First-Order Algorithms for Min-Max Optimization in Geodesic Metric Spaces

Michael I. Jordan **Tianyi Lin** **Emmanouil V. Vlatakis-Gkaragkounis**
University of California, Berkeley
`{jordan@cs,darren_lin@,emvlatakis@}.berkeley.edu`

Abstract

From optimal transport to robust dimensionality reduction, a plethora of machine learning applications can be cast into the min-max optimization problems over Riemannian manifolds. Though many min-max algorithms have been analyzed in the Euclidean setting, it has proved elusive to translate these results to the Riemannian case. Zhang et al. have recently shown that geodesic convex concave Riemannian problems always admit saddle-point solutions. Inspired by this result, we study whether a performance gap between Riemannian and optimal Euclidean space convex-concave algorithms is necessary. We answer this question in the negative—we prove that the Riemannian corrected extragradient (RCEG) method achieves last-iterate convergence at a linear rate in the geodesically strongly-convex-concave case, matching the Euclidean result. Our results also extend to the stochastic or non-smooth case where RCEG and Riemanian gradient ascent descent (RGDA) achieve near-optimal convergence rates up to factors depending on curvature of the manifold.

1 Introduction

Constrained optimization problems arise throughout machine learning, in classical settings such as dimension reduction [2], dictionary learning [3, 4], and deep neural networks [5], but also in emerging problems involving decision-making and multi-agent interactions. While simple convex constraints (such as norm constraints) can be easily incorporated in standard optimization formulations, notably (proximal) gradient descent [6–10], in a range of other applications such as matrix recovery [11, 12], low-rank matrix factorization [13] and generative adversarial nets [14], the constraints are fundamentally nonconvex and are often treated via special heuristics.

Thus, a general goal is to design algorithms that systematically take account of special geometric structure of the feasible set [15–17]. A long line of work in the machine learning (ML) community has focused on understanding the geometric properties of commonly used constraints and how they affect optimization; [see, e.g., 18–26]. A prominent aspect of this agenda has been the re-expression of these constraints through the lens of Riemannian manifolds. This has given rise to new algorithms [27, 28] with a wide range of ML applications, inclduing online principal component analysis (PCA), the computation of Mahalanobis distance from noisy measurements [29], consensus distributed algorithms for aggregation in ad-hoc wireless networks [30] and maximum likelihood estimation for certain non-Gaussian (heavy- or light-tailed) distributions [31].

Going beyond simple minimization problems, the robustification of many ML tasks can be formulated as min-max optimization problems. Well-known examples in this domain include adversarial machine learning [32, 33], optimal transport [34], and online learning [9, 35, 36]. Similar to their minimization counterparts, non-convex constraints have been widely applicable to the min-max optimization as well [37–41]. Recently there has been significant effort in proving tighter results either under more structured assumptions [42–52], and/or obtaining last-iterate convergence guarantees [38, 40, 48, 53–

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

62] for computing min-max solutions in convex-concave settings. Nonetheless, the analysis of the iteration complexity in the general *non-convex non-concave* setting is still in its infancy [63, 64]. In response, the optimization community has recently studied how to extend standard min-max optimization algorithms such as gradient descent ascent (GDA) and extragradient (EG) to the Riemannian setting. In mathematical terms, given two Riemannian manifolds $\mathcal{M}, \mathcal{N}$ and a function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$, the Riemannian min-max optimization (RMMO) problem becomes

$$\min_{x \in \mathcal{M}} \max_{y \in \mathcal{N}} f(x, y).$$

The change of geometry from Euclidean to Riemannian poses several difficulties. Indeed, a fundamental stumbling block has been that this problem may not even have theoretically meaningful solutions. In contrast with minimization where an optimal solution in a bounded domain is always guaranteed [65], existence of such saddle points necessitates typically the application of topological fixed point theorems [66, 67], KKM Theory [68]). For the case of convex-concave f with compact sets $\mathcal{X}$ and $\mathcal{Y}$, Sion [69] generalized the celebrated theorem [70] and guaranteed that a solution $(x^\star, y^\star)$ with the following property exists

$$\min_{x \in \mathcal{X}} f(x, y^\star) = f(x^\star, y^\star) = \max_{y \in \mathcal{Y}} f(x^\star, y).$$

However, at the core of the proof of this result is an ingenuous application of Helly's lemma [71] for the sublevel sets of f, and, until the work of Ivanov [72], it has been unclear how to formulate an analogous lemma for the Riemannian geometry. As a result, until recently have extensions of the min-max theorem been established, and only for restricted manifold families [73–75].

Zhang et al. [1] was the first to establish a min-max theorem for a flurry of Riemannian manifolds equipped with unique geodesics. Notice that this family is not a mathematical artifact since it encompasses many practical applications of RMMO, including Hadamard and Stiefel ones used in PCA [76]. Intuitively, the unique geodesic between two points of a manifold is the analogue of the a linear segment between two points in convex set: For any two points $x_1, x_2 \in \mathcal{X}$, their connecting geodesic is the unique shortest path contained in $\mathcal{X}$ that connects them.

Even when the RMMO is well defined, transferring the guarantees of traditional min-max optimization algorithms like Gradient Ascent Descent (GDA) and Extra-Gradient (EG) to the Riemannian case is non-trivial. Intuitively speaking, in the Euclidean realm the main leitmotif of the last-iterate analyses the aforementioned algorithms is a proof that $\delta_t = \|x_t - x^*\|^2$ is decreasing over time. To achieve this, typically the proof correlates δ_t and δ_{t-1} via a "square expansion," namely:

$$\underbrace{\|x_{t-1} - x^*\|^2}_{\alpha^2} = \underbrace{\|x_t - x^*\|^2}_{\beta^2} + \underbrace{\|x_{t-1} - x_t\|^2}_{\gamma^2} - \underbrace{2\langle x_t - x^*, x_{t-1} - x_t\rangle}_{2\beta\gamma\cos(\hat{A})}. \tag{1}$$

Notice, however that the above expression relies strongly on properties of Euclidean geometry (and the flatness of the corresponding line), namely that the the lines connecting the three points x_t, x_{t-1} and x^* form a triangle; indeed, it is the generalization of the Pythagorean theorem, known also as the law of cosines, for the induced triangle $(ABC) := \{(x_t, x_{t-1}, x^*)\}$. In a uniquely geodesic manifold such triangle may not belong to the manifold as discussed above. As a result, the difference of distances to the equilibrium using the geodesic paths $d^2_{\mathcal{M}}(x_t, x^*) - d^2_{\mathcal{M}}(x_{t-1}, x^*)$ generally cannot be given in a closed form. The manifold's curvature controls how close these paths are to forming a Euclidean triangle. In fact, the phenomenon of *distance distortion*, as it is typically called, was hypothesised by Zhang et al. [1, Section 4.2] to be the cause of exponential slowdowns when applying EG to RMMO problems when compared to their Euclidean counterparts.

Multiple attempts have been made to bypass this hurdle. Huang et al. [77] analyzed the Riemannian GDA (RGDA) for the non-convex non-concave setting. However, they do not present any last-iterate convergence results and, even in the average/best iterate setting, they only derive sub-optimal rates for the geodesic convex-concave setting due to the lack of the machinery that convex analysis and optimization offers they derive sub-optimal rates for the geodesic convex-concave case, which is the problem of our interest. The analysis of Han et al. [78] for Riemannian Hamiltonian Method (RHM), matches the rate of second-order methods in the Euclidean case. Although theoretically faster in terms of iterations, second-order methods are not preferred in practice since evaluating second order derivatives for optimization problems of thousands to millions of parameters quickly becomes prohibitive. Finally, Zhang et al. [1] leveraged the standard averaging output trick in EG to derive a sublinear convergence rate of $O(1/\epsilon)$ for the general geodesically convex-concave Riemannian framework. In addition, they conjectured that the use of a different method could close the exponential gap for the geodesically strongly-convex-strongly-convex scenario and its Euclidean counterpart.

Given this background, a crucial question underlying the potential for successful application of first-order algorithms to Riemannian settings is the following:

Is a performance gap necessary between Riemannian and Euclidean optimal convex-concave algorithms in terms of accuracy and the condition number?

1.1 Our Contributions

Our aim in this paper is to provide an extensive analysis of the Riemannian counterparts of Euclidean optimal first-order methods adapted to the manifold-constrained setting. For the case of the smooth objectives, we consider the *Riemannian corrected extragradient* (RCEG) method while for non-smooth cases, we analyze the textbook *Riemannian gradient descent ascent* (RGDA) method. Our main results are summarized in the following table.

Alg: *RCEG*. Smooth setting with *ℓ-Lipschitz Gradient* (cf. Assumption 2.1, 3.1 and 3.2)

Perf. Measure	**Setting**	**Complexity**	**Theorem**
Last-Iterate	Det. GSCSC	$O\left(\kappa(\sqrt{\tau_0}+\frac{1}{\underline{\xi}_0})\log(\frac{1}{\epsilon})\right)$	**Thm.** 3.1
Last-Iterate	Stoc. GSCSC	$O\left(\kappa(\sqrt{\tau_0}+\frac{1}{\underline{\xi}_0})\log(\frac{1}{\epsilon})+\frac{\sigma^2\bar{\xi}_0}{\mu^2\epsilon}\log(\frac{1}{\epsilon})\right)$	**Thm.** 3.2
Avg-Iterate	Det. GCC	$O\left(\frac{\ell\sqrt{\tau_0}}{\epsilon}\right)$	[1, **Thm.1**]
Avg-Iterate	Stoc. GCC	$O\left(\frac{\ell\sqrt{\tau_0}}{\epsilon}+\frac{\sigma^2\bar{\xi}_0}{\epsilon^2}\right)$	**Thm.** 3.3

Alg: *RGDA*. Nonsmooth setting with *L-Lipschitz Function* (cf. Assumption D.1 and D.2)

Last-Iterate	Det. GSCSC	$O\left(\frac{L^2\bar{\xi}_0}{\mu^2\epsilon}\right)$	**Thm.** D.1
Last-Iterate	Stoc. GSCSC	$O\left(\frac{(L^2+\sigma^2)\bar{\xi}_0}{\mu^2\epsilon}\right)$	**Thm.** D.3
Avg-Iterate	Det. GCC	$O\left(\frac{L^2\bar{\xi}_0}{\epsilon^2}\right)$	**Thm.** D.2
Avg-Iterate	Stoc. GCC	$O\left(\frac{(L^2+\sigma^2)\bar{\xi}_0}{\epsilon^2}\right)$	**Thm.** D.4

For the definition of the acronyms, Det and Stoc stand for deterministic and stochastic, respectively. GSCSC and GCC stand for geodesically strongly-convex-strongly-concave (cf. Assumption 3.1 or Assumption D.1) and geodesically convex-concave (cf. Assumption 3.2 or Assumption D.2). Here $\epsilon \in (0,1)$ is the accuracy, L, ℓ the Lipschitzness of the objective and its gradient, $\kappa = \ell/\mu$ is the condition number of the function, where μ is the strong convexity parameter, $(\tau_0, \underline{\xi}_0, \bar{\xi}_0)$ are curvature parameters (cf. Assumption 2.1), and σ^2 is the variance of a Riemannian gradient estimator.

Our first main contribution is the derivation of a linear convergence rate for RCEG, answering the open conjecture of [1] about the performance gap of single-loop extragradient methods. Indeed, while a direct comparison between $d^2_{\mathcal{M}}(x_t, x^*)$ and $d^2_{\mathcal{M}}(x_{t-1}, x^*)$ is infeasible, we are able to establish a relationship between the iterates via appeal to the duality gap function and obtain a contraction in terms of $d^2_{\mathcal{M}}(x_t, x^*)$. In other words, the effect of Riemannian distance distortion is quantitative (the contraction ratio will depend on it) rather than qualitative (the geometric contraction still remains under a proper choice of constant stepsize). More specifically, we use $d^2_{\mathcal{M}}(x_t, x^\star) + d^2_{\mathcal{N}}(y_t, y^\star)$ and $d^2_{\mathcal{M}}(x_{t+1}, x^\star) + d^2_{\mathcal{N}}(y_{t+1}, y^\star)$ to bound a gap function defined by $f(\hat{x}_t, y^\star) - f(x^\star, \hat{y}_t)$. Since the objective function is geodesically strongly-convex-strongly-concave, we have $f(\hat{x}_t, y^\star) - f(x^\star, \hat{y}_t)$ is lower bounded by $\frac{\mu}{2}(d_{\mathcal{M}}(\hat{x}_t, x^\star)^2 + d_{\mathcal{N}}(\hat{y}_t, y^\star)^2)$. Then, using the relationship between (x_t, y_t) and $(\hat{x}_t, \hat{y}_t)$, we conclude the desired results in Theorem 3.1. Notably, our approach is not affected by the nonlinear geometry of the manifold.

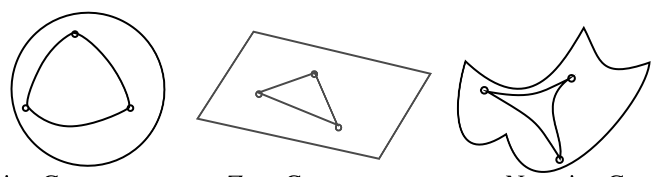
Secondly, we endeavor to give a systematic analysis of aspects of the objective function, including its smoothness, its convexity and oracle access. As we shall see, similar to the Euclidean case, better finite-time convergence guarantees are connected with a geodesic smoothness condition. For the sake of completeness, in the paper's supplement we present the performance of Riemannian GDA for the full spectrum of stochasticity for the non-smooth case. More specifically, for the stochastic setting, the key ingredient to get the optimal convergence rate is to carefully select the step size such that the noise of the gradient estimator will not affect the final convergence rate significantly. As a highlight, such technique has been used for analyzing stochastic RCEG in the Euclidean setting [79] and our analysis can be seen as the extension to the Riemannian setting. For the nonsmooth setting, the analysis is relatively simpler compared to smooth settings but we still need to deal with the issue caused by the nonlinear geometry of manifolds and the interplay between the distortion of Riemannian metrics, the gap function and the bounds of Lipschitzness of our bi-objective. Interestingly, the rates we derive are near optimal in terms of accuracy and condition number of the objective, and analogous to their Euclidean counterparts.

2 Preliminaries and Technical Background

We present the basic setup and optimality conditions for Riemannian min-max optimization. Indeed, we focus on some of key concepts that we need from Riemannian geometry, deferring a fuller presentation, including motivating examples and further discussion of related work, to Appendix A-C.

Riemannian geometry. An n-dimensional manifold $\mathcal{M}$ is a topological space where any point has a neighborhood that is homeomorphic to the n-dimensional Euclidean space. For each $x \in \mathcal{M}$, each tangent vector is tangent to all parametrized curves passing through x and the tangent space $T_x\mathcal{M}$ of a manifold $\mathcal{M}$ at this point is defined as the set of all tangent vectors. A Riemannian manifold $\mathcal{M}$ is a smooth manifold that is endowed with a smooth ("Riemannian") metric $\langle \cdot, \cdot \rangle_x$ on the tangent space $T_x\mathcal{M}$ for each point $x \in \mathcal{M}$. The inner metric induces a norm $\| \cdot \|_x$ on the tangent spaces.

A geodesic can be seen as the generalization of an Euclidean linear segment and is modeled as a smooth curve (map), $\gamma : [0,1] \mapsto \mathcal{M}$, which is locally a distance minimizer. Additionally, because of the non-flatness of a manifold a different relation between the angles and the lengths of an arbitrary geodesic triangle is induced. This distortion can be quantified via the *sectional curvature* parameter $\kappa_{\mathcal{M}}$ thanks to Toponogov's theorem [80, 81]. A constructive consequence of this definition are the



trigonometric comparison inequalities (TCIs) that will be essential in our proofs; see Alimisis et al. [82, Corollary 2.1] and Zhang and Sra [83, Lemma 5] for detailed derivations. Assuming bounded sectional curvature, TCIs provide a tool for bounding Riemannian "inner products" that are more troublesome than classical Euclidean inner products.

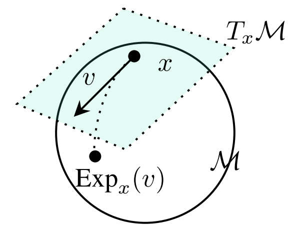
The following proposition summarizes the TCIs that we will need; note that if $\kappa_{\min} = \kappa_{\max} = 0$ (i.e., Euclidean spaces), then the proposition reduces to the law of cosines.

Proposition 2.1 *Suppose that $\mathcal{M}$ is a Riemannian manifold and let Δ be a geodesic triangle in $\mathcal{M}$ with the side length a, b, c and let A be the angle between b and c. Then, we have*

1. *If $\kappa_{\mathcal{M}}$ that is upper bounded by $\kappa_{\max} > 0$ and the diameter of $\mathcal{M}$ is bounded by $\frac{\pi}{\sqrt{\kappa_{\max}}}$, then*
$$a^2 \geq \underline{\xi}(\kappa_{\max}, c) \cdot b^2 + c^2 - 2bc\cos(A),$$
where $\underline{\xi}(\kappa, c) := 1$ for $\kappa \leq 0$ and $\underline{\xi}(\kappa, c) := c\sqrt{\kappa}\cot(c\sqrt{\kappa}) < 1$ for $\kappa > 0$.
2. *If $\kappa_{\mathcal{M}}$ is lower bounded by $\kappa_{\min}$, then*
$$a^2 \leq \overline{\xi}(\kappa_{\min}, c) \cdot b^2 + c^2 - 2bc\cos(A),$$
where $\overline{\xi}(\kappa, c) := c\sqrt{-\kappa}\coth(c\sqrt{-\kappa}) > 1$ if $\kappa < 0$ and $\overline{\xi}(\kappa, c) := 1$ if $\kappa \geq 0$.

Also, in contrast to the Euclidean case, x and $v = \mathrm{grad}_x f(x)$ do not lie in the same space, since $\mathcal{M}$ and $T_x\mathcal{M}$ respectively are distinct entities. The interplay between these dual spaces typically is carried out via the *exponential maps*. An exponential map at a point $x \in \mathcal{M}$ is a mapping from the tangent space $T_x\mathcal{M}$ to $\mathcal{M}$. In particular, $y := \mathrm{Exp}_x(v) \in \mathcal{M}$ is defined such that there exists a geodesic $\gamma : [0,1] \mapsto \mathcal{M}$ satisfying $\gamma(0) = x$, $\gamma(1) = y$ and $\gamma'(0) = v$. The inverse map exists since the manifold has a unique geodesic between any two points, which we denote as $\text{Exp}_x^{-1} : \mathcal{M} \mapsto T_x\mathcal{M}$. Accordingly, we have $d_{\mathcal{M}}(x, y) = \|\text{Exp}_x^{-1}(y)\|_x$ is the Riemannian distance induced by the exponential map.

Finally, in contrast again to Euclidean spaces, we cannot compare the tangent vectors at different points $x, y \in \mathcal{M}$ since these vectors lie in different tangent spaces. To resolve this issue, it suffices to define a transport mapping that moves a tangent vector along the geodesics and also preserves the length and Riemannian metric $\langle \cdot, \cdot \rangle_x$; indeed, we can define a parallel transport $\Gamma_x^y : T_x\mathcal{M} \mapsto T_y\mathcal{M}$ such that the inner product between any $u, v \in T_x\mathcal{M}$ is preserved; i.e., $\langle u, v\rangle_x = \langle \Gamma_x^y(u), \Gamma_x^y(v)\rangle_y$.



Riemannian min-max optimization and function classes. We let $\mathcal{M}$ and $\mathcal{N}$ be Riemannian manifolds with unique geodesic and bounded sectional curvature and assume that the function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ is defined on the product of these manifolds. The regularity conditions that we impose on the function f are as follows.

Definition 2.1 *A function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ is* geodesically L-Lipschitz *if for $\forall x, x' \in \mathcal{M}$ and $\forall y, y' \in \mathcal{N}$, the following statement holds true: $|f(x, y) - f(x', y')| \leq L(d_{\mathcal{M}}(x, x') + d_{\mathcal{N}}(y, y'))$. Additionally, if function f is also differentiable, it is called* geodesically ℓ-smooth *if for $\forall x, x' \in \mathcal{M}$ and $\forall y, y' \in \mathcal{N}$, the following statement holds true,*

$$\begin{aligned} \|\text{grad}_x f(x, y) - \Gamma_{x'}^x \text{grad}_x f(x', y')\| &\leq \ell(d_{\mathcal{M}}(x, x') + d_{\mathcal{N}}(y, y')), \\ \|\text{grad}_y f(x, y) - \Gamma_{y'}^y \text{grad}_y f(x', y')\| &\leq \ell(d_{\mathcal{M}}(x, x') + d_{\mathcal{N}}(y, y')), \end{aligned}$$

where $(\text{grad}_x f(x', y'), \text{grad}_y f(x', y')) \in T_{x'}\mathcal{M} \times T_{y'}\mathcal{N}$ is the Riemannian gradient of f at (x', y'), $\Gamma_{x'}^x$ is the parallel transport of $\mathcal{M}$ from x' to x, and $\Gamma_{y'}^y$ is the parallel transport of $\mathcal{N}$ from y' to y.

Definition 2.2 *A function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ is* geodesically strongly-convex-strongly-concave *with the modulus $\mu > 0$ if the following statement holds true,*

$$\begin{aligned} f(x', y) &\geq f(x, y) + \langle \text{subgrad}_x f(x, y), \text{Exp}_x^{-1}(x')\rangle_x + \tfrac{\mu}{2}(d_{\mathcal{M}}(x, x'))^2, \quad \text{for each } y \in \mathcal{N}, \\ f(x, y') &\leq f(x, y) + \langle \text{subgrad}_y f(x, y), \text{Exp}_y^{-1}(y')\rangle_y - \tfrac{\mu}{2}(d_{\mathcal{N}}(y, y'))^2, \quad \text{for each } x \in \mathcal{M}. \end{aligned}$$

where $(\text{subgrad}_x f(x', y'), \text{subgrad}_y f(x', y')) \in T_{x'}\mathcal{M} \times T_{y'}\mathcal{N}$ is a Riemannian subgradient of f at a point (x', y'). A function f is geodesically convex-concave *if the above holds true with $\mu = 0$.*

Following standard conventions in Riemannian optimization [1, 82, 83], we make the following assumptions on the manifolds and objective functions:[1]

Assumption 2.1 *The objective function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ and manifolds $\mathcal{M}$ and $\mathcal{N}$ satisfy*

1. *The diameter of the domain $\{(x, y) \in \mathcal{M} \times \mathcal{N} : -\infty < f(x, y) < +\infty\}$ is bounded by $D > 0$.*
2. *$\mathcal{M}, \mathcal{N}$ admit unique geodesic paths for any $(x, y), (x', y') \in \mathcal{M} \times \mathcal{N}$.*
3. *The sectional curvatures of $\mathcal{M}$ and $\mathcal{N}$ are both bounded in the range $[\kappa_{\min}, \kappa_{\max}]$ with $\kappa_{\min} \leq 0$. If $\kappa_{\max} > 0$, we assume that the diameter of manifolds is bounded by $\frac{\pi}{\sqrt{\kappa_{\max}}}$.*

Under these conditions, Zhang et al. [1] proved an analog of Sion's minimax theorem [69] in geodesic metric spaces. Formally, we have

$$\max_{y \in \mathcal{N}} \min_{x \in \mathcal{M}} f(x, y) = \min_{x \in \mathcal{M}} \max_{y \in \mathcal{N}} f(x, y),$$

which guarantees that there exists at least one global saddle point $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ such that $\min_{x \in \mathcal{M}} f(x, y^\star) = f(x^\star, y^\star) = \max_{y \in \mathcal{Y}} f(x^\star, y)$. Note that the unicity of geodesics assumption is algorithm-independent and is imposed for guaranteeing that a saddle-point solution always exist. Even though this rules out many manifolds of interest, there are still many manifolds that satisfy such conditions. More specifically, the Hadamard manifold (manifolds with non-positive curvature, $\kappa_{\max} = 0$) has a unique geodesic between any two points. This also becomes a common regularity condition in Riemannian optimization [82, 83]. For any point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$, the duality gap $f(\hat{x}, y^\star) - f(x^\star, \hat{y})$ thus gives an optimality criterion.

[1] In particular, our assumed upper and lower bounds $\kappa_{\min}, \kappa_{\max}$ guarantee that TCIs in Proposition 2.1 can be used in our analysis for proving finite-time convergence.

Algorithm 1 RCEG

Input: initial points (x_0, y_0) and stepsizes $\eta > 0$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
Query $(g_x^t, g_y^t) \leftarrow (\text{grad}_x f(x_t, y_t), \text{grad}_y f(x_t, y_t))$, the Riemannian gradient of f at a point (x_t, y_t)
$\hat{x}_t \leftarrow \text{Exp}_{x_t}(-\eta \cdot g_x^t)$.
$\hat{y}_t \leftarrow \text{Exp}_{y_t}(\eta \cdot g_y^t)$.
Query $(\hat{g}_x^t, \hat{g}_y^t) \leftarrow (\text{grad}_x f(\hat{x}_t, \hat{y}_t), \text{grad}_y f(\hat{x}_t, \hat{y}_t))$, the Riemannian gradient of f at a point $(\hat{x}_t, \hat{y}_t)$
$x_{t+1} \leftarrow \text{Exp}_{\hat{x}_t}(-\eta \cdot \hat{g}_x^t + \text{Exp}_{\hat{x}_t}^{-1}(x_t))$.
$y_{t+1} \leftarrow \text{Exp}_{\hat{y}_t}(\eta \cdot \hat{g}_y^t + \text{Exp}_{\hat{y}_t}^{-1}(y_t))$.
end for

Algorithm 2 SRCEG

Input: initial points (x_0, y_0) and stepsizes $\eta > 0$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
Query (g_x^t, g_y^t) as a **noisy** estimator of Riemannian gradient of f at a point (x_t, y_t).
$\hat{x}_t \leftarrow \text{Exp}_{x_t}(-\eta \cdot g_x^t)$.
$\hat{y}_t \leftarrow \text{Exp}_{y_t}(\eta \cdot g_y^t)$.
Query $(\hat{g}_x^t, \hat{g}_y^t)$ as a **noisy** estimator of Riemannian gradient of f at a point $(\hat{x}_t, \hat{y}_t)$.
$x_{t+1} \leftarrow \text{Exp}_{\hat{x}_t}(-\eta \cdot \hat{g}_x^t + \text{Exp}_{\hat{x}_t}^{-1}(x_t))$.
$y_{t+1} \leftarrow \text{Exp}_{\hat{y}_t}(\eta \cdot \hat{g}_y^t + \text{Exp}_{\hat{y}_t}^{-1}(y_t))$.
end for

Definition 2.3 *A point $(\hat{x}, \hat{y}) \in \mathcal{M} \times \mathcal{N}$ is an ϵ-*saddle point *of a geodesically convex-concave function $f(\cdot, \cdot)$ if $f(\hat{x}, y^\star) - f(x^\star, \hat{y}) \leq \epsilon$ where $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point.*

In the setting where f is geodesically strongly-convex-strongly-concave with $\mu > 0$, it is not difficult to verify the uniqueness of a global saddle point $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$. Then, we can consider the distance gap $(d(\hat{x}, x^\star))^2 + (d(\hat{y}, y^\star))^2$ as an optimality criterion for any point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$.

Definition 2.4 *A point $(\hat{x}, \hat{y}) \in \mathcal{M} \times \mathcal{N}$ is an ϵ-saddle point of a geodesically strongly-convex-strongly-concave function $f(\cdot, \cdot)$ if $(d(\hat{x}, x^\star))^2 + (d(\hat{y}, y^\star))^2 \leq \epsilon$, where $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point. If f is also geodesically ℓ-smooth, we denote $\kappa = \frac{\ell}{\mu}$ as the condition number.*

Given the above definitions, we can ask whether it is possible to find an ϵ-saddle point efficiently or not. In this context, Zhang et al. [1] have answered this question in the affirmative for the setting where f is geodesically ℓ-smooth and geodesically convex-concave; indeed, they derive the convergence rate of Riemannian corrected extragradient (RCEG) method in terms of time-average iterates and also conjecture that *RCEG does not guarantee convergence at a linear rate in terms of last iterates when f is geodesically ℓ-smooth and geodesically strongly-convex-strongly-concave, due to the existence of distance distortion*; see Zhang et al. [1, Section 4.2]. Surprisingly, we show in Section 3 that RCEG with constant stepsize can achieve last-iterate convergence at a linear rate. Moreover, we establish the optimal convergence rates of stochastic RCEG for certain choices of stepsize for both geodesically convex-concave and geodesically strongly-convex-strongly-concave settings.

3 Riemannian Corrected Extragradient Method

In this section, we revisit the scheme of Riemannian corrected extragradient (RCEG) method proposed by Zhang et al. [1] and extend it to a stochastic algorithm that we refer to as *stochastic RCEG*. We present our main results on an optimal last-iterate convergence guarantee for the geodesically strongly-convex-strongly-concave setting (both deterministic and stochastic) and a time-average convergence guarantee for the geodesically convex-concave setting (stochastic). This complements the time-average convergence guarantee for geodesically convex-concave setting (deterministic) [1, Theorem 4.1] and resolves an open problem posted in Zhang et al. [1, Section 4.2].

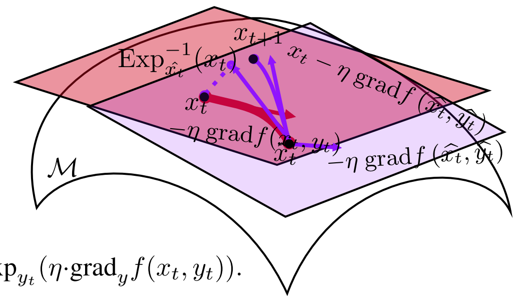
3.1 Algorithmic scheme

The recently proposed *Riemannian corrected extragradient* (RCEG) method [1] is a natural extension of the celebrated extragradient (EG) method to the Riemannian setting. Its scheme resembles that of EG in Euclidean spaces but employs a simple modification in the extrapolation step to accommodate the nonlinear geometry of Riemannian manifolds. Let us provide some intuition how such modifications work.

We start with a basic version of EG as follows, where $\mathcal{M}$ and $\mathcal{N}$ are classically restricted to be convex constraint sets in Euclidean spaces:

$$\begin{array}{rclrcl} \hat{x}_t & \leftarrow & \text{proj}_{\mathcal{M}}(x_t - \eta \cdot \nabla_x f(x_t, y_t)), & \hat{y}_t & \leftarrow & \text{proj}_{\mathcal{N}}(y_t + \eta \cdot \nabla_y f(x_t, y_t)), \\ x_{t+1} & \leftarrow & \text{proj}_{\mathcal{M}}(x_t - \eta \cdot \nabla_x f(\hat{x}_t, \hat{y}_t)), & y_{t+1} & \leftarrow & \text{proj}_{\mathcal{N}}(y_t + \eta \cdot \nabla_y f(\hat{x}_t, \hat{y}_t)). \end{array} \tag{2}$$

Turning to the setting where $\mathcal{M}$ and $\mathcal{N}$ are Riemannian manifolds, the rather straightforward way to do the generalization is to replace the projection operator by the corresponding exponential map and the gradient by the corresponding Riemannian gradient. For the first line of Eq. (2), this approach works and leads to the following updates:



$$\hat{x}_t \leftarrow \mathrm{Exp}_{x_t}(-\eta \cdot \mathrm{grad}_x f(x_t, y_t)), \quad \hat{y}_t \leftarrow \mathrm{Exp}_{y_t}(\eta \cdot \mathrm{grad}_y f(x_t, y_t)).$$

However, we encounter some issues for the second line of Eq. (2): The aforementioned approach leads to some problematic updates, $x_{t+1} \leftarrow \mathrm{Exp}_{x_t}(-\eta \cdot \mathrm{grad}_x f(\hat{x}_t, \hat{y}_t))$ and $y_{t+1} \leftarrow \mathrm{Exp}_{y_t}(\eta \cdot \mathrm{grad}_y f(\hat{x}_t, \hat{y}_t))$; indeed, the exponential maps $\mathrm{Exp}_{x_t}(\cdot)$ and $\mathrm{Exp}_{y_t}(\cdot)$ are defined from $T_{x_t}\mathcal{M}$ to $\mathcal{M}$ and from $T_{y_t}\mathcal{N}$ to $\mathcal{N}$ respectively. However, we have $-\mathrm{grad}_x f(\hat{x}_t, \hat{y}_t) \in T_{\hat{x}_t}\mathcal{M}$ and $\mathrm{grad}_y f(\hat{x}_t, \hat{y}_t) \in T_{\hat{y}_t}\mathcal{N}$. This motivates us to reformulate the second line of Eq. (2) as follows:

$$x_{t+1} \leftarrow \mathrm{proj}_{\mathcal{M}}(\hat{x}_t - \eta \cdot \nabla_x f(\hat{x}_t, \hat{y}_t) + (x_t - \hat{x}_t)), \quad y_{t+1} \leftarrow \mathrm{proj}_{\mathcal{N}}(\hat{y}_t + \eta \cdot \nabla_y f(\hat{x}_t, \hat{y}_t) + (y_t - \hat{y}_t)).$$

In the general setting of Riemannian manifolds, the terms $x_t - \hat{x}_t$ and $y_t - \hat{y}_t$ become $\mathrm{Exp}^{-1}_{\hat{x}_t}(x_t) \in T_{\hat{x}_t}\mathcal{M}$ and $\mathrm{Exp}^{-1}_{\hat{y}_t}(y_t) \in T_{\hat{y}_t}\mathcal{N}$. This observation yields the following updates:

$$x_{t+1} \leftarrow \mathrm{Exp}_{\hat{x}_t}(-\eta \cdot \mathrm{grad}_x f(\hat{x}_t, \hat{y}_t) + \mathrm{Exp}^{-1}_{\hat{x}_t}(x_t)), \quad \hat{y}_t \leftarrow \mathrm{Exp}_{\hat{y}_t}(\eta \cdot \mathrm{grad}_y f(\hat{x}_t, \hat{y}_t) + \mathrm{Exp}^{-1}_{\hat{y}_t}(y_t)).$$

We summarize the resulting RCEG method in Algorithm 1 and present the stochastic extension with noisy estimators of Riemannian gradients of f in Algorithm 2.

3.2 Main results

We present our main results on global convergence for Algorithms 1 and 2. To simplify the presentation, we treat separately the following two cases:

Assumption 3.1 *The objective function f is geodesically ℓ-smooth and geodesically strongly-convex-strongly-concave with $\mu > 0$.*

Assumption 3.2 *The objective function f is geodesically ℓ-smooth and geodesically convex-concave.*

Letting $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ be a global saddle point of f (which exists under either Assumption 3.1 or 3.2), we let $D_0 = (d_{\mathcal{M}}(x_0, x^\star))^2 + (d_{\mathcal{N}}(y_0, y^\star))^2 > 0$ and $\kappa = \ell/\mu$ for geodesically strongly-convex-strongly-concave setting. For simplicity of presentation, we also define a ratio $\tau(\cdot, \cdot)$ that measures how non-flatness changes in the spaces: $\tau([\kappa_{\min}, \kappa_{\max}], c) = \frac{\overline{\xi}(\kappa_{\min}, c)}{\underline{\xi}(\kappa_{\max}, c)} \geq 1$. We summarize our results for Algorithm 1 in the following theorem.

Theorem 3.1 *Given Assumptions 2.1 and 3.1, and letting $\eta = \min\{1/(2\ell\sqrt{\tau_0}), \underline{\xi}_0/(2\mu)\}$, there exists some $T > 0$ such that the output of Algorithm 1 satisfies that $(d(x_T, x^\star))^2 + (d(y_T, y^\star))^2 \leq \epsilon$ (i.e., an ϵ-saddle point of f in Definition 2.4) and the total number of Riemannian gradient evaluations is bounded by*

$$O\left(\left(\kappa\sqrt{\tau_0} + \frac{1}{\underline{\xi}_0}\right)\log\left(\frac{D_0}{\epsilon}\right)\right),$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D) \geq 1$ measures how non-flatness changes in $\mathcal{M}$ and $\mathcal{N}$ and $\underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D) \leq 1$ is properly defined in Proposition 2.1.

Remark 3.1 *Theorem 3.1 illustrates the last-iterate convergence of Algorithm 1 for solving geodesically strongly-convex-strongly-concave problems, thereby resolving an open problem delineated by Zhang et al. [1]. Further, the dependence on κ and $1/\epsilon$ cannot be improved since it matches the lower bound established for min-max optimization problems in Euclidean spaces [84]. However, we believe that the dependence on τ_0 and $\underline{\xi}_0$ is not tight, and it is of interest to either improve the rate or establish a lower bound for general Riemannian min-max optimization.*

Remark 3.2 *The current theoretical analysis covers local geodesic strong-convex-strong-concave settings. The key ingredient is how to define the local region; indeed, if we say the set of* $\{(x, y) : d_{\mathcal{M}}(x, x^\star) \leq \delta, d_{\mathcal{N}}(y_t, y^\star) \leq \delta\}$ *is a local region where the function is geodesic strong-convex-strong-concave. Then, the set of* $\{(x, y) : (d_{\mathcal{M}}(x, x^\star)^2 + d_{\mathcal{N}}(y_t, y^\star)^2) \leq \delta^2\}$ *must be contained in the above local region and the objective function is also geodesic strong-convex-strong-concave. If* $(x_0, y_0) \in \{(x, y) : (d_{\mathcal{M}}(x, x^\star)^2 + d_{\mathcal{N}}(y_t, y^\star)^2) \leq \delta^2\}$*, our theoretical analysis guarantees the last-iterate linear convergence rate. Such argument and definition of local region were standard for min-max optimization in the Euclidean setting; see Liang and Stokes [55, Assumption 2.1]. For an important optimization problem that is globally geodesically strongly-convex-strongly-concave, we refer to Appendix B where Robust matrix Karcher mean problem is indeed the desired one.*

In the scheme of SRECG, we highlight that (g_x^t, g_y^t) and $(\hat{g}_x^t, \hat{g}_y^t)$ are noisy estimators of Riemannian gradients of f at (x_t, y_t) and $(\hat{x}_t, \hat{y}_t)$. It is necessary to impose the conditions such that these estimators are unbiased and has bounded variance. By abuse of notation, we assume that

$$\begin{array}{ll} g_x^t = \mathrm{grad}_x f(x_t, y_t) + \xi_x^t, & g_y^t = \mathrm{grad}_y f(x_t, y_t) + \xi_y^t, \\ \hat{g}_x^t = \mathrm{grad}_x f(\hat{x}_t, \hat{y}_t) + \hat{\xi}_x^t, & \hat{g}_y^t = \mathrm{grad}_y f(\hat{x}_t, \hat{y}_t) + \hat{\xi}_y^t. \end{array} \tag{3}$$

where the noises (ξ_x^t, ξ_y^t) and $(\hat{\xi}_x^t, \hat{\xi}_y^t)$ are independent and satisfy that

$$\begin{array}{lll} \mathbb{E}[\xi_x^t] = 0, & \mathbb{E}[\xi_y^t] = 0, & \mathbb{E}[\|\xi_x^t\|^2 + \|\xi_y^t\|^2] \leq \sigma^2, \\ \mathbb{E}[\hat{\xi}_x^t] = 0, & \mathbb{E}[\hat{\xi}_y^t] = 0, & \mathbb{E}[\|\hat{\xi}_x^t\|^2 + \|\hat{\xi}_y^t\|^2] \leq \sigma^2. \end{array} \tag{4}$$

We are ready to summarize our results for Algorithm 2 in the following theorems.

Theorem 3.2 *Given Assumptions 2.1 and 3.1, letting Eq. (3) and Eq. (4) hold with* $\sigma > 0$ *and letting* $\eta > 0$ *satisfy* $\eta = \min\{\frac{1}{24\ell\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{2\mu}, \frac{2(\log(T) + \log(\mu^2 D_0 \sigma^{-2}))}{\mu T}\}$*, there exists some* $T > 0$ *such that the output of Algorithm 2 satisfies that* $\mathbb{E}[(d(x_T, x^\star))^2 + (d(y_T, y^\star))^2] \leq \epsilon$ *and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\left(\kappa\sqrt{\tau_0} + \frac{1}{\underline{\xi}_0}\right)\log\left(\frac{D_0}{\epsilon}\right) + \frac{\sigma^2 \overline{\xi}_0}{\mu^2 \epsilon}\log\left(\frac{1}{\epsilon}\right)\right),$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D) \geq 1$ *measures how non-flatness changes in* $\mathcal{M}$ *and* $\mathcal{N}$ *and* $\underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D) \leq 1$ *is properly defined in Proposition 2.1.*

Theorem 3.3 *Given Assumptions 2.1 and 3.2 and assume that Eq. (3) and Eq. (4) hold with* $\sigma > 0$ *and let* $\eta > 0$ *satisfies that* $\eta = \min\{\frac{1}{4\ell\sqrt{\tau_0}}, \frac{1}{\sigma}\sqrt{\frac{D_0}{\overline{\xi}_0 T}}\}$*, there exists some* $T > 0$ *such that the output of Algorithm 2 satisfies that* $\mathbb{E}[f(\bar{x}_T, y^\star) - f(x^\star, \bar{y}_T)] \leq \epsilon$ *and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\frac{\ell D_0\sqrt{\tau_0}}{\epsilon} + \frac{\sigma^2 \overline{\xi}_0}{\epsilon^2}\right),$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D)$ *measures how non-flatness changes in* $\mathcal{M}$ *and* $\mathcal{N}$ *and* $\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D) \geq 1$ *is properly defined in Proposition 2.1. The time-average iterates* $(\bar{x}_T, \bar{y}_T) \in \mathcal{M} \times \mathcal{N}$ *can be computed by* $(\bar{x}_0, \bar{y}_0) = (0, 0)$ *and the inductive formula:* $\bar{x}_{t+1} = \mathrm{Exp}_{\bar{x}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}_{\bar{x}_t}^{-1}(\hat{x}_t))$ *and* $\bar{y}_{t+1} = \mathrm{Exp}_{\bar{y}_t}(\frac{1}{t+1} \cdot \mathrm{Exp}_{\bar{y}_t}^{-1}(\hat{y}_t))$ *for all* $t = 0, 1, \ldots, T-1$.

Remark 3.3 *Theorem 3.2 presents the last-iterate convergence rate of Algorithm 2 for solving geodesically strongly-convex-strongly-concave problems while Theorem 3.3 gives the time-average convergence rate when the function* f *is only assumed to be geodesically convex-concave. Note that we carefully choose the stepsizes such that our upper bounds match the lower bounds established for stochastic min-max optimization problems in Euclidean spaces [79, 85, 86], in terms of the dependence on* κ*,* $1/\epsilon$ *and* σ^2*, up to log factors.*

Discussions: The last-iterate linear convergence rate in terms of Riemannian metrics is only limited to geodesically strongly convex-concave cases but other results, e.g., the average-iterate sublinear convergence rate, are derived under more mild conditions. This is consistent with classical results in the Euclidean setting where geodesic convexity reduces to convexity; indeed, the last-iterate linear convergence rate in terms of squared Euclidean norm is only obtained for strongly convex-concave cases. As such, our setting is not restrictive. Moreover, Zhang et al. [1] showed that the existence of a global saddle point is only guaranteed under the geodesically convex-concave assumption. For geodesically nonconvex-concave or geodesically nonconvex-nonconcave cases, a global saddle point might not exist and new optimality notions are required before algorithmic design. This question remains open in the Euclidean setting and is beyond the scope of this paper. However, we remark that an interesting class of robustification problems are nonconvex-nonconcave min-max problems in the Euclidean setting can be geodesically convex-concave in the Riemannian setting; see Appendix B.

4 Experiments

We present numerical experiments on the task of robust principal component analysis (RPCA) for symmetric positive definite (SPD) matrices. In particular, we compare the performance of Algorithm 1 and 2 with different outputs, i.e., the last iterate (x_T, y_T) versus the time-average iterate $(\bar{x}_T, \bar{y}_T)$ (see the precise definition in Theorem 3.3). Note that our implementations of both algorithms are based on the MANOPT package [87]. All the experiments were implemented in MATLAB R2021b on a workstation with a 2.6 GHz Intel Core i7 and 16GB of memory. Due to space limitations, some additional experimental results are deferred to Appendix G.

Experimental setup. The problem of RPCA [88, 89] can be formulated as the Riemannian min-max optimization problem with an SPD manifold and a sphere manifold. Formally, we have

$$\max_{M \in \mathcal{M}^d_{\text{PSD}}} \min_{x \in \mathcal{S}^d} \left\{ -x^\top M x - \frac{\alpha}{n} \sum_{i=1}^n d(M, M_i) \right\}. \tag{5}$$

In this formulation, $\alpha > 0$ denotes the penalty parameter, $\{M_i\}_{i \in [n]}$ is a sequence of given data SPD matrices, $\mathcal{M}^d_{\text{PSD}} = \{M \in \mathbb{R}^{d \times d} : M \succ 0, M = M^\top\}$ denotes the SPD manifold, $\mathcal{S}^d = \{x \in \mathbb{R}^d : \|x\| = 1\}$ denotes the sphere manifold and $d(\cdot, \cdot) : \mathcal{M}^d_{\text{PSD}} \times \mathcal{M}^d_{\text{PSD}} \mapsto \mathbb{R}$ is the Riemannian distance induced by the exponential map on the SPD manifold $\mathcal{M}^d_{\text{PSD}}$. As demonstrated by Zhang et al. [1], the problem of RPCA is nonconvex-nonconcave from a Euclidean perspective but is *locally geodesically strongly-convex-strongly-concave* and satisfies most of the assumptions that we make in this paper. In particular, the SPD manifold is complete with sectional curvature in $[-\frac{1}{2}, 1]$ [90] and the sphere manifold is complete with sectional curvature of 1. Other reasons why we use such example are: (i) it is a classical one in ML; (ii) Zhang et al. [1] also uses this example and observes the linear convergence behavior; (iii) the numerical results show that the unicity of geodesics assumption may not be necessary in practice; and (iv) this is an application where both min and max sides are done on Riemannian manifolds.

Following the previous works of Zhang et al. [1] and Han et al. [78], we generate a sequence of data matrices M_i satisfying that their eigenvalues are in the range of $[0.2, 4.5]$. In our experiment, we fix $\alpha = 1.0$ and also vary the problem dimension $d \in \{25, 50, 100\}$. The evaluation metric is set as gradient norm. We set $n = 40$ and $n = 200$ in Figure 1 and 2. For RCEG, we set $\eta = \frac{1}{2\ell}$ where $\ell > 0$ is selected via grid search. For SRCEG, we set $\eta_t = \min\{\frac{1}{2\ell}, \frac{a}{t}\}$ where $\ell, a > 0$ are selected via grid search. Additional results on the effect of stepsize are summarized in Appendix G.

Experimental results. Figure 1 summarizes the effects of different outputs for RCEG; indeed, RCEG-last and RCEG-avg refer to Algorithm 1 with last iterate and time-average iterate respectively. It is clear that the last iterate of RCEG consistently exhibits linear convergence to an optimal solution in all the settings, verifying our theoretical results in Theorem 3.1. In contrast, the average iterate of RCEG converges much slower than the last iterate of RCEG. The possible reason is that the problem of RPCA is *only* locally geodesically strongly-convex-strongly-concave and averaging with the iterates generated during early stage will significantly slow down the convergence of RCEG.

Figure 2 presents the comparison between SRCEG (with either last iterate or time-average iterate) and RCEG with last-iterate; here, SRCEG-last and SRCEG-avg refer to Algorithm 2 with last iterate and time-average iterate respectively. We observe that SRCEG with either last iterate or average iterate converge faster than RCEG at the early stage and all of them finally converge to an optimal solution. This demonstrates the effectiveness and efficiency of SRCEG in practice. It is also worth

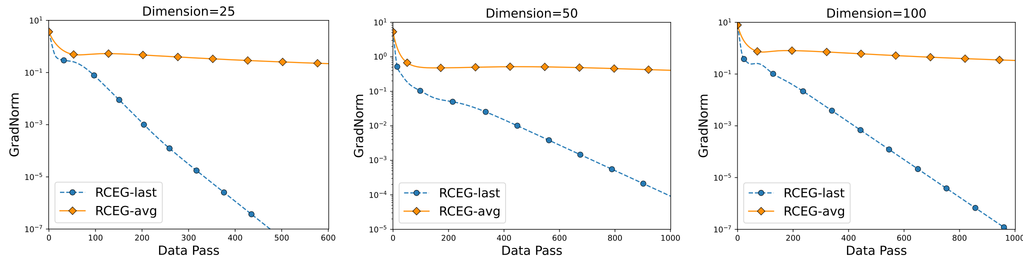


Figure 1: Comparison of last iterate (RCEG-last) and time-average iterate (RCEG-avg) for solving the RPCA problem in Eq. (5) with different problem dimensions $d \in \{25, 50, 100\}$. The horizontal axis represents the number of data passes and the vertical axis represents gradient norm.

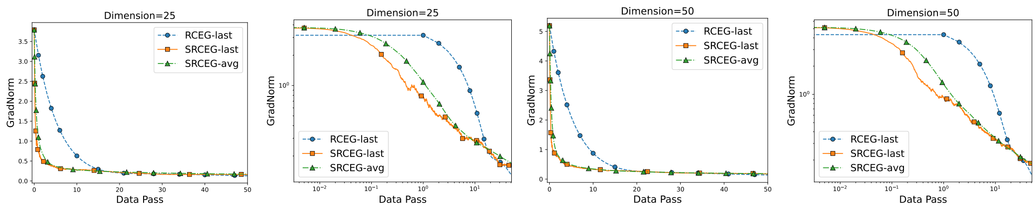


Figure 2: Comparison of RCEG and SRCEG for solving the RPCA problem in Eq. (5) with different problem dimensions $d \in \{25, 50\}$. The horizontal axis is the number of data passes and the vertical axis is gradient norm.

mentioning that the difference between last-iterate convergence and time-average-iterate convergence is not as significant as in the deterministic setting. This is possibly because the technique of averaging help cancels the negative effect of imperfect information [91, 92].

5 Conclusions

Inspired broadly by the structure of the complex competition that arises in many applications of robust optimization in ML, we focus on the problem of min-max optimization in the pure Riemannian setting (where both min and max player are constrained in a smooth manifold). Answering the open question of Zhang et al. [1] for the geodesically (strongly) convex-concave case, we showed that the Riemannian correction technique for EG matches the linear last-iterate complexity of their Euclidean counterparts in terms of accuracy and conditional number of objective for both deterministic and stochastic case. Additionally, we provide near-optimal guarantees for both smooth and non-smooth min-max optimization via Riemannian EG and GDA for the simple convex-concave case.

As a consequence of this work numerous open problems emerge; one immediate open question for future work is to explore whether the dependence on the curvature constant is also tight. Additionally, another generalization of interest would be to consider the performance of RCEG in the case of Riemannian Monotone Variational inequalities (RMVI) and examine the generalization of Zhang et al. [1] existence proof. Finally, there has been recent work in proving last-iterate convergence in the convex-concave setting via Sum-Of-Squares techniques [62]. It would be interesting to examine how one could leverage this machinery in a non-Euclidean but geodesic-metric-friendly framework.

Acknowledgements

We would like to thank the area chair and five anonymous referees for constructive suggestions that improve the paper. This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764 and by the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941. The work of Michael I. Jordan is also partially supported by NSF Grant IIS-1901252. Emmanouil V. Vlatakis-Gkaragkounis is grateful for financial support by the Google-Simons Fellowship, Pancretan Association of America and Simons Collaboration on Algorithms and Geometry. This project was completed while he was a visiting research fellow at the Simons Institute for the Theory of Computing. Additionally, he would like to acknowledge the following series of NSF-CCF grants under the numbers 1763970/2107187/1563155/1814873.

References

[1] P. Zhang, J. Zhang, and S. Sra. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *ArXiv Preprint: 2202.06950*, 2022. 1, 2, 3, 5, 6, 7, 9, 10, 20, 25, 29, 32

[2] N. Boumal and P-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *NIPS*, pages 406–414, 2011. 1, 22

[3] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016. 1, 22

[4] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016. 1, 22

[5] L. Huang, X. Liu, B. Lang, A. Yu, Y. Wang, and B. Li. Orthogonal weight normalization: solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI*, pages 3271–3278, 2018. 1, 22

[6] G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015. 1

[7] A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *COLT*, pages 2147–2148. PMLR, 2021.

[8] A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *NeurIPS*, pages 22655–22666, 2021.

[9] K. Antonakopoulos, E. V. Belmega, and P. Mertikopoulos. Online and stochastic optimization beyond Lipschitz continuity: A Riemannian approach. In *ICLR*, 2020. URL `https://openreview.net/forum?id=rkxZyaNtwB`. 1

[10] E. V. Vlatakis-Gkaragkounis, L. Flokas, T. Lianeas, P. Mertikopoulos, and G. Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *NeurIPS*, pages 1380–1391, 2020. 1

[11] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011. 1

[12] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008. 1

[13] A. Han, B. Mishra, P. K. Jawanpuria, and J. Gao. On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. In *NeurIPS*, pages 8940–8953, 2021. 1

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *NIPS*, pages 2672–2680, 2014. 1

[15] J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *ICML*, pages 7555–7564. PMLR, 2021. 1

[16] S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[17] B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963. 1

[18] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842. PMLR, 2015. 1

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

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

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

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

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

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

[25] C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In *NeurIPS*, pages 5987–5997, 2019. 20

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

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

[28] R. Hosseini and S. Sra. Matrix manifold optimization for Gaussian mixtures. In *NIPS*, pages 910–918, 2015. 1

[29] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. 1, 20

[30] R. Tron, B. Afsari, and R. Vidal. Riemannian consensus for manifolds with bounded curvature. *IEEE Transactions on Automatic Control*, 58(4):921–934, 2012. 1

[31] A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012. 1

[32] A. Kumar, P. Sattigeri, and P. T. Fletcher. Semi-supervised learning with GANs: Manifold invariance with improved inference. In *NIPS*, pages 5540–5550, 2017. 1

[33] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. Smagt. Metrics for deep generative models. In *AISTATS*, pages 1540–1550. PMLR, 2018. 1

[34] T. Lin, C. Fan, N. Ho, M. Cuturi, and M. I. Jordan. Projection robust Wasserstein distance and Riemannian optimization. In *NeurIPS*, pages 9383–9397, 2020. 1, 21

[35] P. Mertikopoulos and W. H. Sandholm. Riemannian game dynamics. *Journal of Economic Theory*, 177:315–364, 2018. 1

[36] I. M. Bomze, P. Mertikopoulos, W. Schachinger, and M. Staudigl. Hessian barrier algorithms for linearly constrained optimization problems. *SIAM Journal on Optimization*, 29(3):2100–2127, 2019. 1

[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6629–6640, 2017. 1

[38] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *NIPS*, pages 9256–9266, 2018. 1

[39] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of N-player differentiable games. In *ICML*, pages 354–363. PMLR, 2018.

[40] P. Mertikopoulos, B. Lecouat, H. Zenati, C-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*, 2019. URL `https://openreview.net/forum?id=Bkg8jjC9KQ`. 1

[41] C. Jin, P. Netrapalli, and M. I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *ICML*, pages 4880–4889. PMLR, 2020. 1

[42] K. K. Thekumprampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *NeurIPS*, pages 12680–12691, 2019. 1

[43] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, pages 14934–14942, 2019.

[44] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

[45] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, pages 2863–2873. PMLR, 2020.

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

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

[48] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, pages 2738–2779. PMLR, 2020. 1

[49] T. Lin, C. Jin, and M. I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, pages 6083–6093. PMLR, 2020.

[50] M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34, 2021.

[51] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4): 2508–2538, 2021.

[52] W. Kong and R. D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021. 1

[53] C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, 2019. 1

[54] L. Adolphs, H. Daneshmand, A. Lucchi, and T. Hofmann. Local saddle point optimization: A curvature exploitation approach. In *AISTATS*, pages 486–495. PMLR, 2019.

[55] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915. PMLR, 2019. 8

[56] G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS*, pages 1802–1811. PMLR, 2019.

[57] E. Mazumdar, L. J. Ratliff, and S. S. Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

[58] M. Liu, Y. Mroueh, J. Ross, W. Zhang, X. Cui, P. Das, and T. Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *ICLR*, 2020. URL https://openreview.net/forum?id=SJxIm0VtwH.

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

[60] E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[61] J. Abernethy, K. A. Lai, and A. Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of Hamiltonian gradient descent and consensus optimization. In *ALT*, pages 3–47. PMLR, 2021.

[62] Y. Cai, A. Oikonomou, and W. Zheng. Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *ArXiv Preprint: 2204.09228*, 2022. 2, 10

[63] E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *NeurIPS*, pages 10450–10461, 2019. 2

[64] E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent. In *NeurIPS*, pages 2373–2386, 2021. 2

[65] J. Fearnley, P. W. Goldberg, A. Hollender, and R. Savani. The complexity of gradient descent: CLS = PPAD$\cap$ PLS. In *STOC*, pages 46–59, 2021. 2

[66] L. E. J. Brouwer. Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1): 97–115, 1911. 2

[67] S. Kakutani. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal*, 8(3):457–459, 1941. 2

[68] B. Knaster, C. Kuratowski, and S. Mazurkiewicz. Ein beweis des fixpunktsatzes für n-dimensionale simplexe. *Fundamenta Mathematicae*, 14(1):132–137, 1929. 2

[69] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. 2, 5, 20

[70] J. V. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. 2

[71] E. D. Helly. Über mengen konvexer körper mit gemeinschaftlichen punkte. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32:175–176, 1923. 2

[72] S. Ivanov. On Helly's theorem in geodesic spaces. *Electronic Research Announcements*, 21: 109, 2014. 2

[73] H. Komiya. Elementary proof for Sion's minimax theorem. *Kodai Mathematical Journal*, 11 (1):5–7, 1988. 2, 20

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

[75] S. Park. Riemannian manifolds are KKM spaces. *Advances in the Theory of Nonlinear Analysis and its Application*, 3(2):64–73, 2019. 2, 20

[76] J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. D. Yoo. Fast and efficient MMD-based fair PCA via optimization over Stiefel manifold. In *AAAI*, pages 7363–7371, 2022. 2

[77] F. Huang, S. Gao, and H. Huang. Gradient descent ascent for min-max problems on Riemannian manifolds. *ArXiv Preprint: 2010.06097*, 2020. 2, 20

[78] A. Han, B. Mishra, P. Jawanpuria, P. Kumar, and J. Gao. Riemannian Hamiltonian methods for min-max optimization on manifolds. *ArXiv Preprint: 2204.11418*, 2022. 2, 9, 20

[79] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022. 4, 8

[80] J. Cheeger and D. G. Ebin. *Comparison Theorems in Riemannian Geometry*, volume 9. North-Holland Amsterdam, 1975. 4

[81] Y. Burago, M. Gromov, and G. Perel'man. A. D. Alexandrov spaces with curvature bounded below. *Russian Mathematical Surveys*, 47(2):1, 1992. 4

[82] F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *AISTATS*, pages 1297–1307. PMLR, 2020. 4, 5, 23

[83] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *COLT*, pages 1617–1638. PMLR, 2016. 4, 5, 19, 23

[84] J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, pages 1–35, 2021. 7

[85] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. 8

[86] A. Fallah, A. Ozdaglar, and S. Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *CDC*, pages 3573–3579. IEEE, 2020. 8

[87] N. Boumal, B. Mishra, P-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(1):1455–1459, 2014. 9

[88] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011. 9

[89] M. Harandi, M. Salzmann, and R. Hartley. Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):48–62, 2017. 9

[90] C. Criscitiello and N. Boumal. An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, pages 1–77, 2022. 9

[91] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL https://openreview.net/forum?id=8gmWwjFyLj. 10

[92] Y. Yazıcı, C-S. Foo, S. Winkler, K-H. Yap, G. Piliouras, and V. Chandrasekhar. The unusual effectiveness of averaging in GAN training. In *ICLR*, 2019. URL https://openreview.net/forum?id=SJgw_sRqFQ. 10

[93] D. Burago, I. D. Burago, Y. Burago, S. Ivanov, S. V. Ivanov, and S. A. Ivanov. *A Course in Metric Geometry*, volume 33. American Mathematical Soc., 2001. 19, 22

[94] J. Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2012. 19

[95] P-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. 19, 20

[96] N. Boumal, P-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019. 19

[97] H. Kasai and B. Mishra. Inexact trust-region algorithms on Riemannian manifolds. In *NeurIPS*, pages 4249–4260, 2018. 19

[98] J. Hu, A. Milzarek, Z. Wen, and Y. Yuan. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(3): 1181–1207, 2018.

[99] J. Hu, B. Jiang, L. Lin, Z. Wen, and Y. Yuan. Structured quasi-Newton methods for optimization with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(4):A2239–A2269, 2019. 19

[100] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013. 19

[101] B. Gao, X. Liu, X. Chen, and Y. Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM Journal on Optimization*, 28(1):302–332, 2018.

[102] H. Liu, A. M-C. So, and W. Wu. Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178(1-2):215–262, 2019. 19

[103] J. Zhang, S. Ma, and S. Zhang. Primal-dual optimization algorithms over Riemannian manifolds: An iteration complexity analysis. *Mathematical Programming*, 184(1):445–490, 2020. 19

[104] P-A. Absil and S. Hosseini. A collection of nonsmooth Riemannian optimization problems. In *Nonsmooth Optimization and Its Applications*, pages 1–15. Springer, 2019. 19

[105] O. P. Ferreira and P. R. Oliveira. Subgradient algorithm on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 97(1):93–104, 1998. 19

[106] G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017. 19, 20

[107] O. P. Ferreira and P. R. Oliveira. Proximal point algorithm on Riemannian manifolds. *Optimization*, 51(2):257–270, 2002. 20

[108] S. Chen, S. Ma, A. M-C. So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020. 20

[109] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. Man-Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021. 20

[110] H. Zhang, S. J. Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *NeurIPS*, pages 4592–4600, 2016. 20

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

[112] G. Becigneul and O-E. Ganea. Riemannian adaptive optimization methods. In *ICLR*, 2019. URL https://openreview.net/forum?id=r1eiqi09K7.

[113] H. Kasai, P. Jawanpuria, and B. Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *ICML*, pages 3262–3271, 2019. 20

[114] Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on Riemannian manifolds. In *NeurIPS*, pages 7274–7284, 2019. 20

[115] B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970. 20

[116] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. 20

[117] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988. 20

[118] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[119] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. 20

[120] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011. 20

[121] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. 20

[122] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. 20

[123] F. Facchinei and J-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007. 20

[124] C. Li, G. López, and V. Martín-Márquez. Monotone vector fields and the proximal point algorithm on Hadamard manifolds. *Journal of the London Mathematical Society*, 79(3): 663–683, 2009. 20

[125] J. H. Wang, G. López, V. Martín-Márquez, and C. Li. Monotone and accretive vector fields on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 146(3):691–708, 2010. 20

[126] O. P. Ferreira, L. R. Pérez, and S. Z. Németh. Singularities of monotone vector fields and an extragradient-type algorithm. *Journal of Global Optimization*, 31(1):133–151, 2005. 20

[127] A. Ben-Tal, L. EL Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009. 20

[128] J. Hu, X. Liu, Z-W. Wen, and Y-X. Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020. 20

[129] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006. 21

[130] P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007. 21

[131] R. Bergmann and R. Herzog. Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds. *SIAM Journal on Optimization*, 29(4):2423–2444, 2019. 21

[132] T. Lin, Z. Zheng, E. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *AISTATS*, pages 262–270. PMLR, 2021. 21

[133] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. 22

[134] P. Jawanpuria and B. Mishra. A unified framework for structured low-rank matrix learning. In *ICML*, pages 2254–2263. PMLR, 2018. 22

[135] M. Bacak. *Convex Analysis and Optimization in Hadamard Spaces*, volume 22. Walter de Gruyter GmbH & Co KG, 2014. 22

[136] P. Petersen. *Riemannian Geometry*, volume 171. Springer, 2006. 23

[137] K. L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954. 32

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theoretical work that does not have any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Proof detailsa re deferred to the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyper-parameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowd-sourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]