THE DYNAMICS OF RIEMANNIAN ROBBINS-MONRO ALGORITHMS

MOHAMMAD REZA KARIMI[⋆,*], YA-PING HSIEH[⋆,*],
PANAYOTIS MERTIKOPOULOS[§], AND ANDREAS KRAUSE[*]

Abstract. Many important learning algorithms, such as stochastic gradient methods, are often deployed to solve nonlinear problems on Riemannian manifolds. Motivated by these applications, we propose a family of Riemannian algorithms generalizing and extending the seminal stochastic approximation framework of Robbins and Monro [60]. Compared to their Euclidean counterparts, Riemannian iterative algorithms are much less understood due to the lack of a global linear structure on the manifold. We overcome this difficulty by introducing an *extended Fermi coordinate* frame which allows us to map the asymptotic behavior of the proposed Riemannian Robbins–Monro (RRM) class of algorithms to that of an associated *deterministic* dynamical system under very mild assumptions on the underlying manifold. In so doing, we provide a general template of almost sure convergence results that mirrors and extends the existing theory for Euclidean Robbins-Monro schemes, albeit with a significantly more involved analysis that requires a number of new geometric ingredients. We showcase the flexibility of the proposed RRM framework by using it to establish the convergence of a retraction-based analogue of the popular optimistic / extra-gradient methods for solving minimization problems and games, and we provide a unified treatment for their convergence.

1. Introduction

Background and motivation. Consider a nonlinear system of equations of the general form

$$\text{Find } z^* \in \mathcal{M} \text{ such that } V(z^*) = 0 \tag{NLS}$$

where $\mathcal{M}$ is a smooth manifold and V is a vector field on $\mathcal{M}$. Root-finding problems of this type play a critical role in many areas of mathematical programming and learning theory, from Riemannian optimization and game theory to reinforcement learning and optimal control – e.g., when designing the optimal path of a robotic arm or employing natural gradient methods [30] over smooth statistical manifolds.

In this paper, we are interested in the case where V is stochastic, i.e., $V(z) = \mathbb{E}[\mathsf{V}(z;\omega)]$ for some random variable ω with unknown distribution. In this case, when $\mathcal{M} = \mathbb{R}^d$, the method of choice for solving (NLS) is the *Robbins–Monro* (RM) algorithm

$$Z_{n+1} = Z_n + \gamma_n \mathsf{V}(Z_n;\omega_n) \tag{RM}$$

where $\gamma_n > 0$ is a variable step-size sequence and ω_n is an i.i.d. sequence of samples (equivalently, this could be thought of as accessing V via a stochastic black-box oracle). This method was introduced in the seminal papers of Robbins and Monro [60] and Kiefer and Wolfowitz [32], and the first general convergence results were obtained by Ljung [45, 46] for gradient problems, i.e., $V = -\nabla f$ for some potential function f on $\mathcal{M}$. This led to

[⋆]Equal contribution.
[*]Institute for Machine Learning, Universitaetstrasse 6, 8092 Zurich, Switzerland.
[§]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France and Criteo AI Lab.
E-mail addresses: mkarimi@inf.ethz.ch, yaping.hsieh@inf.ethz.ch, panayotis.mertikopoulos@imag.fr, krausea@ethz.ch.

substantial activity on the topic, with major contributions by Kushner and co-authors [5, 7, 9, 36, 37, 39], and many others. However, the linear structure of $\mathbb{R}^d$ is deeply ingrained in all these works – as well as the method's very definition – preventing its use to solve many important stochastic approximation (SA) problems on *manifolds*: the d-dimensional torus for a robotic arm with d joints, the Grassman or Stiefel manifolds for robust principal component analysis, the hyperbolic spaces for text and graph embedding, etc.

Our contributions in the context of related work. Our paper seeks to lift this limitation by replacing the "+" operation in (RM) with the Riemannian exponential map $\exp_{Z_n}(\cdot)$ on $\mathcal{M}$ – or, more generally (and often more tractably), a *retraction* on $\mathcal{M}$ based at Z_n. In Riemannian optimization, this approach was pioneered by Bonnabel [10] who examined the case where V is the Riemannian gradient of some objective function f. Subsequent works [13, 16, 42, 67, 69, 73] expanded on the results of [10] for Riemannian stochastic gradient descent, while similar results were obtained in [8, 22, 27, 43] for Riemannian proximal point methods.

All these works focus exclusively on the case where V is a gradient field, so they do not apply to general, non-gradient instances of (NLS) which are crucial for min-max problems, games and multi-agent learning problems. A partial extension to the non-gradient case was provided by a line of works [15, 21, 23, 31, 53, 66], which examined the use of Riemannian extra-gradient methods under the assumption of (geodesic) *monotonicity*. This is a strong, convexity-type assumption which posits that V globally points towards its (necessarily connected) root system in a suitable, geodesic sense; convergence is then obtained following a similar line of reasoning as in the case of monotone operator theory in Euclidean spaces [4, 20].

Our paper does not make any such assumptions and directly examines the dynamics of Riemannian Robbins–Monro methods for *general* vector fields V, gradient and non-gradient, monotone and non-monotone alike. In this regard, our main contributions can be summarized as follows:

(1) We introduce a *generalized Riemannian Robbins–Monro* template which includes as special cases all methods mentioned above (Riemmanian stochastic gradient descent, extra-gradient, proximal point methods, etc.), as well as a number of new SA schemes for (NLS).

(2) We show that, under mild technical conditions on $\mathcal{M}$, the sequence of generated points forms an "approximate solution" – an *asymptotic pseudotrajectory* (APT) to be exact – of an associated *deterministic* dynamical system (Theorem 1), and converges with probability 1 to the so-called *internally chain-transitive* (ICT) sets thereof (Theorem 2).

In gradient and strictly monotone problems, these ICT sets are precisely the roots of V [63], so we immediately recover many of the asymptotic convergence results mentioned above (often under much weaker assumptions). In addition, our framework applies to several interesting settings beyond gradient or monotone systems – such as ordinal potential games, supermodular games, and cooperative dynamics – and covers a significantly wider class of SA schemes.

Tools and techniques. In the absence of a linear structure on $\mathcal{M}$, the major challenge we have to overcome is the lack of a suitable *coordinate frame* with which to analyze the trajectories of Riemannian SA algorithms. This reflects the dichotomy that, unlike the case of $\mathbb{R}^d$, points and vectors on manifolds obey fundamentally different rules and have to be compared using different moving frames. To circumvent this obstacle, we introduce an *extended Fermi coordinate* frame inspired by Manasse and Misner [48], and we use it to

prove that Riemannian SA schemes enjoy similar error bounds as in Euclidean spaces, up to some high-order terms that vanish in the long run. The aggregation and propagation of these errors can then be controlled using arguments from martingale limit theory which ultimately yield the convergence properties mentioned above.

A concurrent approach to establish the APT property in Riemannian SA schemes is due to Shah [61], who assumes the existence of a local diffeomorphism mapping geodesic interpolations to linear interpolations in a Euclidean space. However, the existence of such a diffeomorphism on every point of $\mathcal{M}$ implies that the manifold is globally *flat*, i.e., essentially Euclidean [28]; this assumption is far too restrictive for bona fide Riemannian applications, so the analysis of [61] is not relevant for our purposes. An additional issue is that the error bounds employed by Shah [61, p. 1131] rule out vector fields with a rotational component – such as $V(x, y) = (-y, x)$ on $\mathbb{R}^2$ – further limiting the applicability of their techniques to our setting.

Finally, the recent papers by Durmus et al. [18, 19] also consider a generic version of Robbins–Monro schemes, with both vanishing and constant step-sizes. The analysis of the latter type of schemes cannot lead to convergence with probability 1, so the results of [19] are necessarily ergodic in nature and hence out of our paper's scope. The setting of [18] is closer in spirit to our own, and it also accounts for the effects of bias in the queries to V; however, the results obtained therein concern dynamics that admit a *Lyapunov function* – the so-called "gradient-like" case [5] – so there is no overlap with our analysis.

2. Notation and preliminaries

Throughout our paper, $\mathcal{M}$ will denote a d-dimensional, geodesically complete, smooth manifold equipped with a Riemannian metric $\langle \cdot, \cdot \rangle_z$ with its induced norm $\|\cdot\|_z$; we refer the reader to [40] for standard definitions and notations (such as tangent spaces $\mathcal{T}_z\mathcal{M}$). The Riemannian gradient will be simply denoted by ∇, and the Euclidean 2-norm is denoted by $\|\cdot\|_2$.

For any curve γ, the notation $\dot{\gamma}$ will always denote the derivative with respect to time. Given any points $z, z' \in \mathcal{M}$ and a vector $v \in \mathcal{T}_z\mathcal{M}$, we denote by $\Gamma_{z \to z'}(v) \in \mathcal{T}_{z'}\mathcal{M}$ the vector obtained by parallel transporting v along the minimizing geodesic connecting z and z'; if the minimizing geodesics are not unique, then $\Gamma_{z \to z'}(v)$ is understood as the parallel transport along any of them.

We also say that a vector field V on $\mathcal{M}$ is (geodesically) L-Lipschitz if, for all $z, z' \in \mathcal{M}$,

$$\|\Gamma_{z \to z'}(V(z)) - V(z')\|_{z'} = \|V(z) - \Gamma_{z' \to z}(V(z'))\|_z \leq L \operatorname{dist}(z, z'),$$

where dist denotes the Riemannian distance induced by $\langle \cdot, \cdot \rangle_z$. All vector fields in this paper are assumed to be L-Lipschitz and bounded, i.e., $G := \sup_{z \in \mathcal{M}} \|V(z)\|_z < \infty$.

3. Riemannian Robbins–Monro schemes: Definition, dynamics, and convergence

We begin our analysis by introducing the core algorithmic template of *Riemannian Robbins–Monro* schemes. The material related to asymptotic pseudotrajectories is introduced in Section 3.2; Section 3.3 states all the assumptions and discusses their generality, and our main results are presented in Sections 3.4–3.5.

3.1. The Riemannian Robbins–Monro template.

As we stated before, the main idea behind the *Riemannian Robbins–Monro* (RRM) template is to replace addition along "straight lines" in Euclidean SA schemes with the Riemannian exponential mapping. This leads to the recursive update

$$Z_{n+1} = \exp_{Z_n}(\gamma_n(V(Z_n) + W_n)) \tag{RRM}$$

where

(1) $Z_n \in \mathcal{M}$ denotes the state of the algorithm at each iteration counter $n = 1, 2, \dots$
(2) $W_n \in \mathcal{T}_{Z_n}\mathcal{M}$ is an abstract error term (described in detail below).
(3) $\gamma_n > 0$ is the method's step-size.

In the above, the error term W_n is generated *after* Z_n so W_n is *not* adapted to the natural filtration $\mathcal{F}_n \coloneqq \sigma(Z_1, \dots, Z_n)$ of Z_n (i.e., W_n is *random* relative to $\mathcal{F}_n$). We will also write

$$\hat{V}_n = V(Z_n) + W_n \tag{1}$$

so $\hat{V}_n$ can be seen as a noisy estimator of $V(Z_n)$. To differentiate between "random" (zero-mean) and "systematic" (non-zero-mean) errors, it will be convenient to further decompose W_n as

$$W_n = U_n + b_n \tag{2}$$

where $b_n = \mathbb{E}[W_n \,|\, \mathcal{F}_n]$ represents the systematic component and $U_n = W_n - b_n$ captures the random, zero-mean part. To quantify all this, we will consider the following descriptors for W_n:

$$a) \quad \textit{Bias:} \qquad B_n = \mathbb{E}[\|b_n\|_{Z_n} \,|\, \mathcal{F}_n] \tag{3a}$$
$$b) \quad \textit{Variance:} \qquad \sigma_n^2 = \mathbb{E}[\|U_n\|_{Z_n}^2 | \mathcal{F}_n] \tag{3b}$$

Below, when $B_n = \sigma_n = 0$, we refer to the model as deterministic.

3.2. **Mean dynamics and asymptotic pseudotrajectories.** In analogy to the ODE method of stochastic approximation for ordinary, Euclidean Robbins–Monro schemes [5, 9], we will view (RRM) as a noisy Euler discretization of the *Riemannian mean dynamics*

$$\dot{\theta}(t) = V(\theta(t)). \tag{RMD}$$

To make this analogy precise, we will require a measure of "closeness" between the iterates of (RRM) and the integral curves of (RMD). To this end, let $\tau_n = \sum_{k=1}^{n-1} \gamma_k$ denote the "effective time" that has elapsed till the n-th iteration of (RRM), and define the (continuous-time) *geodesic interpolation* $Z(t)$ of Z_n as

$$Z(t) = \exp_{Z_n}\left(\frac{t - \tau_n}{\tau_{n+1} - \tau_n}\left(\gamma_n[V(Z_n) + W_n]\right)\right) \tag{Int}$$

for all $t \in [\tau_n, \tau_{n+1}]$, $n \geq 1$ (so $Z(\tau_n) = Z_n$ for all n). To compare $Z(t)$ to the solution orbits of (RMD), we will further consider the *flow* $\Theta \colon \mathbb{R}_+ \times \mathcal{M} \to \mathcal{M}$ of (RMD), where $\Theta_h(z)$ is simply the orbit of (RMD) at time $h \in \mathbb{R}_+$ with initial condition $z(0) = z \in \mathcal{M}$. We then have the following notion of "asymptotic closeness" due to Benaïm and Hirsch [7]:

Definition 1 (Asymptotic pseudotrajectories)**.** We say that $Z(t)$ is ((a.s.)) an *asymptotic pseudotrajectory* (APT) of (RMD) if, for all $T > 0$, we have

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \operatorname{dist}(Z(t+h), \Theta_h(Z(t))) = 0 \quad \text{almost surely.} \tag{4}$$

Intuitively, (4) states that, as $t \to \infty$, one cannot distinguish between the interpolation $Z(t+h)$ and the Riemannian mean dynamics with initial condition $Z(t)$. The rigorous connection between (RMD) and (RRM) provided by this notion will become clear in Section 3.4.

3.3. **Technical assumptions and requirements.** We now introduce the basic assumptions of our model and discuss their generality.

Step-size, noise and bias assumptions. Following the literature on stochastic approximation [5, 36], we will make the following standard assumptions for the step-size of (RRM):

Assumption 1. The step-size sequence γ_n, $n = 1, 2, \dots$ of (RRM) satisfies the Robbins–Monro summability conditions $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$.

Our second assumption concerns the error terms that appear in (RRM):

Assumption 2. The bounds on the noise and bias in (3a)-(3b) satisfy

$$B_n \to 0 \quad \text{and} \quad \sup_n \sigma_n^2 < \infty \quad ((\text{a.s.}) \ ) \tag{5}$$

Of the above, Assumption 1 is explicit and is controlled by the algorithm designer. By contrast, Assumption 2 depends on the primitives of the problem (the law providing measurements of V, the specific setup of (RRM), etc.), so it is more delicate; we discuss it in detail in Section 4, where we show that it is satisfied for a range of important instantiations of (RRM).

Boundedness and stability assumptions. To exclude problems where the iterates of stochastic approximation (SA) methods may escape to infinity, a standard practice in the literature is to make this into an explicit assumption. In our manifold setting, this can be stated as follows:

Assumption 3. For some (and hence, all) $p \in \mathcal{M}$, the sequence Z_n generated by (RRM) satisfies

$$\sup_n \operatorname{dist}(p, Z_n) < \infty \quad \text{with probability 1.} \tag{6}$$

Instead of simply imposing Assumption 3 as is commonly done in the literature [5, 11, 39], we propose below a weaker condition that ensures (6), and which is general enough to cover most of the envisioned applications. To state it, we observe first that $\mathcal{M}$ typically falls under one of the following two categories:

(1) The manifold $\mathcal{M}$ is compact: This is the case for problems with simple manifold constraints such as spheres, balls, simplices, hypercubes, etc. Another important example is the *Grassmannian manifold* endowed with certain natural metrics; see [71, 72].

(2) The *sectional curvatures* of $\mathcal{M}$ are non-positive: The most important such instances are matrix manifolds [44, 62] and hyperbolic spaces [54]; additional examples can be found in [14].

In the first case, Assumption 3 is trivially satisfied. For the second case, we propose the following explicit structural replacement:

Assumption 3′. The sectional curvatures of M are non-positive and uniformly bounded from below by $-K_{\min} > -\infty$, and the vector field V is *weakly asymptotically coercive*, i.e.,

$$\big\langle V(z), \nabla \operatorname{dist}^2(p, z) \big\rangle \leq 0 \tag{WAC}$$

for some (and hence all) $p \in \mathcal{M}$ and for all z with sufficiently large $\operatorname{dist}(p, z)$. In addition, we assume that the noise and bias bounds in (3a)-(3b) satisfy

$$\textstyle\sum_{n=1}^{\infty} \mathbb{E}[\gamma_n B_n] < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \mathbb{E}\big[\gamma_n^2 (B_n^2 + \sigma_n^2)\big] < \infty. \tag{7}$$

Remark 1. We shall see later in the proof of Proposition 2 that, for most algorithms, (7) is weaker than (5) and hence can be discarded; for completeness, we also note that the differentiability of the radial function $r^2(z) = \operatorname{dist}^2(p, z)$ follows from Cartan's theorem [40].

The inspiration for (WAC) comes from the Euclidean case where a standard criterion to ensure stability is the coercivity requirement $\lim_{z\to\infty}\langle V(z), z\rangle/\|z\|_2 = -\infty$ [4, 20, 55]. This imposes a stringent growth condition on V which, mutatis mutandis, (WAC) relaxes significantly: under (WAC), it suffices if V does not have a persistent "outward-pointing" component.

In view of all this, our first result shows that, in simply connected manifolds, Assumption 3 can be replaced by Assumption 3′ (which, as we discussed, is much weaker in general):

Proposition 1. *If $\mathcal{M}$ is simply connected, Assumption 3′ implies Assumption 3.*

The most important step in the proof of Proposition 1 is to construct a suitable *energy function* which remains bounded under (WAC). However, due to the nonlinear structure of Riemannian manifolds, the resulting construction is fairly involved, so we defer it to Appendix A.

Conjugacy assumptions. A key notion in the analysis to follow is the so-called *Picard flow* $\lambda : \mathbb{R}^+ \to \mathcal{M}$ associated with the interpolation $Z(t)$, defined here as the solution of the following system of ODEs:

$$\dot\lambda(h) = \Gamma_{Z(t+h)\to\lambda(h)}\left(V(Z(t+h))\right) \tag{PFlow}$$

with initial condition $\lambda(0) = Z(t)$. The term "Picard flow" stems from the fact that, in Euclidean spaces, the integral $\int_0^h V(Z(t+s))\,ds$ is the basic iteration in Picard's method of successive approximations for solving ODEs. In the case of (PFlow), the parallel transport is the extra ingredient required to express the idea of "integrating V along $Z(t)$", so (PFlow) can be seen as a natural generalization of the Picard iteration to Riemannian manifolds.

Now, recall that two points $z \in \mathcal{M}$ and $z' \coloneqq \exp_z(v)$ are said to be *conjugate* along a minimizing geodesic if $\exp_z(\cdot)$ is minimizing but *fails* to be a locally diffeomorphism in a neighborhood of v [40]. Define

$$\mathcal{C}_\lambda(t,T) \coloneqq \{h \in [0,T] \mid Z(t+h) \text{ is conjugate to } \lambda(h)\},$$
$$\mathcal{C}_\theta(t,T) \coloneqq \{h \in [0,T] \mid Z(t+h) \text{ is conjugate to } \theta(h)\},$$

and let $\mathcal{C}(t,T) = \mathcal{C}_\lambda(t,T)\bigcup\mathcal{C}_\theta(t,T)$. We then make the following technical assumption:

Assumption 4. For all t and T, $\mathcal{C}(t,T)$ is a nowhere dense subset of $[0,T]$ with probability 1.

At first sight, Assumption 4 may appear somewhat opaque but, in practice, it is very mild and is satisfied by most applications. Indeed, since the set of points conjugate to $Z(t+h)$ is at most one-dimensional [70], the only way to violate Assumption 4 is if the curves $Z(t+\cdot)$ and $\theta(\cdot)$ (or $\lambda(\cdot)$) simultaneously traverse the same one-dimensional submanifold, which happens with probability 0 if the distribution of the noise U_n is non-singular. It is also worth noting that Assumption 4 is trivially true on negatively curved spaces by the Cartan–Hadamard theorem [40] and, finally, it is straightforward to verify Assumption 4 manually on the manifolds that arise most often in practical applications (such as spheres, balls, Grassmannians, or fixed-rank spectrahedra), cf. [40], [50] and references therein.

3.4. **From discrete to continuous: RRM schemes and APTs.** We are finally in a position to state and discuss our main results concerning the convergence properties of (RRM). Our first result describes the way in which (RRM) can be approximated by (RMD):

Theorem 1. *Suppose that Assumptions 1–4 hold. Then, with probability* 1*, the geodesic interpolation $Z(t)$ of the sequence of iterates of* (RRM) *is an APT of* (RMD)*.*

The proof of Theorem 1 is fairly involved and the geometric scaffolding required is quite delicate, so we first provide here a high-level outline. In brief, the main obstacle that we

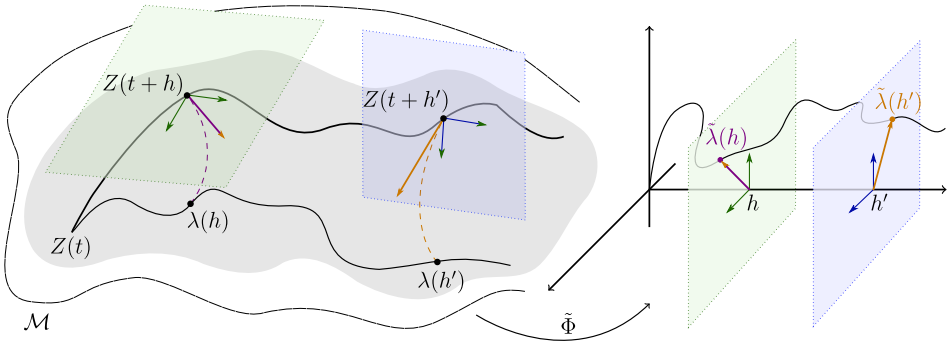

Figure 1: Illustration of an extended Fermi coordinate frame. The moving observer $Z(t+h)$ measures a curve $\lambda(h)$ using time-indexed geodesics and an "inertial frame system", i.e., frames obtained via parallel transport from $Z(t)$. For each h, $\tilde{\lambda}(h) \in \mathbb{R}^d$ is the normal coordinate of $\lambda(h) \in \mathcal{M}$. The space-time map $\tilde{\Phi} : \mathbb{R}^+ \times \mathcal{M} \to \mathbb{R}^+ \times \mathbb{R}^d$ is locally defined on a neighborhood containing $Z(t+h)$ and $\lambda(h)$.

have to overcome in order to extend Euclidean APT theory to the manifold setting is the following dichotomy:

- In view of definition (4), we need a coordinate system to compare the distance between two *points*. This is typically done in *normal coordinates* [40]; on the other hand, comparing *vectors* that belong to different tangent spaces in normal coordinates is a formidable task.
- Since the flow $\Theta_t(z)$ is driven by a vector field, we will also need to compare *vectors* living on different tangent spaces. To this end, the most convenient framework is the *parallel frame system*, described in detail in Appendix B.1. Unfortunately, the parallel frame system cannot coexist with a normal coordinate system unless $\mathcal{M}$ is flat, i.e., the manifold's Riemannian curvature tensor vanishes everywhere [28].

In words, the normal coordinate system is where distances behave as if the space were Euclidean, and the parallel frame system is where vectors behave as in the Euclidean case; unfortunately, the only setting where these two descriptions come together is in flat manifolds, which are "almost" Euclidean to begin with. To circumvent this hurdle, we take the following technical path:

(1) Based on the idea of *Fermi coordinates* [48], which can be intuitively understood as "normal coordinates along a curve" (see Fig. 1), we begin by constructing an *extended Fermi coordinate frame* that allows us to focus on a neighborhood of $Z(t)$ containing all the information we need. [This is a challenging but otherwise purely technical step and can be safely omitted without losing the high-level picture.]

(2) Using the extended Fermi coordinates constructed above, we can reduce the task of comparing the distance between two Riemannian curves to comparing several *Euclidean*, albeit individually intractable, vector fields. This step will incur an error term as compared to the Euclidean analysis; see (17) and (18).

(3) To obtain expressions of vector fields that are more amenable to computation, we will switch from the extended Fermi coordinates to the parallel frame system and bound the difference between the two. This step will pick up another error term as compared to the Euclidean analysis, cf. Eqs. (27) and (28).

(4) Serendipitously – and, perhaps, surprisingly – these additional error terms can be managed without any further assumptions, and a more involved analysis in the spirit of [5] concludes our proof.

Proof of Theorem 1. We will need the following ingredients:

- The concept of *parallel frame system*, reviewed in Appendix B.1.
- A key lemma by Fujita and Kotani [24] and Takahashi and Watanabe [65], which bounds the distortion of velocities measured by a moving observer on a manifold relative to flat spaces; see Lemma B.1.
- A comparison lemma between the differential of exponential map and parallel transport, recited in Lemma B.2.

We now commence the proof. For ease of reading, we break down our proof into several steps:

Step 1: Discrete-to-continuous transformations; noise stability. Recall the "effective time" $\tau_n = \sum_{k=1}^{n-1} \gamma_k$ as the time that has elapsed at the n-th iteration of the discrete-time process Z_n; recall also the definition (Int) of the continuous-time interpolation $Z(t)$ of Z_n as

$$Z(t) = \exp_{Z_n}\left(\frac{t-\tau_n}{\tau_{n+1}-\tau_n}\left(\gamma_n[V(Z_n)+W_n]\right)\right). \tag{Int}$$

We will further require the "continuous-to-discrete" correspondence

$$\mathfrak{m}(t) = \sup\{n \geq 1 : t \geq \tau_n\} \tag{8}$$

which measures the number of iterations required for the effective time τ_n of the process to reach the timestamp t.

Given an arbitrary sequence A_n (which can be numbers or points/vectors, either Euclidean or Riemannian), we will denote its piecewise-constant interpolation as

$$\overline{A}(t) = A_n \quad \text{for all } t \in [\tau_n, \tau_{n+1}),\ n \geq 1. \tag{9}$$

Using this notation, we can express (Int) in a differential form:[1]

$$\dot{Z}(t+h) = \Gamma_{\overline{Z}(t+h)\to Z(t+h)}\left(V\left(\overline{Z}(t+h)\right) + \overline{W}(t+h)\right). \tag{10}$$

Recall also the definition (RMD) and (PFlow):

$$\dot{\theta}(h) = V(\theta(h)), \tag{Flow}$$

$$\dot{\lambda}(h) = \Gamma_{Z(t+h)\to\lambda(h)}\left(V(Z(t+h))\right), \tag{PFlow}$$

both with initial condition $Z(t)$. We also set

$$\begin{aligned}\overline{\gamma}^\star(t) &\coloneqq \sup_{t\leq h\leq t+T}\overline{\gamma}(h),\\ \overline{B}^\star(t) &\coloneqq \sup_{t\leq h\leq t+T}\overline{B}(h).\end{aligned} \tag{11}$$

We will also need a noise stability criterion. Let $\{e_k(n)\}_{k=1}^d$ be an arbitrary sequence of orthonormal basis for $\mathcal{T}_{Z_n}\mathcal{M}$, and let $U_{n,\parallel}$ be the (Eucldiean) noise vector composed of components of the noise U_n in the basis $\{e_k(n)\}_{k=1}^d$:

$$U_{n,\parallel}^k \coloneqq \langle U_n, e_k(n)\rangle_{Z_n}.$$

[1] In the literal sense; it has nothing to do with differential p-forms.

Then it is easy to see that $\mathbb{E}[U_{n,\parallel}|\mathcal{F}_n] = 0$, and

$$\begin{aligned}\mathbb{E}\Big[\|U_{n,\parallel}\|_2^2|\mathcal{F}_n\Big] &= \mathbb{E}\big[\|U_n\|_{Z_n}^2|\mathcal{F}_n\big] \\ &\le \sup_n \sigma_n^2 \\ &< \infty\end{aligned}$$

by Assumption 2. Therefore, $\{U_{n,\parallel}\}_{n=1}^{\infty}$ is a sequence of martingale Euclidean noise vectors that is with zero mean and finite variance. The convergence of such noise sequences is a well-studied subject in stochastic approximation: Define

$$\Delta(t;T) \coloneqq \sup_{0\le h\le T}\left\|\int_t^{t+h} \overline{U}_{\parallel}(u)\,du\right\|_2. \tag{12}$$

It is classical (see e.g., [5]) that

$$\text{For all } T, \quad \lim_{t\to\infty}\Delta(t;T) = 0 \quad \text{(a.s.)} \tag{13}$$

Step 2: Everything is bounded. We first note that, by Proposition 1, $\sup_t r(Z(t)) < \infty$ where $r(\cdot)$ is the radial function defined in Assumption 3′.

We claim that Proposition 1 also implies the boundedness of the Picard flow. To see this, simply note that, since the parallel transport is an isometry,

$$\begin{aligned}\left\|\dot\lambda(h)\right\|_{\lambda(h)} &= \left\|\Gamma_{Z(t+h)\to\lambda(h)}\left(V(Z(t+h))\right)\right\|_{\lambda(h)} \\ &= \left\|V(Z(t+h))\right\|_{Z(t+h)},\end{aligned}$$

so that

$$\sup_{0\le h\le T}\operatorname{dist}(\lambda(0),\lambda(h)) \le T\cdot\sup_{0\le h\le T}\left\|\dot\lambda(h)\right\|_{\lambda(h)} < \infty,$$

which implies $\sup_{0\le h\le T} r(\lambda(h)) < \infty$. On the other hand, the boundedness for the flow follows readily from the L-Lipschitzness of V and the 1-Lipschitzness of dist:[2]

$$\begin{aligned}\frac{d}{dh}r(\theta(h)) &\le \left\|\dot\theta(h)\right\|_{\theta(h)} \\ &= \left\|V(\theta(h))\right\|_{\theta(h)} \\ &\le \left\|V(\theta(h)) - \Gamma_{p\to\theta(h)}\left(V(p)\right)\right\|_{\theta(h)} + \left\|\Gamma_{p\to\theta(h)}\left(V(p)\right)\right\|_{\theta(h)} \\ &\le Lr(\theta(h)) + \left\|V(p)\right\|_p.\end{aligned}$$

An application of the Groïwall's inequality then concludes $\sup_{0\le h\le T} r(\theta(h)) < \infty$. Therefore, all computations in the sequel can be restricted to a compact set. We may thus further assume that the sectional curvatures K are bounded. Since $\gamma_n \to 0$, for t large enough we may assume $\overline{\gamma}^\star(t) < 1$.

To summarize, for some $-\infty < K_{\text{low}} \le K_{\text{up}} < \infty$, we have

$$\begin{aligned}&\overline{\gamma}^\star(t) < 1, \\ &K_{\text{low}} \le K \le K_{\text{up}}.\end{aligned} \tag{14}$$

In addition, there exists an $R = R(T, L, G)$ depending on T, L, and G such that

$$\sup_{0\le h\le T}\Big\{\operatorname{dist}(Z(t+h),\theta(h)), \operatorname{dist}(Z(t+h),\lambda(h))\Big\} \le R < \infty. \tag{15}$$

[2]Due to the smoothness of the flow, the function $r(\theta(h))$ is always differentiable in h in the metric space sense [3], even though $\theta(h)$ might reach the cut locus of p.

Step 3: Constructing the extended Fermi coordinates. Let $\mathcal{V}$ be the neighborhood defined in Appendix B.2 restricted to $[t, t+T]$ (i.e., $\mathcal{V} = \bigcup_{h=0}^{T} \mathcal{U}_h$ where $\mathcal{U}_h$ is defined in (B.4)); evidently, $\mathcal{V}$ contains $\{Z(t+h) : h \in [0,T]\}$. Recall the Fermi coordinate system $\tilde{\Phi}$ in (B.5) restricted to $[t, t+T]$. Recall also that $\tilde{\Phi}$ carries a system of orthonormal frames $\{e_k(h)\}_{k=1}^{d}$, one for each $Z(t+h)$. Below, *we will always work in these frames with no explicit mention.*

Fix $\check{h} \in [0,T]$. Let γ_θ and γ_λ be two minimizing geodesics such that $\gamma_\theta(0) = \gamma_\lambda(0) = Z(t+\check{h})$, $\gamma_\theta(1) = \theta(\check{h})$ and $\gamma_\lambda(1) = \lambda(\check{h})$. Our first goal is to extend $\mathcal{V}$ to an open set of $\mathcal{M}$ that contains the geodesics γ_θ and γ_λ, while retaining the exponential mapping as local diffeomorphisms. This will serve a dual purpose:

(1) It enables us, for a fixed $\check{h}$, to consider the parallel frame systems at $Z(t+\check{h})$ and $\theta(\check{h})$ (or $Z(t+\check{h})$ and $\lambda(\check{h})$), so that we can easily compare the vector fields at these points; see Appendix B.1.

(2) We want to apply Lemma B.1 to the curves $\theta(\cdot)$ and $\lambda(\cdot)$; in order to make sense of $\dot{\theta}(\check{h})$ or $\dot{\lambda}(\check{h})$, $\mathcal{V}$ needs to contain both curves for at least an open time interval that includes $\check{h}$.

This is where Assumption 4 comes into play: Since the conjugate points are precisely where the exponential map ceases to be local diffeomorphisms [40], it is reasonable to expect that, away from the time points where $Z(t+h)$ is conjugate to $\theta(h)$ or $\lambda(h)$, it is always possible to extend $\mathcal{V}$ to include the geodesics connecting $Z(t+h)$ to $\theta(h)$ and $\lambda(h)$. Assumption 4 then simply posits that there cannot be "too many" such conjugate time points.

Consider any $\check{h} \notin \mathcal{C}(t,T)$ where $\mathcal{C}(t,T)$ is defined in (4), and assume also that $t+\check{h} \neq \tau_n$ for all n (i.e., $t+\check{h}$ is not a "corner" of (Int)). Since the exponential mapping is a local diffeomorphism away from conjugate points [40], it follows that

$$\exp_{Z(t+\check{h})}(\cdot) : \mathcal{T}_{Z(t+\check{h})}\mathcal{M} \to \mathcal{M}$$

is a local diffeomorphism at $\tilde{\lambda}(\check{h})$ and $\tilde{\theta}(\check{h})$, where $\tilde{\lambda}(\check{h})$ and $\tilde{\theta}(\check{h})$ are the normal coordinates of $\lambda(\check{h})$ and $\theta(\check{h})$ with center $Z(t+\check{h})$. By continuity of the flow/Picard flow and frame system $\{e_k(\cdot)\}_{k=1}^{d}$, there exists an open interval $(h_{\text{init}}, h_{\text{fin}})$ containing $\check{h}$ such that, for all $h \in (h_{\text{init}}, h_{\text{fin}})$,

$$\exp_{Z(t+h)}(\cdot) : \mathcal{T}_{Z(t+h)}\mathcal{M} \to \mathcal{M}$$

is a local diffeomorphism at $\tilde{\lambda}(h)$ and $\tilde{\theta}(h)$, where $\tilde{\lambda}(h)$ and $\tilde{\theta}(h)$ are the normal coordinates of $\lambda(h)$ and $\theta(h)$ with center $Z(t+h)$. Let γ_θ^h be a family of minimizing geodesics such that $\gamma_\theta^h(0) = Z(t+h)$ and $\gamma_\theta^h(1) = \theta(h)$, and define γ_λ^h similarly. Since both γ_θ^h and γ_λ^h are minimizing geodesics and since $\theta(h)$ and $\lambda(h)$ are not conjugate to $Z(t+h)$, [40, Theorem 10.15] ensures that no point on γ_θ^h or γ_λ^h is conjugate to $Z(t+h)$.

In short, we have shown that the exponential mapping is a local diffeomorphism at any point in the set:

$$\{\gamma_\theta^h\}_{h\in(h_{\text{init}},h_{\text{fin}})} \bigcup \{\gamma_\lambda^h\}_{h\in(h_{\text{init}},h_{\text{fin}})}.$$

The final step in our construction is to consider the union of all such $(h_{\text{init}}, h_{\text{fin}})$ for all $\check{h} \notin \mathcal{C}(t,T)$ and $t+\check{h} \neq \tau_n$; we denote the so obtained set by $\mathcal{H}$. We claim that

- $\mathcal{H}$ is a dense open subset of $[0,T]$, and
- $\mathcal{H}$ can be written as a countable union of disjoint open intervals: $\mathcal{H} = \bigcup_k (h_k, h_{k+1})$.

The first claim follows readily from Assumption 4 and the fact that the set $\{\check{h} : t+\check{h} = \tau_n \text{ for some } n\}$ is countable. To prove the second claim, simply note that, since all open intervals of $\mathbb{R}$ contain at least one rational number, it is impossible for an open set to be the union of uncountably many disjoint open intervals.

To summarize, we can extend $\mathcal{V}$ and $\tilde{\Phi}$ to an open set including

$$\{\gamma_\theta^h\}_{h\in(h_k,h_{k+1})} \bigcup \{\gamma_\lambda^h\}_{h\in(h_k,h_{k+1})},$$

which obviously contains $\bigcup_{h\in(h_k,h_{k+1})}\{\theta(h),\lambda(h)\}$. We call such a pair $(\mathcal{V},\tilde{\Phi})$ the *extended Fermi coordinate*, since it not only contains the central curve $h\to Z(t+h)$ as in the classical case, but also $\theta(h)$ and $\lambda(h)$ for almost every $h\in[0,T]$.

Step 4: Controlling the distance by decomposition. From now on, all computation happens within the scope of the extended Fermi coordinate $(\mathcal{V},\tilde{\Phi})$. By definition of $\theta(\cdot)$, and since $\tilde{\theta}(h)$ is the normal coordinate of $\theta(h)$ with center $Z(t+h)$, we have, for all $h\in\mathcal{H}$,

$$\begin{aligned}\operatorname{dist}(Z(t+h),\Theta_h(Z(t))) &= \operatorname{dist}(Z(t+h),\theta(h))\\ &= \|\tilde{\theta}(h)\|_2\\ &\le \|\tilde{\theta}(h)-\tilde{\lambda}(h)\|_2+\|\tilde{\lambda}(h)\|_2. \end{aligned}\tag{16}$$

Since $\mathcal{H}$ is a dense open subset of $[0,T]$ and since it is a countable union of open intervals, it follows that $\tilde{\theta}(h)$ and $\tilde{\lambda}(h)$ are differentiable except for a measure zero set. We can thus write

$$\|\tilde{\theta}(h)-\tilde{\lambda}(h)\|_2 = \left\|\int_0^h\left(\dot{\tilde{\theta}}(u)-\dot{\tilde{\lambda}}(u)\right)\mathrm{d}u\right\|_2.$$

By Lemma B.1, (RMD), and (PFlow), we have

$$\begin{aligned}\dot{\tilde{\theta}}^k(u) &= \dot{\theta}^k(u)-\dot{Z}^k(t+u)+\mathcal{O}\left(\|\tilde{\theta}(u)\|_2^2\right)\\ &= \tilde{V}^i(u,\tilde{\theta}(u))-\dot{Z}^k(t+u)+\mathcal{O}\left(\|\tilde{\theta}(u)\|_2^2\right),\\ \dot{\tilde{\lambda}}^k(u) &= \dot{\lambda}^k(u)-\dot{Z}^k(t+u)+\mathcal{O}\left(\|\tilde{\lambda}(u)\|_2^2\right)\\ &= \tilde{\Lambda}^k\left(u,\tilde{\lambda}(u)\right)-\dot{Z}^k(t+u)+\mathcal{O}\left(\|\tilde{\lambda}(u)\|_2^2\right),\end{aligned}$$

where $\tilde{V}^k(u,\tilde{\theta}(u))$ and $\tilde{\Lambda}^k\left(u,\tilde{\lambda}(u)\right)$ are, respectively, the k-th components of the vectors $V\big(\theta(u)\big)$ and $\Gamma_{Z(t+u)\to\lambda(u)}\left(V(Z(t+u))\right)$ in the frame induced by the normal coordinate with center $Z(t+u)$ and frame $\{e_k(u)\}_{k=1}^d$, and $\dot{Z}^k(\cdot)$ is defined in Lemma B.1. Denoting by $\tilde{V}(u,\tilde{\theta}(u))$ the (Euclidean) vector with components $\tilde{V}^i(u,\tilde{\theta}(u))$ and, by $\tilde{\Lambda}\left(u,\tilde{\lambda}(u)\right)$ the vector with components $\tilde{\Lambda}^k\left(u,\tilde{\lambda}(u)\right)$, we may thus write

$$\|\tilde{\theta}(h)-\tilde{\lambda}(h)\|_2 \le \left\|\int_0^h\left(\tilde{V}(u,\tilde{\theta}(u))-\tilde{\Lambda}\left(u,\tilde{\lambda}(u)\right)\right)\mathrm{d}u\right\|_2+\int_0^h R_1(u)\mathrm{d}u \tag{17}$$

where the remainder term $R_1(u)$ is of order $\mathcal{O}\left(\|\tilde{\theta}(u)\|_2^2+\|\tilde{\lambda}(u)\|_2^2\right)$. Noting that, by (15),

$$\begin{aligned}\|\tilde{\theta}(u)\|_2 &= \operatorname{dist}\left(Z(t+u),\Theta_u(Z(t))\right)\le R,\\ \|\tilde{\lambda}(u)\|_2 &= \operatorname{dist}\left(Z(t+u),\lambda(t+u)\right)\le R,\end{aligned}$$

we have $\|\tilde{\theta}(u)\|_2^2+\|\tilde{\lambda}(u)\|_2^2\le R\left(\|\tilde{\theta}(u)\|_2+\|\tilde{\lambda}(u)\|_2\right)$, and hence $R_1(u)=\mathcal{O}_R\left(\|\tilde{\theta}(u)\|_2+\|\tilde{\lambda}(u)\|_2\right)$ where $\mathcal{O}_R(\cdot)$ hides constants depending on R.

In the same vein, denoting by $\dot{\tilde{Z}}(t+u)$ the Euclidean vector whose k-th component is $\dot{Z}^k(t+u)$, we have

$$\|\tilde{\lambda}(h)\|_2 \leq \left\| \int_0^h \left(\tilde{\Lambda}\left(u, \tilde{\lambda}(u)\right) - \dot{\tilde{Z}}(t+u) \right) \mathrm{d}u \right\|_2 + \int_0^h R_2(u) \mathrm{d}u \tag{18}$$

where $R_2(u) = \mathcal{O}\left(\|\tilde{\lambda}(u)\|_2^2\right) = \mathcal{O}_R\left(\|\tilde{\lambda}(u)\|_2\right)$.

Step 5: Switching coordinate systems: From Fermi to parallel. So far, we have reduced the proof to comparing between the vectors in (17) and (18). However, these vectors are not amenable to further computation as they are expressed in the frames induced by the normal coordinates, and these frames may not even be orthonormal.

On the other hand, when expressed in the parallel frame system (see Appendix B.1) with a common base point $Z(t+u)$, the vectors $V\big(\theta(u)\big)$ and $\Gamma_{Z(t+u)\to\lambda(u)}\left(V(Z(t+u))\right)$ possess some favorable properties. To see this, parallel transport the frame $\{e_k(u)\}_{k=1}^d$ along the geodesic from $Z(t+u)$ to form an orthonormal frame $\{e'_k(u)\}_{k=1}^d$ of $\mathcal{T}_{\lambda(u)}\mathcal{M}$, and consider the Euclidean vector $\tilde{\Lambda}_{\|}\left(u, \tilde{\lambda}(u)\right)$ whose components are defined as

$$\begin{aligned}\tilde{\Lambda}_{\|}^k\left(u, \tilde{\lambda}(u)\right) &\coloneqq \langle \Gamma_{Z(t+u)\to\lambda(u)}\left(V(Z(t+u))\right), e'_k(u)\rangle_{\lambda(u)} \\ &= \langle V(Z(t+u)), e_k(u)\rangle_{Z(t+u)}.\end{aligned} \tag{19}$$

Similarly, parallel transport the frame $\{e_k(u)\}_{k=1}^d$ along the geodesic from $Z(t+u)$ to form an orthonormal frame $\{e''_k(u)\}_{k=1}^d$ of $\mathcal{T}_{\theta(u)}\mathcal{M}$, and define

$$\tilde{V}_{\|}^k\left(u, \tilde{\theta}(u)\right) \coloneqq \langle V\big(\theta(u)\big), e''_k(u)\rangle_{\theta(u)}. \tag{20}$$

Since the parallel transport is an isometry, we get

$$\begin{aligned}\tilde{V}_{\|}^k\left(u, \tilde{\theta}(u)\right) - \tilde{\Lambda}_{\|}^k\left(u, \tilde{\lambda}(u)\right) &= \left\langle V\big(\theta(u)\big), e''_k(u)\right\rangle_{\theta(u)} - \left\langle \Gamma_{Z(t+u)\to\lambda(u)}\left(V(Z(t+u))\right), e'_k(u)\right\rangle_{\lambda(u)} \\ &= \left\langle \Gamma_{\theta(u)\to Z(t+u)}\left(V\big(\theta(u)\big)\right), e_k(u)\right\rangle_{Z(t+u)} - \langle V(Z(t+u)), e_k(u)\rangle_{Z(t+u)} \\ &= \left\langle \Gamma_{\theta(u)\to Z(t+u)}\left(V\big(\theta(u)\big)\right) - V(Z(t+u)), e_k(u)\right\rangle_{Z(t+u)},\end{aligned}$$

whence

$$\begin{aligned}\left\|\tilde{V}_{\|}\left(u, \tilde{\theta}(u)\right) - \tilde{\Lambda}_{\|}\left(u, \tilde{\lambda}(u)\right)\right\|_2 &= \|\Gamma_{\theta(u)\to Z(t+u)}\left(V\big(\theta(u)\big)\right) - V(Z(t+u))\|_{Z(t+u)} \\ &\leq L\,\mathrm{dist}(\theta(u), Z(t+u)) = L\|\tilde{\theta}(u)\|_2\end{aligned} \tag{21}$$

where we have used the L-Lipschitzness of V and the fact that $\tilde{\theta}(u)$ is the normal coordinate with center $Z(t+u)$.

We would like to therefore replace $\tilde{V}\left(u, \tilde{\theta}(u)\right) - \tilde{\Lambda}\left(u, \tilde{\lambda}(u)\right)$ with $\tilde{V}_{\|}\left(u, \tilde{\theta}(u)\right) - \tilde{\Lambda}_{\|}\left(u, \tilde{\lambda}(u)\right)$ in (17). To this end, consider the difference in the k-th component of $\tilde{V}\left(u, \tilde{\theta}(u)\right)$ and $\tilde{V}_{\|}\left(u, \tilde{\theta}(u)\right)$:

$$\tilde{V}_{\|}^k\left(u, \tilde{\theta}(u)\right) - \tilde{V}^k\left(u, \tilde{\theta}(u)\right) = \left\langle V\big(\theta(u)\big), e''_k(u)\right\rangle_{\theta(u)} - \left\langle V\big(\theta(u)\big), \frac{\partial}{\partial z_k}\Big|_{\theta(u)}\right\rangle_{\theta(u)} \tag{22}$$

where $\frac{\partial}{\partial z_i}\Big|_{\theta(u)}$ is the k-th basis in the frame induced by the normal coordinate with center $Z(t+u)$ and frame $\{e_k(u)\}_{k=1}^d$. More specifically, denote by v the vector $\sum_{k=1}^d \tilde{\theta}^k(u) e_k(u) \in$

$\mathcal{T}_{Z(t+u)}\mathcal{M}$ and consider a family of geodesics

$$\gamma(s, s') \coloneqq \exp_{Z(t+u)}\big(s(v + s' e_k(u))\big). \tag{23}$$

Then

$$\begin{aligned}\frac{\partial}{\partial z_i}\Big|_{\theta(u)} &= \frac{\partial}{\partial s'}\gamma(1,0)\\ &= \mathrm{d}\exp_{Z(t+u)}(v)(e_k(u)).\end{aligned} \tag{24}$$

Invoking Cauchy-Schwartz, (14), Lemma B.2, and using the fact that $\{e_k(u)\}_{k=1}^d$ is orthonormal, we get

$$\begin{aligned}\left|\tilde{V}_{\|}^k\left(u, \tilde{\theta}(u)\right) - \tilde{V}^k\left(u, \tilde{\theta}(u)\right)\right| &= \left|\left\langle V\big(\theta(u)\big), e_k''(u) - \frac{\partial}{\partial z_k}\Big|_{\theta(u)}\right\rangle_{\theta(u)}\right|\\ &\le \|V\big(\theta(u)\big)\|_{\theta(u)} \cdot \left\|e_k''(u) - \frac{\partial}{\partial z_k}\Big|_{\theta(u)}\right\|_{\theta(u)}\\ &\le G \cdot K_{\max} \cdot f_{-K_{\max}}(v) \cdot \|e_k(u)_{\perp}\|_{Z(t+u)}\\ &\le G \cdot K_{\max} \cdot f_{-K_{\max}}(v) = G \cdot K_{\max} \cdot f_{-K_{\max}}(\|\tilde{\theta}(u)\|_2)\end{aligned} \tag{25}$$

where $K_{\max} = \max\left(|K_{\mathrm{up}}|, |K_{\mathrm{low}}|\right)$. Since a standard Taylor expansion argument shows that $\frac{\sinh(x)}{x} - 1 \le \cosh(x) \le e^x$ for all $x \ge 0$, by (B.7), (B.8), and (14), we get

$$\begin{aligned}f_{-K_{\max}}(\|\tilde{\theta}(u)\|_2) &\le \frac{1}{K_{\max}} \exp\left(\sqrt{K_{\max}}\|\tilde{\theta}(u)\|_2\right)\\ &\le \frac{e^{R\sqrt{K_{\max}}}}{R \cdot K_{\max}} \cdot \|\tilde{\theta}(u)\|_2\end{aligned} \tag{26}$$

where the last inequality follows from $\|\tilde{\theta}(u)\|_2 = \mathrm{dist}(Z(t+u), \Theta_u(Z(t))) \le R$. Combining (25) and (26), we thus get

$$\left|\tilde{V}_{\|}^k\left(u, \tilde{\theta}(u)\right) - \tilde{V}^k\left(u, \tilde{\theta}(u)\right)\right| \le \frac{G e^{R\sqrt{K_{\max}}}}{R} \cdot \|\tilde{\theta}(u)\|_2.$$

In short, we have shown that

$$\tilde{V}\left(u, \tilde{\theta}(u)\right) = \tilde{V}_{\|}\left(u, \tilde{\theta}(u)\right) + R_3(u). \tag{27}$$

Here, $R_3(u) = \mathcal{O}_{G, K_{\max}, R}\left(\|\tilde{\theta}(u)\|_2\right)$, where $\mathcal{O}_{G, K_{\max}, R}(\cdot)$ hides constants that depend on $G, K_{\max}$ and R. Exactly the same computation shows that, for some $R_4(u) = \mathcal{O}_{G, K_{\max}, R}\left(\|\tilde{\lambda}(u)\|_2\right)$,

$$\tilde{\Lambda}\left(u, \tilde{\lambda}(u)\right) = \tilde{\Lambda}_{\|}\left(u, \tilde{\lambda}(u)\right) + R_4(u). \tag{28}$$

Step 6: Concluding the proof. We will proceed by bounding (17) and (18).

Using (27), (28), and (21) in (17), we obtain:

$$\begin{aligned}\|\tilde{\theta}(h)-\tilde{\lambda}(h)\|_2 &\leq \left\| \int_0^h \left(\tilde{V}\left(u,\tilde{\theta}(u)\right)-\tilde{\Lambda}\left(u,\tilde{\lambda}(u)\right)\right)\mathrm{d}u\right\|_2 + \int_0^h R_1(u)\mathrm{d}u \\ &\leq \int_0^h \left\|\tilde{V}\left(u,\tilde{\theta}(u)\right)-\tilde{\Lambda}\left(u,\tilde{\lambda}(u)\right)\right\|_2 \mathrm{d}u + \int_0^h R_1(u)\,du \\ &\leq \int_0^h \left\|\tilde{V}_{\shortparallel}\left(u,\tilde{\theta}(u)\right)-\tilde{\Lambda}_{\shortparallel}\left(u,\tilde{\lambda}(u)\right)\right\|_2 \mathrm{d}u + \int_0^h (R_1+R_3+R_4)(u)\,du \\ &\leq L\int_0^h \left\|\tilde{\theta}(u)\right\|_2 \mathrm{d}u + \int_0^h (R_1+R_3+R_4)(u)\,du. \end{aligned} \tag{29}$$

We next turn to (18). Our first task is to obtain an expression for $\dot{\tilde{Z}}(t+u)$; recall that this is the Euclidean vector whose k-th component is $\dot{Z}^k(t+u)$. To this end, fix an iteration count n, and consider all u such that $t+u \in [\tau_n, \tau_{n+1})$ (that is, consider only the interpolation between Z_n and Z_{n+1}). We claim that $\dot{Z}^k(t+u)$ is constant throughout all such u, and, in particular, $\dot{Z}^k(t+u) = \dot{Z}^k(\tau_n)$. This readily follows by noticing that

1) The curve $Z(\cdot)$ is a geodesic segment when restricted to $[\tau_n, \tau_{n+1})$; see (Int).
2) The Fermi coordinate along $Z(\cdot)$, when restricted to $\{Z(s) : s \in [\tau_n, \tau_{n+1})\}$, is simply a parallel frame system: The frame $\{e_k(u)\}_{k=1}^d$ is obtained from parallel transporting $\{e_k(\tau_n)\}_{k=1}^d$ along $Z(\cdot)$ for all u such that $t+u \in [\tau_n, \tau_{n+1})$.

Thus, a simple calculation akin to (B.2) yields, for all such u,

$$\begin{aligned}\dot{Z}^k(t+u) &= \langle \dot{Z}(t+u), e_i(u)\rangle_{Z(t+u)} \\ &= \left\langle \Gamma_{Z(t+u)\to Z(\tau_n)}\left(\left(\dot{Z}(t+u)\right)\right), e_i(\tau_n)\right\rangle_{Z(\tau_n)} \\ &= \langle V(Z(\tau_n)) + W_n, e_i(\tau_n)\rangle_{Z(\tau_n)} = \langle V(Z(\tau_n)) + U_n + b_n, e_i(\tau_n)\rangle_{Z(\tau_n)} \end{aligned} \tag{30}$$

where we have used (10) and the definition of W_n in the last equality.

Armed with the above, we can obtain a succinct expression for $\dot{\tilde{Z}}(t+u)$ as follows. Let $\tilde{\tilde{z}}(u)$ be the normal coordinate of $\overline{Z}(t+u)$ with center $Z(t+u)$ (i.e., $(u,\tilde{\tilde{z}}(u)) = \tilde{\Phi}(u,\overline{Z}(t+u))$), and define a Euclidean vector $\tilde{V}_{\shortparallel}(u,\tilde{\tilde{z}}(u))$ by setting its k-th component to

$$\tilde{V}_{\shortparallel}^k\left(u,\tilde{\tilde{z}}(u)\right) \coloneqq \left\langle V\left(\overline{Z}(t+u)\right), e_i(\mathfrak{m}(t+u))\right\rangle_{\overline{Z}(t+u)} \tag{31}$$

where the mapping $\mathfrak{m}(\cdot)$ is defined in (8). Define also the Euclidean noise and bias vectors $U_{n,\shortparallel}$ and $b_{n,\shortparallel}$ by setting their components to

$$\begin{aligned} U_{n,\shortparallel}^k &\coloneqq \langle U_n, e_i(\mathfrak{m}(t+u))\rangle_{\overline{Z}(t+u)}, \\ b_{n,\shortparallel}^k &\coloneqq \langle b_n, e_i(\mathfrak{m}(t+u))\rangle_{\overline{Z}(t+u)}. \end{aligned}$$

Then (30) states precisely that

$$\dot{\tilde{Z}}(t+u) = \tilde{V}_{\shortparallel}(u,\tilde{\tilde{z}}(u)) + \overline{U}_{\shortparallel}(t+u) + \overline{b}_{\shortparallel}(t+u). \tag{32}$$

Substituting (32) into (18) and invoking (28), (11), and (12), we obtain

$$\begin{aligned}\|\tilde{\lambda}(h)\|_2 &\leq \left\| \int_0^h \left(\tilde{\Lambda}\left(u, \tilde{\lambda}(u)\right) - \tilde{V}_{\shortparallel}(u, \tilde{\tilde{z}}(u)) - \overline{U}_{\shortparallel}(t+u) - \bar{b}_{\shortparallel}(t+u) \right) \mathrm{d}u \right\|_2 + \int_0^h R_2(u)\mathrm{d}u \\ &\leq \int_0^h \left\| \tilde{\Lambda}_{\shortparallel}\left(u, \tilde{\lambda}(u)\right) - \tilde{V}_{\shortparallel}(u, \tilde{\tilde{z}}(u)) \right\|_2 \mathrm{d}u \\ &\qquad + \left\| \int_0^h \overline{U}_{\shortparallel}(t+u)\, du \right\|_2 + \left\| \int_0^h \bar{b}_{\shortparallel}(t+u)\, du \right\|_2 + \int_0^h (R_2+R_4)(u)\mathrm{d}u \\ &\leq \int_0^h \left\| \tilde{\Lambda}_{\shortparallel}\left(u, \tilde{\lambda}(u)\right) - \tilde{V}_{\shortparallel}(u, \tilde{\tilde{z}}(u)) \right\|_2 \mathrm{d}u \\ &\qquad + \Delta(t,T) + \overline{B}^{\star}(t)\cdot h + \int_0^h (R_2+R_4)(u)\mathrm{d}u. \end{aligned} \tag{33}$$

To bound the first term in (33), recall (19) and (31). An identical argument leading to (B.3) shows that

$$\begin{aligned}\left\| \tilde{\Lambda}_{\shortparallel}\left(u, \tilde{\lambda}(u)\right) - \tilde{V}_{\shortparallel}(u, \tilde{\tilde{z}}(u)) \right\|_2 &= \left\| V(Z(t+u)) - \Gamma_{\overline{Z}(t+u)\to Z(t+u)}\left(V(\overline{Z}(t+u))\right) \right\|_{Z(t+u)} \\ &\leq L \cdot \mathrm{dist}\left(\overline{Z}(t+u), Z(t+u)\right). \end{aligned} \tag{34}$$

Since $Z(\cdot)$ is a (not necessarily minimizing) geodesic on $[\mathfrak{m}(t+u), t+u]$, we have, by (32), (11), and (14),

$$\begin{aligned}\mathrm{dist}\left(\overline{Z}(t+u), Z(t+u)\right) &\leq \left\| \int_{\mathfrak{m}(t+u)}^{t+u} \dot{\tilde{Z}}(s)\, ds \right\|_2 \\ &= \left\| \int_{\mathfrak{m}(t+u)}^{t+u} \tilde{V}_{\shortparallel}(s, \tilde{\tilde{z}}(s)) + \overline{U}_{\shortparallel}(s) + \bar{b}_{\shortparallel}(s)\, ds \right\|_2 \\ &\leq \left\| \int_{\mathfrak{m}(t+u)}^{t+u} \tilde{V}_{\shortparallel}(s, \tilde{\tilde{z}}(s))\, ds \right\|_2 + \left\| \int_{\mathfrak{m}(t+u)}^{t+u} \overline{U}_{\shortparallel}(s)\, ds \right\|_2 + \left\| \int_{\mathfrak{m}(t+u)}^{t+u} \bar{b}_{\shortparallel}(s)\, ds \right\|_2 \\ &\leq \left(G + \overline{B}^{\star}(t)\right) \cdot \left(\int_{\mathfrak{m}(t+u)}^{t+u} ds \right) + \left\| \int_{\mathfrak{m}(t+u)}^{t+u} \overline{U}_{\shortparallel}(s)\, ds \right\|_2 \\ &\leq \left(G + \overline{B}^{\star}(t)\right) \overline{\gamma}^{\star}(t) + \left\| \int_{\mathfrak{m}(t+u)}^{t+u} \overline{U}_{\shortparallel}(s)\, ds \right\|_2. \end{aligned} \tag{35}$$

For t large enough, we have $\overline{\gamma}^{\star}(t) < 1$, and hence

$$\begin{aligned}\left\| \int_{\mathfrak{m}(t+u)}^{t+u} \overline{U}_{\shortparallel}(s)\, ds \right\|_2 &\leq \left\| \int_{t-1}^{\mathfrak{m}(t+u)} \overline{U}_{\shortparallel}(s)\, ds \right\|_2 + \left\| \int_{t-1}^{t+u} \overline{U}_{\shortparallel}(s)\, ds \right\|_2 \\ &\leq 2\Delta(t-1, T+1). \end{aligned} \tag{36}$$

Combing (34), (35), and (36), (33) then becomes

$$\begin{aligned}\|\tilde{\lambda}(h)\|_2 &\leq L\cdot h \cdot \left(\left(G + \overline{B}^{\star}(t)\right) \overline{\gamma}^{\star}(t) + 2\Delta(t-1, T+1) \right) + \Delta(t,T) + \overline{B}^{\star}(t)\cdot h + \int_0^h (R_2+R_4)(u)\mathrm{d}u \\ &\leq h \cdot C_{L,G}\left(\overline{B}^{\star}(t) + \overline{\gamma}^{\star}(t) + \Delta(t-1, T+1) \right) + \int_0^h (R_2+R_4)(u)\mathrm{d}u \end{aligned} \tag{37}$$

for some constant $C_{L,G}$ that depends only on L and G. Using (29) and (37), we can then bound (16) as

$$\begin{aligned}\operatorname{dist}(Z(t+h),\Theta_h(Z(t))) &= \|\tilde{\theta}(h)\|_2 \\ &\leq \|\tilde{\theta}(h)\|_2 + \|\tilde{\lambda}(h)\|_2 \qquad (38)\\ &\leq \|\tilde{\theta}(h)-\tilde{\lambda}(h)\|_2 + 2\|\tilde{\lambda}(h)\|_2 \\ &\leq L\int_0^h \left\|\tilde{\theta}(u)\right\|_2 \mathrm{d}u + h\cdot 2C_{L,G}\left(\overline{B}^\star(t)+\overline{\gamma}^\star(t)+\Delta(t-1,T+1)\right) \\ &\qquad + 3\int_0^h (R_1+R_2+R_3+R_4)(u)\,du \qquad (39)\end{aligned}$$

where $(R_1+R_2+R_3+R_4)(u) = \mathcal{O}_{G,K_{\max},R}\left(\left\|\tilde{\theta}(h)\right\|_2 + \left\|\tilde{\lambda}(h)\right\|_2\right)$. Therefore, there exists a constant $C_{L,G,K_{\max},R}$ depending only on $L, G, K_{\max}$, and R such that we may bound (38) as

$$\begin{aligned}\|\tilde{\theta}(h)\|_2 + \|\tilde{\lambda}(h)\|_2 \leq C_{L,G,K_{\max},R}\int_0^h \left(\left\|\tilde{\theta}(h)\right\|_2 + \left\|\tilde{\lambda}(h)\right\|_2\right) du \\ + h\cdot 2C_{L,G}\left(\overline{B}^\star(t)+\overline{\gamma}^\star(t)+\Delta(t-1,T+1)\right).\end{aligned}$$

Grönwall's inequality then implies

$$\|\tilde{\theta}(h)\|_2 + \|\tilde{\lambda}(h)\|_2 \leq h\cdot 2C_{L,G}\left(\overline{B}^\star(t)+\overline{\gamma}^\star(t)+\Delta(t-1,T+1)\right)\cdot e^{h\cdot C_{L,G,K_{\max},R}}. \qquad (40)$$

From (40), we may conclude the proof as follows:

$$\begin{aligned}\lim_{t\to\infty}\sup_{t\leq h\leq T}\operatorname{dist}(Z(t+h),\Theta_h(Z(t))) &\leq \lim_{t\to\infty}\sup_{t\leq h\leq T}\left(\|\tilde{\theta}(h)\|_2 + \|\tilde{\lambda}(h)\|_2\right) \\ &\leq \lim_{t\to\infty} T\cdot 2C_{L,G}\left(\overline{B}^\star(t)+\overline{\gamma}^\star(t)+\Delta(t-1,T+1)\right)\cdot e^{T\cdot C_{L,G,K_{\max},R}} \\ &= 0 \quad \text{a.s.}\end{aligned}$$

since $\lim_{t\to\infty}\overline{B}^\star(t) = \lim_{t\to\infty}\overline{\gamma}^\star(t) = 0$ a.s. by assumption, and $\lim_{t\to\infty}\Delta(t-1,T+1) = 0$ a.s. by (13). ■

3.5. **The Limits of Riemannian Robbins–Monro schemes.** Now, having established that the iterates of (RRM) constitute an APT of (RMD), we proceed to explore the formal link between asymptotic pseudotrajectories and the solution orbits of (RMD). To this end, we will need the important notion of an *internally chain-transitive* (ICT) set.

Definition 2. Let $\mathcal{S}$ be a nonempty compact subset of $\mathcal{M}$ and let Θ be a flow on $\mathcal{M}$. Then:

(1) $\mathcal{S}$ is **invariant** if $\Theta_t(\mathcal{S}) = \mathcal{S}$ for all $t\in\mathbb{R}$, i.e., $\Theta_t(z)\in\mathcal{S}$ for all $z\in\mathcal{S}$.
(2) $\mathcal{S}$ is an **attractor** of Θ if it is invariant and there exists a compact neighborhood $\mathcal{K}$ of $\mathcal{S}$ such that $\lim_{t\to\infty}\operatorname{dist}(\Theta_t(z),\mathcal{S}) = 0$ uniformly in $z\in\mathcal{K}$.
(3) $\mathcal{S}$ is **internally chain-transitive** if it is invariant and $\Theta|_{\mathcal{S}}$ admits no other attractors than $\mathcal{S}$.

In words, ICT sets can be seen as a "terminal object" for the dynamics (RMD): the ensemble of orbits that converges to an ICT set will not be ultimately contained in a smaller subset thereof. Our next result shows that, with probability 1, all limit points of (RRM) lie in some such set:

Theorem 2. *If Assumptions 1–4 hold, then Z_n converges almost surely to an ICT set of (RMD).*

The proof of Theorem 2 leverages Theorem 1 in an essential way and employs a foundational result of Benaïm and Hirsch [7] to establish the link with the ICT sets of (RMD):

Proof of Theorem 2. By Theorem 1, Z_n generates APTs of the mean dynamics (RMD). Now, let $\mathcal{L} = \bigcap_{t\geq 0} \mathrm{cl}(Z(t,\infty))$ be the limit set of $Z(t)$, i.e., the set of limit points of convergent sequences $Z(t_n)$ with $\lim_n t_n = \infty$. Our claim then follows by the limit set theorem of Benaïm and Hirsch [7, Theorem 8.2], which holds for an arbitrary metric (even non-Riemannian) space. ∎

An important message of Theorem 2 is that we can reduce the convergence analysis of (RRM) to the study of a deterministic, continuous-time dynamical system which is significantly simpler than the original *stochastic*, *discrete-time* system. In this sense, Theorem 2 delivers the same high-level message as the classic SA literature for *Euclidean* Robbins–Monro (RM) schemes. In particular, if V admits a potential or is strictly monotone, it is easy to verify that the only ICT sets of (RMD) are the roots of V [63], so we readily recover the series of asymptotic convergence results mentioned in the introduction. In the general case, the roots of V are always contained in its ICT sets, but the inclusion may be strict; however, this again becomes an equivalence under the standard assumptions which have been used in the Euclidean RM literature to obtain global convergence – cooperative dynamics, global asymptotic stability, etc. [5, 11, 36]. In the next section, we shall see that Theorems 1–2 further capture a series of Riemannian algorithms – old and new – in a unified fashion.

4. Applications and implications

We now proceed to show how a wide array of Riemannian algorithms can be seen as special cases of (RRM) – enabling in this way the tandem use of Theorems 1 and 2 to deduce their convergence properties. To simplify the presentation, we will make the following technical assumption:

Assumption 5. The injectivity radius of $\mathcal{M}$ is uniformly bounded from below by $\delta > 0$.

The injectivity radius at a point p on $\mathcal{M}$ is the largest radius for which the exponential map at p is a diffeomorphism, and the injectivity radius of $\mathcal{M}$ is the infimum over all such radii [40]. In this regard, Assumption 5 serves to ensure that exp is invertible at consecutive iterates of (RRM) so as to avoid local topological complications; we will later prove in Proposition 2 that $\exp^{-1}$ is well-defined in all the algorithms below.

4.1. Specific algorithms.

Throughout this section, we adopt a black-box setup [52] with *stochastic first-order oracle* (SFO) feedback. Specifically, when called at $z \in \mathcal{M}$ with random seed $\omega \in \Omega$, an SFO returns a random vector $V(z;\theta) \in \mathcal{T}_z\mathcal{M}$ of the form

$$V(z;\theta) = V(z) + \mathrm{Err}(z;\theta) \tag{SFO}$$

where the error term $\mathrm{Err}(z;\theta) \in \mathcal{T}_z\mathcal{M}$ is zero-mean and finite-variance with respect to the Riemannian metric:

$$\forall z \in \mathcal{M}, \quad \mathbb{E}[\mathrm{Err}(z;\theta)] = 0 \text{ and } \mathbb{E}[\|\mathrm{Err}(z;\theta)\|_z^2] \leq \sigma^2. \tag{41}$$

To facilitate comparisons with the Riemannian optimization literature, we will abusively refer to queries of V as "gradients"; we stress, however, that V *is not assumed to admit a potential*.

Algorithm 1 (Riemannian stochastic gradient methods)**.** The simplest *Riemannian stochastic gradient method* (RSGM) queries an SFO and proceeds as:

$$Z_{n+1} = \exp_{Z_n}(\gamma_n V(Z_n;\theta_n)), \tag{R-SGM}$$

As such, (R-SGM) admits a straightforward RRM representation by taking $W_n = \mathrm{Err}(Z_n;\theta_n)$. ¶

Algorithm 2 (Riemannian proximal point methods)**.** The (deterministic) *Riemannian proximal point method* (RPPM) is an implicit update rule of the form:

$$\exp_{Z_{n+1}}^{-1}(Z_n) = -\gamma_n V(Z_{n+1}). \tag{R-PPM}$$

The RRM representation of (R-PPM) is obtained by taking $W_n = U_n + b_n$ with

$$b_n = \Gamma_{Z_{n+1}\to Z_n}\left(V(Z_{n+1})\right) - V(Z_n)$$

and $U_n = 0$ (recall here that $\Gamma_{z\to z'}(v)$ denotes parallel transport along minimizing geodesics). ¶

Algorithm 3 (Riemannian stochastic extra-gradient)**.** Inspired by the original work of Korpelevich [34], the *Riemannian stochastic extra-gradient* (RSEG) method iterates as

$$\begin{aligned} Z_{n+1/2} &= \exp_{Z_n}\left(\gamma_n V(Z_n;\theta_n)\right), \\ Z_{n+1} &= \exp_{Z_n}\left(\Gamma_{Z_{n+1/2}\to Z_n}\left(\gamma_n V(Z_{n+1/2};\theta_{n+1/2})\right)\right) \end{aligned} \tag{R-SEG}$$

where θ_n and $\theta_{n+1/2}$ are independent random seeds for (SFO). To recast (R-SEG) in the RRM framework, simply take $U_n = \Gamma_{Z_{n+1/2}\to Z_n}\left(\mathrm{Err}(Z_{n+1/2};\theta_{n+1/2})\right)$ and $b_n = \Gamma_{Z_{n+1/2}\to Z_n}\left(V(Z_{n+1/2})\right) - V(Z_n)$. ¶

Algorithm 4 (Riemannian optimistic gradient)**.** Compared to (R-SGM), the scheme (R-SEG) involves two oracle queries per iteration. Inspired by Popov [56], we can instead "recycle" the last oracle query, leading to the Riemannian optimistic gradient method

$$\begin{aligned} Z_{n+1/2} &= \exp_{Z_n}\left(\gamma_n V(Z_{n-1}^+;\theta_{n-1})\right), \\ Z_{n+1} &= \exp_{Z_n}\left(\Gamma_{Z_{n+1/2}\to Z_n}\left(\gamma_n V(Z_{n+1/2};\theta_n)\right)\right). \end{aligned} \tag{R-OG}$$

(R-OG) can be seen as a special case of (RRM) by taking $U_n = \Gamma_{Z_{n+1/2}\to Z_n}\left(\mathrm{Err}(Z_{n+1/2};\theta_n)\right)$ and $b_n = \Gamma_{Z_{n+1/2}\to Z_n}\left(V(Z_{n+1/2})\right) - V(Z_n)$. ¶

In light of Theorems 1–2, to analyze the convergence of Algorithms 1–4 under a specific choice of step-size, we can simply verify Assumptions 1–4. More precisely, it suffices to show that (5) and (7) hold with probability 1. This is the purpose of the next proposition:

Proposition 2. *Under Assumption 5 and a step-size strategy such that* $A/n \leq \gamma_n \leq B/\sqrt{n(\log n)^{1+\varepsilon}}$ *for some* $A, B, \varepsilon > 0$*, the following hold for Algorithms 1–4:*

(1) With probability 1, Z_{n+1} *lies in the injectivity radius of* Z_n *for* n *large enough.*

(2) Equations (5) and (7) hold with probability 1. As a result, under Assumptions 3′–4, Algorithms 1–4 generate APTs of (RMD) and converge to its ICT sets.

Before proving Proposition 2, we first provide a convenient lemma which shows that, almost surely, the effect of the noise is asymptotically annihilated by the step-size:

Lemma 1. *Under the assumptions in Proposition 2, we have, with probability 1,*

$$\lim_{n\to\infty} \|\gamma_n V(Z_n;\theta_n)\|_{Z_n} = 0. \tag{42}$$

Proof of Lemma 1. By definition,

$$\begin{aligned} \|\gamma_n V(Z_n;\theta_n)\|_{Z_n} &= \|\gamma_n V(Z_n) + \mathrm{Err}(Z_n;\theta_n)\|_{Z_n} \\ &\leq \gamma_n G + \gamma_n \|\mathrm{Err}(Z_n;\theta_n)\|_{Z_n}. \end{aligned} \tag{43}$$

The first term goes to 0 by choice of step-sizes. To control the second term, note that by Chebyshev's inequality and (41), we have

$$\mathbb{P}\left(\|\mathrm{Err}(Z_n;\theta_n)\|_{Z_n} \geq \sqrt{n\log^{1+\frac{\epsilon}{2}} n}\right) \leq \frac{\sigma^2}{n\log^{1+\frac{\epsilon}{2}} n} \tag{44}$$

where ε is the same as in our choice of step-size in Proposition 2. In turn, this implies that

$$\sum_{n=2}^{\infty} \mathbb{P}\left(\|\mathrm{Err}(Z_n;\theta_n)\|_{Z_n} \geq \sqrt{n\log^{1+\frac{\epsilon}{2}} n}\right) < \infty$$

so, by the Borel-Cantelli lemma, we have $\|\mathrm{Err}(Z_n;\theta_n)\|_{Z_n} = \mathcal{O}\left(\sqrt{n\log^{1+\frac{\epsilon}{2}} n}\right)$ with probability 1. Hence, by our assumptions for the method's step-size, we get

$$\gamma_n\|\mathrm{Err}(Z_n;\theta_n)\|_{Z_n} = \mathcal{O}\left(\frac{\sqrt{n\log^{1+\frac{\epsilon}{2}} n}}{\sqrt{n\log^{1+\varepsilon} n}}\right) = \mathcal{O}\left(\frac{1}{\log^{\frac{\varepsilon}{4}} n}\right)$$

which, combined with (43), implies (42). ■

We are now ready to prove Proposition 2.

Proof of Proposition 2. The first claim of Proposition 2 is a direct consequence of Lemma 1 and Assumption 5. To prove the second claim, note that $\sum_n \gamma_n^2 < \infty$ by our choice of step-sizes. Thus, in order to prove (7), it suffices to show $\mathbb{E}[B_n] = \mathbb{E}[\|b_n\|] = \mathcal{O}(\gamma_n)$ and $\mathbb{E}[\sigma_n^2] \leq \sigma^2$ for some constant σ.

We next proceed method-by-method:

Algorithm 1: Riemannian stochastic gradient descent(/ascent). For (R-SGM), we have $W_n = U_n = \mathrm{Err}(Z_n;\theta_n)$ and $b_n = 0$, so both (5) and (7) follow from the stated assumptions for (SFO).

Algorithm 2: Riemannian proximal point method. For (R-PPM), we have $U_n = 0$ and

$$\begin{aligned}\|b_n\|_{Z_n} &= \|\Gamma_{Z_{n+1}\to Z_n}(V(Z_{n+1})) - V(Z_n)\|_{Z_n}\\ &\leq L\,\mathrm{dist}(Z_n, Z_{n+1})\\ &= \gamma_n L\|V(Z_n)\|_{Z_n}\\ &\leq \gamma_n LM = \mathcal{O}(\gamma_n)\end{aligned}$$

where we have used the L-Lipschitzness and G-boundedness of V, and the distance-minimizing property of exp within the injectivity radius.

Algorithm 3: Riemannian stochastic extra-gradient. For (R-SEG), we have $U_n = \Gamma_{Z_{n+1/2}\to Z_n}\left(\mathrm{Err}(Z_{n+1/2};\theta_{n+1/2})\right)$ so that

$$\begin{aligned}\mathbb{E}\left[\sigma_n^2\right] &= \mathbb{E}\left[\|U_n\|_{Z_n}^2\right]\\ &= \mathbb{E}\left[\left\|\mathrm{Err}(Z_{n+1/2};\theta_{n+1/2})\right\|_{Z_{n+1/2}}^2\right]\\ &\leq \sigma^2\end{aligned}$$

by (41) and the fact that the parallel transport map is a linear isometry.

To verify (5), by the definition of (R-SEG), we have

$$\begin{aligned}\|b_n\| = \|\Gamma_{Z_{n+1/2}\to Z_n}\left(V(Z_{n+1/2})\right) - V(Z_n)\|_{Z_n} &\leq L\,\mathrm{dist}\left(Z_{n+1/2}, Z_n\right)\\ &= \gamma_n L\|V(Z_n;\theta_n)\|_{Z_n} \to 0\end{aligned}$$

almost surely by Lemma 1.

Algorithm 4: Riemannian optimistic gradient. For (R-OG), we have $U_n = \Gamma_{Z_{n+1/2}\to Z_n}\left(\mathrm{Err}(Z_n;\theta_{n+1/2})\right)$ and $b_n = \Gamma_{Z_{n+1/2}\to Z_n}\left(V(Z_{n+1/2})\right) - V(Z_n)$, so $\mathbb{E}[\sigma_n^2] \le \sigma^2$ again holds by (41). The bias term can then be bounded exactly as in the case of Algorithm 3. ∎

4.2. **Operations that preserve Riemannian Robbins–Monro schemes.** To increase the efficiency of Riemannian iterative schemes, several important operations have been routinely performed on top of the base algorithms in Section 4.1. In this section, we show that the RRM template incorporates these operations in a unified fashion.

Retraction-Based Methods. When the exponential map is expensive to compute, a popular alternative in practice is to use the *retraction map* [1, 12], which is defined as a smooth mapping $\mathcal{R}_\cdot(\cdot) : \mathcal{TM} \to \mathcal{M}$ that mimics the exponential map up to first order: For all $(z, v) \in \mathcal{TM}$,

$$\mathcal{R}_z(0) = z, \quad \frac{d}{dt}\mathcal{R}_z(tv)\Big|_{t=0} = v. \tag{Rtr}$$

It turns out that, to replace the exponential map in Algorithms 1–4 with retraction, we only need to pay a small price on the noise assumption:

Proposition 3. *If, in addition to the assumptions in Proposition 2, the SFO has finite fourth moment:* $\mathbb{E}[\|\mathrm{Err}(z;\theta)\|_z^4] \le \kappa^2 < \infty$, *then replacing* $\exp_{Z_n}(\cdot)$ *with* $\mathcal{R}_{Z_n}(\cdot)$ *in Algorithms 1–4 results in another RRM scheme satisfying* (5) *and* (7) *with probability 1. As a result, under Assumptions 3′–4, they generate APTs of* (RMD) *and converge to its ICT sets.*

Remark 2. The condition $\mathbb{E}[\|\mathrm{Err}(z;\theta)\|_z^4] < \infty$ cannot be relaxed as shown by the Euclidean example of $\mathcal{R}_z(v) = z + (e^v - 1)$ and $\mathrm{Err}(z;\theta) \overset{\text{law}}{=} \sqrt{Y} - \mathbb{E}\left[\sqrt{Y}\right]$, where $\mathbb{P}(Y > y) = y^{-3/2}$ if $y \ge 1$, and $\mathbb{P}(Y > y) = 1$ otherwise.

Remark 3. In many practical settings, the map $\mathcal{R}_z(v)$ satisfies the stronger notion of "second-order retraction" [1, 12]. In this case, the proof technique below shows that it suffices to have $\mathbb{E}[\|\mathrm{Err}(z;\theta)\|_z^3] < \infty$ in Proposition 3.

Proof of Proposition 3. By definition, $\mathcal{R}_z(v)$ is a smooth map and hence satisfies $\lim_{v\to 0}\mathcal{R}_z(v) = z$. Then Lemma 1 readily implies that Z_{n+1} lies in the injectivity radius of Z_n with probability 1 for n large enough.

We first consider the retraction-based Algorithm 1:

$$Z_{n+1} = \mathcal{R}_{Z_n}(\gamma_n V(Z_n;\theta_n)). \tag{45}$$

Let $\tilde{V}_n \in \mathcal{T}_{Z_n}\mathcal{M}$ be the vector such that $\exp_{Z_n}\left(\gamma_n\tilde{V}_n\right) = Z_{n+1}$, i.e.,

$$\gamma_n\tilde{V}_n = \exp_{Z_n}^{-1}\left(\mathcal{R}_{Z_n}(\gamma_n V(Z_n;\theta_n))\right). \tag{46}$$

Then (45) is an RRM scheme with $W_n = \tilde{V}_n - V(Z_n)$ where $\tilde{V}_n$ is defined in (46). We will show that, under the assumption $\mathbb{E}[\|\mathrm{Err}(z;\theta)\|_z^4] < \infty$, the following holds with probability 1:

$$b_n = \mathbb{E}[W_n \,|\, \mathcal{F}_n] \to 0, \quad \sup_n \mathbb{E}\left[\|W_n\|_{Z_n}^2\right] < \infty \tag{47}$$

which obviously implies (5) and (7).

Consider the curve $c(t) \coloneqq \mathcal{R}_{Z_n}(tV(Z_n;\theta_n))$. By Lemma 1, for n large enough, $c(t)$ lies in the injectivity radius of Z_n almost surely for all $t \in [0, \gamma_n]$. Let $\hat{c}(t)$ be the smooth curve of $c(t)$ in the normal coordinate with base Z_n and an arbitrary orthonormal frame, and let $\hat{Z}_{n+1}$ be the normal coordinate of Z_{n+1}. Also, let $\tilde{V}_n^{\mathrm{N}}$ be the (Euclidean) vector of $\tilde{V}_n$

expanded in the chosen orthonormal basis, and define $V^{\mathrm{N}}(Z_n;\theta_n)$ and $\mathrm{Err}^{\mathrm{N}}(Z_n;\theta_n)$ similarly. By definition, $\hat{Z}_{n+1}$ is nothing but $\gamma_n \tilde{V}_n^{\mathrm{N}}$.

Since $Z_n = c(0)$ and $Z_{n+1} = c(\gamma_n)$, by properties of a retraction map we must have

$$\begin{aligned}\gamma_n \tilde{V}_n^{\mathrm{N}} &= \hat{c}(\gamma_n)\\ &= \hat{c}(0) + \gamma_n \dot{\hat{c}}(0) + \mathcal{O}\Big(\gamma_n^2 \|\dot{\hat{c}}(0)\|_2^2\Big)\\ &= \gamma_n V^{\mathrm{N}}(Z_n;\theta_n) + \mathcal{O}\big(\gamma_n^2 \|V(Z_n;\theta_n)\|_{Z_n}^2\big)\\ &=: \gamma_n V^{\mathrm{N}}(Z_n;\theta_n) + \gamma_n \tilde{b}_n \end{aligned} \tag{48}$$

where $\tilde{b}_n = \mathcal{O}\big(\gamma_n \|V(Z_n;\theta_n)\|_{Z_n}^2\big)$. Therefore, since $\mathbb{E}[\|\mathrm{Err}(z;\theta)\|_z^4] < \infty$ for all $z \in \mathcal{M}$, we have

$$\begin{aligned}\mathbb{E}\Big[\|W_n\|_{Z_n}^2\Big] &= \mathbb{E}\left[\left\|\mathrm{Err}^{\mathrm{N}}(Z_n;\theta_n) + \tilde{b}_n\right\|_2^2\right]\\ &< \infty.\end{aligned}$$

On the other hand,

$$\begin{aligned}\|b_n\|_{Z_n} &= \|\mathbb{E}[W_n \,|\, \mathcal{F}_n]\|_{Z_n}\\ &= \left\|\mathbb{E}\Big[\tilde{b}_n \,\Big|\, \mathcal{F}_n\Big]\right\|_2\\ &= \mathcal{O}\big(\gamma_n \|V(Z_n;\theta_n)\|_{Z_n}^2\big).\end{aligned}$$

By Chebyshev's inequality and the fact that $\mathbb{E}[\|\mathrm{Err}(z;\theta)\|_z^4] \le \kappa^2 < \infty$, we have

$$\mathbb{P}\left(\|\mathrm{Err}(Z_n;\theta_n)\|_{Z_n}^2 \ge \sqrt{n \log^{1+\frac{\varepsilon}{2}} n}\right) \le \frac{\kappa^2}{n \log^{1+\frac{\varepsilon}{2}} n}$$

where ε is the same as in our choice of step-size in Proposition 2. Using an calculation identical to Lemma 1, we conclude that

$$\gamma_n \|\mathrm{Err}(Z_n;\theta_n)\|_{Z_n}^2 = \mathcal{O}\left(\frac{\sqrt{n \log^{1+\frac{\varepsilon}{2}} n}}{\sqrt{n \log^{1+\varepsilon} n}}\right) = \mathcal{O}\left(\frac{1}{\log^{\frac{\varepsilon}{4}} n}\right)$$

which concludes the proof of (47).

For retraction-based Algorithms 2–4, by the above analysis, we may replace $\mathcal{R}_z(\gamma_n V(\cdot;\theta_n))$ with $\exp_z\Big(\gamma_n\Big(V(\cdot;\theta_n) + \tilde{b}_n\Big)\Big)$ where $\tilde{b}_n \to 0$ almost surely. The rest is the same as in the proof of Proposition 2. ■

Alternation. Consider a Riemannian (zero-sum) *two-player* game: $\min\max_{X\in\mathcal{M}_1, Y\in\mathcal{M}_2} \ell(X,Y)$ where $\mathcal{M}_1$ and $\mathcal{M}_2$ are smooth manifolds. Instead of performing simultaneous updates as in Algorithms 1–4, a common variant is to *alternately* update the min/max variables:

$$\begin{aligned}X_{n+1} &= \exp_{X_n}(\gamma_n[V_X(X_n,Y_n) + U_{X,n} + b_{X,n}])\\ Y_{n+1} &= \exp_{Y_n}(\gamma_n[V_Y(X_{n+1},Y_n) + U_{Y,n} + b_{Y,n}])\end{aligned} \tag{alt-R-RM}$$

where $(V_X, V_Y) \coloneqq (-\nabla_X \ell, \nabla_Y \ell)$, $U_{X,n}$ and $b_{X,n}$ are the noise and bias of the RRM scheme employed by the X variable, and similarly for $U_{Y,n}$ and $b_{Y,n}$. It is easy to see that alternating Algorithms 1–4 results in another convergent RRM scheme:

Proposition 4. *Let $Z_n \coloneqq (X_n, Y_n)$ where (X_n, Y_n) is generated by using Algorithms 1–4 in (alt-R-RM). Under the same assumptions as in Proposition 2, Z_n is itself an RRM scheme satisfying (5) and (7) with probability 1. As a result, under Assumptions 3′–4, Z_n generates an APT of (RMD) and converges to its ICT sets.*

Remark 4. It is easy to generalize Proposition 4 to any number of players and any type of (not necessarily zero-sum) games.

Remark 5. A simple but useful observation is that compositions of RRM schemes merely "slow down" the algorithm but do not change its asymptotic behavior: For any two RRM schemes RM_1 and RM_2, the update $Z_{n+1} = \mathrm{RM}_2 \circ \mathrm{RM}_1(Z_n)$ is equivalent to a new RRM scheme $\tilde{Z}_n$ where $\tilde{Z}_{2n} = Z_n$ and $\tilde{Z}_{2n+1} = \mathrm{RM}_1(\tilde{Z}_{2n})$. This allows us to "mix-and-match" Propositions 2–4 to prove, for instance, the convergence of alternating (R-SEG) minimizer vs. retraction-based (R-PPM) maximizer in Riemannian two-player games.

Proof of Proposition 4. Similar to Proposition 2, Lemma 1 guarantees that all geodesics are minimizing and invertible. Hence, by the L-Lipschitzness of V and (alt-R-RM), we have, with probability 1,

$$\begin{aligned}\|V_Y(X_{n+1}, Y_n) - V_Y(X_n, Y_n)\|_{Y_n} &\leq L \operatorname{dist}\left(\begin{bmatrix} X_{n+1} \\ Y_n \end{bmatrix}, \begin{bmatrix} X_n \\ Y_n \end{bmatrix}\right) \\ &= L\gamma_n \|V_X(X_n, Y_n) + U_{X,n} + b_{X,n}\|_{X_n} \\ &\to 0 \end{aligned} \tag{49}$$

by Lemma 1 and Proposition 2. Therefore, we may rewrite (alt-R-RM) as

$$Z_{n+1} = \exp_{Z_n}\left(\gamma_n \begin{bmatrix} V_X(X_n, Y_n) + U_{X,n} + b_{X,n} \\ V_Y(X_n, Y_n) + U_{Y,n} + b'_{Y,n} \end{bmatrix}\right)$$

where $b'_{Y,n} = b_{Y,n} + V_Y(X_{n+1}, Y_n) - V_Y(X_n, Y_n)$. By (49) and Proposition 2, $b'_{Y,n}$ satisfies (5) and (7), concluding the proof. ∎

4.3. **Implications.** We collect some immediate consequences of the results in Sections 4.1–4.2.

Riemannian non-convex optimization. When $V(z) = -\nabla f(z)$ for some sufficiently smooth function $f\colon \mathcal{M} \to \mathbb{R}$, Sard's theorem implies that the ICT sets consist of connected components of critical points [63]. Therefore, all the above-mentioned algorithms converge to critical points of f. This is a vast generalization of the main result of Bonnabel [10] which, in the context of this paper, can only cover (retraction-based) Algorithm 1 and bounded (SFO).

Riemannian monotone games. Consider the game setting: $V = [-\nabla_{z_i} \ell_i]$ where $\ell_i : \mathcal{M} = \prod_i \mathcal{M}_i \to \mathbb{R}$ denotes the loss of the i-th player. If V is *monotone* [68] and sufficiently smooth, then combining our results with [35] shows that all the above mentioned algorithms converge to the *Nash–Stampacchia equilibria*, a natural generalization of Nash equilibria to Riemannian manifolds. To the best of our knowledge, most of the algorithms we consider are new in this setting except for the *deterministic* gradient method and extra-gradient mentioned in the Introduction.

Riemannian non-convex potential games. Solving general non-convex games remains an open problem even for Euclidean spaces. However, a special class of games, known as the *Riemannian potential games*, has found interesting applications [33, 51]. For such games, the continuous-time (RMD) has been shown to enjoy many desirable properties such as convergence to critical points or local stability [57–59], whereas we are not aware of any similar result on stochastic RRM schemes. Our theory bridges this gap by showing that the same guarantees in [57–59] are in fact achieved by a wide array of RRM schemes.

Markov potential games via Natural Gradient Descent. An interesting class of games that models collaborative behavior in multi-agent reinforcement learning is the *Markov potential games* (MPGs) [41, 47, 49]. These games can be solved via a variety of learning schemes, notably the *natural policy gradient* (NPG) method of Kakade [30], which has found wide empirical success.

As noticed by Bonnabel [10], NPG, as a particular case of the *Natural Gradient Descent* method [2], is an important instance of retraction-based algorithms. In our framework, NPG is simply a retraction-based Algorithm 1. Since MPGs can be reduced to jointly optimizing a global objective as in Riemannian potential games [47], under mild assumptions on the noise, our theory together with [57–59] then not only implies the convergence of plain NPG, but also its optimistic/extra-gradient variants. To our knowledge, these are new schemes that might be of independent interest.

5. Concluding remarks

Motivated by applications to stochastic approximation and optimization problems on manifolds, our paper proposed a general class of Riemannian iterative algorithms à la Robbins-Monro, and we showed that it captures a wide array of existing and novel algorithms. Our theory provides a unified analysis for the convergence of RRM schemes that might seem vastly different from each other at first sight: By verifying certain simple criteria on the error terms W_n in Assumption 2, we can instead study (RMD) to infer the algorithm's long-run behavior.

At the same time, we should stress that our results offer but a glimpse of the flexibility of (RRM). Specifically, the following important questions are left open (and deferred to future work):

(1) We have only analyzed stochastic *first-order* methods. In practice, however, evaluating Riemannian (stochastic) gradients can be out of reach in online / sequential learning problems. To remedy this issue, various *zeroth-order* or *bandit* algorithms, which only require function evaluations, have been proposed. It is, however, unclear whether a Riemannian analogue of the *Kiefer–Wolfowitz* algorithm can analyzed viea the Robbins–Monro template as in the Euclidean case [26].

(2) The assumption on the *diminishing* step-sizes is indispensable to our analysis, which covers many practically relevant settings. However, another common strategy in practice is the *constant* step-size rule, which is not covered by our theory. It is thus interesting to see if the techniques for analyzing constant step-size SA schemes can be generalized to Riemannian manifolds [11, 38, 39].

(3) Finally, several Riemannian algorithms have been designed to avoid undesirable solutions such as non-minimizing critical points [16, 64]. We conjecture that the avoidance theory of Benaïm and Hirsch [6] can be extended to Riemannian manifolds; this, if true, would imply that many iterative Riemannian methods, including the retraction-based ones, converge *only* to local minimizers ((a.s.)).

Appendix A. Stability Analysis

The purpose of this appendix is to prove:

Proposition 1. *If $\mathcal{M}$ is simply connected, Assumption 3′ implies Assumption 3.*

Fix an arbitrary (base) point $p \in \mathcal{M}$ playing the role of the origin, and define the radial distance $r(z) := \operatorname{dist}(z, p)$ and let $k(z) := \frac{1}{2} r^2(z)$. The following theorem makes it clear under which assumptions is $k(\cdot)$ smooth and provides a control on its Hessian.

Theorem 1 (Jost [29, Theorem 6.6.1]). *Let $p \in \mathcal{M}$ be arbitrary, and suppose that the exponential map $\exp_p$ is a diffeomorphism on $\{v \in T_p\mathcal{M} : \|v\| \le \rho\}$. Moreover, suppose the sectional curvature is nonpositive and bounded below by $-\kappa^2$ on the ball of radius ρ around p.*

Let $r(\cdot)$ and $k(\cdot)$ be defined as above. Then $k(\cdot)$ is smooth on $\mathbb{B}(p,\rho) \coloneqq \{z \neq p : \operatorname{dist}(z,p) \le \rho\}$ and

$$\nabla k(z) = -\exp_p^{-1} z.$$

In addition, $\|\nabla k(z)\| = r(z)$, and one has the following control on Hessian of $k(\cdot)$:

$$\operatorname{Hess} k(z)[v,v] \le \kappa r(z) \operatorname{ctgh}(\kappa r(z)) \|v\|_z^2,$$

for all $z \in \mathbb{B}(p,\rho)$ and $v \in T_z\mathcal{M}$.

Since $\mathcal{M}$ is simply-connected and complete, we may take $\rho = \infty$ in the theorem above [29, Corollary 6.9.1].

Recall that the vector field V satisfies the weak-coercivity condition if for some $R > 0$ and some $p \in \mathcal{M}$ (and hence all $p \in \mathcal{M}$) we have for all $z \in \mathcal{M}$ outside of $\mathbb{B}(p,R)$,

$$\langle V(z), \nabla k(z) \rangle_z \le 0.$$

Our proof relies on constructing a suitable "energy function" that serves as an easy-to-control proxy for the distance of the iterates of (RRM) from the origin. This function will be of the form

$$\Phi(z) = f(r(z))$$

where f is a $\mathcal{C}^\infty$ non-negative function with $f(x) = 0$ for all $x \le R$ and satisfies

$$0 \le f'(x) \le C_1, \quad f''(x) \le C_2 \tag{A.1}$$

for all $x \ge R$. Moreover, we require $f(x) = \Omega(x)$ as $x \to \infty$ so that controlling f implies control of x. One example of such functions is given in Lemma A.2.

We first show that Φ has a bounded Hessian and is smooth.

Lemma A.1. *Let Φ be defined as above. Then, Φ is negatively correlated with V:*

$$\langle \nabla \Phi(z), V(z) \rangle_z \le 0, \quad \forall z \in \mathcal{M},$$

and there is some constant C such that $\operatorname{Hess} \Phi(z)[v,v] \le C\|v\|_z^2$. Moreover, Φ is C-smooth, that is,

$$\Phi(z_2) \le \Phi(z_1) + \langle \nabla \Phi(z_1), \exp_{z_1}^{-1} z_2 \rangle + \frac{C}{2} \operatorname{dist}^2(z_1, z_2).$$

Proof. We first compute the gradient of Φ:

$$\nabla \Phi(z) = \begin{cases} 0 & r(z) \le R, \\ \frac{f'(r(z))}{r(z)} \nabla k(z) & r(z) > R. \end{cases}$$

By assumption, $f'(r(z))/r(z) \ge 0$, and hence, $\langle \nabla \Phi(z), V(z) \rangle_z$ has the same sign as $\langle \nabla k(z), V(z) \rangle_z$ if $r(z) > R$, and is zero if $r(z) \le R$. Thus, Φ and V are negatively correlated.

To compute the Hessian of Φ, notice that $\operatorname{Hess} \Phi(z)[v,v] = \langle \nabla_v \nabla \Phi(z), v \rangle_z$. Hence,

$$\begin{aligned} \operatorname{Hess} \Phi(z)[v,v] &= \nabla_v \frac{f'(r(z))}{r(z)} \cdot \langle \nabla k(z), v \rangle_z + \frac{f'(r(z))}{r(z)} \langle \nabla_v \nabla k(z), v \rangle_z \\ &= \underbrace{\left\langle \nabla \frac{f'(r(z))}{r(z)}, v \right\rangle \cdot \langle \nabla k(z), v \rangle_z}_{\text{ⓐ}} + \underbrace{\frac{f'(r(z))}{r(z)} \operatorname{Hess} k(z)[v,v]}_{\text{ⓑ}}. \end{aligned}$$

Here we use the same notation for directional derivative of a scalar function and the covariant derivative. We first compute ⓐ. We have

$$\nabla \frac{f'(r(z))}{r(z)} = \left(f''(r(z)) - \frac{f'(r(z))}{r(z)} \right) \frac{1}{r^2(z)} \nabla k(z),$$

hence,

$$\begin{aligned} ⓐ &= \left(f''(r(z)) - \frac{f'(r(z))}{r(z)} \right) \frac{1}{r^2(z)} \langle \nabla k(z), v \rangle_z^2 \\ &\leq \frac{C_2}{r^2(z)} \|\nabla k(z)\|_z^2 \|v\|_z^2 \\ &= C_2 \|v\|_z^2. \end{aligned}$$

For ⓑ, as $x \operatorname{ctgh} x \leq 1 + x$ for $x \geq 0$, we obtain

$$ⓑ \leq \frac{f'(r(z))}{r(z)} (1 + \kappa r(z)) \|v\|_z^2 \leq C_1 \left(\tfrac{1}{R} + \kappa \right) \|v\|_z^2.$$

Summing up everything, we obtain

$$\operatorname{Hess} \Phi(z)[v, v] \leq (C_2 + C_1/R + C_1\kappa) \|v\|_z^2 =: C \|v\|_z^2,$$

that is, Φ has bounded Hessian. Moreover, Φ is smooth as a composition of smooth functions. Let $z_1, z_2 \in \mathcal{M}$ be arbitrary, and let $\gamma : [0,1] \to \mathcal{M}$ be a geodesic connecting the two. By Taylor's remainder theorem, there exists some $t \in (0,1)$ such that

$$\Phi(z_2) = \Phi(z_1) + \langle \nabla \Phi(z_1), \dot{\gamma}(0) \rangle_{z_1} + \frac{1}{2} \operatorname{Hess} \Phi(\gamma(t))[\dot{\gamma}, \dot{\gamma}].$$

Thus, by the Hessian upper bound, and noticing that $\|\dot{\gamma}\| = \operatorname{dist}(z_1, z_2)$ and $\dot{\gamma}(0) = \exp_{z_1}^{-1} z_2$, we get

$$\Phi(z_2) \leq \Phi(z_1) + \langle \nabla \Phi(z_1), \exp_{z_1}^{-1} z_2 \rangle + \frac{C}{2} \operatorname{dist}^2(z_1, z_2),$$

as desired. ∎

We now proceed to the main argument, where we show how to use Φ to control the iterates Z_n. Letting $\Phi_n = \Phi(Z_n)$ and using Lemma A.1 we get

$$\begin{aligned} \Phi_{n+1} &= \Phi \left(\exp_{Z_n}(\gamma_n \hat{V}_n) \right) \\ &\leq \Phi_n + \gamma_n \langle \nabla \Phi(Z_n), \hat{V}_n \rangle_{Z_n} + \frac{C\gamma_n^2}{2} \|\hat{V}_n\|_{Z_n}^2 \\ &\leq \Phi_n + \gamma_n \langle \nabla \Phi(Z_n), U_n + b_n \rangle_{Z_n} + \frac{3C\gamma_n^2}{2} \left[\|V(Z_n)\|_{Z_n}^2 + \|U_n\|_{Z_n}^2 + \|b_n\|_{Z_n}^2 \right] \end{aligned}$$

where the second line follows from negative correlation of Φ and V, the definition (1) of $\hat{V}_n$, and the Cauchy-Schwartz inequality. Conditioning on $\mathcal{F}_n$ and taking expectations, and invoking Cauchy-Schwartz and the fact that $\|\nabla \Phi(Z_n)\| \leq \frac{C_1}{r(Z_n)} \|\nabla k(Z_n)\| = C_1$, we obtain:

$$\mathbb{E}[\Phi_{n+1} \,|\, \mathcal{F}_n] \leq \Phi_n + \gamma_n C_1 B_n + \tfrac{3}{2} C \gamma_n^2 \left[G^2 + B_n^2 + \sigma_n^2 \right], \tag{A.2}$$

where we have bounded the second moments by their respective upper bounds.

To proceed, let $\varepsilon_n = \gamma_n C_1 B_n + (3/2) C \gamma_n^2 \left[G^2 + B_n^2 + \sigma_n^2 \right]$ denote the "residual" term in (A.2). Notice that

$$\sum_{n=1}^{\infty} \varepsilon_n \leq C_1 \sum_{n=1}^{\infty} \gamma_n B_n + \frac{3C}{2} \sum_{n=1}^{\infty} \gamma_n^2 (G^2 + B_n^2 + \sigma_n^2),$$

and hence, by (7) and dominated convergence theorem, we have that $\mathbb{E}[\sum_n \varepsilon_n] < \infty$. Next, consider the auxiliary process $\hat{\Phi}_n = \Phi_n + \mathbb{E}[\sum_{k=n}^{\infty} \varepsilon_k \,|\, \mathcal{F}_n]$, adapted to the same filtration. By (A.2), we have $\mathbb{E}[\hat{\Phi}_{n+1} \,|\, \mathcal{F}_n] \le \Phi_n + \mathbb{E}[\sum_{k=n}^{\infty} \varepsilon_k \,|\, \mathcal{F}_n] = \hat{\Phi}_{n-1}$, i.e., $\hat{\Phi}_n$ is a supermartingale with respect to $\mathcal{F}_n$. This shows that $\mathbb{E}[\hat{\Phi}_n] \le \mathbb{E}[\hat{\Phi}_1] < \infty$, i.e., $\hat{\Phi}_n$ is uniformly bounded in L^1. Hence, by Doob's supermartingale convergence theorem [25, Theorem 2.5], it follows that $\hat{\Phi}_n$ converges with probability 1 to some finite random limit $\hat{\Phi}_\infty$. In turn, since $\sum_n \varepsilon_n < \infty$ (a.s.), this implies that $\Phi_n = \hat{\Phi}_n - \mathbb{E}[\sum_{k=n}^{\infty} \varepsilon_n \,|\, \mathcal{F}_n]$ also converges to some (random) finite limit (a.s.). From this and the fact that $\Phi_n = \Omega(r(Z_n))$, we deduce $\limsup_n r(Z_n) < \infty$ as claimed.

Lemma A.2. *Let $h\colon \mathbb{R} \to \mathbb{R}$ be the function*

$$h(x) = \begin{cases} 0 & x \le 0 \\ \frac{e^{-1/x}}{e^{-1/x} + e^{-1/(1-x)}} & x \in (0,1) \\ 1 & x \ge 1 \end{cases}.$$

It is easy to see that h is C^∞. For $R > 0$ define

$$f(x) = \int_0^x h(s - R)\, ds.$$

Then f is C^∞, satisfies the conditions (A.1) with $C_1 = 1$ and $C_2 = 2$. In addition, one has $f(x) \ge x - (R+1)$, and hence $f(x) = \Omega(x)$.

Proof. As $h(x) \in [0,1]$, we obtain that $f'(x) \in [0,1]$. By straight-forward computation, one observes that h has bounded first derivative $0 \le h'(x) \le 2$. Hence,

$$f''(x) = h'(x - R) \le 2.$$

Notice that for $x \ge R+1$, $f(x) = \int_0^x h(s-R)\, ds \ge \int_{R+1}^x 1\, ds = x - (R+1)$. ∎

Appendix B. Preliminaries for the Proof of Theorem 1

In this appendix, we collect the necessary tools for the proof of Theorem 1.

B.1. **The parallel frame system.** Fix any two points $z, z' \in \mathcal{M}$, and consider two arbitrary vectors $v \in \mathcal{T}_z\mathcal{M}$ and $v' \in \mathcal{T}_{z'}\mathcal{M}$. There is a convenient frame system (i.e., a set of bases for $\mathcal{T}_z\mathcal{M}$ and $\mathcal{T}_{z'}\mathcal{M}$) for comparing v and v', defined as follows: Pick an arbitrary orthonormal frame $\{e_k\}_{k=1}^d$ for $\mathcal{T}_z\mathcal{M}$. Since the parallel transport map is an isometry, the vectors $\{e'_k\}_{k=1}^d := \{\Gamma_{z\to z'}(e_k)\}_{k=1}^d$ form an orthonormal basis for $\mathcal{T}_{z'}\mathcal{M}$. Consider the *component vectors* of v, v' in these two frames:

$$v_\parallel \in \mathbb{R}^d,\ v_\parallel^k := \langle v, e_k\rangle_z; \quad v'_\parallel \in \mathbb{R}^d,\ v_\parallel'^k := \langle v', e'_k\rangle_{z'}. \tag{B.1}$$

We shall call (B.1) the *parallel frame system* for vectors v and v' (the dependence on the initial frame $\{e_k\}_{k=1}^d$ is suppressed).

By virtue of parallel transport, in the parallel frame system we have

$$\begin{aligned} \Gamma_{z\to z'}(v)_\parallel^k &= \langle \Gamma_{z\to z'}(v), e'_k\rangle_{z'} \\ &= \langle v, \Gamma_{z'\to z}(e'_k)\rangle_z \\ &= v_\parallel^k. \end{aligned} \tag{B.2}$$

In the same vein, we have $\Gamma_{z'\to z}(v')_{\shortparallel}^k = v_{\shortparallel}'^k$. Thus, since $\{e_k\}_{k=1}^d$ is orthonormal, we may write:

$$\begin{aligned}\|v' - \Gamma_{z\to z'}(v)\|_z &= \|v - \Gamma_{z'\to z}(v')\|_{z'} \\ &= \|v_{\shortparallel} - v_{\shortparallel}'\|_2.\end{aligned}$$

In other words, in the parallel frame system, the comparison of vectors living on different tangent spaces is reduced to simply comparing the Euclidean norms of their components. For instance, the L-Lipschitzness of V can be rephrased as:

$$\forall z, z' \in \mathcal{M}, \quad \|V_{\shortparallel}(z') - V_{\shortparallel}(z)\|_2 \le L\,\mathrm{dist}(z, z'). \tag{B.3}$$

B.2. **The Fermi coordinate system.** For any h, let $\mathcal{U}_h \subset \mathcal{T}_{Z(t+h)}\mathcal{M} \simeq \mathbb{R}^d$ be a neighborhood of 0 on which the mapping

$$\exp_{Z(t+h)}(\cdot) : \mathcal{U}_h \to \mathcal{M} \tag{B.4}$$

is a diffeomorphism between $\mathcal{U}_h$ and $\exp_{Z(t+h)}(\mathcal{U}_h)$. It is well-known that such a neighborhood exists, and that the exponential map $\exp_{Z(t+h)}$ along with an arbitrary orthonormal frame at $\mathcal{T}_{Z(t+h)}\mathcal{M}$ induces a local coordinate system on $\exp_{Z(t+h)}(\mathcal{U}_h)$, called the *normal coordinate with center* $Z(t+h)$ [40]. Normal coordinates are best suitable for comparing *distances* of points on manifolds. For instance, if $\tilde{z}'$ is the normal coordinate of z' with center z, then $\mathrm{dist}(z, z') = \|\tilde{z}'\|_2$.

The *Fermi coordinate* [48], roughly speaking, is a system of normal coordinates "along a curve". To define it, fix an arbitrary orthonormal frame $\{e_k(0)\}_{k=1}^d$ for $\mathcal{T}_{Z(t)}\mathcal{M}$. We can obtain a system of orthonormal frames $\{e_k(h)\}_{k=1}^d$ by parallel transporting $\{e_k(0)\}_{k=1}^d$ from $\mathcal{T}_{Z(t)}\mathcal{M}$ to $\mathcal{T}_{Z(t+h)}\mathcal{M}$ along the curve $h \mapsto Z(t+h)$. Let $\mathcal{U}_h \subset \mathcal{T}_{Z(t+h)}\mathcal{M}$ be a neighborhood of 0 defined as in (B.4), and set $\mathcal{V} \coloneqq \bigcup_h \{\exp_{Z(t+h)}(\mathcal{U}_h)\} \subset \mathcal{M}$. Consider the mapping

$$\tilde{\Phi} : \mathbb{R}^+ \times \mathcal{V} \to \mathbb{R}^+ \times \mathbb{R}^d \tag{B.5}$$

by sending a point $(h, z) \in \mathbb{R}^+ \times \mathcal{V}$ to $(h, \tilde{z}) \in \mathbb{R}^+ \times \mathbb{R}^d$, where $\tilde{z}$ is the normal coordinate of z with center $Z(t+h)$ and frame $\{e_k(h)\}_{k=1}^d$. By virtue of the normal coordinate, we know that $\tilde{\Phi}$ is a diffeomorphism between $\mathbb{R}^+ \times \mathcal{V}$ and a neighborhood of $\mathbb{R}^+ \times \{0\}$. The mapping $\tilde{\Phi}$ and its inverse is called the *Fermi coordinate system* along the curve $h \mapsto Z(t+h)$. In the sequel, we will abuse the notation and simply call it the Fermi coordinate along $Z(\cdot)$.

The following property of the Fermi coordinate system plays a key role in our analysis:

Lemma B.1 ([24, 65])**.** *Let γ be a differentiable curve on $\mathcal{M}$ such that $\gamma(h) \in \mathcal{U}_h$ for all $h \in \mathbb{R}^+$, and let $\tilde{\gamma}$ be the curve of γ in the Fermi coordinate system along $Z(\cdot)$ (i.e., $(h, \tilde{\gamma}(h)) = \tilde{\Phi}(h, \gamma(h))$. Then*

$$\dot{\tilde{\gamma}}^k(h) = \dot{\gamma}^k(h) - \dot{Z}^k(t+h) + \mathcal{O}\left(\|\tilde{\gamma}(h)\|_2^2\right).$$

Here, $\dot{Z}^k(t+h) \coloneqq \langle \dot{Z}(t+h), e_k(h)\rangle_{Z(t+h)}$ is the k-th component of $\dot{Z}(t+h)$ in the frame $\{e_k(h)\}_{k=1}^d$, and $\dot{\gamma}^k(h)$ is the k-th component of $\dot{\gamma}(h)$ in the (possibly non-orthonormal) frame induced by the normal coordinate with center $Z(t+h)$ and frame $\{e_k(h)\}_{k=1}^d$.

B.3. **Comparing the differential of** exp **and parallel transport.** As will become clear in the proof, the parallel frame system is convenient for comparing *vectors* at different points, whereas the Fermi coordinate system is best suitable for comparing the *distance* between curves, both features being essential to our proof. There is, however, a dichotomy: It is known that if the Fermi coordinate in (B.5) is simultaneously a parallel frame system for all points nearby the curve $Z(\cdot)$, then the underlying manifold $\mathcal{M}$ must be flat; i.e., the Riemannian curvature tensor vanishes everywhere [28].

Therefore, we would like to work with parallel frame and Fermi coordinate systems separately, and compare the difference between the two whenever needed. To this end, we will need the following technical lemma, whose proof can be found in [42, Theorem 3.12] or [17, Proposition A.1]:

Lemma B.2 (Comparing $\mathrm{d}\exp$ and parallel transport)**.** *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\mathrm{low}}, K_{\mathrm{up}}]$, and let $K = \max(|K_{\mathrm{low}}|, |K_{\mathrm{up}}|)$. For $v \in \mathcal{T}_z\mathcal{M}$, consider the geodesic $\gamma(t) = \exp_z(tv)$. If γ is defined and has no interior conjugate point on the interval $[0,1]$, then*

$$\forall v' \in \mathcal{T}_z\mathcal{M}, \qquad \|T_v(v') - \Gamma_v(v')\|_{\gamma(1)} \leq K \cdot f_{K_{\mathrm{low}}}(\|v\|_z) \cdot \|v'_\perp\|_z \tag{B.6}$$

where $v'_\perp \coloneqq v' - \frac{\langle v, v'\rangle_z}{\langle v, v\rangle_z} v$ is the component of v' orthogonal to v, $T_v = \mathrm{d}\exp_z(v)$ is the differential of the exponential map, and Γ_v denotes the parallel transport along γ from $\gamma(0)$ to $\gamma(1)$. The function $f_{K_{\mathrm{low}}}$ in (B.6) is defined as

$$f_{K_{\mathrm{low}}}(a) = \begin{cases} \frac{r^2}{6} & , \ K_{\mathrm{low}} = 0 \\ r^2\left(1 - \frac{\sin(a/r)}{a/r}\right) & , \ K_{\mathrm{low}} = \frac{1}{r^2} > 0 \\ r^2\left(\frac{\sinh(a/r)}{a/r} - 1\right) & , \ K_{\mathrm{low}} = -\frac{1}{r^2} < 0 \end{cases} . \tag{B.7}$$

Moreover, the function $f_{K_{\mathrm{low}}}$ is dominated by the case $K_{\mathrm{low}} < 0$: For all $K \geq |K_{\mathrm{low}}|$ and $a \in \mathbb{R}^+$,

$$f_{K_{\mathrm{low}}}(a) \leq f_{-K}(a). \tag{B.8}$$

Acknowledgments

This research was supported by the SNSF grant 407540_167212 through the NRP 75 Big Data program, and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant aggreement No 815943. YPH acknowledges funding through an ETH Foundations of Data Science (ETH-FDS) postdoctoral fellowship. Part of this work was done while PM was visiting the Simons Institute for the Theory of Computing. PM is also grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the grant ALIAS (ANR-19-CE48-0018-01).

References

[1] Absil PA, Mahony R, Sepulchre R (2009) Optimization algorithms on matrix manifolds. Princeton University Press

[2] Amari SI (1998) Natural gradient works efficiently in learning. Neural computation 10(2):251–276

[3] Ambrosio L, Gigli N, Savaré G (2005) Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media

[4] Bauschke HH, Combettes PL (2017) Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, New York, NY, USA

[5] Benaïm M (1999) Dynamics of stochastic approximation algorithms. In: Azéma J, Émery M, Ledoux M, Yor M (eds) Séminaire de Probabilités XXXIII, Lecture Notes in Mathematics, vol 1709, Springer Berlin Heidelberg, pp 1–68

[6] Benaïm M, Hirsch MW (1995) Dynamics of Morse-Smale urn processes. Ergodic Theory and Dynamical Systems 15(6):1005–1030

[7] Benaïm M, Hirsch MW (1996) Asymptotic pseudotrajectories and chain recurrent flows, with applications. Journal of Dynamics and Differential Equations 8(1):141–176

[8] Bento GC, Ferreira OP, Melo JG (2017) Iteration-complexity of gradient, subgradient and proximal point methods on Riemannnian manifolds. Journal of Optimization Theory and Applications 173(2):548–562

[9] Benveniste A, Métivier M, Priouret P (1990) Adaptive Algorithms and Stochastic Approximations. Springer

[10] Bonnabel S (2013) Stochastic gradient descent on Riemannnian manifolds. IEEE Transactions on Automatic Control 58(9):2217–2229

[11] Borkar VS (2008) Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press and Hindustan Book Agency

[12] Boumal N (2020) An introduction to optimization on smooth manifolds. Available online, Aug

[13] Boumal N, Absil PA, Cartis C (2019) Global rates of convergence for nonconvex optimization on manifolds. IMA Journal of Numerical Analysis 39(1):1–33

[14] Bridson MR, Haefliger A (2013) Metric spaces of non-positive curvature, vol 319. Springer Science & Business Media

[15] Chen J, Liu S, Chang X (2021) Modified tseng's extragradient methods for variational inequality on Hadamard manifolds. Applicable Analysis 100(12):2627–2640

[16] Criscitiello C, Boumal N (2019) Efficiently escaping saddle points on manifolds. Advances in Neural Information Processing Systems 32:5987–5997

[17] Criscitiello C, Boumal N (2020) An accelerated first-order method for non-convex optimization on manifolds. arXiv preprint arXiv:200802252

[18] Durmus A, Jiménez P, Moulines É, Said S, Wai HT (2020) Convergence analysis of Riemannnian stochastic approximation schemes. arXiv preprint arXiv:200513284

[19] Durmus A, Jiménez P, Moulines É, Said S (2021) On Riemannnian stochastic approximation schemes with fixed step-size. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 1018–1026

[20] Facchinei F, Pang JS (2003) Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer Series in Operations Research, Springer

[21] Fan J, Qin X, Tan B (2020) Tseng's extragradient algorithm for pseudomonotone variational inequalities on Hadamard manifolds. Applicable Analysis pp 1–14

[22] Ferreira O, Oliveira P (2002) Proximal point algorithm on Riemannnian manifolds. Optimization 51(2):257–270

[23] Ferreira OP, Pérez LL, Németh SZ (2005) Singularities of monotone vector fields and an extragradient-type algorithm. Journal of Global Optimization 31(1):133–151

[24] Fujita T, Kotani Si (1982) The onsager-machlup function for diffusion processes. Journal of Mathematics of Kyoto University 22(1):115–130

[25] Hall P, Heyde CC (1980) Martingale Limit Theory and Its Application. Probability and Mathematical Statistics, Academic Press, New York

[26] Hsieh YP, Mertikopoulos P, Cevher V (2021) The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In: International Conference on Machine Learning, PMLR, pp 4337–4348

[27] Huang W, Wei K (2021) Riemannnian proximal gradient methods. Mathematical Programming pp 1–43

[28] Iliev BZ (2006) Handbook of normal frames and coordinates, vol 42. Springer Science & Business Media

[29] Jost J (2017) Riemannian geometry and geometric analysis, 7th edn. Springer, DOI https://doi.org/10.1007/978-3-319-61860-9

[30] Kakade SM (2001) A natural policy gradient. Advances in neural information processing systems 14

[31] Khammahawong K, Kumam P, Chaipunya P, Yao JC, Wen CF, Jirakitpuwapat W (2020) An extragradient algorithm for strongly pseudomonotone equilibrium problems on Hadamard manifolds. Thai Journal of Mathematics 18(1):350–371

[32] Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23(3):462–466

[33] Klavins E, Koditschek DE (2002) Phase regulation of decentralized cyclic robotic systems. The International Journal of Robotics Research 21(3):257–275

[34] Korpelevich GM (1976) The extragradient method for finding saddle points and other problems. Èkonom i Mat Metody 12:747–756

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

[36] Kushner H, Yin GG (2003) Stochastic approximation and recursive algorithms and applications, vol 35. Springer Science & Business Media

[37] Kushner HJ, Clark DS (1978) Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer

[38] Kushner HJ, Huang H (1981) Asymptotic properties of stochastic approximations with constant coefficients. SIAM Journal on Control and Optimization 19(1):87–105

[39] Kushner HJ, Yin GG (1997) Stochastic approximation algorithms and applications. Springer-Verlag, New York, NY

[40] Lee JM (2006) Riemannnian manifolds: an introduction to curvature, vol 176. Springer Science & Business Media

[41] Leonardos S, Overman W, Panageas I, Piliouras G (2021) Global convergence of multi-agent policy gradient in markov potential games. arXiv preprint arXiv:210601969

[42] Lezcano-Casado M (2020) Curvature-dependant global convergence rates for optimization on manifolds of bounded geometry. arXiv preprint arXiv:200802517

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

[44] Lin Z (2019) Riemannnian geometry of symmetric positive definite matrices via cholesky decomposition. SIAM Journal on Matrix Analysis and Applications 40(4):1353–1370

[45] Ljung L (1977) Analysis of recursive stochastic algorithms 22(4):551–575

[46] Ljung L (1978) Strong convergence of a stochastic approximation algorithm. Annals of Statistics 6(3):680–696

[47] Macua SV, Zazo J, Zazo S (2018) Learning parametric closed-loop policies for markov potential games. arXiv preprint arXiv:180200899

[48] Manasse F, Misner CW (1963) Fermi normal coordinates and some basic concepts in differential geometry. Journal of mathematical physics 4(6):735–745

[49] Marden JR (2012) State based potential games. Automatica 48(12):3075–3088

[50] Massart E, Hendrickx JM, Absil PA (2019) Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the bures–wasserstein metric. In: International Conference on Geometric Science of Information, Springer, pp 739–748

[51] Muhammad A, Egerstedt M (2005) Decentralized coordination with local interactions: Some new directions. In: Cooperative Control, Springer, pp 153–170

[52] Nesterov Y (2004) Introductory Lectures on Convex Optimization: A Basic Course. No. 87 in Applied Optimization, Kluwer Academic Publishers

[53] Neto JC, Santos P, Soares P (2016) An extragradient method for equilibrium problems on Hadamard manifolds. Optimization Letters 10(6):1327–1336

[54] Nickel M, Kiela D (2017) Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems 30:6338–6347

[55] Phelps RR (1993) Convex Functions, Monotone Operators and Differentiability, 2nd edn. Lecture Notes in Mathematics, Springer-Verlag

[56] Popov LD (1980) A modification of the Arrow–Hurwicz method for search of saddle points. Mathematical Notes of the Academy of Sciences of the USSR 28(5):845–848

[57] Ratliff LJ, Burden SA, Sastry SS (2013) Characterization and computation of local nash equilibria in continuous games. In: 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, pp 917–924

[58] Ratliff LJ, Burden SA, Sastry SS (2014) Genericity and structural stability of non-degenerate differential nash equilibria. In: 2014 American Control Conference, IEEE, pp 3990–3995

[59] Ratliff LJ, Burden SA, Sastry SS (2016) On the characterization of local nash equilibria in continuous games. IEEE transactions on automatic control 61(8):2301–2307

[60] Robbins H, Monro S (1951) A stochastic approximation method. Annals of Mathematical Statistics 22:400–407

[61] Shah SM (2021) Stochastic approximation on riemannian manifolds. Applied Mathematics & Optimization 83(2):1123–1151

[62] Sra S, Hosseini R (2015) Conic geometric optimization on the manifold of positive definite matrices. SIAM Journal on Optimization 25(1):713–739

[63] Sternberg S (1999) Lectures on differential geometry, vol 316. American Mathematical Soc.

[64] Sun Y, Flammarion N, Fazel M (2019) Escaping from saddle points on riemannian manifolds. Advances in Neural Information Processing Systems 32

[65] Takahashi Y, Watanabe S (1981) The probability functionals (onsager-machlup functions) of diffusion processes. In: Stochastic Integrals, Springer, pp 433–463

[66] Tang Gj, Huang Nj (2012) Korpelevich's method for variational inequality problems on Hadamard manifolds. Journal of Global Optimization 54(3):493–509

[67] Tripuraneni N, Flammarion N, Bach F, Jordan MI (2018) Averaging stochastic gradient descent on Riemannnian manifolds. In: Conference On Learning Theory, PMLR, pp 650–687

[68] Wang J, López G, Martín-Márquez V, Li C (2010) Monotone and accretive vector fields on riemannian manifolds. Journal of optimization theory and applications 146(3):691–708

[69] Wang X, Tu Z, Hong Y, Wu Y, Shi G (2021) No-regret online learning over Riemannnian manifolds. In: Thirty-Fifth Conference on Neural Information Processing Systems

[70] Warner FW (1965) The conjugate locus of a riemannian manifold. American Journal of Mathematics 87(3):575–604

[71] Wong YC (1967) Differential geometry of grassmann manifolds. Proceedings of the National Academy of Sciences of the United States of America 57(3):589

[72] Wong YC (1968) Sectional curvatures of grassmann manifolds. Proceedings of the National Academy of Sciences of the United States of America 60(1):75

[73] Zhang H, Sra S (2016) First-order methods for geodesically convex optimization. In: Conference on Learning Theory, PMLR, pp 1617–1638