right]$$
$$\quad - \sum_{i\in\mathcal{I}}t\left[d^{\mathrm{N}}_{y_k,\mu_k i}\right]_i\mathrm{Hess}\hat{g}_{ix_k}(0_{x_k})\left[d^{\mathrm{N}}_{\omega_k,\mu_k}\right]$$
$$\quad - \sum_{i\in\mathcal{I}}\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\left(\mathrm{grad}\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{grad}\hat{g}_{ix_k}(0_{x_k})\right)$$
$$L := \left[\left(y_{ki} + t\left[d^{\mathrm{N}}_{y_k,\mu_k i}\right]_i\right)\left\langle\mathrm{grad}\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{grad}\hat{g}_{ix_k}(0_{x_k}), d^{\mathrm{N}}_{x_k,\mu_k}\right\rangle_{x_k}\right.$$
$$\quad + t\left[d^{\mathrm{N}}_{y_k,\mu_k i}\right]_i\left\langle\mathrm{grad}\hat{g}_{ix_k}(0_{x_k}), d^{\mathrm{N}}_{x_k,\mu_k}\right\rangle_{x_k}$$
$$\quad + \left.\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\left(\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \hat{g}_{ix_k}(0_{x_k})\right)\right]_{i=1,\dots,m}.$$

Therefore, letting $\kappa > 0$ be sufficiently large, we have that, for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently

small,

$$\left\|\left(\mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}\left(td^{\mathrm{N}}_{\omega_k,\mu_k}\right) - \mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k})\right)\left[d^{\mathrm{N}}_{\omega_k,\mu_k}\right]\right\|$$

$$\leq \left\|\mathrm{Hess}\hat{f}_{x_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{Hess}\hat{f}_{x_k}(0_{x_k})\right\|_{\mathrm{op}}\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|_{x_k}$$

$$+ \sum_{i\in\mathcal{I}}\left|y_{ki} + t\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\left\|\mathrm{Hess}\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{Hess}\hat{g}_{ix_k}(0_{x_k})\right\|_{\mathrm{op}}\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|_{x_k}$$

$$+ \sum_{i\in\mathcal{I}}t\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\left\|\mathrm{Hess}\hat{g}_{ix_k}(0_{x_k})\right\|_{\mathrm{op}}\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|_{x_k}$$

$$+ \sum_{i\in\mathcal{I}}\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\left\|\mathrm{grad}\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{grad}\hat{g}_{ix_k}(0_{x_k})\right\|$$

$$+ \sum_{i\in\mathcal{I}}\left|\left(y_{ki} + t\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right)\left\langle\mathrm{grad}\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \mathrm{grad}\hat{g}_{ix_k}(0_{x_k}), d^{\mathrm{N}}_{x_k,\mu_k}\right\rangle_{x_k}\right|$$

$$(5.23)\quad + \sum_{i\in\mathcal{I}}t\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\left\langle\mathrm{grad}\hat{g}_{ix_k}(0_{x_k}), d^{\mathrm{N}}_{x_k,\mu_k}\right\rangle_{x_k}\right|$$

$$+ \sum_{i\in\mathcal{I}}\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\hat{g}_{ix_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right) - \hat{g}_{ix_k}(0_{x_k})\right|$$

$$\leq \hat{L}^f_{\mathrm{H}}\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|^2_{x_k} + \sum_{i\in\mathcal{I}}\hat{L}^{g_i}_{\mathrm{H}}\left(|y_{ki}| + \left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\right)\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|^2_{x_k}$$

$$+ \sum_{i\in\mathcal{I}}\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\left\|\mathrm{Hess}g_i(x_k)\right\|_{\mathrm{op}}\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|_{x_k} + \sum_{i\in\mathcal{I}}\hat{L}^{g_i}_{\mathrm{g}}\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|_{x_k}$$

$$+ \sum_{i\in\mathcal{I}}\beta^{g_i}_{RL}\left(|y_{ki}| + \left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\right)\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|^2_{x_k}$$

$$+ \sum_{i\in\mathcal{I}}\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\left\|\mathrm{grad}g_i(x_k)\right\|_{x_k}\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|_{x_k}$$

$$+ \sum_{i\in\mathcal{I}}L^{g_i}\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right|\mathrm{dist}\left(\mathrm{R}_{x_k}\left(td^{\mathrm{N}}_{x_k,\mu_k}\right), x_k\right) \leq \kappa\left\|d^{\mathrm{N}}_{\omega_k,\mu_k}\right\|^2_{x_k} \leq \kappa c^2_{\mathrm{N}}\mu^2_{k-1},$$

where the second inequality follows from (5.2), (5.3), (2.5), (5.22) with ξ_x replaced by $td^{\mathrm{N}}_{x_k,\mu_k}$, (2.4), (5.21) with ξ_x replaced by $td^{\mathrm{N}}_{x_k,\mu_k}$, the radially L-$C^1$ property of $\{g_i\}_{i\in\mathcal{I}}$ around x^*, and $t \leq 1$, the third one from $\left|\left[d^{\mathrm{N}}_{y_k,\mu_k}\right]_i\right| \leq \left\|d^{\mathrm{N}}_{\omega_k,\mu_k}\right\|_{\omega_k}$, $\left\|d^{\mathrm{N}}_{x_k,\mu_k}\right\|_{x_k} \leq \left\|d^{\mathrm{N}}_{\omega_k,\mu_k}\right\|_{\omega_k}$, the boundedness of $\{y_k\}_k$ and $\left\{\left\|d^{\mathrm{N}}_{y_k,\mu_k}\right\|\right\}_k$, the smoothness of $\left\{\|\mathrm{Hess}g_i(\cdot)\|_{\mathrm{op}}\right\}_{i\in\mathcal{I}}$ and $\{\|\mathrm{grad}g_i(\cdot)\|.\}_{i\in\mathcal{I}}$, and Lemma 2.3, and the fourth one from Lemma 5.5. From (5.18), we obtain

$$\left\|\widetilde{\Psi}^{\mu_k}_{\omega_k}\left(d^{\mathrm{N}}_{\omega_k,\mu_k}\right)\right\|_{\omega_k} = \left\|\Psi(\omega_k;\mu_k) + \mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k})\left[d^{\mathrm{N}}_{\omega_k,\mu_k}\right]\right.$$

$$(5.24)\quad + \left.\int_0^1\left(\mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}\left(td^{\mathrm{N}}_{\omega_k,\mu_k}\right) - \mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k})\right)\left[d^{\mathrm{N}}_{\omega_k,\mu_k}\right]\mathrm{d}t\right\|_{\omega_k}$$

$$\leq \int_0^1\left\|\left(\mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}\left(td^{\mathrm{N}}_{\omega_k,\mu_k}\right) - \mathrm{D}\widetilde{\Psi}^{\mu_k}_{\omega_k}(0_{\omega_k})\right)\left[d^{\mathrm{N}}_{\omega_k,\mu_k}\right]\right\|_{\omega_k}\mathrm{d}t \leq \kappa c^2_{\mathrm{N}}\mu^2_{k-1},$$

where the first inequality from (5.6), (5.19), and (5.20), and the second one from (5.23). Hence, we have that, for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small,

$$
\frac{1}{\mu_k}\left\|\operatorname{grad} f\left(\mathrm{R}_{x_k}\left(d_{x_k,\mu_k}^{\mathrm{N}}\right)\right) - \sum_{i\in\mathcal{I}}\left(y_{ki} + d_{y_k,\mu_k}^{\mathrm{N}}\right)\operatorname{grad} g_i\left(\mathrm{R}_{x_k}\left(d_{x_k,\mu_k}^{\mathrm{N}}\right)\right)\right\|_{x_k}
$$

$$
\leq \frac{c}{\mu_k}\left\|\operatorname{grad}(f\circ \mathrm{R}_{x_k})\left(d_{x_k,\mu_k}^{\mathrm{N}}\right) - \sum_{i\in\mathcal{I}}\left(y_{ki} + d_{y_k,\mu_k}^{\mathrm{N}}\right)\operatorname{grad}(g_i\circ \mathrm{R}_{x_k})\left(d_{x_k,\mu_k}^{\mathrm{N}}\right)\right\|_{x_k}
$$

$$
\leq \frac{c}{\mu_k}\left\|\widetilde{\Psi}_{\omega_k}^{\mu_k}\left(d_{\omega_k,\mu_k}^{\mathrm{N}}\right)\right\|_{\omega_k} \leq \frac{c\kappa c_{\mathrm{N}}^2\mu_{k-1}^2}{\mu_k},
$$

where the first inequality follows from (5.13) with some $c > 1$ and the third one from (5.24). Since the right-hand side converges to zero as μ_{k-1} tends to zero under Assumption C.3, equation (5.14) holds for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small. Similarly, from (5.24), we have

$$
\frac{1}{\mu_k}\left\|G\left(\mathrm{R}_{x_k}\left(d_{x_k,\mu_k}^{\mathrm{N}}\right)\right)\left(y_{ki} + d_{y_k,\mu_k}^{\mathrm{N}}\right) - \mu_k\mathbf{1}\right\| \leq \frac{\kappa}{\mu_k}\left\|\widetilde{\Psi}_{\omega_k}^{\mu_k}\left(d_{\omega_k,\mu_k}^{\mathrm{N}}\right)\right\|_{\omega_k} \leq \frac{\kappa c_{\mathrm{N}}^2\mu_{k-1}^2}{\mu_k}
$$

for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small. The right-hand side converges to zero as μ_{k-1} tends to zero under Assumption C.3, which implies that (5.15) holds for any k-th iteration of Algorithm 3.1 satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small. The proof is complete. $\square$

In the following lemma, we prove the positive definiteness of $H(\omega)$ around ω^*; see (3.8) for the definition of $H(\omega)$. In the proof, we use the parallel transport along the minimizing geodesic since the operator $H(\omega)$ is defined over the space $T_x\mathcal{M}$ that varies depending on $x \in \mathcal{M}$, which is peculiar to the Riemannian setting.

LEMMA 5.10. *Under Assumptions* C.1 *and* C.6, $H(\omega)$ *is positive definite for all* $\omega \in \operatorname{str}\mathcal{F}\times\mathbb{R}_{++}^m$ *sufficiently close to* $\omega^* \in \mathcal{F}\times\mathbb{R}_+^m$.

Proof. We first derive auxiliary bounds to prove the positive definiteness of H around ω^*. First, from [23, Theorem 3], the SOSC (2.13) implies the existence of $\tau, \varepsilon \in \mathbb{R}_{++}$ such that, for any $\xi_{x^*} \in T_{x^*}\mathcal{M}\backslash\{0_{x^*}\}$,

$$
(5.25)\quad
\begin{aligned}
&\langle\operatorname{Hess}_x\mathcal{L}(\omega^*)\xi_{x^*}, \xi_{x^*}\rangle_{x^*} + \sum_{i\in\mathcal{I}}\tau\langle\operatorname{grad} g_i(x^*), \xi_{x^*}\rangle_{x^*}^2 \\
&\geq \langle\operatorname{Hess}_x\mathcal{L}(\omega^*)\xi_{x^*}, \xi_{x^*}\rangle_{x^*} + \sum_{i\in\mathcal{A}(x^*)}\tau\langle\operatorname{grad} g_i(x^*), \xi_{x^*}\rangle_{x^*}^2 \geq \varepsilon\|\xi_{x^*}\|_{x^*}^2.
\end{aligned}
$$

Second, equations (5.4) and (5.5) imply that there exists $L_{\mathrm{H}}^{\mathcal{L}} > 0$ such that

$$
(5.26)\quad
\begin{aligned}
&\left\|\operatorname{Hess}_x\mathcal{L}(\omega) - \mathrm{PT}_{x\leftarrow x^*}\circ\operatorname{Hess}_x\mathcal{L}(x^*)\circ\mathrm{PT}_{x^*\leftarrow x}\right\|_{\mathrm{op}} \\
&\leq \left(L_{\mathrm{H}}^f + \sum_{i\in\mathcal{I}}y_i L_{\mathrm{H}}^{g_i}\right)\operatorname{dist}(x, x^*) \leq L_{\mathrm{H}}^{\mathcal{L}}\operatorname{dist}(x, x^*)
\end{aligned}
$$

for any $\omega \in \mathcal{M} \times \mathbb{R}^m$ sufficiently close to ω^*. Third, for each $i \in \mathcal{I}$, since $\|\mathrm{Hess}g_i(\cdot)\|_{\mathrm{op}}$ is bounded around x^* by the twice continuous differentiability of $\{g_i\}_{i \in \mathcal{I}}$, there exists $L_{\mathrm{H}}^{g_i} > 0$ such that

$$(5.27) \qquad \left\|\mathrm{grad}g_i(x) - \mathrm{PT}_{x \leftarrow x^*}[\mathrm{grad}g_i(x^*)]\right\|_x \leq L_{\mathrm{H}}^{g_i} \mathrm{dist}(x, x^*)$$

for any $x \in \mathcal{M}$ sufficiently close to x^* by [11, Corollary 10.47]. In addition, since the parallel transport is isometric and the functions $\{g_i\}_{i \in \mathcal{I}}$ are continuously differentiable, there exists $\kappa_{\mathrm{g}} > 0$ such that, for any $x \in \mathcal{M}$ sufficiently close to x^*,

$$(5.28) \qquad \begin{aligned} &\|\mathrm{grad}g_i(x)\|_x + \left\|\mathrm{PT}_{x \leftarrow x^*}[\mathrm{grad}g_i(x^*)]\right\|_x \\ &= \|\mathrm{grad}g_i(x)\|_x + \|\mathrm{grad}g_i(x^*)\|_{x^*} \leq \kappa_{\mathrm{g}}. \end{aligned}$$

Now, we prove the positive definiteness of $H(\omega)$. Let $\xi_x \in T_x\mathcal{M}\backslash\{0_x\}$ be any nonzero vector. It follows that

$$\begin{aligned} \langle H(\omega)[\xi_x], \xi_x \rangle_x &= \langle \mathrm{Hess}_x\mathcal{L}(\omega)[\xi_x], \xi_x \rangle_x + \sum_{i \in \mathcal{I}} \frac{y_i}{g_i(x)} \langle \mathrm{grad}g_i(x), \xi_x \rangle_x^2 \\ (5.29) &\geq \langle \mathrm{Hess}_x\mathcal{L}(\omega)[\xi_x], \xi_x \rangle_x + \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{grad}g_i(x^*), \mathrm{PT}_{x^* \leftarrow x}[\xi_x] \rangle_{x^*}^2 \\ &\quad - \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{grad}g_i(x^*), \mathrm{PT}_{x^* \leftarrow x}[\xi_x] \rangle_{x^*}^2 + \sum_{i \in \mathcal{A}(x^*)} \frac{y_i}{g_i(x)} \langle \mathrm{grad}g_i(x), \xi_x \rangle_x^2, \end{aligned}$$

where the inequality follows from $i \in \mathcal{A}(x^*) \subseteq \mathcal{I}$. We derive the bound on the first two terms as follows: for all $\omega \in \mathrm{str}\,\mathcal{F} \times \mathbb{R}_{++}^m$ sufficiently close to ω^*, we have

$$\begin{aligned} &\langle \mathrm{Hess}_x\mathcal{L}(\omega)[\xi_x], \xi_x \rangle_x + \sum_{i \in \mathcal{I}} \tau \langle \mathrm{grad}g_i(x^*), \mathrm{PT}_{x^* \leftarrow x}[\xi_x] \rangle_{x^*}^2 \\ &= \langle \mathrm{Hess}_x\mathcal{L}(\omega)[\xi_x], \xi_x \rangle_x - \langle \mathrm{PT}_{x \leftarrow x^*} \circ \mathrm{Hess}_x\mathcal{L}(\omega^*) \circ \mathrm{PT}_{x^* \leftarrow x}[\xi_x], \xi_x \rangle_x \\ &\quad + \langle \mathrm{PT}_{x \leftarrow x^*} \circ \mathrm{Hess}_x\mathcal{L}(\omega^*) \circ \mathrm{PT}_{x^* \leftarrow x}[\xi_x], \xi_x \rangle_x \\ (5.30) &\quad + \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{grad}g_i(x^*), \mathrm{PT}_{x^* \leftarrow x}[\xi_x] \rangle_{x^*}^2 \\ &\geq -\left\|\mathrm{Hess}_x\mathcal{L}(\omega) - \mathrm{PT}_{x \leftarrow x^*} \circ \mathrm{Hess}_x\mathcal{L}(\omega^*) \circ \mathrm{PT}_{x^* \leftarrow x}\right\|_{\mathrm{op}} \|\xi_x\|_x^2 \\ &\quad + \langle \mathrm{Hess}_x\mathcal{L}(\omega^*) \circ \mathrm{PT}_{x^* \leftarrow x}[\xi_x], \mathrm{PT}_{x^* \leftarrow x}[\xi_x] \rangle_{x^*} \\ &\quad + \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{grad}g_i(x^*), \mathrm{PT}_{x^* \leftarrow x}[\xi_x] \rangle_{x^*}^2 \geq -L_{\mathrm{H}}^{\mathcal{L}} \mathrm{dist}(x, x^*)\|\xi_x\|_x^2 + \varepsilon\|\xi_x\|_x^2, \end{aligned}$$

where the first inequality follows from the adjoint property of the parallel transport

and the second one from (5.25) and (5.26). As for the last two terms, we have

$$
\begin{aligned}
&- \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{grad}\, g_i(x^*), \mathrm{PT}_{x^* \leftarrow x}[\xi_x] \rangle_{x^*}^2 + \sum_{i \in \mathcal{A}(x^*)} \frac{y_{ki}}{g_i(x)} \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2 \\
&= \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2 - \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{PT}_{x \leftarrow x^*}[\mathrm{grad}\, g_i(x^*)], \xi_x \rangle_x^2 \\
&\quad - \sum_{i \in \mathcal{A}(x^*)} \tau \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2 + \sum_{i \in \mathcal{A}(x^*)} \frac{y_{ki}}{g_i(x)} \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2 \\
&= \sum_{i \in \mathcal{A}(x^*)} \tau \Big(\big\langle \mathrm{grad}\, g_i(x) + \mathrm{PT}_{x \leftarrow x^*}[\mathrm{grad}\, g_i(x^*)], \xi_x \big\rangle_x \\
&\qquad\qquad \big\langle \mathrm{grad}\, g_i(x) - \mathrm{PT}_{x \leftarrow x^*}[\mathrm{grad}\, g_i(x^*)], \xi_x \big\rangle_x \Big) \\
&\quad + \sum_{i \in \mathcal{A}(x^*)} \left(\frac{y_{ki}}{g_i(x)} - \tau \right) \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2 \\
&\geq - \sum_{i \in \mathcal{A}(x^*)} \tau \Big(\|\mathrm{grad}\, g_i(x)\|_x + \big\| \mathrm{PT}_{x \leftarrow x^*}[\mathrm{grad}\, g_i(x^*)] \big\|_x \Big) \\
&\qquad \big\| \mathrm{grad}\, g_i(x) - \mathrm{PT}_{x \leftarrow x^*}[\mathrm{grad}\, g_i(x^*)] \big\|_x \|\xi_x\|_x^2 \\
&\quad + \sum_{i \in \mathcal{A}(x^*)} \left(\frac{y_{ki}}{g_i(x)} - \tau \right) \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2 \\
&\geq - \sum_{i \in \mathcal{A}(x^*)} \tau \kappa_{\mathrm{g}} L_{\mathrm{H}}^{g_i} \, \mathrm{dist}(x, x^*) \|\xi_x\|_x^2 + \sum_{i \in \mathcal{A}(x^*)} \left(\frac{y_i}{g_i(x)} - \tau \right) \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2,
\end{aligned}
$$

(5.31)

where the second inequality follows from (5.27) and (5.28). Therefore, combining (5.29) with (5.30) and (5.31) yields that, for any $\omega \in \mathrm{str}\, \mathcal{F} \times \mathbb{R}_{++}^m$ sufficiently close to ω^*,

$$
\begin{aligned}
\langle H(\omega)[\xi_x], \xi_x \rangle_x &\geq \big(\varepsilon - \big(L_{\mathrm{H}}^{\mathcal{L}} + \tau \kappa_{\mathrm{g}} L_{\mathrm{H}}^{g_i} \big) \mathrm{dist}(x, x^*) \big) \|\xi_x\|_x^2 \\
&\quad + \sum_{i \in \mathcal{A}(x^*)} \left(\frac{y_{ki}}{g_i(x)} - \tau \right) \langle \mathrm{grad}\, g_i(x), \xi_x \rangle_x^2.
\end{aligned}
$$

Since $\mathrm{dist}(x, x^*)$ and $\left\{ \frac{y_{ki}}{g_i(x)} \right\}_{i \in \mathcal{A}(x^*)}$ can be made arbitrarily small and large by retaking $\omega \in \mathrm{str}\, \mathcal{F} \times \mathbb{R}_{++}^m$ sufficiently close to ω^*, respectively, the right-hand side can be positive for any $\xi_x \in T_x \mathcal{M} \setminus \{0_{x^*}\}$, which completes the proof. $\quad\square$

Now, we prove that, for any k-th iteration satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small, Algorithm 3.2 terminates in one iteration if we employ the exact step as the search direction.

PROPOSITION 5.11. *Suppose Assumptions C.1-C.5. If 2ND_ORDER is True in Algorithm 3.1, suppose Assumption C.6 additionally. Then, the point $\left(\mathrm{R}_{x_k}\big(d_{x_k}^* \big), y_k + d_{y_k}^* \right) \in \mathrm{str}\, \mathcal{F} \times \mathbb{R}_{++}^m$ satisfies the stopping conditions (3.3) with $\mu_k > 0$ for any k-th iteration satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small.*

Proof. Recall that, from Lemma 5.2 and Proposition 5.3, the first iterate of Algorithm 3.2, that is, $\left(\mathrm{R}_{x_k}\big(d_{x_k}^* \big), y_k + d_{y_k}^* \right) \in \mathrm{str}\, \mathcal{F} \times \mathbb{R}_{++}^m$, is equivalent to the Newton

step $\left(\mathrm{R}_{x_k}\left(d^{\mathrm{N}}_{x_k,\mu_k}\right), y_k + d^{\mathrm{N}}_{y_k,\mu_k}\right)$ for any k-th iteration satisfying that ω_k is sufficiently close to ω^* and μ_{k-1} is sufficiently small. The iterate satisfies (3.3a) and (3.3b) by Lemma 5.9. Under 2ND_ORDER being True in Algorithm 3.1 and Assumption C.6 additionally, since H_k is positive definite around ω^* by Lemma 5.10, condition (3.3c) is fulfilled for any k-th iteration satisfying that ω_k is sufficiently close to ω^*. Equation (3.3d) follows from Lemmas 5.6 and 5.7. The proof is complete. □

5.2. Local near-quadratic convergence of RIPTRM. In this subsection, we prove the local near-quadratic convergence of RIPTRM when using the exact step and a specific update for the sequence of barrier parameters. To this end, we additionally assume the following:

ASSUMPTION C.7. *Algorithm* 3.1 *uses Algorithm* 3.2 *to compute* $\{\omega_k\}_k$. *In addition, Algorithm* 3.2 *uses the exact step as the search direction at every iteration.*

ASSUMPTION C.8. $\mu_{k+1} = o(\mu_k)$ *holds.*

In the following lemma, we provide an update rule for the sequence of barrier parameters that satisfies Assumptions C.3, C.4, and C.8 and will be used for RIPTRM to achieve the local near-quadratic convergence.

LEMMA 5.12. *The update rule*

$$(5.32) \qquad \mu_{k+1} \leftarrow c\mu_k^{1+r} \ \text{ with } 0 < c < 1, 0 < r < 1, \ \text{ and } 0 < \mu_0 \leq 1$$

satisfies Assumptions C.3, C.4, *and* C.8.

Proof. Note that $\{\mu_k\}_k$ is monotonically decreasing by definition. It follows that

$$\lim_{k \to \infty} \frac{\mu_k^2}{\mu_{k+1}} = \lim_{k \to \infty} \frac{\mu_k^{1-r}}{c} = 0 \ \text{and} \ \lim_{k \to \infty} \frac{\mu_{k+1}}{\mu_k} = \lim_{k \to \infty} c\mu_k^r = 0,$$

which imply Assumptions C.3 and C.8, respectively. □

Let $(\mathcal{U}, \varphi_x)$ be a chart of $\mathcal{M}$ with $x \in \mathcal{U}$ and $(\mathbb{R}^m, \varphi_y)$ be that of $\mathbb{R}^m$, where φ_y is the identity map. Define the chart of the product manifold $\mathcal{M} \times \mathbb{R}^m$ as $\varphi(\omega) := (\varphi_x(x), \varphi_y(y))$, $\overline{\omega} = (\overline{x}, \overline{y})$, and the coordinate expressions of the barrier KKT vector field as

$$
\begin{aligned}
&F(\overline{\omega}, \mu) \\
(5.33) &:= \begin{bmatrix} \mathrm{D}\varphi_x\left(\varphi_x^{-1}(\overline{x})\right)\left[\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega})\right)\right] \\ \mathrm{D}\varphi_y\left(\varphi_y^{-1}(\overline{y})\right)\left[G(\overline{x})y - \mu\mathbf{1}\right] \end{bmatrix} = \begin{bmatrix} \mathrm{D}\varphi_x\left(\varphi_x^{-1}(\overline{x})\right)\left[\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega})\right)\right] \\ G(\overline{x})y - \mu\mathbf{1} \end{bmatrix},
\end{aligned}
$$

respectively. Note that $F(\overline{\omega^*}, 0) = 0$ holds by definition. We also define

$$(5.34) \qquad\qquad \tilde{F}\left(\overline{\omega}, \overline{l}, \overline{\xi}\right) := F(\overline{\omega}, 0) - \begin{bmatrix} \overline{l} \\ \overline{\xi} \end{bmatrix}$$

for $\overline{l} \in \mathbb{R}^d$ and $\overline{\xi} \in \mathbb{R}^m$. Given $\mu > 0$, we write $F_\mu(\overline{\omega})$ for the restricted function $\overline{\omega} \mapsto F(\overline{\omega}, \mu)$. We first prove the nonsingularity of $\mathrm{D}F_\mu$ at $\overline{\omega^*}$ using Lemma 5.1:

LEMMA 5.13. *Let* $\mu \in \mathbb{R}$ *be an arbitrary scalar. Under Assumption* C.1, $\mathrm{D}F_\mu\left(\overline{\omega^*}\right)$ *is nonsingular and independent of the value of* μ.

Proof. See Appendix C.6 □

Then, we prove the existence and the uniqueness of a solution of $\tilde{F}\left(\cdot, \overline{l}, \overline{\xi}\right) = 0$ using the implicit function theorem and Lemma 5.13:

LEMMA 5.14. *Under Assumption C.1, let $\overline{\omega}(\overline{l}, \overline{\xi})$ be a solution of $\tilde{F}(\cdot, \overline{l}, \overline{\xi}) = 0$. Then, for some $\varepsilon > 0$,*

1. *the solution $\overline{\omega}(\overline{l}, \overline{\xi})$ exists and is unique with respect to $(\overline{l}, \overline{\xi})$ in the neighborhood $\mathcal{N}(\varepsilon) := \{(\overline{l}, \overline{\xi}) \in \mathbb{R}^d \times \mathbb{R}^m \colon \|\overline{l}\| + \|\overline{\xi}\| \leq \varepsilon\}$. Moreover, $\overline{\omega}(\overline{l}, \overline{\xi})$ is a continuously differentiable function of $(\overline{l}, \overline{\xi})$ in the neighborhood $\mathcal{N}(\varepsilon)$, and*
2. *for $(\overline{l}_1, \overline{\xi}_1), (\overline{l}_2, \overline{\xi}_2) \in \mathcal{N}(\varepsilon)$ sufficiently small, we have*

$$(5.35) \qquad \left\|\overline{\omega}(\overline{l}_1, \overline{\xi}_1) - \overline{\omega}(\overline{l}_2, \overline{\xi}_2)\right\| = \Theta\left(\left\|\overline{l}_1 - \overline{l}_2\right\| + \left\|\overline{\xi}_1 - \overline{\xi}_2\right\|\right).$$

Proof. See Appendix C.6 $\qquad\qquad\square$

Next, we derive an upper bound on the Euclidean distance between the $(k+1)$-th iteration and the accumulation point.

LEMMA 5.15. *Suppose Assumptions C.1-C.5 and C.7. If 2ND_ORDER is True in Algorithm 3.1, additionally suppose Assumption C.6. Then,*

$$(5.36) \qquad\qquad \left\|\overline{\omega_{k+1}} - \overline{\omega^*}\right\| = \mathcal{O}(\mu_k).$$

Proof. Let $k \in \mathbb{N}$ be an index where $\omega_{k+1} \in \operatorname{str}\mathcal{F} \times \mathbb{R}_+^m$ is sufficiently close to ω^* and $\mu_k > 0$ is sufficiently small. We have

$$(5.37) \qquad \left\|\overline{\omega_{k+1}} - \overline{\omega^*}\right\| \leq \left\|\overline{\omega_{k+1}} - \overline{\omega}(0, \mu_k\mathbf{1})\right\| + \left\|\overline{\omega}(0, \mu_k\mathbf{1}) - \overline{\omega^*}\right\|.$$

In the following, we derive the bound on each term of the right-hand side of (5.37). For the first term, it holds by Item 1 of Lemma 5.14 that the point

$$\overline{\omega}\left(\mathrm{D}\varphi_x\left(\varphi_x{}^{-1}(\overline{x_{k+1}})\right)\left[\operatorname{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega_{k+1}})\right)\right], G(\overline{x_{k+1}})y_{k+1}\right)$$

is well-defined and identical to $\overline{\omega_{k+1}}$ by the uniqueness of $\overline{\omega}(\cdot, \cdot)$. Therefore, there exist $c_1, c_2 \in \mathbb{R}_{++}$ such that, for any $k \in \mathbb{N}_0$ where ω_{k+1} is sufficiently close to ω^* and $\mu_k > 0$ is sufficiently small,

$$(5.38)$$
$$\left\|\overline{\omega_{k+1}} - \overline{\omega}(0, \mu_k\mathbf{1})\right\|$$
$$= \left\|\overline{\omega}\left(\mathrm{D}\varphi_x\left(\varphi_x{}^{-1}(\overline{x_{k+1}})\right)\left[\operatorname{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega_{k+1}})\right)\right], G(\overline{x_{k+1}})y_{k+1}\right) - \overline{\omega}(0, \mu_k\mathbf{1})\right\|$$
$$\leq c_1\left(\left\|\mathrm{D}\varphi_x\left(\varphi_x{}^{-1}(\overline{x_{k+1}})\right)\left[\operatorname{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega_{k+1}})\right)\right]\right\| + \left\|G(\overline{x_{k+1}})y_{k+1} - \mu_k\mathbf{1}\right\|\right)$$
$$\leq c_1\left(\left\|\mathrm{D}\varphi_x\left(\varphi_x{}^{-1}(\overline{x_{k+1}})\right)\right\|_{\mathrm{op}}\left\|\operatorname{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega_{k+1}})\right)\right\| + \left\|G(\overline{x_{k+1}})y_{k+1} - \mu_k\mathbf{1}\right\|\right)$$
$$\leq c_1\left(\left\|\mathrm{D}\varphi_x\left(\varphi_x{}^{-1}(\overline{x_{k+1}})\right)\right\|_{\mathrm{op}}\varepsilon_{\mathcal{L}}(\mu_k) + \varepsilon_C(\mu_k)\right) \leq c_2\mu_k,$$

where the first inequality follows from Item 2 of Lemma 5.14, the third one from (3.3a) and (3.3b), and the last one from the boundedness of $\mathrm{D}\varphi_x(\cdot)$ around x^* and (5.1). Moreover, since $\overline{\omega}(0,0) = \overline{\omega^*}$ holds by definition, the second term in the right-hand side of (5.37) can be bounded as

$$(5.39) \qquad \left\|\overline{\omega}(0, \mu_k\mathbf{1}) - \overline{\omega^*}\right\| = \left\|\overline{\omega}(0, \mu_k\mathbf{1}) - \overline{\omega}(0,0)\right\| \leq c_3\mu_k$$

for some $c_3 > 0$, where the inequality follows from Item 2 of Lemma 5.14 again. Therefore, combining (5.37) with (5.38) and (5.39) yields

$$(5.40) \qquad\qquad \left\|\overline{\omega_{k+1}} - \overline{\omega^*}\right\| \leq (c_2 + c_3)\mu_k.$$

In the following, we regard the manipulation $y + d_y$ as a retraction on $\mathbb{R}^m$; that is, we define $R_y(d_y) := y + d_y$. We also define the retraction on the product manifold $\mathcal{M} \times \mathbb{R}^m$ as $R_\omega(d_\omega) = (R_x(d_x), R_y(d_y))$. We have

$$\mathrm{dist}(\omega_{k+2}, \omega^*) \leq \mathrm{dist}(\omega_{k+2}, \omega_{k+1}) + \mathrm{dist}(\omega_{k+1}, \omega^*)$$

(5.41)
$$= \mathrm{dist}\left(R_{\omega_{k+1}}\left(d^{\mathrm{N}}_{\omega_{k+1}, \mu_{k+1}}\right), R_{\omega_{k+1}}\left(0_{\omega_{k+1}}\right)\right) + \mathrm{dist}(\omega_{k+1}, \omega^*)$$

$$\leq a_2 \left\| d^{\mathrm{N}}_{\omega_k, \mu_k} \right\|_{\omega_{k+1}} + a_3 \left\| \overline{\omega_{k+1}} - \overline{\omega^*} \right\| \leq (a_2 c_{\mathrm{N}} + a_3(c_2 + c_3))\mu_k$$

for some $a_2, a_3 \in \mathbb{R}_+$, where the equality follows from Proposition 5.11 and (2.2a), the second inequality from Lemmas 2.1 and 2.3, and the third one from Lemma 5.5 and (5.40). Since, without loss of generality, the right-hand side can be made arbitrarily small, equation (5.41) implies that the point ω_{k+2} remains in a sufficiently small neighborhood of ω^*. Therefore, using the argument above inductively, we conclude that equation (5.36) holds. $\square$

Using Lemma 5.15, we derive the lower bound on the Euclidean distance between $\overline{\omega_{k+1}}$ and $\overline{\omega^*}$, and provide the tight bound.

LEMMA 5.16. *Suppose Assumptions C.1-C.5 and C.7. If 2ND_ORDER is True in Algorithm 3.1, additionally suppose Assumption C.6. Then,*

$$\left\| \overline{\omega_{k+1}} - \overline{\omega^*} \right\| = \Theta(\mu_k).$$

Proof. Recall that, from Lemma 5.15, the point ω_{k+1} can be made arbitrarily close to ω^* by taking $\mu_k > 0$ sufficiently small. We have

$$F(\overline{\omega_{k+1}}, \mu_k) = F(\overline{\omega^*}, 0) + \mathrm{D}F(\cdot, \mu_k)(\overline{\omega^*})\left[\overline{\omega_{k+1}} - \overline{\omega^*}\right] + \mathrm{D}F(\overline{\omega^*}, \cdot)(0)[\mu_k - 0]$$

(5.42)
$$+ \int_0^1 \mathrm{D}\left(F\left((\overline{\omega^*}, 0) + t(\overline{\omega_{k+1}}, \mu_k)\right) - F(\overline{\omega^*}, 0)\right)\left[\left(\overline{\omega_{k+1}} - \overline{\omega^*}, \mu_k - 0\right)\right] \mathrm{d}t$$

$$= \mathrm{D}F(\cdot, \mu_k)(\overline{\omega^*})\left[\overline{\omega_{k+1}} - \overline{\omega^*}\right] + \begin{bmatrix} 0 \\ -\mu_k \mathbf{1} \end{bmatrix} + r,$$

where the second equality follows from $F(\overline{\omega^*}, 0) = 0$ and

(5.43)
$$r := \int_0^1 \mathrm{D}\left(F\left((\overline{\omega^*}, 0) + t(\overline{\omega_{k+1}}, \mu_k)\right) - F(\overline{\omega^*}, 0)\right)\left[\left(\overline{\omega_{k+1}} - \overline{\omega^*}, \mu_k - 0\right)\right] \mathrm{d}t.$$

We derive bounds on each term of (5.42) as follows: First, we consider the first term. Let

(5.44)
$$\dot{\omega}(0) := \mathrm{D}F(\cdot, \mu_k)^{-1}(\overline{\omega^*}) \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix} \in \mathbb{R}^{d+m+1} \setminus \{0\}.$$

Since $\mathrm{D}F(\cdot, \mu_k)(\overline{\omega^*})$ is nonsingular and independent of μ_k by Lemma 5.13, we also define $\upsilon := \left\| \mathrm{D}F(\cdot, \mu_k)^{-1}(\overline{\omega^*}) \right\|_{\mathrm{op}} > 0$ regardless of the barrier parameter. Since $\|\mathrm{D}\varphi_x(\cdot)\|_{\mathrm{op}}$ is bounded around x^* due to the continuous differentiability of F, we obtain

$$\left\| \mathrm{D}F(\cdot, \mu_k)^{-1}(\overline{\omega^*}) F(\overline{\omega_{k+1}}, \mu_k) \right\| \leq \upsilon \|F(\overline{\omega_{k+1}}, \mu_k)\|$$

(5.45)
$$\leq \upsilon \left(\left\| \mathrm{D}\varphi_x\left(\varphi_x^{-1}(\overline{x})\right) \right\|_{\mathrm{op}} \left\| \mathrm{grad}_x \mathcal{L}\left(\varphi^{-1}(\overline{\omega})\right) \right\| + \|G(\overline{x})y - \mu\mathbf{1}\| \right)$$

$$\leq \frac{\mu_k}{2} \|\dot{\omega}(0)\|$$

for any $k \in \mathbb{N}_0$ satisfying that μ_k is sufficiently small, where the second inequality follows from (5.33) and the third one from Lemma 5.9 with sufficiently small coefficients $c_{\mathcal{L}}, c_c \in \mathbb{R}_{++}$. For the third term of (5.42), we have

$$
(5.46) \quad \begin{aligned}
\|r\| &\leq \int_0^1 \left\| \mathrm{D}\big(F\big((\overline{\omega^*},0) + t(\overline{\omega_{k+1}},\mu_k)\big) - F(\overline{\omega^*},0)\big)\right\|_{\mathrm{op}} \left\|(\overline{\omega_{k+1}} - \overline{\omega^*}, \mu_k)\right\| \mathrm{d}t \\
&= o\big(\|(\overline{\omega_{k+1}} - \overline{\omega^*}, \mu_k)\|\big) = o\big(\|\overline{\omega_{k+1}} - \overline{\omega^*}\| + \mu_k\big) = o(\mu_k),
\end{aligned}
$$

where the first inequality follows from (5.43), the first equality from the continuous differentiability of F, and the last equality from (5.36). Therefore, for any $k \in \mathbb{N}_0$ satisfying that μ_k is sufficiently small, we have

$$
(5.47)
$$
$$
\begin{aligned}
&\left\|\overline{\omega_{k+1}} - \overline{\omega^*}\right\| \\
&= \left\| \mathrm{D}F(\cdot,\mu_k)^{-1}(\overline{\omega^*})\begin{bmatrix} 0 \\ \mu_k\mathbf{1} \end{bmatrix} + \mathrm{D}F(\cdot,\mu_k)^{-1}(\overline{\omega^*})[F(\overline{\omega_{k+1}},\mu_k)] - \mathrm{D}F(\cdot,\mu_k)^{-1}(\overline{\omega^*})r \right\| \\
&\geq \left\| \mathrm{D}F(\cdot,\mu_k)^{-1}(\overline{\omega^*})\begin{bmatrix} 0 \\ \mu_k\mathbf{1} \end{bmatrix} \right\| - \left\| \mathrm{D}F(\cdot,\mu_k)^{-1}(\overline{\omega^*})F(\overline{\omega_{k+1}},\mu_k) \right\| \\
&\quad - \left\| \mathrm{D}F(\cdot,\mu_k)^{-1}(\overline{\omega^*})r \right\| \geq \mu_k\|\dot{\omega}(0)\| - \frac{\mu_k}{2}\|\dot{\omega}(0)\| - v\|r\| = \frac{\mu_k}{2}\|\dot{\omega}(0)\| - v\|r\| > 0,
\end{aligned}
$$

where the equality follows from (5.42) and the nonsingularity of $\mathrm{D}F(\cdot,\mu_k)$ at $\overline{\omega^*}$, the second inequality from (5.44) and (5.45), and the last one from (5.46). Hence, from (5.44), (5.46), and (5.47), we have $\|\overline{\omega_{k+1}} - \overline{\omega^*}\| = \Omega(\mu_k)$, which, together with Lemma 5.15, completes the proof. □

Using Lemma 5.16, we establish the local convergence of RIPTRM. We first prove its local superlinear convergence property and then local near-quadratic convergence when using the update rule (5.32). We refer readers to [1, Definition 4.5.2] for the definition of local convergence in Riemannian optimization.

THEOREM 5.17. *Suppose Assumptions C.1-C.5, C.7, and C.8. If 2ND_ORDER is True in Algorithm 3.1, additionally suppose Assumption C.6. Then, the sequence $\{\omega_k\}_k$ superlinearly converges to ω^*. Moreover, the sequence $\{\omega_k\}_k$ converges near-quadratically[1] if the sequence of the barrier parameters $\{\mu_k\}_k$ is updated according to (5.32).*

Proof. By Lemma 5.16, we have $\frac{\|\overline{\omega_{k+1}} - \overline{\omega^*}\|}{\|\overline{\omega_k} - \overline{\omega^*}\|} = \Theta\left(\frac{\mu_k}{\mu_{k-1}}\right)$, which, together with Assumption C.8, implies the superlinear convergence, that is,

$$
\lim_{k\to\infty} \frac{\|\overline{\omega_{k+1}} - \overline{\omega^*}\|}{\|\overline{\omega_k} - \overline{\omega^*}\|} = 0.
$$

If we employ the update (5.32) with any $r \in (0,1)$, it follows that

$$
\frac{\|\overline{\omega_{k+1}} - \overline{\omega^*}\|}{\|\overline{\omega_k} - \overline{\omega^*}\|^{1+r}} = \Theta\left(\frac{\mu_{k+1}}{\mu_k^{1+r}}\right) = \Theta(1),
$$

which is actually the local near-quadratic convergence. The proof is complete. □

---

[1]We say that the generated sequence $\{\omega_k\}_k$ converges near-quadratically if, for any given $r \in \mathbb{R}$ with $0 < r < 1$, the generated sequence $\{\omega_k\}_k$ converges to ω^* with order at least $1 + r$. We refer [1, Definition 4.5.2] for the definition of the order of the convergence.

6. Numerical experiments. In this section, we present numerical experiments on stable linear system identification and the minimization of the Rosenbrock function on the Grassmann manifold, demonstrating the efficiency of RIPTRM. For comparison, we also solve these problems using other Riemannian algorithms. All the experiments are implemented in Python with Pymanopt 2.2.0 [64] and executed on a MacBook Pro with an Apple M1 Max chip and 64GB of memory.

6.1. Problem settings.

6.1.1. Stable linear system identification. Linear system identification is the problem of estimating a linear system from observed data and prior knowledge. The problem setting is from [50] with modifications to the dimension of the system and the initial points.

Define $\mathrm{Skew}(n) := \left\{ X \in \mathbb{R}^{n \times n} \colon X = -X^\top \right\}$ and $\mathrm{Sym}_{++}(n) := \{ X \in \mathbb{R}^{n \times n} \colon X \succ 0 \}$. Let $\mathcal{M}_{\mathrm{stab}} := \mathrm{Skew}(n) \times \mathrm{Sym}_{++}(n) \times \mathrm{Sym}_{++}(n)$ be the product manifold representing the set of the stable matrices via the parameterization $A = (J - R)Q$ with $(J, R, Q) \in \mathcal{M}_{\mathrm{stab}}$. Given the noisy observations $\{x_j\}_{j=1}^N \subseteq \mathbb{R}^n$ from the true stable system $A^{\mathrm{true}} \in \mathbb{R}^{n \times n}$, we estimate the linear system to match the observed data. Specifically, we identify the system to satisfy the stability via manifold formulation and adhere to the prior knowledge expressed as nonlinear constraints. In the experiment, we consider the box-type constraints. Formally, we address the following optimization problem:

$$\underset{(J, R, Q) \in \mathcal{M}_{\mathrm{stab}}}{\mathrm{minimize}} \quad \frac{1}{N\|x_0\|} \sum_{j=1}^{N-1} \|x_{j+1} - (I + h(J - R)Q)x_j\|_{\mathrm{F}}$$

(6.1a) $\qquad$ subject to $\qquad l_{ij} \leq e_i^\top (J - R)Q e_j \leq r_{ij}, \quad (i, j) \in \mathcal{I}_1 \cup \mathcal{I}_2,$

(6.1b) $\qquad\qquad\qquad\qquad k_{ij}^2 \leq \left(e_i^\top (J - R)Q e_j - c_{ij} \right)^2, \quad (i, j) \in \mathcal{I}_2,$

where $h > 0$ is the sampling interval, $\mathcal{I}_1, \mathcal{I}_2 \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$ are disjoint index sets, $e_i \in \mathbb{R}^n$ is the i-th standard basis, and $l_{ij}, r_{ij}, k_{ij}, c_{ij} \in \mathbb{R}_{++}$ satisfy $l_{ij} \leq A_{ij}^{\mathrm{true}} \leq r_{ij}$ for $(i, j) \in \mathcal{I}_1$ and $A_{ij}^{\mathrm{true}} \in [l_{ij}, c_{ij} - k_{ij}] \cup [c_{ij} + k_{ij}, r_{ij}]$ for $(i, j) \in \mathcal{I}_2$.

Input. Our implementation follows that in [50] with modifications. We consider the case $n = 5$, $h = 0.02$, $N = 20$, $|\mathcal{I}_1| = \lfloor 0.2n^2 \rfloor$, and $|\mathcal{I}_2| = \lfloor 0.1n^2 \rfloor$. For the initial points, we use interior points numerically obtained in advance: we solve $\mathrm{minimize}_{(J,R,Q) \in \mathcal{M}_{\mathrm{stab}}} 0$ subject to (6.1a) and (6.1b) using RALM until a solution with a residual of 10^{-2} is obtained. We initialize the dual variables with the vector of ones. We randomly generate 20 initial points.

6.1.2. Rosenbrock function minimization. The Rosenbrock function is a well-known benchmark in the field of mathematical optimization. We formulate the following optimization problem on the Grassmann manifold:

$$\underset{X \in \mathrm{Gr}(n, k)}{\mathrm{minimize}} \quad \sum_{m=1}^{nk-1} \alpha \left([\mathrm{vec}(X)]_{m+1} - [\mathrm{vec}(X)]_m \right)^2 + \left(1 - [\mathrm{vec}(X)]_m \right)^2$$

$$\text{subject to} \quad X \geq c \cdot \mathbf{1}\mathbf{1}^\top,$$

where $\mathrm{vec} \colon \mathbb{R}^{n \times k} \to \mathbb{R}^{nk}$ is the vectorization operator that rearranges the rows of a matrix into a vector, and $\mathrm{Gr}(n, k) := \left\{ \mathrm{span}(X) \colon X \in \mathbb{R}^{n \times k}, X^\top X = I_k \right\}$ is the Grassmann manifold.

Input. We consider the case $(n, k) = (5, 3)$, $\alpha = 10^7$, and $c = -0.01$. The initial point is set to $[I_k | \mathbf{0}]^\top \in \mathbb{R}^{n \times k}$. We initialize the dual variables with the vector of ones.

6.2. Experimental environment. We compare the following algorithms:
- **RIPTRM (tCG)**: Riemannian interior point trust region method (Algorithms 3.1 and 3.2) with the search direction obtained by the truncated conjugate gradient (tCG) method.
- **RIPTRM (exact)**: Riemannian interior point trust region method (Algorithms 3.1 and 3.2) with the exact step.
- RIPM: Riemannian interior point method [40, Algorithm 4].
- RSQO: Riemannian sequential quadratic optimization [49].
- RALM: Riemannian augmented Lagrangian method [44].

In addition to our RIPTRMs, we implemented RIPM, RSQO, and RALM in Python following their original MATLAB implementations. We will explain the details of the search direction by tCG and the exact steps in RIPTRM at the end of this subsection. To measure the deviation of an iterate from the set of KKT points, we introduce the residual defined as

$$\text{KKT\_residual}(\omega)$$

$$:= \sqrt{\left\|\text{grad}_x \mathcal{L}(\omega)\right\|_x^2 + \sum_{i \in \mathcal{I}}\left(\min(0, y_i)^2 + \min(0, g_i(x))^2 + (y_i g_i(x))^2\right) + \text{Manvio}(x)^2},$$

where the first two terms are from the KKT conditions $=(2.10)$, and Manvio: $\mathcal{M} \to \mathbb{R}$ measures the manifold constraint violations. In this experiments, we consider $\text{Manvio}(J, R, Q) = \|J + J^\top\|_\text{F} + \|R - R^\top\|_\text{F} + \|Q - Q^\top\|_\text{F} + \tau_\succ(R) + \tau_\succ(Q)$ for $\mathcal{M}_\text{stab}$, where $\tau_\succ(X) := +\infty$ if the matrix X has a negative real eigenvalue and otherwise 0. For Grassmann manifold, we define $\text{Manvio}(X) := +\infty$ if $\dim \text{span } X \neq k$ and otherwise 0. In practice, these violations are negligible in Riemannian optimization due to the use of the retractions. The stopping criterion is based on runtime limits; each algorithm is run for 240 seconds.

Let $s > 0$ be the scale of the manifold implemented in Pymanopt. For RIPTRM, we set the parameters as $\Delta_0^\text{init} = \frac{s}{8}, \Delta_\text{min}^\text{init} = 10^{-15}, \Delta_\text{max} = 10, \varepsilon_\mathcal{L}(\mu) = \mu, \varepsilon_C(\mu) = 0.001\mu, \varepsilon_S(\mu) = \mu, \rho' = 0.1, \gamma = 0.25$. We apply the rule in (5.32) with $\mu_0 = 0.1$, $c = 0.5$, and $r = 0.01$ to generate the sequence of the barrier parameters. We employ the clipping in (4.34) with $\underline{c} = 0.5$ and $\tilde{c} = 10^{20}$ to update the dual variables. For the parameters of RIPM, RSQO, and RALM, we use their default settings except that we employ the gradient descent as the subsolver of RALM instead of limited-memory BFGS, the subsolver used in the original MATLAB implementation, as it is not implemented in Pymanopt 2.2.0.

Implementation details of RIPTRM. In this paper, our RIPTRM employs the two search directions. The first is obtained by the tCG method [11, Algorithm 6.4]. The quality of the search direction is guaranteed to be better than that of the Cauchy step; that is, the direction obtained by tCG satisfies Assumption B.1 with $\kappa_C = \frac{1}{2}$ [11, Exercise 6.26]. Nevertheless, this direction may not satisfy Assumption B.11 [11, Exercise 6.27]. We use the tCG method implementation available in the Riemannian trust region method for unconstrained optimization in Pymanopt with its default setting. We also set 2ND_ORDER=False for this setting.
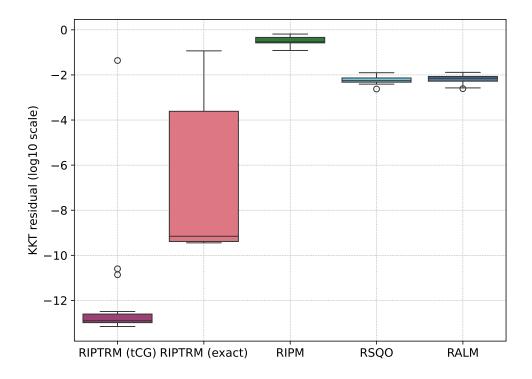
FIG. 1. *Box plots of best residuals among 20 instances for stable linear system identification*

On the other hand, computing the exact step is challenging since the trust region subproblem (3.10) is generally nonconvex. To compute the exact step, we employ the algorithm by Adachi et al. [3], which is based on the generalized eigenvalue problem. This direction satisfies Assumptions B.1 and B.11 with $\kappa_C = \frac{1}{2}$ and $\kappa_E = \frac{1}{2}$ and also satisfies Assumption C.7. We implemented a Python version of the algorithm since the original code is written in MATLAB.[2] We also set 2ND_ORDER=True for this setting.

6.3. Results and discussion.

6.3.1. Stable linear system identification. We applied the algorithms to (6.1), starting from 20 initial points. Figure 1 shows box plots of the minimum residuals computed by the solvers across 20 outputs. RIPTRM (tCG) robustly solved the instances with the highest accuracy. While RIPTRM (exact) also performed well, its robustness appeared to be inferior to RIPTRM (tCG). Figure 2 illustrates a representative example of a residual over time for the algorithms. Here, we omit the inner iterations of RIPTRMs and only plot the outer iterations for clarity. We observe that RIPTRM (tCG) successfully solved the instance with the highest accuracy. Compared with RIPTRM and RIPM, these results indicate that the trust region strategy provide accurate solutions than the line search method in Riemannian constrained optimization.

---

[2]https://people.maths.ox.ac.uk/nakatsukasa/codes.htm

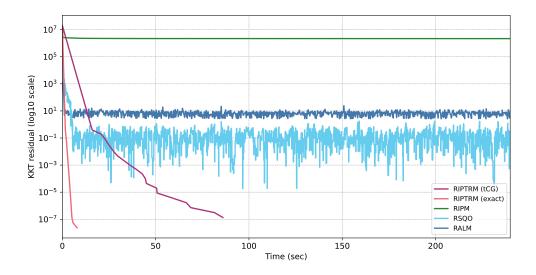FIG. 2. *Residuals over time on a representative case of stable linear system identification.*

FIG. 3. *Residual over time for Rosenbrock function minimization*

6.3.2. Rosenbrock function minimization. We applied the algorithms to the artificial instance of (6.2). Figure 3 shows the residuals of the algorithms over time. We see that RIPTRM (exact) successfully solved the instance with the fastest speed and the highest accuracy. In addition, we also measured the second-order stationarity (2.15) in the following manner; the i-th inequality constraint is regarded as active if it satisfies $g_i(X) < 10^{-6}$. We then identify $\mathcal{C}_{\mathrm{w}}$ (2.16) at X and compute the minimum eigenvalue of the Hessian of the Lagrangian over $\mathcal{C}_{\mathrm{w}}(X)$ as the second-order stationarity. Note that, for this instance, the minimum eigenvalue at the initial point is -2.000×10^7, indicating that the Hessian of the Lagrangian has the large negative

Fig. 4. *Second-order stationarity over time for Rosenbrock function minimization*

eigenvalue in this problem. Figure 4 illustrates the second-order stationarity using an arc-tangent scale. For clarity, we omit the inner iterations of RIPTRMs and plot only their outer iterations. For RIPM, RSQO, and RALM, values are plotted every 15 iterations to improve readability. The black dashed line in the figure represents zero, indicating that values above the line correspond to a positive minimum eigenvalue and signify that the second-order stationarity condition is satisfied. We observe that RIPTRM (exact) rapidly achieved and maintained the minimum eigenvalues above the line, indicating successful computation of an SOSP. This may suggest that the exact steps are robust and efficient for problem instances where the Hessian of the Lagrangian has a large negative eigenvalue, as exact steps can incorporate the information of the negative eigenvalues. In contrast, the second-order stationarity of RIPTRM (tCG), RALM, and RSQO oscillated between positive and negative ranges, and that of RIPM stays negative. This behavior may be attributed to the lack of second-order convergence properties in these algorithms.

7. Concluding remarks. In this paper, we proposed RIPTRM, composed of Algorithms 3.1 and 3.2, for solving RICO (1.1). We analyzed the limiting behavior of Algorithm 3.1 in Theorem 4.2, and established its global convergence of Algorithm 3.1 to an AKKT point and an SOSP in Theorem 4.4 and Corollary 4.5, respectively. For Algorithm 3.2, we proved its consistency and the global convergence to a point satisfying the stopping conditions in Proposition 4.10 and Theorem 4.19. In Section 4.4, we introduced the clipping technique (4.33) for the updating the dual variables in Algorithm 3.2. In Theorem 5.17, we established the local near-quadratic convergence of RIPTRM. In Section 6, we presented numerical experiments. The results indicate that RIPTRMs find solutions more accurately compared to an existing Riemannian interior point method and other algorithms. We also introduced an eigenvalue-based subsolver for RIPTRM to obtain the exact search direction by solving the trust region subproblems. We observed that its performance is promising in an instance where the Hessian of the Lagrangian has a large negative eigenvalue.

In closing, we discuss future directions for further development of RIPTRM:

1. **Treatment of equality constraints and slack variables:** In Euclidean optimization, several IPTRMs have been designed to manage both inequality and equality constraints although their guarantees of the global convergence guarantees are first-order [14, 15, 72]. Furthermore, these methods incorporate slack variables, which allow the algorithms to start from infeasible initial points. Integrating such techniques into our method could broaden its applicability.

2. **Use of filter or funnel method:** Our RIPTRM currently uses the log barrier function as a merit function to ensure the global convergence. In the Euclidean optimization, filter [66, 60] and funnel methods [21] are also employed as alternative globalization strategies, demonstrating improved practical performance [8, 39]. Extending these strategies to the Riemannian setting would enhance the efficiency and robustness of our method.

3. **Handle of degenerate problems:** In this paper, we assume the LICQ in Assumptions C.1 and A.3 for the global convergence to an SOSP and the local convergence, respectively. In the Euclidean setting, the LICQ has been relaxed for IPMs [70, 67]. Extending these techniques to the Riemannian case is one of the future work.

4. **Local convergence to non-strict local minima:** In our local convergence analysis, we assume the SOSC condition as stated in Assumption C.1, which implies the strict local optimality. Considering local convergence to non-strict local minima presents a challenging direction for future research. In the unconstrained setting, the local convergence properties of trust region methods to such minima significantly depend on the choice of the search direction. Rebjock and Boumal [52] demonstrated that trust region methods can fail to achieve local convergence when using the exact step as the search direction. In contrast, the direction computed by the tCG method retains the local convergence properties of trust region methods [51] under assumptions. Extending these results to the constrained setting and improving the local convergence properties when using the exact step would be an interesting avenue for future work.

5. **Complexity analysis:** Non-asymptotic complexity analysis for algorithms in constrained Euclidean optimization is an active area of research [17]. However, complexity analyses for IPMs are still limited to the cases where the barrier parameter is fixed due to the non-Lipschitzness of the log barrier function [34, 7]. The overall complexity of IPMs for nonlinear optimization problems remains an open question even in the Euclidean context. It would be an significant contribution to provide the overall complexity for IPMs and extend it to the Riemannian case.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, Princeton, 2008.

[2] A. Acuaviva, V. Makam, H. Nieuwboer, D. Pérez-García, F. Sittner, M. Walter, and F. Witteveen, *The minimal canonical form of a tensor network*, in IEEE Symposium on Foundations of Computer Science, 2023, pp. 328–362.

[3] S. Adachi, S. Iwata, Y. Nakatsukasa, and A. Takeda, *Solving the trust-region subproblem by a generalized eigenvalue problem*, SIAM Journal on Optimization, 27 (2017), pp. 269–291.

[4] N. Agarwal, N. Boumal, B. Bullins, and C. Cartis, *Adaptive regularization with cubics on*

manifolds, Mathematical Programming, 188 (2021), pp. 85–134.

[5] R. ANDREANI, K. R. COUTO, O. P. FERREIRA, AND G. HAESER, *Constraint qualifications and strong global convergence properties of an augmented Lagrangian method on Riemannian manifolds*, SIAM Journal on Optimization, 34 (2024), pp. 1799–1825.

[6] R. ANDREANI, K. R. COUTO, O. P. FERREIRA, G. HAESER, AND L. F. PRUDENTE, *Global convergence of an augmented Lagrangian method for nonlinear programming via Riemannian optimization*, 2024, https://optimization-online.org/?p=27595.

[7] S. ARAHATA, T. OKUNO, AND A. TAKEDA, *Complexity analysis of interior-point methods for second-order stationary points of nonlinear semidefinite optimization problems*, Computational Optimization and Applications, 86 (2023), pp. 555–598.

[8] H. Y. BENSON, R. J. VANDERBEI, AND D. F. SHANNO, *Interior-point methods for nonconvex nonlinear programming: filter methods and merit functions*, Computational Optimization and Applications, 23 (2002), pp. 257–272.

[9] R. BERGMANN AND R. HERZOG, *Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds*, SIAM Journal on Optimization, 29 (2010), pp. 2423–2444.

[10] E. G. BIRGIN, O. P. FERREIRA, G. HAESER, N. MACULAN, L. M. RAMIREZ, AND L. F. PRUDENTE, *Smoothing ℓ_1-exact penalty methiod for intrinsically constarined Riemannian optimization problems*, 2025, https://optimization-online.org/?p=28986.

[11] N. BOUMAL, *An introduction to optimization on smooth manifolds*, Cambridge University Press, Cambridge, 2023.

[12] N. BOUMAL, P.-A. ABSIL, AND C. CARTIS, *Global rates of convergence for nonconvex optimization on manifolds*, IMA Journal of Numerical Analysis, 39 (2019), pp. 1–33.

[13] S. BROSSETTE, A. ESCANDE, AND A. KHEDDAR, *Multicontact postures computation on manifolds*, IEEE Transactions on Robotics, 34 (2018), pp. 1252–1265.

[14] R. H. BYRD, J. C. GILBERT, AND J. NOCEDAL, *A trust region method based on interior point techniques for nonlinear programming*, Mathematical Programming, 89 (2000), pp. 149–185.

[15] R. H. BYRD, G. LIU, AND J. NOCEDAL, *On the local behavior of an interior point method for nonlinear programming*, in Numerical Analysis 1997, Chapman and Hall/CRC, Harlow, 1997, pp. 37–56.

[16] R. CANARY, D. EPSTEIN, AND A. MARDEN, *Fundamentals of hyperbolic manifolds: selected expositions*, Cambridge University Press, Cambridge, 2006.

[17] C. CARTIS, N. I. M. GOULD, AND P. L. TOIN, *Evaluation complexity of algorithms for nonconvex optimization*, Society for Industrial and Applied Mathematics, Philadelphia, 2022.

[18] A. R. CONN, N. I. M. GOULD, D. ORBAN, AND P. L. TOINT, *A primal-dual trust-region algorithm for non-convex nonlinear programming*, Mathematical Programming, 87 (2000), pp. 215–249.

[19] A. R. CONN, N. I. M. GOULD, AND P. L. TOINT, *Trust region methods*, Society for Industrial and Applied Mathematics, Philadelphia, 2000.

[20] C. CRISCITIELLO AND N. BOUMAL, *Efficiently escaping saddle points on manifolds*, in Advances in Neural Information Processing Systems, 2019, pp. 5987–5997.

[21] F. E. CURTIS, N. I. M. GOULD, D. P. ROBINSON, AND P. L. TOINT, *An interior-point trust-funnel algorithm for nonlinear optimization*, Mathematical Programming, 161 (2017), pp. 73–134.

[22] F. R. DE OLIVEIRA AND O. P. FERREIRA, *Newton method for finding a singularity of a special class of locally Lipschitz continuous vector fields on Riemannian manifolds*, Journal of Optimization Theory and Applications, 185 (2020), pp. 522–539.

[23] G. DEBREU, *Definite and semidefinite quadratic forms*, Econometrica, 20 (1952), pp. 295–300.

[24] K. DENG, J. HU, J. WU, AND Z. WEN, *Oracle complexities of augmented Lagrangian methods for nonsmooth manifold optimization.* arXiv:2404.05121, 2024.

[25] K. DENG AND Z. PENG, *A manifold inexact augmented Lagrangian method for nonsmooth optimization on Riemannian submanifolds in Euclidean space*, IMA Journal of Numerical Analysis, 43 (2023), pp. 1653–1684.

[26] M. P. DO CARMO, *Riemannian Geometry*, Birkhäuser, Basel, 1992.

[27] T. A. FERNANDES, O. P. FERREIRA, AND J. YUAN, *On the superlinear convergence of Newton's method on Riemannian manifolds*, Journal of Optimization Theory and Applications, 173 (2017), pp. 828–843.

[28] A. FORSGREN, P. E. GILL, AND M. H. WRIGHT, *Interior methods for nonlinear optimization*, SIAM Review, 44 (2002), pp. 525–597.

[29] K. A. GALLIVAN, C. QI, AND P.-A. ABSIL, *A Riemannian Dennis-Moré condition*, in High-Performance Scientific Computing, Springer, London, 2012, pp. 281–293.

[30] C. GEIERSBACH, T. SUCHAN, AND K. WELKER, *Stochastic augmented Lagrangian method in*

Riemannian shape manifolds, Journal of Optimization Theory and Applications, (2024). to appear.

[31] N. I. M. GOULD, D. ORBAN, A. SARTENAER, AND P. L. TOINT, *Superlinear convergence of primal-dual interior point algorithms for nonlinear programming*, SIAM Journal on Optimization, 11 (2001), pp. 974–1002.

[32] N. I. M. GOULD, D. ORBAN, AND P. L. TOINT, *Numerical methods for large-scale nonlinear optimization*, Acta Numerica, 14 (2005), pp. 299–361.

[33] F. GOYENS AND C. W. ROYER, *Riemannian trust-region methods for strict saddle functions with complexity guarantees.* arXiv:2402.07614v2, 2024.

[34] O. HINDER AND Y. YE, *Worst-case iteration bounds for log barrier methods on problems with nonconvex constraints*, Mathematics of Operations Research, (2023). to appear.

[35] H. HIRAI, H. NIEUWBOER, AND M. WALTER, *Interior-point methods on manifolds: theory and applications*, in IEEE Symposium on Foundations of Computer Science, 2023, pp. 2021–2030.

[36] W. HUANG, P.-A. ABSIL, AND K. A. GALLIVAN, *A Riemannian symmetric rank-one trust-region method*, Mathematical Programming, 150 (2015), pp. 179–216.

[37] W. HUANG, P.-A. ABSIL, AND K. A. GALLIVAN, *A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems*, SIAM Journal on Optimization, 28 (2018), pp. 470–495.

[38] H. KASAI AND B. MISHRA, *Inexact trust-region algorithms on Riemannian manifolds*, in Advances in Neural Information Processing Systems, 2018, pp. 4252–4265.

[39] D. KIESSLING, S. LEYFFER, AND C. VANARET, *A unified funnel restoration SQP algorithm.* arXiv:2404.09208, 2024.

[40] Z. LAI AND A. YOSHISE, *Riemannian interior point methods for constrained optimization on manifolds*, Journal of Optimization Theory and Applications, 201 (2024), pp. 433–469.

[41] J. M. LEE, *Introduction to smooth manifolds*, Springer, New York, second ed., 2012.

[42] J. M. LEE, *Introduction to Riemannian manifolds*, Springer, New York, second ed., 2018.

[43] E. LEVIN, J. KILEEL, AND N. BOUMAL, *The effect of smooth parameterizations on nonconvex optimization landscapes*, Mathematical Programming, (2024). to appear.

[44] C. LIU AND N. BOUMAL, *Simple algorithms for optimization on Riemannian manifolds with constraints*, Applied Mathematics & Optimization, 82 (2020), pp. 949–981.

[45] Y. LUO, X. LI, AND A. ZHANG, *Nonconvex factorization and manifold formulations are almost equivalent in low-rank matrix optimization*, INFORMS Journal on Optimization, (2024). to appear.

[46] Y. LUO AND N. G. TRILLOS, *Nonconvex matrix factorization is geodesically convex: global landscape analysis for fixed-rank matrix optimization from a Riemannian perspective.* arXiv:2209.15130, 2022.

[47] Y. NARUSHIMA, S. NAKAYAMA, M. TAKEMURA, AND H. YABE, *Memoryless quasi-Newton methods based on the spectral-scaling Broyden family for Riemannian optimization*, Journal of Optimization Theory and Applications, 197 (2023), pp. 639–664.

[48] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer, New York, 2006.

[49] M. OBARA, T. OKUNO, AND A. TAKEDA, *Sequential quadratic optimization for nonlinear optimization problems on Riemannian manifolds*, SIAM Journal on Optimization, 32 (2022), pp. 822–853.

[50] M. OBARA, K. SATO, H. SAKAMOTO, T. OKUNO, AND A. TAKEDA, *Stable linear system identification with prior knowledge by Riemannian sequential quadratic optimization*, IEEE Transactions on Automatic Control, 69 (2024), pp. 2060–2066.

[51] Q. REBJOCK AND N. BOUMAL, *Fast convergence of trust-regions for non-isolated minima via analysis of CG on indefinite matrices*, Mathematical Programming, (2024). to appear.

[52] Q. REBJOCK AND N. BOUMAL, *Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions*, Mathematical Programming, (2024). to appear.

[53] H. SAKAI AND H. IIDUKA, *Modified memoryless spectral-scaling Broyden family on Riemannian optimization*, Journal of Optimization Theory and Applications, 202 (2024), pp. 834–853.

[54] H. SATO, *Riemannian optimization and its applications*, Springer, Cham, 2021.

[55] H. SATO, *Riemannian conjugate gradient methods: general framework and specific algorithms with convergence analyses*, SIAM Journal on Optimization, 32 (2022), pp. 2690–2717.

[56] H. SATO AND K. SATO, *Riemannian optimal system identification algorithm for linear MIMO systems*, IEEE Control Systems Letters, 1 (2017), pp. 376 – 381.

[57] K. SATO AND H. SATO, *Structure-preserving H^2 optimal model reduction based on the Riemannian trust-region method*, IEEE Transactions on Automatic Control, 63 (2017), pp. 505–512.

[58] K. SATO, H. SATO, AND T. DAMM, *Riemannian optimal identification method for linear systems with symmetric positive-definite matrix*, IEEE Transactions on Automatic Control,

65 (2020), pp. 4493–4508.

[59] A. Schiela and J. Ortiz, *An SQP method for equality constrained optimization on Hilbert manifolds*, SIAM Journal on Optimization, 31 (2021), pp. 949–981.

[60] R. Silva, M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming: new filter optimality measures and computational results*, 2008, http://hdl.handle.net/10316/11218.

[61] S. Sra and R. Hosseini, *Conic geometric optimization on the manifold of positive definite matrices*, SIAM Journal on Optimization, 25 (2015), pp. 713–739.

[62] S. Sra, N. K. Vishnoi, and O. Yildiz, *On geodesically convex formulations for the Brascamp-Lieb constant*, in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, vol. 116, 2018, pp. 25:1–25:15.

[63] Y. Sun, N. Flammarion, and M. Fazel, *Escaping from saddle points on Riemannian manifolds*, in Advances in Neural Information Processing Systems, 2019, pp. 7276–7286.

[64] J. Townsend, N. Koep, and S. Weichwald, *Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation*, Journal of Machine Learning Research, 17 (2016), pp. 1—-5.

[65] P. Tseng, *Convergent infeasible interior-point trust-region methods for constrained minimization*, SIAM Journal on Optimization, 13 (2002), pp. 432–469.

[66] M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming*, Mathematical Programming, 100 (2004), pp. 379–410.

[67] L. Vicente and S. J. Wright, *Local convergence of a primal-dual method for degenerate nonlinear programming*, Computational Optimization and Applications, 22 (2002), pp. 311–328.

[68] A. Wächter and L. T. Biegler, *Failure of global convergence for a class of interior point methods for nonlinear programming*, Mathematical Programming, 88 (2000), pp. 565–574.

[69] M. Weber and S. Sra, *Global optimality for Euclidean CCCP under Riemannian convexity*, in International Conference on Machine Learning, 2023, pp. 202:36709–36803.

[70] S. J. Wright and D. Orban, *Properties of the log-barrier function on degenerate nonlinear programs*, Mathematics of Operations Research, 27 (2002), pp. 585–613.

[71] Y. Yamakawa and H. Sato, *Sequential optimality conditions for nonlinear optimization on Riemannian manifolds and a globally convergent augmented Lagrangian method*, Computational Optimization and Applications, 81 (2022), pp. 397–421.

[72] H. Yamashita, H. Yabe, and T. Tanabe, *A globally and superlinearly convergent primal-dual interior point trust region method for large scale constrained optimization*, Mathematical Programming, 102 (2005), pp. 111–151.

[73] W. H. Yang, L.-H. Zhang, and R. Song, *Optimality conditions for the nonlinear programming problems on Riemannian manifolds*, Pacific Journal of Optimization, 10 (2014), pp. 415–434.

[74] K. Ye, K. S.-W. Wong, and L.-H. Lim, *Optimization on flag manifolds*, Mathematical Programming, 194 (2022), pp. 621–660.

[75] R. Zass and A. Shashua, *Nonnegative sparse PCA*, in Advances in Neural Information Processing Systems, 2006, pp. 1561–1568.

[76] J. Z. S. Zhang, *A cubic regularized Newton's method over Riemannian manifolds*. arXiv:1805.05565, 2018.

[77] Y. Zhou, C. Bao, C. Ding, and J. Zhu, *A semismooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds*, Mathematical Programming, 201 (2022), pp. 1–61.

Appendix A. Proofs of Lemmas 4.9 and 4.14. In this section, we derive the proofs of the key lemmas, Lemmas 4.9 and 4.14, that estimate the gaps between the predicted and actual reductions.

To prove the lemmas, we first express the gaps in the following form: under $T_{td_x}(T_x\mathcal{M}) \simeq T_x\mathcal{M}$, by Lemma 4.6, we have that, for all $d_x \in T_x\mathcal{M}$ sufficiently small,

$$
\begin{aligned}
&\left|\operatorname{pred}_{\omega,\mu}(d_x) - \operatorname{ared}_\mu(d_x)\right| \\
&= \left|\hat{P}_{\mu_x}(0_x) - \hat{P}_{\mu_x}(d_x) - (m_{\omega,\mu}(0_x) - m_{\omega,\mu}(d_x))\right| \\
&= \left| -\mathrm{D}\hat{P}_{\mu_x}(0_x)[d_x] - \int_0^1 (1-t)\mathrm{D}^2\hat{P}_{\mu_x}(0_x + td_x)[d_x, d_x]\,\mathrm{d}t \right. \\
&\qquad \left. + \frac{1}{2}\langle H(\omega)[d_x], d_x\rangle_x + \langle c_\mu(x), d_x\rangle_x \right| \\
&= \left| \int_0^1 (1-t)\Big(\langle H(\omega)[d_x], d_x\rangle_x - \mathrm{D}^2\hat{P}_{\mu_x}(td_x)[d_x, d_x]\Big)\,\mathrm{d}t \right| \\
&\leq \int_0^1 (1-t)\Bigg(\left|\Big\langle \big(\operatorname{Hess}f(x) - \operatorname{Hess}\hat{f}_x(td_x)\big)[d_x], d_x\Big\rangle_x\right| \\
&\qquad + \sum_{i\in\mathcal{I}}\left|\Big\langle \Big(-y_i\operatorname{Hess}g_i(x) + \frac{\mu}{\hat{g}_{i_x}(td_x)}\operatorname{Hess}\hat{g}_{i_x}(td_x)\Big)[d_x], d_x\Big\rangle_x\right| \\
&\qquad + \sum_{i\in\mathcal{I}}\left|\frac{y_i}{g_i(x)}\langle\operatorname{grad}g_i(x), d_x\rangle_x^2 - \frac{\mu}{\hat{g}_{i_x}(td_x)^2}\langle\operatorname{grad}\hat{g}_{i_x}(td_x), d_x\rangle_x^2\right|\Bigg)\,\mathrm{d}t,
\end{aligned}
$$

(A-1)

where the second equality follows from Taylor's theorem, the third one from (4.12), and the inequality from (4.14). In the following, we will derive the bounds on the right-hand side of (A-1) under the settings of Lemma 4.9, Item 1 of Lemma 4.14, and Item 2 of Lemma 4.14.

Proof of Lemma 4.9. It follows from (A-1) that, for all $d_x \in T_x\mathcal{M}$ sufficiently small,

$$
\begin{aligned}
\left|\operatorname{pred}_{\omega,\mu}(d_x) - \operatorname{ared}_\mu(d_x)\right| &\leq \int_0^1 (1-t)\Bigg(\left\|\operatorname{Hess}f(x)\right\|_{\mathrm{op}} + \left\|\operatorname{Hess}\hat{f}_x(td_x)\right\|_{\mathrm{op}} \\
&\quad + \sum_{i\in\mathcal{I}}\Big(y_i\|\operatorname{Hess}g_i(x)\|_{\mathrm{op}} + \frac{\mu}{\hat{g}_{i_x}(td_x)}\left\|\operatorname{Hess}\hat{g}_{i_x}(td_x)\right\|_{\mathrm{op}} \\
&\quad + \frac{y_i}{g_i(x)}\|\operatorname{grad}g_i(x)\|_{\mathrm{op}} + \frac{\mu}{\hat{g}_{i_x}(td_x)^2}\left\|\operatorname{grad}\hat{g}_{i_x}(td_x)\right\|_{\mathrm{op}}\Big)\Bigg)\|d_x\|_x^2\,\mathrm{d}t.
\end{aligned}
$$

Let $\alpha > 0$ be sufficiently large. Since the sequences $\left\|\operatorname{Hess}\hat{f}_x(\cdot)\right\|_{\mathrm{op}}$, $\{\hat{g}_{i_x}(\cdot)\}_{i\in\mathcal{I}}$, $\left\{\left\|\operatorname{Hess}\hat{g}_{i_x}(\cdot)\right\|_{\mathrm{op}}\right\}_{i\in\mathcal{I}}$, and $\{\|\operatorname{grad}\hat{g}_{i_x}(\cdot)\|\}_{i\in\mathcal{I}}$ are all continuous, the right-hand side can be bounded above by $\alpha\|d_x\|_x^2$ for all $d_x \in T_x\mathcal{M}$ sufficiently small. The proof is complete. □

We proceed to bound (A-1). Under Assumptions B.2 and B.4, the first term of the right-hand side of (A-1) is bounded as follows: if $d_x \in T_x\mathcal{M}$ satisfies $\|d_x\|_x \leq \delta_{RL2}^f$,

then we have

(A-2)
$$\left| \left\langle \left(\operatorname{Hess} f(x) - \operatorname{Hess} \hat{f}_x(td_x) \right)[d_x], d_x \right\rangle_x \right|$$
$$= \left| \left\langle \left(\operatorname{Hess} \hat{f}_x(0_x) - \operatorname{Hess} \hat{f}_x(td_x) \right)[d_x], d_x \right\rangle_x \right| \le \beta^f_{RL2} \|d_x\|^3_{x^\ell}$$

for $t \in [0, 1]$, where the equality follows from Proposition 2.4 and the inequality from $t\|d_x\|_x \le \|d_x\|_x \le \delta^f_{RL2}$, and (2.7) with $t \le 1$. For the third term of the right-hand side of (A-1), we obtain the following bound: for each $i \in \mathcal{I}$, all $x \in \operatorname{str}\mathcal{F}$, $t \in [0, 1]$, and all $d_x \in T_x\mathcal{M}$ sufficiently small,

(A-3)
$$\left| \frac{y_i}{g_i(x)} \langle \operatorname{grad} g_i(x), d_x \rangle^2_x - \frac{\mu}{\hat{g}_{i_x}(td_x)^2} \langle \operatorname{grad} \hat{g}_{i_x}(td_x), d_x \rangle^2_x \right|$$
$$\le \left| \frac{y_i}{g_i(x)} - \frac{\mu}{\hat{g}_{i_x}(td_x)^2} \right| \langle \operatorname{grad} g_i(x), d_x \rangle^2_x$$
$$+ \frac{\mu}{\hat{g}_{i_x}(td_x)^2} \left| \langle \operatorname{grad} g_i(x), d_x \rangle^2_x - \langle \operatorname{grad} \hat{g}_{i_x}(td_x), d_x \rangle^2_x \right|$$
$$= \left| \frac{y_i}{g_i(x)} - \frac{\mu}{\hat{g}_{i_x}(td_x)^2} \right| \langle \operatorname{grad} g_i(x), d_x \rangle^2_x$$
$$+ \frac{\mu}{\hat{g}_{i_x}(td_x)^2} \left(\left| \langle \operatorname{grad} g_i(x), d_x \rangle_x - \langle \operatorname{grad} \hat{g}_{i_x}(td_x), d_x \rangle_x \right| \right.$$
$$\left. \cdot \left| \langle \operatorname{grad} \hat{g}_{i_x}(td_x) - \operatorname{grad} g_i(x) + 2\operatorname{grad} g_i(x), d_x \rangle_x \right| \right),$$

where the inequality follows from $\mu > 0$. In the following, we provide the proofs of Lemma 4.14 by developing these bounds, together with that of the second term of the right-hand side of (A-1), under the settings of Items 1 and 2, respectively.

Proof of Item 1 of Lemma 4.14. Define

$$\Delta' := \min\left\{ 1, \delta', \delta^f_{RL2}, \{\delta^{g_i}_{RL}\}_{i\in\mathcal{I}}, \{\delta^{g_i}_{RL2}\}_{i\in\mathcal{I}} \right\} > 0,$$

where $\delta' > 0$ is the threshold associated with $\varepsilon' > 0$ in Item 2 of Lemma 4.12 and $\delta^f_{RL2} > 0$, $\{\delta^{g_i}_{RL}\}_{i\in\mathcal{I}}$, and $\{\delta^{g_i}_{RL2}\}_{i\in\mathcal{I}}$ are the ones for (2.6) and (2.7), respectively. Let $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ be a search direction satisfying $\|d_{x^\ell}\|_{x^\ell} \le \Delta'$.

As for the second term of the right-hand side of (A-1), we have that, for each $i \in \mathcal{I}$,

(A-4)
$$\left| \left\langle \left(-y^\ell_i \operatorname{Hess} g_i(x^\ell) + \frac{\mu}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})} \operatorname{Hess} \hat{g}_{i_{x^\ell}}(td_{x^\ell}) \right)[d_{x^\ell}], d_{x^\ell} \right\rangle_{x^\ell} \right|$$
$$\le \left| \left(-y^\ell_i + \frac{\mu}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})} \right) \langle \operatorname{Hess} g_i(x^\ell)[d_{x^\ell}], d_{x^\ell} \rangle_{x^\ell} \right|$$
$$+ \left| \frac{\mu}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})} \langle \left(-\operatorname{Hess} \hat{g}_{i_{x^\ell}}(0_{x^\ell}) + \operatorname{Hess} \hat{g}_{i_{x^\ell}}(td_{x^\ell}) \right)[d_{x^\ell}], d_{x^\ell} \rangle_{x^\ell} \right|$$
$$\le \left(y^\ell_i + \frac{\mu}{\varepsilon'} \right) \|\operatorname{Hess} g_i(x^\ell)\|_{op} \|d_{x^\ell}\|^2_{x^\ell} + \frac{\mu \beta^{g_i}_{RL2}}{\varepsilon'} \|d_{x^\ell}\|^3_{x^\ell},$$

where the first inequality follows from Proposition 2.4 and the second one from (2.7) with $t \le 1$, $\mu, y_i^\ell \in \mathbb{R}_{++}$, and $\hat{g}_{ix^\ell}(td_{x^\ell}) > \varepsilon'$ by Item 2 of Lemma 4.12. Using (A-3), the third term of the right-hand side of (A-1) can be bounded as

$$
\begin{aligned}
&\left| \frac{y_i^\ell}{g_i(x^\ell)} \left\langle \mathrm{grad}\, g_i(x^\ell), d_{x^\ell} \right\rangle_{x^\ell}^2 - \frac{\mu}{\hat{g}_{ix^\ell}(td_{x^\ell})^2} \left\langle \mathrm{grad}\, \hat{g}_{ix^\ell}(td_{x^\ell}), d_{x^\ell} \right\rangle_{x^\ell}^2 \right| \\
&\le \left(\frac{y_i^\ell}{g_i(x^\ell)} + \frac{\mu}{\hat{g}_{ix^\ell}(td_{x^\ell})^2} \right) \left\langle \mathrm{grad}\, g_i(x^\ell), d_{x^\ell} \right\rangle_{x^\ell}^2 \\
&\quad + \frac{\mu}{\hat{g}_{ix^\ell}(td_{x^\ell})^2} \left| \left\langle \mathrm{grad}\, \hat{g}_{ix^\ell}(0_{x^\ell}) - \mathrm{grad}\, \hat{g}_{ix^\ell}(td_{x^\ell}), d_{x^\ell} \right\rangle_{x^\ell} \right| \\
&\qquad \left(\left| \left\langle \mathrm{grad}\, \hat{g}_{ix^\ell}(td_{x^\ell}) - \mathrm{grad}\, \hat{g}_{ix^\ell}(0_{x^\ell}), d_{x^\ell} \right\rangle_{x^\ell} \right| + 2 \left| \left\langle \mathrm{grad}\, g_i(x^\ell), d_{x^\ell} \right\rangle_{x^\ell} \right| \right) \\
&\le \left(\frac{y_i^\ell}{\varepsilon'} + \frac{\mu}{\varepsilon'^2} \right) \left\| \mathrm{grad}\, g_i(x^\ell) \right\|_{x^\ell}^2 \left\| d_{x^\ell} \right\|_{x^\ell}^2 \\
&\quad + \frac{\mu(\beta_{RL}^{g_i})^2}{\varepsilon'^2} \left\| d_{x^\ell} \right\|_{x^\ell}^4 + \frac{2\mu\beta_{RL}^{g_i}}{\varepsilon'^2} \left\| \mathrm{grad}\, g_i(x^\ell) \right\|_{x^\ell} \left\| d_{x^\ell} \right\|_{x^\ell}^3,
\end{aligned}
$$
(A-5)

where the first inequality follows from $y_i^\ell, \mu \in \mathbb{R}_{++}$, $g_i(x^\ell) > 0$, and (2.4) and the second one from (2.6) with $t \le 1$ and $g_i(x^\ell) > \varepsilon'$ and $\hat{g}_{ix^\ell}(td_{x^\ell}) > \varepsilon'$ by Item 2 of Lemma 4.12. By combining (A-1) with (A-2), (A-4), and (A-5), we obtain

$$
\begin{aligned}
&\left| \mathrm{pred}^\ell - \mathrm{ared}^\ell \right| \\
&\le \int_0^1 (1-t) \Bigg(\beta_{RL2}^f \left\| d_{x^\ell} \right\|_{x^\ell}^3 \\
&\quad + \sum_{i \in \mathcal{I}} \left(\left(y_i^\ell + \frac{\mu}{\varepsilon'} \right) \left\| \mathrm{Hess}\, g_i(x^\ell) \right\|_{\mathrm{op}} \left\| d_{x^\ell} \right\|_{x^\ell}^2 + \frac{\mu\beta_{RL2}^{g_i}}{\varepsilon'} \left\| d_{x^\ell} \right\|_{x^\ell}^3 \right) \\
&\quad + \sum_{i \in \mathcal{I}} \left(\left(\frac{y_i^\ell}{\varepsilon'} + \frac{\mu}{\varepsilon'^2} \right) \left\| \mathrm{grad}\, g_i(x^\ell) \right\|_{x^\ell}^2 \left\| d_{x^\ell} \right\|_{x^\ell}^2 \right. \\
&\quad \left. + \frac{\mu(\beta_{RL}^{g_i})^2}{\varepsilon'^2} \left\| d_{x^\ell} \right\|_{x^\ell}^4 + \frac{2\mu\beta_{RL}^{g_i}}{\varepsilon'^2} \left\| \mathrm{grad}\, g_i(x^\ell) \right\|_{x^\ell} \left\| d_{x^\ell} \right\|_{x^\ell}^3 \right) \Bigg) \, \mathrm{d}t \\
&\le \frac{1}{2} \Bigg(\beta_{RL2}^f + \sum_{i \in \mathcal{I}} \left(\left(y_i^\ell + \frac{\mu}{\varepsilon'} \right) \left\| \mathrm{Hess}\, g_i(x^\ell) \right\|_{\mathrm{op}} + \frac{\mu\beta_{RL2}^{g_i}}{\varepsilon'} \right) \\
&\quad + \sum_{i \in \mathcal{I}} \left(\left(\frac{y_i^\ell}{\varepsilon'} + \frac{\mu}{\varepsilon'^2} \right) \left\| \mathrm{grad}\, g_i(x^\ell) \right\|_{x^\ell}^2 \right. \\
&\quad \left. + \frac{\mu(\beta_{RL}^{g_i})^2}{\varepsilon'^2} + \frac{2\mu\beta_{RL}^{g_i}}{\varepsilon'^2} \left\| \mathrm{grad}\, g_i(x^\ell) \right\|_{x^\ell} \right) \Bigg) \left\| d_{x^\ell} \right\|_{x^\ell}^2,
\end{aligned}
$$

where the second inequality follows from $\left\| d_{x^\ell} \right\|_{x^\ell} \le \Delta' \le 1$. From Assumption B.5.2 and Assumption B.8, all the terms in the coefficients of the right-hand side are bounded above. Thus, by taking $\beta > 0$ sufficiently large, we complete the proof of Item 1 of Lemma 4.14. □

Proof of Item 2 *of Lemma* 4.14. From Assumption B.5.2 and Assumption B.8, there exist positive scalars $\kappa_{\mathrm{g}}, \kappa_{\mathrm{H}}, \kappa_y \in \mathbb{R}_{++}$ such that

$$(\text{A-6}) \qquad \left\|\operatorname{grad}g_i\!\left(x^\ell\right)\right\|_{x^\ell} \le \kappa_{\mathrm{g}}, \quad \left\|\operatorname{Hess}g_i\!\left(x^\ell\right)\right\|_{x^\ell} \le \kappa_{\mathrm{H}}, \quad y_i^\ell \le \kappa_y \text{ for all } i \in \mathcal{I}.$$

Let $\gamma > 0$ be any positive scalar, and $\varepsilon', \delta' \in \mathbb{R}_{++}$ be the positive values from Item 2 of Lemma 4.12, respectively. Under Assumption B.7, we choose $\delta > 0$ so that, for any $\ell \in \mathbb{N}_0$ and all $\xi_{x^\ell} \in T_{x^\ell}\mathcal{M}$ with $\|\xi_{x^\ell}\|_{x^\ell} \le \delta$,

$$(\text{A-7}) \qquad \left|\hat{g}_{i\,x^\ell}(\xi_{x^\ell}) - g_i\!\left(x^\ell\right)\right| \le \frac{2\gamma\varepsilon'^2}{9m} \min\left\{\frac{1}{\mu\kappa_{\mathrm{H}}}, \frac{1}{3\kappa_y\kappa_{\mathrm{g}}^2}, \frac{\varepsilon'}{3\mu\kappa_{\mathrm{g}}^2}\right\}.$$

From Assumption B.9, we denote by $K' \in \mathbb{N}_0$ an index satisfying that, for all $\ell \ge K'$,

$$(\text{A-8}) \qquad \left\|y^\ell - \mu G^{-1}\!\left(x^\ell\right)\mathbf{1}\right\| \le \frac{2\gamma}{9m} \min\left\{\frac{1}{\kappa_{\mathrm{H}}}, \frac{\varepsilon'}{3\kappa_{\mathrm{g}}^2}\right\}.$$

Let

$$
\begin{aligned}
(\text{A-9}) \quad \Delta'' := \min\Bigg\{ & \delta', \delta, \delta_{RL2}^f, \{\delta_{RL}^{g_i}\}_{i\in\mathcal{I}}, \{\delta_{RL2}^{g_i}\}_{i\in\mathcal{I}}, \\
& \left\{\frac{2\gamma\varepsilon'}{9\mu m\beta_{RL2}^{g_i}}\right\}_{i\in\mathcal{I}}, \left\{\frac{\gamma\varepsilon'^2}{9\mu m\beta_{RL}^{g_i}\kappa_{\mathrm{g}}}\right\}_{i\in\mathcal{I}}, \left\{\frac{\varepsilon'\sqrt{2\gamma}}{3m\beta_{RL}^{g_i}\sqrt{\mu}}\right\}_{i\in\mathcal{I}}, \frac{2\gamma}{3\beta_{RL2}^f} \Bigg\} > 0
\end{aligned}
$$

and let $d_{x^\ell} \in T_{x^\ell}\mathcal{M}$ be a search direction satisfying $\left\|d_{x^\ell}\right\|_{x^\ell} \le \Delta''$. We derive the bounds on each term of the right-hand side of (A-1) under the assumptions. Note that, under Assumptions B.2 and B.4, the bound (A-2) also holds. As for the second term of the right-hand side of (A-1), we have the following inequality: for each $i \in \mathcal{I}$, it holds that

$$
\begin{aligned}
(\text{A-10}) \quad & \left|\left\langle\left(-y_i^\ell\operatorname{Hess}g_i\!\left(x^\ell\right) + \frac{\mu}{\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right)}\operatorname{Hess}\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right)\right)\left[d_{x^\ell}\right], d_{x^\ell}\right\rangle_{x^\ell}\right| \\
& \le \left(\left|\frac{\mu}{g_i\!\left(x^\ell\right)} - y_i^\ell\right| + \left|\frac{\mu\left(\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right) - g_i\!\left(x^\ell\right)\right)}{\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right)g_i\!\left(x^\ell\right)}\right|\right)\left|\left\langle\operatorname{Hess}g_i\!\left(x^\ell\right)\left[d_{x^\ell}\right], d_{x^\ell}\right\rangle_{x^\ell}\right| \\
& \quad + \frac{\mu}{\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right)}\left|\left\langle\left(-\operatorname{Hess}\hat{g}_{i\,x^\ell}(0_{x^\ell}) + \operatorname{Hess}\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right)\right)\left[d_{x^\ell}\right], d_{x^\ell}\right\rangle_{x^\ell}\right| \\
& \le \left(\frac{2\gamma}{9m\kappa_{\mathrm{H}}} + \frac{2\gamma}{9m\kappa_{\mathrm{H}}}\right)\left\|\operatorname{Hess}g_i\!\left(x^\ell\right)\right\|_{\mathrm{op}}\left\|d_{x^\ell}\right\|_{x^\ell}^2 + \frac{\mu\beta_{RL2}^{g_i}}{\varepsilon'}\Delta''\left\|d_{x^\ell}\right\|_{x^\ell}^2 \\
& \le \frac{2\gamma}{3m}\left\|d_{x^\ell}\right\|_{x^\ell}^2,
\end{aligned}
$$

where the first inequality follows from Proposition 2.4, $\mu > 0$, and $\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right) > 0$ by Item 2 of Lemma 4.12, the second one from (A-8), (A-7), (2.7) with $t \le 1$, $\left\|d_{x^\ell}\right\|_{x^\ell} \le \Delta''$, and $g_i\!\left(x^\ell\right) > \varepsilon'$ and $\hat{g}_{i\,x^\ell}\!\left(td_{x^\ell}\right) > \varepsilon'$ by Item 2 of Lemma 4.12 again, and the last one from (A-6) and (A-9). Next, we provide a bound on the third term of the

right-hand side of (A-1). Using (A-3), we have
(A-11)
$$\left| \frac{y_i^\ell}{g_i(x^\ell)} \langle \operatorname{grad} g_i(x^\ell), d_{x^\ell} \rangle_{x^\ell}^2 - \frac{\mu}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})^2} \langle \operatorname{grad} \hat{g}_{i_{x^\ell}}(td_{x^\ell}), d_{x^\ell} \rangle_{x^\ell}^2 \right|$$

$$\leq \left| \frac{y_i^\ell}{g_i(x^\ell)} - \frac{\mu}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})^2} \right| \langle \operatorname{grad} g_i(x^\ell), d_{x^\ell} \rangle_{x^\ell}^2$$

$$+ \frac{\mu}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})^2} \left| \langle \operatorname{grad} \hat{g}_{i_{x^\ell}}(0_{x^\ell}) - \operatorname{grad} \hat{g}_{i_{x^\ell}}(td_{x^\ell}), d_{x^\ell} \rangle_{x^\ell} \right|$$

$$\left(\left| \langle \operatorname{grad} \hat{g}_{i_{x^\ell}}(td_{x^\ell}) - \operatorname{grad} \hat{g}_{i_{x^\ell}}(0_{x^\ell}), d_{x^\ell} \rangle_{x^\ell} \right| + 2 \left| \langle \operatorname{grad} g_i(x^\ell), d_{x^\ell} \rangle_{x^\ell} \right| \right)$$

$$\leq \left(\frac{y_i^\ell |\hat{g}_{i_{x^\ell}}(td_{x^\ell}) - g_i(x^\ell)|}{\hat{g}_{i_{x^\ell}}(td_{x^\ell}) g_i(x^\ell)} + \frac{1}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})} \left| y_i^\ell - \frac{\mu}{g_i(x^\ell)} \right| + \frac{\mu |\hat{g}_{i_{x^\ell}}(td_{x^\ell}) - g_i(x^\ell)|}{\hat{g}_{i_{x^\ell}}(td_{x^\ell})^2 g_i(x^\ell)} \right)$$

$$\|\operatorname{grad} g_i(x^\ell)\|_{x^\ell}^2 \|d_{x^\ell}\|_{x^\ell}^2 + \frac{\mu(\beta_{RL}^{g_i})^2}{\varepsilon'^2} \|d_{x^\ell}\|_{x^\ell}^4 + \frac{2\mu \beta_{RL}^{g_i}}{\varepsilon'^2} \|\operatorname{grad} g_i(x^\ell)\|_{x^\ell} \|d_{x^\ell}\|_{x^\ell}^3$$

$$\leq \left(\frac{2\gamma}{27m\kappa_{\mathrm{g}}^2} + \frac{2\gamma}{27m\kappa_{\mathrm{g}}^2} + \frac{2\gamma}{27m\kappa_{\mathrm{g}}^2} \right) \|\operatorname{grad} g_i(x^\ell)\|_{x^\ell}^2 \|d_{x^\ell}\|_{x^\ell}^2$$

$$+ \frac{\mu(\beta_{RL}^{g_i})^2}{\varepsilon'^2} (\Delta'')^2 \|d_{x^\ell}\|_{x^\ell}^2 + \frac{2\mu \beta_{RL}^{g_i} \|\operatorname{grad} g_i(x^\ell)\|_{x^\ell}}{\varepsilon'^2} \Delta'' \|d_{x^\ell}\|_{x^\ell}^2$$

$$\leq \left(\frac{2\gamma}{27m} + \frac{2\gamma}{27m} + \frac{2\gamma}{27m} \right) \|d_{x^\ell}\|_{x^\ell}^2 + \frac{2\gamma}{9m} \|d_{x^\ell}\|_{x^\ell}^2 + \frac{2\gamma}{9m} \|d_{x^\ell}\|_{x^\ell}^2 = \frac{2\gamma}{3m} \|d_{x^\ell}\|_{x^\ell}^2,$$

where the first inequality follows from (2.4) and $\mu > 0$, the second one from (2.6) with $t \leq 1$, $y_i^\ell, \mu \in \mathbb{R}_{++}$, and $g_i(x^\ell) > \varepsilon' > 0$ and $\hat{g}_{i_{x^\ell}}(td_{x^\ell}) > \varepsilon' > 0$ by Item 2 of Lemma 4.12, the third one from (A-6)–(A-8), Item 2 of Lemma 4.12 again, and $\|d_{x^\ell}\|_{x^\ell} \leq \Delta''$, and the last one from (A-6) and (A-9). By combining (A-1) with (A-2), (A-10), and (A-11), we obtain

$$\left| \operatorname{pred}^\ell - \operatorname{ared}^\ell \right|$$

$$\leq \int_0^1 (1-t) \left(\beta_{RL2}^f \Delta'' \|d_{x^\ell}\|_{x^\ell}^2 + \sum_{i \in \mathcal{I}} \frac{2\gamma}{3m} \|d_{x^\ell}\|_{x^\ell}^2 + \sum_{i \in \mathcal{I}} \frac{2\gamma}{3m} \|d_{x^\ell}\|_{x^\ell}^2 \right) dt$$

$$\leq \frac{1}{2} \left(\frac{2\gamma}{3} + \frac{2\gamma}{3} + \frac{2\gamma}{3} \right) \|d_{x^\ell}\|_{x^\ell}^2 = \gamma \|d_{x^\ell}\|_{x^\ell}^2,$$

where the first inequality holds by $\|d_{x^\ell}\|_{x^\ell} \leq \Delta''$ and the second one from (A-9). The proof is complete. □

Appendix B. Sufficient conditions for Assumptions B.2-B.7 and B.10.
In this section, we provide detailed discussions of the sufficient conditions for Assumptions B.2-B.7 and B.10 in Section 4.3.

As discussed in Section 2, Assumption B.2 is not restrictive. From Lemma 2.5, Assumption B.3 is fulfilled if the generated sequence $\{x^\ell\}_\ell$ is bounded. Moreover, it follows from Lemma 2.5 again that Assumption B.4 holds if the generated sequence $\{x^\ell\}_\ell$ is bounded and all functions $f, \{g_i\}_{i \in \mathcal{I}}$ are of class C^3. Assumption B.5 and Assumption B.6.2 hold if $\{x^\ell\}_\ell$ is bounded. Assumption B.6.1 is fulfilled if f itself

is bounded below, $\{x^\ell\}_\ell$ is bounded, or $\mathcal{M}$ is a compact manifold. As for Assumption B.7, we provide sufficient conditions as follows:

LEMMA B.1. *The following hold:*
1. *Suppose* R $=$ Exp *and that, for every* $i \in \mathcal{I}$, *the function* $g_i \colon \mathcal{M} \to \mathbb{R}$ *is* L^i*-Lipschitz continuous; that is, there exists* $L^i > 0$ *such that*

$$|g_i(x_1) - g_i(x_2)| \leq L^i \operatorname{dist}(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{M}.$$

   Then, Assumption B.7 holds.
2. *If* $\{x^\ell\}_\ell$ *is bounded, then Assumption B.7 holds.*

Proof. We first prove Item 1. Notice that, due to the completeness of $\mathcal{M}$, the domain of the exponential map at x is the whole of $T_x\mathcal{M}$ for all $x \in \mathcal{M}$. It follows from [11, Proposition 10.41] that L^i-Lipschitz continuity is equivalent to $\left|g_i\big(\operatorname{Exp}_x(\xi_x)\big) - g_i(x)\right| \leq L^i\|\xi_x\|_x$ for all $x \in \mathcal{M}$ and all $\xi_x \in T_x\mathcal{M}$. Therefore, for any $\varepsilon > 0$, we attain the conclusion by setting $\delta := \min_{i\in\mathcal{I}} \frac{\varepsilon}{L^i}$.

Next, we consider Item 2. We regard the tangent bundle $T\mathcal{M}$ as a Riemannian manifold endowed with the Riemannian distance

$$\operatorname{dist}_{T\mathcal{M}}((x_1, \xi_{x_1}), (x_2, \xi_{x_2})) := \inf_{\gamma\in\Gamma_{x_1\leftarrow x_2}} \left\{ \sqrt{\bar{\ell}(\gamma)^2 + \left\|\operatorname{PT}^\gamma_{x_1\leftarrow x_2}[\xi_{x_2}] - \xi_{x_1}\right\|_{x_1}^2} \right\}$$

for any $(x_1, \xi_{x_1}), (x_2, \xi_{x_2}) \in T\mathcal{M}$, where $\Gamma_{x_1\leftarrow x_2}$ denotes the set of all piecewise regular curve segments in $\mathcal{M}$ joining x_2 to x_1 and $\bar{\ell}(\gamma)$ is the length of $\gamma \in \Gamma_{x_1\leftarrow x_2}$; see [42, pp.33–34,108] and [11, Section 10.1] for the topics related to curve segments and [22, Section 2] and [16, Appendix II.A.2] for the Riemannian distance on tangent bundles. Let $\mathcal{Q} \subseteq \mathcal{M}$ be a compact subset with $\{x^\ell\}_\ell \subseteq \mathcal{Q}$ and $\mathcal{T} := \left\{(x, \xi_x) \in T\mathcal{M} \colon x \in \mathcal{Q} \text{ and } \|\xi_x\|_x \leq \tilde{\delta}\right\}$ with some $\tilde{\delta} > 0$. Note that $\mathcal{T}$ is also compact [11, Exercise 10.31]. For each $i \in \mathcal{I}$, we consider a composite function $g_i \circ \operatorname{R} \colon T\mathcal{M} \to \mathbb{R}$. From the continuities of g_i and R, the restriction of $g_i \circ \operatorname{R}$ to $\mathcal{T}$ is uniformly continuous by the Heine-Cantor theorem. Namely, for any $\varepsilon > 0$, there exists $\delta > 0$ such that $\delta \leq \tilde{\delta}$ and, for all $(x_1, \xi_{x_1}), (x_2, \xi_{x_2}) \in \mathcal{T}$ with $\operatorname{dist}_{T\mathcal{M}}((x_1, \xi_{x_1}), (x_2, \xi_{x_2})) \leq \delta$, $|g_i \circ \operatorname{R}(x_1, \xi_1) - g_i \circ \operatorname{R}(x_2, \xi_2)| \leq \varepsilon$ holds. Therefore, for all $\ell \in \mathbb{N}_0$ and any $\xi_{x^\ell} \in T_{x^\ell}\mathcal{M}$ with $\|\xi_{x^\ell}\|_{x^\ell} \leq \delta$, we conclude that the statement is true by substituting $\big(x^\ell, \xi_{x^\ell}\big), \big(x^\ell, 0_{x^\ell}\big)$ for $(x_1, \xi_{x_1}), (x_2, \xi_{x_2})$, together with $\operatorname{dist}_{T\mathcal{M}}\big(\big(x^\ell, \xi_{x^\ell}\big), \big(x^\ell, 0_{x^\ell}\big)\big) = \|\xi_{x^\ell}\|_{x^\ell}$. $\square$

We also provide sufficient conditions for Assumption B.10 in the following lemma:

LEMMA B.2. *The following hold:*
1. *If* $\{x^\ell\}_\ell$ *is bounded, then Assumption B.10 holds.*
2. *If* $\mathcal{M}$ *is compact, then Assumption B.10 holds.*

Proof. Under Item 1 or Item 2, let $\mathcal{Q} \subseteq \mathcal{M}$ be a compact subset with $\{x^\ell\}_\ell \subseteq \mathcal{Q}$. For any $\delta_R > 0$, define $\mathcal{T} := \left\{(x, \xi_x) \in T\mathcal{M} \colon x \in \mathcal{Q} \text{ and } \|\xi_x\|_x \leq \delta_R\right\}$. Note that $\mathcal{T}$ is compact [11, Exercise 10.31]. Thus, from the smoothness of R and the continuity of the operator norm, there exists $L_R > 0$ such that $\|\operatorname{D} \operatorname{R}_x(\xi_x)\|_{\text{op}} \leq L_R$ for all $(x, \xi_x) \in \mathcal{T}$. For any $(x, \xi_x) \in \mathcal{T}$, consider the curve $c(t) = \operatorname{R}_x(t\xi_x)$ and let $\bar{\ell}(c) := \int_0^1 \|c'(t)\|_{c(t)} \, dt$ be the length of the curve c on the interval $[0, 1]$. Then, we

have

$$\text{dist}(x, R_x(\xi_x)) \leq \bar{\ell}(c) = \int_0^1 \|D R_x(t\xi_x)[\xi_x]\|_{c(t)} \, dt$$

$$\leq \int_0^1 \|D R_x(t\xi_x)\|_{\text{op}} \|\xi_x\|_x \, dt \leq \int_0^1 L_R \|\xi_x\|_x \, dt = L_R \|\xi_x\|_x.$$

The proof is complete. $\square$

As in Lemma 3.4, the eigenstep satisfies Assumption B.11 with $\kappa_E = \frac{1}{2}$, and so does the exact step (3.16).

Notice that these assumptions are standard in the literature; for example, Assumption B.2 is mentioned in [12, Lemma 3.3] and [11, A6.2]. Assumption B.3 is made in [1, Section 7.4.1]. The smoothness of the functions and the boundedness of the generated sequences, which are stronger than Assumptions B.3-B.8 and B.10, are assumed in [40, Assumptions 3.(C1), 3.(C2)]. The lower boundedness of f, which is a sufficient condition of Assumption B.6.1, is made in [4, Assumption A.1] and [11, A6.5]. Assumptions B.8 and B.9 are made in [18, AS.6, AS.10]. Assumption B.10 is made in [1, Equation (7.25)]. Assumption B.11 is made in [11, A6.4].

Appendix C. Complete proofs of Lemmas 2.5, 2.6, 4.3, 4.6-4.8, 4.13, 5.1, 5.2, 5.8, 5.13, and 5.14 and Proposition 5.3.

C.1. Proofs of Lemmas 2.5 and 2.6. In this subsection, we provide the proofs of lemmas in Section 2.

Proof of Lemma 2.5. Let $\delta_{RL}^\theta > 0$ be any scalar. If θ is of class C^2, it follows from [11, Lemma 10.57] that there exists $L_1 > 0$ such that, for any $x \in \mathcal{U}$ and all $\zeta_x \in T_x\mathcal{M}$ with $\|\zeta_x\|_x \leq \delta_{RL}^\theta$, $\left\|\text{grad}\hat{\theta}_x(\zeta_x) - \text{grad}\hat{\theta}_x(0_x)\right\|_x \leq L_1\|\zeta_x\|_x$ holds. Therefore, for any $x \in \mathcal{U}$ and all $t \geq 0, \xi_x \in T_x\mathcal{M}$ with $t\|\xi_x\|_x \leq \delta_{RL}^\theta$, it holds that

$$\left|\left\langle \text{grad}\hat{\theta}_x(t\xi_x) - \text{grad}\hat{\theta}_x(0_x), \xi_x \right\rangle_x\right|$$

$$\leq \left\|\text{grad}\hat{\theta}_x(t\xi_x) - \text{grad}\hat{\theta}_x(0_x)\right\|_x \|\xi_x\|_x \leq L_1 t\|\xi_x\|_x^2.$$

By setting $\beta_{RL}^\theta = L_1$, we complete the proof of the sufficient condition for radially L-C^1 property.

Similarly, if θ is of class C^3, it follows again from [11, Lemma 10.57] that there exists $L_2 > 0$ such that, for any $x \in \mathcal{U}$ and all $\zeta_x \in T_x\mathcal{M}$ with $\|\zeta_x\|_x \leq \delta_{RL2}^\theta$, $\left\|\text{Hess}\hat{\theta}_x(\zeta_x) - \text{Hess}\hat{\theta}_x(0_x)\right\|_{\text{op}} \leq L_2\|\zeta_x\|_x$ holds. Thus, for any $x \in \mathcal{U}$ and all $t \geq 0, \xi_x \in T_x\mathcal{M}$ with $t\|\xi_x\|_x \leq \delta_{RL2}^\theta$, it holds that

$$\left|\left\langle \left(\text{Hess}\hat{\theta}_x(t\xi_x) - \text{Hess}\hat{\theta}_x(0_x)\right)[\xi_x], \xi_x \right\rangle_x\right|$$

$$\leq \left\|\text{Hess}\hat{\theta}_x(t\xi_x) - \text{Hess}\hat{\theta}_x(0_x)\right\|_{\text{op}} \|\xi_x\|_x^2 \leq L_2 t\|\xi_x\|_x^3.$$

By setting $\beta_{RL2}^\theta = L_2$, we complete the proof of the sufficient condition for the radially L-C^2 property. $\square$

Proof of Lemma 2.6. Let $\mathcal{P}_\theta := \left\{ x \in \mathcal{M} \colon \mathrm{dist}(x, x^*) \le \frac{1}{2}\mathrm{inj}(x^*) \right\}$ be the ball centered at x^*, and define the product distance on $\mathcal{M} \times T_{x^*}\mathcal{M}$ as

$$\mathrm{dist}_{\mathcal{M} \times T_{x^*}\mathcal{M}}((x_1, \xi_{x^*}), (x_2, \zeta_{x^*})) := \sqrt{\mathrm{dist}(x_1, x_2)^2 + \|\xi_{x^*} - \zeta_{x^*}\|_{x^*}^2}$$

for $x_1, x_2 \in \mathcal{M}$ and $\xi_{x^*}, \zeta_{x^*} \in T_{x^*}\mathcal{M}$. Let $h \colon \mathcal{P}_\theta \times T_{x^*}\mathcal{M} \to T_{x^*}\mathcal{M} \colon (x, \xi_{x^*}) \mapsto \mathrm{PT}_{x^* \leftarrow x} \circ \mathrm{Hess}\theta(x) \circ \mathrm{PT}_{x \leftarrow x^*}[\xi_{x^*}]$. Since θ is of class C^3, h is of class C^1. Hence, there exists $s_\theta > 0$ such that, for any $x \in \mathcal{P}_\theta$ and $\xi_{x^*} \in T_{x^*}\mathcal{M}$ with $\|\xi_{x^*}\|_{x^*} \le 1$,

(C-1)
$$\begin{aligned}
&\left\| \left(\mathrm{Hess}\theta(x^*) - \mathrm{PT}_{x^* \leftarrow x} \circ \mathrm{Hess}\theta(x) \circ \mathrm{PT}_{x \leftarrow x^*} \right)[\xi_{x^*}] \right\|_{x^*} \\
&= \|h(x, \xi_{x^*}) - h(x^*, \xi_{x^*})\|_{x^*} \le s_\theta\, \mathrm{dist}_{\mathcal{M} \times T_{x^*}\mathcal{M}}((x, \xi_{x^*}), (x^*, \xi_{x^*})) \\
&= s_\theta\, \mathrm{dist}(x, x^*).
\end{aligned}$$

For every $x \in \mathcal{P}_\theta$, there exists a vector ζ_{x^*} with $\|\zeta_{x^*}\|_{x^*} \le 1$ such that

(C-2)
$$\begin{aligned}
&\left\| \left(\mathrm{Hess}\theta(x^*) - \mathrm{PT}_{x^* \leftarrow x} \circ \mathrm{Hess}\theta(x) \circ \mathrm{PT}_{x \leftarrow x^*} \right)[\zeta_{x^*}] \right\|_{x^*} \\
&= \left\| \mathrm{Hess}\theta(x^*) - \mathrm{PT}_{x^* \leftarrow x} \circ \mathrm{Hess}\theta(x) \circ \mathrm{PT}_{x \leftarrow x^*} \right\|_{\mathrm{op}}
\end{aligned}$$

since the set $\left\{ \xi_{x^*} \in T_{x^*}\mathcal{M} \colon \|\xi_{x^*}\|_{x^*} \le 1 \right\}$ is compact. Combining (C-2) with (C-1) yields

$$\left\| \mathrm{Hess}\theta(x^*) - \mathrm{PT}_{x^* \leftarrow x} \circ \mathrm{Hess}\theta(x) \circ \mathrm{PT}_{x \leftarrow x^*} \right\|_{\mathrm{op}} \le s_\theta\, \mathrm{dist}(x, x^*)$$

for every $x \in \mathcal{P}_\theta$. The proof is complete. □

C.2. Proof of Lemma 4.3. In this section, we provide the proof of the auxiliary lemmas in Section 4.1.

Proof of Lemma 4.3. We argue by contradiction. Suppose that, for any $\mathcal{N} \subseteq \mathcal{M}$ with $x^* \in \mathcal{N}$, there exists $x \in \mathcal{N}$ such that $\{\mathrm{grad}g_i(x)\}_{i \in \mathcal{A}(x^*)}$ are linearly dependent. For any $r \in \mathbb{N}_0$, we let $\mathcal{N}_{\frac{1}{r}} := \left\{ x \in \mathcal{M} \colon \mathrm{dist}(x, x^*) < \frac{1}{r} \right\}$. There exist $x_r \in \mathcal{N}_{\frac{1}{r}}$ and a nonzero vector $v^r \in \mathbb{R}^{|\mathcal{A}(x^*)|}$ such that $\|v^r\| = 1$ and $\sum_{i \in \mathcal{A}(x^*)} v_i^r \mathrm{grad}g_i(x_r) = 0$. Considering $r = 1, 2, 3, \dots$ yields that $\{x_r\}_r$ converges to x^* as $r \to \infty$, and there exists $v^* \in \mathbb{R}^{|\mathcal{A}(x^*)|}$ such that $\|v^*\| = 1$ and $\sum_{i \in \mathcal{A}(x^*)} v_i^* \mathrm{grad}g_i(x^*) = 0$, which contradicts Assumption A.3. The proof is complete. □

C.3. Proofs of Lemmas 4.6-4.8. In this subsection, we provide the proof of the auxiliary lemmas in Section 4.2.

Proof of Lemma 4.6. Define $v := \min_{i \in \mathcal{I}} \frac{1}{2}g_i(x) > 0$. It follows from the continuity of R and $\{g_i\}_{i \in \mathcal{I}}$ that $\{\hat{g}_{i_x}\}_{i \in \mathcal{I}}$ are all continuous. Therefore, for each $i \in \mathcal{I}$ and any $x \in \mathrm{str}\,\mathcal{F}$, there exists $\delta_x^i > 0$ such that $|\hat{g}_{i_x}(\xi_x) - \hat{g}_{i_x}(0_x)| \le v$ for $\xi_x \in T_x\mathcal{M}$ with $\|\xi_x\|_x \le \delta_x^i$, which implies

$$\hat{g}_{i_x}(\xi_x) \ge \hat{g}_{i_x}(0_x) - v \ge \frac{1}{2}g_i(x) > 0,$$

where the second inequality follows from (2.2a) and the definition of v. Define $\delta_x := \min_{i \in \mathcal{I}} \delta_x^i$. Then, the statement holds for any $\|\xi_x\|_x \le \delta_x$, which completes the proof.□

Proof of Lemma 4.7. Equation (4.12) directly follows from

$$\mathrm{D}\hat{P}_{\mu_x}(0_x)[\xi_x] = \mathrm{D}P_\mu(\mathrm{R}_x(0_x))[\mathrm{D}\,\mathrm{R}_x(0_x)[\xi_x]] = \mathrm{D}P_\mu(x)[\xi_x]$$

$$(\text{C-3}) = \mathrm{D}f(x)[\xi_x] - \mu\sum_{i\in\mathcal{I}}\frac{1}{g_i(x)}\mathrm{D}g_i(x)[\xi_x] = \left\langle \mathrm{grad}\,f(x) - \mu\mathcal{G}_x\left[G(x)^{-1}\mathbf{1}\right], \xi_x\right\rangle_x$$

$$= \langle c_\mu(x), \xi_x\rangle_x,$$

where the first equality follows from the chain rule, the second one from (2.2), the fourth one from (2.1), and the last one from (3.9).

Next, we prove (4.13). We have

$$\mathrm{D}\hat{P}_{\mu_x}(0_x)\left[d_x^*\right] = \left\langle c_\mu(x), d_x^*\right\rangle_x = -\left\langle \left(H(\omega) + \nu\,\mathrm{id}_{T_x\mathcal{M}}\right)\left[d_x^*\right], d_x^*\right\rangle_x,$$

where the first equality follows from (C-3) with $\xi_x = d_x^*$ and the second one from (3.16a). The proof is complete. □

Proof of Lemma 4.8. For each $i \in \mathcal{I}$, all $\zeta_x \in T_x\mathcal{M}$, and any $\xi_x \in T_x\mathcal{M}$ with $\hat{g}_{i_x}(\xi_x) \neq 0$, it follows that

$$\left\langle \mathrm{grad}\log\hat{g}_{i_x}(\xi_x), \zeta_x\right\rangle_x = \mathrm{D}\left(\log\hat{g}_{i_x}\right)(\xi_x)[\zeta_x]$$

$$= \mathrm{D}\log\left(\hat{g}_{i_x}(\xi_x)\right)\left[\mathrm{D}\hat{g}_{i_x}(\xi_x)[\zeta_x]\right] = \left\langle \frac{1}{\hat{g}_{i_x}(\xi_x)}\mathrm{grad}\hat{g}_{i_x}(\xi_x), \zeta_x\right\rangle_x,$$

where the second equality holds by the chain rule on $T_x\mathcal{M}$ and the third one by (2.1). Therefore, we have

$$\mathrm{grad}\log\hat{g}_{i_x}(\xi_x) = \frac{1}{\hat{g}_{i_x}(\xi_x)}\mathrm{grad}\hat{g}_{i_x}(\xi_x),$$

which implies

$$\mathrm{Hess}\left(\log\hat{g}_{i_x}\right)(\xi_x) = \mathrm{D}\left(\mathrm{grad}\log\hat{g}_{i_x}\right)(\xi_x) = \mathrm{D}\left(\frac{1}{\hat{g}_{i_x}(\cdot)}\mathrm{grad}\hat{g}_{i_x}(\cdot)\right)(\xi_x)$$

$$(\text{C-4})\quad = -\frac{\mathrm{grad}\hat{g}_{i_x}(\xi_x)}{\hat{g}_{i_x}(\xi_x)^2}\mathrm{D}\hat{g}_{i_x}(\xi_x) + \frac{1}{\hat{g}_{i_x}(\xi_x)}\mathrm{D}\left(\mathrm{grad}\hat{g}_{i_x}\right)(\xi_x)$$

$$= -\frac{\mathrm{grad}\hat{g}_{i_x}(\xi_x)}{\hat{g}_{i_x}(\xi_x)^2}\mathrm{D}\hat{g}_{i_x}(\xi_x) + \frac{1}{\hat{g}_{i_x}(\xi_x)}\mathrm{Hess}\hat{g}_{i_x}(\xi_x).$$

Using (C-4), we derive (4.14) as follows:

$$\mathrm{D}^2\hat{P}_{\mu_x}(\xi_x)[\zeta_x, \eta_x] = \mathrm{D}^2\hat{f}_x(\xi_x)[\zeta_x, \eta_x] - \mu\sum_{i\in\mathcal{I}}\mathrm{D}^2\left(\log\hat{g}_{i_x}\right)(\xi_x)[\zeta_x, \eta_x]$$

$$= \left\langle\left(\mathrm{Hess}\hat{f}_x(\xi_x) - \mu\sum_{i\in\mathcal{I}}\mathrm{Hess}\left(\log\hat{g}_{i_x}\right)(\xi_x)\right)[\zeta_x], \eta_x\right\rangle_x$$

$$= \left\langle\left(\mathrm{Hess}\hat{f}_x(\xi_x) - \sum_{i\in\mathcal{I}}\frac{\mu}{\hat{g}_{i_x}(\xi_x)}\mathrm{Hess}\hat{g}_{i_x}(\xi_x)\right)[\zeta_x], \eta_x\right\rangle_x$$

$$+ \sum_{i\in\mathcal{I}}\frac{\mu}{\hat{g}_{i_x}(\xi_x)^2}\left\langle\mathrm{grad}\hat{g}_{i_x}(\xi_x), \zeta_x\right\rangle_x\left\langle\mathrm{grad}\hat{g}_{i_x}(\xi_x), \eta_x\right\rangle_x,$$

where the third equality follows from (2.1). The proof is complete. □

C.4. Proof of Lemma 4.13. In this subsection, we provide the proof of the auxiliary lemmas in Section 4.3.

Proof of Lemma 4.13. For any $\ell \in \mathbb{N}_0$ and all $\xi_{x^\ell} \in T_{x^\ell}\mathcal{M}$ with $\|\xi_{x^\ell}\|_{x^\ell} \leq 1$, it holds that

$$
\begin{aligned}
\left\|H^\ell[\xi_{x^\ell}]\right\|_{x^\ell} &\leq \left\|\mathrm{Hess}\,f(x^\ell)\right\|_{\mathrm{op}}\|\xi_{x^\ell}\|_{x^\ell} \\
&\quad + \sum_{i\in\mathcal{I}} y_i^\ell\left\|\mathrm{Hess}\,g_i(x^\ell)\right\|_{\mathrm{op}}\|\xi_{x^\ell}\|_{x^\ell} + \sum_{i\in\mathcal{I}}\frac{y_i^\ell}{g_i(x^\ell)}\left\|\mathrm{grad}\,g_i(x^\ell)\right\|_{x^\ell}^2\|\xi_{x^\ell}\|_{x^\ell}^2 \\
&\leq \left\|\mathrm{Hess}\,f(x^\ell)\right\|_{\mathrm{op}} + \sum_{i\in\mathcal{I}} y_i^\ell\left\|\mathrm{Hess}\,g_i(x^\ell)\right\|_{\mathrm{op}} + \sum_{i\in\mathcal{I}}\frac{y_i^\ell}{g_i(x^\ell)}\left\|\mathrm{grad}\,g_i(x^\ell)\right\|_{x^\ell}^2,
\end{aligned}
$$

where the first inequality follows from (3.8) and the second one from $\|\xi_{x^\ell}\|_{x^\ell} \leq 1$. Under Assumption B.5, Assumption B.8, Assumption B.6.2, and Item 1 of Lemma 4.12, the right-hand side is bounded above. The proof is complete. $\qquad\square$

C.5. Proof of Theorem 4.20. In this subsection, we provide the proof of Theorem 4.20 in Section 4.4.

Proof of Theorem 4.20. First, we consider Item 1. Since the dual variable is updated when the primal iterate is successful, we focus on the successful iterations. From Item 1 of Lemma 4.12, there exists $\underline{\varepsilon} > 0$ such that $\frac{1}{\underline{\varepsilon}} \geq \frac{1}{g_i(x^\ell)}$ for any $i \in \mathcal{I}$ and all $\ell \in \mathbb{N}_0$, which, together with (4.33), implies that

$$
y_i^\ell \leq \max\left\{\tilde{c}, y_i^0, \frac{\tilde{c}}{\mu}, \frac{\tilde{c}}{\underline{\varepsilon}}\right\}
$$

for each $i \in \mathcal{I}$. Thus, Item 1 holds.

Next, we consider Item 2. Recall that $\mathcal{S}$ denotes the set of the successful iterations and $\{x^{\ell_j}\}_j$ is the ordered sequence of successful iterates. If $|\mathcal{S}|$ is finite, then it follows from Proposition 4.10 that $\Psi(\omega^\ell; \mu) = 0$ for all $\ell \in \mathbb{N}_0$ sufficiently large. Thus, we have $\|y^\ell - \mu G(x^\ell)^{-1}\mathbf{1}\| = 0$ for all $\ell \in \mathbb{N}_0$ sufficiently large, meaning that Assumption B.9 holds. In the following, we consider the case where $|\mathcal{S}|$ is infinite. It follows from $\hat{g}_{i_{x^{\ell_j}}}\left(d_{x^{\ell_j}}\right) = g_i(x^{\ell_j+1})$, (4.17), and (4.35) that

$$
\tag{C-5} \lim_{j\to\infty}\left|g_i(x^{\ell_j}) - g_i(x^{\ell_j+1})\right| = 0
$$

for all $i \in \mathcal{I}$. Therefore, we have

$$
\begin{aligned}
&\left\|y^{\ell_j} + d_{y^{\ell_j}} - \mu G(x^{\ell_j+1})^{-1}\mathbf{1}\right\| \\
&\leq \left\|y^{\ell_j} + d_{y^{\ell_j}} - \mu G(x^{\ell_j})^{-1}\mathbf{1}\right\| + \mu\left\|G(x^{\ell_j})^{-1}\mathbf{1} - G(x^{\ell_j+1})^{-1}\mathbf{1}\right\| \\
&\leq \sum_{i\in\mathcal{I}}\left|\frac{y_i^{\ell_j}}{g_i(x^{\ell_j})}\left\langle\mathrm{grad}\,g_i(x^{\ell_j}), d_{x^{\ell_j}}\right\rangle_{x^{\ell_j}}\right| + \sum_{i\in\mathcal{I}}\frac{\mu\left|g_i(x^{\ell_j}) - g_i(x^{\ell_j+1})\right|}{g_i(x^{\ell_j})g_i(x^{\ell_j+1})} \\
&\leq \sum_{i\in\mathcal{I}}\frac{\left|y_i^{\ell_j}\right|}{\underline{\varepsilon}}\left\|\mathrm{grad}\,g_i(x^{\ell_j})\right\|_{x^{\ell_j}}\left\|d_{x^{\ell_j}}\right\|_{x^{\ell_j}} + \sum_{i\in\mathcal{I}}\frac{\mu\left|g_i(x^{\ell_j}) - g_i(x^{\ell_j+1})\right|}{\underline{\varepsilon}^2},
\end{aligned}
$$

where the second inequality follows from (3.7) and the third one from Item 1 of Lemma 4.12. By Item 1, Assumption B.5.2, (C-5), and (4.35), the right-hand side converges to zero as $n \to \infty$. Thus, under Assumption B.5.1, letting $\kappa_g > 0$ be the scalar satisfying $g_i(x^{\ell_j}) \le \kappa_g$ for any $i \in \mathcal{I}$, we obtain

$$\left| y^{\ell_j} + d_{y^{\ell_j}} - \mu G(x^{\ell_j+1})^{-1} \mathbf{1} \right| \le \min\{1 - \underline{c}, \tilde{c} - 1\} \frac{\mu}{\kappa_g}$$

for any $i \in \mathcal{I}$ and all $n \in \mathbb{N}_0$ sufficiently large, which implies

$$y_i^{\ell_j} + \left[d_{y^{\ell_j}} \right]_i$$
$$\ge -(1 - \underline{c}) \frac{\mu}{\kappa_g} + \frac{\mu}{g_i(x^{\ell_j+1})} \ge -(1 - \underline{c}) \frac{\mu}{g_i(x^{\ell_j+1})} + \frac{\mu}{g_i(x^{\ell_j+1})} = \frac{\underline{c}\mu}{g_i(x^{\ell_j+1})},$$

$$y_i^{\ell_j} + \left[d_{y^{\ell_j}} \right]_i \le (\tilde{c} - 1) \frac{\mu}{\kappa_g} + \frac{\mu}{g_i(x^{\ell_j+1})} \le (\tilde{c} - 1) \frac{\mu}{g_i(x^{\ell_j+1})} + \frac{\mu}{g_i(x^{\ell_j+1})} = \frac{\tilde{c}\mu}{g_i(x^{\ell_j+1})}.$$

for any $i \in \mathcal{I}$. By the update rule (4.34), $y^{\ell_j+1} = y^{\ell_j} + d_{y^{\ell_j}}$ holds for all $n \in \mathbb{N}_0$ sufficiently large, which implies

(C-6) $\qquad G(x^{\ell_j+1}) y^{\ell_j+1} = G(x^{\ell_j+1}) G(x^{\ell_j})^{-1} \left(\mu \mathbf{1} - Y^{\ell_j} \mathcal{G}_{x^{\ell_j}}^* \left[d_{x^{\ell_j}} \right] \right),$

where we write Y^{ℓ_j} for $\mathrm{diag}(y^{\ell_j})$ and the equality follows from (3.7). Here, it follows from Item 1 of Lemma 4.12 that

$$\left\| G(x^{\ell_j+1}) G(x^{\ell_j})^{-1} - I_m \right\| \le \sum_{i \in \mathcal{I}} \left| \frac{g_i(x^{\ell_j+1})}{g_i(x^{\ell_j})} - 1 \right| \le \sum_{i \in \mathcal{I}} \frac{\left| g_i(x^{\ell_j+1}) - g_i(x^{\ell_j}) \right|}{\underline{\varepsilon}},$$

where $I_m \in \mathbb{R}^{m \times m}$ is the identity matrix, and the right-hand side converges to zero as $\ell_j \to \infty$ by (C-5). Therefore, equations (4.35) and (C-6) and the boundedness of $\{y^{\ell_j}\}_j$ and $\left\{ \left\| \mathrm{grad} g_i(x^{\ell_j}) \right\|_{x^{\ell_j}} \right\}_{j,i}$ imply

(C-7) $\qquad\qquad\qquad\qquad \lim_{j \to \infty} G(x^{\ell_j+1}) y^{\ell_j+1} = \mu \mathbf{1}.$

Since $\omega^{\ell_j+1} = \omega^{\ell_{j+1}}$ holds and all iterates between ℓ_j and ℓ_{j+1} are unsuccessful for all $j \in \mathbb{N}_0$, equation (C-7) extends to all the iterates; that is, it holds that $\lim_{\ell \to \infty} G(x^\ell) y^\ell = \mu \mathbf{1}$. From $x^\ell \in \mathrm{str} \mathcal{F}$, for all $\ell \in \mathbb{N}_0$, we have $\lim_{\ell \to \infty} \| y^\ell - \mu G(x^\ell)^{-1} \mathbf{1} \| = 0$. The proof is complete. $\qquad\square$

C.6. Proof of Lemmas 5.1, 5.2, 5.8, 5.13, and 5.14 and Proposition 5.3. In this section, we provide the proof of the auxiliary lemmas in Section 5.

Proof of Lemma 5.1. Let $d_{\omega^*} \in T_{\omega^*} \mathcal{M}$ satisfy $\mathrm{J}_\Psi(\omega^*)[d_{\omega^*}] = 0_{\omega^*}$; that is, under $T_y \mathbb{R}^m \simeq \mathbb{R}^m$,

(C-8) $\qquad\qquad \mathrm{Hess}_x \mathcal{L}(\omega^*)[d_{x^*}] - \sum_{i \in \mathcal{I}} d_{y_i^*} \mathrm{grad} g_i(x^*) = 0_{x^*},$

(C-9) $\qquad\qquad y_i^* \langle \mathrm{grad} g_i(x^*), d_{x^*} \rangle_{x^*} + d_{y_i^*} g_i(x^*) = 0$ for all $i \in \mathcal{I}$.

We now prove that such d_{ω^*} is actually the zero vector. For each $i \notin \mathcal{A}(x^*)$, $y_i^* = 0$ and $g_i(x^*) > 0$ hold by the complementarity condition in (2.10), which, together with (C-9), implies $d_{y_i^*} = 0$. On the other hand, for each $i \in \mathcal{A}(x^*)$, $y_i^* > 0$ and $g_i(x^*) = 0$ hold by the SC, which, together with (C-9) again, implies $\langle \mathrm{grad}g_i(x^*), d_{x^*} \rangle_{x^*} = 0$. Combining these with (C-8) yields $d_{x^*} \in \mathcal{C}(\omega^*)$ and $\langle \mathrm{Hess}_x\mathcal{L}(\omega^*)[d_{x^*}], d_{x^*} \rangle_{x^*} = 0_{x^*}$. Therefore, $d_{x^*} = 0_{x^*}$ holds by the SOSC. Substituting $d_{x^*} = 0_{x^*}$ and $d_{y_i^*} = 0$ for $i \notin \mathcal{A}(x^*)$ into (C-8), we obtain $\sum_{i \in \mathcal{A}(x^*)} d_{y_i^*}\mathrm{grad}g_i(x^*) = 0_{x^*}$. Thus, by the LICQ, $d_{y_i^*} = 0$ holds for $i \in \mathcal{A}(x^*)$, and hence we have $d_{\omega^*} = 0_{\omega^*}$. The proof is complete. □

Proof of Lemma 5.2. Since $\mathrm{J}_\Psi(\omega)$ is continuous, it follows from Lemma 5.1 that, for any $\omega \in \mathcal{F} \times \mathbb{R}_+^m$ sufficiently close to ω^*, J_Ψ is nonsingular and

$$\left\| d_{\omega,\mu}^{\mathrm{N}} \right\|_\omega \leq \left\| \mathrm{J}_\Psi(\omega)^{-1} \right\|_{\mathrm{op}} \|\Psi(\omega; \mu)\|_\omega \leq r\left(\|\Psi(\omega)\|_\omega + \sqrt{m}\mu \right)$$

for some $r > 0$. Since the point ω^* is the solution of $\Psi(\omega) = 0_{\omega^*}$ and Ψ is continuous, the right-hand side can be made arbitrarily small by choosing a sufficiently small μ. The proof is complete. □

Proof of Proposition 5.3. By the definition of d_y^*, it follows that

(C-10) $$Y\mathcal{G}_x^*\left[d_x^*\right] + G(x)d_y^* = -(G(x)y - \mu\mathbf{1}).$$

Recall that, by Proposition 3.5, there exists $\nu \geq 0$ such that d_x^* and ν satisfy (3.16). From (3.16b) and $\left\| d_x^* \right\|_x < \Delta$, we have $\nu = 0$. Thus, it follows that

$$\mathrm{Hess}_x\mathcal{L}(\omega)\left[d_x^*\right] - \mathcal{G}_x\left[d_y^*\right]$$

(C-11) $$= \mathrm{Hess}_x\mathcal{L}(\omega)\left[d_x^*\right] + \mathcal{G}_x\left[YG(x)^{-1}\mathcal{G}^*\left[d_x^*\right]\right] - \mathcal{G}_x\left[\mu G(x)^{-1}\mathbf{1}\right] + \mathcal{G}_x[y]$$

$$= -\mathrm{grad}f(x) + \mathcal{G}_x[y] = -\mathrm{grad}_x\mathcal{L}(\omega),$$

where the first equality follows from (3.7), the second one from (3.8), (3.9), and (3.16a), and the third one from (2.8). By (C-10) and (C-11), we have that $d_\omega^* = \left(d_x^*, d_y^*\right)$ is the solution of (3.4). Since the solution is unique due to the nonsingularity of $\mathrm{J}_\Psi(\omega)$, we conclude that d_ω^* is equivalent to the Newton step (5.6). □

Proof of Lemma 5.8. Let $(\varphi, \mathcal{U})$ be a chart with $x^* \in \mathcal{U} \subseteq \mathbb{R}^d$, and let e_j be the j-th standard basis of $\mathbb{R}^d$. For any $\vartheta \in \mathfrak{F}(\mathcal{M})$, any $x \in \mathcal{U}$ and all $\xi_x \in T_x\mathcal{M}$, we have

$$\partial_i \hat{\vartheta}_x(\xi_x) = \sum_j \partial_j \vartheta(\mathrm{R}_x(\xi_x))A_i^j(\xi_x),$$

where $\partial_i \hat{\vartheta}_x(\xi_x) : T_x\mathcal{M} \to \mathbb{R}$ is the partial derivative of $\hat{\vartheta}_x$ at ξ_x with respect to the i-th variable, $\partial_j \vartheta(\cdot) := \lim_{t\downarrow 0} \frac{\vartheta \circ \varphi^{-1}(\varphi(\cdot) + te_j) - \vartheta \circ \varphi^{-1}(\varphi(\cdot))}{t}$, $A(\xi_x)$ denotes the differential of R_x at $\xi_x \in T_x\mathcal{M}$, and $A_i^j(\xi_x)$ is its (i,j)-th element. Let $M(x) \in \mathbb{R}^{d \times d}$ be the coordinate expression of the Riemannian metric at $x \in \mathcal{U}$, and let $M_{ij}(x)$ be its (i,j)-th element. Note that $M(x)$ is positive-definite. Then,

$$\left\| \sum_{t=1}^n a_t \mathrm{grad}\hat{\theta}_x^t(\xi_x) \right\|_x^2 = \sum_{t^1, t^2, i, j} a_{t^1} a_{t^2} \partial_i \hat{\theta}_x^{t^1}(\xi_x) M_{ij}(x) \partial_j \hat{\theta}_x^{t^2}(\xi_x)$$

(C-12) $$= \sum_{t^1, t^2, i, j, k, l} a_{t^1} a_{t^2} \partial_k \theta^{t^1}(\mathrm{R}_x(\xi_x)) A_i^k(\xi_x) M_{ij}(x) A_j^l(\xi_x) \partial_l \theta^{t^2}(\mathrm{R}_x(\xi_x))$$

$$= a^\top \partial\theta(\mathrm{R}_x(\xi_x))^\top A(\xi_x) M(x) A(\xi_x) \partial\theta(\mathrm{R}_x(\xi_x))a,$$

where $\partial\theta(\cdot) \in \mathbb{R}^{d\times n}$ is a matrix whose (i,j)-th element is $\partial_i\theta^j(\cdot)$. We also have

(C-13)
$$
\begin{aligned}
&\left\|\sum_{t=1}^n a_t\mathrm{grad}\theta^t(\mathrm{R}_x(\xi_x))\right\|_x^2 \\
&= \sum_{t^1,t^2,i,j} a_{t^1}a_{t^2}\partial_i\theta^{t^1}(\mathrm{R}_x(\xi_x))M_{ij}(\mathrm{R}_x(\xi_x))\partial_j\theta^{t^2}(\mathrm{R}_x(\xi_x)) \\
&= a^\top\partial\theta(\mathrm{R}_x(\xi_x))^\top M(x)\partial\theta(\mathrm{R}_x(\xi_x))a.
\end{aligned}
$$

It follows from (C-12) and (C-13) that

$$
\begin{aligned}
&c^2\left\|\sum_{t=1}^n a_t\mathrm{grad}\hat{\theta}_x^t(\xi_x)\right\|_x^2 - \left\|\sum_{t=1}^n a_t\mathrm{grad}\theta^t(\mathrm{R}_x(\xi_x))\right\|_x^2 \\
&= a^\top\partial\theta(\mathrm{R}_x(\xi_x))^\top\left(c^2A(\xi_x)M(x)A(\xi_x) - M(x)\right)\partial\theta(\mathrm{R}_x(\xi_x))a.
\end{aligned}
$$

Since $A(0_{x^*}) = \mathrm{id}_{T_{x^*}\mathcal{M}}$ holds by (2.2b), we have

$$
c^2A(0_{x^*})M(x^*)A(0_{x^*}) - M(x^*) = \left(c^2 - 1\right)M(x^*) \succ 0,
$$

implying that $c\left\|\sum_{t=1}^n a_t\mathrm{grad}\hat{\theta}_{x^*}^t(0_{x^*})\right\|_{x^*} \geq \left\|\sum_{t=1}^n a_t\mathrm{grad}\theta^t(\mathrm{R}_{x^*}(0_{x^*}))\right\|_{x^*}$. Let $\mathcal{P} \subseteq \mathcal{M}$ denote a closed neighborhood of x^*. Since the functions R and $\{\theta^t\}_t$ are continuously differentiable and the set $\left\{(x,\xi_x)\colon x \in \mathcal{P}, \xi_x \in T_x\mathcal{M}, \|\xi_x\|_x \leq \delta\right\}$ is compact, we obtain the positive definiteness of $c^2A(\xi_x)M(x)A(\xi_x) - M(x)$ for all $x \in \mathcal{P}$ by taking $\mathcal{P}$ sufficiently small and δ sufficiently small if necessary. This implies (5.13) for all $x \in \mathcal{P}$, any $a \in \mathbb{R}^n$, and all $\xi_x \in T_x\mathcal{M}$ with $\|\xi_x\|_x \leq \delta$. The proof is complete. $\square$

Proof of Lemma 5.13. For any $v = (v_x, v_y) \in T_x\mathcal{M} \times \mathbb{R}^m$, we have

(C-14)
$$
\mathrm{D}F_\mu(\overline{\omega})[\overline{v}] = \begin{bmatrix} \mathrm{D}\left(\mathrm{D}\varphi_x\left(\varphi_x^{-1}(\cdot_x)\right)\left[\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\cdot)\right)\right]\right)(\overline{\omega})[\overline{v}] \\ \mathrm{D}\left(G\left(\varphi_x^{-1}(\cdot_x)\right)\varphi_y^{-1}(\cdot_y) - \mu\mathbf{1}\right)(\overline{\omega})[\overline{v}] \end{bmatrix},
$$

where $\varphi_x^{-1}(\cdot_x)$ and $\varphi_y^{-1}(\cdot_y)$ denote the maps $\omega \mapsto \varphi_x^{-1}(x)$ and $\omega \mapsto \varphi_y^{-1}(y) = y$, respectively. Note that equation (C-14) is independent of the value of μ. In the following, we analyze each component of (C-14). For the first component, it follows from the product rule that

(C-15)
$$
\begin{aligned}
&\mathrm{D}\left(\mathrm{D}\varphi_x\left(\varphi_x^{-1}(\cdot_x)\right)\left[\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\cdot)\right)\right]\right)(\overline{\omega})[\overline{v}] \\
&= \mathrm{D}\left(\mathrm{D}\varphi_x\left(\varphi_x^{-1}(\cdot_x)\right)\right)(\overline{\omega})[\overline{v}]\left[\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega})\right)\right] \\
&\quad + \mathrm{D}\varphi_x\left(\varphi_x^{-1}(\overline{x})\right)\left[\mathrm{D}\left(\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\cdot)\right)\right)(\overline{\omega})[\overline{v}]\right] \\
&= \mathrm{D}\left(\mathrm{D}\varphi_x\left(\varphi_x^{-1}(\cdot_x)\right)\right)(\overline{\omega})[\overline{v}]\left[\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{\omega})\right)\right] \\
&\quad + \mathrm{D}\varphi_x\left(\varphi_x^{-1}(\overline{x})\right)\left[\mathrm{D}\left(\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\cdot_x,\overline{y})\right)\right)(\overline{x})[\overline{v_x}]\right] \\
&\quad + \mathrm{D}\varphi_x\left(\varphi_x^{-1}(\overline{x})\right)\left[\mathrm{D}\left(\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{x},\cdot_y)\right)\right)(\overline{y})[\overline{v_y}]\right],
\end{aligned}
$$

where $\varphi^{-1}(\cdot_x,\overline{y})$ and $\varphi^{-1}(\overline{x},\cdot_y)$ denote the maps $\overline{x} \mapsto \varphi^{-1}((\overline{x},\overline{y})) = (x,y)$ and $\overline{y} \mapsto \varphi^{-1}((\overline{x},\overline{y})) = (x,y)$, respectively. We also have

(C-16)
$$
\begin{aligned}
&\mathrm{D}\left(\mathrm{grad}_x\mathcal{L}\left(\varphi^{-1}(\overline{x},\cdot_y)\right)\right)(\overline{y})[\overline{v_y}] \\
&= -\sum_{i\in\mathcal{I}}\left[\mathrm{D}\varphi_y^{-1}(\overline{y})[\overline{v_y}]\right]_i\mathrm{grad}g_i\left(\varphi_x^{-1}(\overline{x})\right) = -\mathcal{G}_{\varphi_x^{-1}(\overline{x})}[\overline{v_y}],
\end{aligned}
$$

where the second equality follows since $\mathrm{D}\varphi_y^{-1}(\overline{y})$ is the identity map. Here, it follows from $\mathrm{grad}_x\mathcal{L}(\omega^*) = 0_{x^*}$ and [1, Equation (5.7)] that

$$
\begin{aligned}
\text{(C-17)} \quad & \mathrm{D}\big(\mathrm{grad}_x\mathcal{L}\big(\varphi^{-1}(\cdot_x, \overline{y}^*)\big)\big)\big(\overline{x}^*\big)\big[\overline{v_{x^*}}\big] \\
&\overset{}{=} \nabla_{\mathrm{D}\varphi_x^{-1}(x^*)[\overline{v_{x^*}}]}\mathrm{grad}_x\mathcal{L}\big(\varphi^{-1}(\overline{\omega}^*)\big) = \mathrm{Hess}_x\mathcal{L}\big(\varphi^{-1}(\overline{\omega}^*)\big) \circ \mathrm{D}\varphi_x^{-1}(\overline{x}^*)[\overline{v_x}],
\end{aligned}
$$

where the second equality follows from the definition of the Riemannian Hessian. Combining (C-16) and (C-17) with (C-15) and $\mathrm{grad}_x\mathcal{L}(\omega^*) = 0_{x^*}$, again, yields

$$
\begin{aligned}
\text{(C-18)} \quad & \mathrm{D}\big(\mathrm{D}\varphi_x\big(\varphi_x^{-1}(\cdot_x)\big)\big[\mathrm{grad}_x\mathcal{L}\big(\varphi^{-1}(\cdot_\omega)\big)\big]\big)\big(\overline{\omega}^*\big)[\overline{v}] \\
&= \mathrm{D}\varphi_x\big(\varphi_x^{-1}(\overline{x}^*)\big) \circ \mathrm{Hess}_x\mathcal{L}\big(\varphi^{-1}(\overline{\omega}^*)\big) \circ \mathrm{D}\varphi_x^{-1}(\overline{x}^*)[\overline{v_x}] \\
&\quad - \mathrm{D}\varphi_x\big(\varphi_x^{-1}(\overline{x}^*)\big)\Big[\mathcal{G}_{\varphi_x^{-1}(\overline{x}^*)}[\overline{v_y}]\Big].
\end{aligned}
$$

Next, we consider the second component of (C-14): for each $i \in \mathcal{I}$,

$$
\begin{aligned}
& \Big[\mathrm{D}\big(G\big(\varphi_x^{-1}(\cdot_x)\big)\varphi_y^{-1}(\cdot_y) - \mu\mathbf{1}\big)(\overline{\omega})[\overline{v}]\Big]_i \\
&= \mathrm{D}\big(g_i\big(\varphi_x^{-1}(\cdot_x)\big)\big[\varphi_y^{-1}(\cdot_y)\big]_i - \mu\mathbf{1}\big)(\overline{\omega})[\overline{v}] \\
&= \mathrm{D}g_i\big(\varphi_x^{-1}(\overline{x})\big)\big[\mathrm{D}\varphi_x^{-1}(\overline{x})[\overline{v_x}]\big] \cdot y_i + g_i\big(\varphi_x^{-1}(\overline{x})\big) \cdot \big[\mathrm{D}\varphi_y^{-1}(\overline{y})[\overline{v_y}]\big]_i \\
&= y_i\big\langle \mathrm{grad}g_i\big(\varphi_x^{-1}(\overline{x})\big), \mathrm{D}\varphi_x^{-1}(\overline{x})[\overline{v_x}]\big\rangle_{\varphi_x^{-1}(\overline{x})} + g_i\big(\varphi_x^{-1}(\overline{x})\big) \cdot [\overline{v_y}]_i,
\end{aligned}
$$

which implies

$$
\begin{aligned}
\text{(C-19)} \quad & \mathrm{D}\big(G\big(\varphi_x^{-1}(\cdot_x)\big)\varphi_y^{-1}(\cdot_y) - \mu\mathbf{1}\big)(\overline{\omega})[\overline{v}] \\
&= Y\mathcal{G}^*_{\varphi_x^{-1}(\overline{x})}\big[\mathrm{D}\varphi_x^{-1}(\overline{x})[\overline{v_x}]\big] + G\big(\varphi_x^{-1}(\overline{x})\big)\overline{v_y}.
\end{aligned}
$$

Substituting (C-18) and (C-19) into (C-14) yields

$$
\begin{aligned}
\mathrm{D}F_\mu(\overline{\omega}^*) &= \mathrm{D}\varphi\big(\varphi^{-1}(\overline{\omega}^*)\big) \circ \begin{bmatrix} \mathrm{Hess}_x\mathcal{L}\big(\varphi^{-1}(\overline{\omega}^*)\big) & -\mathcal{G}_{\varphi_x^{-1}(\overline{x}^*)} \\ Y^*\mathcal{G}^*_{\varphi_x^{-1}(\overline{x}^*)} & G\big(\varphi_x^{-1}(\overline{x}^*)\big) \end{bmatrix} \circ \mathrm{D}\varphi^{-1}(\overline{\omega}^*) \\
&= \mathrm{D}\varphi\big(\varphi^{-1}(\overline{\omega}^*)\big) \circ \mathrm{J}_\Psi\big(\varphi^{-1}(\overline{\omega}^*)\big) \circ \mathrm{D}\varphi^{-1}(\overline{\omega}^*),
\end{aligned}
$$

where we write $Y^* \in \mathbb{R}^{m \times m}$ for $\mathrm{diag}(y^*)$ and

$$
\begin{aligned}
\mathrm{D}\varphi\big(\varphi^{-1}(\overline{\omega}^*)\big) &: T_{x^*}\mathcal{M} \times \mathbb{R}^m \to \mathbb{R}^d \times \mathbb{R}^m \\
(v_{x^*}, v_{y^*}) &\mapsto \begin{bmatrix} \mathrm{D}\varphi_x\big(\varphi_x^{-1}(\overline{x}^*)\big)[v_{x^*}] \\ v_{y^*} \end{bmatrix}, \\
\mathrm{D}\varphi^{-1}(\overline{\omega}^*) &: \mathbb{R}^d \times \mathbb{R}^m \to T_{x^*}\mathcal{M} \times \mathbb{R}^m \\
(\overline{v_{x^*}}, \overline{v_{y^*}}) &\mapsto \begin{bmatrix} \mathrm{D}\varphi_x\big(\varphi_x^{-1}(\overline{x}^*)\big)[\overline{v_{x^*}}] \\ v_{y^*} \end{bmatrix}.
\end{aligned}
$$

Note that, since the maps $\mathrm{D}\varphi\big(\varphi^{-1}(\overline{\omega}^*)\big)$ and $\mathrm{D}\varphi^{-1}(\overline{\omega}^*)$ are bijective and the operator $\mathrm{J}_\Psi\big(\varphi^{-1}(\overline{\omega}^*)\big)$ is nonsingular by Lemma 5.1, $\mathrm{D}F_\mu(\overline{\omega}^*)$ is also nonsingular. $\qquad\square$

Proof of Lemma 5.14. We first consider Item 1. Since $\tilde{F}(\overline{\omega}, 0, 0) = F(\overline{\omega}, 0)$ holds by definition, it follows that $\tilde{F}(\overline{\omega}^*, 0, 0) = F(\overline{\omega}^*, 0) = 0$ and $\mathrm{D}\tilde{F}(\cdot, 0, 0)(\overline{\omega}^*) = \mathrm{D}F_0(\overline{\omega}^*)$ is nonsingular by Lemma 5.13. Thus, by the implicit function theorem [41, Theorem C.40], there exist a positive scalar $\varepsilon > 0$ and a continuously differentiable

function $\overline{\omega}(\overline{l}, \overline{\xi}) \colon \mathcal{N}(\varepsilon) \to \mathbb{R}^d$ such that $\overline{\omega}(0,0) = \overline{\omega^*}$ and $\tilde{F}\big(\overline{\omega}(\overline{l}, \overline{\xi}), \overline{l}, \overline{\xi}\big) = 0$ for any $(\overline{l}, \overline{\xi}) \in \mathcal{N}(\varepsilon)$. Notice that the implicit function theorem also ensures the existence and uniqueness of $\overline{\omega}(\overline{l}, \overline{\xi})$ in $\mathcal{N}(\varepsilon)$. The proof of Item 1 is complete.

Next, we consider Item 2. By taking $\varepsilon > 0$ smaller if necessary, we have

$$
\begin{aligned}
&\overline{\omega}\big(\overline{l}_2, \overline{\xi}_2\big) \\
\text{(C-20)} \quad &= \overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big) + \int_0^1 \mathrm{D}\overline{\omega}\big(\big(\overline{l}_1, \overline{\xi}_1\big) + t\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big)\big[\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big] \, \mathrm{d}t \, .
\end{aligned}
$$

for all $\overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big), \overline{\omega}\big(\overline{l}_2, \overline{\xi}_2\big) \in \mathcal{N}(\varepsilon)$. We first consider the upper bound on the norm of the difference between $\overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)$ and $\overline{\omega}\big(\overline{l}_2, \overline{\xi}_2\big)$. There exists $\tilde{c}' > 0$ such that

$$
\begin{aligned}
&\big\|\overline{\omega}\big(\overline{l}_2, \overline{\xi}_2\big) - \overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big\| \\
\text{(C-21)} \quad &\leq \int_0^1 \big\|\mathrm{D}\overline{\omega}\big(\big(\overline{l}_1, \overline{\xi}_1\big) + t\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big)\big\|_{\mathrm{op}}\big\|\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big\| \, \mathrm{d}t \\
&\leq \tilde{c}'\big\|\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big\| \leq \tilde{c}'\big(\big\|\overline{l}_2 - \overline{l}_1\big\| + \big\|\overline{\xi}_2 - \overline{\xi}_1\big\|\big),
\end{aligned}
$$

where the second inequality follows from the continuity of $\mathrm{D}\overline{\omega}(\cdot, \cdot)$. We next derive the lower bound on $\big\|\overline{\omega}\big(\overline{l}_2, \overline{\xi}_2\big) - \overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big\|$. To this end, we derive auxiliary bounds as follows: by (5.34) and the definition of $\overline{\omega}(\overline{l}, \overline{\xi})$, it follows that $F_0\big(\overline{\omega}(\overline{l}, \overline{\xi})\big) = \begin{bmatrix} \overline{l} \\ \overline{\xi} \end{bmatrix}$. Differentiating it yields

$$
\mathrm{D}F_0\big(\overline{\omega}(\overline{l}, \overline{\xi})\big) = \mathrm{id},
$$

which implies

$$
\mathrm{D}F_0\big(\overline{\omega}(\overline{l}, \overline{\xi})\big) \circ \mathrm{D}\overline{\omega}(\overline{l}, \overline{\xi}) = \mathrm{id} \, .
$$

Since $\mathrm{D}F_0(\cdot)$ is nonsingular at $\overline{\omega^*}$ and is continuous by definition, $\mathrm{D}F_0(\cdot)$ remains nonsingular around $\overline{\omega^*}$ and hence

$$
\text{(C-22)} \qquad \mathrm{D}\overline{\omega}(\overline{l}, \overline{\xi}) = \mathrm{D}F_0^{-1}\big(\overline{\omega}(\overline{l}, \overline{\xi})\big)
$$

for any $(\overline{l}, \overline{\xi}) \in \mathcal{N}(\varepsilon)$ by taking $\varepsilon > 0$ smaller if necessary. Therefore, there exists $\underline{c}' > 0$ such that

$$
\begin{aligned}
&\big\|\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big\| \\
&= \big\|\mathrm{D}F_0\big(\overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big) \circ \mathrm{D}F_0^{-1}\big(\overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big)\big[\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big]\big\| \\
\text{(C-23)} \quad &\leq \big\|\mathrm{D}F_0\big(\overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big)\big\|_{\mathrm{op}}\big\|\mathrm{D}F_0^{-1}\big(\overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big)\big[\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big]\big\| \\
&\leq \frac{1}{(1 + \sqrt{2})\underline{c}'}\big\|\mathrm{D}\overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big[\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big]\big\|,
\end{aligned}
$$

where the second inequality follows from (C-22) and the boundedness of $\mathrm{D}F_0(\cdot)$ around $\overline{\omega^*}$. By taking $\varepsilon > 0$ smaller again if necessary, we also have

$$
\text{(C-24)} \qquad \big\|\mathrm{D}\big(\overline{\omega}\big(\big(\overline{l}_1, \overline{\xi}_1\big) + t\big(\overline{l}_2 - \overline{l}_1, \overline{\xi}_2 - \overline{\xi}_1\big)\big) - \overline{\omega}\big(\overline{l}_1, \overline{\xi}_1\big)\big)\big\|_{\mathrm{op}} \leq \underline{c}'
$$

for any $(\bar{l}, \bar{\xi}) \in \mathcal{N}(\varepsilon)$ by the continuous differentiability of $\bar{\varpi}(\cdot, \cdot)$. Now, we derive the lower bound on $\|\bar{\varpi}(\bar{l}_2, \bar{\xi}_2) - \bar{\varpi}(\bar{l}_1, \bar{\xi}_1)\|$:

$$\|\bar{\varpi}(\bar{l}_2, \bar{\xi}_2) - \bar{\varpi}(\bar{l}_1, \bar{\xi}_1)\| \geq \|\mathrm{D}\bar{\varpi}(\bar{l}_1, \bar{\xi}_1)[(\bar{l}_2 - \bar{l}_1, \bar{\xi}_2 - \bar{\xi}_1)]\| - \|\bar{\varpi}(\bar{l}_2, \bar{\xi}_2) - \bar{\varpi}(\bar{l}_1, \bar{\xi}_1)$$
$$- \mathrm{D}\bar{\varpi}(\bar{l}_1, \bar{\xi}_1)[(\bar{l}_2 - \bar{l}_1, \bar{\xi}_2 - \bar{\xi}_1)]\| \geq \left(1 + \sqrt{2}\right)\underline{c}'\|(\bar{l}_2 - \bar{l}_1, \bar{\xi}_2 - \bar{\xi}_1)\|$$
$$\text{(C-25)} \quad - \int_0^1 \|\mathrm{D}\left(\bar{\varpi}((\bar{l}_1, \bar{\xi}_1) + t(\bar{l}_2 - \bar{l}_1, \bar{\xi}_2 - \bar{\xi}_1)) - \bar{\varpi}(\bar{l}_1, \bar{\xi}_1)\right)\|_{\mathrm{op}}\|(\bar{l}_2 - \bar{l}_1, \bar{\xi}_2 - \bar{\xi}_1)\|\,\mathrm{d}t$$
$$\geq \sqrt{2}\underline{c}'\|(\bar{l}_2 - \bar{l}_1, \bar{\xi}_2 - \bar{\xi}_1)\| \geq \underline{c}'\left(\|\bar{l}_2 - \bar{l}_1\| + \|\bar{\xi}_2 - \bar{\xi}_1\|\right)$$

for any $(\bar{l}, \bar{\xi}) \in \mathcal{N}(\varepsilon)$, where the second inequality follows from (C-20) and (C-23), the third one from (C-24), and the fourth one from

$$\left\|(\bar{l}_2 - \bar{l}_1, \bar{\xi}_2 - \bar{\xi}_1)\right\|^2 = \|\bar{l}_2 - \bar{l}_1\|^2 + \|\bar{\xi}_2 - \bar{\xi}_1\|^2 \geq \|\bar{l}_2 - \bar{l}_1\|^2 + \|\bar{\xi}_2 - \bar{\xi}_1\|^2$$
$$- \frac{1}{2}\left(\|\bar{l}_2 - \bar{l}_1\| - \|\bar{\xi}_2 - \bar{\xi}_1\|\right)^2 = \frac{1}{2}\left(\|\bar{l}_2 - \bar{l}_1\| + \|\bar{\xi}_2 - \bar{\xi}_1\|\right)^2.$$

From (C-21) and (C-25), equation (5.35) holds. The proof of Item 2 is complete. $\square$