Curvature-Independent Last-Iterate Convergence for Games on Riemannian Manifolds

Yang Cai
Yale University
yang.cai@yale.edu

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Tianyi Lin
University of California, Berkeley
darren_lin@berkeley.edu

Argyris Oikonomou
Yale University
argyris.oikonomou@yale.edu

Emmanouil V. Vlatakis-Gkaragkounis
University of California, Berkeley
emvlatakis@berkeley.edu

Abstract

Numerous applications in machine learning and data analytics can be formulated as equilibrium computation over Riemannian manifolds. Despite the extensive investigation of their Euclidean counterparts, the performance of Riemannian gradient-based algorithms remain opaque and poorly understood. We revisit the original scheme of Riemannian gradient descent (RGD) and analyze it under a *geodesic monotonicity* assumption, which includes the well-studied *geodesically convex-concave* min-max optimization problem as a special case. Our main contribution is to show that, despite the phenomenon of *distance distortion*, the RGD scheme, with a step size that is agnostic to the manifold's curvature, achieves a *curvature-independent* and *linear last-iterate* convergence rate in the geodesically strongly monotone setting. To the best of our knowledge, the possibility of curvature-independent rates and/or last-iterate convergence in the Riemannian setting has not been considered before.

1 Introduction

Game-theoretic concepts permeate a multitude of machine learning (ML) problem domains, ranging from conventional constrained optimization [7, 69, 91, 95] and feature selection tasks [4, 34, 53] to burgeoning fields such as adversarial training [45, 68, 81] and multi-agent system dynamics [18, 117]. In these research thrusts, the prevailing approach involves formulating the optimal solution of the pertinent task as a Nash equilibrium [89, 103] of some carefully constructed zero-sum, cooperative, or general-sum N-player game. A deeper exploration of the optimization literature, however, reveals that to secure tight convergence guarantees, the prevailing methods tend to subtly posit the assumption of a "nearly" concave/monotone game structure whose strategy space is constrained to be some convex set [20, 31, 32, 96].

This is problematic, because in many real-world applications, spanning from optimal transport [78, 79, 97] to bounded value problems in environmental engineering [64, 92] and robotics especially those involving complex interactions among multiple agents [9, 10, 99, 107], the game can exhibit nonconvexities both in the players' utilities and their strategy spaces, posing significant challenges for conventional approaches [23, 27, 84, 118], and often compelling recourse to ad hoc heuristics [17, 33, 51, 114]. Thus, a key objective is to devise algorithms that systematically exploit the unique geometric structure of the feasible set [80, 82, 98]. Attempts at providing such geometric foundations can be found in a number of recent paperes [see, e.g., 3, 5, 25, 35, 41, 42, 48, 54, 55, 83, 91, 101, 108, 119]. A notable outcome of this line of work is the interpretation of constraints in game theory and equilibrium computation through the lens of *Riemannian manifolds*.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

The Riemannian framework, commonly utilizing Hadamard and Stiefel manifolds, has been productive in a wide range of other statistical tasks, from online principal component analysis [71] to diffusion tensor data processing [74], and maximum likelihood estimation for certain heavy-tailed non-Gaussian distributions [125]. It remains a significant challenge, however, to bring Riemannian methods to bear in game-theoretic scenarios, especially those characterized by nonconvex strategy spaces, including multi-agent robotic systems where the feasible joint angles of robotic arms use a SO(3) rotation manifold [47, 113, 121], environmental pollution control games where the physical transmission of pollutants along bounded surface areas represents a manifold constraint [88, 106], and optimal transport problems via robust Wasserstein barycenters [24, 63, 109]. Some of the specific challenges arising in such domains include:

1. **Existence of Solution Concept**. One fundamental hurdle is that the problem may not possess solutions. This contrasts starkly with minimization tasks where, given a bounded domain, the existence of an optimal solution is always assured [36]. The establishment of Nash equilibria, on the other hand, typically calls for the application of topological fixed-point theorems [15, 58], whose core requirement is the convexity of the strategy space. Formally, the Nash Equilibrium problem in an N-player game with (non-Euclidean) strategy spaces $(\mathcal{M}_i)_{i\in[N]}$ is formulated by identifying a point $y^* = (y_1^*, \cdots, y_N^*)$ satisfying the following condition:

$$\text{loss}_i(y^*) \leq \text{loss}_i(y_i', y_{-i}^*) \quad \forall i \in [N] \quad \forall y_i' \in \mathcal{M}_i,$$

   where we employ the standard shorthand (y_i, y_{-i}) to denote the action of the i-th player and the actions of all other players in the game.

   A guarantee for the existence of Nash equilibria in geodesically concave games over Hadamard manifolds was introduced by Németh [90] and extended by Li [76], employing the framework of variational inequalities. Following this, [100] further broadened the scope by introducing the notion of local Nash equilibria in games situated on either finite-dimensional differentiable manifolds or infinite-dimensional Hilbert spaces. In a more recent advance, [130] proved a strong duality theorem for min-max optimization, applicable across a variety of Riemannian manifolds that are equipped with unique geodesics.

2. **Convergence Guarantees**. A general understanding of the efficiency and convergence properties of game dynamics over the Riemannian manifolds remains elusive. To fill this lacuna, an emerging body of research is devoted to extending standard game dynamics and optimization algorithms such as Gradient Descent Ascent (GDA)[49] and Extragradient (EG)[56, 130] to variational inequalities, and tailoring the resulting algorithms to the Riemannian setting.

   Most of the aforementioned research focuses on min-max scenarios. For instance, [49] examined the Riemannian Gradient Descent Ascent (RGDA) for nonconvex-nonconcave settings, but their findings only offer sub-optimal rates and merely best/average-iterate convergence. [46] analyzed the Riemannian Hamiltonian Method (RHM), equating its performance with second-order methods in Euclidean settings, but the practical application of such methods is limited due to the computational expense of handling second-order derivatives in large-scale optimization problems. On a promising note, [57] demonstrated that the Riemannian Corrected Extragradient (RCEG) method achieves last-iterate convergence at a linear rate in the geodesically strongly-convex-concave case, thereby matching the results in the Euclidean regime. Their findings also extend to the stochastic or non-smooth case, where RCEG and RGDA achieve near-optimal convergence rates, albeit with factors dependent on the curvature of the manifold. For a more comprehensive discussion of related work in both Euclidean and Riemannian settings, we refer the interested reader to our supplementary material.

When investigating multi-player games where strategy spaces are represented as manifolds, several significant challenges arise. *Firstly*, we need to understand the appropriate solution concept in this nonconvex regime. *Secondly*, it is crucial to identify meaningful structure in the game so that the solution concepts are tractable using computationally lightweight methods—e.g., first-order methods—in the face of nonconvexity.[1] *Lastly,* from both the optimization and the equilibrium perspective, it is desirable to construct algorithms or game dynamics that remain uncoupled and more importantly model-free; that is, they remain unaffected by transformations that alter the curvature of the underlying manifold.

[1]Note that the problem is intractable without any structural assumption on the game [30].

1.1 Our Contributions

Our work aims to shed new light on the problem of finding game-theoretic equilibria on Riemannian manifolds. Specifically, we consider finding Nash equilibria in *geodesically monotone Riemannian games* (Definition 2)—a new family of games proposed in this paper that captures both the well-studied *monotone games* [102] and geodesically convex-concave min-max optimization as special cases. In a geodesically monotone Riemannian game, players choose their actions from a Riemannian manifold, and their loss functions jointly satisfy the geodesic monotonicity condition.

We revisit the Riemannian Gradient Descent (RGD) Dynamics, the leading method in Riemannian optimization, in geodesically monotone Riemannian games.

> **Main Contribution:** For geodesically strongly monotone Riemannian games, we show that the RGD exhibits *curvature-independent* and *linear last-iterate convergence rate* to the Nash equilibrium. Additionally, our instantiation of RGD (Algorithm 1) is completely agnostic to the curvature of the manifold.

To the best of our knowledge, this is the first curvature-independent linear convergence rate for even the special case of min-max manifold optimization, and hence answers an open question raised in [57]. We showcase our finding in both the deterministic and stochastic/mini-batch setting (Theorem 1).

Organization. In Section 2, we provide some necessary background on manifold theory. In Section 3, we supply the definition of geodesically monotone Riemannian games and present our performance function. Section 4 is devoted to the analysis of RGD in geodesically strongly monotone Riemannian games. Finally, Section 5 concludes our work with a discussion of future directions and challenges. We defer the full proofs of our results as well as further discussion on applications to the Appendix. In Appendix A we describe some limitations of our work.

1.2 Related work on games on Riemmanian manifolds

Focusing on min-max problems on Riemannian manifolds, existing research has mainly been devoted to: (1) studying the existence and uniqueness of equilibria; (2) computation of equilibrium concepts. The former results have been established for the special case of Hadamard manifolds [65, 67, 93] and then generalized to finite-dimensional manifolds with unique geodesics [130]. Similarly, algorithms for computing the equilibrium were first developed for the special case of hyperbolic Hadamard manifolds with negative curvature [38, 75, 120]. However, only asymptotic convergence guarantees were provided. It was not until recently that a nonasymptotic convergence rate guarantee has been derived for Riemannian gradient-based algorithms when applied to min-max optimization problems. In the geodesically smooth and strongly-convex-strongly-concave setting, the best known method achieves a last-iterate and **curvature-dependent** global rate [57]. In the geodesically convex-concave setting, the best known method only achieves a **time-average** global rate [57, 130]. To the best of our knowledge, we provide the first set of results that fill in a gap of the equilibrium computation for games on Riemannian manifolds, where the possibility of curvature-independent and/or last-iterate convergence rates for gradient-based algorithms has been unexplored before.

In the supplementary material we provide a further discussion of related theoretical work and applications.

2 Preliminaries & Notation

The topological definitions of manifolds and their analytic constructions have given rise to a vast literature which we cannot hope to review here; for a basic introduction, see [16, 70] and references therein. Here, we review the key definitions and tools needed for the study of optimization on manifolds:

A n-dimensional *smooth manifold* $\mathcal{M}$ is a topological space that is locally homeomorphic to $\mathbb{R}^n$ and whose transition maps—the local homeomorphisms that cover the manifold—are smooth functions, meaning they are infinitely differentiable. A *Riemannian manifold* $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ equipped with a *Riemannian metric* g, which is a smoothly varying positive-definite symmetric bilinear form on the tangent spaces of $\mathcal{M}$. Intuitively, for each point $y \in \mathcal{M}$, the Riemannian metric

g assigns an inner product $\langle \cdot, \cdot \rangle_y$ and correspondingly a n-dimensional real vector space $\mathcal{T}_y\mathcal{M}$, called the *tangent space* at y. The tangent space $\mathcal{T}_y\mathcal{M}$ at a point $y \in \mathcal{M}$ is a real vector space consisting of all tangent vectors to smooth curves in $\mathcal{M}$ passing through y. The dimension of the tangent space $\mathcal{T}_y\mathcal{M}$ is equal to the dimension of the manifold $\mathcal{M}$. We denote it by $\mathcal{TM} = \cup_{y \in \mathcal{M}} \mathcal{T}_y\mathcal{M}$. Tangent spaces provide a means to analyze the local geometry of a manifold near a given point.

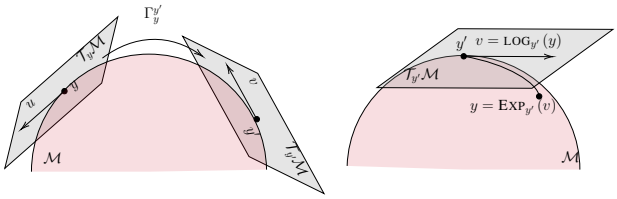

Figure 1: Illustative example of tangent spaces T_pM, Parallel transport and $\text{EXP}_q(\cdot)$ and $\text{LOG}_q(\cdot)$ mapping.

2.1 Geodesics, Parallel Transport, Exponential and Logarithmic Maps

A geodesic is a smooth curve $\gamma : [0, 1] \to \mathcal{M}$ in a Riemannian manifold $\mathcal{M}$ that locally minimizes distance and has the property that its tangent vector remains parallel along the curve with respect to the Riemannian connection.

Parallel transport is a means of transporting vectors along curves while preserving their length and direction with respect to the Riemannian metric g. More formally, given a smooth curve $\gamma : [0, 1] \to \mathcal{M}$, starting at $\gamma(0) = y$, ending at $\gamma(1) = y'$ and a tangent vector $v \in \mathcal{T}_y\mathcal{M}$ at the starting point y, parallel transport defines a smooth vector field $\Gamma_y^{\gamma(t)}(v)$ along the curve γ such that $\Gamma_y^y(v) = v$ and the covariant derivative of $\Gamma_y^{\gamma(t)}(v)$ with respect to the curve's tangent vector field is zero, i.e., $\nabla_{\dot{\gamma}(t)}\Gamma_y^{\gamma(t)}(v) = 0$ for all $t \in [0, 1]$. The parallel transport of v along the curve γ to the endpoint $\gamma(1)$ is then given by the vector $\Gamma_y^{y'}(v) \in \mathcal{T}_{y'}\mathcal{M}$. Parallel transport is dependent on both the curve γ and the Riemannian metric g, and it provides a natural way to compare tangent vectors at different points in the manifold, while taking the manifold's geometry into account.

The exponential map $\text{EXP}_y(\cdot) : \mathcal{T}_y\mathcal{M} \to \mathcal{M}$ and the logarithmic map $\text{LOG}_y(\cdot) : \mathcal{M} \to \mathcal{T}_y\mathcal{M}$ are essential tools in connecting the local geometry of a Riemannian manifold $(\mathcal{M}, g)$ with the tangent space $\mathcal{T}_y\mathcal{M}$. The exponential map maps a tangent vector $v \in \mathcal{T}_y\mathcal{M}$ to a point in the manifold along the geodesic starting at y. Mathematically, if $\gamma_v : [0, 1] \to \mathcal{M}$ is the geodesic with $\gamma_v(0) = y$ and $\dot{\gamma}_v(0) = v$, then $\text{EXP}_y(v) = \gamma_v(1)$. The logarithmic map $\text{LOG}_y(\cdot)$ is the inverse operation, mapping a point $y' \in \mathcal{M}$ to the tangent vector $v \in \mathcal{T}_y\mathcal{M}$ such that $\text{EXP}_y(v) = y'$. Note that the logarithmic map is well-defined when the manifold is equipped with a unique geodesic. The norm of $\text{LOG}_y(y')$ corresponds to the geodesic distance between points y and y: $|\text{LOG}_y(y')| = |\text{LOG}_{y'}(y)| = d_{\mathcal{M}}(y, y')$.

Throughout this paper, we only consider Riemannian manifolds $\mathcal{M}$, such that for any pair of points y, y' in $\mathcal{M}$, there exists a unique minimizing geodesic within $\mathcal{M}$ that connects them. We also make use of the following properties of the parallel transport and logarithmic map:

- For any triplet of points $y, y', y'' \in \mathcal{M}$ that lie in the same geodesic, and any vector u in the tangent space $\mathcal{T}_y$ of y, we have $\Gamma_{y'}^{y''}\left(\Gamma_y^{y'}u\right) = \Gamma_y^{y''}u$.
- For any pair of points $y, y' \in \mathcal{M}$, we have $\Gamma_y^{y'}\text{LOG}_y(y') = -\text{LOG}_{y'}(y)$.

Moreover, for a point $y \in \mathcal{M}$ and a radius $D > 0$, we use the notation $d_{\mathcal{M}}(y, D) = \{y' \in \mathcal{M} : d_{\mathcal{M}}(y, y') \leq D\}$.

2.2 Geodesic Convex Analysis

We begin with the definition of a geodesically convex set and function, respectively:

- The function f is called *geodesically strongly convex* with the modulus $\mu > 0$ if the following statement holds:

$$f(y') \geq f(y) + \langle \text{grad}_y f(y), \text{LOG}_y(y') \rangle_y + \tfrac{\mu}{2}(d_{\mathcal{M}}(y, y'))^2, \text{ for each } y, y' \in \mathcal{M}, \tag{1}$$

where $\text{grad}_y f(y) \in \mathcal{T}_y\mathcal{M}$ is a Riemannian gradient of f at a point y.

Finally, in our work we adopt the following smoothness conditions for the utilities of our games, conventionally used in Riemannian optimization [2, 129, 130]:

- A function $f : \mathcal{M} \to \mathbb{R}$ is *geodesically L-Lipschitz* if for $\forall y, y' \in \mathcal{M}$ it holds that
$$|f(y) - f(y')| \leq L d_{\mathcal{M}}(y, y') \tag{2}$$
- Additionally, if function f is also differentiable, it is called *geodesically L-smooth* if for $\forall y, y' \in \mathcal{M}$ it holds that
$$\| \text{grad}_y f(y) - \Gamma^y_{y'} \text{grad}_y f(y') \| \leq L \cdot d_{\mathcal{M}}(y, y'), \tag{3}$$
where $\text{grad}_y f(y'), \in \mathcal{T}_{y'}\mathcal{M}$ is the Riemannian gradient of f at y'.

3 Geodesically Monotone Riemannian Games

In this section, we consider games where players' actions lie on a Riemannian manifold, and we extend Rosen's monotone games to this setting. Rosen's original formulation considers players whose actions lie in Euclidean spaces [102]. This new formulation allows us to study games where the action space for player i has a geometric structure, represented by the Riemannian manifold $\mathcal{M}_i$. We extend this concept to accommodate more diverse geometries, broadening the range of potential applications, as illustrated in Section 3.2. The missing proofs in this section are presented in Appendix D.

Definition 1 (Games on Riemannian Manifolds)**.** A game on Riemannian manifold consists of the following components:

1. A finite set of players $\mathcal{N} = 1, 2, \ldots, N$.
2. For each player $i \in \mathcal{N}$, there is a Riemannian manifold $\mathcal{M}_i$. Player i's action set is $\mathcal{M}_i$, and we denote the joint action profile by $\mathcal{M} = \Pi_{i \in \mathcal{N}} \mathcal{M}_i$,.
3. For each player $i \in \mathcal{M}$, a loss function $l_i : \mathcal{M} \to \mathbb{R}$ is defined, which assigns a real-valued loss to every combination of strategies in the joint strategy space $\mathcal{M}$. We let $\mathcal{L} = \{l_i\}_{i \in \mathcal{N}}$.

A Riemannian game is smooth if all loss functions l_i's are smooth with respect to the Riemannian metric on the manifold $\mathcal{M}$. We denote by $F(y_1, y_2, \ldots, y_N) = (\text{grad}_{y_1} l_1(y), \text{grad}_{y_2} l_2(y), \ldots, \text{grad}_{y_N} l_N(y)) : \mathcal{M} \to \mathcal{T}\mathcal{M}$ the vector of concatenated gradients of the players' loss functions with respect to the Riemannian manifold at the joint strategy profile $y \in \mathcal{M}$.

The notion of a Nash equilibrium in a Riemannian game remains similar to the one in the Euclidean space: a strategy profile $y^* \in \mathcal{M}$ is a Nash Equilibrium if no player can unilaterally decrease their loss by deviating from their current strategy, given the strategies of the other players. Formally, y^* is an equilibrium if
$$l_i(y^*) \leq l_i(y'_i, y^*_{-i}) \tag{4}$$
for all players $i \in \mathcal{N}$, all strategies $y'_i \in \mathcal{M}_i$. In the following lemma, we provide a necessery condition for a strategy profile to be a Nash equilibrium of a arbitrary Riemannian games.

Lemma 1. *A necessery condition for the strategy profile $y^* \in \mathcal{M}$ to be a Nash equilibrium of a smooth Riemannian game, is that $\|F(y^*)\|_{y^*} = 0$.*

Definition 2 (Geodesically Monotone Riemannian Games)**.** A geodesically μ-strongly monotone Riemannian game satisfies the following monotonicity condition: for any two action profiles $y, y' \in \mathcal{M}$, $\langle \Gamma^y_{y'} F(y') - F(y), \text{LOG}_y(y') \rangle_y \geq \mu \cdot d_{\mathcal{M}}(y, y')^2$, where $d_{\mathcal{M}}$ is the Riemannian distance on manifold $\mathcal{M}$. When $\mu = 0$, we refer to the game as a geodesically monotone Riemannian game.

In the next lemma, we show that geodesically monotone Riemannian games are a generalization of geodesically convex, geodesic concave min-max optimization.

Lemma 2. *Consider a two player Riemannian game, where $y_1 \in \mathcal{M}_1$ and $y_2 \in \mathcal{M}_2$ are the actions of player 1 and 2 respectively. The loss of player 1 is $u(y_1, y_2) : \mathcal{M}_1 \times \mathcal{M}_2 \to \mathbb{R}$ and the loss of player 2 is $-u(y_1, y_2)$. If u is μ-strongly geodesically convex in y_1, and μ-strongly geodesically concave in y_2, then the Riemannian game is μ-strongly geodesically monotone.*

Consider a strategy profile $y \in \mathcal{M}$ and a radius $D > 0$. The standard measures of proximity to a Nash equilibrium in games in Euclidean spaces are the gap function and the total gap function. The gap function is inspired by the variational characterization of the game's solution, as presented in Corollary 1, and the total gap function can be interpreted as the cumulative suboptimality gaps of the players. Note that a total gap of value zero indicates that the action profile is a Nash equilibrium.

Definition 3 (Duality Gap and Total Gap in Riemannian Games)**.** The duality gap and total gap for a radius $D > 0$ are defined as follows:

$$gap_D(y) = \max_{y' \in B_{\mathcal{M}}(y,D)} \langle F(y), -\text{LOG}_y(y') \rangle_y,$$

$$Tgap_D(y) = \sum_{i \in \mathcal{N}} \left(l_i(y) - \min_{y_i' \in B_{\mathcal{M}_i}(y_i,D)} l_i(y_i', y_{-i}) \right)$$

In the subsequent lemma, we establish that the duality gap provides an upper bound for the total gap. One notable consequence of this lemma is the characterization of solutions for monotone Riemannian games.

Lemma 3 (Adapted from [43, 44])**.** *For any monotone Riemannian game, the following inequality hold for all $D > 0$ and $y \in \mathcal{M}$:*

$$Tgap_D(y) \leq gap_{\sqrt{N} \cdot D}(y).$$

Corollary 1. *A strategy profile $y^* \in \mathcal{M}$ is a Nash equilibrium of a monotone Riemannian game, if and only if $\|F(y^*)\|_{y^*} = 0$.*

3.1 Measuring proximity to Nash equilibrium in a monotone Riemannian game

We measure the proximity of a strategy profile y to a Nash equilibrium by evaluating the norm of the loss function's gradient $\|F(y)\|_y$. In the following lemma, we show that the norm of the gradient bounds the gap function, which therefore also bounds the total gap due to Lemma 3.

Lemma 4. *For any monotone Riemannian games, the following inequality hold for all $D > 0$ and $y \in \mathcal{M}$,*

$$gap_D(y) \leq D \cdot \|F(y)\|_y.$$

3.2 Applications of Riemmanian Games

We provide two examples of Riemmanian games to give a sense of their expressivity. One of the examples is a generic model from the optimization and game theory literature [87, 104] and the other is a formalization of a class of applied problems in economics and statistical machine learning [39, 60, 94].

Example 3.1 (Potential game on Riemannian manifolds)**.** We refer to a Riemmanian game $(\mathcal{N}, \mathcal{S} = \prod_{i=1}^N \mathcal{S}_i, \{u_i\}_{i=1}^N)$ as a potential game if there exists a potential function $f : \mathcal{S} \mapsto \mathbb{R}$ such that

$$u_i(y_i, y_{-i}) - u_i(\tilde{y}_i, y_{-i}) = f(y_i, y_{-i}) - f(\tilde{y}_i, y_{-i}),$$

for all $i \in \mathcal{N}$, all $y \in \mathcal{S}$ and all $\tilde{y}_i \in \mathcal{S}_i$. If the potential function f is geodesically strongly concave, we have $(\mathcal{N}, \mathcal{S}, \{u_i\}_{i=1}^N)$ is a geodesically strongly monotone game.

Example 3.2 (Robust matrix Karcher mean problem)**.** We consider a robust version of classical matrix Karcher mean problem. More specifically, the Karcher mean of N symmetric positive definite matrices $\{A_i\}_{i=1}^N$ is defined as the matrix $X \in \mathcal{M} = \{X \in \mathbb{R}^{n \times n} : X \succ 0, X = X^\top\}$ that minimizes the sum of squared distance induced by the Riemannian metric:

$$d(X, Y) = \|\log(X^{-1/2} Y X^{-1/2})\|_F.$$

The loss function is thus defined by

$$f(X; \{A_i\}_{i=1}^N) = \sum_{i=1}^N (d(X, A_i))^2,$$

which is known to be nonconvex in Euclidean spaces but is geodesically strongly convex. Then, the robust version of classical matrix Karcher mean problem aims at solving the following problem:

$$\min_{X\in\mathcal{M}}\max_{Y_i\in\mathcal{M}} f(X;\{Y_i\}_{i=1}^N) - \gamma\left(\sum_{i=1}^N (d(Y_i,A_i))^2\right),$$

where $\gamma > 0$ encodes the tradeoff between the computation of Karcher mean over a set of $\{Y_i\}_{i=1}^N$ and the difference between the observed samples $\{A_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$. It is clear that the above problem is a geodesically strongly-convex-strongly-concave min-max optimization problem.

In addition to these examples, it is worth mentioning that Riemmanian games contain all general games in Euclidean spaces and both minimization and maximization optimization problems in geodesic spaces. A typical class of examples consists in min-max optimization in geodesic spaces which abstracts many machine learning problems, e.g,. principal component analysis [13], dictionary learning [110, 111], deep neural networks (DNNs) [50] and low-rank matrix learning [52, 116]. In particular, the problem of principal component analysis can be formulated as optimization on a Grassmann manifold.

4 Curvature-Independent Convergence Rates for Gradient Descent

In this section, we show that the Riemannian Gradient Descent, formally defined in Algorithm 1, attains a curvature-independent linear convergence rate for strongly monotone and smooth Riemannian games. This holds true in two cases: (i) when we have exact access to the gradients of the loss functions, and (ii) when our access to the gradients is only stochastic but we can use mini-batches to reduce the variance.[2] The assumptions we place on the Riemannian game under consideration in this section are summarized in Assumption 1. Note that we do not need to assume that a Nash equilibrium exists, because the existence is already implied by Assumption 1, as shown in Lemma 6. We postpone the proofs in this section to Appendix E.

Assumption 1. Throughout this section we make the following assumptions:

- The Riemannian game is μ-strongly monotone and L-smooth, i.e., each player i's loss function l_i is L-smooth for any pair of action profiles y, y' in $\mathcal{M}$, there exists a unique minimizing geodesic within $\mathcal{M}$ that connects them. We denote the condition number of the game by $\kappa = \frac{\mu}{L}$.
- The Riemannian manifold and the metric $(d_{\mathcal{M}}(\cdot,\cdot) : \mathcal{M}\to\mathbb{R}, \mathcal{M})$ form a complete metric space.
- In the stochastic setting, we have an upper bound $B \geq d_{\mathcal{M}}(y_0, y^*)$ for the initial strategy y_0.

Algorithm 1: (Stochastic) Riemannian Gradient Descent

Require:

- Riemannian game $(\mathcal{N}, \mathcal{M}, \mathcal{L})$.
- Access to an unbiased estimator of the gradient $F : \mathcal{M} \to \mathcal{T}\mathcal{M}$ with variance σ^2.
- Initial strategy profile $= y_0 \in \mathcal{M}$, maximum number of iterations K, step size schedule $\{\eta_k \geq 0\}_{k\in[K]}$, mini-batch schedule $\{m_k \in \mathbb{N}\}_{k\in[K]}$.

for $k = 0, 1, \ldots, K-1$ **do**
  Query an unbiased estimator of the gradient at y_k for m_k times, and let $g^{(k)} \in \mathcal{T}_{y_k}\mathcal{M}$ be the average of the output of the m_k queries.
  Perform Riemannian gradient descent with step size η_k, $y_{k+1} \leftarrow \mathrm{Exp}_{y_k}\left(-\eta_k \cdot g^{(k)}\right) \in \mathcal{M}$.
end for
return the final strategy profile y_K.

The core of our proof rests on a descent inequality, detailed in Lemma 5. The descent inequality shows that the norm of the gradient contracts in both the deterministic case and the stochastic setting

[2] Consider m independent and identically distributed (i.i.d.) samples drawn from an unbiased estimator of F, with a variance denoted by σ^2. The empirical average of these samples has variance $\frac{\sigma^2}{m}$.

with a carefully tuned batch-size schedule. For clarity of exposition, we only include the proof for the deterministic case here and defer the proof for the stochastic case to the Appendix.

The transition to a stochastic setting presents an additional layer of complexity: bounding the term $\left\langle F(y_k) - g^{(k)}, \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) \right\rangle_{y_k}$. We employ Young's inequality to completely avoid the distance distortion. Contrasted with conventional approaches for Euclidean space, our methodology exacerbates the dependence on the condition number, introducing an additional factor of $\frac{1}{\kappa^2}$ as seen in Theorem 1, but allows our convergence rate to be curvature-independent.[3]

A slightly sharper bound can be derived for the deterministic case, as evidenced in the proof of Lemma 5. However, for the sake of clarity and coherence, we opted to present a unified bound applicable to both deterministic and stochastic settings.

Lemma 5. *For $\eta_k \leq \frac{2\mu}{L^2}$, conditioning on y_k, the iterates of Algorithm 1 satisfy the following descent inequality for every $k < K$:*

$$\left(1 - \frac{2\eta_k\mu - (\eta_k L)^2}{2 - 2\eta_k\mu + (\eta_k L)^2}\right) \|F(y_k)\|_{y_k}^2 + \frac{4 \cdot \sigma^2}{m_k \cdot (2\eta_k\mu - (\eta_k)^2)} \geq \mathbb{E}\left[\|F(y_{k+1})\|_{y_{k+1}}^2\right].$$

Proof. As the game is μ-strongly monotone, we have

$$\begin{aligned} &-2\left\langle \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}), F(y_k) \right\rangle_{y_k} + 2\|F(y_k)\|_{y_k}^2 \\ =&\frac{2}{\eta_k}\left\langle \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) - F(y_k)), \operatorname{LOG}_{y_k}(y_{k+1}) \right\rangle_{y_k} \\ \geq&\frac{2\mu}{\eta_k}\|\operatorname{LOG}_{y_k}(y_{k+1})\|_{y_k}^2 = 2\eta_k\mu\|F(y_k)\|_{y_k}^2. \end{aligned}$$

By the smoothness of the loss functions in $\mathcal{L}$ we have

$$\begin{aligned} (\eta_k L)^2 \cdot \|F(y_k)\|_{y_k}^2 =&L^2 \cdot \|\operatorname{LOG}_{y_k}(y_{k+1})\|_{y_k}^2 \\ \geq& \|F(y_k) - \Gamma_{y_{k+1}}^{y_k} F(y_{k+1})\|_{y_k}^2 \\ =& \|F(y_k)\|_{y_k}^2 - 2\left\langle F(y_k), \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) \right\rangle_{y_k} + \|\Gamma_{y_{k+1}}^{y_k} F(y_{k+1})\|_{y_k}^2 \\ =& \|F(y_k)\|_{y_k}^2 - 2\left\langle F(y_k), \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) \right\rangle_{y_k} + \|F(y_{k+1})\|_{y_{k+1}}^2 \end{aligned}$$

Combining the two inequalities we have

$$\left(1 - 2\eta_k\mu + (\eta_k L)^2\right)\|F(y_k)\|_{y_k}^2 \geq \|F(y_{k+1})\|_{y_{k+1}}^2.$$

Since $\mu \leq L$, we have $2 - 2\eta_k\mu + (\eta_k L)^2 \geq 2 - 2\eta_k\mu + (\eta_k\mu)^2 = 1 + (\eta_k\mu - 1)^2 \geq 1$, which implies that

$$\left(1 - \frac{2\eta_k\mu - (\eta_k L)^2}{2 - 2\eta_k\mu + (\eta_k L)^2}\right)\|F(y_k)\|_{y_k}^2 \geq \left(1 - 2\eta_k\mu + (\eta_k L)^2\right)\|F(y_k)\|_{y_k}^2.$$

■

Next, we show that we can prove the existence of a Nash equilibrium by combining Lemma 5 and Lemma 3. In other words, the existence of a Nash equilibrium is guaranteed by Assumption 1.

Lemma 6. *A Riemannian game that satisfies Assumption 1 has a Nash equilibrium.*

[3] A way to bound this term is to introduce an additional point $\widetilde{y}_{k+1} = \operatorname{EXP}_{y_k}(-\eta_k F(y_k))$. Since $\mathbb{E}\left[\langle g^{(k)} - F(y_k), \Gamma_{y_{k+1}}^{y_k} F(\widetilde{y}_{k+1})\rangle_{y_k}\right] = 0$, we can use Cauchy-Schwartz and smoothness of the loss functions to bound $\langle g^{(k)} - F(y_k), \Gamma_{y_{k+1}}^{y_k} F(y_{k+1})\rangle_{y_k} = \langle g^{(k)} - F(y_k), \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) - \Gamma_{y_{k+1}}^{y_k} F(\widetilde{y}_{k+1})\rangle_{y_k}$. However this approach depends on the term $d_{\mathcal{M}}(y_{k+1}, \widetilde{y}_{k+1})$ which suffers from distance distortion.

We summarize the convergence rates in Theorem 1.

Theorem 1. *Under Assumption 1, the RGD (Algorithm 1) with constant step size* $\left\{\eta_k = \frac{\mu}{L^2}\right\}_{k\in[1,K]}$ *and batch-size schedule* $\left\{m_k = \frac{16\sigma^2}{\kappa^4(LB)^2}e^{\frac{\kappa^2}{4}(k-1)}\right\}_{k\in[K]}$ *has the following convergence rate with respect to any target* $\epsilon \geq 0$.

- *In the deterministic case (i.e.,* $\sigma^2 = 0$*) for* $K^* = \left\lceil \frac{4\cdot log\left(\frac{L\cdot d_{\mathcal{M}}(y_0,y^*)}{\epsilon}\right)}{\kappa^2} \right\rceil$,

$$\|F(y_{K^*})\|_{y_{K^*}} \leq \epsilon, \quad gap_D(y_{k^*}) \leq D\cdot\epsilon.$$

- *In the stochastic setting, for* $K^* = \left\lceil \frac{8\cdot log\left(\frac{L\cdot B}{\epsilon}\right)}{\kappa^2} \right\rceil$,

$$\mathbb{E}\left[\|F(y_{K^*})\|_{y_{K^*}}\right] \leq \epsilon, \quad \mathbb{E}\left[gap_D(y_{K^*})\right] \leq D\cdot\epsilon.$$

Moreover, the total number of times a stochastic gradient is queried is at most $\frac{109\sigma^2}{\kappa^6\epsilon^2}$.

Comparison with the Riemannian Corrected Extragradient method [130]. Geodesically strongly convex-concave min-max optimization is a special instance of strongly monotone Riemannian games (see Lemma 2). In this setting, [57] studied the Riemannian Corrected Extragradient method (RCEG) [130] and used the distance to equilibrium, $d_{\mathcal{M}}(y_{K^*}, y^*)$, as the proximity measure. By the smoothness of the loss functions, $\|F(y_{K^*})\|_{y_{K^*}} \leq L d_{\mathcal{M}}(y_{K^*}, y^*)$, and thus we can directly compare our bounds using the norm of the gradient. While our analysis and choice of step size is independent of the curvature of the manifold, for their analysis [57], they require a step size that depends on the curvature of the manifold.

To output a strategy with gradient norm at most ϵ in the deterministic case, both algorithms require the number of iterations that scales logarithmically in $d_{\mathcal{M}}(y_0, y^*), L, \epsilon$. The main difference is that for the analysis of RCEG, the number of iterations scales with the product of $\frac{1}{\kappa}$ and a metric that quantifies the distance distortion caused by the manifold, while our bound depends on $\frac{1}{\kappa^2}$ but is independent of the curvature of the manifold. We observe a similar tradeoff in the stochastic case.

5 Conclusions

We have introduced the Riemannian game framework, a generalization of min-max optimization over Riemannian manifolds. We show that the Riemannian Gradient Descent, arguably the simplest first-order method in manifold optimization, can achieve curvature-independent and linear convergence rate with a step size that is agnostic to the manifold's curvature. We have also extended the result to the stochastic setting.

Our work raises several open questions. The first is whether we can obtain best-of-both-worlds convergence rate in geodesically strongly monotone Riemannian games, achieving both curvature independence and linear dependence on the condition number. Another intriguing challenge is to obtain curvature-independent and last-iterate convergence rate for monotone Riemannian games. Finally, can we design first-order methods that converge in structural but geodesically nonmonotone Riemannian games?

Acknowledgement: Yang Cai and Argyris Oikonomou were supported by a Sloan Foundation Research Fellowship and the NSF Award CCF-1942583 (CAREER). Emmanouil V. Vlatakis-Gkaragkounis is grateful for financial support by the Post-Doctoral FODSI- Simons Fellowship, Pancretan Association of America and Simons Collaboration on Algorithms and Geometry and Onassis Doctoral Fellowship. MJ and TL were supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764 and by the Vannevar Bush Faculty Fellowship pro- gram under grant number N00014-21-1-2941.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[2] F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *AISTATS*, pages 1297–1307. PMLR, 2020.

[3] A. Anandkumar and R. Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *COLT*, pages 81–102. PMLR, 2016.

[4] F. Bach. The "η-trick" or the effectiveness of reweighted least-squares, 2019.

[5] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

[6] G. Becigneul and O.-E. Ganea. Riemannian adaptive optimization methods. In *ICLR*, 2019.

[7] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[8] G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

[9] D. Berenson, S. S. Srinivasa, D. Ferguson, and J. J. Kuffner. Manipulation planning on constraint manifolds. In *2009 IEEE International Conference on Robotics and Automation*, pages 625–632. IEEE, 2009.

[10] S. Bhattacharya, N. Michael, and V. Kumar. Distributed coverage and exploration in unknown non-convex environments. In *Distributed Autonomous Robotic Systems: The 10th International Symposium*, pages 61–75. Springer, 2013.

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

[12] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[13] N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *NIPS*, pages 406–414, 2011.

[14] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

[15] L. E. J. Brouwer. Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115, 1911.

[16] D. Burago, I. D. Burago, Y. Burago, S. Ivanov, S. V. Ivanov, and S. A. Ivanov. *A Course in Metric Geometry*, volume 33. American Mathematical Soc., 2001.

[17] S. Burer and A. N. Letchford. Non-convex mixed-integer nonlinear programming: A survey. *Surveys in Operations Research and Management Science*, 17(2):97–106, 2012.

[18] L. Buşoniu, R. Babuška, and B. De Schutter. Multi-agent reinforcement learning: An overview. *Innovations in Multi-Agent Systems and Applications-1*, pages 183–221, 2010.

[19] Y. Cai, A. Oikonomou, and W. Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. June 2022. arXiv:2206.05248 [cs, math].

[20] Y. Cai, A. Oikonomou, and W. Zheng. Finite-time last-iterate convergence for learning in multi-player games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33904–33919. Curran Associates, Inc., 2022.

[21] Y. Cai and W. Zheng. Accelerated Single-Call Methods for Constrained Min-Max Optimization. Feb. 2023.

[22] Y. Cai and W. Zheng. Doubly optimal no-regret learning in monotone games, Jan. 2023. arXiv:2301.13120 [cs].

[23] X. Chen and X. Deng. Settling the complexity of two-player Nash equilibrium. In *FOCS '06: Proceedings of the 2006 IEEE Symposium on Foundations of Computer Scince*, 2006.

[24] S. Chewi, T. Maunu, P. Rigollet, and A. J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

[25] C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In *NeurIPS*, pages 5987–5997, 2019.

[26] C. Criscitiello and N. Boumal. An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, pages 1–77, 2022.

[27] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a Nash equilibrium. In *STOC '06: Proceedings of the 38th annual ACM symposium on the Theory of Computing*, 2006.

[28] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR*, 2018.

[29] C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, 2019.

[30] C. Daskalakis, S. Skoulakis, and M. Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, pages 1466–1478, New York, NY, USA, June 2021. Association for Computing Machinery.

[31] J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *COLT*, pages 1428–1451. PMLR, 2020.

[32] J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In A. Banerjee and K. Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2746–2754. PMLR, 2021.

[33] S. Diamond, R. Takapoui, and S. Boyd. A general system for heuristic minimization of convex functions over non-convex sets. *Optimization Methods and Software*, 33(1):165–193, 2018.

[34] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

[35] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, B. Póczos, and A. Singh. Gradient descent can take exponential time to escape saddle points. In *NIPS*, pages 1067–1077, 2017.

[36] J. Fearnley, P. W. Goldberg, A. Hollender, and R. Savani. The complexity of gradient descent: CLS = PPAD∩ PLS. In *STOC*, pages 46–59, 2021.

[37] O. P. Ferreira and P. R. Oliveira. Subgradient algorithm on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 97(1):93–104, 1998.

[38] O. P. Ferreira, L. R. Pérez, and S. Z. Németh. Singularities of monotone vector fields and an extragradient-type algorithm. *Journal of Global Optimization*, 31(1):133–151, 2005.

[39] P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007.

[40] G. G., H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

[41] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842. PMLR, 2015.

[42] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *ICML*, pages 1233–1242. PMLR, 2017.

[43] N. Golowich, S. Pattathil, and C. Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS*, pages 20766–20778, 2020.

[44] N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT*, pages 1758–1784. PMLR, 2020.

[45] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

[46] A. Han, B. Mishra, P. Jawanpuria, P. Kumar, and J. Gao. Riemannian Hamiltonian methods for min-max optimization on manifolds. *ArXiv Preprint: 2204.11418*, 2022.

[47] S. Hong, J.-H. Kim, and H.-W. Park. Real-time constrained nonlinear model predictive control on so (3) for dynamic legged locomotion. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3982–3989. IEEE, 2020.

[48] Y.-G. Hsieh, K. Antonakopoulos, V. Cevher, and P. Mertikopoulos. No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation. In *NeurIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

[49] F. Huang and S. Gao. On riemannian gradient-based methods for minimax problems. *arXiv preprint arXiv:2010.06097*, 2020.

[50] L. Huang, X. Liu, B. Lang, A. Yu, Y. Wang, and B. Li. Orthogonal weight normalization: solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI*, pages 3271–3278, 2018.

[51] P. Jain, P. Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

[52] P. Jawanpuria and B. Mishra. A unified framework for structured low-rank matrix learning. In *ICML*, pages 2254–2263. PMLR, 2018.

[53] A. Jiménez-Cordero, J. M. Morales, and S. Pineda. A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. *European Journal of Operational Research*, 293(1):24–35, 2021.

[54] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *ICML*, pages 1724–1732. PMLR, 2017.

[55] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

[56] J. Jin and S. Sra. Understanding riemannian acceleration via a proximal extragradient framework. In *Conference on Learning Theory*, pages 2924–2962. PMLR, 2022.

[57] M. I. Jordan, T. Lin, and E.-V. Vlatakis-Gkaragkounis. First-order algorithms for min-max optimization in geodesic metric spaces. In *NeurIPS*, pages 6557–6574, 2022.

[58] S. Kakutani. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal*, 8(3):457–459, 1941.

[59] H. Kasai, P. Jawanpuria, and B. Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *ICML*, pages 3262–3271. PMLR, 2019.

[60] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.

[61] D. Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1-2):57–87, 2021.

[62] J. Kim and I. Yang. Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. In *ICML*, pages 11255–11282. PMLR, 2022.

[63] Y.-H. Kim and B. Pass. Wasserstein barycenters over riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017.

[64] D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

[65] H. Komiya. Elementary proof for Sion's minimax theorem. *Kodai Mathematical Journal*, 11(1):5–7, 1988.

[66] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[67] A. Kristály. Nash-type equilibria on Riemannian manifolds: A variational approach. *Journal de Mathématiques Pures et Appliquées*, 101(5):660–688, 2014.

[68] A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. *Advances in neural information processing systems*, 30, 2017.

[69] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458, 2012.

[70] J. Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2012.

[71] J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. D. Yoo. Fast and efficient MMD-based fair PCA via optimization over Stiefel manifold. In *AAAI*, pages 7363–7371, 2022.

[72] S. Lee and D. Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *NeurIPS*, pages 22588–22600, 2021.

[73] Q. Lei, S. G. Nagarajan, I. Panageas, and X. Wang. Last iterate convergence in no-regret learning: Constrained min-max optimization for convex-concave landscapes. In *AISTATS*, pages 1441–1449. PMLR, 2021.

[74] C. Lenglet. Geometric and variational methods for diffusion tensor mri processing. (méthodes géométriques et variationnelles pour le traitement d'irm du tenseur de diffusion). 2006.

[75] C. Li, G. López, and V. Martín-Márquez. Monotone vector fields and the proximal point algorithm on Hadamard manifolds. *Journal of the London Mathematical Society*, 79(3):663–683, 2009.

[76] S.-L. Li, C. Li, Y.-C. Liou, and J.-C. Yao. Existence of solutions for variational inequalities on riemannian manifolds. *Nonlinear Analysis: Theory, Methods & Applications*, 71(11):5695–5706, 2009.

[77] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915. PMLR, 2019.

[78] T. Lin, N. Ho, M. Cuturi, and M. I. Jordan. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research*, 23(65):1–43, 2022.

[79] T. Lin, Z. Zheng, E. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pages 262–270. PMLR, 2021.

[80] S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aux Dérivées Partielles*, 117:87–89, 1963.

[81] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[82] J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *ICML*, pages 7555–7564. PMLR, 2021.

[83] P. Mertikopoulos, Y.-P. Hsieh, and V. Cevher. A unified stochastic approximation framework for learning in games. https://arxiv.org/abs/2206.03922, 2022.

[84] P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

[85] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, pages 1497–1507. PMLR, 2020.

[86] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.

[87] D. Monderer and L. S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996.

[88] A. Nagurney, K. K. Dhanda, and J. K. Stranlund. General multi-product, multi-pollutant market pollution permit model: a variational inequality approach. *Energy Economics*, 19(1):57–76, 1997.

[89] J. F. Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, September 1951.

[90] S. Németh. Variational inequalities on hadamard manifolds. *Nonlinear Analysis: Theory, Methods & Applications*, 52(5):1491–1498, 2003.

[91] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[92] J. Oliger and A. Sundström. Theoretical and practical aspects of some initial boundary value problems in fluid dynamics. *SIAM Journal on Applied Mathematics*, 35(3):419–446, 1978.

[93] S. Park. Riemannian manifolds are KKM spaces. *Advances in the Theory of Nonlinear Analysis and its Application*, 3(2):64–73, 2019.

[94] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[95] A. Petcu and B. Faltings. Dpop: A scalable method for multiagent constraint optimization. In *IJCAI 05*, number CONF, pages 266–271, 2005.

[96] T. Pethick, O. Fercoq, P. Latafat, P. Patrinos, and V. Cevher. Solving stochastic weak minty variational inequalities without increasing batch size. *CoRR*, abs/2302.09029, 2023.

[97] G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[98] B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[99] S. Quinlan. Efficient distance computation between non-convex objects. In *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*, pages 3324–3329. IEEE, 1994.

[100] L. J. Ratliff, S. A. Burden, and S. S. Sastry. Characterization and computation of local nash equilibria in continuous games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 917–924. IEEE, 2013.

[101] S. Reddi, M. Zaheer, S. Sra, B. Poczos, F. Bach, R. Salakhutdinov, and A. Smola. A generic approach for escaping saddle points. In *AISTATS*, pages 1233–1242. PMLR, 2018.

[102] J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3):520–534, 1965.

[103] H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *Computer Journal*, 3(3):175–184, 1960.

[104] W. H. Sandholm. Potential games with continuous player sets. *Journal of Economic Theory*, 97(1):81–108, 2001.

[105] H. Sato, H. Kasai, and B. Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019.

[106] L. Scrimali. A variational inequality formulation of the environmental pollution control problem. *Optimization Letters*, 4:259–274, 2010.

[107] M. Spong, K. Khorasani, and P. Kokotovic. An integral manifold approach to the feedback control of flexible joint robots. *IEEE Journal on Robotics and Automation*, 3(4):291–300, 1987.

[108] S. Sra and R. Hosseini. Geometric optimization in machine learning. In *Algorithmic Advances in Riemannian Geometry and Applications*, pages 73–91. Springer, 2016.

[109] M. Staib, S. Claici, J. M. Solomon, and S. Jegelka. Parallel streaming wasserstein barycenters. *Advances in Neural Information Processing Systems*, 30, 2017.

[110] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

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

[112] Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on Riemannian manifolds. In *NeurIPS*, pages 7274–7284, 2019.

[113] C. J. Taylor and D. J. Kriegman. Minimization on the Lie group SO(3) and related manifolds. 1994.

[114] R. Toscano and P. Lyonnet. A new heuristic approach for non-convex optimization problems. *Information Sciences*, 180(10):1955–1966, 2010.

[115] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *COLT*, pages 650–687, 2018.

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

[117] E. V. Vlatakis-Gkaragkounis, L. Flokas, T. Lianeas, P. Mertikopoulos, and G. Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *NeurIPS*, pages 1380–1391, 2020.

[118] E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *NeurIPS*, pages 10450–10461, 2019.

[119] E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent. In *NeurIPS*, pages 2373–2386, 2021.

[120] J. H. Wang, G. López, V. Martín-Márquez, and C. Li. Monotone and accretive vector fields on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 146(3):691–708, 2010.

[121] M. Watterson, S. Liu, K. Sun, T. Smith, and V. Kumar. Trajectory optimization on manifolds with applications to quadrotor systems. *The International Journal of Robotics Research*, 39(2-3):303–320, 2020.

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

[123] M. Weber and S. Sra. Riemannian optimization via Frank-Wolfe methods. *Mathematical Programming*, To appear:1–32, 2022.

[124] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR*, 2021.

[125] A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.

[126] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with o(1/kˆ 2) rate on squared gradient norm. In *ICML*, pages 12098–12109. PMLR, 2021.

[127] G. Zhang and Y. Yu. Convergence of gradient methods on bilinear zero-sum games. In *ICLR*, 2020.

[128] H. Zhang, S. J. Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *NeurIPS*, pages 4592–4600, 2016.

[129] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *COLT*, pages 1617–1638. PMLR, 2016.

[130] P. Zhang, J. Zhang, and S. Sra. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *ArXiv Preprint: 2202.06950*, 2022.

A Limitations of our Work

As a theoretical contribution, our results are valid provided that the assumptions we have made, summarized in Assumption 1, are met. Please note that this paper offers no guarantees if any of these conditions are violated.

B Additional Related Work

We cannot survey all the increasing stream of works on the geometric properties of Riemannian manifolds but refer the reader to the books [16, 70] and the references therein.

The minimization on Riemannian manifolds. Riemannian optimization [1, 12] is an extensive and active area of research, for which one aspires to develop Riemannian gradient-based algorithms that share analogous properties to the more broadly studied Euclidean methods, including deterministic [8, 14, 26, 37, 62, 129], projection-free [122, 123], stochastic and adaptive [6, 11, 59], variance-reduced [105, 115, 128], and saddle-point-escaping [25, 112], among others. In this context, the curvature-independent and last-iterate convergence rate can be achieved by Riemannian gradient descent method [129].

The last-iterate convergence for games in Euclidean spaces. In the unconstrained setting, if we further assume that either the game is strongly monotone or the payoff matrix in a bilinear game has all singular values bounded away from zero, the linear convergence rate is known for gradient-based algorithms [28, 40, 77, 85, 127]. The results for the constrained setting are relatively scare. Indeed, most of convergence results for monotone games are asymptotic [29, 73] and we are aware of [124] which provided a linear convergence rate of optimistic gradient method for bilinear games when the constraint set is a polytope. However, the bound depends on a problem-dependent parameter that can be arbitrarily close to zero, making their results invalid for all smooth and monotone games. Recently, [20] has shown the first last-iterate convergence rates for all smooth and monotone games, matching the lower bounds [43, 44].

From a technical viewpoint, the choice of performance measures is crucial to proving last-iterate convergence. It is known that classical gradient-based algorithms have a time-average convergence rate of $O(1/t)$ in terms of the gap function for smooth monotone games [66, 86, 91]. Other than the gap function, one can measure the convergence using the norm of the operator if the setting is unconstrained [61, 72, 126], or the natural residual or similar notions if the setting is constrained and further satisfies the additional cocoercive condition [31]. More recently, a line of work [19, 21, 22] show how to obtain the optimal $O(1/t)$ last-iterate convergence rate in constrained settings without relying on the cocoercive assumption, and these results can be further extended to a structural non-monotone setting known as the comonotone setting.

C Additional Preliminaries

D Missing Details of Section 3

Proof of Lemma 1: The proof follows from the descent property of Riemannian gradient descent (A4.2 in [12]). Assume, for sake of contradiction, that $\|F(y^*)\|_{y^*} \neq 0$. Then there exists player i such that $\|\operatorname{grad}_{y_i} l_i(y^*)\|_{y_i^*} \neq 0$, and consider the following deviation for player i parameterized by $\eta > 0$,

$$y_i' = \operatorname{EXP}_{y_i^*}\left(-\eta \operatorname{grad}_{y_i} l_i(y^*)\right),$$

Assume that the loss function l_i is L-smooth, then for sufficiently small η we have the sufficient decrease property (Proposition 4.8 in [12]),

$$\begin{aligned} l_i(y_i', y_{-i}^*) \leq & l_i(y^*) + \langle \operatorname{grad}_{y_i} l_i(y^*), \operatorname{LOG}_{y_i^*}(y_i') \rangle_{y_i^*} + \frac{L}{2} d(y_i', y_i^*)^2 \\ = & l_i(y^*) - \eta \|\operatorname{grad}_{y_i} l_i(y^*)\|_{y_i^*}^2 + \frac{\eta^2 L \|\operatorname{grad}_{y_i} l_i(y^*)\|_{y_i^*}^2}{2} \\ < & l_i(y^*) \end{aligned}$$

where the last inequality follows for sufficiently small η, reaching a contradiction that strategy profile y^* is a Nash equilibrium. ■

Proof of Lemma 2: The proof follows from the standard argument that convex-concave games are monotone. Observe that the gradient is $F(y_1, y_2) = (\text{grad}_{y_1} u(y_1, y_2), -\text{grad}_{y_2} u(y_1, y_2))$. Now we show that F is geodesically monotone. Since function $u(y_1, y_2)$ is μ-strongly geodesically convex in y_1 and μ-strongly geodesically concave in y_2, for any $y_1, y_1' \in \mathcal{M}_1$ and $y_2, y_2' \in \mathcal{M}_2$, we have

$$\begin{aligned}
u(y_1', y_2) \geq& u(y_1, y_2) + \langle \text{grad}_{y_1} u(y_1, y_2), \text{LOG}_{y_1}(y_1') \rangle_{y_1} + \frac{\mu}{2} d_{\mathcal{M}_1}(y_1, y_1')^2 \\
u(y_1, y_2') \geq& u(y_1', y_2') + \langle \text{grad}_{y_1} u(y_1', y_2'), \text{LOG}_{y_1'}(y_1) \rangle_{y_1'} + \frac{\mu}{2} d_{\mathcal{M}_1}(y_1, y_1')^2 \\
-u(y_1, y_2') \geq& -u(y_1, y_2) - \langle \text{grad}_{y_2} u(y_1, y_2), \text{LOG}_{y_2}(y_2') \rangle_{y_2} + \frac{\mu}{2} d_{\mathcal{M}_2}(y_2, y_2')^2 \\
-u(y_1', y_2) \geq& -u(y_1', y_2') - \langle \text{grad}_{y_2} u(y_1', y_2'), \text{LOG}_{y_2'}(y_2) \rangle_{y_2'} + \frac{\mu}{2} d_{\mathcal{M}_2}(y_2, y_2')^2.
\end{aligned}$$

By adding all four inequalities, we conclude that the Riemannian game is μ-strongly monotone,

$$\begin{aligned}
&- \langle \text{grad}_{y_1} u(y_1, y_2), \text{LOG}_{y_1}(y_1') \rangle_{y_1} - \langle \text{grad}_{y_1} u(y_1', y_2'), \text{LOG}_{y_1'}(y_1) \rangle_{y_1'} \\
&\quad + \langle \text{grad}_{y_2} u(y_1, y_2), \text{LOG}_{y_2}(y_2') \rangle_{y_2} + \langle \text{grad}_{y_2} u(y_1', y_2'), \text{LOG}_{y_2'}(y_2) \rangle_{y_2'} \geq \mu \cdot d_{\mathcal{M}}((y_1, y_2), (y_1', y_2'))^2.
\end{aligned}$$

Using the parallel transport and that $\Gamma_{y_1'}^{y_1} \text{LOG}_{y_1'}(y_1) = -\text{LOG}_{y_1}(y_1')$,

$$\begin{aligned}
\langle \text{grad}_{y_1} u(y_1', y_2'), \text{LOG}_{y_1'}(y_1) \rangle_{y_1'} =& \langle \Gamma_{y_1'}^{y_1} \text{grad}_{y_1} u(y_1', y_2'), \Gamma_{y_1'}^{y_1} \text{LOG}_{y_1'}(y_1) \rangle_{y_1} \\
=& -\langle \Gamma_{y_1'}^{y_1} \text{grad}_{y_1} u(y_1', y_2'), \text{LOG}_{y_1}(y_1') \rangle_{y_1}.
\end{aligned}$$

Similarly, we can show that $\langle \text{grad}_{y_2} u(y_1', y_2'), \text{LOG}_{y_2'}(y_2) \rangle_{y_2'} = -\langle \Gamma_{y_2'}^{y_2} \text{grad}_{y_2} u(y_1', y_2'), \text{LOG}_{y_2}(y_2') \rangle_{y_2}$. Thus,

$$\begin{aligned}
&\langle \Gamma_{y_1'}^{y_1} \text{grad}_{y_1} u(y_1', y_2') - \text{grad}_{y_1} u(y_1, y_2), \text{LOG}_{y_1}(y_1') \rangle_{y_1} - \langle \Gamma_{y_2'}^{y_2} \text{grad}_{y_2} u(y_1', y_2') - \text{grad}_{y_2} u(y_1, y_2), \text{LOG}_{y_2}(y_2') \rangle_{y_2} \\
&\qquad \geq \mu \cdot d_{\mathcal{M}}((y_1, y_2), (y_1', y_2'))^2.
\end{aligned}$$

After further simplification, the above inequality exactly states that the Riemannian game is μ-strongly monotone.

$$\langle \Gamma_{(y_1', y_2')}^{(y_1, y_2)} F(y_1', y_2') - F(y_1, y_2), \text{LOG}_{(y_1, y_2)}((y_1', y_2')) \rangle_{(y_1, y_2)} \geq \mu \cdot d((y_1, y_2), (y_1', y_2'))^2.$$

■

Proof of Lemma 3: Let $y_i^* = \arg\min_{y_i' \in B_{\mathcal{M}_i}(y_i, D)} l(y_i', y_{-i})$. For each player $i \in \mathcal{N}$, consider the geodesic $y'(t) : [0, 1] \to \mathcal{M}$, that connects y to (y_i^*, y_{-i}). Since the game is monotone, we have that for every $t \geq t' \in [0, 1]$,

$$\begin{aligned}
&\langle \Gamma_{y'(t)}^{y'(t')} F(y'(t)) - F(y'(t')), \text{LOG}_{y'(t')}(y'(t)) \rangle_{y'(t')} \\
=& \langle \Gamma_{y'(t)}^{y} F(y'(t)) - \Gamma_{y'(t')}^{y} F(y'(t')), \Gamma_{y'(t')}^{y} \text{LOG}_{y'(t')}(y'(t)) \rangle_{y} \\
=& \langle \Gamma_{y_i'(t)}^{y_i} \text{grad}_{y_i} l_i(y'(t)) - \Gamma_{y_i'(t')}^{y_i} \text{grad}_{y_i} l_i(y'(t')), \Gamma_{y_i'(t')}^{y_i} \text{LOG}_{y_i'(t')}(y_i'(t)) \rangle_{y_i} \\
=& (t - t') \cdot \langle \Gamma_{y_i'(t)}^{y_i} \text{grad}_{y_i} l_i(y'(t)) - \Gamma_{y_i'(t)}^{y_i} \text{grad}_{y_i} l_i(y'(t')), \text{LOG}_{y_i}(y_i^*) \rangle_{y_i} \\
\geq& 0.
\end{aligned}$$

The first equality holds as $y, y'(t)$, and $y'(t')$ all lie in the same geodesic, which implies that $\Gamma_{y'(t)}^{y} F(y'(t)) = \Gamma_{y'(t')}^{y} \Gamma_{y'(t)}^{y'(t')} F(y'(t)) = \Gamma_{y'(t)}^{y} F(y'(t))$. The second equality follows from $\text{LOG}_{y'_{-i}(t')}(y'_{-i}(t)) = 0$. The third equality is due to the fact that y_i, $y_i'(t)$ and $y_i(t')$ lie on the same geodesic, which implies that $\Gamma_{y_i'(t')}^{y_i} \text{LOG}_{y_i'(t')}(y_i'(t)) = (t - t') \cdot \text{LOG}_{y_i}(y_i^*)$.

Since $t \geq t'$, we further simplify the inequality above as follows

$$\langle \Gamma_{y_i'(t)}^{y_i} \text{grad}_{y_i} l_i(y'(t)), \text{LOG}_{y_i}(y_i^*) \rangle_{y_i} \geq \langle \Gamma_{y_i'(t)}^{y_i} \text{grad}_{y_i} l_i(y'(t')), \text{LOG}_{y_i}(y_i^*) \rangle_{y_i}. \tag{D.1}$$

Moreover, since on the geodesic $y_i'(\cdot)$, at point $y_i'(t)$, the direction we are moving is $\Gamma_{y_i}^{y_i'(t)} \text{LOG}_{y_i}(y_i^*)$, by the chain rule,

$$\frac{dl_i(y_i(t), y_{-i})}{dt} = \langle \text{grad}_{y_i} l(y'(t)), \Gamma_{y_i}^{y_i'(t)} \text{LOG}_{y_i}(y_i^*) \rangle_{y_i'(t)} = \langle \Gamma_{y_i'(t)}^{y_i} \text{grad}_{y_i} l(y'(t)), \text{LOG}_{y_i}(y_i^*) \rangle_{y_i}.$$

Hence, Equation (D.1) implies that the function $\widetilde{l}_i(t) = l_i(y'(t))$ is convex in $t \in [0,1]$, which further implies that

$$l_i(y_i^*, y_{-i}) = l_i(y_i'(1), y_{-i}) \geq l_i(y) + \langle \text{grad}_{y_i} l_i(y), \text{LOG}_{y_i}(y_i^*) \rangle_{y_i}.$$

We conclude the proof by the following chain of inequalities,

$$\begin{aligned}
Tgap_D(y) =& \sum_{i \in \mathcal{N}} (l_i(y) - l_i(y_i^*, y_{-i})) \\
\leq& \sum_{i \in \mathcal{N}} \langle \text{grad}_{y_i} l_i(y), -\text{LOG}_{y_i}(y_i^*) \rangle_{y_i} \\
=& \langle F(y), -\text{LOG}_y(y^*) \rangle_y \\
\leq& \max_{y' \in B_{\mathcal{M}}(y, \sqrt{N} \cdot D)} \langle F(y), -\text{LOG}_y(y') \rangle_y \\
=& gap_{\sqrt{N} \cdot D}(y).
\end{aligned}$$

■

Proof of Corollary 1: By Lemma 1, the condition that $\|F(y^*)\|_{y^*} = 0$ is necessary for a strategy y^* to be a Nash equilibrium. Now we show that for monotone Riemmanian games, this is also a sufficient condition. If $\|F(y^*)\|_{y^*} = 0$, then for any $D > 0$ by $gap_D(y^*) = 0$. Moreover by Lemma 3, this implies that $Tgap_D(y^*) = 0$, which implies that y^* is a Nash equilibrium. ■

Proof of Lemma 4: The proof follows by Cauchy-Schwartz inequality,

$$\begin{aligned}
gap_D(z) =& \max_{y' \in B_{\mathcal{M}}(y, D)} \langle F(y), -\text{LOG}_y(y') \rangle_y \\
\leq& D \cdot \|F(y)\|_y
\end{aligned}$$

■

E Missing Details of Section 4

Proof of Lemma 5: We use the following four inequalities in our proof,

$$\begin{aligned}
\|\Gamma_{y_{k+1}}^{y_k} F(y_{k+1})\|_{y_k}^2 =& \langle \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}), \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) \rangle_{y_k} \\
=& \langle \Gamma_{y_k}^{y_{k+1}} \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}), \Gamma_{y_k}^{y_{k+1}} \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) \rangle_{y_{k+1}} \\
=& \langle F(y_{k+1}), F(y_{k+1}) \rangle_{y_{k+1}} \\
=& \|F(y_{k+1})\|_{y_{k+1}}^2, \qquad \text{(E.1)} \\
\mathbb{E}\left[\left\langle F(y_k), g^{(k)} \right\rangle_{y_k} \right] =& \|F(y_k)\|_{y_k}^2, \qquad \text{(E.2)} \\
\mathbb{E}\left[\|g^{(k)}\|_{y_k}^2 \right] =& \mathbb{E}\Big[\|g^{(k)} - F(y_k)\|_{y_k}^2 + \|F(y_k)\|_{y_k}^2 \\
& \qquad + 2 \left\langle g^{(k)} - F(y_k), F(y_k) \right\rangle_{y_k} \Big] \\
=& \mathbb{E}\left[\|g^{(k)} - F(y_k)\|_{y_k}^2 \right] + \|F(y_k)\|_{y_k}^2, \qquad \text{(E.3)} \\
-2 \cdot \left\langle g^{(k)} - F(y_k), \Gamma_{y_{k+1}}^{y_k} F(y_{k+1}) \right\rangle_{y_k} \leq& \frac{2}{2\eta_k \mu - (\eta_k L)^2} \|g^{(k)} - F(y_k)\|_{y_k}^2 \\
& + \frac{2\eta_k \mu - (\eta_k L)^2}{2} \|\Gamma_{y_{k+1}}^{y_k} F(y_{k+1})\|_{y_k}^2
\end{aligned}$$

$$= \frac{2}{2\eta_k\mu - (\eta_k L)^2}\|g^{(k)} - F(y_k)\|^2_{y_k} + \frac{2\eta_k\mu - (\eta_k L)^2}{2}\|F(y_{k+1})\|^2_{y_{k+1}}. \quad \text{(E.4)}$$

We use properties of the parallel transport for the first inequality. To prove the last inequality, we used Equation (E.1) and that $\eta_k \leq \frac{2\mu}{L^2} \Rightarrow 2\eta_k\mu - (\eta_k L)^2 \geq 0$. By monotonicity, and the use of Equation (E.2), Equation (E.3) we get,

$$\begin{aligned}
&2\mathbb{E}\left[-\left\langle \Gamma^{y_k}_{y_{k+1}} F(y_{k+1}), g^{(k)}\right\rangle_{y_k}\right] + 2\|F(y_k)\|^2_{y_k} \\
=&2\mathbb{E}\left[-\left\langle \Gamma^{y_k}_{y_{k+1}} F(y_{k+1}), g^{(k)}\right\rangle_{y_k}\right] + 2\mathbb{E}\left[\left\langle F(y_k), g^{(k)}\right\rangle_{y_k}\right] \\
=&\frac{2}{\eta_k}\mathbb{E}\left[\left\langle \Gamma^{y_k}_{y_{k+1}} F(y_{k+1}) - F(y_k), -\eta_k g^{(k)}\right\rangle_{y_k}\right] \\
=&\frac{2}{\eta_k}\mathbb{E}\left[\left\langle \Gamma^{y_k}_{y_{k+1}} F(y_{k+1}) - F(y_k), \mathrm{LOG}_{y_k}(y_{k+1})\right\rangle_{y_k}\right] \\
\geq&\frac{2}{\eta_k}\mathbb{E}\left[\mu\|\mathrm{LOG}_{y_k}(y_{k+1})\|^2_{y_k}\right] \\
=&2\eta_k\mu\mathbb{E}\left[\|g^{(k)}\|^2_{y_k}\right] \\
=&2\eta_k\mu\|F(y_k)\|^2_{y_k} + 2\eta_k\mu\mathbb{E}\left[\|g^{(k)} - F(y_k)\|^2_{y_k}\right]
\end{aligned}$$

Thus,

$$(2 - 2\eta_k\mu)\|F(y_k)\|^2_{y_k} + 2\mathbb{E}\left[-\left\langle \Gamma^{y_k}_{y_{k+1}} F(y_{k+1}), g^{(k)}\right\rangle_{y_k}\right] - 2\eta_k\mu\mathbb{E}\left[\|g^{(k)} - F(y_k)\|^2_{y_k}\right] \geq 0 \quad \text{(E.5)}$$

By smoothness of the loss functions $\mathcal{L}$, Equation (E.1), Equation (E.2), we have

$$\begin{aligned}
&(\eta_k L)^2\|F(y_k)\|^2_{y_k} + (\eta_k L)^2\mathbb{E}\left[\|g^{(k)} - F(y_k)\|^2_{y_k}\right] \\
=&(\eta_k L)^2\mathbb{E}\left[\|g^{(k)}\|^2_{y_k}\right] \\
=&L^2\mathbb{E}\left[\|\eta_k g^{(k)}\|^2_{y_k}\right] \\
=&\mathbb{E}\left[L^2\|\mathrm{LOG}_{y_k}(y_{k+1})\|^2_{y_k}\right] \\
\geq&\mathbb{E}\left[\|F(y_k) - \Gamma^{y_k}_{y_{k+1}} F(y_{k+1})\|^2_{y_k}\right] \\
=&\|F(y_k)\|^2_{y_k} - 2\mathbb{E}\left[\left\langle F(y_k), \Gamma^{y_k}_{y_{k+1}} F(y_{k+1})\right\rangle_{y_k}\right] + \mathbb{E}\left[\|\Gamma^{y_k}_{y_{k+1}} F(y_{k+1})\|^2_{y_k}\right] \\
=&\|F(y_k)\|^2_{y_k} - \mathbb{E}\left[2\left\langle F(y_k), \Gamma^{y_k}_{y_{k+1}} F(y_{k+1})\right\rangle^2_{y_k}\right] + \mathbb{E}\left[\|F(y_{k+1})\|^2_{y_{k+1}}\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
&(-1 + (\eta_k L)^2)\|F(y_k)\|^2_{y_k} + (\eta_k L)^2\mathbb{E}\left[\|g^{(k)} - F(y_k)\|^2_{y_k}\right] + \mathbb{E}\left[2\left\langle F(y_k), \Gamma^{y_k}_{y_{k+1}} F(y_{k+1})\right\rangle_{y_k}\right] \\
\geq&\mathbb{E}\left[\|F(y_{k+1})\|^2_{y_{k+1}}\right]. \quad \text{(E.6)}
\end{aligned}$$

By summing Equation (E.5) and Equation (E.6) and grouping together the terms, then using Equation (E.4), then using the fact that $\eta_k \leq \frac{2\mu}{L^2} \Leftrightarrow -2\mu\eta_k + (\eta_k L)^2 \leq 0$, and finally by the fact that $2 - 2\eta_k\mu + (\eta_k L)^2 \geq 2 - 2\eta_k\mu + (\eta_k\mu)^2 = 1 + (\eta_k\mu - 1)^2 \geq 1$ we get,

$$\left(1 - 2\eta_k\mu + (\eta_k L)^2\right) \cdot \|F(y_k)\|^2_{y_k} + \left(-2\mu\eta_k + (\eta_k L)^2\right) \cdot \mathbb{E}\left[\|g^{(k)} - F(y_k)\|^2_{y_k}\right]$$

$$
\begin{aligned}
&- 2 \cdot \mathbb{E}\left[\left\langle g^{(k)} - F(y_k), \Gamma_{y_{k+1}}^{y_k} F(y_{k+1})\right\rangle_{y_k}\right] \geq \mathbb{E}\left[\|F(y_{k+1})\|_{y_{k+1}}^2\right] \\
\Rightarrow &\left(1 - 2\eta_k\mu + (\eta_k L)^2\right) \cdot \|F(y_k)\|_{y_k}^2 + \left(\frac{2}{2\eta_k\mu - (\eta_k L)^2} - 2\mu\eta_k + (\eta_k L)^2\right) \cdot \mathbb{E}\left[\|g^{(k)} - F(y_k)\|_{y_k}^2\right] \\
&\geq \left(1 - \frac{2\eta_k\mu - (\eta_k L)^2}{2}\right) \mathbb{E}\left[\|F(y_{k+1})\|_{y_{k+1}}^2\right] \\
\Rightarrow &\left(1 - 2\eta_k\mu + (\eta_k L)^2\right) \cdot \|F(y_k)\|_{y_k}^2 + \frac{2}{2\eta_k\mu - (\eta_k L)^2} \cdot \mathbb{E}\left[\|g^{(k)} - F(y_k)\|_{y_k}^2\right] \\
&\geq \left(1 - \frac{2\eta_k\mu - (\eta_k L)^2}{2}\right) \mathbb{E}\left[\|F(y_{k+1})\|_{y_{k+1}}^2\right] \\
\Rightarrow &\left(1 - \frac{2\eta_k\mu - (\eta_k L)^2}{2 - 2\eta_k\mu + (\eta L)^2}\right) \cdot \|F(y_k)\|_{y_k}^2 + \frac{4}{(2 - 2\eta_k\mu + (\eta_k L)^2) \cdot (2\eta_k\mu - (\eta_k L)^2)} \cdot \mathbb{E}\left[\|g^{(k)} - F(y_k)\|_{y_k}^2\right] \\
&\geq \mathbb{E}\left[\|F(y_{k+1})\|_{y_{k+1}}^2\right] \\
\Rightarrow &\left(1 - \frac{2\eta_k\mu - (\eta_k L)^2}{2 - 2\eta_k\mu + (\eta L)^2}\right) \cdot \|F(y_k)\|_{y_k}^2 + \frac{4}{2\eta_k\mu - (\eta_k L)^2} \cdot \mathbb{E}\left[\|g^{(k)} - F(y_k)\|_{y_k}^2\right] \\
&\geq \mathbb{E}\left[\|F(y_{k+1})\|_{y_{k+1}}^2\right].
\end{aligned}
$$

■

Proof of Lemma 6: Consider an exectuion of Algorithm 1 with exact access to the gradient (e.g., $\sigma^2 = 0$ and $g^{(k)} = F(y_k)$), and constant step-size $\eta_k = \frac{\mu}{L^2}$ initialized at an arbitrary strategy profile $y_0 \in \mathcal{M}$. Lemma 5 implies that for $k \geq 1$,

$$
\begin{aligned}
\|F(y_k)\|_{y_k}^2 \leq &\left(1 - \frac{\kappa^2}{2 - \kappa^2}\right)^k \|F(y_0)\|_{y_0}^2 \\
\leq &e^{-\frac{\kappa^2 \cdot k}{2}} \|F(y_0)\|_{y_0}^2.
\end{aligned}
$$

Observe that $d_{\mathcal{M}}(y_{k+1}, y_k) = \eta \|F(y_k)\|_{y_k}$, which further implies that

$$
\sum_{k \geq 1} d_{\mathcal{M}}(y_{k+1}, y_k) \leq \left(\sum_{k \geq 1} e^{-\frac{\kappa^2 k}{2}}\right) \eta \|F(y_0)\|_{y_0}^2.
$$

Note that the RHS of the inequaltity above converges to an absolute constant C. Since the metric $(d_{\mathcal{M}}, \mathcal{M})$ is a complete metric space, the iterates of the algorithm $\{y_k\}_{k \geq 1}$ form a Cauchy sequance and converge to a strategy profile $y^* = \lim_{k \to +\infty} y_k$ in a closed area of radius C (e.g., $y^* \in B(y_0, C)$). By smoothness of loss functions in $\mathcal{L}$ and the fact that the sequence $\{y_k\}_{k \geq 1}$ converges,

$$
0 = \lim_{k \to +\infty} d_{\mathcal{M}}(y_{k+1}, y_k) = \lim_{k \to +\infty} \eta^2 \|F(y_k)\|_{y_k} = \eta^2 \|F(y^*)\|_{y^*},
$$

which implies that $\|F(y^*)\|_{y^*} = 0$. By Corollary 1, y^* is a Nash equilibrium. ■

Proof of Theorem 1: The following is true for both the deterministic and stochastic setting. By Corollary 1 $F(y^*) = 0$, and by smoothness of the loss functions in $\mathcal{L}$ we have that,

$$
\|F(y_0)\|_{y_0}^2 \leq L^2 \cdot d_{\mathcal{M}}(y_0, y^*)^2.
$$

First, we prove the deterministic case. By the descent inequality in Lemma 5 and the chosen step-size $\eta_k = \frac{\mu}{L^2}$, we know that for $k \geq 1$,

$$
\|F(y_K)\|_{y_K}^2 \leq \left(1 - \frac{\left(\frac{\mu}{L}\right)^2}{2 - \left(\frac{\mu}{L}\right)^2}\right)^K \|F(y_0)\|_{y_0}^2
$$

$$\leq \left(1 - \frac{\kappa^2}{2}\right)^K \cdot L^2 d_{\mathcal{M}}(y_0, y^*)^2$$
$$\leq e^{-\frac{\kappa^2 \cdot K}{2}} (L \cdot d_{\mathcal{M}}(y_0, y^*))^2.$$

The deterministic case then follows from Lemma 3, Lemma 4 and our choice of $K^* = \left\lceil \frac{4\log\left(\frac{L \cdot d_{\mathcal{M}}(y_0, y^*)}{\epsilon}\right)}{\kappa^2} \right\rceil$,

$$\|F(y_{K^*})\|^2_{y_{K^*}} \leq e^{-\frac{\kappa^2 \cdot K^*}{2}} (L \cdot d_{\mathcal{M}}(y_0, y^*))^2 \leq \epsilon^2.$$

Now we turn our attention to the stochastic setting and prove the following inequality by induction.

$$\mathbb{E}\left[\|F(y_k)\|^2_{y_k}\right] \leq e^{-\frac{\kappa^2}{4}k} \cdot (LB)^2.$$

The base case is clear. By the inductive hypothesis, Lemma 5, and the choice of $m_{k-1} = \frac{16\sigma^2}{\kappa^4 (LB)^2} e^{\frac{\kappa^2}{4}(k-1)}$ and $\eta_{k-1} = \frac{\mu}{L^2}$ we have,

$$\begin{aligned}
\mathbb{E}\left[\|F(y_k)\|^2_{y_k}\right] &\leq \left(1 - \frac{\kappa^2}{2 - \kappa^2}\right) \mathbb{E}\left[\|F(y_{k-1})\|^2_{y_{k-1}}\right] + \frac{4 \cdot \sigma^2}{\kappa^2 m_{k-1}} \\
&\leq \left(1 - \frac{\kappa^2}{2}\right) e^{-\frac{\kappa^2}{4}(k-1)} (L \cdot B)^2 + \frac{\kappa^2}{4} e^{-\frac{\kappa^2}{4}(k-1)} (L \cdot B)^2 \\
&= \left(1 - \frac{\kappa^2}{4}\right) e^{-\frac{\kappa^2}{4}(k-1)} (L \cdot B)^2 \\
&\leq e^{-\frac{\kappa^2}{4}k} (L \cdot B)^2.
\end{aligned}$$

The proof in the stochastic setting follows by Lemma 3, Lemma 4, Jensen's inequality, and noting that for $K^* = \left\lceil \frac{8\log\left(\frac{L \cdot B}{\epsilon}\right)}{\kappa^2} \right\rceil$,

$$\mathbb{E}\left[\|F(y_{K^*})\|_{y_{K^*}}\right] \leq \sqrt{\mathbb{E}\left[\|F(y_{K^*})\|^2_{y_{K^*}}\right]} \leq \sqrt{e^{-\frac{\kappa^2}{4}K^*} (L \cdot B)^2} \leq \epsilon.$$

Moreover, the total number of stochastic queries is at most

$$\begin{aligned}
\sum_{k=0}^{K^*} m_k \leq \frac{16\sigma^2}{\kappa^4 (LB)^2} \sum_{k=0}^{\frac{8\log\left(\frac{L \cdot B}{\epsilon}\right)}{\kappa^2}+1} e^{\frac{\kappa^2 \cdot k}{4}} &= \frac{16\sigma^2}{\kappa^4 (LB)^2} \frac{e^{\frac{\kappa^2}{2}} \cdot \left(\frac{L \cdot B}{\epsilon}\right)^2 - 1}{e^{\frac{\kappa^2}{4}} - 1} \\
&\leq \frac{109\sigma^2}{\kappa^6 \epsilon^2},
\end{aligned}$$

where in the last inequality we used that $e^x - 1 \geq x$, and that $\kappa \leq 1 \Rightarrow e^{\frac{\kappa^2}{2}} \leq 1.7$. ■