Riemannian Bilevel Optimization

Sanchayan Dutta   dutta@ucdavis.edu
UC Davis, Davis, CA, USA

Xiang Cheng   chengx@mit.edu
Massachusetts Institute of Technology, Cambridge, USA

Suvrit Sra   s.sra@tum.de
TU Munich, Garching, Germany

Abstract

We develop new algorithms for Riemannian bilevel optimization. We focus in particular on batch and stochastic gradient-based methods, with the explicit goal of avoiding second-order information such as Riemannian hyper-gradients. We propose and analyze RF^{2}SA, a method that leverages first-order gradient information to navigate the complex geometry of Riemannian manifolds efficiently. Notably, RF^{2}SA is a single-loop algorithm, and thus easier to implement and use. Under various setups, including stochastic optimization, we provide explicit convergence rates for reaching ϵ-stationary points. We also address the challenge of optimizing over Riemannian manifolds with constraints by adjusting the multiplier in the Lagrangian, ensuring convergence to the desired solution without requiring access to second-order derivatives.

1 Introduction

We investigate Riemannian bilevel optimization problems described by:

$$\min_{x\in\mathcal{M}} \ F(x) := f\left(x, y^*(x)\right) \quad \text{s.t.} \quad y^*(x) \in \operatorname{argmin}_{y\in\mathcal{N}} g(x,y), \tag{P}$$

where $\mathcal{M}$ and $\mathcal{N}$ are d_x- and d_y-dimensional complete Riemannian manifolds, respectively, and f and g are smooth functions. The function F serves as the outer objective, with g as the inner objective and $y^*(x)$ as the optimal solution for the inner problem.

Bilevel optimization provides a useful model for hierarchical decision-making, and is thus of great value to various fields such as machine learning, economics, operations research, and engineering. In machine learning it is directly relevant to applications such as meta-learning (Rajeswaran et al., 2019; Hospedales et al., 2021; Pham et al., 2020; Ravi and Larochelle, 2016), hyper-parameter optimization (Franceschi et al., 2018; Bao et al., 2021; Pedregosa, 2016), model selection (Kunapuli et al., 2008; Giovannelli et al., 2021), architecture search (Liu et al., 2018; Wang et al., 2022; Zhang et al., 2021), and reinforcement learning (Konda and Tsitsiklis, 1999; Sutton and Barto, 2018; Hong et al., 2023). Algorithms like the two-timescale stochastic approximation (TTSA) (Hong et al., 2023) highlight the ongoing development of efficient solutions for these problems.

Riemannian optimization arises in several applications, e.g., policy optimization, where algorithms utilize the Fisher Information manifold (Ding et al., 2020; Kakade, 2001; Cen et al., 2022), and matrix factorization, where problems are reformulated over suitable matrix manifolds

(Ahn and Suarez, 2021; Hou et al., 2020; Li et al., 2021). Hyperbolic manifolds have also been applied in neural network architecture design (Peng et al., 2021) and image segmentation (Ghadimi and Wang, 2018).

Toward solving (**P**) we build on recent progress in (Euclidean) bilevel optimization (Kwon et al., 2023), and develop Riemannian "fully first-order" batch and stochastic methods. Indeed, while some aspects of the Euclidean analysis translate directly into the Riemannian setting, curvature causes distortion that poses unique challenges in developing the analysis.

1.1 Related Work

Riemannian optimization

The development of efficient stochastic gradient-based optimization algorithms is crucial across various domains, with first-order methods preferred for their computational efficiency. Several first-order stochastic methods have been adapted to the Riemannian setting, such as Riemannian stochastic gradient descent (RSGD) (Bonnabel, 2013), and AMSGRAD (Bécigneul and Ganea, 2018). Convergence analysis in geodesically convex cases has been proposed (Zhang and Sra, 2016), and nonconvex Riemannian optimization methods have been developed to handle complex landscapes (Kasai et al., 2018; Hu et al., 2024). Saddle-point problems, including geodesic min-max formulations, have also been explored (Zhang et al., 2023).

Bilevel optimization

The motivation for bilevel optimization on manifolds arises from addressing lower-level problems that are non-strongly convex, frequently encountered outside traditional Euclidean spaces. Chen et al. (2023) highlights the necessity for geometry-aware optimization techniques, particularly useful in meta-learning and hyperparameter tuning scenarios. Additional research focuses on constrained bilevel problems in machine learning, where model parameters are defined on specific manifolds (Lin and Zha, 2008; Franceschi et al., 2018; Tabealhojeh et al., 2023). Many of these problems involve non-convex constraints, which existing methods often cannot adequately address. Moreover, even those methods that can handle such constraints may not fully exploit the computational advantages provided by a geometry-aware approach (Beck et al., 2023).

Riemannian bilevel optimization

Li and Ma (2024) investigated hypergradient calculations for bilevel optimization on Riemannian manifolds, proposing deterministic and stochastic algorithms (RieBO and RieSBO). Similarly, Han et al. (2024) developed a framework for bilevel optimization, offering several hypergradient estimation strategies and conducting convergence and complexity analyses. Both works emphasize the importance of hypergradient information. In contrast, our approach is fully first-order, avoiding second-order information like Riemannian hypergradients, presenting a novel direction in this field.

Useful Riemannian bilevel problems often involve manifold constraints, such as machine learning with manifold constraints or robust PCA on the Stiefel manifold (Yao et al., 2024; Hong et al., 2023; Xu and Zhu, 2023; Khanduri et al., 2023; Podosinnikova et al., 2014). Our paper contributes tools for manifold-based bilevel optimization, opening avenues in optimal transport

and geometric ODEs where manifold structures are crucial (Figalli et al., 2011; Udriste and Tevy, 2020).

1.2 Main Contributions

The structure of our analysis parallels the fully-first order Euclidean approach (Kwon et al., 2023). We reformulate the optimization problem (**P**) into a constrained, single-level problem:

$$\min_{x \in \mathcal{M}, y \in \mathcal{N}} f(x,y) \quad \text{s.t.} \quad g(x,y) - g^*(x) \le 0, \tag{P'}$$

where $g^*(x) = g(x, y^*(x))$. The associated Lagrangian $\mathcal{L}_\lambda(x,y) = f(x,y) + \lambda(g(x,y) - g^*(x))$ uses a multiplier $\lambda > 0$. To optimize $\mathcal{L}_\lambda$, we employ Riemannian gradient descent, calculating gradients with first-order derivatives, akin to the Euclidean approach.

The main challenge in addressing (**P'**) is choosing λ. The optimal solution $x^* = \operatorname{argmin}_x F(x)$ is found as $\lambda \to \infty$. However, a high λ makes $\mathcal{L}_\lambda(x,y)$ non-smooth, affecting gradient-descent efficacy.

To address this, we begin with $\lambda = \lambda_0 > 0$ and increment it gradually. Each iteration k sets $\lambda_k = O\left(k^b\right)$ for $b \in (0,1]$, balancing bias removal and nonsmoothness increase rate. This strategy is vital for converging to an ϵ-stationary point of F without needing second derivatives.

One primary contribution is the derivation of an optimal growth rate for λ_k, ensuring non-asymptotic convergence to an ϵ-stationary point of F without needing second-order derivatives. Algorithm RF^{2}SA advances Riemannian bilevel optimization by employing a first-order gradient method that navigates curvature complexities such as varying sectional curvatures, parallel transport, and the geometry of geodesics. These complexities impact the behavior of gradients and necessitate specialized techniques to ensure efficient convergence. Our method converges efficiently to an ϵ-stationary solution, as outlined in Theorem 2.

Theorem 1 (Informal). *There exist choices of hyperparameters of Algorithm* RF^{2}SA *such that the following stationarity guarantees hold:*

1. *If noise is present in* $\operatorname{grad} f$ *and* $\operatorname{grad} g$, *then* $\mathbb{E}\left[\|\operatorname{grad} F(x_K)\|^2\right] = \tilde{O}(K^{-2/7})$;

2. *If noise is present only in* $\operatorname{grad} f$, *then* $\mathbb{E}\left[\|\operatorname{grad} F(x_K)\|^2\right] = \tilde{O}(K^{-2/5})$; *and*

3. *If both* $\operatorname{grad} f$ *and* $\operatorname{grad} g$ *are exact, then* $\mathbb{E}\left[\|\operatorname{grad} F(x_K)\|^2\right] = \tilde{O}(K^{-2/3})$.

Theorem 1 is an informal version of our convergence guarantee; the formal version is Theorem 2.

Below, we summarize the key aspects of our result:

- **Stochastic First-Order Algorithm Without Hypergradient Computations**: Algorithm RF^{2}SA uniquely employs only gradient computations, avoiding the complex hypergradient calculations seen in prior works like (Ghadimi and Wang, 2018). This simplification is especially beneficial in large-scale machine learning tasks, where data completeness is not guaranteed, reducing the need for extensive iterations and streamlining the optimization process in stochastic settings.

- **Convergence Rate Guarantees in Stochastic Scenarios**: The convergence rates for Algorithm RF^{2}SA in stochastic gradient scenarios are $\tilde{O}(\epsilon^{-3.5})$ when both $\operatorname{grad} f$ and $\operatorname{grad} g$ are noisy, improving to $\tilde{O}(\epsilon^{-2.5})$ when only $\operatorname{grad} f$ is noisy, and $\tilde{O}(\epsilon^{-1.5})$ under exact gradients. These rates

are tight and align with those expected in Euclidean optimization, adapted to the complexities of Riemannian manifolds.

- **Achieving ϵ-Stationarity**: The algorithm effectively converges to an ϵ-stationary point, where the norm of the gradient is below ϵ. This capability is critical for assessing the efficacy of optimization algorithms under various gradient noise conditions.

- **Modular Analysis of λ**: Our analysis reveals how different adjustments in λ affect step size, noise variance, and bias, providing insights that help optimize algorithm performance on Riemannian manifolds. This modular approach allows for the strategic modification of λ, enhancing both computational efficiency and algorithm robustness, paving the way for future advancements in manifold-based optimization algorithm design.

2 Mathematical Background

2.1 Some Precepts of Riemannian Geometry

The Hessian of a function f at a point p on a manifold $\mathcal{M}$ with a metric g is defined using the Levi-Civita connection ∇. It is a bilinear form that can be expressed in local coordinates as:

$$\text{Hess}_f(X, Y) = X(Y(f)) - (\nabla_X Y)(f), \tag{2.1}$$

where X and Y are vector fields on $\mathcal{M}$, and $\nabla_X Y$ is the covariant derivative of Y in the direction of X.

The term ∇^2_{xy} can be related to the components of the Hess in local coordinates. Specifically, if X and Y are coordinate vector fields corresponding to coordinates x and y respectively, then $\nabla^2_{xy} f$ would correspond to the (x, y)-component of the Hessian matrix of f, which is:

$$\text{Hess}_f(X, Y) = \nabla_X \nabla_Y f - \nabla_{\nabla_X Y} f \tag{2.2}$$

In local coordinates, this would be written as:

$$\text{Hess}_f(\partial_x, \partial_y) = \frac{\partial^2 f}{\partial x \partial y} - \Gamma^k_{xy} \frac{\partial f}{\partial k} \tag{2.3}$$

where Γ^k_{xy} are the Christoffel symbols of the second kind, which encode the manifold's connection and hence its curvature.

2.2 Main Definitions and Assumptions

Definition 1 (ϵ-stationary point). A point $x \in \mathcal{M}$ is called ϵ-stationary if $\| \text{grad } F(x) \|^2_x \leq \epsilon$. A stochastic algorithm is said to achieve an ϵ-stationary point in K iterations if $\mathbb{E}\left[\| \text{grad } F(x_K) \|^2_{x_K} \right] \leq \epsilon$, where the expectation is over the algorithm's stochasticity.

Notation: $O_P(\cdot)$ denotes the order of constants dependent on instance-specific parameters (e.g., Lipschitz constants, strong convexity, and smoothness conditions). The notation $a_k \asymp b_k$ indicates that a_k and b_k decrease or increase at the same rate as $k \to \infty$, i.e., $\lim_{k \to \infty} a_k / b_k = \Theta(1)$. The norm $\| \cdot \|_x$ is induced by the metric at x, reflecting the geometry of $\mathcal{M}$.

To outline the class of problems (**P**) of interest, we assume the outer-level objective's optimal value on $\mathcal{M}$ is bounded below by $F^* := \arg\min_{x \in \mathcal{M}} F(x) > -\infty$.

Assumption 1 (Objective Functions Properties). *The objective functions f and g exhibit the following properties:*

1. *f is continuously differentiable on $\mathcal{M}$. Its gradient satisfies $l_{f,1}$- smoothness, meaning that for any two points x, y on $\mathcal{M}$,*

$$\|PT_{y \leftarrow x} \nabla f(x) - \nabla f(y)\|_{\mathcal{M}} \leq l_{f,1} d_{\mathcal{M}}(x, y). \tag{2.4}$$

2. *g is continuously differentiable on $\mathcal{N}$. Its gradient satisfies $l_{g,1}$-smoothness, implying,*

$$\|PT_{y \leftarrow x} \nabla g(x) - \nabla g(y)\|_{\mathcal{M}} \leq l_{g,1} d_{\mathcal{M}}(x, y). \tag{2.5}$$

3. *For every $\bar{x} \in \mathcal{M}$, the magnitude of the gradient $\|\nabla_y f(\bar{x}, y)\|_{\mathcal{M}}$ is bounded by $l_{f,0}$ for all y.*

Assumption 2 (Lower-level Objective Properties). *For the lower-level objective g on $\mathcal{N}$:*

1. *For every $\bar{x} \in \mathcal{M}$, the function $g(\bar{x}, y)$ is μ_g-strongly convex in y for some $\mu_g > 0$ on $\mathcal{N}$.*

2. *g is twice continuously differentiable on $\mathcal{N}$, and its Hessian $\nabla^2 g$ satisfies $l_{g,2}$-Lipschitz continuity,*

$$\|PT_{y \leftarrow x}^{\gamma} \nabla^2 g(y) PT_{x \leftarrow y}^{\gamma} - \nabla^2 g(x)\|_{\mathcal{N}} \leq l_{g,2} d_{\mathcal{N}}(x, y). \tag{2.6}$$

Assumption 3 (Gradient Access). *Access to the gradients of the objective functions f and g is provided via unbiased estimators $\operatorname{grad} f(x, y; \zeta)$ and $\operatorname{grad} g(x, y; \phi)$, where:*

$$\begin{aligned} \mathbb{E}[\operatorname{grad} f(x, y; \zeta)] &= \operatorname{grad} f(x, y), \\ \mathbb{E}[\operatorname{grad} g(x, y; \phi)] &= \operatorname{grad} g(x, y), \end{aligned} \tag{2.7}$$

and the variances of the stochastic gradient estimators are bounded:

$$\begin{aligned} \mathbb{E}\left[\|\operatorname{grad} f(x, y; \zeta) - \operatorname{grad} f(x, y)\|_x^2\right] &\leq \sigma_f^2, \\ \mathbb{E}\left[\|\operatorname{grad} g(x, y; \phi) - \operatorname{grad} g(x, y)\|_x^2\right] &\leq \sigma_g^2, \end{aligned} \tag{2.8}$$

where $\|\cdot\|_x$ denotes the norm induced by the metric at point x.

Assumption 4 (Gradient Boundedness). *The gradients with respect to x for f and g are bounded for every $\bar{y}$, with $\|\operatorname{grad}_x f(x, \bar{y})\|$ and $\|\operatorname{grad}_x g(x, \bar{y})\|$ bounded by $l_{f,0}$ and $l_{g,0}$, respectively, for all x.*

Assumption 5 (Second-order Smoothness of f). *f is twice continuously differentiable, with its Hessian $\operatorname{Hess} f$ being $l_{f,2}$-Lipschitz continuous in the sense over the product of the manifold's tangent spaces at (x, y).*

The assumptions 1 through 5 are necessary for guaranteeing the smoothness of $y_\lambda^*(x)$ and the efficacy of the inner iterations throughout all outer iterations. These assumptions align the analysis with the inherent curvature and metrics of $\mathcal{M}$ and $\mathcal{N}$. They are essential for our proof of Theorem 2.

2.3 Computing the Hypergradient via Perturbation Analysis

In this section, we utilize first-order perturbations in the variables x and y to derive the hypergradient $\text{grad}\, F(x)$ of F at x. This formulation of the hyper-gradient is crucial for the proofs of the foundational lemmas and theorems that follow.

Considering an infinitesimal perturbation δv within the tangent space $T_x \mathcal{M}$, we transition to a new manifold point $x' = \text{Exp}_x(\delta v)$. Similarly, perturbing the solution $y^*(x)$ by δu in $T_y \mathcal{N}$ leads to $y' = \text{Exp}_y(\delta u)$.

The first-order Taylor expansion of g around the point $(x, y^*(x))$ is expressed as:

$$\text{grad}_y\, g(x', y^*(x)) \approx \text{grad}_y\, g(x, y^*(x)) + \text{grad}_x(\text{grad}_y\, g)(x, y^*(x))[\delta v]. \tag{2.9}$$

Incorporating the perturbation δu in $T_y \mathcal{N}$, we refine our approximation to:

$$\text{grad}_{y'}\, g(x', y') \approx \text{grad}_y\, g(x', y^*(x)) + \text{Hess}_{yy}\, g(x, y^*(x))[\delta u]. \tag{2.10}$$

To satisfy the optimality condition $\text{grad}_{y'}\, g(x', y') = 0$ for y' as the new minimizer, we establish a linkage between δu and δv:

$$\text{grad}_x(\text{grad}_y\, g)(x', y^*(x))[\delta v] + \text{Hess}_{yy}\, g(x, y^*(x))[\delta u] = 0. \tag{2.11}$$

Solving for δu, we invert the Hessian of g with respect to y, obtaining:

$$[\delta u] = -(\text{Hess}_{yy}\, g(x', y^*(x)))^{-1}\, \text{grad}_x(\text{grad}_y\, g)(x', y^*(x))[\delta v]. \tag{2.12}$$

Finally, the gradient of F at x, influenced by the movements δv and δu, is concisely articulated as:

$$\text{grad}\, F(x) = \text{grad}_x\, f(x, y^*(x)) - \text{Hess}_{xy}\, g(x, y^*(x))(\text{Hess}_{yy}\, g(x, y^*(x)))^{-1}\, \text{grad}_y\, f(x, y^*(x)), \tag{2.13}$$

culminating our systematic approach to compute the hyper-gradient via perturbation analysis.

3 Algorithm Design and Step-Size Calculations

3.1 Algorithm

We devise an algorithm to find a stationary point of the bilevel problem, specifically, a point where $F(x) = f(x, y^*(x))$ is stationary, using gradients of f and g. Considering the formulation (**P'**) and aiming to bypass second-order derivatives, we assess the gradient of $\mathcal{L}_\lambda$:

$$\begin{aligned}
\text{grad}_x\, \mathcal{L}_\lambda(x, y) &= \text{grad}_x\, f(x, y) + \lambda \left(\text{grad}_x\, g(x, y) - \text{grad}\, g^*(x)\right), \\
\text{grad}_y\, \mathcal{L}_\lambda(x, y) &= \text{grad}_y\, f(x, y) + \lambda\, \text{grad}_y\, g(x, y).
\end{aligned} \tag{3.1}$$

On the manifold, the gradient of $g^*(x)$ simplifies to $\text{grad}\, g^*(x) = \text{grad}_x\, g(x, y^*(x))$ due to g's optimality at $y^*(x)$. To optimize $\mathcal{L}_\lambda(x, y)$, we introduce an auxiliary variable z, approximating $y^*(x)$, and consider an alternative bilevel formulation (**P**) with the outer-level objective $\mathcal{L}_\lambda(x', z)$, where $x' = (x, y)$ is the outer variable, and z is the inner variable. This modification alters $F(x)$'s landscape, introducing a bias that must be managed carefully to not affect the function $\mathcal{L}_\lambda$'s smoothness, which is crucial for step-size and noise variance.

To manage the bias, we explore the relation between $\mathcal{L}_\lambda$ and $F(x)$ through an auxiliary function $\mathcal{L}_\lambda^*$ defined as:

$$\mathcal{L}_\lambda^*(x) := \min_y \mathcal{L}_\lambda(x,y). \tag{3.2}$$

For $\lambda > 2l_{f,1}/\mu_g$, $\mathcal{L}_\lambda(x,y)$ becomes strongly convex in y, ensuring a unique minimizer $y_\lambda^*(x)$:

$$y_\lambda^*(x) := \mathrm{argmin}_y \mathcal{L}_\lambda(x,y). \tag{3.3}$$

Given $F(x) = \lim_{\lambda\to\infty} \mathcal{L}_\lambda^*(x)$ for any x in X, $\mathcal{L}_\lambda^*(x)$ effectively approximates $F(x)$ for a sufficiently large λ. This approach is underpinned by a lemma adapted for manifolds.

Algorithm 1 RF^{2}SA- Riemannian First-order Fast Stochastic Approximation

---

1: **Input:** step sizes: $\{\alpha_k, \gamma_k\}$, multiplier difference sequence: $\{\delta_k\}$, inner-loop iteration count: T,
2: step-size ratio: ξ, initializations: λ_0, x_0, y_0, z_0
3: For $k = 0$ to $K - 1$ do
4:      $z_{k,0} \leftarrow z_k$, $y_{k,0} \leftarrow y_k$
5:      For $t = 0$ to $T - 1$ do
6:          $z_{k,t+1} \leftarrow \mathrm{Exp}_{z_{k,t}}(-\gamma_k h_{gz}^{k,t})$
7:          $y_{k,t+1} \leftarrow \mathrm{Exp}_{y_{k,t}}(-\alpha_k(h_{fy}^{k,t} + \lambda_k h_{gy}^{k,t}))$
8:      EndFor
9:      $z_{k+1} \leftarrow z_{k,T}$, $y_{k+1} \leftarrow y_{k,T}$
10:      $x_{k+1} \leftarrow \mathrm{Exp}_{x_k}(-\xi\alpha_k(h_{fx}^k + \lambda_k(h_{gxy}^k - h_{gxz}^k)))$
11:      $\lambda_{k+1} \leftarrow \lambda_k + \delta_k$
12: EndFor

Lemma 1. *For any $x \in X$ and $\lambda \geq 2l_{f,1}/\mu_g$, the gradient of $\mathcal{L}_\lambda^*(x)$ is*

$$\mathrm{grad}_x \mathcal{L}_\lambda(x, y_\lambda(x)) = \mathrm{grad}_x f(x, y_\lambda(x)) + \lambda \left(\mathrm{grad}_x g(x, y_\lambda(x)) - \mathrm{grad}_x g(x, y(x))\right). \tag{3.4}$$

Furthermore, the norm of the difference between the gradients of $F(x)$ and $\mathcal{L}_\lambda^*(x)$ is bounded by

$$|\mathrm{grad}\, F(x) - \mathrm{grad}\, \mathcal{L}_\lambda^*(x)| \leq \frac{C_\lambda}{\lambda}. \tag{3.5}$$

where $C_\lambda := \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right)$.

The gradient $\mathrm{grad}\, \mathcal{L}_\lambda^*(x)$ is computable with first-order derivatives of both f and g. Thus, any first-order method locating a stationary point of $\mathcal{L}_\lambda^*(x)$ approximates the trajectory of x updated with $\mathrm{grad}\, F(x)$, with a bias of $O(1/\lambda)$.

We use $\mathrm{grad}\, \mathcal{L}_\lambda^*(x)$ as a proxy for $\mathrm{grad}\, F(x)$ to produce a sequence of iterates $\{x_k\}$. Concurrently, we generate sequences $\{y_k\}$ and $\{z_k\}$ to approximate the solutions $y_{\lambda_k}^*(x_k)$ and $y^*(x_k)$, respectively, incrementing λ_k with k to ensure the bias in $\{x_k\}$ diminishes to zero.

Our Fully First-order Stochastic Approximation (F^{2}SA) method, adapted for the manifold, employs stochastic gradients as unbiased estimators:

$$
\begin{aligned}
h_{gz}^{k,t} &:= \mathrm{grad}_y g(x_k, z_{k,t}; \phi_z^{k,t}), & h_{fy}^{k,t} &:= \mathrm{grad}_y f(x_k, y_{k,t}; \zeta_y^{k,t}), \\
h_{gy}^{k,t} &:= \mathrm{grad}_y g(x_k, y_{k,t}; \phi_y^{k,t}), & h_{gxy}^k &:= \mathrm{grad}_x g(x_k, y_{k+1}; \phi_{xy}^k), \\
h_{fx}^k &:= \mathrm{grad}_x f(x_k, y_{k+1}; \zeta_x^k), & h_{gxz}^k &:= \mathrm{grad}_x g(x_k, z_{k+1}; \phi_{xz}^k).
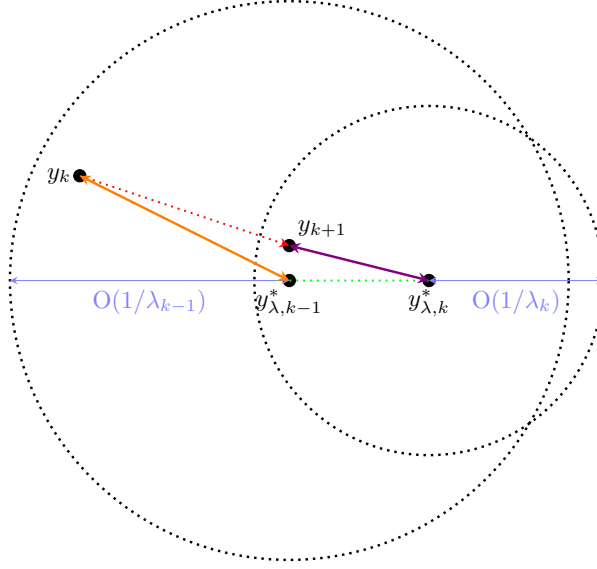\end{aligned} \tag{3.6}
$$

7

Figure 1: y_k should move faster than $y^*_{\lambda_k}(x_k)$, remaining within an $O(1/\lambda_k)$-ball around $y^*_{\lambda_k}(x_k)$.

With $T = 1$ and an appropriate choice of ξ, Algorithm RF^{2}SA enables a fully single-loop update of all variables, tailored to the manifold's geometry. The step-size design for RF^{2}SA, as outlined in Algorithm 1, adapts to this setup.

3.2 Step-Size Design Principle

We tailor the step-sizes for Algorithm 1 to ensure convergence to an ϵ-stationary point of F. This involves meeting several geometric conditions, considering the curvature. For instance, if $\mathcal{L}_{\lambda_k}$ is $(\lambda_k \mu_g/2)$-strongly convex along y's geodesics, then updating $y_{k,t}$ resembles a geodesic contraction towards $y^*_{\lambda,k}$, with a rate of $1 - O(\mu_g \beta_k)$.

Here, $\beta_k = \alpha_k \lambda_k$ is the effective step-size for y_k. We simplify notation by denoting $y^*_{\lambda,k} = \mathrm{Exp}^{-1}_{x_k}(y^*_{\lambda_k}(x_k))$ and $y^*_k = \mathrm{Exp}^{-1}_{x_k}(y^*(x_k))$, where $\mathrm{Exp}^{-1}_{x_k}(\cdot)$ represents the unique inverse of the exponential map at x_k, mapping points in the manifold back to the tangent space at x_k.

For updating x_k, the step-size $\xi \alpha_k$ should decay no slower than $\Omega(1/k)$. The step-size β_k is limited to $O(1/l_{g,1})$, implying a polynomial growth in λ_k with k.

The manifold distance $\mathrm{d}(\cdot, \cdot)$ between x_{k+1} and x_k depends on several factors. Ideally, λ_k's growth rate ensures $\mathrm{d}(y_k, y^*_{\lambda,k})$ is roughly λ_k^{-2}, suggesting λ_k grows inversely to $\beta_k^{1/4}$.

Efficiency in Algorithm 1 relies on how quickly y_k and z_k can track their targets as x_k and λ_k evolve. We will explore how $y^*_\lambda(x)$ adapts to changes in λ and x.

Lemma 2. *For any points x_1, x_2 on a manifold X and multipliers $\lambda_2 \geq \lambda_1 \geq 2l_{f,1}/\mu_g$, the distance between optimal solutions for these multipliers is bounded by*

$$\mathrm{d}(y^*_{\lambda_1}(x_1), y^*_{\lambda_2}(x_2)) \leq \frac{2(\lambda_2 - \lambda_1)}{\lambda_1 \lambda_2} \frac{l_{f,0}}{\mu_g} + l_{\lambda,0} \mathrm{d}(x_2, x_1), \tag{3.7}$$

with some constant $l_{\lambda,0} \leq 3l_{g,1}/\mu_g$.

In the algorithmic setting, it's crucial that y_k's update moves it sufficiently close to the current target $y^*_{\lambda,k}$ each iteration, surpassing the target's movement due to updates in x_k and λ_k. Ideally, in expectation,

$$d(y_{k+1}, y^*_{\lambda,k}) < d(y_k, y^*_{\lambda,k-1}). \tag{3.8}$$

Considering the geometry, the squared distance $d(y_{k+1}, y^*_{\lambda,k})^2$ contracts with T-steps of $1 - O(\mu_g \beta_k)$, starting from y_k, leading to the requirement

$$(1 - O(T\mu_g\beta_k))d(y_k, y^*_{\lambda,k})^2 < d(y_k, y^*_{\lambda,k-1})^2. \tag{3.9}$$

Utilizing Lemma 2 and the geometry, the minimal condition is

$$d(y_{\lambda,k-1}, y_{\lambda,k}) \leq \left(\frac{l_{f,0}}{\mu_g}\right)\left(\frac{\delta_k}{\lambda_k^2}\right) + l_{\lambda,0}d(x_k, x_{k-1}) \quad \leq T\mu_g\beta_k d(y_k, y^*_{\lambda,k-1}). \tag{3.10}$$

The rate at which $d(y_{k+1}, y^*_{\lambda,k})$ decreases must surpass λ_k^{-1}, maintaining controlled bias in x_k updates. Additionally, $d(x_k, x_{k-1})$ should align with $\xi\beta_k d(y_k, y^*_{\lambda,k-1})$.

Two key conditions emerge:

$$\frac{\delta_k}{\lambda_k} \leq O_P(1) \cdot \beta_k, \quad \frac{\xi}{T} < O_P(1), \tag{3.11}$$

where $O_P(1)$ denotes constants dependent on the problem instance. If λ_k increases polynomially, then $\delta_k/\lambda_k = O(1/k)$, satisfying the first condition if $\beta_k = \Omega(1/k)$. The second condition concerns the inner iterations T needed per outer iteration, allowing for a single-loop algorithm with $T = 1$ and an adequately small ξ, or setting $\xi = 1$ and adjusting $T > 1$ for specific instance parameters.

4 Non-Asymptotic Convergence Analysis

We discussed a number of assumptions on the regularity of the optimization objective and the underlying manifold in Section 2.2. We now present our main convergence result in Theorem 2. Corollary 3 provides explicit iteration complexity bounds for each setting, further elucidating the efficiency and applicability of our results in various contexts.

Convergence Analysis Results

Theorem 2 (Alexandrov Space Version). *Given that the assumptions from Section 2.2 hold within an Alexandrov space with curvature bounded by κ, analogous to manifolds $\mathcal{M}, \mathcal{N}$, and assuming appropriate selection of parameters and step-sizes such that $\lambda_0 \geq 2l_{f,1}/\mu_g$ and*

$$\beta_k \leq \gamma_k \leq \min\left(\frac{1}{4l_{g,1}}, \frac{1}{4T\mu_g}\right), \alpha_k \leq \min\left(\frac{1}{8l_{f,1}}, \frac{1}{2\xi l_{F,1}}\right), \tag{4.1a}$$

$$\frac{\xi}{T} < c_\xi\mu_g \cdot \max\left(l_{g,1}l_{*,0}^2, l_{*,1}\sqrt{\mathcal{M}}\right)^{-1}, \frac{\delta_k}{\lambda_k} \leq \frac{T\mu_g\beta_k}{16}. \tag{4.1b}$$

9

for all $k \geq 0$, with c_ξ being a suitably chosen constant. Then, over $K \geq 1$ iterations within the Alexandrov space—a generalization allowing curvature bounds without necessitating smoothness—the outcomes for Algorithm 1 adhere to

$$\sum_{k=0}^{K-1} \xi \alpha_k \mathbb{E}\left[\|\text{grad } F(x_k)\|_{x_k}^2\right] \leq O_P(1) \cdot \sum_k \xi \alpha_k \lambda_k^{-2} + O_P\left(\sigma_f^2\right) \cdot \sum_k \alpha_k^2 \lambda_k + O_P\left(\sigma_g^2\right) \cdot \sum_k \gamma_k^2 \lambda_k + O_P(1),$$

(4.2)

where $\text{grad } F(x)$ symbolizes a generalized notion of gradient in Alexandrov spaces, and $\|\cdot\|_{x_k}$ signifies the distance measure at point x_k, aligning with the space's metric structure.

The proof of Theorem 2 is deferred to Appendix B. Therein we also explain why the effect of the curvature κ is negligible on the final inequality 4.2.

Our analysis examines the expected decrease of the potential function $\mathbb{V}_k$, defined as

$$\mathbb{V}_k := (F(x_k) - F^*) + l_{g,1} \lambda_k d\left(y_k, y_{\lambda_k}^*(x_k)\right)^2 + \frac{\lambda_k l_{g,1}}{2} d(z_k, y^*(x_k))^2,$$

(4.3)

where F^* is the minimum value of F, and y_λ^* and y^* correspond to solutions. Monitoring the distance between y_k and $y_{\lambda_k}^*(x_k)$ is essential for computing the true gradient of F at x_k using only gradients. The proof will also show that the correct scaling factor for these errors is proportional to λ_k.

For step-size design, we maintain conditions similar to 4.1a for gradient-based methods. The conditions 4.1b address the double-loop nature of the problem. Aligning with the step-size design rule (3), we propose:

$$T = \max\left(32, \left(c_\xi \mu_g\right)^{-1} \max\left(l_{g,1} l_{*,0}^2, \sqrt{M} l_{*,1}\right)\right),$$

$$\xi = 1, \quad \alpha_k = \frac{c_\alpha}{(k+k_0)^a}, \quad \gamma_k = \frac{c_\gamma}{(k+k_0)^c},$$

(4.4)

and for the Lagrange multiplier increase sequence $\{\delta_k\}$,

$$\delta_k = \min\left(\frac{T\mu_g}{16} \alpha_k \lambda_k^2, \frac{\gamma_k}{2\alpha_k} - \lambda_k\right).$$

(4.5)

Rate constants $a, c \in [0,1]$ with $a \geq c$, the initial value of the Lagrange multiplier λ_0, and constants for the context are established as:

$$k_0 \geq \frac{4}{\mu_g} \max\left(\frac{\xi l_{F,1}}{2}, T l_{g,1}, l_{f,1}\right), \lambda_0 \geq \frac{2 l_{f,1}}{\mu_g},$$

$$c_\gamma = \frac{1}{\mu_g k_0^{1-c}}, \quad c_\alpha = \frac{1}{2\lambda_0 \mu_g k_0^{1-a}}.$$

(4.6)

These specifications streamline convergence rate analysis, with the framework accommodating various other choices that comply with conditions 4.1a and 4.1b, facilitating the delineation of the convergence rate across different stochastic noise regimes.

In the following corollary, we present a more interpretable version of Theorem 2, in terms of the iteration complexity guarantees under different settings. This interpretation allows for a clearer understanding of how our theoretical findings translate into practical implications for convergence rates in various scenarios.

10

Corollary 3. *Assume the stipulations of Theorem 2 are upheld, with step-sizes delineated as in equations 4.4, 4.5, and 4.6. Let R signify a random variable uniformly distributed over $\{0, \ldots, K-1\}$. Under these premises, after K iterations, the ensuing convergence outcomes are derived:*

(a) *In the presence of stochastic noise in both objectives f and g ($\sigma_f^2, \sigma_g^2 > 0$), setting $a = 5/7$ and $c = 4/7$, we achieve a convergence rate of $\mathbb{E}\left[\|\operatorname{grad} F(x_R)\|^2\right] \asymp \frac{\log K}{K^{2/7}}$.*

(b) *If stochastic noise is solely in f ($\sigma_f^2 > 0$, $\sigma_g^2 = 0$), setting $a = 3/5$ and $c = 2/5$, we attain $\mathbb{E}\left[\|\operatorname{grad} F(x_R)\|^2\right] \asymp \frac{\log K}{K^{2/5}}$.*

(c) *In scenarios with exact gradients ($\sigma_f^2 = \sigma_g^2 = 0$), appointing $a = 1/3$ and $c = 0$, it follows that $\|\operatorname{grad} F(x_K)\|^2 \asymp \frac{\log K}{K^{2/3}}$.*

These findings illustrate that convergence rates improve when stochastic noise affects fewer components of the problem. Specifically, the rate improves from $O(k^{-2/7})$ to $O(k^{-2/5})$ with noise only in f, and to $O(k^{-2/3})$ in fully deterministic contexts. This compares to the $O(k^{-1})$ rates that can be obtained by second-order methods as in Li and Ma (2024); Han et al. (2024).

5 Limitations and Future Work

In general, fully first-order stochastic algorithms, although competitive with their second-order counterparts exhibit certain limitations like higher iteration complexity (Kwon et al., 2023). This gap highlights the need for further investigation into the theoretical limits of first-order methods with respect to second-order methods for Riemannian bilevel optimization. Additionally, RF²SA's application is predominantly restricted to well-conditioned lower-level problems. It remains open to study RF²SA's potential in a broader range of problem classes.

Future research directions include utilizing our framework to address a wide array of real-world applications, where the natural settings of problems involve varying geometric structures at different decision levels. An exciting future direction is to explore applications of our algorithm in hyperparameter optimization, meta-learning, and reinforcement learning, where manifold optimization ideas have provided significant advantages. (Jaquier and Rozo, 2020; Tabealhojeh et al., 2023; Xu et al., 2016; Jaquier et al., 2020)

Investigating two-player games with states represented as matrices or other manifold-valued objects could further enrich the bilevel optimization landscape, offering novel insights into game theory and decision-making processes on complex geometries. Moreover, the integration of operator-valued optimization tasks and the development of algorithms that consider the manifold's curvature effects more explicitly would refine our understanding and application of Riemannian optimization techniques. (Domingo-Enrich et al., 2020; Cai et al., 2023)

Moreover, in general, the concept of "bilevel" formulations is becoming increasingly significant in the context of Riemannian problems, where many issues seem to naturally incorporate a two-tier optimization process. One pertinent example is the k-sparse barycenter problem (e.g., in Do et al. (2023)), where the goal is to approximate a covariance matrix X (represented as an ellipsoid) using a sparse combination of given covariance matrices $A_1, \ldots, A_N$. Specifically, one aims to find a k-sparse weight vector $q(X)$ that minimizes the distance between X and the Wasserstein barycenter of the selected matrices. Formally, $q(X) := \arg\min_{q \in \mathcal{Q}} \operatorname{dist}^2(X, \operatorname{BaryCenter}(q, A_1, \ldots, A_N))$, where $\mathcal{Q} = \{q \in \mathbb{R}^N \mid q \geq 0, \|q\|_0 \leq k, \sum_{i=1}^N q_i = 1\}$. The barycenter is computed as

BaryCenter$(q, A_1, \ldots, A_N) = \arg\min_{Y \in \mathcal{M}} \sum_{i=1}^{N} q_i \text{dist}^2(Y, A_i)$, where $\mathcal{M}$ is the manifold of symmetric positive definite matrices. Consequently, $X \approx \text{BaryCenter}(q(X), A_1, \ldots, A_N)$, with X approximated using at most k of the covariance matrices $A_1, \ldots, A_N$.

6 Conclusion

We have presented a novel and fully first-order approach to Riemannian bilevel optimization. This opens new avenues for addressing non-strongly convex lower-level problems and provides a geometrically aware framework for complex optimization challenges involving manifold constraints.

References

Kwangjun Ahn and Felipe Suarez. Riemannian perspective on matrix factorization. *arXiv preprint arXiv:2102.00937*, 2021.

Fan Bao, Guoqiang Wu, Chongxuan Li, Jun Zhu, and Bo Zhang. Stability and generalization of bilevel programming in hyperparameter optimization. *Advances in neural information processing systems*, 34:4529–4541, 2021.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018.

Yasmine Beck, Daniel Bienstock, Martin Schmidt, and Johannes Thürauf. On a computationally ill-behaved bilevel problem with a continuous and nonconvex lower level. *Journal of Optimization Theory and Applications*, 198(1):428–447, 2023.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Yang Cai, Michael I Jordan, Tianyi Lin, Argyris Oikonomou, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. Curvature-independent last-iterate convergence for games on riemannian manifolds. *arXiv preprint arXiv:2306.16617*, 2023.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Minh-Hieu Do, Jean Feydy, and Olga Mula. Approximation and structured prediction with sparse wasserstein barycenters. *arXiv preprint arXiv:2302.05356*, 2023.

Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. *Advances in neural information processing systems*, 33:20215–20226, 2020.

Alessio Figalli, Ludovic Rifford, and Cédric Villani. Necessary and sufficient conditions for continuity of optimal transport maps on riemannian manifolds. *Tohoku Mathematical Journal, Second Series*, 63(4):855–876, 2011.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pages 1568–1577. PMLR, 2018.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Tommaso Giovannelli, Griffin Kent, and Luis Nunes Vicente. Bilevel stochastic methods for optimization and machine learning: Bilevel stochastic descent and darts. *arXiv preprint arXiv:2110.00604*, 2021.

Andi Han, Bamdev Mishra, Pratik Jawanpuria, and Akiko Takeda. A framework for bilevel optimization on riemannian manifolds. *arXiv preprint arXiv:2402.03883*, 2024.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Fast global convergence for low-rank matrix recovery via riemannian gradient descent with random initialization. *arXiv preprint arXiv:2012.15467*, 2020.

Jiang Hu, Ruicheng Ao, Anthony Man-Cho So, Minghan Yang, and Zaiwen Wen. Riemannian natural gradient methods. *SIAM Journal on Scientific Computing*, 46(1):A204–A231, 2024.

Noémie Jaquier and Leonel Rozo. High-dimensional bayesian optimization via nested riemannian manifolds. *Advances in Neural Information Processing Systems*, 33:20939–20951, 2020.

Noémie Jaquier, Leonel Rozo, Sylvain Calinon, and Mathias Bürger. Bayesian optimization meets riemannian manifolds in robot learning. In *Conference on Robot Learning*, pages 233–246. PMLR, 2020.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic recursive gradient algorithm. In *International conference on machine learning*, pages 2516–2524. PMLR, 2018.

Prashant Khanduri, Ioannis Tsaknakis, Yihua Zhang, Jia Liu, Sijia Liu, Jiawei Zhang, and Mingyi Hong. Linearly constrained bilevel optimization: A smoothed implicit gradient approach. In *International Conference on Machine Learning*, pages 16291–16325. PMLR, 2023.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Gautam Kunapuli, K Bennett, Jing Hu, and Jong-Shi Pang. Bilevel model selection for support vector machines. In *CRM proceedings and lecture notes*, volume 45, pages 129–158, 2008.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages

18083–18113. PMLR, 2023.

Jiaxiang Li and Shiqian Ma. Riemannian bilevel optimization. *arXiv preprint arXiv:2402.02019*, 2024.

Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man-Cho So. Weakly convex optimization over stiefel manifold using riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

Tong Lin and Hongbin Zha. Riemannian manifold learning. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):796–809, 2008.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 44 (12):10023–10044, 2021.

Quang Pham, Chenghao Liu, Doyen Sahoo, and HOI Steven. Contextual transformation networks for online continual learning. In *International Conference on Learning Representations*, 2020.

Anastasia Podosinnikova, Simon Setzer, and Matthias Hein. Robust pca: Optimization of the robust reconstruction error over the stiefel manifold. In *German Conference on Pattern Recognition*, pages 121–131. Springer, 2014.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Hadi Tabealhojeh, Peyman Adibi, Hossein Karshenas, Soumava Kumar Roy, and Mehrtash Harandi. Rmaml: Riemannian meta-learning with orthogonality constraints. *Pattern Recognition*, 140:109563, 2023.

Constantin Udriste and Ionel Tevy. Geometric dynamics on riemannian manifolds. *Mathematics*, 8 (1):79, 2020.

Xiaoxing Wang, Wenxuan Guo, Jianlin Su, Xiaokang Yang, and Junchi Yan. Zarts: On zero-order optimization for neural architecture search. *Advances in Neural Information Processing Systems*, 35: 12868–12880, 2022.

Siyuan Xu and Minghui Zhu. Efficient gradient approximation method for constrained bilevel optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12509–12517, 2023.

Xin Xu, Zhenhua Huang, Lei Zuo, and Haibo He. Manifold-based reinforcement learning via locally linear reconstruction. *IEEE transactions on neural networks and learning systems*, 28(4): 934–947, 2016.

Wei Yao, Chengming Yu, Shangzhi Zeng, and Jin Zhang. Constrained bi-level optimization:

Proximal lagrangian value function approach and hessian-free algorithm. *arXiv preprint arXiv:2401.16164*, 2024.

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

Miao Zhang, Steven W Su, Shirui Pan, Xiaojun Chang, Ehsan M Abbasnejad, and Reza Haffari. idarts: Differentiable architecture search with stochastic implicit gradients. In *International Conference on Machine Learning*, pages 12557–12566. PMLR, 2021.

Peiyuan Zhang, Jingzhao Zhang, and Suvrit Sra. Sion's minimax theorem in geodesic metric spaces and a riemannian extragradient algorithm. *SIAM Journal on Optimization*, 33(4):2885–2908, 2023.

Appendix / Supplemental Material

Symbol	Meaning	Less than
$l_{f,0}$	Bound of $\|\nabla_x f\|, \|\nabla_y f\|$	.
$l_{f,1}$	Smoothness of f	.
$l_{g,0}$	Bound of $\|\nabla_x g\|$	.
$l_{g,1}$	Smoothness of g	.
μ_g	Strong-convexity of g	.
$l_{g,2}$	Hessian-continuity of g	.
M_f	Second-order moment of $\nabla f(x,y;\zeta)$	$l_{f,0}^2 + \sigma_f^2$
M_g	Second-order moment of $\nabla g(x,y;\phi)$	$l_{g,0}^2 + \sigma_g^2$
$l_{f,2}$	Hessian-continuity of f	.
$l_{F,1}$	Smoothness of $F(x)$	$l_{*,0}\left(l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{2l_{f,0}l_{g,1}l_{g,2}}{\mu_g^2}\right)$
$l_{\lambda,0}$	Lipschitzness of $y_\lambda^*(x)$ (for all $\lambda \geq 2l_{f,1}/\mu_g$)	$\frac{3l_{g,1}}{\mu_g}$
$l_{\lambda,1}$	Smoothness of $y_\lambda^*(x)$ (for $\lambda \geq 2l_{f,1}/\mu_g$)	$32(l_{g,2} + \lambda^{-1} \cdot l_{f,2})\frac{l_{g,1}^2}{\mu_g^3}$
$l_{*,0}$	$= 1 + \max_{\lambda \geq 2l_{f,1}/\mu_g} l_{\lambda,0}$	.
$l_{*,1}$	$= \max_{\lambda \geq 2l_{f,1}/\mu_g} l_{\lambda,1}$	.

Table 1: Meaning of Constants

In Table 1, we list the main symbols used in the following proofs, their interpretations, and the inequalities that they satisfy (where applicable).

To simplify the representation of variable movements, we define q_k^x, $q_{k,t}^y$, and $q_{k,t}^z$ as follows:

$$q_k^x := \operatorname{grad}_x f(x_k, y_{k+1}) + \lambda_k(\operatorname{grad}_x g(x_k, y_{k+1}) - \operatorname{grad}_x g(x_k, z_{k+1})),$$
$$q_{k,t}^y := \operatorname{grad}_y f(x_k, y_{k,t}) + \lambda_k \operatorname{grad}_y g(x_k, y_{k,t}),$$
$$q_{k,t}^z := \operatorname{grad}_y g(x_k, z_{k,t}).$$

These quantities represent the expected movements of x_k, $y_k^{(t)}$, and $z_k^{(t)}$ in the absence of stochastic noise in the gradient oracles.

A   Detailed Proofs of Lemmas 1 and 2

A.1   Lemma A.1

This lemma establishes a bound on the difference between the gradient of a function at two points x_2 and x_1, taking into account the effects of parallel transport.

$F(x) = f(x, y^*(x))$ is $l_{F,1}$-smooth where

$$l_{F,1} \leq l_{*,0}\left(l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{2l_{f,0}l_{g,1}l_{g,2}}{\mu_g^2}\right).$$

Proof. We recall from equation (2.13) that the gradient $\operatorname{grad} F(x)$ of the function F, defined in (**P**), is given by the expression:

$$\operatorname{grad} F(x) = \operatorname{grad}_x f(x, y^*(x)) - \operatorname{Hess}_{xy} g(x, y^*(x)) \left(\operatorname{Hess}_{yy} g(x, y^*(x))\right)^{-1} \operatorname{grad}_y f(x, y^*(x)),$$

where $y^*(x)$ denotes the solution to the inner-level optimization problem associated with x.

The bound for the difference between the parallel transported gradient at x_2 to x_1 and the gradient at x_1 is given by:

$$\|\operatorname{PT}_{x_2 \to x_1} \operatorname{grad} F(x_2) - \operatorname{grad} F(x_1)\|$$
$$\leq \left(l_{f,1} + \frac{l_{f,0}}{\mu_g} l_{g,2} + \frac{l_{g,1}}{\mu_g} l_{g,1}\right) \left(\|d_{\mathcal{M}}(x_1, x_2)\| + \|d_{\mathcal{N}}(y^*(x_1), y^*(x_2))\|\right)$$
$$+ l_{g,1} l_{f,0} \left\|\operatorname{PT}_{x_2 \to x_1} \left(\nabla_{yy}^2 g(x_2, y^*(x_2))^{-1}\right) \operatorname{PT}_{x_1 \to x_2} - \nabla_{yy}^2 g(x_1, y^*(x_1))^{-1}\right\|.$$

To simplify this inequality, we employ the assumptions on the smoothness and strong convexity of the functions f and g, alongside the triangle inequality. From our assumptions in Section 2.2, we require in particular:

1. **Smoothness of f:** f is $l_{f,1}$-smooth, which provides a bound on the gradient differences of f at two points.

2. **Smoothness of g:** g is $l_{g,1}$-smooth, enabling us to bound the gradient differences of g.

3. **Strong Convexity of g:** The μ_g-strong convexity of g facilitates relating the Hessian of g to its inverse, critical for bounding the differences in Hessian inverses.

4. **Lipschitz Continuity of the Hessian of g:** The $l_{g,2}$-Lipschitz continuity of the Hessian of g aids in bounding the differences in Hessians at two points.

$$\|\operatorname{PT}_{x_2 \to x_1} \operatorname{grad} F(x_2) - \operatorname{grad} F(x_1)\|$$
$$\leq \|\operatorname{PT}_{x_2 \to x_1} \operatorname{grad} f(x_2, y^*(x_2)) - \operatorname{grad} f(x_1, y^*(x_1))\|$$
$$+ \|\operatorname{PT}_{x_2 \to x_1} \operatorname{grad} g(x_2, y^*(x_2)) - \operatorname{grad} g(x_1, y^*(x_1))\|$$
$$\cdot \max_x \|\operatorname{Hess}_{yy} g(x, y^*(x))^{-1}\| \cdot \max_x \|\operatorname{grad}_y f(x, y^*(x))\|$$
$$+ \|\operatorname{PT}_{x_2 \to x_1} (\operatorname{Hess}_{xy} g(x_2, y^*(x_2)) \operatorname{PT}_{x_1 \to x_2} - \operatorname{Hess}_{xy} g(x_1, y^*(x_1)))\|$$
$$\cdot \max_x \|\operatorname{Hess}_{yy} g(x, y^*(x))^{-1}\| \cdot \max_x \|\operatorname{grad}_y f(x, y^*(x))\|$$
$$+ \max_x \|\operatorname{Hess}_{xy} g(x, y^*(x))\| \cdot \max_x \|\operatorname{grad}_y f(x, y^*(x))\|$$
$$\cdot \|\operatorname{PT}_{x_2 \to x_1} \operatorname{Hess}_{yy}^{-1}(x_2, y^*(x_2)) \operatorname{PT}_{x_1 \to x_2} - \operatorname{Hess}_{yy}^{-1}(x_1, y^*(x_1))\|$$
$$\leq (l_{f_1} + \frac{l_{f_0}}{\mu_g} l_{g_2} + \frac{l_{g_1}}{\mu_g} l_{f_1})(d_{\mathcal{M}}(x_1, x_2) + d_{\mathcal{N}}(y^*(x_1), y^*(x_2)))$$
$$+ \frac{l_{g_1} l_{f_0}}{\mu_g^2} l_{g,2}(d_{\mathcal{M}}(x_1, x_2) + d_{\mathcal{N}}(y^*(x_1), y^*(x_2))).$$

where to upper bound $\|\operatorname{PT}_{x_2 \to x_1} \operatorname{Hess}_{yy}^{-1}(x_2, y^*(x_2)) \operatorname{PT}_{x_1 \to x_2} - \operatorname{Hess}_{yy}^{-1}(x_1, y^*(x_1))\|$ we used the following result on bounding the norm of the difference of the inverses of two matrices, A and

B, leveraging the Neumann series to express the inverse of a matrix in terms of its perturbation. Specifically, we have:

$$\|A^{-1} - B^{-1}\| \le \|\Delta A\| \cdot \|A^{-1}\| \cdot \|B^{-1}\|,$$

where $\Delta A = A - B$, which is instrumental in establishing the final bound on the gradient difference.

The lemma concludes with:

$$l_{F,1} \le l_{*,0} \left(l_{f,1} + \frac{l_{f,0} l_{g,2} + l_{g,1}^2}{\mu_g} + \frac{l_{f,0} l_{g,1} l_{g,2}}{\mu_g^2} \right)$$
$$\le l_{*,0} \left(l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{2 l_{f,0} l_{g,1} l_{g,2}}{\mu_g^2} \right),$$

where the last inequality utilizes the condition that $l_{g,1}/\mu_g \ge 1$. $\qquad\square$

A.2 Lemma A.2

This lemma establishes a bound on the difference between the gradient of a function $F(x)$ and the gradient of a Lagrangian $\mathcal{L}_\lambda(x,y)$ with respect to x, adjusted for the effects of parallel transport.

For any x, y, λ, the following holds:

$$\left\| \mathrm{grad} F(x) - \mathrm{grad}_x \mathcal{L}_\lambda(x,y) + \mathrm{PT}_{y \to y^*} \mathrm{Hess}_{xy} g(x,y^*) \left(\mathrm{Hess}_{yy} g(x,y^*) \right)^{-1} \mathrm{PT}_{y^* \to y} \mathrm{grad}_y \mathcal{L}(x,y) \right\|$$
$$\le 2 \left(\frac{l_{g,1}}{\mu_g} \right) d_{\mathcal{M}}(y,y^*) \left(l_{f,1} + \lambda \cdot \min \left(2 l_{g,1}, l_{g,2} d_{\mathcal{N}}(y,y^*) \right) \right).$$

Proof. Given the Lagrangian $\mathcal{L}_\lambda(x,y)$, we consider the gradients with respect to variables x and y, expressed as:

$$\mathrm{grad}_x \mathcal{L}_\lambda(x,y) = \mathrm{grad}_x f(x,y) + \lambda \left(\mathrm{grad}_x g(x,y) - \mathrm{PT}_{y^*(x) \to y} \mathrm{grad}_x g(x,y^*(x)) \right),$$
$$\mathrm{grad}_y \mathcal{L}_\lambda(x,y) = \mathrm{grad}_y f(x,y) + \lambda \, \mathrm{grad}_y g(x,y).$$

Here, $\mathrm{PT}_{y^*(x) \to y}$ denotes the parallel transport operation that moves vectors along geodesics from the tangent space at $y^*(x)$ to the tangent space at y, ensuring that the comparison of vectors is meaningful.

The discrepancy between the gradient of F and the gradient of the Lagrangian with respect to x is detailed as follows:

$$\mathrm{grad}\, F(x) - \mathrm{grad}_x \mathcal{L}_\lambda(x,y) = \mathrm{grad}_x f(x,y^*) - \mathrm{PT}_{y \to y^*} \mathrm{grad}_x f(x,y)$$
$$- \mathrm{Hess}_{xy} g(x,y^*) \left(\mathrm{Hess}_{yy} g(x,y^*)^{-1} \mathrm{grad}_y f(x,y) \right)$$
$$- \lambda \left(\mathrm{grad}_x g(x,y) - \mathrm{PT}_{y^* \to y} \mathrm{grad}_x g(x,y^*) \right).$$

We can rearrange terms for $\mathrm{grad}_x g(x,y) - \mathrm{PT}_{y^* \to y} \mathrm{grad}_x g(x,y^*)$ as the following:

$$\mathrm{grad}_x g(x,y) - \mathrm{PT}_{y^* \to y} \mathrm{grad}_x g(x,y^*) = \mathrm{grad}_x g(x,y) - \mathrm{PT}_{y^* \to y} \mathrm{grad}_x g(x,y^*)$$
$$- \mathrm{PT}_{y^* \to y} \mathrm{Hess}_{xy} g(x,y^*) \mathrm{PT}_{y \to y^*} \mathrm{Exp}_y^{-1}(y^*)$$
$$+ \mathrm{PT}_{y^* \to y} \mathrm{Hess}_{xy} g(x,y^*) \mathrm{PT}_{y \to y^*} \mathrm{Exp}_y^{-1}(y^*).$$

Note that from the optimality condition for y^*, we have $\text{grad}_y g(x,y^*) = 0$. From the gradient of the Lagrangian $\mathcal{L}$, we have $\text{grad}_y \mathcal{L}(x,y) = \text{grad}_y f(x,y) + \lambda \text{grad}_y g(x,y)$. We can express the equivalent of $y - y^*$ using the inverse exponential map and the Hessian as follows:

$$\text{Exp}_y^{-1}(y^*) = -\left(\text{Hess}_{yy} g(x,y^*)\right)^{-1} \left(\text{grad}_y g(x,y) - \text{PT}_{y^* \to y} \text{grad}_y g(x,y^*)\right.$$

$$-\text{PT}_{y^* \to y} \text{Hess}_{yy} g(x,y^*) \text{PT}_{y \to y^*} \text{Exp}_y^{-1}(y^*)\Big)$$

$$+ \frac{1}{\lambda} \left(\text{Hess}_{yy} g(x,y^*)\right)^{-1} \left(\text{grad}_y \mathcal{L}(x,y) - \text{grad}_y f(x,y)\right).$$

This approximation is based on the Taylor expansion in the setting, where $\text{grad}_y g(x,y^*) = 0$ because y^* is an optimal point (assuming g is minimized at y^* with respect to y).

$$\text{grad}_y g(x,y) - \text{PT}_{y^* \to y} \text{grad}_y g(x,y^*) \approx \text{PT}_{y^* \to y} \text{Hess}_{yy} g(x,y^*) \text{PT}_{y \to y^*} \text{Exp}_y^{-1}(y)$$

$\text{grad} F(x) - \text{grad}_x \mathcal{L}_\lambda(x,y)$

$$= (\text{grad}_x f(x,y^*) - \text{grad}_x f(x,y))$$

$$- \text{Hess}_{xy} g(x,y^*) \left(\text{Hess}_{yy} g(x,y^*)\right)^{-1} \left(\text{grad}_y f(x,y^*) - \text{PT}_{y \to y^*} \text{grad}_y f(x,y)\right)$$

$$- \text{Hess}_{xy} g(x,y^*) \left(\text{Hess}_{yy} g(x,y^*)\right)^{-1} \text{PT}_{y \to y^*} \text{grad}_y \mathcal{L}(x,y)$$

$$- \lambda \left(\text{grad}_x g(x,y) - \text{PT}_{y^* \to y} \text{grad}_x g(x,y^*) - \text{PT}_{y^* \to y} \text{Hess}_{xy} g(x,y^*) \text{Exp}_y^{-1}(y^*)\right)$$

$$+ \lambda \text{Hess}_{xy} g(x,y^*) \left(\text{Hess}_{yy} g(x,y^*)\right)^{-1} \left(\text{grad}_y g(x,y) - \text{PT}_{y^* \to y} \text{grad}_y g(x,y^*)\right.$$

$$\left.- \text{PT}_{y^* \to y} \text{Hess}_{yy} g(x,y^*) \text{PT}_{y \to y^*} \text{Exp}_y^{-1}(y^*)\right).$$

To simplify this, we will require the following facts:

$$\left\|\text{grad}_y g(x,y) - \text{PT}_{y^* \to y} \text{grad}_y g(x,y^*) - \text{PT}_{y^* \to y} \text{Hess}_{yy} g(x,y^*) \text{PT}_{y \to y^*} \text{Exp}_{y^*}^{-1}(y)\right\| \leq l_{g,2} \|\text{Exp}_{y^*}^{-1}(y)\|^2$$

$$\left\|\text{grad}_y g(x,y) - \text{PT}_{y^* \to y} \text{grad}_y g(x,y^*) - \text{PT}_{y^* \to y} \text{Hess}_{yy} g(x,y^*) \text{PT}_{y \to y^*} \text{Exp}_{y^*}^{-1}(y)\right\| \leq 2l_{g,1} d_{\mathcal{M}}(y,y^*)$$

$$\left\|\text{grad}_x g(x,y) - \text{PT}_{y^* \to y} \text{grad}_x g(x,y^*) - \text{PT}_{y^* \to y} \text{Hess}_{xy} g(x,y^*) \text{PT}_{y \to y^*} \text{Exp}_{y^*}^{-1}(y)\right\|$$

$$\leq \min\left(l_{g,2} d_{\mathcal{M}}(y,y^*)^2, 2l_{g,1} d_{\mathcal{M}}(y,y^*)\right).$$

$$\left\|\text{grad}_x f(x,y^*) - \text{PT}_{y \to y^*} \text{grad}_x f(x,y)\right\| \leq l_{f,1} d_{\mathcal{N}}(y,y^*),$$

$$\left\|\text{grad}_y f(x,y^*) - \text{PT}_{y \to y^*} \text{grad}_y f(x,y)\right\| \leq l_{f,1} d_{\mathcal{N}}(y,y^*).$$

With this, our final result is:

$$\left\|\text{grad} F(x) - \text{grad}_x \mathcal{L}_\lambda(x,y) + \text{PT}_{y \to y^*} \text{Hess}_{xy} g(x,y^*) \left(\text{Hess}_{yy} g(x,y^*)\right)^{-1} \text{PT}_{y^* \to y} \text{grad}_y \mathcal{L}(x,y)\right\|$$

$$\leq l_{f,1} \left(1 + \frac{l_{g,1}}{\mu_g}\right) d_{\mathcal{N}}(y,y^*) + \lambda \left(1 + \frac{l_{g,1}}{\mu_g}\right) d_{\mathcal{N}}(y,y^*) \min\left(l_{g,2} d_{\mathcal{M}}(y,y^*)^2, 2l_{g,1}\right).$$

We know that $l_{g,1}/\mu_g \geq 1$ and thus, we have:

$$\left\|\text{grad} F(x) - \text{grad}_x \mathcal{L}_\lambda(x,y) + \text{PT}_{y \to y^*} \text{Hess}_{xy} g(x,y^*) \left(\text{Hess}_{yy} g(x,y^*)\right)^{-1} \text{PT}_{y^* \to y} \text{grad}_y \mathcal{L}(x,y)\right\|$$

$$\leq 2\left(\frac{l_{g,1}}{\mu_g}\right) d_{\mathcal{N}}(y,y^*) \left(l_{f,1} + \lambda \cdot \min\left(2l_{g,1}, l_{g,2} d_{\mathcal{N}}(y,y^*)\right)\right).$$

$\square$

A.3 Lemma A.3

Under Assumptions 1, 2 and 3, and $\lambda > 2l_{f,1}/\mu_g$, a function $y_\lambda^*(x)$ is $l_{\lambda,1}$-smooth: for any $x_1, x_2 \in X$, we have

$$\|\operatorname{grad} y_\lambda^*(x_1) - \operatorname{grad} y_\lambda^*(x_2)\| \leq l_{\lambda,1} d_{\mathcal{M}}(x_1, x_2)$$

where $l_{\lambda,1} \leq 32(l_{g,2} + \lambda^{-1}l_{f,2})l_{g,1}^2/\mu_g^3$.

Proof. The Lipschitz continuity of $y_\lambda^*(x)$ directly follows from Lemma 2, considering the manifold's intrinsic geometry. By the optimality condition for $\nabla y_\lambda^*(x)$ in the manifold setting, we obtain

$$\nabla_y \mathcal{L}_\lambda(x, y_\lambda^*(x)) = \nabla_y f(x, y_\lambda^*(x)) + \lambda \nabla_y g(x, y_\lambda^*(x)) = 0,$$

where ∇_y denotes the gradient with respect to y. Differentiating with respect to x along the manifold yields

$$\left(\operatorname{Hess}_{yy} f(x, y_\lambda^*(x)) + \lambda \operatorname{Hess}_{yy} g(x, y_\lambda^*(x))\right) \nabla y_\lambda^*(x) = -\left(\operatorname{Hess}_{xy} f(x, y_\lambda^*(x)) + \lambda \operatorname{Hess}_{xy} g(x, y_\lambda^*(x))\right),$$

where Hess_{yy} and Hess_{xy} represent the Hessians with respect to y and the mixed partial derivative along the manifold, respectively. Given $\lambda > 2l_{f,1}/\mu_g$, the left-hand side exhibits a positive definiteness with a minimum eigenvalue greater than $\lambda\mu_g/2$. Thus,

$$\nabla y_\lambda^*(x) = -\left(\frac{1}{\lambda}\operatorname{Hess}_{yy} f(x, y_\lambda^*(x)) + \operatorname{Hess}_{yy} g(x, y_\lambda^*(x))\right)^{-1} \left(\frac{1}{\lambda}\operatorname{Hess}_{xy} f(x, y_\lambda^*(x)) + \operatorname{Hess}_{xy} g(x, y_\lambda^*(x))\right).$$

To derive the smoothness property, we compare the expression at two points x_1 and x_2 on the manifold:

$$\frac{\lambda\mu_g}{2}\|\nabla y_\lambda^*(x_1) - \nabla y_\lambda^*(x_2)\| \leq (l_{f,2} + \lambda l_{g,2})(d_{\mathcal{M}}(x_1, x_2) + d_{\mathcal{N}}(y_\lambda^*(x_1), y_\lambda^*(x_2))) \max_{x \in X}\|\nabla y_\lambda^*(x)\|$$

$$+ (l_{f,2} + \lambda l_{g,2})(d_{\mathcal{M}}(x_1, x_2) + d_{\mathcal{N}}(y_\lambda^*(x_1), y_\lambda^*(x_2)))$$

$$\leq (l_{f,2} + \lambda l_{g,2})(1 + l_{\lambda,0})^2 d_{\mathcal{M}}(x_1, x_2).$$

Rearranging, we obtain

$$\|\nabla y_\lambda^*(x_1) - \nabla y_\lambda^*(x_2)\| \leq 32\left(\frac{l_{f,2}}{\lambda} + l_{g,2}\right)\frac{l_{g,1}^2}{\mu_g^3} d_{\mathcal{M}}(x_1, x_2).$$

$\square$

A.4 Lemma A.4

For any fixed $\lambda > 2l_{f,1}/\mu_g$, at every k iteration conditioned on $\mathcal{F}_k$, we have

$$\mathbb{E}[\|d_{\mathcal{N}}(y^*(x_{k+1}), y^*(x_k))^2 | \mathcal{F}_k] \leq \xi^2 l_{*,0}^2 \left(\alpha_k^2 \mathbb{E}[\|q_k^x\|^2 | \mathcal{F}_k] + \alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right).$$

Proof. The result directly follows from the Lipschitz continuity established in Lemma 2, taking the limit as $\lambda_1 = \lambda_2$ approaches infinity on a manifold.

Given the structure of a manifold, we assess the changes in the optimal solution y^* between consecutive points x_{k+1} and x_k through the geodesic distance, conditioned on the filtration $\mathcal{F}_k$.

This approach quantifies the modifications in y^* as we traverse from one location to another on the manifold. Specifically, we express the expectation of the squared geodesic distance between $y^*(x_{k+1})$ and $y^*(x_k)$ as follows:

$$\mathbb{E}\left[d_{\mathcal{N}}\left(y^*(x_{k+1}), y^*(x_k)\right)^2 \mid \mathcal{F}_k\right] \leq l_{*,0}^2 \mathbb{E}\left[d_{\mathcal{M}}\left(x_{k+1}, x_k\right)^2 \mid \mathcal{F}_k\right],$$

The inequality captures the bounded change in y^* in response to movements in x across the manifold, leveraging the Lipschitz property of y^* relative to x.

Further, by incorporating the step-sizes and the stochastic gradients' variances, we refine this inequality to:

$$\leq l_{*,0}^2 \xi^2 \alpha_k^2 \left(\mathbb{E}\left[\|q_k^x\|^2 \mid \mathcal{F}_k\right] + \alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right),$$

where ξ^2, α_k, σ_f, and σ_g encapsulate the effect of the algorithm's parameters and the inherent randomness of the optimization problem within the manifold setting. The expression $\mathbb{E}\left[\|q_k^x\|^2 \mid \mathcal{F}_k\right]$ reflects the expected squared norm of the search direction on the tangent space. $\quad \square$

A.5 Lemma A.5

At every k^{th} iteration, conditioned on $\mathcal{F}_k$, let v_k be a random vector decided before updating x_k. Then for any $\eta_k > 0$, we have

$$\mathbb{E}[\langle v_k, y^*(x_{k+1}) - y^*(x_k)\rangle | F_k] \leq \left(\xi \alpha_k \eta_k + M\xi^2 l_{*,1}^2 \beta_k^2\right) \mathbb{E}[\|v_k\|^2 | F_k]$$
$$+ \left(\frac{\xi \alpha_k l_{*,0}^2}{4\eta_k} + \frac{\xi^2 \alpha_k^2}{4}\right) \mathbb{E}[\|q_k^x\|^2 | \mathcal{F}_k] + \frac{\xi^2}{4}(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2),$$

where $M := \max\left(l_{f,0}^2 + \sigma_f^2, l_{g,0}^2 + \sigma_g^2\right)$.

Proof. Utilizing the smoothness property of $y^*(x)$ as discussed in Chen et al. (2021), which is essential for controlling the noise variance induced by updating x, we proceed as follows on a manifold:

Consider the inner product on the tangent space of the manifold at point x_k, which respects the manifold's geometry. For two vectors u, v in the tangent space at x_k, their inner product is denoted by $\langle u, v\rangle_{x_k}$. We can then express the expectation involving this inner product as follows:

$$\langle v_k, \text{Exp}_{x_k}^{-1}(y_{k+1}^*) - \text{Exp}_{x_k}^{-1}(y_k^*)\rangle_{x_k} = \langle v_k, \nabla y^*(x_k)(\text{Exp}_{x_k}^{-1}(x_{k+1}))\rangle_{x_k}$$
$$+ \langle v_k, \text{Exp}_{x_k}^{-1}(y^*(x_{k+1})) - \text{Exp}_{x_k}^{-1}(y^*(x_k)) - \nabla y^*(x_k)(\text{Exp}_{x_k}^{-1}(x_{k+1}))\rangle_{x_k}.$$

For the first term, applying the expectation and the Cauchy-Schwarz inequality on the manifold, we get:

$$\mathbb{E}[\langle v_k, \nabla y^*(x_k)(\text{Exp}_{x_k}^{-1}(x_{k+1}))\rangle_{x_k} \mid \mathcal{F}_k] = -\xi \alpha_k \mathbb{E}[\langle v_k, \nabla y^*(x_k)q_k^x\rangle_{x_k} \mid \mathcal{F}_k]$$

$$\leq \xi \alpha_k \eta_k \mathbb{E}[\|v_k\|_{x_k}^2 \mid \mathcal{F}_k] + \frac{\xi \alpha_k}{4\eta_k} \mathbb{E}[\|\nabla y^*(x_k)q_k^x\|_{x_k}^2 \mid \mathcal{F}_k]$$

$$\leq \xi \alpha_k \eta_k \mathbb{E}[\|v_k\|_{x_k}^2 \mid \mathcal{F}_k] + \frac{\xi \alpha_k l_{*,0}^2}{4\eta_k} \mathbb{E}[\|q_k^x\|_{x_k}^2 \mid \mathcal{F}_k]$$

For the second term, leveraging the smoothness of $y^*(x)$ on the manifold, we have:

$$\mathbb{E}[\langle v_k, \mathrm{Exp}_{x_k}^{-1}(y^*(x_{k+1})) - \mathrm{Exp}_{x_k}^{-1}(y^*(x_k)) - \nabla y^*(x_k)(\mathrm{Exp}_{x_k}^{-1}(x_{k+1}))\rangle_{x_k} \mid \mathcal{F}_k]$$

$$\leq \frac{l_{*,1}}{2} \mathbb{E}[d_{\mathcal{M}}(x_{k+1}, x_k)^2 \mid \mathcal{F}_k]$$

$\square$

A.6   Lemma A.6

Under Assumptions 1-5, at every k^{th} iteration, conditioned on $\mathcal{F}_k$, let v_k be a random vector decided before updating x_k. Then for any $\eta_k > 0$, we have

$$\mathbb{E}[\langle v_k, y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_k)\rangle | \mathcal{F}_k] \leq (\delta_k/\lambda_k + \xi \alpha_k \eta_k + M\xi^2 l_{\lambda_k,1}^2 \beta_k^2) \, \mathbb{E}[\|v_k\|^2 | \mathcal{F}_k]$$

$$+ \left(\frac{\xi \alpha_k l_{*,0}^2}{4\eta_k} + \frac{\xi^2 \alpha_k^2}{4} \right) \mathbb{E}[\|q_k^x\|^2 | \mathcal{F}_k] + \frac{\xi^2}{4}(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2) + \frac{\delta_k l_{f,0}^2}{\lambda_k^3 \mu_g^2},$$

where $M := \max\left(l_{f,0}^2 + \sigma_f^2, l_{g,0}^2 + \sigma_g^2 \right)$.

Proof. We start with the following decomposition:

$$\langle v_k, \mathrm{Exp}_{x_k}^{-1}(y_{\lambda_{k+1}}^*(x_{k+1})) - \mathrm{Exp}_{x_k}^{-1}(y_{\lambda_k}^*(x_k))\rangle =$$
$$\langle v_k, \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_{k+1}}^*(x_{k+1})) - \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_k}^*(x_{k+1}))\rangle$$
$$+ \langle v_k, \nabla y_{\lambda_k}^*(x_k)(\mathrm{Exp}_{x_k}^{-1}(x_{k+1}))\rangle$$
$$+ \langle v_k, \mathrm{Exp}_{x_k}^{-1}(y_{\lambda_k}^*(x_{k+1})) - \mathrm{Exp}_{x_k}^{-1}(y_{\lambda_k}^*(x_k)) - \nabla y_{\lambda_k}^*(x_k)(\mathrm{Exp}_{x_k}^{-1}(x_{k+1}))\rangle.$$

For the second and third terms, the smoothness of $y_\lambda(x)$ is applied similarly to the proof in A.5, considering the manifold's intrinsic geometry.

Regarding the first term, taking expectation and using the inequality $\langle a, b \rangle \leq c\|a\|^2 + \frac{1}{4c}\|b\|^2$ adapted for the tangent space, we get:

$$\mathbb{E}[\langle v_k, \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_{k+1}}^*(x_{k+1})) - \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_k}^*(x_{k+1}))\rangle \mid \mathcal{F}_k]$$

$$\leq c\mathbb{E}[\|v_k\|^2] + \frac{1}{4c}\mathbb{E}[\| \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_{k+1}}^*(x_{k+1})) - \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_k}^*(x_{k+1}))\|^2]$$

$$\leq c\mathbb{E}[\|v_k\|^2] + \frac{1}{c}\frac{\delta_k^2}{\lambda_k^2 \lambda_{k+1}^2}\frac{l_{f,0}^2}{\mu_g^2},$$

where the expectation and norms are understood within the context of the manifold's geometry. By selecting $c = \frac{\delta_k}{\lambda_k}$, we derive:

$$\mathbb{E}[\langle v_k, \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_{k+1}}^*(x_{k+1})) - \mathrm{Exp}_{x_{k+1}}^{-1}(y_{\lambda_k}^*(x_{k+1}))\rangle \mid \mathcal{F}_k] \leq \frac{\delta_k}{\lambda_k}\mathbb{E}[\|v_k\|^2] + \frac{l_{f,0}^2 \delta_k}{\mu_g^2 \lambda_k^3}.$$

Combining this with bounds on the other two terms, we conclude the lemma. $\square$

22

Proof of Lemma 1

Let $y_\lambda^*(x) := \arg\min_y \mathcal{L}_\lambda(x,y)$. Note that $\mathrm{grad}_y \mathcal{L}_\lambda(x, y_\lambda^*(x)) = 0$, and thus

$$\mathrm{grad}\mathcal{L}_\lambda^*(x) = \mathrm{grad}_x\mathcal{L}_\lambda(x, y_\lambda^*(x)) + \mathrm{grad}_x y_\lambda^*(x)^T\mathrm{grad}_y\mathcal{L}_\lambda(x, y_\lambda^*(x)) = \mathrm{grad}_x\mathcal{L}_\lambda(x, y_\lambda^*(x)).$$

To compare this to $\mathrm{grad}F(x)$, we can invoke Lemma A.2 which gives

$$\|\mathrm{grad}F(x) - \mathrm{grad}_x\mathcal{L}_\lambda(x, y_\lambda^*(x))\|$$
$$\leq 2\left(l_{g,1}/\mu_g\right)d_\mathcal{N}(y_\lambda^*(x), y^*(x))\left(l_{f,1} + \lambda \cdot \min\left(2l_{g,1}, l_{g,2}d_\mathcal{N}(y^*(x), y_\lambda^*(x))\right)\right).$$

From a version of Lemma 2 (A.6), we use $d_\mathcal{N}(y_\lambda^*(x), y^*(x)) \leq \frac{2l_{f,0}}{\lambda\mu_g}$, and get

$$\|\mathrm{grad}F(x) - \mathrm{grad}_x\mathcal{L}_\lambda(x, y_\lambda^*(x))\| \leq \frac{1}{\lambda} \cdot \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right). \quad \square$$

Here, $\mathrm{grad}_x y_\lambda^*(x)^T$ represents the transpose of the gradient of $y_\lambda^*(x)$ with respect to x.

Proof of Lemma 2

Note that on a manifold, the function $\mathcal{L}_\lambda(x,y)$ is at least $\frac{\lambda\mu_g}{2}$ strongly-convex in y with respect to the metric once $\lambda \geq 2l_{f,1}\mu_g$. To see this,

$$\mathcal{L}_\lambda(x,y) = f(x,y) + \lambda(g(x,y) - g^*(x))$$

which is at least $-l_{f,1} + \lambda\mu_g$-strongly convex in y with respect to the metric. If $\lambda > 2l_{f,1}/\mu_g$, this implies at least $\lambda\mu_g/2$ strong-convexity of $\mathcal{L}_\lambda(x,y)$ in y.

By the optimality condition at $y_{\lambda_1}^*(x_1)$ with x_1, λ_1, we have

$$\mathrm{grad}_y f(x_1, y_{\lambda_1}^*(x_1)) + \lambda_1\mathrm{grad}_y g(x_1, y_{\lambda_1}^*(x_1)) = 0,$$

which also implies that $d_M(g(x_1, y_{\lambda_1}^*(x_1)), 0) \leq l_{f,0}/\lambda_1$. Observe that

$$\mathrm{grad}_y f(x_2, y_{\lambda_1}^*(x_1)) + \lambda_2\mathrm{grad}_y g(x_2, y_{\lambda_1}^*(x_1))$$
$$= \left(\mathrm{grad}_y f(x_2, y_{\lambda_1}^*(x_1)) - \mathrm{PT}_{x_1\to x_2}(\mathrm{grad}_y f(x_1, y_{\lambda_1}^*(x_1)))\right) + \mathrm{PT}_{x_1\to x_2}(\mathrm{grad}_y f(x_1, y_{\lambda_1}^*(x_1)))$$
$$+ \lambda_2\left(\mathrm{grad}_y g(x_2, y_{\lambda_1}^*(x_1)) - \mathrm{PT}_{x_1\to x_2}(\mathrm{grad}_y g(x_1, y_{\lambda_1}^*(x_1)))\right) + \lambda_2\mathrm{PT}_{x_1\to x_2}(\mathrm{grad}_y g(x_1, y_{\lambda_1}^*(x_1)))$$
$$= \left(\mathrm{grad}_y f(x_2, y_{\lambda_1}^*(x_1)) - \mathrm{PT}_{x_1\to x_2}(\mathrm{grad}_y f(x_1, y_{\lambda_1}^*(x_1)))\right)$$
$$+ \lambda_2\left(\mathrm{grad}_y g(x_2, y_{\lambda_1}^*(x_1)) - \mathrm{PT}_{x_1\to x_2}(\mathrm{grad}_y g(x_1, y_{\lambda_1}^*(x_1)))\right)$$
$$+ (\lambda_2 - \lambda_1)\mathrm{PT}_{x_1\to x_2}(\mathrm{grad}_y g(x_1, y_{\lambda_1}^*(x_1))),$$

where in the last equality, we applied the optimality condition for $y_{\lambda_1}^*(x_1)$. Then applying the Lipschitzness of $\mathrm{grad}_y f$ and $\mathrm{grad}_y g$ in x, we have

$$d_M(\mathrm{grad}_y f(x_2, y_{\lambda_1}^*(x_1)) + \lambda_2\mathrm{grad}_y g(x_2, y_{\lambda_1}^*(x_1)), 0) \leq l_{f,1}d_M(x_1, x_2) + l_{g,1}\lambda_2 d_M(x_2, x_1) + (\lambda_2 - \lambda_1)\frac{l_{f,0}}{\lambda_1}.$$

Since $\mathcal{L}_{\lambda_2}(x_2, y)$ is $\lambda_2 \mu_g/2$-strongly convex in y with respect to the metric, from the coercivity property of strongly-convex functions, along with the optimality condition with $y^*_{\lambda_2}(x_2)$, we have

$$
\begin{aligned}
\frac{\lambda_2 \mu_g}{2} & d_{\mathcal{M}}(y^*_{\lambda_1}(x_1), y^*_{\lambda_2}(x_2)) \\
& \leq \|\nabla_y \mathcal{L}_{\lambda_2}(x_2, y^*_{\lambda_1}(x_1)) - \nabla_y \mathcal{L}_{\lambda_2}(x_2, y^*_{\lambda_1}(x_2))\| \\
& \leq (l_{f,1} + \lambda_2 l_{g,1}) d_{\mathcal{M}}(x_1, x_2) + \frac{\lambda_2 - \lambda_1}{\lambda_1} l_{f,0}.
\end{aligned}
$$

$$
\implies \frac{\lambda_2 \mu_g}{2} d_{\mathcal{N}}(y^*_{\lambda_1}(x_1), y^*_{\lambda_2}(x_2)) \leq \|\nabla_y \mathcal{L}_{\lambda_2}(x_2, y^*_{\lambda_1}(x_1))\| \leq (l_{f,1} + \lambda_2 l_{g,1}) d_{\mathcal{M}}(x_1, x_2) + \frac{\lambda_2 - \lambda_1}{\lambda_1} l_{f,0}.
$$

Dividing both sides by $\lambda_2 \mu_g/2$ concludes the first part of the proof. Note that $y^*(x) = \lim_{\lambda \to \infty} y^*_\lambda(x)$. Thus, for any x and finite $\lambda \geq 2 l_{f,1}/\mu_g$,

$$
d_{\mathcal{N}}(y^*_\lambda(x), y^*(x)) \leq \frac{2 l_{f,0}}{\lambda \mu_g}. \quad \square
$$

Remarks.

The Cauchy-Schwarz inequality in the context of manifolds states that for any two tangent vectors u, v at a point, the following holds:

$$
\langle u, v \rangle \leq \|u\| \cdot \|v\|,
$$

where $\langle \cdot, \cdot \rangle$ denotes the metric and $\| \cdot \|$ is the norm induced by this metric.

Applying this to the inequality for strong convexity:

$$
\langle \operatorname{grad}_x f - \operatorname{PT}_{yx}(\operatorname{grad}_y f), \operatorname{Exp}_x^{-1}(y) \rangle \geq \mu \| \operatorname{Exp}_x^{-1}(y)\|^2,
$$

we get:

$$
\|\operatorname{grad}_x f - \operatorname{PT}_{yx}(\operatorname{grad}_y f)\| \cdot \| \operatorname{Exp}_x^{-1}(y)\| \geq \langle \operatorname{grad}_x f - \operatorname{PT}_{yx}(\operatorname{grad}_y f), \operatorname{Exp}_x^{-1}(y) \rangle.
$$

Since the right-hand side of this inequality is the same as the left-hand side of the strong convexity inequality, we can substitute it in, yielding:

$$
\|\operatorname{grad}_x f - \operatorname{PT}_{yx}(\operatorname{grad}_y f)\| \cdot \| \operatorname{Exp}_x^{-1}(y)\| \geq \mu \| \operatorname{Exp}_x^{-1}(y)\|^2.
$$

This form of the inequality highlights the relationship between the difference in gradients (after parallel transport) and the geodesic distance between points x and y.

B Proofs of Theorem 2 and Corollary 3

In this section, we prove our central result in Theorem 2. The crux of this analysis revolves around determining the upper boundary of $V_{k+1} - V_k$ concerning the potential function V_k, as elucidated in equation 4.3, tailored to our specific setting.

About x_k and y_k, as characterized in RF^{2}SA within the manifold framework, we introduce the following notations:

$$
I_k := d_{\mathcal{N}}(y_k, y^*_{\lambda,k})^2, \quad J_k := d_{\mathcal{N}}(z_k, y^*_k)^2, \tag{B.1}
$$

where $y_{\lambda,k}^* := y_{\lambda_k}^*(x_k)$, $y_k^* := y^*(x_k)$, and $x^* = \mathrm{argmin}_x F(x)$, all situated. Here, $d_{\mathcal{N}}(\cdot, \cdot)$ signifies the distance metric on manifold $\mathcal{N}$.

Leveraging these notations, we redefine the potential function V_k as:

$$V_k := (F(x_k) - F(x^*)) + \lambda_k l_{g,1} I_k + \frac{\lambda_k l_{g,1}}{2} J_k, \tag{B.2}$$

for each $k \in \mathbb{N}$. In the ensuing subsections, our aim is to delineate the upper limit of $V_{k+1} - V_k$ vis-à-vis I_k and J_k, giving due consideration to the manifold's geometry and curvature characteristics. The proof for the Theorem 2, aptly adapted to this scenario, will be explicated in Section B.4.

B.1 Estimation of $F(x_{k+1}) - F(x_k)$

The selection of the step size α_k is carefully chosen to fit the context of the manifold:

$$\text{(step-size rule):} \quad \alpha_k \leq \frac{1}{2\xi \tilde{l}_{F,1}}, \tag{B.3}$$

where $\tilde{l}_{F,1}$ is appropriately adjusted to match the manifold's geometric properties. This adjustment is crucial for including the negative term $-\frac{\xi \alpha_k}{4} \|\mathrm{grad}F(x_k)\|_{x_k}^2$ in our analysis. This term is key in the demonstration of Theorem 2, as discussed in Section B.4.

Moreover, we also stipulate:

$$\text{(step-size rule):} \quad \frac{\xi}{T} \leq \frac{\mu_g}{96 \tilde{l}_{g,1}}. \tag{B.4}$$

The metrics $d_{\mathcal{N}}^2(y_{k+1}, y_{\lambda,k}^*)$ and $d_{\mathcal{N}}^2(z_{k+1}, y_k^*)$, will be integral to deriving our upper bound estimates, as detailed in Propositions 3 and 5, respectively.

Proposition 1. Under the step-size rules given in equations B.3 and B.4, and $\lambda_k \geq 2l_{f,1}/\mu_g$, it holds that for each $k \in \mathbb{N}$

$$
\begin{aligned}
\mathbb{E}\left[F(x_{k+1}) - F(x_k) \mid \mathcal{F}_k\right] \leq &- \frac{\xi \alpha_k}{4} \left(2 \|\nabla F(x_k)\|^2 + \|q_k^x\|^2\right) \\
&+ \frac{T \mu_g \alpha_k \lambda_k^2}{32} \left(2d_{\mathcal{N}}^2(y_{k+1}, y_{\lambda,k}^*) + d_{\mathcal{N}}^2(z_{k+1}, y_k^*)\right) \\
&+ \frac{\xi^2 l_{F,1}}{2} \left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right) + \frac{\xi \alpha_k}{2} \cdot 3C_\lambda^2 \lambda_k^{-2}
\end{aligned}
$$

where q_k^x is given in equation 1, and $C_\lambda := \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right)$.

Proof. From the smoothness of F,

$$\mathbb{E}\left[F(x_{k+1}) - F(x_k) \mid \mathcal{F}_k\right] \leq \mathbb{E}\left[\langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{l_{F,1}}{2} d_{\mathcal{M}}^2(x_{k+1}, x_k) \,\middle|\, \mathcal{F}_k\right]$$

As q_k^x satisfies $\mathbb{E}\left[x_{k+1} - x_k \mid \mathcal{F}_k\right] = \alpha_k q_k^x$,

$$
\begin{aligned}
\mathbb{E}\left[F(x_{k+1}) - F(x_k) \mid \mathcal{F}_k\right] &= -\xi \alpha_k \langle \nabla_x F(x_k), q_k^x \rangle + \frac{l_{F,1}}{2} \mathbb{E}\left[d_{\mathcal{M}}^2(x_{k+1}, x_k) \mid \mathcal{F}_k\right] \\
&= -\frac{\xi \alpha_k}{2} \left(\|\nabla F(x_k)\|^2 + \|q_k^x\|^2 - \|\nabla F(x_k) - q_k^x\|^2\right) + \frac{l_{F,1}}{2} \mathbb{E}\left[d_{\mathcal{M}}^2(x_{k+1}, x_k) \mid \mathcal{F}_k\right]
\end{aligned}
$$

25

Note that

$$\mathbb{E}\left[d_{\mathcal{M}}^2\left(x_{k+1}, x_k\right)\right] \leq \xi^2 \alpha_k^2 \mathbb{E}\left[\|q_k^x\|^2 + \xi^2\left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right)\right]$$

and thus with B.3 we have

$$\mathbb{E}\left[F\left(x_{k+1}\right) - F\left(x_k\right) \mid \mathcal{F}_k\right] \leq -\frac{\xi \alpha_k}{2}\|\nabla F\left(x_k\right)\|^2 - \frac{\xi \alpha_k}{4}\|q_k^x\|^2$$
$$+ \frac{\xi \alpha_k}{2}\|\nabla F\left(x_k\right) - q_k^x\|^2 + \frac{\xi^2 l_{F,1}}{2}\left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right).$$

Next, we bound $\left\|\nabla F\left(x_k\right) - q_k^x\right\|$ using the triangle inequality:

$$\|\nabla F\left(x_k\right) - q_k^x\| \leq \left\|\nabla_x f\left(x_k, y_{k+1}\right) - \nabla_x f\left(x_k, y_{\lambda,k}^*\right)\right\| + \lambda_k \left\|\nabla_x g\left(x_k, y_{k+1}\right) - \nabla_x g\left(x_k, y_{\lambda,k}^*\right)\right\|$$
$$+ \lambda_k \left\|\nabla_x g\left(x_k, z_{k+1}\right) - \nabla_x g\left(x_k, y_k^*\right)\right\| + \left\|\nabla \mathcal{L}_{\lambda_k}^*\left(x_k\right) - \nabla F\left(x_k\right)\right\|$$

From Lemma 1, the term $\left\|\nabla \mathcal{L}_{\lambda_k}^*\left(x_k\right) - \nabla F\left(x_k\right)\right\|$ is bounded by C_λ / λ_k. Combining with the regularity of f and g yields the following:

$$\|\nabla F\left(x_k\right) - q_k^x\| \leq 2 l_{g,1} \lambda_k d_{\mathcal{N}}(y_{k+1}, y_{\lambda,k}^*) + l_{g,1} \lambda_k d_{\mathcal{N}}(z_{k+1}, y_k^*) + C_\lambda / \lambda_k.$$

Finally, from the Cauchy-Schwartz inequality $(a + b + c)^2 \leq 3\left(a^2 + b^2 + c^2\right)$, we get

$$\mathbb{E}\left[F\left(x_{k+1}\right) - F\left(x_k\right) \mid \mathcal{F}_k\right] \leq -\frac{\xi \alpha_k}{2}\|\nabla F\left(x_k\right)\|^2 - \frac{\xi \alpha_k}{4}\|q_k^x\|^2$$
$$+ \frac{\xi \alpha_k}{2} \cdot 3 C_\lambda^2 \lambda_k^{-2} + 3 \xi \alpha_k l_{g,1} \lambda_k^2 d_{\mathcal{N}}(z_{k+1}, y_k^*)^2 + 6 \xi \alpha_k l_{g,1} \lambda_k^2 d_{\mathcal{N}}(y_{k+1}, y_{\lambda,k}^*)^2 + \frac{\xi^2 l_{F,1}}{2}\left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right)$$

The step-size condition B.4 concludes our claim. $\qquad\square$

B.2 Descent Lemma for y_k towards $y_{\lambda,k}^*$

In this section, we provide the upper bounds of $\mathcal{I}_{k+1}$ and $d_{\mathcal{N}}(y_{k+1}, y_{\lambda,k}^*)$ in the context of a manifold. The following step-size rule is adapted for the manifold's geometry:

$$\text{(step-size rule):} \quad \frac{\delta_k}{\lambda_k} \leq \frac{T \beta_k \mu_g}{32}, \text{ and } 2 \xi^2 M l_{*,1}^2 \beta_k^2 < \frac{T \beta_k \mu_g}{16} \tag{B.5}$$

Proposition 2. Given the step-size rule B.7, for each $k \in \mathbb{N}$, we have

$$\mathbb{E}\left[\mathcal{I}_{k+1} \mid \mathcal{F}_k\right] \leq \left(\left[\frac{\sqrt{|\kappa|} d_{\mathcal{N}}(y_{\lambda,k+1}^*, y_{\lambda,k}^*)}{\tanh(\sqrt{|\kappa|} d_{\mathcal{N}}(y_{\lambda,k+1}^*, y_{\lambda,k}^*))}\right] + 2\delta_k / \lambda_k + T\beta_k \mu_g / 8 + 2M\xi^2 l_{*,1}^2 \beta_k^2\right) \mathbb{E}\left[d_{\mathcal{N}}^2(y_{k+1}, y_{\lambda,k}^*)\right]$$
$$+ O\left(\frac{\xi^2 l_{*,0}^2 \alpha_k^2}{\mu_g T \beta_k}\right) \mathbb{E}\left[\|\tau_{x_k}(q_k^x)\|^2\right] + O\left(\frac{\delta_k}{\lambda_k^3} \frac{l_{f,0}^2}{\mu_g^2}\right)$$
$$+ O\left(\xi^2 l_{*,0}^2\right) \cdot \left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right)$$

Here, $\mathcal{I}_k$ is adapted to consider the geodesic distance, and q_k^x is calculated in the tangent space of the manifold.

Proof. We start from the version of the distance and the inner product, considering the curvature κ:

$$d^2_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k+1}) = \underbrace{d^2_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k})}_{(i)} + \underbrace{d^2_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k})}_{(ii)}$$
$$- \underbrace{2 d_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k}) d_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k}) \cos(\angle(y_{k+1}, y^*_{\lambda,k}, y^*_{\lambda,k+1}))}_{(iii)}$$

Incorporating the curvature κ, we apply the Alexandrov space cosine law (Zhang and Sra, 2016):

$$d^2_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k+1})$$
$$\leq \underbrace{\frac{\sqrt{|\kappa|} d_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k})}{\tanh(\sqrt{|\kappa|} d_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k}))} \left(d^2_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k}) \right)}_{(i)}$$
$$+ \underbrace{d^2_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k})}_{(ii)} - \underbrace{2 d_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k}) d_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k}) \cos(\angle(y_{k+1}, y^*_{\lambda,k}, y^*_{\lambda,k+1}))}_{(iii)}$$

The upper bound of (i) is given in Proposition 3 below. To bound (ii), we invoke Lemma 2, yielding

$$(ii) \ \mathbb{E}\left[d^2_{\mathcal{N}}\left(y^*_{\lambda,k+1}, y^*_{\lambda,k}\right) \mid \mathcal{F}_k\right] \leq \frac{4\delta^2_k}{\lambda^2_k \lambda^2_{k+1}} \frac{l^2_{f,0}}{\mu^2_g} + l^2_{*,0} \mathbb{E}\left[d^2_{\mathcal{N}}\left(x_{k+1}, x_k\right) \mid \mathcal{F}_k\right]$$

$$\leq \frac{4\delta^2_k}{\lambda^4_k} \frac{l^2_{f,0}}{\mu^2_g} + \xi^2 l^2_{*,0} \left(\alpha^2_k \mathbb{E}\left[\|\tau_{x_k}(q^x_k)\|^2\right] + \alpha^2_k \sigma^2_f + \beta^2_k \sigma^2_g \right)$$

where $\tau_{x_k}(q^x_k)$ denotes the parallel transport of the search direction q^x_k at the point x_k along the manifold.

For (iii), considering the smoothness property of $y^*_\lambda(x)$ as per a generalized version of Lemma A.3, and thus Lemma A.6, we set $v = \mathrm{Exp}^{-1}_{y^*_{\lambda,k}}(y_{k+1})$ and $\eta_k = T\mu_g \lambda_k/(16\xi)$, we obtain

$$(iii) \leq \left(2\delta_k/\lambda_k + T\beta_k \mu_g/8 + 2M\xi^2 l^2_{*,1}\beta^2_k\right) \mathbb{E}\left[d^2_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k}) \mid \mathcal{F}_k\right]$$
$$+ \xi^2 \left(\frac{\alpha^2_k}{2} + \frac{8\alpha^2_k l^2_{*,0}}{\mu_g T\beta_k}\right) \mathbb{E}\left[\|\tau_{x_k}(q^x_k)\|^2\right] + \frac{\xi^2}{2}\left(\alpha^2_k \sigma^2_f + \beta^2_k \sigma^2_g\right) + \frac{2\delta_k}{\lambda^3_k} \frac{l^2_{f,0}}{\mu^3_g}$$

We sum up the $(i), (ii), (iii)$ to conclude

$$\mathbb{E}\left[\mathcal{I}_{k+1} \mid \mathcal{F}_k\right] \leq \left(\left[\frac{\sqrt{|\kappa|} d_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k})}{\tanh(\sqrt{|\kappa|} d_{\mathcal{N}}(y^*_{\lambda,k+1}, y^*_{\lambda,k}))}\right] + 2\delta_k/\lambda_k + T\beta_k \mu_g/8 + 2M\xi^2 l^2_{*,1}\beta^2_k\right) \mathbb{E}\left[d^2_{\mathcal{N}}(y_{k+1}, y^*_{\lambda,k})\right]$$
$$+ O\left(\frac{\xi^2 l^2_{*,0}\alpha^2_k}{\mu_g T\beta_k}\right) \mathbb{E}\left[\|\tau_{x_k}(q^x_k)\|^2\right] + O\left(\frac{\delta_k}{\lambda^3_k} \frac{l^2_{f,0}}{\mu^2_g}\right)$$
$$+ O\left(\xi^2 l^2_{*,0}\right) \cdot \left(\alpha^2_k \sigma^2_f + \beta^2_k \sigma^2_g\right)$$

Lastly, the step-size rule B.5 yields our conclusion. $\qquad\square$

Next, we note that α_k and β_k are chosen to satisfy

$$\text{(step size rules):} \quad \alpha_k \leq \frac{1}{8l_{f,1}} \quad \text{and} \quad \beta_k \leq \frac{1}{8l_{g,1}} \tag{B.6}$$

Note that $\beta_k \leq \frac{1}{8l_{g,1}}$ is given from the step-size condition (3a), and $\alpha_k \leq \frac{1}{8l_{g,1}\lambda_k} \leq \frac{1}{8l_{f,1}}$ since $\lambda_k \geq l_{f,1}/\mu_g$.

Proposition 3. Under the given step-size rules, it holds that for each $k \in \mathbb{N}$

$$\mathbb{E}\left[d_{\mathcal{N}}^2\left(y_{k+1}, y_{\lambda,k}^*\right) \mid \mathcal{F}_k\right] \leq \left(\left(\frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_k^{(t+1)}, y_k^{(t)}\right)}{\tanh(\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_k^{(t+1)}, y_k^{(t)}\right))}\right) - 3T\mu_g\beta_k/4\right)\mathcal{I}_k + T\left(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right)$$

Proof. Since the expected value of the difference between successive iterations in a manifold can be expressed as $\mathbb{E}\left[\text{Exp}_{y_k^{(t+1)}}^{-1}(y_k^{(t)}) \mid \mathcal{F}_k\right] = -\alpha_k\nabla_y q_k^{(t)} = -\alpha_k\nabla_y \mathcal{L}_{\lambda_k}(x_k, y_k^{(t)})$, we have

$$\mathbb{E}\left[d_{\mathcal{N}}^2\left(y_k^{(t+1)}, y_{\lambda,k}^*\right) \mid \mathcal{F}_k\right]$$

$$\leq \frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_k^{(t+1)}, y_k^{(t)}\right)}{\tanh(\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_k^{(t+1)}, y_k^{(t)}\right))}\left(d_{\mathcal{N}}^2\left(y_k^{(t)}, y_{\lambda,k}^*\right)\right) + d_{\mathcal{N}}^2\left(y_k^{(t+1)}, y_k^{(t)}\right)$$

$$- 2d_{\mathcal{N}}\left(y_k^{(t)}, y_{\lambda,k}^*\right)d_{\mathcal{N}}\left(y_k^{(t+1)}, y_k^{(t)}\right)\cos(\angle(y_k^{(t)}, y_{\lambda,k}^*, y_k^{(t+1)}))$$

Given that $\lambda_0 \geq 2\mu_f/\mu_g$, and all $\mathcal{L}_k$ is strongly convex in y, the following inequality holds

$$\max\left(\frac{\lambda_k\mu_g}{2}d_{\mathcal{N}}^2\left(y_k^{(t)}, y_{\lambda,k}^*\right), \frac{1}{l_{f,1} + \lambda_k l_{g,1}}\|\nabla_y q_k^{(t)}\|^2\right) \leq \langle\nabla_y q_k^{(t)}, \text{Exp}_{y_k^{(t)}}^{-1}(y_{\lambda,k}^*)\rangle$$

Using the Alexandrov space result to approximate the expected squared distance, we have

$$\mathbb{E}\left[d_{\mathcal{N}}^2\left(y_k^{(t+1)}, y_{\lambda,k}^*\right) \mid \mathcal{F}_k\right] \leq \left(\frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_k^{(t+1)}, y_k^{(t)}\right)}{\tanh(\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_k^{(t+1)}, y_k^{(t)}\right))} - \frac{3\mu_g\beta_k}{4}\right)d_{\mathcal{N}}^2\left(y_k^{(t)}, y_{\lambda,k}^*\right) + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2,$$

where $\alpha_k\left(l_{f,1} + \lambda_k l_{g,1}\right) = \alpha_k l_{f,1} + \beta_k l_{g,1} \leq 1/4$ under condition B.6. Repeating this process T times as per the algorithm leads to the conclusion in Proposition 3, where $y_{k+1} = y_k^{(T)}$ and $y_k = y_k^{(0)}$. □

B.3   Descent Lemma for z_k towards y_k^*

Similar to the previous section, we provide the upper bound of $\mathcal{J}_{k+1}$ first and then estimate $d_{\mathcal{N}}(z_{k+1}, y_k^*)$ that appears in the upper bound. We work with the following step-size condition:

$$\text{(step-size rule):} \quad 2Ml_{*,1}^2\xi^2\beta_k^2 \leq T\mu_g\gamma_k/16 \tag{B.7}$$

This condition holds since $\beta_k \leq \gamma_k$, and $\beta_k \leq \frac{1}{4T\mu_g}$ and $\frac{\xi^2}{T^2} \leq \frac{\mu_g^2}{8}\left(Ml_{*,1}^2\right)^{-1}$.

Proposition 4. Under the step-size rule B.7, at each kth iteration, the following holds:

$$\mathbb{E}\left[\mathcal{J}_{k+1} \mid \mathcal{F}_k\right] \leq \left(\frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right)}{\tanh(\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right))} + \frac{3T\gamma_k\mu_g}{8}\right) \cdot \mathbb{E}\left[d_{\mathcal{N}}^2\left(z_{k+1}, y_k^*\right) \mid \mathcal{F}_k\right]$$

$$+ O\left(\frac{\xi^2\alpha_k^2 l_{*,0}^2}{T\mu_g\gamma_k}\right)\|q_k^x\|^2 + O\left(\xi^2 l_{*,0}^2\right)\left(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right)$$

Proof. We estimate each term in the following decomposition.

$$d_{\mathcal{N}}\left(z_{k+1}, y_{k+1}^*\right)^2 \leq \underbrace{\frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right)}{\tanh(\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right))}\left(d_{\mathcal{N}}\left(z_{k+1}, y_k^*\right)^2\right)}_{(i)}$$

$$+ \underbrace{d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right)^2}_{(ii)} - \underbrace{2d_{\mathcal{N}}\left(z_{k+1}, y_k^*\right)d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right)\cos(\angle(z_{k+1}, y_k^*, y_{k+1}^*))}_{(iii)}$$

Lemma 2 (A.6) implies that

$$(ii)\ \mathbb{E}\left[\|y_{k+1}^* - y_k^*\|^2 \mid \mathcal{F}_k\right] \leq l_{*,0}^2\xi^2\left(\alpha_k^2\|\nabla_x q_k\|^2 + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right)$$

For (iii), we recall Lemma A.5 with $v_k = z_{k+1} - y_k^*$ and $\eta_k = T\mu_g\gamma_k/(8\xi\alpha_k)$, we have

$$(iii)\ \langle\text{Exp}_{y_k^*}^{-1}(z_{k+1}), \text{Exp}_{y_k^*}^{-1}(y_{k+1}^*)\rangle \leq \left(T\gamma_k\mu_g/8 + M\xi^2 l_{*,1}^2\beta_k^2\right)\mathbb{E}\left[d_{\mathcal{N}}^2(z_{k+1}, y_k^*) \mid \mathcal{F}_k\right]$$

$$+ \left(\frac{\xi^2\alpha_k^2}{4} + \frac{2\xi^2\alpha_k^2 l_{*,0}^2}{T\mu_g\gamma_k}\right)\|\nabla_x q_k\|^2 + \frac{\xi^2}{4}\left(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right)$$

The above bounds and Proposition 5 imply that

$$\mathbb{E}\left[\mathcal{J}_{k+1} \mid \mathcal{F}_k\right] \leq \left(\frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right)}{\tanh(\sqrt{|\kappa|}d_{\mathcal{N}}\left(y_{k+1}^*, y_k^*\right))} + \frac{T\gamma_k\mu_g}{4} + 2M\xi^2 l_{*,1}^2\beta_k^2\right) \cdot \mathbb{E}\left[d_{\mathcal{N}}\left(z_{k+1} - y_k^*\right)^2 \mid \mathcal{F}_k\right]$$

$$+ \xi^2\alpha_k^2 \cdot \left(l_{*,0}^2 + \frac{4l_{*,0}^2}{T\mu_g\gamma_k} + \frac{1}{2}\right)\|q_k^x\|^2 + \xi^2 \cdot \left(\frac{1}{2} + l_{*,0}^2\right)\left(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right)$$

Using B.7, we conclude. □

Next, γ_k is chosen to satisfy the following step-size rules:

$$\text{(step-size rule):}\quad l_{g,1}\gamma_k \leq 1/4, \quad T\mu_g\gamma_k \leq 1/4 \tag{B.8}$$

which directly comes from 4.1a.

Proposition 5. If B.8 holds, then for each $k \in \mathbb{N}$, the following holds:

$$\mathbb{E}\left[d_{\mathcal{N}}\left(z_{k+1}, y_k^*\right)^2 \mid \mathcal{F}_k\right] \leq \left(C - 3T\mu_g\gamma_k/4\right)\mathcal{J}_k + T\gamma_k^2\sigma_g^2$$

Proof. We analyze one step iteration of the inner loop: for each $t = 0, \cdots, T-1$.

Using the Alexandrov space cosine law:

$$d_{\mathcal{N}}\left(z_k^{(t+1)}, y_k^*\right)^2 \le \frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right)}{\tanh\left(\sqrt{|\kappa|}d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right)\right)} d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right)^2 + \gamma_k^2 d_{\mathcal{N}}\left(h_{gz}^{k,t}\right)^2$$
$$- 2\gamma_k d_{\mathcal{N}}\left(h_{gz}^{k,t}\right) d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right) \cos(\angle(h_{gz}^{k,t}, z_k^{(t)}, y_k^*)),$$

Here, $z_{k+1} = z_k^{(T)}$ and $z_k = z_k^{(0)}$. Note that $\mathbb{E}\left[h_{gz}^{k,t}\right] = \nabla_y g\left(x_k, z_k^{(t)}\right) = \nabla_y g_k\left(z_k^{(t)}\right)$ where $g_k\left(z_k^{(t)}\right) := g\left(x_k, z_k^{(t)}\right)$. Taking expectation,

$$\mathbb{E}\left[d_{\mathcal{N}}\left(z_k^{(t+1)}, y_k^*\right)^2 \mid \mathcal{F}_k\right] \le \frac{\sqrt{|\kappa|}d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right)}{\tanh\left(\sqrt{|\kappa|}d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right)\right)} d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right)^2$$
$$+ \gamma_k^2 \sigma_g^2 - \gamma_k(1 - l_{g,1}\gamma_k)d_{\mathcal{N}}\left(\nabla g_k\left(z_k^{(t)}\right)\right) d_{\mathcal{N}}\left(z_k^{(t)}, y_k^*\right),$$

The strong convexity and smoothness of g_k imply the coercivity and co-coercivity (Nesterov et al., 2018), that is,

$$\max\left(\mu_g d_{\mathcal{N}}^2(z_k^{(t)}, y_k^*), \frac{1}{l_{g,1}}d_{\mathcal{N}}^2\left(\nabla g_k\left(z_k^{(t)}\right), \nabla g_k\left(y_k^*\right)\right)\right) \le d_{\mathcal{N}}\left(\nabla g_k\left(z_k^{(t)}\right), \nabla g_k\left(y_k^*\right), z_k^{(t)}, y_k^*\right)$$

Note that y_k^* minimizes $g_k(y)$. Use this to cancel out $\gamma_k^2 d_{\mathcal{N}}\left(\nabla g_k\left(z_k^{(t)}\right)\right)^2$, yielding

$$\mathbb{E}\left[d_{\mathcal{N}}^2(z_k^{(t+1)}, y_k^*) \mid \mathcal{F}_k\right] \le \frac{\sqrt{|\kappa|}d_{\mathcal{N}}(z_k^{(t)}, y_k^*)}{\tanh\left(\sqrt{|\kappa|}d_{\mathcal{N}}(z_k^{(t)}, y_k^*)\right)} d_{\mathcal{N}}^2(z_k^{(t)}, y_k^*)$$
$$+ \gamma_k^2 \sigma_g^2 - \gamma_k(1 - l_{g,1}\gamma_k)d_{\mathcal{N}}\left(\nabla g_k\left(z_k^{(t)}\right), \nabla g_k\left(y_k^*\right), z_k^{(t)}, y_k^*\right)$$
$$\le \left(\frac{\sqrt{|\kappa|}d_{\mathcal{N}}(z_k^{(t)}, y_k^*)}{\tanh\left(\sqrt{|\kappa|}d_{\mathcal{N}}(z_k^{(t)}, y_k^*)\right)} - \frac{3\mu_g\gamma_k}{4}\right) d_{\mathcal{N}}^2(z_k^{(t)}, y_k^*) + \gamma_k^2 \sigma_g^2.$$

For this to hold we need step-size condition B.8. We can repeat this T times and get the result. Here we're using $\frac{\sqrt{|\kappa|}d_{\mathcal{N}}(z_k^{(t)}, y_k^*)}{\tanh\left(\sqrt{|\kappa|}d_{\mathcal{N}}(z_k^{(t)}, y_k^*)\right)} < C$ for some $C > 0$ for all values of t. $\square$

B.4 Proof of Theorem 2

Let us revisit the potential function V_k within the Riemannian context:

$$V_{k+1} - V_k = F(x_{k+1}) - F(x_k) + \lambda_{k+1}l_{g,1}\mathcal{I}_{k+1} - \lambda_k l_{g,1}\mathcal{I}_k$$
$$+ \frac{\lambda_{k+1}l_{g,1}}{2}\mathcal{J}_{k+1} - \frac{\lambda_k l_{g,1}}{2}\mathcal{J}_k,$$

Utilizing an adaptation of Proposition 1 and reorganizing terms, we obtain:

$$\mathbb{E}[V_{k+1} - V_k \mid \mathcal{F}_k] \leq -\frac{\xi\alpha_k}{2}\|\operatorname{grad}F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\mathbb{E}[\|q_k^x\|^2 \mid \mathcal{F}_k] + \frac{\xi\alpha_k}{2}\cdot 3C_\lambda^2\lambda_k^{-2}$$
$$+ \frac{\xi^2 l_{F,1}}{2}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2)$$
$$+ l_{g,1}\underbrace{\mathbb{E}[\lambda_{k+1}\mathcal{I}_{k+1} + \frac{\lambda_k T\beta_k\mu_g}{16}d_\mathcal{N}(y_{k+1}, y_{\lambda,k}^*)^2 - \lambda_k\mathcal{I}_k \mid \mathcal{F}_k]}_{(i)}$$
$$+ \frac{l_{g,1}}{2}\underbrace{\mathbb{E}[\lambda_{k+1}\mathcal{J}_{k+1} + \frac{\lambda_k T\gamma_k\mu_g}{32}d_\mathcal{N}(z_{k+1}, y_k^*)^2 - \lambda_k\mathcal{J}_k \mid \mathcal{F}_k]}_{(ii)},$$

From proposition 2 and $\lambda_{k+1} = \lambda_k + \delta_k$, we get:

$$(i) \leq \lambda_k\left(\left[\frac{\sqrt{|\kappa|}d_\mathcal{N}(y_{\lambda,k+1}^*, y_{\lambda,k}^*)}{\tanh(\sqrt{|\kappa|}d_\mathcal{N}(y_{\lambda,k+1}^*, y_{\lambda,k}^*))}\right] + \frac{5T\beta_k\mu_g}{16} + \frac{\delta_k}{\lambda_k}\right)\mathbb{E}\left[d_\mathcal{N}^2(y_{k+1}, y_{\lambda,k}^*) \mid \mathcal{F}_k\right] - \lambda_k\mathcal{I}_k$$
$$\underbrace{+ O\left(\xi^2 l_{\lambda,0}^2\right)\frac{\lambda_k\alpha_k^2}{\mu_g T\beta_k}d^2(q_k^x, 0) + O\left(\xi^2 l_{*,0}^2\right)\lambda_k(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + O\left(\frac{l_{f,0}^2}{\mu_g^3}\right)\frac{\delta_k}{\lambda_k^2}}_{(iii)}.$$

Given the step-size rules B.7, we obtain:

$$(i) \leq \lambda_k\left(\left[\frac{\sqrt{|\kappa|}d_\mathcal{N}(y_{\lambda,k+1}^*, y_{\lambda,k}^*)}{\tanh(\sqrt{|\kappa|}d_\mathcal{N}(y_{\lambda,k+1}^*, y_{\lambda,k}^*))}\right] + \frac{T\beta_k\mu_g}{2}\right)\mathbb{E}\left[d_\mathcal{N}^2(y_{k+1}, y_{\lambda,k}^*) \mid \mathcal{F}_k\right] - \lambda_k\mathcal{I}_k + (iii).$$

Leveraging Proposition 3 within the framework to estimate $d_\mathcal{N}^2(y_{k+1}, y_{\lambda,k}^*)$, we derive:

$$(i) \leq -\frac{\lambda_k T\mu_g\beta_k}{4}\mathcal{I}_k + O\left(\xi^2 l_{*,0}^2\right)\frac{\alpha_k}{\mu_g T} + (iii)$$
$$= -\frac{\lambda_k T\mu_g\beta_k}{4}\mathcal{I}_k + O\left(\xi^2 l_{*,0}^2\right)\frac{\alpha_k}{\mu_g T} + O\left(T + \xi^2 l_{*,0}^2\right)\lambda_k(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + O\left(\frac{l_{f,0}^2}{\mu_g^3}\right)\frac{\delta_k}{\lambda_k^2}.$$

Given the inequality $(1 + a/2)(1 - 3a/4) \leq 1 - a/4$ for $a > 0$, we estimate the term (ii) using Proposition 4:

$$(ii) \leq \lambda_k\left(\frac{\sqrt{|\kappa|}d_\mathcal{N}(y_{k+1}^*, y_k^*)}{\tanh(\sqrt{|\kappa|}d_\mathcal{N}(y_{k+1}^*, y_k^*))} + \frac{\delta_k}{\lambda_k} + \frac{3T\gamma_k\mu_g}{8} + \frac{\lambda_k T\beta_k\mu_g}{32}\right)\mathbb{E}\left[d_\mathcal{N}(z_{k+1}, y_k^*)^2 \mid \mathcal{F}_k\right] - \lambda_k\mathcal{J}_k$$
$$\underbrace{+ O\left(\xi^2 l_{*,0}^2\right)\frac{\lambda_{k+1}\alpha_k^2}{T\mu_g\gamma_k}\|q_k^x\|^2 + O\left(\xi^2\lambda_{k+1}l_{*,0}^2\right)\left(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right)}_{(iv)}.$$

Assuming $\beta_k \leq \gamma_k$, and thus $\delta_k/\lambda_k < T\mu_g\gamma_k/32$, we have:

$$(ii) \leq \lambda_k \left(\frac{\sqrt{|\kappa|} d_{\mathcal{N}}(y_{k+1}^*, y_k^*)}{\tanh(\sqrt{|\kappa|} d_{\mathcal{N}}(y_{k+1}^*, y_k^*))} + \frac{T \gamma_k \mu_g}{2} \right) \mathbb{E} \left[d\left(z_{k+1}, y_k^* \right)^2 \mid \mathcal{F}_k \right] - \lambda_k \mathcal{J}_k + (iv).$$

Following the argument for (i), Proposition 5 provides:

$$(ii) \leq -\frac{\lambda_k T \mu_g \gamma_k}{4} \mathcal{J}_k + O\left(\xi^2 l_{*,0}^2 \right) \frac{\alpha_k \beta_k}{T \mu_g \gamma_k} \|q_k^x\|^2 + O\left(\xi^2 \lambda_k l_{*,0}^2 \right) \left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2 \right) + O\left(\lambda_k \right) T \gamma_k^2 \sigma_g^2.$$

Upon combining the bounds for (i) and (ii) and rearranging terms, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[V_{k+1} - V_k \mid \mathcal{F}_k\right] \leq &-\frac{\xi \alpha_k}{2} \|\nabla F(x_k)\|^2 + \frac{\xi \alpha_k}{2} \cdot 3 C_\lambda^2 \lambda_k^{-2} + \frac{\xi^2 l_{F,1}}{2} \left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2 \right) \\
&- \frac{\xi \alpha_k}{4} \left(1 - O\left(\frac{\xi l_{g,1} l_{*,0}^2 \beta_k}{\mu_g T \gamma_k} \right) - O\left(\frac{\xi l_{g,1} l_{*,0}^2}{\mu_g T} \right) \right) \mathbb{E}\left[\|q_k^x\|^2 \mid \mathcal{F}_k \right] \\
&- \frac{\lambda_k l_{g,1} T \mu_g \beta_k}{4} \mathcal{I}_k - \frac{\lambda_k l_{g,1} T \mu_g \gamma_k}{4} \mathcal{J}_k \\
&+ O\left(T + \xi^2 l_{*,0}^2 \right) \cdot l_{g,1} \lambda_k \left(\alpha_k^2 \sigma_f^2 + \left(\beta_k^2 + \gamma_k^2 \right) \sigma_g^2 \right) + O\left(\frac{l_{g,1} l_{f,0}^2}{\mu_g^3} \right) \frac{\delta_k}{\lambda_k^2}.
\end{aligned}
$$

A key requirement is that terms driven by $\mathbb{E}\left[d^2(q_k^x, 0) \right]$ remain negative. To ensure this, we impose:

$$
\text{(Step-size rules):} \quad \begin{aligned} \xi l_{g,1} l_{*,0}^2 \beta_k &\leq c_1 \mu_g T \gamma_k, \\ \xi l_{g,1} l_{*,0}^2 &\leq c_2 \mu_g T, \end{aligned} \tag{B.9}
$$

for some absolute constants $c_1, c_2 > 0$, which are achievable given $\beta_k \leq \gamma_k$ and condition (3b) with sufficiently small $c_\xi > 0$. Upon satisfying these conditions, we derive that:

$$
\begin{aligned}
\mathbb{E}\left[V_{k+1} - V_k \mid \mathcal{F}_k\right] \leq &-\frac{\xi \alpha_k}{2} \operatorname{grad}^2 F(x_k) - \frac{\lambda_k T \mu_g \gamma_k}{4} d_{\mathcal{N}}^2(z_k, y_k^*) - \frac{\lambda_k T \mu_g \beta_k}{4} d_{\mathcal{N}}^2(y_k, y_{\lambda,k}^*) \\
&+ O\left(\xi C_\lambda^2 \right) \frac{\alpha_k}{\lambda_k^2} + O\left(\frac{l_{g,1} l_{f,0}^2}{\mu_g^3} \right) \frac{\delta_k}{\lambda_k^2} + O\left(\xi^2 l_{F,1} \right) \left(\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2 \right) \\
&+ O\left(T + \xi^2 l_{*,0}^2 \right) \cdot l_{g,1} \lambda_k (\alpha_k^2 \sigma_f^2 + (\beta_k^2 + \gamma_k^2) \sigma_g^2).
\end{aligned}
$$

Summing over $k = 0$ to $K - 1$, and focusing on the dominant terms, given that $\sum_k \delta_k / \lambda_k^2 = O(1)$ (due to $\delta_k / \lambda_k = O(1/k)$ and $\lambda_k = \operatorname{poly}(k)$), leads us to the theorem conclusion.

Note: The effect of the sectional curvature κ here is negligible because we're implicitly using the approximation $x / \tanh(x) \approx 1 + x^2/3 + O(x^4)$. Terms like $\frac{\sqrt{|\kappa|} d(w_k, w_{k+1})}{\tanh(\sqrt{|\kappa|} d(w_k, w_{k+1}))} \approx 1 + |\kappa| d^2(w_k, w_{k+1})/3 + O(\cdot)$ appear in the Alexandrov space cosine law. Summing from $k = 0$ to $K - 1$ and focusing on dominant terms, the curvature κ's effect on the final result is negligible, since negative terms like $-\operatorname{grad}^2 F(x_k)$, $-d_{\mathcal{N}}^2(z_k, y_k^*)$, and $-d_{\mathcal{N}}^2(y_k, y_{\lambda,k}^*)$ do not affect the final inequality.

B.5 Proof of Corollary 3

We begin by establishing that the step-size design within the theorem ensures $\lambda_k = \gamma_k/(2\alpha_k)$ for all k. This follows from the initial condition $\lambda_0 = \gamma_0/(2\alpha_0)$ and, by mathematical induction, we derive:

$$\frac{T\mu_g}{16}\alpha_k\lambda_k^2 = \frac{T}{32}\frac{c_\gamma}{2c_\alpha}(k+k_0)^{-2c+a}$$

and

$$\frac{c_\gamma}{2c_\alpha}\left((k+k_0+1)^{a-c} - (k+k_0)^{a-c}\right) \leq \frac{(a-c)c_\gamma}{2c_\alpha}(k+k_0)^{-1-c+a}.$$

Given that $c \leq 1$ and $T \geq 32$, it holds that

$$\lambda_{k+1} = \frac{c_\gamma}{2c_\alpha}(k+k_0+1)^{a-c} = \frac{\gamma_{k+1}}{2\alpha_{k+1}}.$$

By applying the step-size designs to a manifold, we obtain:

$$\sum_{k=0}^{K-1} \frac{\mathbb{E}[\|\operatorname{grad} F(x_k)\|^2]}{(k+k_0)^a} \leq O_P(1) \cdot \sum_k \frac{1}{(k+k_0)^{3a-2c}}$$
$$+ O_P(\sigma_f^2) \cdot \sum_k \frac{1}{(k+k_0)^{a+c}}$$
$$+ O_P(\sigma_g^2) \cdot \sum_k \frac{1}{(k+k_0)^{3c-a}} + O_P(1).$$

The choices of rates $a, c \in [0,1]$ depend on the specific stochasticity of the gradients. Letting $b = a - c$, and with the step-size design, $\lambda_k = \gamma_k/(2\alpha_k) = O(k^b)$. Considering a random variable R uniformly distributed over $\{0, 1, ..., K\}$, the inequality is reframed as:

$$\frac{K}{(K+k_0)^a}\mathbb{E}[\|\operatorname{grad} F(x_R)\|^2] \geq K^{1-a} \cdot \mathbb{E}[\|\operatorname{grad} F(x_R)\|^2]$$

We examine three scenarios based on the stochasticity in the upper and lower-level objectives:

1. **Stochastic in both objectives ($\sigma_f^2, \sigma_g^2 > 0$):** Setting $a = 5/7, c = 4/7$ leads to $\lambda_k = O(k^{1/7})$. The dominating term becomes $O(\log K)$, resulting in:

$$\mathbb{E}[\|\operatorname{grad} F(x_R)\|^2] = O\left(\frac{\log K}{K^{2/7}}\right).$$

2. **Stochastic only in the upper-level ($\sigma_f^2 > 0, \sigma_g^2 = 0$):** Here, $a = 3/5, c = 2/5$ is chosen, simplifying to:

$$\mathbb{E}[\|\operatorname{grad} F(x_R)\|^2] = O\left(\frac{\log K}{K^{2/5}}\right).$$

3. **Deterministic case ($\sigma_f^2 = 0, \sigma_g^2 = 0$):** With $a = 1/3, c = 0$, we find:

$$\| \operatorname{grad} F(x_K) \|^2 = O\left(\frac{\log K}{K^{2/3}}\right).$$

This proof adaptation ensures that the step-size and λ_k designs are tailored for the geometric complexities of Riemannian manifolds, thereby facilitating convergence under various stochastic settings.