Riemannian Accelerated Optimization: Handling Constraints to Bound Geometric Penalties

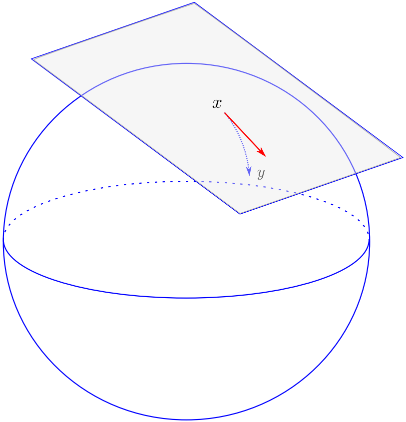
David Martínez-Rubio, Sebastian Pokutta

Technische Universität Berlin, Zuse Institute Berlin

Riemannian Optimization

For a Riemannian manifold $\mathcal{M}$:

$$\min_{x \in \mathcal{M}} f(x).$$



- Spheres, hyperbolic spaces.
- *SPD* matrices.
- $\mathrm{SO}(n)$ (real orthogonal matrices with $\det(A) = 1$).
- Stiefel manifold $V_k(\mathbb{R}^n)$ (ordered orthonormal basis of a k-dim vector space).
- ...

Riemannian Optimization

For a Riemannian manifold $\mathcal{M}$:

$$\min_{x \in \mathcal{M}} f(x).$$

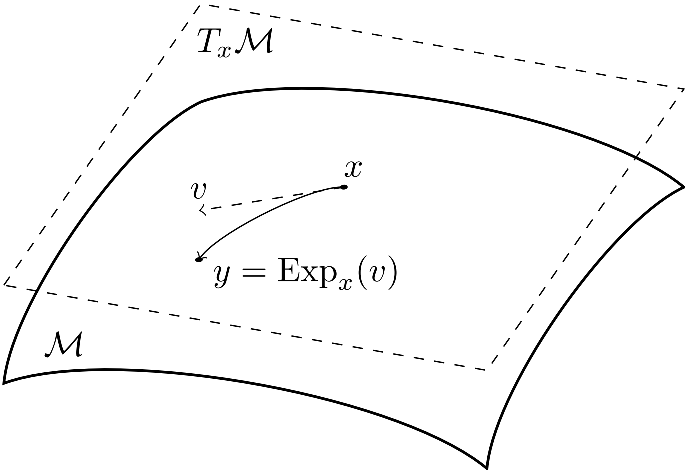
- Constrained $\rightarrow$ unconstrained.
- A function can be non-convex in the Euclidean case but geodesically convex on a manifold with the right metric $\rightarrow$ Fast algorithms.

Riemannian Optimization - Applications

- **Principal Components Analysis** (Jolliffe et al., 2003; Genicot et al., 2015; Huang and Wei, 2019).
- **Low-rank matrix completion** (Cambier and Absil, 2016; Heidel and Schulz, 2018; Mishra and Sepulchre, 2014; Tan et al., 2014; Vandereycken, 2013).
- **Dictionary learning** (Cherian and Sra, 2017; Sun et al., 2017).
- **Optimization under orthogonality constraints** (Edelman et al., 1998)
- **Robust covariance estimation in Gaussian distributions** (Wiesel, 2012).
- **Gaussian mixture models** (Hosseini and Sra, 2015).
- **Operator scaling** (Allen-Zhu et al., 2018).
- **Wasserstein Barycenters** (Hosseini and Sra, 2020)
- **Many more...**

Geodesic Convexity

Notation: Let $\mathcal{M}$ be a Riemannian manifold. Given $x, y \in \mathcal{M}$ and $w \in T_x\mathcal{M}$ we use $\langle w, y - x \rangle \stackrel{\text{def}}{=} -\langle w, x - y \rangle \stackrel{\text{def}}{=} \langle w, \mathrm{Exp}_x^{-1}(y) \rangle_x$.



Geodesic Convexity

Notation: Let $\mathcal{M}$ be a Riemannian manifold. Given $x, y \in \mathcal{M}$ and $w \in T_x\mathcal{M}$ we use $\langle w, y - x \rangle \stackrel{\text{def}}{=} -\langle w, x - y \rangle \stackrel{\text{def}}{=} \langle w, \mathrm{Exp}_x^{-1}(y) \rangle_x$.

- μ**-strongly geodesic convexity of** $F : \mathcal{M} \to \mathbb{R}$**:**

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2} d(x, y)^2, \textbf{ for } \mu > 0, \forall x, y \in \mathcal{M}.$$

- L**-smoothness:**

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2} d(x, y)^2, \quad \forall x, y \in \mathcal{M}.$$

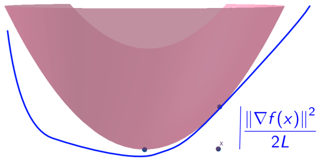
If F satisfies the μ-strong convexity inequality for $\mu = 0$ we say F is geodesically convex (g-convex).

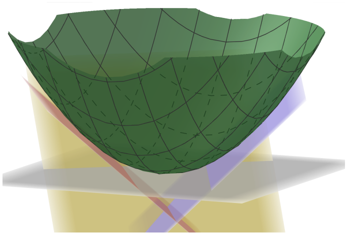
- A set $\mathcal{X}$ is uniquely geodesically convex if there is one and only one geodesic between two points, and it remains in $\mathcal{X}$.

Nesterov's Accelerated Gradient Descent (AGD) Methods

- Optimal first-order method for the minimization of Euclidean convex (resp. μ-strongly convex) and L-smooth functions.

	$\mu = 0$	$\mu > 0$ $[\kappa \stackrel{\text{def}}{=} L/\mu]$
Gradient Descent	$O(L/\varepsilon)$	$O(\kappa \log 1/\varepsilon)$
Accelerated Gradient Descent	$O(\sqrt{L/\varepsilon})$	$O(\sqrt{\kappa} \log 1/\varepsilon)$





Accelerated Gradient Descent can be seen as a combination of Gradient Descent and an online learning algorithm that have, respectively, progress and instantaneous regret that are proportional to each other (proportional to $\|\nabla f(x)\|^2$ in the unconstrained case).

Problem

Can a Riemannian first-order method enjoy the same rates as Nesterov's AGD does in the Euclidean space?

This work:

- Yes, for a wide class of Hadamard manifolds, up to log factors and geometric constants.

Problem

Can a Riemannian first-order method enjoy the same rates as Nesterov's AGD does in the Euclidean space?

This work:

- Yes, for a wide class of Hadamard manifolds, up to log factors and geometric constants.
- Crucially, we can enforce the iterates to stay in a pre-specified bounded set.

Problem

Can a Riemannian first-order method enjoy the same rates as Nesterov's AGD does in the Euclidean space?

This work:

- Yes, for a wide class of Hadamard manifolds, up to log factors and geometric constants.
- Crucially, we can enforce the iterates to stay in a pre-specified bounded set.
- We develop an inexact Riemannian proximal point method and a way to implement it via first-order methods.

Problem

Can a Riemannian first-order method enjoy the same rates as Nesterov's AGD does in the Euclidean space?

This work:

- Yes, for a wide class of Hadamard manifolds, up to log factors and geometric constants.
- Crucially, we can enforce the iterates to stay in a pre-specified bounded set.
- We develop an inexact Riemannian proximal point method and a way to implement it via first-order methods.
- We boost convergence by using a ball-optimization oracle argument.

Problem

Can a Riemannian first-order method enjoy the same rates as Nesterov's AGD does in the Euclidean space?

This work:

- Yes, for a wide class of Hadamard manifolds, up to log factors and geometric constants.
- Crucially, we can enforce the iterates to stay in a pre-specified bounded set.
- We develop an inexact Riemannian proximal point method and a way to implement it via first-order methods.
- We boost convergence by using a ball-optimization oracle argument.

In our follow-up (MRC+23): all Hadamard manifolds. Better geometric constants.

A Riemannian accelerated inexact proximal point method

- Estimate a regularized lower bound on $f(x^*)$ by "moving" the bound and aggregating.

A Riemannian accelerated inexact proximal point method

- Estimate a regularized lower bound on $f(x^*)$ by "moving" the bound and aggregating.

- The Moreau envelope $M_{\lambda,f}(x) \mapsto \min_{x\in\mathcal{X}}\{f(x) + \frac{1}{2\lambda}d(x,\hat{x})^2\}$ is g-convex in Hadamard manifolds. Use it to get greater descent than with RGD.

A Riemannian accelerated inexact proximal point method

- Estimate a regularized lower bound on $f(x^*)$ by "moving" the bound and aggregating.

- The Moreau envelope $M_{\lambda,f}(x) \mapsto \min_{x\in\mathcal{X}}\{f(x) + \frac{1}{2\lambda}d(x,\hat{x})^2\}$ is g-convex in Hadamard manifolds. Use it to get greater descent than with RGD.

- Inexact proximal point method. Subproblems are strongly g-convex and smooth with condition number $O(\zeta)$ (independent from the conditioning of f).

A Riemannian accelerated inexact proximal point method

- Estimate a regularized lower bound on $f(x^*)$ by "moving" the bound and aggregating.

- The Moreau envelope $M_{\lambda,f}(x) \mapsto \min_{x\in\mathcal{X}}\{f(x) + \frac{1}{2\lambda}d(x,\hat{x})^2\}$ is g-convex in Hadamard manifolds. Use it to get greater descent than with RGD.

- Inexact proximal point method. Subproblems are strongly g-convex and smooth with condition number $O(\zeta)$ (independent from the conditioning of f).

- In a ball $\bar{B}(x_0, D)$ (D not dependent on the condition number of f) the pull-back of the prox function to the Euclidean space via $f(\mathrm{Exp}_{x_0}(x)) + \frac{1}{2\lambda}d(\mathrm{Exp}_{x_0}(x),\hat{x})^2$ is strongly convex with condition number $O(\zeta)$.

Proximal subproblem and Ball Optimization Oracle

- After using a warm start, we can approximate the prox with linear rates.
- Using this procedure we can implement an approximate ball optimization oracle.
- Distance to x^* with an exact ball optimization oracle does not increase and the distance is controlled with an approximate ball optimization oracle.
- $\widetilde{O}(\zeta)$ ball optimization iterations suffice to optimize.

Comparison with Related Work

Method	**g-convex**	**μ-st. g-cvx**	**K?**	**G?**	**F?**	**C?**
(Nes05)	$O(\sqrt{\frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\sqrt{\kappa})$	0	✓	✓	✓
(ZS18)	-	$\widetilde{O}(\sqrt{\kappa})$	R	L	✓	✗
(AS20)	-	$\widetilde{O}(\kappa)$	R	✓	✗	✗
(Mar22)	$\widetilde{O}(\zeta^{\frac{3}{2}}\sqrt{\zeta + \frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\zeta^{\frac{3}{2}}\sqrt{\kappa})$	c	✓	✓	✓
(CB22)	-	$O(\sqrt{\kappa})$	R*	L'	✓	✓
(KY22)	$O(\zeta\sqrt{\frac{LR^2}{\varepsilon}})$	$O(\zeta\sqrt{\kappa})$	R	✓	✓	✗
This work	$\widetilde{O}(\zeta^2\sqrt{\zeta + \frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\zeta^2\sqrt{\kappa})$	H*	✓	✓	✓
This work**	$\widetilde{O}(\zeta\sqrt{\zeta + \frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\zeta\sqrt{\kappa})$	H	✓	✓	✓
(MRC+23)	$\widetilde{O}(\zeta^{\frac{3}{2}}\sqrt{\zeta + \frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\zeta^{\frac{3}{2}}\sqrt{\kappa})$	H	✓	✓	✓
(MRC+23)**	$\widetilde{O}(\zeta^{\frac{1}{2}}\sqrt{\zeta + \frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\sqrt{\zeta\kappa} + \zeta)$	H	✓	✓	✓

* $\|\nabla \mathfrak{R}\| = 0$. Most applications satisfy this. Bounded by a constant works.

** Requires possibly hard projection. But useful for grad. oracle complexity.

K? = curvature;

G?=global?

F?= fully accelerated?

C? = enforces some constraints?

$\kappa \stackrel{\text{def}}{=} L/\mu$.

H = Hadamard.

R = Riemannian.

c = ctant. curv.

Lower bound: $\widetilde{\Omega}(\zeta + \sqrt{\kappa})$

Future work

- Other manifolds: positive curvature, bounded curvature.
- Remove extra logarithmic factors.
- Can geometric constants be reduced? Maybe we need better lower bounds.
- Stochastic case.