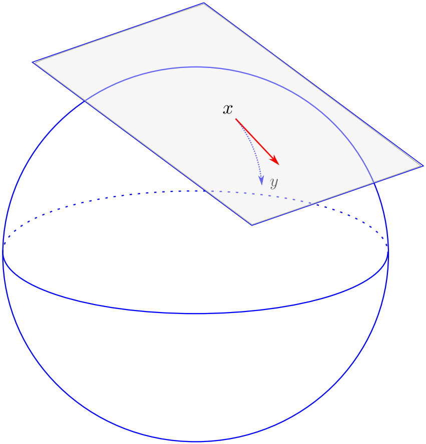
Global Riemannian Acceleration in Hyperbolic and Spherical Spaces

David Martínez-Rubio

Department of Computer Science - University of Oxford (Now at Zuse Institute Berlin)

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

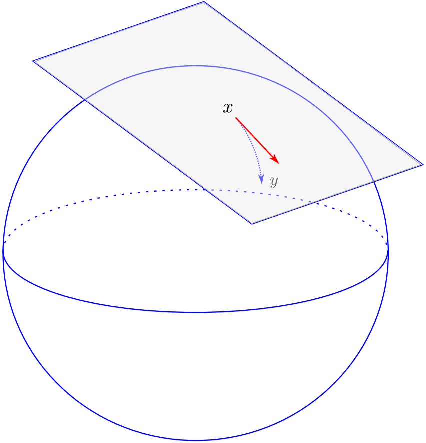
$$\min_{x \in \mathcal{M}} f(x).$$



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 $\rightarrow$ Efficient optimization.

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 $\rightarrow$ Efficient optimization.

Many first-order methods have analogous Riemannian counterparts:

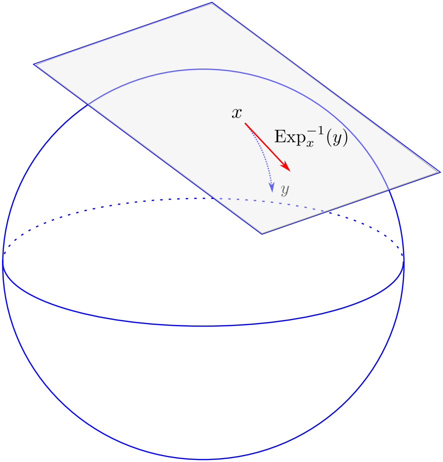
- **Deterministic** (de Carvalho Bento et al., 2017; Zhang and Sra, 2016).
- **Stochastic** (Hosseini and Sra, 2017; Khuzani and Li, 2017; Tripuraneni et al., 2018).
- **Variance reduced** (Sato et al., 2017, 2019; Zhang et al., 2016).
- **Adaptive** (Kasai et al., 2019).
- **Saddle-point escaping** (Criscitiello and Boumal, 2019; Sun et al., 2019; Zhang et al., 2018; Zhou et al., 2019; Criscitiello and Boumal, 2020).
- **Projection free** (Weber and Sra, 2017, 2019).

Riemannian Optimization - Applications

- **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)
  - **Some applications to RNNs** (Lezcano-Casado and **M-R.**, 2019).
- **Robust covariance estimation in Gaussian distributions** (Wiesel, 2012).
- **Gaussian mixture models** (Hosseini and Sra, 2015).
- **Operator scaling** (Allen-Zhu et al., 2018).
- **Sparse principal component analysis** (Jolliffe et al., 2003; Genicot et al., 2015; Huang and Wei, 2019).
- **Many more...**

Geodesic Convexity

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



Geodesic Convexity

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

- 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).

Nesterov's Accelerated Gradient Descent (AGD) Methods

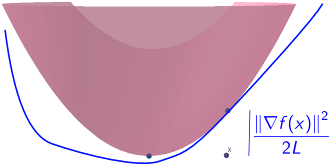
- Optimal first-order method for the minimization of Euclidean convex (resp. μ-strongly convex) and L-smooth functions.

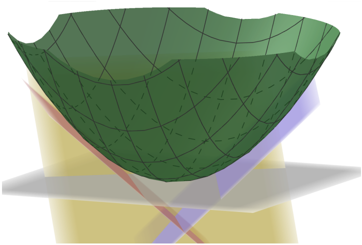
	$\mu > 0$ [$\kappa \stackrel{\text{def}}{=} L/\mu$]	$\mu = 0$
Accelerated Gradient Descent	$O(\sqrt{\kappa}\log 1/\varepsilon)$	$O(\sqrt{L/\varepsilon})$
Gradient Descent	$O(\kappa \log 1/\varepsilon)$	$O(L/\varepsilon)$

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.

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





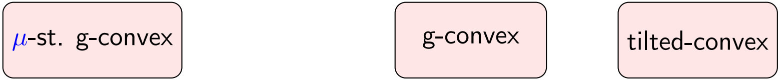
Accelerated Gradient Descent can be seen as a combination of Gradient Descent and an online learning algorithm that have, respectively, progress and instantaneous regret that are proportional to each other (proportional to $\|\nabla f(x)\|^2$ in the unconstrained case).

Problem

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

This work:

- Yes, for functions defined on manifolds of constant sectional curvature K, up to log factors and constants depending on K and the initial distance R to a minimizer.
- We reduce the problem to a ***constrained tilted-convex*** problem and optimize it in an accelerated way. The problem is non-convex and Euclidean. We provide some reductions in the Riemannian case:

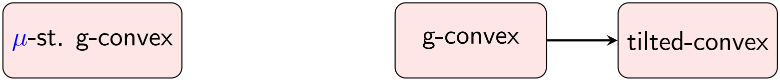


Problem

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

This work:

- Yes, for functions defined on manifolds of constant sectional curvature K, up to log factors and constants depending on K and the initial distance R to a minimizer.
- We reduce the problem to a ***constrained tilted-convex*** problem and optimize it in an accelerated way. The problem is non-convex and Euclidean. We provide some reductions in the Riemannian case:

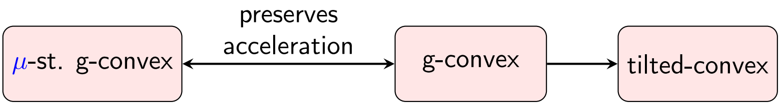


Problem

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

This work:

- Yes, for functions defined on manifolds of constant sectional curvature K, up to log factors and constants depending on K and the initial distance R to a minimizer.
- We reduce the problem to a ***constrained tilted-convex*** problem and optimize it in an accelerated way. The problem is non-convex and Euclidean. We provide some reductions in the Riemannian case:



Related Work

Method	$\mu > 0$ [$\kappa \stackrel{\text{def}}{=} L/\mu$]	$\mu = 0$
AGD in $\mathbb{R}^n$	$O(\sqrt{\kappa}\log(1/\varepsilon))$	$O(\sqrt{L/\varepsilon})$
[ZS18]	$O(\sqrt{\kappa}\log(1/\varepsilon))$ (locally: starts $O(\kappa^{-3/4})$-close)	–
[AS20]	$O^*(\kappa + \sqrt{\kappa}\log(1/\varepsilon))$	–
RGD+[ZS18]	$O^*(\kappa + \sqrt{\kappa}\log(1/\varepsilon))$	–
This work	$O^*(\sqrt{\kappa}\log(1/\varepsilon))$	$\widetilde{O}(\sqrt{L/\varepsilon})$

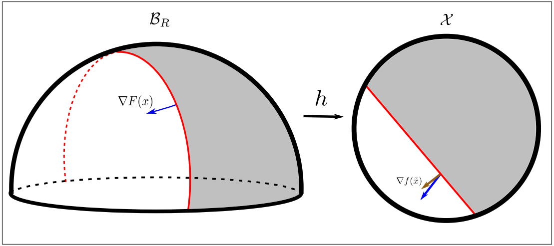
- [ZS18] Hongyi Zhang and Suvrit Sra. An Estimate Sequence for Geodesically Convex Optimization. COLT 2018.
- [AS20] Kwangjun Ahn and Suvrit Sra. From Nesterov's Estimate Sequence to Riemannian Acceleration. COLT 2020.

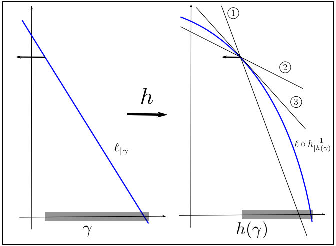
Previous works: bounded curvature.
Our work: constant curvature.

Tilted Convexity and Geodesic Maps

A function $f : \mathbb{R}^d \to \mathbb{R}$ is tilted-convex if $\exists\ \gamma_n, \gamma_p \in (0, 1]$ such that:

$$f(\tilde{x}) + \frac{1}{\gamma_n}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \leq f(\tilde{y}) \quad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \leq 0, \text{(grey area)}$$
$$f(\tilde{x}) + \gamma_p\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \leq f(\tilde{y}) \quad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \geq 0.$$



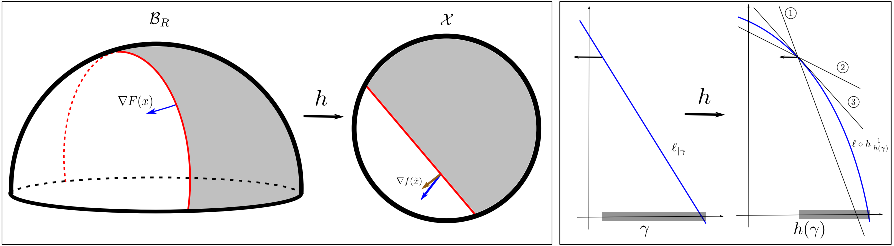


Tilted Convexity and Geodesic Maps

A function $f : \mathbb{R}^d \to \mathbb{R}$ is tilted-convex if $\exists\, \gamma_n, \gamma_p \in (0,1]$ such that:

$$f(\tilde{x}) + \frac{1}{\gamma_n}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \le f(\tilde{y}) \quad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \le 0, \text{(grey area)}$$

$$f(\tilde{x}) + \gamma_p\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \le f(\tilde{y}) \quad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \ge 0.$$



Theorem

For closed convex $Q \subseteq \mathbb{R}^d$, an L-smooth, and (γ_n, γ_p)-tilted-convex function $f : \mathbb{R}^d \to \mathbb{R}$ and $x^* \in Q$ s.t. $\nabla f(x^*) = 0$, we can find x_t s.t. $f(x_t) - f(x^*) < \varepsilon$ using $\widetilde{O}(\sqrt{L/(\gamma_n^2\gamma_p\varepsilon)})$ queries to $\nabla f(\cdot)$.

The Approximate Duality Gap Technique (ADGT)

- We obtain continuous dynamics and use an implicit Euler discretization.
- By tilted convexity we have lower bounds that are looser by a factor of $\frac{1}{\gamma_n}$, but they can be aggregated:

$$f(\tilde{x}^*) \geq \frac{\int_{t_0}^{t} f(\tilde{x}_\tau) d\alpha_\tau}{A_t} + \frac{\int_{t_0}^{t} \frac{1}{\gamma_n} \langle \nabla f(\tilde{x}_\tau), \tilde{x}^* - \tilde{x}_\tau \rangle d\alpha_\tau}{A_t}.$$

The Approximate Duality Gap Technique (ADGT)

- We obtain continuous dynamics and use an implicit Euler discretization.
- By tilted convexity we have lower bounds that are looser by a factor of $\frac{1}{\gamma_n}$, but they can be aggregated:

$$f(\tilde{x}^*) \geq \frac{\int_{t_0}^{t} f(\tilde{x}_\tau) d\alpha_\tau}{A_t} + \frac{\int_{t_0}^{t} \frac{1}{\gamma_n} \langle \nabla f(\tilde{x}_\tau), \tilde{x}^* - \tilde{x}_\tau \rangle d\alpha_\tau}{A_t}.$$

We conclude the continuous trajectory of an accelerated method should follow the differential equation:

$$\dot{\tilde{z}}_t = -\frac{1}{\gamma_n} \dot{\alpha}_t \nabla f(\tilde{x}_t); \quad \dot{\tilde{x}}_t = \frac{1}{\gamma_n} \dot{\alpha}_t \frac{\nabla \psi^*(\tilde{z}_t) - \tilde{x}_t}{\alpha_t}; \quad \tilde{z}_{t_0} = \nabla \psi^*(\tilde{x}_{t_0}), \tilde{x}_{t_0} \in \mathcal{X}.$$

Thus, we would like to have an approximate implementation of the implicit method:

$$\tilde{x}_{i+1} = \lambda_i \tilde{x}_i + (1 - \lambda_i) \nabla \psi^*(\tilde{z}_i - \frac{a_{i+1}}{\gamma_n} \nabla f(\tilde{x}_{i+1})), \quad \lambda_i \in [0, 1].$$

Discretization

Use two fixed-point iterations that approximates implicit Euler, adjusted to deal with tilted convexity:

$$\begin{cases} \tilde{\chi}_i = \lambda_i \tilde{x}_i + (1-\lambda_i)\nabla\psi^*(\tilde{z}_i); & \tilde{\zeta}_i = \tilde{z}_i - \frac{a_{i+1}}{\gamma_n}\nabla f(\tilde{\chi}_i) \\ \tilde{x}_{i+1} = \lambda_i \tilde{x}_i + (1-\lambda_i)\nabla\psi^*(\tilde{\zeta}_i); & \tilde{z}_{i+1} = \tilde{z}_i - \frac{a_{i+1}}{\gamma_n}\nabla f(\tilde{x}_{i+1}) \end{cases}$$

For a parameter $\lambda_i \in [0,1]$ depending on a value $\hat{\gamma}_i \in [\gamma_p, 1/\gamma_n]$ that we require to satisfy:

$$f(\tilde{x}_{i+1}) - f(\tilde{x}_i) \leq \hat{\gamma}_i \langle \nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i \rangle + \hat{\varepsilon},$$

- Double dependency $\tilde{x}_{i+1}(\hat{\gamma}_i)$, $\hat{\gamma}_i(\tilde{x}_{i+1})$.
- We can solve it with a binary search.

Conclusion

- **Globally** accelerated algorithm in (non-Euclidean) manifolds.
- We optimize both strongly g-convex problems as well as **g-convex** problems.
- Fast **constrained optimization of tilted-convex** problems (Euclidean, non-convex).
- Some other things:
  - Some Riemannian optimization reductions.
  - Tight lower bound on the condition number for functions defined in our manifolds.
- **Future directions:**
  - Generalization to bounded curvature.
  - Improve the dependence on curvature bounds and on the diameter of the feasible set.