Riemannian Bilevel Optimization

Jiaxiang Li [*] Shiqian Ma [†]

Feb 2, 2024

Abstract

In this work, we consider the bilevel optimization problem on Riemannian manifolds. We inspect the calculation of the hypergradient of such problems on general manifolds and thus enable the utilization of gradient-based algorithms to solve such problems. The calculation of the hypergradient requires utilizing the notion of Riemannian cross-derivative and we inspect the properties and the numerical calculations of Riemannian cross-derivatives. Algorithms in both deterministic and stochastic settings, named respectively RieBO and RieSBO, are proposed that include the existing Euclidean bilevel optimization algorithms as special cases. Numerical experiments on robust optimization on Riemannian manifolds are presented to show the applicability and efficiency of the proposed methods.

1 Introduction

Bilevel optimization has drawn attentions from various fields in optimization and machine learning communities, due to its wide range of applications including meta learning (Rajeswaran et al., 2019; Ji et al., 2020), hyperparameter optimization (Okuno et al., 2021; Yu and Zhu, 2020), reinforcement learning (Konda and Tsitsiklis, 1999; Hong et al., 2020) and signal processing (Kunapuli et al., 2008; Flamary et al., 2014). In this work, we focus on the manifold-constrained bilevel optimization problem, which can be formulated as:

$$\begin{aligned} &\min_{x\in\mathcal{M}} \Phi(x) := f(x, y^*(x)) \\ &\text{s.t. } y^*(x) = \operatorname*{argmin}_{y\in\mathcal{N}} g(x,y), \end{aligned} \tag{1.1}$$

where $\mathcal{M}$ and $\mathcal{N}$ are m and n-dimensional complete Riemannian manifolds, respectively. We also consider the stochastic bilevel optimization which is in the following form:

$$\begin{aligned} &\min_{x\in\mathcal{M}} \Phi(x) = f\left(x, y^*(x)\right) := \mathbb{E}_{\xi}\left[F\left(x, y^*(x); \xi\right)\right] \\ &\text{s.t. } y^*(x) = \operatorname*{argmin}_{y\in\mathcal{N}} g(x,y) := \mathbb{E}_{\zeta}[G(x,y;\zeta)], \end{aligned} \tag{1.2}$$

where ξ and ζ are random variables that usually represent the randomness from the data. Such a framework allows us to utilize the stochastic gradient methods to get a desired convergence result with only a noisy estimate of the gradients for f and g. Here we also assume that g is a (geodesically)

[*] Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. `li003755@umn.edu`

[†] Department of Computational Applied Math and Operations Research, Rice University. Research supported in part by NSF grants DMS-2243650, CCF-2308597, CCF-2311275 and ECCS-2326591, and a startup fund from Rice University. `sqma@rice.edu`

strongly convex function with respect to y in both (1.1) and (1.2) so that the solution to the lower level problem $y^*(x)$ is well-defined.

Notice that the original (Euclidean) bilevel optimization is a special case of (1.1) by taking the manifolds as the Euclidean spaces with the same dimensions:

$$\begin{aligned} &\min_{x\in\mathbb{R}^m} \Phi(x) := f(x, y^*(x)) \\ &\text{s.t. } y^*(x) = \operatorname*{argmin}_{y\in\mathbb{R}^n} g(x, y), \end{aligned} \tag{1.3}$$

where f and g are assumed to be continuously differentiable. It is worth noticing that the objective function $\Phi(x)$ is still nonconvex even if we impose convexity assumptions on f, which makes such a problem hard to tackle, let alone the more complicated manifold-constraint problems, namely (1.1) and (1.2).

There has been an extensive study on the Euclidean bilevel optimization (Ji et al., 2021; Hong et al., 2020; Chen et al., 2021b; Ghadimi and Wang, 2018). On the algorithmic sense, the bilevel optimization seeks to obtain a first-order ϵ-stationary point (Definition 4.1 and 5.1 for deterministic and stochastic cases, respectively) with the access to the gradient oracle of f and g, as well as the Jacobian- and Hessian-vector product, i.e. $\nabla_x\nabla_y g(x, y)v$ and $\nabla_y^2 g(x, y)v$, respectively. To find an ϵ-stationary point, we denote the number of calls to the gradient oracle of f and g as $\mathrm{Gc}(f, \epsilon)$ and $\mathrm{Gc}(g, \epsilon)$, correspondingly; similarly we have the notation JV and HV for the number of oracle calls for the Jacobian- and Hessian-vector product. In the Euclidean setting, We have Tables 1 and 2 as the summary for the oracle calls to achieve an ϵ-stationary point for deterministic and stochastic cases, correspondingly (and we denote κ as the condition number of the lower level strongly convex problem).

Algorithm	BA	AID-BiO	ITD-BiO*
y-update	GD	GD	GD
$\mathrm{Gc}(f, \epsilon)$	$\mathcal{O}(\kappa^4\epsilon^{-1})$	$\mathcal{O}(\kappa^3\epsilon^{-1})$	$\mathcal{O}(\kappa^3\epsilon^{-1})$
$\mathrm{Gc}(g, \epsilon)$	$\mathcal{O}(\kappa^5\epsilon^{-5/4})$	$\mathcal{O}(\kappa^4\epsilon^{-1})$	$\mathcal{O}(\kappa^4\epsilon^{-1})$
$\mathrm{JV}(g, \epsilon)$	$\mathcal{O}(\kappa^4\epsilon^{-1})$	$\mathcal{O}(\kappa^3\epsilon^{-1})$	$\mathcal{O}(\kappa^4\epsilon^{-1})$
$\mathrm{HV}(g, \epsilon)$	$\mathcal{O}(\kappa^{4.5}\epsilon^{-1})$	$\mathcal{O}(\kappa^{3.5}\epsilon^{-1})$	$\mathcal{O}(\kappa^4\epsilon^{-1})$

* Require explicit assumption on the sequence.

Table 1: Summary of the convergence results for different algorithms for **deterministic** Euclidean bilevel optimization, including BA (Ghadimi and Wang, 2018), AID-BiO (Ji et al., 2021) and ITD-BiO (Ji et al., 2021).

1.1 Main results

In this work, we first analyze the method of calculating and estimating the hypergradient for bilevel problems on Riemannian manifolds. Our propositions include the Euclidean bilevel problems as special cases and involve the calculation of Riemannian cross derivatives, which are of independent interests to the Riemannian optimization field.

Our contribution also lies in proposing two algorithms (RieBO and RieSBO) for both the problems (1.1) and (1.2) correspondingly. For the deterministic problem (1.1), our analysis shows that with a multi-step inner loop and a single-step outer loop, one could yield the similar gradient complexities $\mathrm{Gc}(f, \epsilon)$, $\mathrm{Gc}(g, \epsilon)$, Jacobian- and Hessian-vector product complexities $\mathrm{JV}(g, \epsilon)$ and $\mathrm{HV}(g, \epsilon)$ same as the Euclidean counterparts in Ji et al. (2021), as presented in Table 3, al well as

Algorithm	BSA	Stoc-BiO	TTSA*	ALSET	STABLE*
batch size	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
y-update	$\mathcal{O}(\epsilon^{-1})$ steps SGD	SGD	SGD	SGD	correction
$\mathrm{Gc}(F,\epsilon)$	$\mathcal{O}(\kappa^6\epsilon^{-2})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2.5})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2})$
$\mathrm{Gc}(G,\epsilon)$	$\mathcal{O}(\kappa^9\epsilon^{-2})$	$\mathcal{O}(\kappa^9\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2.5})$	$\mathcal{O}(\kappa^9\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2})$
$\mathrm{JV}(G,\epsilon)$	$\mathcal{O}(\kappa^6\epsilon^{-2})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2.5})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2})$
$\mathrm{HV}(G,\epsilon)$	$\mathcal{O}(\kappa^6\epsilon^{-2})$	$\mathcal{O}(\kappa^6\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2.5})$	$\mathcal{O}(\kappa^6\epsilon^{-2})$	$\mathcal{O}(\mathrm{poly}(\kappa)\epsilon^{-2})$

* For algorithms that did not specify the dependence on condition number κ, we use the notation $\mathrm{poly}(\kappa)$ to summarize the κ dependence.

Table 2: Summary of the convergence results for different algorithms for **stochastic** Euclidean bilevel optimization, including BSA (Ghadimi and Wang, 2018), Stoc-BiO (Ji et al., 2021), TTSA (Hong et al., 2020), ALSET (Chen et al., 2021b) and STABLE (Chen et al., 2021a). For the batch size we only include the ϵ dependency.

for the stochastic problem (1.2) (Chen et al., 2021b). It is worth noticing that for the stochastic problem, we adopt the framework of Chen et al. (2021b) onto Riemannian manifolds so that the batch-size of the hypergradient estimate can be $\mathcal{O}(1)$, significantly smaller than $\mathcal{O}(\epsilon^{-1})$ as in Ji et al. (2021).

Algorithm	RieBO (Algorithm 1)	RieSBO (Algorithm 2)
batch size	No batch	$\mathcal{O}(1)$
y-update	GD	SGD
$\mathrm{Gc}(F,\epsilon)$	$\mathcal{O}(\kappa^3\epsilon^{-1})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$
$\mathrm{Gc}(G,\epsilon)$	$\mathcal{O}(\kappa^4\epsilon^{-1})$	$\mathcal{O}(\kappa^9\epsilon^{-2})$
$\mathrm{JV}(G,\epsilon)$	$\mathcal{O}(\kappa^3\epsilon^{-1})$	$\mathcal{O}(\kappa^5\epsilon^{-2})$
$\mathrm{HV}(G,\epsilon)$	$\mathcal{O}(\kappa^{3.5}\epsilon^{-1})$	$\mathcal{O}(\kappa^6\epsilon^{-2})$

Table 3: Summary of the convergence results for the proposed algorithms in this paper, where all the oracles are with respect to Riemannian gradients and Riemannian second-order derivatives. RieBO (Algorithm 1) solves (1.1), and RieSBO (Algorithm 2) solves (1.2).

Finally, we implement the proposed method in the manifold-constrained bilevel optimization problems, namely the distributionally robust optimization on Riemannian manifolds with two specific examples: robust maximum likelihood estimation and robust Karcher mean problem on the manifold of positive definite matrices. These numerical results demonstrate the efficiency and potential applicability of the proposed methods.

1.2 Related works

Bilevel optimization. Bilevel optimization problem, also known as nested optimization problem, whose origin dates back to the 50s and 70s (Stackelberg and Peacock, 1952; Bracken and McGill, 1973). Since then, extensive studies have been conducted for solving the bilevel optimization problem (Shi et al., 2005; Moore, 2010). Recently, gradient-based algorithms for solving bilevel optimization problems draw attention because of their applications in machine learning, see, e.g. Domke (2012); Pedregosa (2016); Gould et al. (2016); Maclaurin et al. (2015); Franceschi et al. (2018); Liao et al. (2018); Shaban et al. (2019); Liu et al. (2020); Li et al. (2020); Grazzi et al. (2020); Lorraine et al. (2020); Ji and Liang (2021); Ghadimi and Wang (2018); Hong et al. (2020); Ji et al. (2021); Chen et al. (2021b,a), among which there has been discussions about the rate of convergence for specific algorithms (Ghadimi and Wang, 2018; Hong et al., 2020; Ji et al., 2021; Chen et al., 2021a,b). These well-established convergence rate results are summarized in Tables 1 and 2. Recently, a line of work (Khanduri et al., 2021; Yang et al., 2023) which utilizes the momentum-based stochastic algorithms can achieve a better oracle complexity of $\mathcal{O}(\epsilon^{-1.5})$ for the Euclidean version of the stochastic problem (1.2). We did not include this line of work since the Riemannian counterparts of these works would rely on the utilization of parallel/vector transport in the algorithm updates. We deliberately avoid these complicated operations in algorithm design and postpone them for future works.

It is worth mentioning that minimax saddle point problems $\min_x \max_y f(x, y)$ are special cases of bilevel optimization problems by taking $g = -f$. Minimax problems are of great interests to the machine learning community (Daskalakis and Panageas, 2018; Mokhtari et al., 2020; Yoon and Ryu, 2021; Lin et al., 2020b). The analysis of this paper relates to the nonconvex-strongly-concave minimax problem on Riemannian manifolds as in Huang et al. (2020), which showed that the Riemannian gradient descent ascent (RGDA) achieves oracle calls with orders $\mathcal{O}(\kappa^2\epsilon^{-1})$ for the deterministic case and $\mathcal{O}(\kappa^3\epsilon^{-2})$ for the stochastic case. These results match our convergence results in terms of the order of ϵ, but has better κ dependence. This makes sense because our proposed method is a multi-y step GDmax algorithm (see (Nouiehed et al., 2019; Jin et al., 2020)) when applied to the minimax problem and naturally has a larger κ dependence. Recently, the authors of Cai et al. (2023) considered the minimax game on Riemannian manifolds under the assumption of geodesic-strongly-monotone (a generalization of strongly-convex-strongly-concave minimax game) and provided a stochastic Riemannian gradient descent-ascent approach which enjoys linear rate of convergence – similar to its Euclidean counterpart. We point out that our work considers nonconvex upper level problems, which is different from the setting in Cai et al. (2023).

Other bilevel-related ongoing research topics include decentralized bilevel optimization Chen et al. (2022, 2023b); Dong et al. (2023), federate bilevel optimization Tarzanagh et al. (2022), bilevel without lower strongly convexity Chen et al. (2023a), to name a few.

Optimization on Riemannian manifolds. Optimization on Riemannian manifolds draws lots of attention recently due to its applications in various fields, including low-rank matrix completion (Boumal and Absil, 2011; Vandereycken, 2013), phase retrieval (Bendory et al., 2017; Sun et al., 2018), dictionary learning (Cherian and Sra, 2016; Sun et al., 2016), dimensionality reduction (Harandi et al., 2017; Tripuraneni et al., 2018; Mishra et al., 2019) and manifold regression (Lin et al., 2017, 2020a). The manifold optimization usually transforms an manifold constrained problem into an unconstrained problem by viewing the manifold as the ambient space and using proper retraction to deal with the loss of linearity, thus achieves better convergence results. For smooth Riemannian optimization, it can be shown that Riemannian gradient descent method require $\mathcal{O}(1/\epsilon)$ iterations to converge to an ϵ-stationary point (i.e. bounding the norm square of the gradient by ϵ) (Boumal et al., 2018). Stochastic algorithms were also studied for smooth Riemannian optimization (Bonnabel, 2013; Zhou et al., 2019; Weber and Sra, 2019; Zhang et al., 2016; Kasai et al., 2018).

The combination of bilevel optimization with Riemannian optimization is largely blank. Bonnel et al. (2015) considered a semi-vectorial bilevel optimization model over Riemannian manifolds, which deals with the situation where the lower level problem does not have unique solutions and necessary optimality conditions are provided for their surrogate model. To the best of our knowledge, there lacks convergence analysis for Riemannian bilevel optimization in the literature.

2 Preliminaries on Riemannian Optimization

In this part, we briefly review the basic tools we use for optimization on Riemannian manifolds (Lee, 2006; Tu, 2011; Boumal, 2023). Suppose $\mathcal{M}$ is an m-dimensional differentiable manifold. The tangent space $\mathrm{T}_x\,\mathcal{M}$ at $x \in \mathcal{M}$ is a linear subspace that consists of the derivatives of all differentiable curves on $\mathcal{M}$ passing through x: $\mathrm{T}_x\,\mathcal{M} := \{\gamma'(0) : \gamma(0) = x, \gamma([-\delta, \delta]) \subset \mathcal{M}$ for some $\delta > 0, \gamma$ is differentiable$\}$. Notice that for every vector $\gamma'(0) \in \mathrm{T}_x\,\mathcal{M}$, it can be defined in a coordinate-free sense via the operation over smooth functions: $\forall f \in C^\infty(\mathcal{M})$, $\gamma'(0)(f) := \frac{df\circ\gamma(t)}{dt}\mid_{t=0}$. The notion of Riemannian manifold is defined as follows.

Definition 2.1 (Riemannian manifold)**.** *Manifold $\mathcal{M}$ is a Riemannian manifold if it is equipped with an* ***inner product*** *on the tangent space, $\langle\cdot,\cdot\rangle_x : \mathrm{T}_x\,\mathcal{M} \times \mathrm{T}_x\,\mathcal{M} \to \mathbb{R}$, that varies smoothly on $\mathcal{M}$. The $(0,2)$-tensor field g is usually referred to as Riemannian metric.*

We also review the notion of the differential between manifolds here.

Definition 2.2 (Differential and Riemannian gradients)**.** *Let $F : \mathcal{M} \to \mathcal{N}$ be a C^∞ map between two differential manifolds. At each point $x \in \mathcal{M}$, the differential of F is a mapping (also known as the push-forward):*

$$\mathrm{D}F : \mathrm{T}_x\,\mathcal{M} \to \mathrm{T}_{F(x)}\,\mathcal{N},$$

such that $\forall \xi \in \mathrm{T}_x\,\mathcal{M}$, $\mathrm{D}F(\xi) \in \mathrm{T}_x\,\mathcal{N}$ is given by

$$(\mathrm{D}F(\xi))(f) := \xi(f \circ F) \in \mathbb{R},\ \forall f \in C^\infty_{F(x)}(\mathcal{M}).$$

If $\mathcal{N} = \mathbb{R}$, i.e. $f \in C^\infty(\mathcal{M})$, the differential of f is usually denoted as $\mathrm{d}f$. For a Riemannian manifold with Riemannian metric $\langle\cdot,\cdot\rangle$, the Riemannian gradient for $f \in C^\infty(\mathcal{M})$ is the unique tangent vector $\mathsf{grad}f(x) \in \mathrm{T}_x\,\mathcal{M}$ satisfying

$$\mathrm{d}f(\xi) = \langle \mathsf{grad}f, \xi\rangle_x,\ \forall \xi \in \mathrm{T}_x\,\mathcal{M}.$$

For the convergence analysis, we also need the notion of exponential mapping and parallel transport. To this end, we need to first recall the definition of a geodesic.

Definition 2.3 (Geodesic and exponential mapping)**.** *Given $x \in \mathcal{M}$ and $\xi \in \mathrm{T}_x\,\mathcal{M}$, the geodesic is the curve $\gamma : I \to \mathcal{M}$, $0 \in I \subset \mathbb{R}$ is an open set, so that $\gamma(0) = x$, $\dot\gamma(0) = \xi$ and $\nabla_{\dot\gamma}\dot\gamma = 0$ where $\nabla : \mathrm{T}_x\,\mathcal{M} \times \mathrm{T}_x\,\mathcal{M} \to \mathrm{T}_x\,\mathcal{M}$ is the Levi-Civita connection defined by metric g. In local coordinate sense, γ is the unique solution of the following second-order differential equations:*

$$\frac{d^2\gamma^k}{dt^2} + \Gamma^k_{i,j}\frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt} = 0,$$

under Einstein summation convention, where $\Gamma^k_{i,j}$ are Christoffel symbols, again defined by metric tensor. The exponential mapping Exp_x is defined as a mapping from $\mathrm{T}_x\,\mathcal{M}$ to $\mathcal{M}$ s.t. $\mathsf{Exp}_x(\xi) := \gamma(1)$ with γ being the geodesic with $\gamma(0) = x$, $\dot\gamma(0) = \xi$. A natural corollary is $\mathsf{Exp}_x(t\xi) := \gamma(t)$ for $t \in [0,1]$. Another useful fact is $\mathrm{dist}(x, \mathsf{Exp}_x(\xi)) = \|\xi\|_x$ since $\gamma'(0) = \xi$ which preserves the speed. Here dist is the geodesic distance which connects the two points by the minimum geodesic.

Throughout this paper, we always assume that $\mathcal{M}$ is complete, so that Exp_x is always defined for every $\xi \in \mathrm{T}_x\,\mathcal{M}$. For any $x, y \in \mathcal{M}$, the inverse of the exponential mapping $\mathsf{Exp}_x^{-1}(y) \in \mathrm{T}_x\,\mathcal{M}$ is called the logarithm mapping, and we have $\mathrm{dist}(x, y) = \|\mathsf{Exp}_x^{-1}(y)\|_x$, which derives directly from $\mathrm{dist}(x, \mathsf{Exp}_x(\xi)) = \|\xi\|_x$.

With the notion of geodesic, we have the following definition of geodesic convexity and strong convexity, which are the generalizations of their Euclidean counterparts.

Definition 2.4 (Geodesic (strong) convexity)**.** *A geodesic convex set $\Omega \subset \mathcal{M}$ is a set such that for any two points in the set, there exists a geodesic connecting them that lies entirely in Ω. A function $h : \Omega \to \mathbb{R}$ is called geodesically convex if for any $p, q \in \Omega$, we have $h(\gamma(t)) \leq (1-t)h(p) + th(q)$, where γ is a geodesic in Ω with $\gamma(0) = p$ and $\gamma(1) = q$. It is called μ-geodesically strongly convex if we have $h(\gamma(t)) \leq (1-t)h(p) + th(q) - \frac{\mu t(1-t)}{2} \operatorname{dist}(p,q)^2$.*

If h is a continuously differentiable function, then it is geodesically convex if and only if (see Boumal (2023, Chapter 11)) $h(y) \geq h(x) + \langle \mathsf{grad} h(x), \mathsf{Exp}_x^{-1}(y)\rangle_x$, and is geodesically strongly convex if and only if $h(y) \geq h(x) + \langle \mathsf{grad} h(x), \mathsf{Exp}_x^{-1}(y)\rangle_x + \frac{\mu}{2}\operatorname{dist}(x,y)^2$.

If h is a twice continuously differentiable function, then it is geodesically convex if and only if (see Boumal (2023, Chapter 11)) $\frac{d^2 h(\gamma(t))}{dt^2} \geq 0$, and is geodesically strongly convex if and only if $\frac{d^2 h(\gamma(t))}{dt^2} \geq \mu$.

We also present the definition of parallel transport, which is used in the assumption and the convergence analysis, but not explicitly used in the algorithm updates.

Definition 2.5 (Parallel transport)**.** *Given a Riemannian manifold $(\mathcal{M}, g)$ and two points $x, y \in \mathcal{M}$, the parallel transport $P_{x\to y} : \mathrm{T}_x \mathcal{M} \to \mathrm{T}_y \mathcal{M}$ is a linear operator which keeps the inner product: $\forall \xi, \zeta \in \mathrm{T}_x \mathcal{M}$, we have $\langle P_{x\to y}\xi, P_{x\to y}\zeta\rangle_y = \langle \xi, \zeta\rangle_x$.*

Notice that the existence of parallel transport depends on the curve connecting x and y, which is not a problem for complete Riemannian manifold since we always take the unique geodesic that connects x and y. Parallel transport is useful in our convergence proofs since the Lipschitz condition for the Riemannian gradient requires moving the gradients in different tangent spaces "parallel" to the same tangent space.

We also have the following definition of Lipschitz smoothness on the manifolds.

Definition 2.6 (Geodesic Lipschitz smoothness)**.** *A function $h : \Omega \to \mathbb{R}$ is called geodesic-Lipschitz smooth if we have:*

$$\|\mathsf{grad} h(y) - P_{x\to y}\mathsf{grad} h(x)\| \leq \ell_h \operatorname{dist}(x,y). \tag{2.1}$$

Moreover, we have (see Zhang et al. (2016))

$$h(y) \leq h(x) + \langle \mathsf{grad} h(x), \mathsf{Exp}_x^{-1}(y)\rangle_x + \frac{\ell_h}{2}\operatorname{dist}(x,y)^2. \tag{2.2}$$

To proceed to the bilevel hypergradient estimation, we need the notion of Riemannian Hessian and Riemannian cross-derivatives (Jacobians) (see Han et al. (2023)).

Definition 2.7 (Riemannian Hessian)**.** *For function $f : \mathcal{M} \to \mathbb{R}$, the Riemannian Hessian is a symmetric 2-form $H(f) : \mathrm{T}\mathcal{M} \times \mathrm{T}\mathcal{M} \to \mathbb{R}$ defined as: $\forall \xi, \eta \in \mathrm{T}\mathcal{M}$,*

$$H(f)(\xi, \eta) = \langle \nabla_\xi \mathsf{grad} f, \eta\rangle,$$

where ∇ here is the Levi-Civita connection (see Lee (2006)). H can also be interpreted as a linear map $H(f) : \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$, $\forall \xi \in \mathrm{T}_x \mathcal{M}$,

$$H(f)(\xi) = \nabla_\xi \mathsf{grad} f.$$

Definition 2.8 (Riemannian cross-derivatives)**.** *For a smooth function defined on product manifold $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$, the Riemannian cross-derivatives is defined as a linear mapping $\mathsf{grad}^2_{x,y}(f) : \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{N}$ such that $\forall \xi \in \mathrm{T}_x \mathcal{M}$,*

$$\mathsf{grad}^2_{x,y}(f)[\xi] = \mathrm{D}_x \mathsf{grad}_y f(x,y)[\xi],$$

where D_x is the differential with respect to variable x. $\mathsf{grad}^2_{y,x}(f)$ is defined similarly.

A useful fact is that $\mathsf{grad}^2_{x,y}(f)$ and $\mathsf{grad}^2_{y,x}(f)$ are adjoint operators.

Proposition 2.1. *$\mathrm{grad}^2_{x,y}$ and $\mathrm{grad}^2_{y,x}$ are adjoints, i.e.*

$$\langle \eta, \mathrm{grad}^2_{x,y} f(x,y)[\xi]\rangle_y = \langle \mathrm{grad}^2_{y,x} f(x,y)[\eta], \xi\rangle_x, \forall \xi \in \mathrm{T}_x \mathcal{M} \text{ and } \forall \eta \in \mathrm{T}_y \mathcal{N},$$

where $f \in \mathcal{C}^1(\mathcal{M})$ is any continuously differentiable function over $\mathcal{M}$.

Proof. Note

$$\langle \eta, \mathrm{grad}^2_{x,y} f(x,y)[\xi]\rangle_y = \xi(\langle \eta, \mathrm{grad}_y f(x,y)\rangle_y) = \xi(\eta(f)),$$

and similarly

$$\langle \mathrm{grad}^2_{y,x} f(x,y)[\eta], \xi\rangle_x = \eta(\xi(f)).$$

Note that here ξ and η are actually acting on different coordinates of f. We can extend $\tilde{\xi}(x,y) = (\xi(x), 0) \in \mathrm{T}_x \mathcal{M} \times \mathrm{T}_y \mathcal{N}$ and similarly $\tilde{\eta}(x,y) = (0, \eta(y)) \in \mathrm{T}_x \mathcal{M} \times \mathrm{T}_y \mathcal{N}$. Now subtracting the above equations we have

$$\langle \eta, \mathrm{grad}^2_{x,y} f(x,y)[\xi]\rangle_y - \langle \mathrm{grad}^2_{y,x} f(x,y)[\eta], \xi\rangle_x = [\tilde{\xi}, \tilde{\eta}](f),$$

where $[\tilde{\xi}, \tilde{\eta}] = \tilde{\xi}\tilde{\eta} - \tilde{\eta}\tilde{\xi}$ is the Lie bracket. It is easy to verify in local coordinates that $[\tilde{\xi}, \tilde{\eta}]$ is zero since $\tilde{\xi}$ and $\tilde{\eta}$ act on disjoint local coordinates. □

3 Bilevel hypergradient estimation on Riemannian manifolds

We first inspect the calculation of the hypergradient for problem (1.1), namely the Riemannian gradient $\mathsf{grad}\Phi(x)$. Notice that y^* is actually a map $\mathcal{M} \to \mathcal{N}$, thus we need to follow the notion of the differential of maps between manifolds. We can calculate the Riemannian gradient $\mathsf{grad}\Phi(x)$ as follows.

Proposition 3.1. *The Riemannian gradient $\mathsf{grad}\Phi(x)$ is given by:*

$$\mathsf{grad}\Phi(x) = \mathsf{grad}_x f(x, y^*(x)) - \mathsf{grad}^2_{y,x} g(x, y^*(x))[v^*(x)], \tag{3.1}$$

where $v^*(x) \in T_{y^*(x)}\mathcal{N}$ is the solution of the following equation:

$$H_y(g(x, y^*(x)))(v) = \mathsf{grad}_y f(x, y^*(x)), \tag{3.2}$$

where H_y is the Riemannian Hessian for the y variable.

Proof. By chain rule,

$$\mathrm{d}\Phi(x) = \mathrm{d}_x f(x, y^*(x)) + \mathrm{d}_y f(x, y^*(x)) \circ (\mathrm{D}y^*(x)), \tag{3.3}$$

where d and D represent the differential operators. Notice that the above equation holds in the cotangent space. Since the Riemannian gradients are defined as $\mathsf{grad}\Phi(x) \in \mathrm{T}_x \mathcal{M}$ s.t. $\forall \xi \in \mathrm{T}_x \mathcal{M}$, $\mathrm{d}\,\Phi(x)(\xi) = \langle \mathsf{grad}\Phi(x), \xi\rangle$, we get from the above equality that

$$\langle \mathsf{grad}\Phi(x), \xi\rangle = \langle \mathsf{grad}_x f(x, y^*(x)), \xi\rangle + \langle \mathsf{grad}_y f(x, y^*(x)), \mathrm{D}y^*(x)(\xi)\rangle. \tag{3.4}$$

Now we have the following optimality condition from the y lower-level problem:

$$\mathsf{grad}_y g(x, y^*(x)) = 0.$$

By taking the differential for x on both sides of the above formula we get: $\forall \xi \in \mathrm{T}_x \mathcal{M}$,

$$\mathsf{grad}^2_{x,y} g(x, y^*(x))(\xi) + H_y(g(x, y^*(x)))(\mathrm{D}y^*(x)(\xi)) = 0. \tag{3.5}$$

Now taking the inner-product of both sides of the above equation with $v^*(x)$, we get

$$\langle v^*(x), \mathsf{grad}^2_{x,y} g(x, y^*(x))(\xi)\rangle + \langle \mathsf{grad}_y f(x, y^*(x)), \mathrm{D}y^*(x)(\xi)\rangle = 0.$$

Therefore we get the final result by plugging back the above equation to (3.4) and applying Proposition 2.1. □

When both $\mathcal{M}$ and $\mathcal{N}$ are embedded submanifolds (of two different Euclidean spaces $\mathbb{R}^M$ and $\mathbb{R}^N$), and $\bar{f} : \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}$ which restricts to $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ naturally. The Riemannian gradients of f are simply projections of the Euclidean gradients onto the tangent spaces:

$$\mathsf{grad}_x f(x, y) = \mathsf{proj}_{\mathrm{T}_x \mathcal{M}}(\nabla_x \bar{f}(x, y)),\ \mathsf{grad}_y f(x, y) = \mathsf{proj}_{\mathrm{T}_y \mathcal{N}}(\nabla_y \bar{f}(x, y)), \tag{3.6}$$

and the cross-derivatives are calculated as follows as a matrix:

$$\mathsf{grad}^2_{x,y} f(x, y) = P_y(\nabla^2_{x,y} \bar{f}(x, y))P_x, \tag{3.7}$$

where $P_x = \mathsf{proj}_{\mathrm{T}_x \mathcal{M}} \in \mathbb{R}^{M\times M}$ and $P_y = \mathsf{proj}_{\mathrm{T}_y \mathcal{N}} \in \mathbb{R}^{N\times N}$ are projection matrices onto tangent spaces, and $\nabla^2_{x,y} \bar{f}(x, y) \in \mathbb{R}^{N\times M}$ is the regular partial gradient, namely $[\nabla^2_{x,y} \bar{f}(x, y)]_{j,i} = \frac{\partial^2 \bar{f}}{\partial y_j \partial x_i}$.

In practice we cannot solve the inner minimization and (3.2) exactly. Suppose we have a point $y \in \mathcal{N}$, we can solve the approximate problem of (3.2):

$$H_y(g(x, y))[v] = \mathsf{grad}_y f(x, y) \tag{3.8}$$

with an N-step conjugate gradient method, yielding $\hat{v}^N(x, y)$, then we can estimate

$$\mathsf{grad}\Phi(x) \approx \mathsf{grad}_x f(x, y) - \mathsf{grad}^2_{y,x} g(x, y)[\hat{v}^N(x, y)], \tag{3.9}$$

which we further refer to as the approximate implicit differentiation (AID) estimate of (3.1). For the rest of the paper we denote $h_g^{k,t} := \mathsf{grad}_y g(x^k, y^{k,t})$ and

$$h_\Phi(x, y) := \mathsf{grad}_x f(x, y) - \mathsf{grad}^2_{y,x} g(x, y)[\hat{v}^N(x, y)]. \tag{3.10}$$

We abbreviate the notation by $h^k_\Phi := h_\Phi(x^k, y^k)$ in the algorithms.

For the stochastic problem (1.2), we have an estimate of the gradient of the stochastic function $G(x, y; \zeta)$ described as follows (see Hong et al. (2020); Chen et al. (2021b)). We first update the inner problem $y^t \leftarrow \mathsf{Exp}_{y^{t-1}}(-\alpha \mathsf{grad}_y G(x, y^{t-1}; \zeta^{t-1}))$ for $t = 1, ..., T$. Meanwhile the estimate for the gradient of F and the second-order gradient of G will require us to further have independent samples ξ and $\zeta_{(q)}$, $q = 0, 1, ..., Q$, so that we define the stochastic gradient estimator as:

$$\mathsf{grad}\Phi\,(x) \approx \mathsf{grad}_x F(x, y^T; \xi) - \mathsf{grad}^2_{y,x} G(x, y^T; \zeta_0)[v_Q(x, y^T)], \tag{3.11}$$

where v_Q is the approximation of (3.2), defined as (see (Hong et al., 2020, Lemma 1)):

$$v_Q(x, y) := \eta Q \prod_{q=1}^{Q'} (I - \eta H_y(G(x, y; \zeta_{(q)})))[\mathsf{grad}_y F\,(x, y; \xi)], \tag{3.12}$$

where Q' is drawn uniformly from $\{0, 1, ..., Q-1\}$ and the extra parameter η will be later determined to ensure a better approximation to (3.2), motivated by the Neumann series $\sum_{i=0}^{\infty} U^i = (I-U)^{-1}$.

From now on we denote $\tilde{h}_g^{k,t} = \mathsf{grad}_y G(x^k, y^{k,t}; \zeta_{k,t})$ and

$$\tilde{h}_\Phi^k := \mathsf{grad}_x F(x^k, y^k; \xi_k) - \mathsf{grad}^2_{y,x} G(x^k, y^k; \zeta_{k,(0)})[v_Q^k], \tag{3.13}$$

where

$$v_Q^k := \eta Q \prod_{q=1}^{Q'} (I - \eta H_y(G(x^k, y^k; \zeta_{k,(q)})))[\mathsf{grad}_y F(x^k, y^k; \xi_k)].$$

4 Deterministic Algorithm RieBO and Its Convergence

We propose RieBO (Algorithm 1) for the deterministic bilevel manifold optimization (1.1). The algorithm is a generalization of its Euclidean counterpart proposed in Ji et al. (2021) where we employ conjugate gradient method to solve the hypergradient estimation problem (3.10).

For the deterministic case, we utilize the following notion of stationarity:

Definition 4.1. *A point $x \in \mathcal{M}$ is called an ϵ-stationary point for (1.1) if $\|\nabla\Phi(x)\|^2 \le \epsilon$.*

We need the following assumptions for the convergence analysis.

Assumption 4.1. *The manifolds $\mathcal{M}$ and $\mathcal{N}$ are complete Riemannian manifolds. Moreover, $\mathcal{N}$ is a Hadamard manifold whose sectional curvature is lower bounded by $\iota < 0$ [1].*

We use the notation $\langle \cdot, \cdot \rangle_x$ and $\langle \cdot, \cdot \rangle_y$ to represent their Riemannian metrics, for $x \in \mathcal{M}$ and $y \in \mathcal{N}$. The corresponding norms are $\|\cdot\|_x$ and $\|\cdot\|_y$. Note that from now on we may omit the subscript since the corresponding manifold and tangent space can be identified by the vector in the "$\cdot$" position.

Following Zhang and Sra (2016), it is very important that the quantity $\tau(\iota, \operatorname{dist}(y^{k,t}, y^*(x^k)))$ is bounded during the lower level update, where

$$\tau(\iota, c) := \frac{\sqrt{|\iota| c}}{\tanh(\sqrt{|\iota| c})}. \tag{4.1}$$

Therefore we need the following assumption.

Assumption 4.2. *The lower level objective function $g(x, y)$ is μ-geodesically strongly convex with respect to y. Note that the total objective $\Phi(x) = f(x, y^*(x))$ may still be nonconvex. Moreover, we assume that the quantity $\tau(\iota, \operatorname{dist}(y^{k,t}, y^*(x^k)))$ is always upper bounded by τ for all k and t [2].*

Assumption 4.3 (Smoothness). *For simplicity denote $z = (x, y)$ and $z' = (x', y')$, also denote $\operatorname{dist}(z, z') = \sqrt{\operatorname{dist}(x, x')^2 + \operatorname{dist}(y, y')^2}$ (note that these two distances are on different manifolds). Moreover, f and g satisfy the following assumptions.*

- *f satisfies $\ell_{f,0}$-Lipschitzness:*

$$f(z) - f(z') \le \ell_{f,0} \operatorname{dist}(z, z').$$

[1] We make this assumption so that the lower function g can be geodesically strongly convex, see Zhang and Sra (2016).

[2] Note that this assumption is satisfied if the lower level problem is conducted in a compact subset in $\mathcal{N}$ and assuming that the iterates of the algorithm stay in this compact region, see Zhang and Sra (2016).

- $\mathsf{grad} f = [\mathsf{grad}_x f, \mathsf{grad}_y f]$ *and* $\mathsf{grad} g = [\mathsf{grad}_x g, \mathsf{grad}_y g]$ *are* $\ell_{f,1}$ *and* $\ell_{g,1}$*Lipschitz, i.e.,*

$$\|\mathsf{grad} f(z) - P_{z'\to z}\mathsf{grad} f(z')\| \le \ell_{f,1}\,\mathrm{dist}(z,z')$$
$$\|\mathsf{grad} g(z) - P_{z'\to z}\mathsf{grad} g(z')\| \le \ell_{g,1}\,\mathrm{dist}(z,z'),$$

where $P_{z'\to z}$ *is the parallel transport on the product manifold* $\mathcal{M}\times\mathcal{N}$ *and the norm on the left hand side is also induced by the product Riemannian metric on* $\mathcal{M}\times\mathcal{N}$ [3].

Assumption 4.4 (Hessian Smoothness)**.** *The second-order derivatives* $\mathsf{grad}^2_{x,y} g(z)$ *and* $H_y(g(z))$ *are* $\ell_{g,2}$ *Lipschitz (we use the same constant here for simplicity), i.e. for* $z=(x,y)$ *and* $z'=(x',y')$*, we have*

$$\left\|\mathsf{grad}^2_{x,y} g(z) - P_{y^*(x')\to y^*(x)}\circ \mathsf{grad}^2_{x,y} g(z')\circ P_{x'\to x}\right\|_{\mathrm{op}} \le \ell_{g,2}\,\mathrm{dist}(z,z')$$
$$\left\|H_y(g(z)) - P_{y^*(x')\to y^*(x)}\circ H_y(g(z'))\circ P_{y^*(x)\to y^*(x')}\right\|_{\mathrm{op}} \le \ell_{g,2}\,\mathrm{dist}(z,z').$$

Here the norms on the left hand side are the operator norms. We will keep the subscript op *whenever it comes to the operator norm, in order to distinguish it from the norms on the tangent spaces.*

We have the following smoothness lemma under the above assumptions.

Lemma 4.1. *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4 hold, then functions* $y^*(x)$ *and* $\Phi(x) := f(x, y^*(x))$ *satisfy:* $\forall x, x' \in \mathcal{M}$

$$\mathrm{dist}(y^*(x), y^*(x')) \le \kappa\,\mathrm{dist}(x,x'),\ \ \kappa = \frac{\ell_{g,1}}{\mu} \tag{4.2a}$$
$$\|\mathrm{D}y^*(x) - P_{y^*(x')\to y^*(x)}\circ \mathrm{D}y^*(x')\circ P_{x'\to x}\|_{\mathrm{op}} \le L_{y^*}\,\mathrm{dist}(x,x') \tag{4.2b}$$
$$\|\mathsf{grad}\Phi(x) - P_{x'\to x}\mathsf{grad}\Phi(x')\| \le L_\Phi\,\mathrm{dist}(x,x'), \tag{4.2c}$$

where

$$L_{y^*} := \left(1+\frac{\ell_{g,2}}{\mu}\right)\frac{\ell_{g,2}}{\mu}\sqrt{1+\kappa^2} = \mathcal{O}(\kappa^2), \tag{4.3}$$

and

$$L_\Phi := \ell_{f,1}\sqrt{1+\kappa^2} + \ell_{g,2}\frac{\ell_{f,0}}{\mu} + \ell_{g,1}\left(\frac{\ell_{f,0}\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2}+\frac{\ell_{f,1}}{\mu}\right) = \mathcal{O}(\kappa^3). \tag{4.4}$$

Proof. For (4.2a), by (3.5) and our Assumptions 4.2, 4.3 and 4.4, we have

$$\|\mathrm{D}y^*(x)\|_{\mathrm{op}} = \|(H_y(g(x,y^*(x))))^{-1}\circ \mathsf{grad}^2_{x,y} g(x,y^*(x))\|_{\mathrm{op}} \le \frac{\ell_{g,1}}{\mu}.$$

We thus obtain (4.2a) by a mean value theorem argument in local coordinates (see this link for a detailed proof).

[3] It is worth noticing that in our convergence result, we need $\tau < \ell_{g,1}/2$. We comment here that this assumption is not a big issue since if g is Lipschitz smooth with parameter $\ell_{g,1}$, then it is also Lipschitz smooth with any parameters $\ell > \ell_{g,1}$. We could always pick up a parameter that satisfies $\tau < \ell_{g,1}/2$, with some sacrifice of the convergence speed. Note that in the Euclidean case $\tau = 0$ so $\tau < \ell_{g,1}/2$ holds naturally.

For (4.2b), we have

$$\begin{aligned}
&\|\mathrm{D}y^*(x) - P_{y^*(x')\to y^*(x)} \circ \mathrm{D}y^*(x') \circ P_{x'\to x}\| \\
=&\|(H_y(g(x,y^*(x))))^{-1} \circ \mathsf{grad}^2_{x,y} g(x,y^*(x)) - P_{y^*(x')\to y^*(x)} \circ (H_y(g(x',y^*(x'))))^{-1} \circ \mathsf{grad}^2_{x,y} g(x',y^*(x')) \circ P_{x'\to x}\| \\
\leq&\|(H_y(g(x,y^*(x))))^{-1} \circ \mathsf{grad}^2_{x,y} g(x,y^*(x)) \\
&\qquad\qquad - (H_y(g(x,y^*(x))))^{-1} \circ P_{y^*(x')\to y^*(x)} \circ \mathsf{grad}^2_{x,y} g(x',y^*(x')) \circ P_{x'\to x}\| \\
&+ \|(H_y(g(x,y^*(x))))^{-1} \circ P_{y^*(x')\to y^*(x)} \circ \mathsf{grad}^2_{x,y} g(x',y^*(x')) \\
&\qquad\qquad - P_{y^*(x')\to y^*(x)} \circ (H_y(g(x',y^*(x'))))^{-1} \circ \mathsf{grad}^2_{x,y} g(x',y^*(x'))\| \\
\leq&\|(H_y(g(x,y^*(x))))^{-1}\|\|\mathsf{grad}^2_{x,y} g(x,y^*(x)) - P_{y^*(x')\to y^*(x)} \circ \mathsf{grad}^2_{x,y} g(x',y^*(x')) \circ P_{x'\to x}\| \\
&+ \|(H_y(g(x,y^*(x))))^{-1} - P_{y^*(x')\to y^*(x)} \circ (H_y(g(x',y^*(x'))))^{-1} \circ P_{y^*(x)\to y^*(x')}\|\|\mathsf{grad}^2_{x,y} g(x',y^*(x'))\| \\
\leq&\frac{\ell_{g,2}}{\mu}\sqrt{\mathrm{dist}(x,x')^2 + \mathrm{dist}(y^*(x),y^*(x'))^2} \\
&+ \ell_{g,1}\|(H_y(g(x,y^*(x))))^{-1} - P_{y^*(x')\to y^*(x)} \circ (H_y(g(x',y^*(x'))))^{-1} \circ P_{y^*(x)\to y^*(x')}\| \\
\leq&\frac{\ell_{g,2}}{\mu}\sqrt{1+\kappa^2}\,\mathrm{dist}(x,x') + \ell_{g,1}\|(H_y(g(x,y^*(x))))^{-1} - P_{y^*(x')\to y^*(x)} \circ (H_y(g(x',y^*(x'))))^{-1} \circ P_{y^*(x)\to y^*(x')}\|_{\mathrm{op}}
\end{aligned}$$

where in the second last inequality we used Assumptions 4.2, 4.3 and 4.4, and we used (4.2a) for the last inequality. Denote $H_1 = H_y(g(x,y^*(x)))$, $P = P_{y^*(x')\to y^*(x)}$ so that $P_{y^*(x)\to y^*(x')} = P^{-1}$ and $H_2 = H_y(g(x',y^*(x')))$, then the last term in the above formula becomes:

$$\begin{aligned}
&\|H_1^{-1} - PH_2^{-1}P\|_{\mathrm{op}} = \|H_1^{-1}P^{-1}(H_2 - P^{-1}H_1P)H_2^{-1}P^{-1}\|_{\mathrm{op}} \\
\leq&\frac{1}{\mu^2}\|H_2 - P^{-1}H_1P\|_{\mathrm{op}} \leq \frac{\ell_{g,2}}{\mu^2}\sqrt{\mathrm{dist}(x,x')^2 + \mathrm{dist}(y^*(x),y^*(x'))^2} \\
\leq&\frac{\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2}\,\mathrm{dist}(x,x').
\end{aligned} \tag{4.5}$$

Here we used Assumptions 4.2, 4.4, (4.2a) and the fact that the parallel transport $P = P_{y^*(x')\to y^*(x)}$ is an isometry, i.e., $\|P\|_{\mathrm{op}} = 1$. Plugging this back we get

$$\|\mathrm{D}y^*(x) - P_{y^*(x')\to y^*(x)} \circ \mathrm{D}y^*(x') \circ P_{x'\to x}\|_{\mathrm{op}} \leq \left(\frac{\ell_{g,2}}{\mu} + \frac{\ell_{g,1}\ell_{g,2}}{\mu^2}\right)\sqrt{1+\kappa^2}\,\mathrm{dist}(x,x'),$$

which gives (4.2b).

We now show (4.2c). Using (3.1), we get

$$\begin{aligned}
&\|\mathsf{grad}\Phi(x) - P_{x'\to x}\mathsf{grad}\Phi(x')\| \leq \|\mathsf{grad}_x f(x,y^*(x)) - P_{x'\to x}\mathsf{grad}_x f(x',y^*(x'))\| \\
&+ \|\mathsf{grad}^2_{y,x} g(x,y^*(x))[v^*(x)] - P_{x'\to x}\mathsf{grad}^2_{y,x} g(x',y^*(x'))[v^*(x')]\|.
\end{aligned} \tag{4.6}$$

For the first term on the right hand side of (4.6), by Assumption 4.3 and (4.2a), we have

$$\begin{aligned}
&\|\mathsf{grad}_x f(x,y^*(x)) - P_{x'\to x}\mathsf{grad}_x f(x',y^*(x'))\| \\
\leq&\ell_{f,1}\sqrt{\mathrm{dist}(x,x')^2 + \mathrm{dist}(y^*(x),y^*(x'))^2} \leq \ell_{f,1}\sqrt{1+\kappa^2}\,\mathrm{dist}(x,x').
\end{aligned}$$

For the second term on the right hand side of (4.6), we have

$$
\begin{aligned}
&\|\mathsf{grad}^2_{y,x} g(x, y^*(x))[v^*(x)] - P_{x' \to x}\mathsf{grad}^2_{y,x} g(x', y^*(x'))[v^*(x')]\| \\
=&\|\mathsf{grad}^2_{y,x} g(x, y^*(x))[v^*(x)] - P_{x' \to x}\mathsf{grad}^2_{y,x} g(x', y^*(x'))\Big[P_{x \to x'}P_{x' \to x}[v^*(x')]\Big]\| \\
\leq&\|\mathsf{grad}^2_{y,x} g(x, y^*(x))[v^*(x)] - P_{x' \to x}\mathsf{grad}^2_{y,x} g(x', y^*(x'))P_{x \to x'}[v^*(x)]\| \\
&+ \|P_{x' \to x}\mathsf{grad}^2_{y,x} g(x', y^*(x'))P_{x \to x'}[v^*(x)] - P_{x' \to x}\mathsf{grad}^2_{y,x} g(x', y^*(x'))P_{x \to x'}P_{x' \to x}[v^*(x')]\| \\
\leq&\|\mathsf{grad}^2_{y,x} g(x, y^*(x)) - P_{x' \to x}\mathsf{grad}^2_{y,x} g(x', y^*(x'))P_{x \to x'}\|_{\mathrm{op}}\|v^*(x)\| \\
&+ \|\mathsf{grad}^2_{y,x} g(x', y^*(x'))\|_{\mathrm{op}}\|v^*(x) - P_{x' \to x}v^*(x')\|.
\end{aligned}
$$

Since $v^*(x)$ is the solution of (3.2), also by Assumptions 4.2 and 4.3, we have $\|v^*(x)\| \leq \ell_{f,0}/\mu$, also

$$
\begin{aligned}
&\|v^*(x) - P_{x' \to x}v^*(x')\| \\
=&\|(H_y(g(x, y^*(x))))^{-1}\mathsf{grad}_y f(x, y^*(x)) - P_{x' \to x}(H_y(g(x', y^*(x'))))^{-1}\mathsf{grad}_y f(x', y^*(x'))\| \\
=&\|(H_y(g(x, y^*(x))))^{-1}\mathsf{grad}_y f(x, y^*(x)) - P_{x' \to x}(H_y(g(x', y^*(x'))))^{-1}P_{x \to x'}P_{x' \to x}\mathsf{grad}_y f(x', y^*(x'))\| \\
\leq&\|(H_y(g(x, y^*(x))))^{-1} - P_{x' \to x}(H_y(g(x', y^*(x'))))^{-1}P_{x \to x'}\|_{\mathrm{op}}\|\mathsf{grad}_y f(x, y^*(x))\| \\
&+ \|(H_y(g(x', y^*(x'))))^{-1}\|_{\mathrm{op}}\|\mathsf{grad}_y f(x, y^*(x)) - P_{x' \to x}\mathsf{grad}_y f(x', y^*(x'))\| \\
\leq&\left(\frac{\ell_{f,0}\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2} + \frac{\ell_{f,1}}{\mu}\right)\mathrm{dist}(x, x'),
\end{aligned}
\tag{4.7}
$$

where we used (4.5), Assumptions 4.2 and 4.3.

Combining the above bounds and plugging it to (4.6), we get

$$
\begin{aligned}
&\|\mathsf{grad}\Phi(x) - P_{x' \to x}\mathsf{grad}\Phi(x')\| \\
\leq&\ell_{f,1}\sqrt{1+\kappa^2}\,\mathrm{dist}(x, x') + \ell_{g,2}\frac{\ell_{f,0}}{\mu}\,\mathrm{dist}(x, x') + \ell_{g,1}\left(\frac{\ell_{f,0}\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2} + \frac{\ell_{f,1}}{\mu}\right)\mathrm{dist}(x, x'),
\end{aligned}
$$

which proves (4.2c). □

Algorithm 1: Algorithm for **Rie**mannian (deterministic) **B**ilevel **O**ptimization (**RieBO**)

```
input  : $K$, $T$, $N$(steps for conjugate gradient), stepsize $\{\alpha_k, \beta_k\}$, initializations
         $x^0 \in \mathcal{M}, y^0 \in \mathcal{N}$
for $k = 0, 1, 2, ..., K-1$ do
    Set $y^{k,0} = y^{k-1}$;
    for $t = 0, ..., T-1$ do
        Update $y^{k,t+1} \leftarrow \mathsf{Exp}_{y^{k,t}}(-\beta_k h_g^{k,t})$ with $h_g^{k,t} := \mathsf{grad}_y g(x^k, y^{k,t})$ ;
    end
    Set $y^k \leftarrow y^{k,T}$;
    Update $x^{k+1} \leftarrow \mathsf{Exp}_{x^k}(-\alpha_k h_\Phi^k)$ as in (3.10), where $\hat{v}^N(x^k, y^k)$ is given by an $N$-step
     conjugate gradient update, with $\hat{v}^0(x^k, y^k) = P_{y^{k-1} \to y^k}\hat{v}^N(x^{k-1}, y^{k-1})$;
end
```

Theorem 4.1. *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4 hold, and take the parameters* $\beta_k = \beta \leq \frac{1}{\ell_{g,1}}$, $\alpha_k = \alpha \leq \frac{1}{8L_\Phi}$, $T \geq \mathcal{O}(\kappa)$ *and conjugate gradient iteration number* $N \geq \mathcal{O}(\sqrt{\kappa})$*. Then RieBO (Algorithm 1) satisfies:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\mathsf{grad}\Phi(x^k)\|^2 \leq \mathcal{O}\left(\frac{L_\Phi}{K}\right). \tag{4.8}$$

The specific choice parameters are given in the proof for the simplicity of the statement. In order to achieve an ϵ*-accurate stationary point, the complexity is given by:*

- *Gradients:* $\mathrm{Gc}(f,\epsilon) = \mathcal{O}(\kappa^3\epsilon^{-1})$, $\mathrm{Gc}(g,\epsilon) = \mathcal{O}(\kappa^4\epsilon^{-1})$*;*
- *Jacobian and Hessian-vector products:* $\mathrm{JV}(g,\epsilon) = \mathcal{O}(\kappa^3\epsilon^{-1})$, $\mathrm{HV}(g,\epsilon) = \mathcal{O}(\kappa^{3.5}\epsilon^{-1})$.

To prove this theorem, we need the following lemmas. The first lemma quantifies the error when optimizing (3.8) with N-step conjugate gradient method, see Grazzi et al. (2020, Equation (17))[4].

Lemma 4.2. *Suppose we solve (3.8) with* N*-step conjugate gradient method with the initial point* $\hat{v}^0(x,y)$ *and output* $\hat{v}^N(x,y)$*, then we have*

$$\|\hat{v}^N(x,y) - \tilde{v}\| \leq \sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N \|\hat{v}^0(x,y) - \tilde{v}\|,$$

where $\tilde{v}$ *is the exact solution of (3.8).*

The next lemma quantifies the error of the inner loop, i.e. the T steps where we do Riemannian gradient descent for the lower problem in RieBO (Algorithm 1).

Lemma 4.3. *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4 hold, and we take* $\beta_k = \beta = 1/\ell_{g,1}$ *as a constant, then RieBO satisfies:*

$$\mathrm{dist}(y^{k,T}, y^*(x^k))^2 \leq (1 - 2\mu\tau\beta^2)^T\,\mathrm{dist}(y^{k,0}, y^*(x^k))^2. \tag{4.9}$$

Proof. For simplicity, we denote $h(y) = g(x^k, y)$, so that $y^*(x^k)$ is the optimal solution of h. We also omit k in this proof, i.e., the update becomes:

$$y^{t+1} \leftarrow \mathsf{Exp}_{y^t}(-\beta_k \mathsf{grad}h(y^t)).$$

By the notion of geodesic smoothness and geodesic convexity, we have

$$\begin{aligned}
&h(y^{t+1}) - h(y^*) = h(y^{t+1}) - h(y^t) + h(y^t) - h(y^*)\\
\leq&\langle \mathsf{grad}h(y^t), \mathsf{Exp}_{y^t}(y^{t+1})\rangle + \frac{\ell_{g,1}}{2}\|\mathsf{Exp}_{y^t}(y^{t+1})\|^2 - \langle \mathsf{grad}h(y^t), \mathsf{Exp}_{y^t}(y^*)\rangle - \frac{\mu}{2}\,\mathrm{dist}(y^t, y^*)^2\\
=&-(\beta - \frac{\beta^2\ell_{g,1}}{2})\|\mathsf{grad}h(y^t)\|^2 - \langle \mathsf{grad}h(y^t), \mathsf{Exp}_{y^t}(y^*)\rangle - \frac{\mu}{2}\,\mathrm{dist}(y^t, y^*)^2,
\end{aligned}$$

i.e.,

$$(\beta - \frac{\beta^2\ell_{g,1}}{2})\|\mathsf{grad}h(y^t)\|^2 - \langle \mathsf{grad}h(y^t), \mathsf{Exp}_{y^t}(y^*)\rangle \leq -\frac{\mu}{2}\,\mathrm{dist}(y^t, y^*)^2. \tag{4.10}$$

[4] Note that here the Hessian matrix $H_y(g(x,y))$ is full-rank in the tangent space. However if it is an embedded submanifold then $H_y(g(x,y))$ is actually rank-deficient matrix in the ambient Euclidean space. This is not a concern for showing the linear rate of convergence since we can always conduct CG steps only on the tangent spaces (as Euclidean subspaces of the ambient Euclidean space). It is known that the convergence is still linear even if $H_y(g(x,y))$ is rank-deficient, see Hayami (2018) for a detailed inspection.

Now by Zhang and Sra (2016, Corollary 8), we have,

$$\begin{aligned}\operatorname{dist}(y^{t+1},y^*)^2 \leq& \operatorname{dist}(y^t,y^*)^2 + 2\beta\langle \mathsf{grad}h(y^t), \mathsf{Exp}_{y^t}(y^*)\rangle + \tau\beta^2\|\mathsf{grad}h(y^t)\|^2\\ \leq&(1-2\mu\tau\beta^2)\operatorname{dist}(y^t,y^*)^2,\end{aligned}$$

where the last inequality is by (4.10) and $\beta = 1/\ell_{g,1}$. The proof is done by repeatedly applying the above inequality from $t = T-1$ back to $t = 0$. □

The next lemma quantifies the error between our estimation h_Φ^k and the true upper level gradient $\mathsf{grad}\Phi(x^k)$.

Lemma 4.4. *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4 hold, then RieBO satisfies:*

$$\begin{aligned}\|h_\Phi^k - \mathsf{grad}\Phi(x^k)\| \leq& \Gamma(1-2\mu\tau\beta^2)^{T/2}\operatorname{dist}(y^*(x^k),y^{k-1})\\ &+\ell_{g,1}\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N\|\hat{v}^0(x^k,y^k)-P_{y^*(x^k)\to y^k}v^*(x^k)\|,\end{aligned}\tag{4.11}$$

where h_Φ^k is the estimate from (3.10) and we have the parameters:

$$\begin{aligned}&\tilde{v}^k = (H_y(g(x^k,y^k)))^{-1}\mathsf{grad}_y f(x^k,y^k)\\ &\Gamma = \ell_{f,1}+\frac{\ell_{f,0}\ell_{g,2}}{\mu}+\ell_{g,1}\left(1+\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N\right)\left(\ell_{g,2}\ell_{f,0}+\frac{\ell_{f,1}}{\mu}\right).\end{aligned}\tag{4.12}$$

Proof. We first restate the expression (3.1) for $\mathsf{grad}\Phi(x)$ and (3.10) for h_Φ^k:

$$\begin{aligned}&\mathsf{grad}\Phi(x^k) = \mathsf{grad}_x f(x^k,y^*(x^k)) - \mathsf{grad}^2_{y,x}g(x^k,y^*(x^k))[v^*(x^k)],\\ &h_\Phi(x^k,y^k) := \mathsf{grad}_x f(x^k,y^k) - \mathsf{grad}^2_{y,x}g(x^k,y^k)[\hat{v}^N(x^k,y^k)].\end{aligned}$$

Thus,

$$\begin{aligned}&\|h_\Phi^k - \mathsf{grad}\Phi(x^k)\| \leq \|\mathsf{grad}_x f(x^k,y^*(x^k)) - \mathsf{grad}_x f(x^k,y^k)\|\\ &\quad+\|\mathsf{grad}^2_{y,x}g(x^k,y^*(x^k))[v^*(x^k)] - \mathsf{grad}^2_{y,x}g(x^k,y^k)[\hat{v}^N(x^k,y^k)]\|\\ \leq&\|\mathsf{grad}_x f(x^k,y^*(x^k)) - \mathsf{grad}_x f(x^k,y^k)\|\\ &\quad+\|\mathsf{grad}^2_{y,x}g(x^k,y^*(x^k))[v^*(x^k)] - \mathsf{grad}^2_{y,x}g(x^k,y^k)[P_{y^*(x^k)\to y^k}v^*(x^k)]\|\\ &\quad+\|\mathsf{grad}^2_{y,x}g(x^k,y^k)[P_{y^*(x^k)\to y^k}v^*(x^k)] - \mathsf{grad}^2_{y,x}g(x^k,y^k)[\hat{v}^N(x^k,y^k)]\|\\ \leq&\ell_{f,1}\operatorname{dist}(y^*(x^k),y^k)\\ &\quad+\|\mathsf{grad}^2_{y,x}g(x^k,y^*(x^k)) - \mathsf{grad}^2_{y,x}g(x^k,y^k)\circ P_{y^*(x^k)\to y^k}\|_{\mathrm{op}}\|v^*(x^k)\|\\ &\quad+\|\mathsf{grad}^2_{y,x}g(x^k,y^k)\|_{\mathrm{op}}\|P_{y^*(x^k)\to y^k}v^*(x^k) - \hat{v}^N(x^k,y^k)\|\\ \leq&\left(\ell_{f,1}+\frac{\ell_{f,0}\ell_{g,2}}{\mu}\right)\operatorname{dist}(y^*(x^k),y^k)+\ell_{g,1}\|P_{y^*(x^k)\to y^k}v^*(x^k)-\hat{v}^N(x^k,y^k)\|.\end{aligned}$$

Following Lemma 4.2, we have

$$\begin{aligned}&\|P_{y^*(x^k)\to y^k}v^*(x^k)-\hat{v}^N(x^k,y^k)\| \leq \|P_{y^*(x^k)\to y^k}v^*(x^k)-\tilde{v}^k\| + \|\tilde{v}^k - \hat{v}^N(x^k,y^k)\|\\ \leq&\|P_{y^*(x^k)\to y^k}v^*(x^k)-\tilde{v}^k\| + \sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N\|\hat{v}^0(x^k,y^k)-\tilde{v}^k\|\\ \leq&\left(1+\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N\right)\|P_{y^*(x^k)\to y^k}v^*(x^k)-\tilde{v}^k\| + \sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N\|\hat{v}^0(x^k,y^k)-P_{y^*(x^k)\to y^k}v^*(x^k)\|.\end{aligned}$$

For $\|P_{y^*(x^k)\to y^k} v^*(x^k) - \tilde{v}^k\|$, by the definitions of $\tilde{v}_k$ and v_k^*, we have

$$\begin{aligned}
&\|P_{y^*(x^k)\to y^k} v^*(x^k) - \tilde{v}^k\| \\
=&\|P_{y^*(x^k)\to y^k}(H_y(g(x^k,y^*(x^k))))^{-1}\mathsf{grad}_y f(x^k,y^*(x^k)) - (H_y(g(x^k,y^k)))^{-1}\mathsf{grad}_y f(x^k,y^k)\| \\
\le&\|P_{y^*(x^k)\to y^k}(H_y(g(x^k,y^*(x^k))))^{-1}\mathsf{grad}_y f(x^k,y^*(x^k)) - (H_y(g(x^k,y^k)))^{-1}P_{y^*(x^k)\to y^k}\mathsf{grad}_y f(x^k,y^*(x^k))\| \\
&+ \|(H_y(g(x^k,y^k)))^{-1}P_{y^*(x^k)\to y^k}\mathsf{grad}_y f(x^k,y^*(x^k)) - (H_y(g(x^k,y^k)))^{-1}\mathsf{grad}_y f(x^k,y^k)\| \\
\le&\|(H_y(g(x^k,y^*(x^k))))^{-1} - P_{y^k\to y^*(x^k)}(H_y(g(x^k,y^k)))^{-1}P_{y^*(x^k)\to y^k}\|_{\mathrm{op}}\|\mathsf{grad}_y f(x^k,y^*(x^k))\| \\
&+ \|(H_y(g(x^k,y^k)))^{-1}\|_{\mathrm{op}}\|P_{y^*(x^k)\to y^k}\mathsf{grad}_y f(x^k,y^*(x^k)) - \mathsf{grad}_y f(x^k,y^k)\| \\
\le&\left(\ell_{g,2}\ell_{f,0} + \frac{\ell_{f,1}}{\mu}\right)\mathrm{dist}(y^k,y^*(x^k)).
\end{aligned} \tag{4.13}$$

Therefore, we get

$$\begin{aligned}
\|h_\Phi^k - \mathsf{grad}\Phi(x^k)\| \le& \left(\ell_{f,1} + \frac{\ell_{f,0}\ell_{g,2}}{\mu}\right)\mathrm{dist}(y^*(x^k),y^k) \\
&+ \ell_{g,1}\left(1+\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N\right)\|P_{y^*(x^k)\to y^k}v^*(x^k) - \tilde{v}^k\| \\
&+ \ell_{g,1}\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N \|\hat{v}^0(x^k,y^k) - P_{y^*(x^k)\to y^k}v^*(x^k)\| \\
\le&\left(\ell_{f,1} + \frac{\ell_{f,0}\ell_{g,2}}{\mu} + \ell_{g,1}\left(1+\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N\right)\left(\ell_{g,2}\ell_{f,0}+\frac{\ell_{f,1}}{\mu}\right)\right)\mathrm{dist}(y^*(x^k),y^k) \\
&+ \ell_{g,1}\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^N \|\hat{v}^0(x^k,y^k) - P_{y^*(x^k)\to y^k}v^*(x^k)\|.
\end{aligned}$$

We obtain the desired result by applying Lemma 4.3 to the above inequality. □

Lemma 4.5. *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4 hold, then RieBO satisfies:*

$$\mathrm{dist}(y^{k,0},y^*(x^k))^2 + \|P_{y^*(x^k)\to y^k}v^*(x^k) - \hat{v}^0(x^k,y^k)\|^2 \le \left(\frac{1}{2}\right)^k \Delta_0 + \Omega\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\left\|\mathsf{grad}\Phi\left(x^j\right)\right\|^2, \tag{4.14}$$

with the following choice of parameters:

$$\begin{aligned}
&T \ge \log\left(2\Big(7+8\kappa^2\alpha^2\Gamma^2\Big)\left(\ell_{g,2}\ell_{f,0}+\frac{\ell_{f,1}}{\mu}\right)^2\right)/(2\log\left(\frac{1}{1-2\mu\tau\beta^2}\right)) = \Theta(\kappa), \\
&N \ge \log\left((4+16\kappa^2\alpha^2\ell_{g,1}^2)\kappa\right)/(2\log\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)) = \Theta(\sqrt{\kappa}), \\
&\Omega = \left[2\Big(\frac{\ell_{f,0}\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2}+\frac{\ell_{f,1}}{\mu}\Big)^2 + 4\kappa^2\right]\alpha^2, \\
&\Delta_0 = \mathrm{dist}(y^{0,0},y^*(x^0))^2 + \|P_{y^*(x^0)\to y^0}v^*(x^0) - \hat{v}^0(x^0,y^0)\|^2.
\end{aligned} \tag{4.15}$$

Proof. Since $y^{k,0} = y^{k-1,T}$, we have

$$\operatorname{dist}(y^{k,0}, y^*(x^k))^2 \leq 2\operatorname{dist}(y^{k-1,T}, y^*(x^{k-1}))^2 + 2\operatorname{dist}(y^*(x^{k-1}), y^*(x^k))^2.$$

Here the first term is again bounded by $(1-2\mu\tau\beta^2)^T \operatorname{dist}(y^{k-1,0}, y^*(x^{k-1}))^2$ by Lemma 4.3, and the second term is bounded by the Lipschitzness of y^* (Lemma 4.1) and by the update in the following way:

$$\operatorname{dist}(y^*(x^{k-1}), y^*(x^k))^2 \leq \kappa^2 \operatorname{dist}(x^{k-1}, x^k)^2 = \kappa^2\alpha^2 \|h_\Phi^{k-1}\|^2.$$

Thus,

$$\begin{aligned}
&\operatorname{dist}(y^{k,0}, y^*(x^k))^2 \leq 2\operatorname{dist}(y^{k-1,T}, y^*(x^{k-1}))^2 + 2\operatorname{dist}(y^*(x^{k-1}), y^*(x^k))^2\\
\leq& 2(1-2\mu\tau\beta^2)^T \operatorname{dist}(y^{k-1,0}, y^*(x^{k-1}))^2 + 2\kappa^2\alpha^2\|h_\Phi^{k-1}\|^2\\
\leq& 2(1-2\mu\tau\beta^2)^T \operatorname{dist}(y^{k-1,0}, y^*(x^{k-1}))^2 + 4\kappa^2\alpha^2\|h_\Phi^{k-1} - \mathsf{grad}\Phi(x^{k-1})\|^2 + 4\kappa^2\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2\\
\leq& \left(2 + 8\kappa^2\alpha^2\Gamma^2\right)(1-2\mu\tau\beta^2)^T \operatorname{dist}(y^*(x^{k-1}), y^{k-1,0})^2\\
&+ 8\kappa^2\alpha^2\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N} \|\hat{v}^0(x^{k-1}, y^{k-1}) - \tilde{v}^{k-1}\|^2 + 4\kappa^2\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2\\
\leq& \left(2 + 8\kappa^2\alpha^2\Gamma^2\right)(1-2\mu\tau\beta^2)^T \operatorname{dist}(y^*(x^{k-1}), y^{k-1,0})^2 + 4\kappa^2\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2\\
&+ 16\kappa^2\alpha^2\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N} \|\hat{v}^0(x^{k-1}, y^{k-1}) - P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1})\|^2\\
&+ 16\kappa^2\alpha^2\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N} \|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \tilde{v}^{k-1}\|^2,
\end{aligned}$$

where the third inequality is by Lemma 4.4. For the last term, by (4.13) we have

$$\|P_{y^*(x^{k-1})\to y^k}v^*(x^{k-1}) - \tilde{v}^{k-1}\|^2 \leq \left(\ell_{g,2}\ell_{f,0} + \frac{\ell_{f,1}}{\mu}\right)^2 \operatorname{dist}(y^{k-1}, y^*(x^{k-1}))^2. \tag{4.16}$$

Thus we have

$$\begin{aligned}
&\operatorname{dist}(y^{k,0}, y^*(x^k))^2\\
\leq& \left(2 + 8\kappa^2\alpha^2\Gamma^2\right)(1-2\mu\tau\beta^2)^T \operatorname{dist}(y^*(x^{k-1}), y^{k-1,0})^2 + 4\kappa^2\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2\\
&+ 16\kappa^2\alpha^2\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N} \|\hat{v}^0(x^{k-1}, y^{k-1}) - P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1})\|^2\\
&+ 16\kappa^2\alpha^2\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N} \left(\ell_{g,2}\ell_{f,0} + \frac{\ell_{f,1}}{\mu}\right)^2 \operatorname{dist}(y^{k-1}, y^*(x^{k-1}))^2\\
\leq& \left[2 + 8\kappa^2\alpha^2\Gamma^2 + 16\kappa^2\alpha^2\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\left(\ell_{g,2}\ell_{f,0} + \frac{\ell_{f,1}}{\mu}\right)^2\right](1-2\mu\tau\beta^2)^T \operatorname{dist}(y^*(x^{k-1}), y^{k-1,0})^2\\
&+ 16\kappa^2\alpha^2\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N} \|\hat{v}^0(x^{k-1}, y^{k-1}) - P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1})\|^2 + 4\kappa^2\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2.
\end{aligned}$$

Now we bound $\|P_{y^*(x^k)\to y^k}v^*(x^k) - \hat{v}^0(x^k, y^k)\|^2$. We have

$$
\begin{aligned}
&\|P_{y^*(x^k)\to y^k}v^*(x^k) - \hat{v}^0(x^k, y^k)\|^2 = \|P_{y^*(x^k)\to y^k}v^*(x^k) - P_{y^{k-1}\to y^k}\hat{v}^N(x^{k-1}, y^{k-1})\|^2\\
\le& 2\|P_{y^*(x^k)\to y^k}v^*(x^k) - P_{y^*(x^{k-1})\to y^k}v^*(x^{k-1})\|^2 + 2\|P_{y^*(x^{k-1})\to y^k}v^*(x^{k-1}) - P_{y^{k-1}\to y^k}\hat{v}^N(x^{k-1}, y^{k-1})\|^2\\
\le& 2\|P_{y^*(x^k)\to y^*(x^{k-1})}v^*(x^k) - v^*(x^{k-1})\|^2 + 4\|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \tilde{v}^{k-1}\|^2 + 4\|\tilde{v}^{k-1} - \hat{v}^N(x^{k-1}, y^{k-1})\|^2\\
\le& 4\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|\tilde{v}^{k-1} - \hat{v}^0(x^{k-1}, y^{k-1})\|^2\\
&+ 2\|P_{y^*(x^k)\to y^*(x^{k-1})}v^*(x^k) - v^*(x^{k-1})\|^2 + 4\|P_{y^*(x^{k-1})\to y^k}v^*(x^{k-1}) - \tilde{v}^{k-1}\|^2\\
\le& 4\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|\tilde{v}^{k-1} - P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1})\|^2\\
&+ 4\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \hat{v}^0(x^{k-1}, y^{k-1})\|^2\\
&+ 2\|P_{y^*(x^k)\to y^*(x^{k-1})}v^*(x^k) - v^*(x^{k-1})\|^2 + 4\|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \tilde{v}^{k-1}\|^2\\
=& 4\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \hat{v}^0(x^{k-1}, y^{k-1})\|^2\\
&+ 2\|P_{y^*(x^k)\to y^*(x^{k-1})}v^*(x^k) - v^*(x^{k-1})\|^2 + 4\left(\kappa(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2N} + 1\right)\|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \tilde{v}^{k-1}\|^2,
\end{aligned}
\tag{4.17}
$$

where in the second last inequality we again used Lemma 4.2. Now we inspect the two terms in the last line above. Note that the last term is bounded in (4.16). For the first term, by (4.7) we have

$$
\|P_{y^*(x^k)\to y^*(x^{k-1})}v^*(x^k) - v^*(x^{k-1})\|^2 \le \left(\frac{\ell_{f,0}\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2} + \frac{\ell_{f,1}}{\mu}\right)^2\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2.
$$

Now plugging everything back to (4.17) we get

$$
\begin{aligned}
&\|P_{y^*(x^k)\to y^k}v^*(x^k) - \hat{v}^0(x^k, y^k)\|^2\\
\le& 4\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \hat{v}^0(x^{k-1}, y^{k-1})\|^2\\
&+ 2\left(\frac{\ell_{f,0}\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2} + \frac{\ell_{f,1}}{\mu}\right)^2\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2\\
&+ 4\left(\kappa(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2N} + 1\right)\left(\ell_{g,2}\ell_{f,0} + \frac{\ell_{f,1}}{\mu}\right)^2(1-2\mu\tau\beta^2)^T\operatorname{dist}(y^{k-1,0}, y^*(x^{k-1}))^2,
\end{aligned}
\tag{4.18}
$$

where we also used Lemma 4.3 in the last inequality. Now summing up the bound for $\operatorname{dist}(y^{k,0}, y^*(x^k))^2$ and $\|P_{y^*(x^k)\to y^k}v^*(x^k) - \hat{v}^0(x^k, y^k)\|^2$, we get:

$$
\begin{aligned}
&\operatorname{dist}(y^{k,0}, y^*(x^k))^2 + \|P_{y^*(x^k)\to y^k}v^*(x^k) - \hat{v}^0(x^k, y^k)\|^2\\
\le& C_1(1-2\mu\tau\beta^2)^T\operatorname{dist}(y^{k-1,0}, y^*(x^{k-1}))^2\\
&+ C_2\|P_{y^*(x^{k-1})\to y^{k-1}}v^*(x^{k-1}) - \hat{v}^0(x^{k-1}, y^{k-1})\|^2\\
&+ \left[2\left(\frac{\ell_{f,0}\ell_{g,2}}{\mu^2}\sqrt{1+\kappa^2} + \frac{\ell_{f,1}}{\mu}\right)^2 + 4\kappa^2\right]\alpha^2\|\mathsf{grad}\Phi(x^{k-1})\|^2,
\end{aligned}
$$

with

$$C_1 = \left(6 + 8\kappa^2\alpha^2\Gamma^2 + C_2\right)\left(\ell_{g,2}\ell_{f,0} + \frac{\ell_{f,1}}{\mu}\right)^2$$
$$C_2 = \left(4 + 16\kappa^2\alpha^2\ell_{g,1}^2\right)\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}.$$

Now consider the choice of T and N in the statement of this lemma, we can guarantee that $C_1, C_2 \leq 1/2$, thus

$$\begin{aligned}&\operatorname{dist}(y^{k,0}, y^*(x^k))^2 + \|P_{y^*(x^k)\to y^k} v^*(x^k) - \hat{v}^0(x^k, y^k)\|^2 \\ \leq&\frac{1}{2}(\operatorname{dist}(y^{k-1,0}, y^*(x^{k-1}))^2 + \|P_{y^*(x^k)\to y^k} v^*(x^k) - \hat{v}^0(x^k, y^k)\|^2) + \Omega\|\mathsf{grad}\Phi(x^{k-1})\|^2.\end{aligned}$$

The final result is obtained by taking the telescoping sum of the above inequality. □

Lemma 4.6. *Suppose the parameters are set the same as in Lemma 4.5, then we have*

$$\|h_\Phi^k - \mathsf{grad}\Phi(x^k)\|^2 \leq \delta_{T,N}\left(\frac{1}{2}\right)^k \Delta_0 + \delta_{T,N}\Omega\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\left\|\mathsf{grad}\Phi\left(x^j\right)\right\|^2, \tag{4.19}$$

where

$$\delta_{T,N} = 2\Gamma^2(1-2\mu\tau\beta^2)^T\ell_{g,1}^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}. \tag{4.20}$$

Proof. By Lemma 4.4 and $ab + cd \leq (a+c)(b+d)$ for any positive a, b, c, d, we have

$$\|h_\Phi^k - \mathsf{grad}\Phi(x^k)\|^2 \leq \delta_{T,N}\left(\operatorname{dist}(y^*(x^k), y^{k-1})^2 + \|\hat{v}^0(x^k, y^k) - P_{y^*(x^k)\to y^k} v^*(x^k)\|^2\right).$$

The proof is completed by applying Lemma 4.5. □

Now we proceed to the proof of Theorem 4.1.

Proof. [Proof of Theorem 4.1] By Lemma 4.1, we have

$$\begin{aligned}\Phi(x^{k+1}) &\leq \Phi(x^k) + \langle\mathsf{grad}\Phi(x^k), \mathsf{Exp}_{x^k}^{-1}(x^{k+1})\rangle_{x^k} + \frac{L_\Phi}{2}\operatorname{dist}(x^k, x^{k+1})^2 \\ &= \Phi(x^k) - \alpha\langle\mathsf{grad}\Phi(x^k), h_\Phi^k\rangle_{x^k} + \frac{L_\Phi\alpha^2}{2}\|h_\Phi^k\|_{x^k}^2 \\ &\leq \Phi(x^k) - (\frac{\alpha}{2} - \alpha^2 L_\Phi)\|\mathsf{grad}\Phi(x^k)\|_{x^k}^2 + (\frac{\alpha}{2} + \alpha^2 L_\Phi)\|\mathsf{grad}\Phi(x^k) - h_\Phi^k\|_{x^k}^2.\end{aligned}$$

Now by using Lemma 4.6, we get

$$\begin{aligned}\Phi(x^{k+1}) \leq&\Phi(x^k) - (\frac{\alpha}{2} - \alpha^2 L_\Phi)\|\mathsf{grad}\Phi(x^k)\|_{x^k}^2 \\ &+ (\frac{\alpha}{2} + \alpha^2 L_\Phi)\left[\delta_{T,N}\left(\frac{1}{2}\right)^k\Delta_0 + \delta_{T,N}\Omega\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\left\|\mathsf{grad}\Phi\left(x^j\right)\right\|^2\right].\end{aligned}$$

Now by taking the telescoping sum of the above inequality over k from 0 to $K-1$, we have

$$\left(\frac{\alpha}{2}-\alpha^2 L_\Phi\right)\sum_{k=0}^{K-1}\left\|\mathsf{grad}\Phi\left(x^k\right)\right\|^2 \leq \Phi\left(x_0\right)-\inf_{x\in\mathcal{M}}\Phi(x)+\left(\frac{\alpha}{2}+\alpha^2 L_\Phi\right)\delta_{T,N}\Delta_0 + \left(\frac{\alpha}{2}+\alpha^2 L_\Phi\right)\delta_{T,N}\Omega\sum_{k=1}^{K-1}\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\left\|\mathsf{grad}\Phi\left(x^j\right)\right\|^2.$$

By the fact that

$$\sum_{k=1}^{K-1}\sum_{j=0}^{k-1}\left(\frac{1}{2}\right)^{k-1-j}\left\|\mathsf{grad}\Phi\left(x^j\right)\right\|^2 \leq \sum_{k=0}^{K-1}\frac{1}{2^k}\sum_{k=0}^{K-1}\left\|\mathsf{grad}\Phi\left(x^k\right)\right\|^2 \leq 2\sum_{k=0}^{K-1}\left\|\mathsf{grad}\Phi\left(x^k\right)\right\|^2,$$

we have

$$\left(\frac{\alpha}{2}-\alpha^2 L_\Phi-(\alpha+2\alpha^2 L_\Phi)\delta_{T,N}\Omega\right)\sum_{k=0}^{K-1}\left\|\mathsf{grad}\Phi\left(x^k\right)\right\|^2 \leq \Phi\left(x_0\right)-\inf_{x\in\mathcal{M}}\Phi(x)+\left(\frac{\alpha}{2}+\alpha^2 L_\Phi\right)\delta_{T,N}\Delta_0.$$

Choosing $N \geq \Theta(\sqrt{\kappa})$ and $D \geq \Theta(\kappa)$ as in Lemma 4.5, we are able to ensure that

$$\Omega\left(1+2\alpha L_\Phi\right)\delta_{T,N} \leq \frac{1}{4}, \quad \delta_{T,N} \leq 1.$$

As a result, we get

$$\left(\frac{\alpha}{4}-\alpha^2 L_\Phi\right)\sum_{k=0}^{K-1}\left\|\mathsf{grad}\Phi\left(x^k\right)\right\|^2 \leq \Phi\left(x_0\right)-\inf_{x\in\mathcal{M}}\Phi(x)+\left(\frac{\alpha}{2}+\alpha^2 L_\Phi\right)\Delta_0.$$

Thus, with $\alpha \leq \frac{1}{8L_\Phi}$ we get

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\mathsf{grad}\Phi\left(x^k\right)\right\|^2 \leq \frac{64L_\Phi\left(\Phi\left(x_0\right)-\inf_x\Phi(x)\right)+5\Delta_0}{K}.$$

Now we inspect the oracle complexities. To ensure $\frac{1}{K}\sum_{k=0}^{K-1}\left\|\mathsf{grad}\Phi\left(x^k\right)\right\|^2 \leq \epsilon$, we need $K=\mathcal{O}(\frac{\kappa^3}{\epsilon})$, so that $\mathrm{Gc}(f,\epsilon)=\mathcal{O}(\frac{\kappa^3}{\epsilon})$. Since in each outer iteration, we need $D=\mathcal{O}(\kappa)$ iterations, so $\mathrm{Gc}(g,\epsilon)=\mathcal{O}(\frac{\kappa^4}{\epsilon})$. The Jacobian-vector product count is the same as the iteration number K since it is only conducted once for every iteration. The Hessian-vector product is conducted for $N=\mathcal{O}(\sqrt{\kappa})$ times for each iteration. Thus we have the previously described complexities. □

5 Stochastic Algorithm RieSBO and Its Convergence

In this section, we propose RieSBO (Algorithm 2) for stochastic bilevel manifold optimization (1.2). The algorithm is a generalization of its counterpart in the Euclidean space as in Hong et al. (2020); Chen et al. (2021b), where we employ the Neumann series estimation for the hypergradient as in (3.13).

For the stochastic case, we utilize the following notion of stationarity.

Algorithm 2: Algorithm for **Rie**mannian **S**tochastic **B**ilevel **O**ptimization (**RieSBO**)

```
input  : $K$, $T$, $Q$, stepsize $\{\alpha_k, \beta_k\}$, initializations $x^0 \in \mathcal{M}, y^0 \in \mathcal{N}$
for $k = 0, 1, 2, ..., K-1$ do
    Set $y^{k,0} = y^{k-1}$;
    for $t = 0, ..., T-1$ do
        Update $y^{k,t+1} \leftarrow \mathsf{Exp}_{y^{k,t}}(-\beta_k \tilde{h}_g^{k,t})$ with $\tilde{h}_g^{k,t} := \mathsf{grad}_y G(x^k, y^{k,t}; \zeta_{k,t})$;
    end
    Set $y^k \leftarrow y^{k,T}$;
    Update $x^{k+1} \leftarrow \mathsf{Exp}_{x^k}(-\alpha_k \tilde{h}_\Phi^k)$, where $\tilde{h}_\Phi^k$ is as defined in (3.13);
end
```

Definition 5.1. *A random point $x \in \mathcal{M}$ is called an ϵ-stationary point for* (1.2) *if $\mathbb{E}\|\nabla\Phi(x)\|^2 \le \epsilon$.*

We now proceed to the convergence analysis for the Riemannian stochastic bilevel optimization (RieSBO, Algorithm 2). For RieSBO, we need the following additional assumption over the mean and variance of the estimators.

Assumption 5.1. *The stochastic gradients satisfy $\mathsf{grad}F(x, y; \xi) = [\mathsf{grad}_x F(x, y; \xi), \mathsf{grad}_y F(x, y; \xi)]$ and $\mathsf{grad}G(x, y; \zeta) = [\mathsf{grad}_x G(x, y; \zeta), \mathsf{grad}_y G(x, y; \zeta)]$. The second order gradients $\mathsf{grad}^2_{x,y} G(x, y; \zeta)$, $H_y(G(x, y; \zeta))$ are all unbiased estimators of the corresponding deterministic quantities of f and g. Their variances are all bounded by σ^2 (in tangent space norms and operator norms, respectively for the Riemannian gradient and Riemannian Hessian).*

Note that we do not need to assume the smoothness or strong-convexity of the stochastic functions F and G.

Now we are ready to state the following convergence result.

Theorem 5.1. *Suppose Assumptions 4.1, 4.2, 4.3, 4.4 and 5.1 hold. If we take the stepsizes $\alpha_k = \alpha = \frac{1}{\kappa^{5/2}\sqrt{K}}$, $\beta_k = \beta = \min\{\frac{1}{\kappa^{7/4}\sqrt{K}}, \frac{1}{\ell_{g,1}}\}$, also $\eta = 1/\ell_{g,1}, Q = \mathcal{O}(\kappa \log K)$ and $T = \mathcal{O}(\kappa^4)$. Also suppose that the random variables for all iterations ζ_k^t, $\zeta_{k,(q)}$, ξ_k are i.i.d. samples, then RieSBO (Algorithm 2) satisfies*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2] \le \mathcal{O}\left(\frac{\kappa^{2.5}}{\sqrt{K}}\right).$$

Here the expectation is taken with respect to all the random samples. In order to obtain an ϵ-stationary point, i.e., $\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2] \le \epsilon$, the oracle complexities needed are given by:

- *Gradients:* $\mathrm{Gc}(f, \epsilon) = \mathcal{O}(\kappa^5 \epsilon^{-2})$, $\mathrm{Gc}(g, \epsilon) = \mathcal{O}(\kappa^9 \epsilon^{-2})$;
- *Jacobian and Hessian-vector products:* $\mathrm{JV}(g, \epsilon) = \mathcal{O}(\kappa^5 \epsilon^{-2})$, $\mathrm{HV}(g, \epsilon) = \mathcal{O}(\kappa^6 \epsilon^{-2})$.

To prove this theorem, we need the following lemmas. For simplicity, denote $\mathcal{U}_k$ the σ-algebra generated by all the random samples up to the $(k-1)$-th iterate, and denote $\bar{h}_\Phi^k := \mathbb{E}[\tilde{h}_\Phi^k \mid \mathcal{U}_k]$, i.e., the expectation only with respect to the samples of the current iterate.

Lemma 5.1. *Suppose we estimate the hypergradient $\tilde{h}_\Phi^k$ via (3.13) with $\eta \le \frac{1}{\ell_{g,1}}$, then we have the following bounds.*

$$\mathbb{E}[\|\tilde{h}_\Phi^k - \bar{h}_\Phi^k\|^2 \mid \mathcal{U}_k] \le \tilde{\sigma}^2, \tag{5.1}$$

and

$$\|\mathsf{grad}\hat{\Phi}(x^k) - \bar{h}_\Phi^k\|^2 \leq b_k^2, \tag{5.2}$$

where

$$\begin{aligned} \tilde{\sigma}^2 &:= 2\sigma^2 + 6\left(\sigma^2(\sigma^2 + \ell_{f,0}^2) + \ell_{g,1}^2(\sigma^2 + \ell_{f,0}^2) + \ell_{g,1}^2\sigma^2\right)\max\{\frac{1}{\mu^2}, \frac{d_1^2}{\eta^2\mu^2}\} = \mathcal{O}(\kappa^2), \\ b_k &:= \ell_{f,0}\frac{\ell_{g,1}}{\mu}(1 - \frac{\mu}{\ell_{g,1}})^Q, \end{aligned} \tag{5.3}$$

and $\hat{\Phi}(x) = f(x, y^T(x))$ *which is the approximate function after* T *steps of the inner loop.*

Further, we have the following bound on the second moment:

$$\mathbb{E}[\|\tilde{h}_\Phi^k\|^2 \mid \mathcal{U}_k] \leq 2\tilde{\sigma}^2 + 4b_k^2 + 4\ell_{f,0}^2(1+\kappa)^2 =: \tilde{C}^2 = \mathcal{O}(\kappa^2). \tag{5.4}$$

Proof. [Proof of Lemma 5.1] By the expression (3.1) for $\mathsf{grad}\Phi(x)$ and (3.11) for $\tilde{h}_\Phi^k$, we have:

$$\begin{aligned} \mathsf{grad}\Phi(x^k) &= \mathsf{grad}_x f(x^k, y^*(x^k)) - \mathsf{grad}_{y,x}^2 g(x^k, y^*(x^k))[v^*(x^k)], \\ \mathsf{grad}\hat{\Phi}(x^k) &= \mathsf{grad}_x f(x^k, y^k) - \mathsf{grad}_{y,x}^2 g(x^k, y^k)[\tilde{v}^k], \\ \tilde{h}_\Phi^k &= \mathsf{grad}_x F(x^k, y^k; \xi_k) - \mathsf{grad}_{y,x}^2 G(x^k, y^k; \zeta_{k,(0)})[v_Q^k], \end{aligned}$$

where again $\tilde{v}^k := (H_y(g(x^k, y^{k,T})))^{-1}\mathsf{grad}_y f(x^k, y^{k,T})$.

For (5.1), denote

$$\bar{v}_Q^k = \mathbb{E}[v_Q^k] = \eta\sum_{q=1}^{Q}(I - \eta H_y(g(x^k, y^k)))^q[\mathsf{grad}_y f(x^k, y^k)].$$

We have

$$\begin{aligned} &\mathbb{E}[\|\tilde{h}_\Phi^k - \bar{h}_\Phi^k\|^2 \mid \mathcal{U}_k] \\ \leq & 2\mathbb{E}[\|\mathsf{grad}_x f\left(x^k, y^{k,T}\right) - \mathsf{grad}_x F\left(x^k, y^{k,T}; \xi_k\right)\|^2 \mid \mathcal{U}_k] \\ &+ 2\mathbb{E}[\|\mathsf{grad}_{y,x}^2 G(x^k, y^k; \zeta_{k,(0)})[v_Q^k] - \mathsf{grad}_{y,x}^2 g(x^k, y^k)[\bar{v}_Q^k]\|^2 \mid \mathcal{U}_k] \\ \leq & 2\sigma^2 + 2\mathbb{E}[\|\mathsf{grad}_{y,x}^2 G(x^k, y^k; \zeta_{k,(0)})[v_Q^k] - \mathsf{grad}_{y,x}^2 g(x^k, y^k)[\bar{v}_Q^k]\|^2 \mid \mathcal{U}_k]. \end{aligned} \tag{5.5}$$

We now inspect the last term above. Denote

$$H^k := \eta Q\prod_{q=1}^{Q'}(I - \eta H_y(G(x^k, y^k; \zeta_{k,(q)}))),$$

which is our estimation of the Riemannian Hessian at the k-th outer iteration, and we have that

$$\begin{aligned} &\mathsf{grad}_{y,x}^2 G(x^k, y^k; \zeta_{k,(0)})[v_Q^k] - \mathsf{grad}_{y,x}^2 g(x^k, y^k)[\bar{v}_Q^k] \\ =&\mathsf{grad}_{y,x}^2 G(x^k, y^k; \zeta_{k,(0)})\left[H^k[\mathsf{grad}_y F(x^k, y^k; \xi_k)]\right] - \mathsf{grad}_{y,x}^2 g(x^k, y^k)\left[\mathbb{E}[H^k[\mathsf{grad}_y F(x^k, y^k; \xi_k)]]\right] \\ =&\left\{\mathsf{grad}_{y,x}^2 G(x^k, y^k; \zeta_{k,(0)}) - \mathsf{grad}_{y,x}^2 g(x^k, y^k)\right\}\left[H^k[\mathsf{grad}_y F(x^k, y^k; \xi_k)]\right] \\ &+ \mathsf{grad}_{y,x}^2 g(x^k, y^k)\left[\left\{H^k - \mathbb{E}[H^k]\right\}[\mathsf{grad}_y F(x^k, y^k; \xi_k)]\right] \\ &+ \mathsf{grad}_{y,x}^2 g(x^k, y^k)\mathbb{E}[H^k]\left\{\mathsf{grad}_y F(x^k, y^k; \xi_k) - \mathsf{grad}_y f(x^k, y^k)\right\}. \end{aligned}$$

Since

$$\begin{aligned}&\mathbb{E}[\|\mathsf{grad}_y F(x^k,y^k;\xi_k)\|^2]\\=&\mathbb{E}[\|\mathsf{grad}_y F(x^k,y^k;\xi_k)-\mathsf{grad}_y f(x^k,y^k)\|^2]+\mathbb{E}[\|\mathsf{grad}_y f(x^k,y^k)\|^2]\le\sigma^2+\ell_{f,0}^2,\end{aligned}$$

we have that

$$\begin{aligned}&\mathbb{E}[\|\mathsf{grad}^2_{y,x}G(x^k,y^k;\zeta_{k,(0)})[v_Q^k]-\mathsf{grad}^2_{y,x}g(x^k,y^k)[\bar v_Q^k]\|^2\mid\mathcal{U}_k]\\\le&3\sigma^2(\sigma^2+\ell_{f,0}^2)\mathbb{E}\|H^k\|_{\mathrm{op}}^2+3\ell_{g,1}^2(\sigma^2+\ell_{f,0}^2)\mathbb{E}\|H^k-\mathbb{E}[H^k]\|_{\mathrm{op}}^2+3\ell_{g,1}^2\sigma^2\|\mathbb{E}[H^k]\|_{\mathrm{op}}^2.\end{aligned}$$

It remains to bound $\mathbb{E}\|H^k\|_{\mathrm{op}}^2$ and $\|\mathbb{E}[H^k]\|_{\mathrm{op}}$. For $\mathbb{E}\|H^k\|_{\mathrm{op}}^2$, using Hong et al. (2020, Lemma 12), we have that

$$\mathbb{E}\|H^k\|_{\mathrm{op}}^2\le\frac{d_1}{\eta\mu},$$

where $d_1>0$ is some absolute constant. On the other hand $\|\mathbb{E}[H^k]\|_{\mathrm{op}}$ can be easily calculated as (since $\mu\eta<\mu/\ell_{g,1}<1$)

$$\begin{aligned}\|\mathbb{E}[H^k]\|_{\mathrm{op}}=&\eta\|\sum_{q=1}^{Q}(I-\eta H_y(g(x^k,y^k)))^q\|_{\mathrm{op}}\\\le&\|H^{-1}\|_{\mathrm{op}}\|I-\eta H_y(g(x^k,y^k))\|_{\mathrm{op}}\le\frac{1}{\mu}.\end{aligned}$$

Therefore, we finally have

$$\begin{aligned}&\mathbb{E}[\|\mathsf{grad}^2_{y,x}G(x^k,y^k;\zeta_{k,(0)})[v_Q^k]-\mathsf{grad}^2_{y,x}g(x^k,y^k)[\bar v_Q^k]\|^2\mid\mathcal{U}_k]\\\le&3\left(\sigma^2(\sigma^2+\ell_{f,0}^2)+\ell_{g,1}^2(\sigma^2+\ell_{f,0}^2)+\ell_{g,1}^2\sigma^2\right)\max\{\frac{1}{\mu^2},\frac{d_1^2}{\eta^2\mu^2}\}.\end{aligned}$$

Plugging the above equation to (5.5) we get (5.1).

Now for (5.2), since

$$\begin{aligned}\bar h_\Phi^k:=&\mathbb{E}\left[\mathsf{grad}_x F(x^k,y^k;\xi_k)-\mathsf{grad}^2_{y,x}G(x^k,y^k;\zeta_{k,(0)})[v_Q^k]\right]\\=&\,\mathsf{grad}_x f(x^k,y^k)-\mathsf{grad}^2_{y,x}g(x^k,y^k)[\bar v_Q^k],\end{aligned}$$

we have

$$\begin{aligned}&\|\mathsf{grad}\hat\Phi(x^k)-\bar h_\Phi^k\|^2\le\|\mathsf{grad}^2_{y,x}g(x^k,y^k)\|_{\mathrm{op}}^2\|\tilde v^k-\bar v_Q^k\|^2\le\ell_{g,1}^2\|\tilde v^k-\bar v_Q^k\|^2\\=&\ell_{g,1}^2\|(H_y(g(x^k,y^k)))^{-1}[\mathsf{grad}_y f(x^k,y^k)]-\eta\sum_{q=1}^{Q}(I-\eta H_y(g(x^k,y^k)))^q[\mathsf{grad}_y f(x^k,y^k)]\|^2\\\le&\ell_{g,1}^2\ell_{f,0}^2\|(H_y(g(x^k,y^k)))^{-1}-\eta\sum_{q=1}^{Q}(I-\eta H_y(g(x^k,y^k)))^q\|_{\mathrm{op}}^2\\\le&\ell_{f,0}^2\frac{\ell_{g,1}^2}{\mu^2}(1-\frac{\mu}{\ell_{g,1}})^{2Q}=b_k^2,\end{aligned}$$

where the last line is by Ghadimi and Wang (2018, Lemma 3.2). Note that we take $\eta\le\frac{1}{\ell_{g,1}}$ so that the Neumann sequence converges.

Now for the moment $\mathbb{E}[\|\tilde{h}_\Phi^k\|^2 \mid \mathcal{U}_k]$, we have

$$\begin{aligned}\mathbb{E}[\|\tilde{h}_\Phi^k\|^2 \mid \mathcal{U}_k] \leq & 2\mathbb{E}[\|\tilde{h}_\Phi^k - \bar{h}_\Phi^k\|^2 \mid \mathcal{U}_k] + 4\|\bar{h}_\Phi^k - \mathsf{grad}\hat{\Phi}(x^k)\|^2 + 4\|\mathsf{grad}\hat{\Phi}(x^k)\|^2 \\ \leq & 2\tilde{\sigma}^2 + 4b_k^2 + 4\|\mathsf{grad}\hat{\Phi}(x^k)\|^2.\end{aligned}$$

Since

$$\begin{aligned}\|\mathsf{grad}\hat{\Phi}(x^k)\| = & \|\mathsf{grad}_x f(x^k, y^k) - \mathsf{grad}_{y,x}^2 g(x^k, y^k)[\tilde{v}^k]\| \\ \leq & \|\mathsf{grad}_x f(x^k, y^k)\| + \|\mathsf{grad}_{y,x}^2 g(x^k, y^k)\|_{\text{op}}\|\tilde{v}^k\| \leq \ell_{f,0} + \ell_{g,1}\frac{\ell_{f,0}}{\mu} = \ell_{f,0}(1+\kappa),\end{aligned}$$

we have

$$\mathbb{E}[\|\tilde{h}_\Phi^k\|^2 \mid \mathcal{U}_k] \leq 2\tilde{\sigma}^2 + 4b_k^2 + 4\ell_{f,0}^2(1+\kappa)^2.$$

This completes the proof. □

Lemma 5.2. *Suppose we have the sequence $\{y^{k,t}\}$ by RieSBO with stepsize $\beta_k = \beta \leq \frac{1}{\ell_{g,1}}$, then the following inequalities hold:*

$$\mathbb{E}\,\mathrm{dist}(y^{k,T}, y^*(x^k))^2 \leq (1 - 2\mu\tau\beta^2)^T\,\mathrm{dist}(y^{k,0}, y^*(x^k))^2 + \tau\beta^2\sigma^2 T, \tag{5.6}$$

and

$$\begin{aligned}&\mathbb{E}[\mathrm{dist}(y^{k,T}, y^*(x^{k+1}))^2] \\ \leq & 2(1 - 2\mu\tau\beta^2)^T\,\mathrm{dist}(y^{k,0}, y^*(x^k))^2 + 2\tau\beta^2\sigma^2 T + 4\tau\kappa^2\alpha^2\|\bar{h}_\Phi^k\|_{x^k}^2 + 4\tau\kappa^2\alpha^2\tilde{\sigma}^2.\end{aligned} \tag{5.7}$$

Proof. [Proof of Lemma 5.2] For simplicity, all the expectations are conditioned on $\mathcal{U}_k$ in this proof.

First we have by Zhang and Sra (2016, Corollary 8) that

$$\begin{aligned}&\mathbb{E}_{\zeta_{k,t}}\,\mathrm{dist}(y^{k,t+1}, y^*(x^k))^2 \\ \leq & \mathrm{dist}(y^{k,t}, y^*(x^k))^2 + 2\beta\langle \mathsf{grad} g(x^k, y^k), \mathsf{Exp}_{y^{k,t}}(y^*(x^k))\rangle + \tau\beta^2\mathbb{E}_{\zeta_{k,t}}\|\tilde{h}_g^{k,t}\|^2 \\ \leq & \mathrm{dist}(y^{k,t}, y^*(x^k))^2 + 2\beta\langle \mathsf{grad} g(x^k, y^k), \mathsf{Exp}_{y^{k,t}}(y^*(x^k))\rangle + \tau\beta^2\|\mathsf{grad} g(x^k, y^k)\|^2 + \tau\beta^2\sigma^2 \\ \leq & (1 - 2\mu\tau\beta^2)\,\mathrm{dist}(y^{k,t}, y^*(x^k))^2 + \tau\beta^2\sigma^2,\end{aligned}$$

where in the last line we used the same trick as the proof of Lemma 4.3. Note that in the above formulas the expectation is only taken with respect to the random variables in $\tilde{h}_g^{k,t}$, i.e., $\zeta_{k,t}$. Repeating this for T times yields (5.6). Now for the second inequality (5.7), we have

$$\begin{aligned}&\mathbb{E}[\mathrm{dist}(y^{k,T}, y^*(x^{k+1}))^2] \\ \leq & 2\mathbb{E}[\mathrm{dist}(y^{k,T}, y^*(x^k))^2] + 2\mathbb{E}\,\mathrm{dist}(y^*(x^k), y^*(x^{k+1}))^2 \\ \leq & 2(1 - 2\mu\tau\beta^2)^T\,\mathrm{dist}(y^{k,0}, y^*(x^k))^2 + 2\tau\beta^2\sigma^2 T + 2\tau\kappa^2\mathbb{E}\,\mathrm{dist}(x^k, x^{k+1})^2,\end{aligned} \tag{5.8}$$

where the last inequality is by (5.6) and Lemma 4.1. For $\mathbb{E}d(x^{k+1}, x^k)^2$ we have the bound:

$$\begin{aligned}&\mathbb{E}d(x^{k+1}, x^k)^2 = \alpha^2\mathbb{E}\|\tilde{h}_\Phi^k\|_{x^k}^2 \\ = & \alpha^2\mathbb{E}\|\tilde{h}_\Phi^k - \bar{h}_\Phi^k + \bar{h}_\Phi^k\|_{x^k}^2 \leq 2\alpha^2(\|\bar{h}_\Phi^k\|_{x^k}^2 + \tilde{\sigma}^2),\end{aligned}$$

which completes the proof. □

Now we turn to the proof of Theorem 5.1.

Proof. [Proof of Theorem 5.1] Denote $V_k := \Phi(x^k) + \kappa \operatorname{dist}(y^{k-1,T}, y^*(x^k))^2$. By Lemma 4.1 and Lemma 5.1, we have

$$
\begin{aligned}
&\mathbb{E}[\Phi(x^{k+1}) \mid \mathcal{U}_k] \leq \Phi(x^k) + \mathbb{E}[\langle \mathsf{grad}\Phi(x^k), \mathsf{Exp}_{x^k}^{-1}(x^{k+1})\rangle_{x^k} \mid \mathcal{U}_k] + \frac{L_\Phi}{2}\mathbb{E}[\operatorname{dist}(x^k, x^{k+1})^2 \mid \mathcal{U}_k] \\
=&\Phi(x^k) - \alpha \mathbb{E}[\langle \mathsf{grad}\Phi(x^k), \tilde{h}_\Phi^k\rangle_{x^k} \mid \mathcal{U}_k] + \frac{L_\Phi \alpha^2}{2}\|\tilde{h}_\Phi^k\|_{x^k}^2 \\
=&\Phi(x^k) - \frac{\alpha}{2}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2 \mid \mathcal{U}_k] - (\frac{\alpha}{2} - \frac{\alpha^2 L_\Phi}{2})\|\bar{h}_\Phi^k\|^2 + \frac{\alpha}{2}\|\mathsf{grad}\Phi(x^k) - \bar{h}_\Phi^k\|^2 \\
&+ \frac{\alpha^2 L_\Phi}{2}\mathbb{E}[\|\tilde{h}_\Phi^k - \bar{h}_\Phi^k\|^2 \mid \mathcal{U}_k] \\
\leq&\Phi(x^k) - \frac{\alpha}{2}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2 \mid \mathcal{U}_k] - (\frac{\alpha}{2} - \frac{\alpha^2 L_\Phi}{2})\|\bar{h}_\Phi^k\|^2 + \frac{\alpha}{2}\|\mathsf{grad}\Phi(x^k) - \bar{h}_\Phi^k\|^2 + \frac{\alpha^2 L_\Phi}{2}\tilde{\sigma}^2.
\end{aligned}
$$

Now we decompose the bias term $\|\mathsf{grad}\Phi(x^k) - \bar{h}_\Phi^k\|$ as:

$$
\begin{aligned}
\|\mathsf{grad}\Phi(x^k) - \bar{h}_\Phi^k\|^2 =&2\|\mathsf{grad}\Phi(x^k) - \mathsf{grad}\hat{\Phi}(x^k)\|^2 + 2\|\mathsf{grad}\hat{\Phi}(x^k) - \bar{h}_\Phi^k\|^2 \\
\leq&2\Gamma_0^2 \operatorname{dist}(y^{k,T}, y^*(x^k))^2 + 2b_k^2,
\end{aligned}
\tag{5.9}
$$

where we use a similar process as the proof of Lemma 4.5 to bound $\|\mathsf{grad}\Phi(x^k) - \mathsf{grad}\hat{\Phi}(x^k)\|$ and $\Gamma_0 = \ell_{f,1} + \frac{\ell_{f,0}\ell_{g,2}}{\mu} + \ell_{g,1}(\ell_{g,2}\ell_{f,0} + \frac{\ell_{f,1}}{\mu}) = \mathcal{O}(\kappa)$. Thus we have

$$
\begin{aligned}
\mathbb{E}[\Phi(x^{k+1}) \mid \mathcal{U}_k] \leq&\Phi(x^k) - \frac{\alpha}{2}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2 \mid \mathcal{U}_k] - (\frac{\alpha}{2} - \frac{\alpha^2 L_\Phi}{2})\|\bar{h}_\Phi^k\|^2 \\
&+ \alpha\Gamma_0^2 \operatorname{dist}(y^{k,T}, y^*(x^k))^2 + \alpha b_k^2 + \frac{\alpha^2 L_\Phi}{2}\tilde{\sigma}^2.
\end{aligned}
\tag{5.10}
$$

Now we have

$$
\begin{aligned}
\mathbb{E}[V_{k+1}] - \mathbb{E}[V_k] =& \mathbb{E}[\Phi(x^{k+1})] - \mathbb{E}[\Phi(x^k)] + \kappa\mathbb{E}\operatorname{dist}(y^{k,T}, y^*(x^{k+1}))^2 - \kappa\mathbb{E}\operatorname{dist}(y^{k-1,T}, y^*(x^k))^2 \\
\leq& -\frac{\alpha}{2}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2 \mid \mathcal{U}_k] - (\frac{\alpha}{2} - \frac{\alpha^2 L_\Phi}{2})\mathbb{E}\|\bar{h}_\Phi^k\|^2 + \alpha b_k^2 + \frac{\alpha^2 L_\Phi}{2}\tilde{\sigma}^2 \\
&+ \kappa\mathbb{E}\operatorname{dist}(y^{k,T}, y^*(x^{k+1}))^2 - \kappa\mathbb{E}\operatorname{dist}(y^{k-1,T}, y^*(x^k))^2 + \alpha\Gamma_0^2\mathbb{E}\operatorname{dist}(y^{k,T}, y^*(x^k))^2 \\
\leq& -\frac{\alpha}{2}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2 \mid \mathcal{U}_k] - (\frac{\alpha}{2} - \frac{\alpha^2 L_\Phi}{2})\mathbb{E}\|\bar{h}_\Phi^k\|^2 + \alpha b_k^2 + \frac{\alpha^2 L_\Phi}{2}\tilde{\sigma}^2 \\
&+ \kappa\mathbb{E}\Big(2(1 - 2\mu\tau\beta^2)^T \operatorname{dist}(y^{k,0}, y^*(x^k))^2 + 2\tau\beta^2\sigma^2 T + 4\tau\kappa^2\alpha^2\|\bar{h}_\Phi^k\|_{x^k}^2 + 4\tau\kappa^2\alpha^2\tilde{\sigma}^2\Big) \\
&- \kappa\mathbb{E}\operatorname{dist}(y^{k-1,T}, y^*(x^k))^2 + \alpha\Gamma_0^2\Big((1 - 2\mu\tau\beta^2)^T\mathbb{E}\operatorname{dist}(y^{k,0}, y^*(x^k))^2 + \tau\beta^2\sigma^2 T\Big) \\
=& -\frac{\alpha}{2}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2 \mid \mathcal{U}_k] - \Big(\frac{\alpha}{2} - \frac{\alpha^2 L_\Phi}{2} - 4\tau\kappa^3\alpha^2\Big)\mathbb{E}\|\bar{h}_\Phi^k\|^2 \\
&+ \Big((2\kappa + \alpha\Gamma_0^2)(1 - 2\mu\tau\beta^2)^T - \kappa\Big)\mathbb{E}\operatorname{dist}(y^{k-1,T}, y^*(x^k))^2 \\
&+ \Big(2\kappa + \alpha\Gamma_0^2\Big)\tau\beta^2\sigma^2 T + \alpha b_k^2 + (\frac{L_\Phi}{2} + 4\tau\kappa^2)\alpha^2\tilde{\sigma}^2,
\end{aligned}
$$

where the first inequality is by (5.10) and the second inequality is by Lemma 5.2, as well as the fact that $y^{k+1,0} = y^{k,T}$. To make the coefficients negative, notice that by taking

$$\begin{aligned}
&\alpha \leq \frac{1}{L_\Phi + 8\tau\kappa^2} \\
&T \geq \log\left(\frac{1}{1-2\mu\tau\beta^2}\right) / \log\left(\frac{2\kappa+\alpha\Gamma_0^2}{\kappa}\right),
\end{aligned}$$

we can guarantee

$$\begin{aligned}
&\frac{\alpha}{2} - \frac{\alpha^2 L_\Phi}{2} - 4\tau\kappa^3\alpha^2 \geq 0 \\
&(2\kappa + \alpha\Gamma_0^2)(1-2\mu\tau\beta^2)^T - \kappa \leq 0.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[V_{k+1}] - \mathbb{E}[V_k] \leq &- \frac{\alpha}{2}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2 \mid \mathcal{U}_k] \\
&+ \left(2\kappa + \alpha\Gamma_0^2\right)\tau\beta^2\sigma^2 T + \alpha b_k^2 + (\frac{L_\Phi}{2} + 4\tau\kappa^2)\alpha^2\tilde{\sigma}^2.
\end{aligned} \tag{5.11}$$

Note that here we do not need an increasing T.

Now taking the telescoping sum of the above inequality for $k = 0, ..., K-1$, we get

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2] \leq \frac{2V_0}{\alpha K} + \frac{2}{K}\sum_{k=0}^{K-1} b_k^2 + \left(\frac{4\kappa}{\alpha} + 2\Gamma_0^2\right)\tau\beta^2\sigma^2 T + (L_\Phi + 8\tau\kappa^2)\alpha\tilde{\sigma}^2.$$

Now since $b_k^2 = \ell_{f,0}^2\kappa^2(1-\frac{1}{\kappa})^{2Q}$, the term $\frac{2}{K}\sum_{k=0}^{K-1} b_k^2 = \mathcal{O}(\frac{1}{\sqrt{K}})$ if $Q = \mathcal{O}(\kappa\log(K))$, following the inequality $(1-x)^n \leq e^{-nx}$. If we also select $\alpha_k = \alpha = \frac{1}{\kappa^{5/2}\sqrt{K}}$, $\beta_k = \beta = \min\{\frac{1}{\kappa^{7/4}\sqrt{K}}, \frac{1}{\ell_{g,1}}\}$ and $T = \mathcal{O}(\kappa^4)$, we are able to get:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2] \leq \mathcal{O}(\frac{\kappa^{2.5}}{\sqrt{K}}).$$

Now we inspect the oracle complexities. To ensure $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\mathsf{grad}\Phi(x^k)\|^2] \leq \epsilon$, we need $K = \mathcal{O}(\kappa^5\epsilon^{-2})$, thus $\mathrm{Gc}(F,\epsilon) = \mathcal{O}(\kappa^5\epsilon^{-2})$; Also $\mathrm{Gc}(G,\epsilon) = KT = \mathcal{O}(\kappa^9\epsilon^{-2})$. □

Remark 5.1. *Note that the trick for estimating the Hessian-vector product (3.12) can also be applied to the deterministic case, without using conjugate gradient method, leading to an easier implementation. We just need to replace the stochastic functions in (3.12) by their deterministic versions. In the experiments, we always use (3.12) in this way instead of solving (3.8) which uses N-step conjugate gradient method, while still achieving reasonable results numerically.*

6 Numerical experiments on robust optimization on manifolds

Consider the robust optimization on manifolds:

$$\min_{x\in\mathcal{M}}\max_{y\in\Delta_n}\sum_{i=1}^{n} y_i\ell\left(x;\xi_i\right) - \lambda\left\|y - \frac{\mathbf{1}}{n}\right\|^2, \tag{6.1}$$

where $\Delta_n := \{y \in \mathbb{R}^n : \sum_{i=1}^n y_i = 1, y_i \geq 0\}$ is the probability simplex, and ℓ is geodesically convex. This problem minimizes n loss function by dynamically assigning different weights to them, and making sure that the larger loss has larger weights (see Chen et al. (2017); Huang et al. (2020)). By minimax theorem we can exchange the min and max of the problem, thus it can be equivalently formulated as a bilevel optimization as follows:

$$\begin{aligned} &\min_{y \in \Delta_n} \ \lambda \left\| y - \frac{\mathbf{1}}{n} \right\|^2 - \sum_{i=1}^n y_i \ell\left(x; \xi_i\right) \\ &\text{s.t. } \ x \in \operatorname*{argmin}_{x \in \mathcal{M}} \ \sum_{i=1}^n y_i \ell\left(x; \xi_i\right). \end{aligned} \tag{6.2}$$

It is worth noticing that having an constraint set in the upper level problem is not covered in our theoretical analysis due to the fact that existing constrained Riemannian optimization techniques such as (stochastic) Riemannian Frank-Wolfe (see Weber and Sra (2022, 2023)) require a mini-batch sampling technique, i.e., using (3.13) multiple times and taking the average of these estimators to estimate $\mathsf{grad}\Phi(x^k)$ to reduce the variance, which is not desirable in practice. Instead, we point out that since the upper level in (6.2) is a constrained optimization in a Euclidean space, one could utilize the analysis in Hong et al. (2020) to achieve a similar convergence result as the unconstrained case in (1.1). Therefore we simply add a projection step for the upper level update, and we still observed reasonable convergence results. We present the algorithm we use for the numerical experiments in Algorithm 3. Note that here the variables in the upper and lower level problems are respectively denoted by y and x, which is different from the previous algorithms. It is also worth noticing that the convergence criteria are altered due to the existence of the projection step: in Algorithm 3, we simply measure the norm of the quantity

$$\mathcal{G}^k := \frac{1}{\alpha_k}(y^k - y^{k+1}) = \frac{1}{\alpha_k}(y^k - \mathsf{proj}_{\Delta_n}(y^k - \alpha_k h_\Phi^k)),$$

which we refer to as the approximate gradient mapping. This quantity can be used for approximately measuring the stationarity since if we do not have a constraint,

$$\mathcal{G}^k = \frac{1}{\alpha_k}(y^k - y^{k+1}) = h_\Phi^k \approx \mathsf{grad}\Phi(y^k),$$

based on Lemma 5.1.

Algorithm 3: Bilevel algorithm for robust manifold optimization problem (6.2)

```
input  : $K$, $T$, $N$(steps for conjugate gradient), stepsize $\{\alpha_k, \beta_k\}$, initializations
         $y^0 \in \Delta_n, x^0 \in \mathcal{N}$
for $k = 0, 1, 2, ..., K-1$ do
    Set $x^{k,0} = x^{k-1}$;
    for $t = 0, ..., T-1$ do
        Update $x^{k,t+1} \leftarrow \mathsf{Exp}_{x^{k,t}}(-\beta_k h_g^{k,t})$ with $h_g^{k,t} := \mathsf{grad}_y g(y^k, x^{k,t})$ ;
    end
    Set $x^k \leftarrow x^{k,T}$;
    Update $y^{k+1} \leftarrow \mathsf{proj}_{\Delta_n}(y^k - \alpha_k h_\Phi^k)$ as in (3.12), in the view of Remark 5.1;
end
```

We test our proposed algorithms on two concrete examples that lie in this scope: the robust Karcher mean problem and the robust covariance matrix estimation problems. In both experiments we only consider the deterministic function to test the efficacy of the proposed algorithm framework. In both experiments, we use RieBO (Algorithm 1) while utilizing the trick in Remark 5.1 to estimate the Hessian-vector products.

6.1 Robust Karcher mean problem

For the robust Karcher mean problem, one seeks to solve

$$\begin{aligned} &\min_{y \in \Delta_n} \left\| y - \frac{\mathbf{1}}{n} \right\|^2 - \sum_{i=1}^n y_i \operatorname{dist}(S, A_i)^2 \\ &\text{s.t. } S \in \operatorname*{argmin}_{S \in \mathbb{S}^d_{++}} \sum_{i=1}^n y_i \operatorname{dist}(S, A_i)^2, \end{aligned} \tag{6.3}$$

where A_i's are the symmetric positive definite data matrices, and $\operatorname{dist}(A, B) := \|\log(A^{-1/2} B A^{-1/2})\|_F$ is the geodesic distance of two positive definite matrices (see Bhatia (2009, Chapter 6)). The squared geodesic distance guarantees the geodesic strong convexity of the lower level problem (see Zhang and Sra (2016)), which further ensures that the bilevel problem (6.3) is well-defined. For the function $h(S) := \operatorname{dist}(S, A)^2$, we have the Euclidean and Riemannian gradients as (see Ferreira et al. (2019), Bhatia (2009, Chapter 6)):

$$\begin{aligned} \nabla_S h(S) &= S^{-1/2} \log(S^{1/2} A^{-1} S^{1/2}) S^{-1/2}, \\ \mathsf{grad}_S h(S) &= S \nabla_S h(S) S = S^{1/2} \log(S^{1/2} A^{-1} S^{1/2}) S^{1/2}. \end{aligned}$$

The Euclidean and Riemannian Hessian of $h(S) := \operatorname{dist}(S, A)^2$ are less straightforward to calculate, and to the best of our knowledge, they do not exist in the literature. Here we propose an implementable way to calculate it: first notice that (see Bhatia (2009, Chapter 6))

$$\nabla_S h(S) = S^{-1/2} \log(S^{1/2} A^{-1} S^{1/2}) S^{-1/2} = S^{-1} A^{1/2} \log(A^{-1/2} S A^{-1/2}) A^{-1/2}.$$

For any symmetric matrix V, we have

$$\langle \nabla_S h(S), V \rangle = \operatorname{tr}(V S^{-1} A^{1/2} \log(A^{-1/2} S A^{-1/2}) A^{-1/2}).$$

To take the derivative of this, notice that S appears twice in $\langle \nabla_S h(S), V \rangle$. Denote $\tilde{h}(S_1, S_2) = \operatorname{tr}(V S_1^{-1} A^{1/2} \log(A^{-1/2} S_2 A^{-1/2}) A^{-1/2})$, we know that (see Petersen et al. (2008))

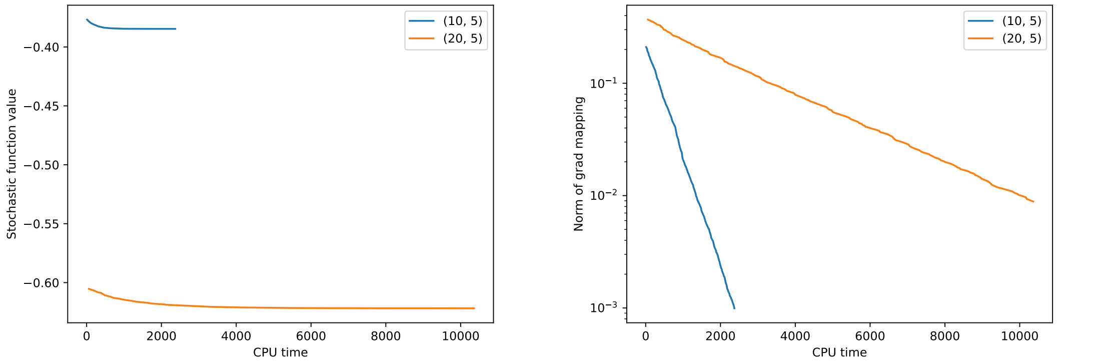
$$\frac{\partial \tilde{h}}{\partial S_1} = -S_1^{-1} V A^{-1/2} \log(A^{-1/2} S_2 A^{-1/2}) A^{1/2} S_1^{-1}.$$

It remains to calculate $\partial \tilde{h} / \partial S_2$, which takes a form of $l(S) := \operatorname{tr}(C \log(PSQ))$. Denote $Y = PSQ$ and $L = \log(Y)$, we have

$$\begin{bmatrix} L & dL \\ 0 & L \end{bmatrix} = \log\left(\begin{bmatrix} Y & dY \\ 0 & Y \end{bmatrix} \right) = \log\left(\begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} S & dS \\ 0 & S \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \right).$$

Therefore, we get

$$dl = \left\langle \begin{bmatrix} 0 & C \\ 0 & 0 \end{bmatrix}, \log\left(\begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} S & dS \\ 0 & S \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \right) \right\rangle,$$



(a) Function value

(b) Norm of the approximate gradient mapping

Figure 1: The convergence curve of applying Algorithm 3 to the robust Karcher mean problem (6.3). The CPU time is in seconds.

where the inner product is simply the Euclidean inner product. We can plug dS as standard Euclidean basis to obtain an representation of dl/dS, which will take $\mathcal{O}(d^2)$ number of times to cover all the entries. Nevertheless, this provides an implementable way to calculate the Euclidean and Riemannian Hessian.

To summarize, the Euclidean and Riemannian Hessian of h can be calculated as follows.

$$\begin{aligned}
&\nabla^2_S h(S)[V] = -S^{-1}VA^{-1/2}\log(A^{-1/2}SA^{-1/2})A^{1/2}S^{-1} + L,\\
&H_S(h(S))[V] = S\nabla^2_S h(S)[V]S + \mathrm{sym}(S\nabla_S h(S)V),
\end{aligned} \tag{6.4}$$

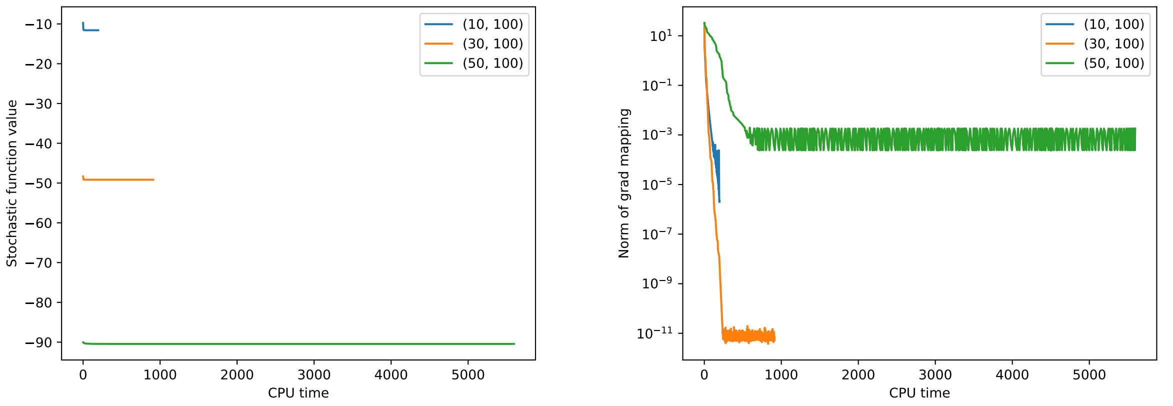
where each entry of matrix L is calculated as follows:

$$L_{i,j} = \left\langle \begin{bmatrix} 0 & C \\ 0 & 0 \end{bmatrix}, \log\left(\begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} S & E_{i,j} \\ 0 & S \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \right) \right\rangle,$$

Here the (i,j)-th entry of $E_{i,j} \in \mathbb{R}^{d\times d}$ is one, and all other entries are zeros. Moreover,

$$\begin{aligned}
P &= A^{-1/2},\\
Q &= A^{-1/2},\\
C &= A^{-1/2}VS^{-1}A^{1/2}.
\end{aligned}$$

In the experiment, we test RieBO (Algorithm 1) with $d \in \{10, 20\}$ and $n = 5$. We repeat each dimension settings for 5 times and plot the average. The algorithm is terminated with $K = 200$ rounds of outer iterations, and the inner iteration is also taken to be $T = 200$ (the value which we observe a good inner iteration convergence). We take $\alpha_k = 10^{-2}$ and $\beta_k = 10^{-1}$. Figure 1 shows the results of the robust Karcher mean problem (6.3). It can be seen from Figure 1 that Algorithm 3 can efficiently decrease both the function values and the norm of gradient mappings. We point out here that the computation of the Riemannian Hessian is time consuming by (6.4) (which is also the reason why we cannot try larger dimensions), yet we remind the reader that this is currently the only formula for calculating it.



(a) Function value

(b) Norm of the approximate gradient mapping

Figure 2: The convergence curve of Algorithm 3 applying to the robust covariance matrix maximum likelihood estimation problem (6.5) with different choice of (d, n). The CPU time is in seconds.

6.2 Robust maximum likelihood estimation

For the robust maximum likelihood estimation of the covariance matrix, one seeks to solve:

$$\begin{aligned} &\min_{y \in \Delta_n} \left\| y - \frac{\mathbf{1}}{n} \right\|^2 - \sum_{i=1}^{n} y_i \mathcal{L}(S; x_i) \\ &\text{s.t. } X \in \operatorname*{argmin}_{S \in \mathbb{S}^d_{++}} \sum_{i=1}^{n} y_i \mathcal{L}(S; x_i), \end{aligned} \tag{6.5}$$

where $\mathcal{L}(S; x)$ is the log likelihood of the Gaussian distribution, namely

$$\mathcal{L}(S; \mathcal{D}) := \frac{1}{2} \operatorname{logdet}(S) + \frac{x^\top S^{-1} x}{2}. \tag{6.6}$$

Note that this lower level problem is geodesically strictly convex (see Sra and Hosseini (2015)), and thus has a unique solution. The calculations of the Riemannian gradient, Hessian-vector product and cross-derivatives all have closed form solutions (following Petersen et al. (2008)).

In the experiment, we test our algorithm with $d \in \{10, 30, 50\}$ and $n = 100$. We repeat each dimension settings for 5 times and plot the average. The algorithm is terminated with $K = 1000$ rounds of outer iterations, and the inner iteration is still taken to be $T = 200$ (again a value which we observe a good inner iteration convergence). We take $\alpha_k = 10^{-2}$ and $\beta_k = 10^{-1}$. Figure 2 shows the results when applying RieBO to the above robust MLE problem with different choices of dimensions. It can be seen from Figure 2 that Algorithm 3 can efficiently decrease both the function values and the norm of gradient mappings. Also, here we are able to test and present the results for a much larger dimension due to much faster calculations of Riemannian gradients, Hessian-vector products and cross-derivatives.

7 Conclusion

We introduced the Riemannian bilevel optimization, a generalization of the traditional Euclidean bilevel optimization. We show that the Riemannian counterparts of Euclidean algorithms in Chen et al. (2021b); Ji et al. (2021) can achieve the same rate of convergence.

Our work raises several open questions. The first is how we can make the convergence independent of the sectional curvature of the manifold, similar to the results in Cai et al. (2023). It is also worth exploring the last iterate convergence of Riemannian bilevel problem. Last, it still needs investigation to see if there are efficient algorithms that can overcome the difficulty we mentioned in the numerical experiment part to efficiently calculate the Riemannian Hessian-vector product thus enabling large-scale implementation of algorithms for solving the Riemannian bilevel optimization problems.

References

T. Bendory, Y. C. Eldar, and N. Boumal. Non-convex phase retrieval from STFT measurements. *IEEE Transactions on Information Theory*, 64(1):467–484, 2017. (Cited on page 4.)

R. Bhatia. *Positive definite matrices*. Princeton university press, 2009. (Cited on page 27.)

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. (Cited on page 4.)

H. Bonnel, L. Todjihoundé, and C. Udrişte. Semivectorial bilevel optimization on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 167(2):464–486, 2015. (Cited on page 4.)

N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. (Cited on pages 5 and 6.)

N. Boumal and P. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in neural information processing systems*, pages 406–414, 2011. (Cited on page 4.)

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2018. (Cited on page 4.)

J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973. (Cited on page 3.)

Y. Cai, M. I. Jordan, T. Lin, A. Oikonomou, and E.-V. Vlatakis-Gkaragkounis. Curvature-independent last-iterate convergence for games on Riemannian manifolds. *arXiv preprint arXiv:2306.16617*, 2023. (Cited on pages 4 and 30.)

L. Chen, J. Xu, and J. Zhang. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023a. (Cited on page 4.)

R. Chen, B. Lucier, Y. Singer, and V. Syrgkanis. Robust optimization for non-convex objectives. *arXiv preprint arXiv:1707.01047*, 2017. (Cited on page 26.)

T. Chen, Y. Sun, and W. Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021a. (Cited on pages 3 and 4.)

T. Chen, Y. Sun, and W. Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. *arXiv preprint arXiv:2106.13781*, 2021b. (Cited on pages 2, 3, 4, 8, 19, and 30.)

X. Chen, M. Huang, and S. Ma. Decentralized bilevel optimization. *arXiv preprint arXiv:2206.05670*, 2022. (Cited on page 4.)

X. Chen, M. Huang, S. Ma, and K. Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International Conference on Machine Learning*, pages 4641–4671. PMLR, 2023b. (Cited on page 4.)

A. Cherian and S. Sra. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE transactions on neural networks and learning systems*, 28(12):2859–2871, 2016. (Cited on page 4.)

C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 4.)

J. Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012. (Cited on page 3.)

Y. Dong, S. Ma, J. Yang, and C. Yin. A single-loop algorithm for decentralized bilevel optimization. *arXiv preprint arXiv:2311.08945*, 2023. (Cited on page 4.)

O. P. Ferreira, M. S. Louzeiro, and L. Prudente. Gradient method for optimization on Riemannian manifolds with lower bounded curvature. *SIAM Journal on Optimization*, 29(4):2517–2541, 2019. (Cited on page 27.)

R. Flamary, A. Rakotomamonjy, and G. Gasso. Learning constrained task similarities in graph regularized multi-task learning. *Regularization, Optimization, Kernels, and Support Vector Machines*, 103, 2014. (Cited on page 1.)

L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018. (Cited on page 3.)

S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018. (Cited on pages 2, 3, 4, and 22.)

S. Gould, B. Fernando, A. Cherian, P. Anderson, R. S. Cruz, and E. Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016. (Cited on page 3.)

R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020. (Cited on pages 3 and 13.)

A. Han, B. Mishra, P. Jawanpuria, P. Kumar, and J. Gao. Riemannian Hamiltonian methods for min-max optimization on manifolds. *SIAM Journal on Optimization*, 33(3):1797–1827, 2023. (Cited on page 6.)

M. Harandi, M. Salzmann, and R. Hartley. Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):48–62, 2017. (Cited on page 4.)

K. Hayami. Convergence of the conjugate gradient method on singular systems. *arXiv preprint arXiv:1809.00793*, 2018. (Cited on page 13.)

M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020. (Cited on pages 1, 2, 3, 4, 8, 19, 22, and 26.)

F. Huang, S. Gao, and H. Huang. Gradient descent ascent for min-max problems on Riemannian manifolds. *arXiv preprint arXiv:2010.06097*, 2020. (Cited on pages 4 and 26.)

K. Ji and Y. Liang. Lower bounds and accelerated algorithms for bilevel optimization. *arXiv preprint arXiv:2102.03926*, 2021. (Cited on page 3.)

K. Ji, J. D. Lee, Y. Liang, and H. V. Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33: 11490–11500, 2020. (Cited on page 1.)

K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021. (Cited on pages 2, 3, 4, 9, and 30.)

C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020. (Cited on page 4.)

H. Kasai, H. Sato, and B. Mishra. Riemannian stochastic recursive gradient algorithm. In *International Conference on Machine Learning*, pages 2516–2524, 2018. (Cited on page 4.)

P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021. (Cited on page 4.)

V. Konda and J. Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999. (Cited on page 1.)

G. Kunapuli, K. P. Bennett, J. Hu, and J.-S. Pang. Classification model selection via bilevel programming. *Optimization Methods & Software*, 23(4):475–489, 2008. (Cited on page 1.)

J. M. Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006. (Cited on pages 5 and 6.)

J. Li, B. Gu, and H. Huang. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020. (Cited on page 3.)

R. Liao, Y. Xiong, E. Fetaya, L. Zhang, K. Yoon, X. Pitkow, R. Urtasun, and R. Zemel. Reviving and improving recurrent back-propagation. In *International Conference on Machine Learning*, pages 3082–3091. PMLR, 2018. (Cited on page 3.)

L. Lin, B. St. Thomas, H. Zhu, and D. B. Dunson. Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association*, 112(519):1261–1273, 2017. (Cited on page 4.)

L. Lin, D. Lazar, B. Sarpabayeva, and D. B. Dunson. Robust optimization and inference on manifolds. *arXiv preprint arXiv:2006.06843*, 2020a. (Cited on page 4.)

T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020b. (Cited on page 4.)

R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning*, pages 6305–6315. PMLR, 2020. (Cited on page 3.)

J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020. (Cited on page 3.)

D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015. (Cited on page 3.)

B. Mishra, H. Kasai, P. Jawanpuria, and A. Saroop. A Riemannian gossip approach to subspace learning on Grassmann manifold. *Machine Learning*, pages 1–21, 2019. (Cited on page 4.)

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020. (Cited on page 4.)

G. M. Moore. *Bilevel programming algorithms for machine learning model selection*. Rensselaer Polytechnic Institute, 2010. (Cited on page 3.)

M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on page 4.)

T. Okuno, A. Takeda, A. Kawana, and M. Watanabe. On lp-hyperparameter learning via bilevel nonsmooth optimization. *Journal of Machine Learning Research*, 22(245):1–47, 2021. (Cited on page 1.)

F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016. (Cited on page 3.)

K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7 (15):510, 2008. (Cited on pages 27 and 29.)

A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019. (Cited on page 1.)

A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019. (Cited on page 3.)

C. Shi, J. Lu, and G. Zhang. An extended Kuhn–Tucker approach for linear bilevel programming. *Applied Mathematics and Computation*, 162(1):51–63, 2005. (Cited on page 3.)

S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015. (Cited on page 29.)

H. v. Stackelberg and A. T. Peacock. The theory of the market economy. *(No Title)*, 1952. (Cited on page 3.)

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere ii: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016. (Cited on page 4.)

J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. (Cited on page 4.)

D. A. Tarzanagh, M. Li, C. Thrampoulidis, and S. Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179. PMLR, 2022. (Cited on page 4.)

N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *Conference On Learning Theory*, pages 650–687, 2018. (Cited on page 4.)

L. W. Tu. *An Introduction to Manifolds*. Springer Science & Universitext, 2011. (Cited on page 5.)

B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. (Cited on page 4.)

M. Weber and S. Sra. Nonconvex stochastic optimization on manifolds via Riemannian Frank-Wolfe methods. *arXiv preprint arXiv:1910.04194*, 2019. (Cited on page 4.)

M. Weber and S. Sra. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *IMA Journal of Numerical Analysis*, 42(4):3241–3271, 2022. (Cited on page 26.)

M. Weber and S. Sra. Riemannian optimization via Frank-Wolfe methods. *Mathematical Programming*, 199(1-2):525–556, 2023. (Cited on page 26.)

Y. Yang, P. Xiao, and K. Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OzjBohmLvE. (Cited on page 4.)

T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, pages 12098–12109. PMLR, 2021. (Cited on page 4.)

T. Yu and H. Zhu. Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*, 2020. (Cited on page 1.)

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638. PMLR, 2016. (Cited on pages 9, 14, 23, and 27.)

H. Zhang, S. J. Reddi, and S. Sra. Fast stochastic optimization on Riemannian manifolds. *ArXiv e-prints*, pages 1–17, 2016. (Cited on pages 4 and 6.)

P. Zhou, X. Yuan, S. Yan, and J. Feng. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 2019. (Cited on page 4.)