Bounding Geometric Penalties in First-Order Riemannian Optimization

David Martínez-Rubio, Christophe Roux, Christopher Criscitiello, Sebastian Pokutta

Technische Universität Berlin, Zuse Institute Berlin, École Polytechnique Fédérale de Lausanne

Collaborators

Christophe Roux
(ZIB, TU Berlin)

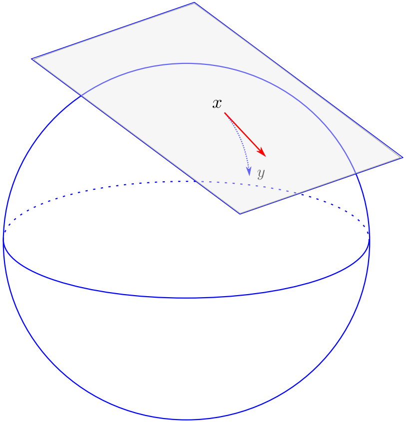
Christopher Criscitiello
(EPFL)

Sebastian Pokutta
(ZIB, TU Berlin)

Riemannian Optimization

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

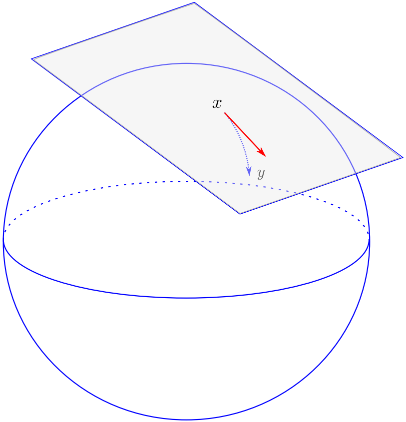
$$\min_{x \in \mathcal{M}} f(x).$$



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 - 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).
  - **Some applications to RNNs** (Lezcano-Casado and **Martínez-Rubio**, 2019).
- **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...**

Riemannian Optimization

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

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

- Constrained $\rightarrow$ unconstrained.

Riemannian Optimization

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

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

- Constrained $\to$ unconstrained.
- **Sometimes:** Euclidean non-convex $\to$ Riemannian geodesically convex.

Riemannian Optimization

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

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

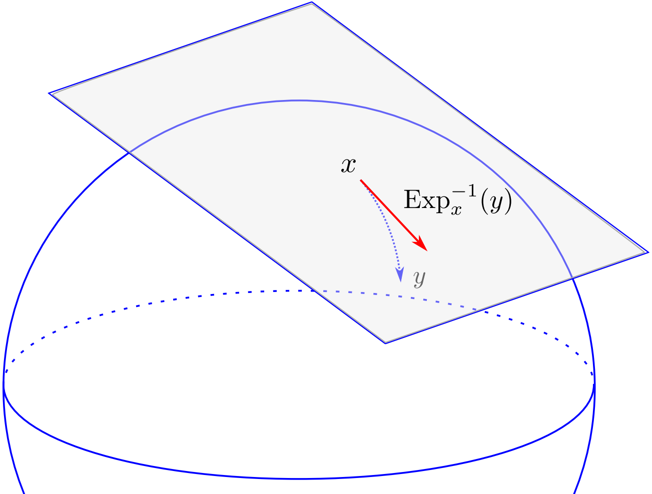
- Constrained $\rightarrow$ unconstrained.
- **Sometimes:** Euclidean non-convex $\rightarrow$ Riemannian geodesically convex.

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).
- **Accelerated** (Zhang and Sra, 2018; Ahn and Sra, 2020; Kim and Yang, 2022).
- **Min-max** (Zhang et al., 2022; Jordan et al., 2022).

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 $\langle v, y - x\rangle \overset{\text{def}}{=} -\langle v, x - y\rangle \overset{\text{def}}{=} \langle v, \mathrm{Exp}_x^{-1}(y)\rangle_x$.

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

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

  If $\mu = 0$, F is geodesically convex (g-convex).

- L**-smoothness:**

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

- G**-Lipschitzness:**

$$\|\nabla F(y)\| \leq G \text{ for all } y \in \mathcal{M}.$$

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

Distance squared and cosine inequalities

- Sectional curvature in $[K_{\min}, K_{\max}]$. Assume wlog $|K_{\min}| = 1$.
- $\Phi_x(y) \stackrel{\text{def}}{=} \frac{1}{2}d(x,y)^2$.
- $\mathcal{X} \subset \mathcal{M}$ compact, g-convex set of diameter D.

$$\nabla \Phi_x(y) = -\operatorname{Exp}_y^{-1}(x) \qquad \text{and} \qquad \delta \|v\|^2 \leq \operatorname{Hess}\Phi_x(y)[v,v] \leq \zeta \|v\|^2 \text{ for all } x, y \in \mathcal{X}.$$

where

$$\zeta \stackrel{\text{def}}{=} D\sqrt{|K_{\min}|}\coth(D\sqrt{|K_{\min}|}) = \Theta(D\sqrt{|K_{\min}|}+1) \qquad \text{if } K_{\min} < 0 \text{ else } 1.$$

$$\delta \stackrel{\text{def}}{=} D\sqrt{K_{\max}}\cot(D\sqrt{K_{\max}}) \qquad \text{if } K_{\max} > 0 \text{ else } 1.$$

Distance squared and cosine inequalities

- Sectional curvature in $[K_{\min}, K_{\max}]$. Assume wlog $|K_{\min}| = 1$.
- $\Phi_x(y) \stackrel{\text{def}}{=} \frac{1}{2} d(x,y)^2$.
- $\mathcal{X} \subset \mathcal{M}$ compact, g-convex set of diameter D.

$$\nabla \Phi_x(y) = -\operatorname{Exp}_y^{-1}(x) \qquad \text{and} \qquad \delta \|v\|^2 \leq \operatorname{Hess} \Phi_x(y)[v,v] \leq \zeta \|v\|^2 \text{ for all } x, y \in \mathcal{X}.$$

where

$$\zeta \stackrel{\text{def}}{=} D\sqrt{|K_{\min}|} \coth(D\sqrt{|K_{\min}|}) = \Theta(D\sqrt{|K_{\min}|} + 1) \qquad \text{if } K_{\min} < 0 \text{ else } 1.$$

$$\delta \stackrel{\text{def}}{=} D\sqrt{K_{\max}} \cot(D\sqrt{K_{\max}}) \qquad \text{if } K_{\max} > 0 \text{ else } 1.$$

Cosine inequalities: Let $x, y, z \in \mathcal{X}$. We have:

$$2\langle \operatorname{Exp}_x^{-1}(y), \operatorname{Exp}_x^{-1}(z) \rangle \leq \zeta d(x,y)^2 + d(x,z)^2 - d(y,z)^2,$$

$$2\langle \operatorname{Exp}_x^{-1}(y), \operatorname{Exp}_x^{-1}(z) \rangle \geq \delta d(x,y)^2 + d(x,z)^2 - d(y,z)^2.$$

In neg. curvature: minimum condition number of any L-smooth μ-strongly convex function is $\approx \zeta_D$!!

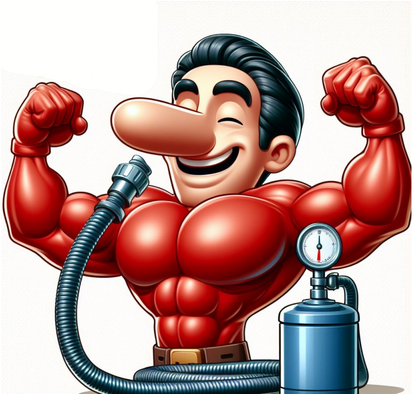
Bound what's gotta be bounded!

"*Showing that a method converges assuming iterates remain bounded is compatible with the algorithm **diverging**.*"

A. Matthem Attishen

Ha ha ha!

I proved convergence!



Bound what's gotta be bounded!

"*Showing that a method converges assuming iterates remain bounded is compatible with the algorithm **diverging**.*"

A. Matthem Attishen

*Even worse, if you assume your algorithm knows the bound **a priori**, uses its value and the **iterates depend on it**. Circularity!*

Let's do better than that.

Bound what's gotta be bounded!

"*Showing that a method converges assuming iterates remain bounded is compatible with the algorithm **diverging**.*"

A. Matthem Attishen

*Even worse, if you assume your algorithm knows the bound **a priori**, uses its value and the **iterates depend on it**. Circularity!*

Let's do better than that.

Aim of papers in my talk: Show convergence without unreasonable assumptions.

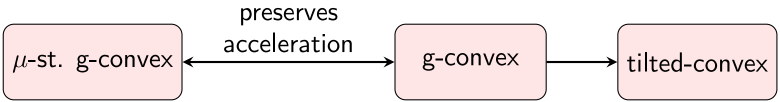
Techniques to guarantee iterates are bounded, to deal with in-manifold constraints, new rates are discovered, some times very different algorithms, etc.

You won't Believe these 7 Techniques to Bound your Riemannian Iterates!

#5 will blow up your mind!

1. Mapping to Euclidean space (I): Constant curvature solution (Ref.)

We reduce the problem to a non-convex, Euclidean ***constrained*** problem.



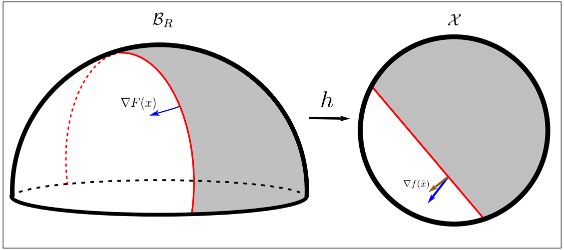
1. Mapping to Euclidean space (I): Constant curvature solution (Ref.)

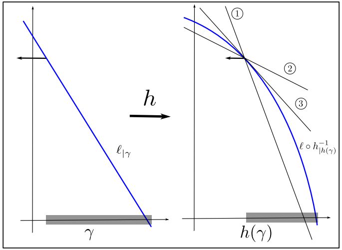
We reduce the problem to a non-convex, Euclidean ***constrained*** problem.

$$\mu\text{-st. g-convex} \xleftrightarrow{\text{preserves acceleration}} \text{g-convex} \longrightarrow \text{tilted-convex}$$

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

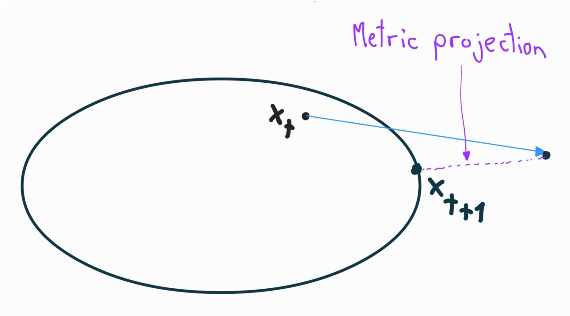
$$
\begin{aligned}
f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle &\leq f(\tilde{y}) \qquad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \leq 0, \text{(grey area)} \\
f(\tilde{x}) + \gamma_{\mathrm{p}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle &\leq f(\tilde{y}) \qquad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x}\rangle \geq 0.
\end{aligned}
$$





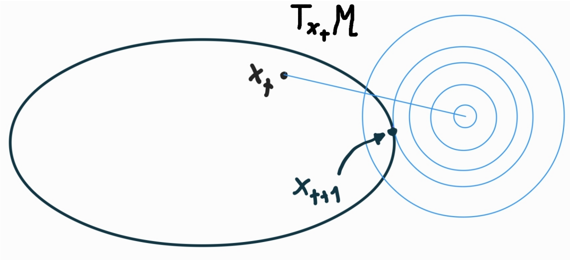
2. Metric-Projected Riemannian Gradient Descent (Ref.)

- PRGD works in **Hadamard**: $x_{t+1} = \Pi_{\mathcal{X}}(\mathrm{Exp}_{x_t}(-\eta \nabla f(x_t)))$.
- Metric projection: $\Pi_{\mathcal{X}}(x) \leftarrow \mathrm{argmin}_{y\in\mathcal{X}}\{d(y,x)\}$ for closed g-convex $\mathcal{X}$.
- Easy to implement if the constraint is a ball.
- Convergence for **Lipschitz** functions: easy.
- For **smooth** problems: not so easy.
- We show convergence and pay a ζ_R factor, where $R = G/L$ (Lipschitzness over smoothness).



3. Another Projected Riemannian Gradient Descent (Ref.)

- Minimize, in $T_{x_t}\mathcal{M}$, the quadratic upper model given by smoothness.
- $x_{t+1} = \operatorname{argmin}_{x\in\mathcal{X}}\{f(x_t) + \langle \nabla f(x_t), \operatorname{Exp}_{x_t}^{-1}(x)\rangle + \frac{L}{2}d(x, x_t)^2\}$.
- Works regardless of the curvature.
- Possibly a non-convex problem. Implementable at least in constant curvature.
- Gives better information theoretical upper bound wrt number of gradient oracle queries.



4. Proximal point algorithm (Ref.)

1. **Known:** nonexpansive operator in Hadamard manifolds.

2. **We showed:** quasi-nonexpansive, i.e., for minimizers x^* it is $d(x_t, x^*) \leq d(x_{t-1}, x^*)$ in the **general Riemannian case**.

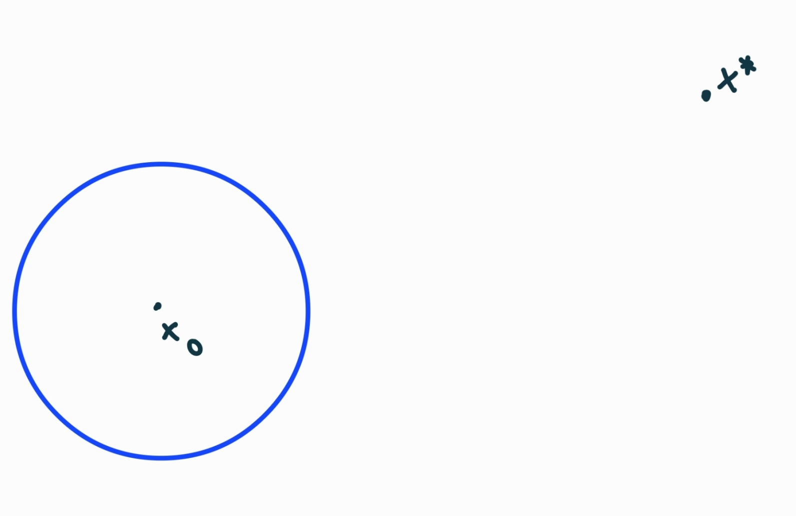
3. Approximate versions of this algorithm work and are almost quasi-nonexpansive.

4. For L-smooth functions and $\lambda = 1/L$ we get a condition number of ζ_{R_0} in $B(x, R_0)$. Only depends on the geometry!

$$x_t \leftarrow \operatorname{argmin} \left\{ f(x) + \frac{1}{\lambda} d(x, x_{t-1})^2 \right\}$$

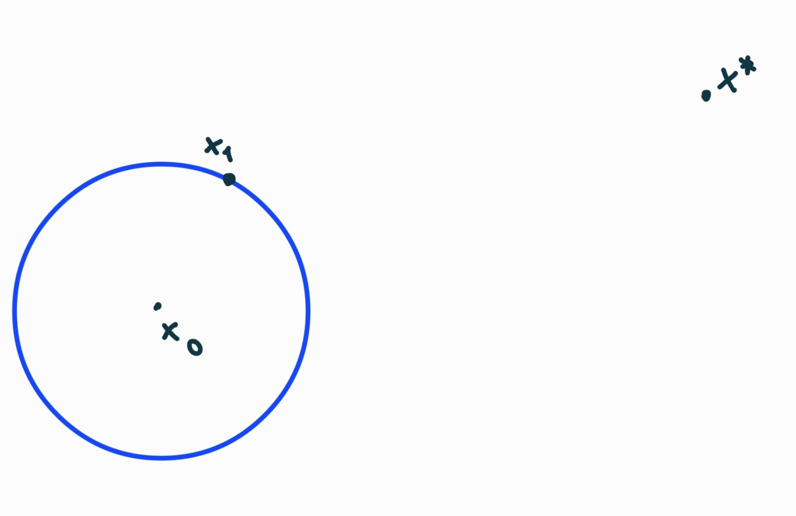
5. Ball optimization oracle (Ref. 1), (Ref. 2)

Sequentially optimize with linear rates in a ball of radius $O(1)$.
If done $O(\zeta_{R_0})$ times, you optimize globally. Initial distance: $R_0 = d(x_0, x^*)$.



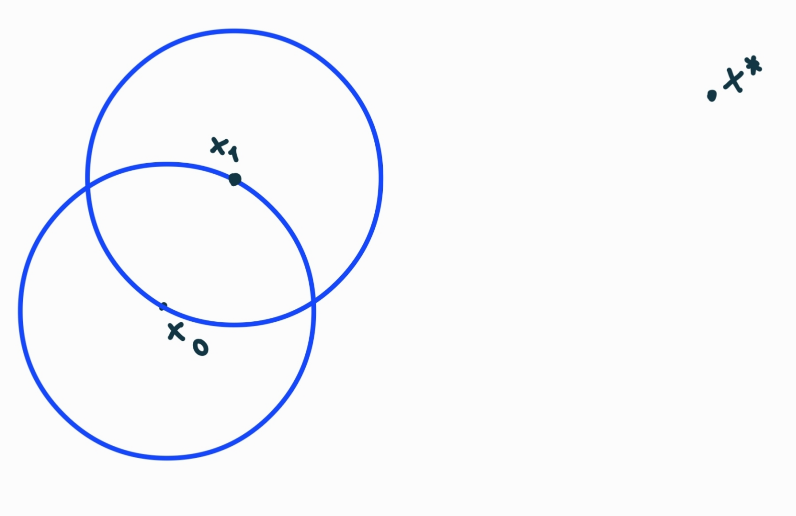
5. Ball optimization oracle (Ref. 1), (Ref. 2)

Sequentially optimize with linear rates in a ball of radius $O(1)$.
If done $O(\zeta_{R_0})$ times, you optimize globally. Initial distance: $R_0 = d(x_0, x^*)$.



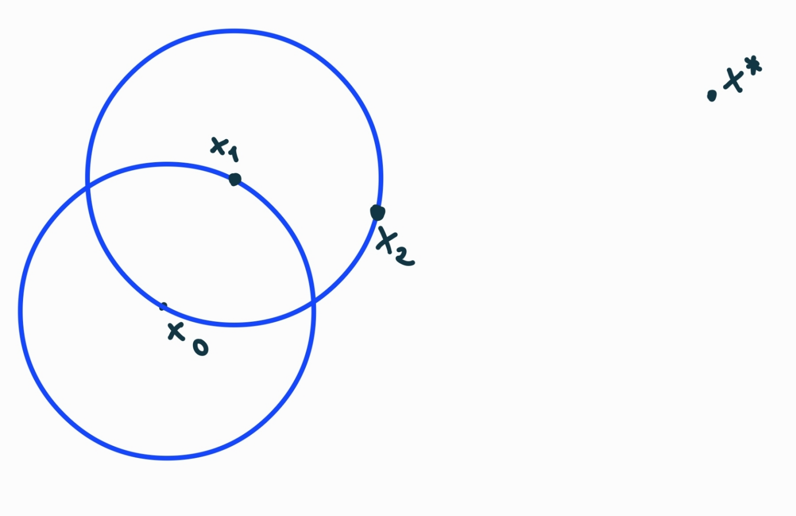
5. Ball optimization oracle (Ref. 1), (Ref. 2)

Sequentially optimize with linear rates in a ball of radius $O(1)$.
If done $O(\zeta_{R_0})$ times, you optimize globally. Initial distance: $R_0 = d(x_0, x^*)$.



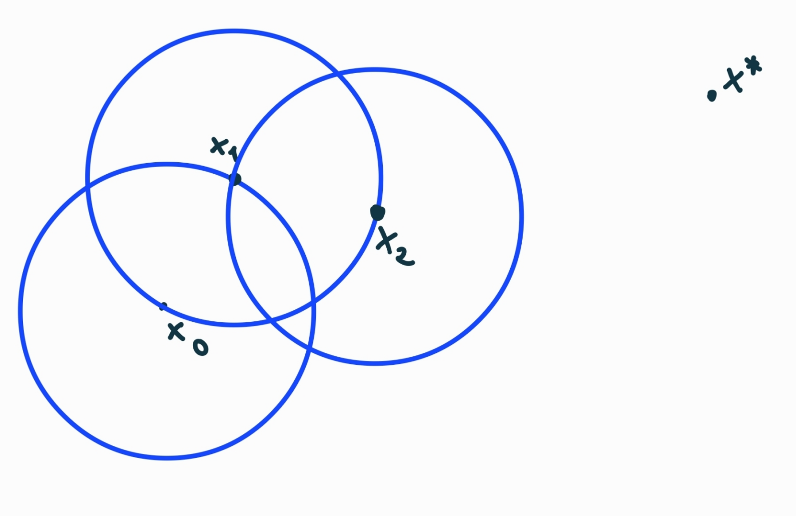
5. Ball optimization oracle (Ref. 1), (Ref. 2)

Sequentially optimize with linear rates in a ball of radius $O(1)$.
If done $O(\zeta_{R_0})$ times, you optimize globally. Initial distance: $R_0 = d(x_0, x^*)$.



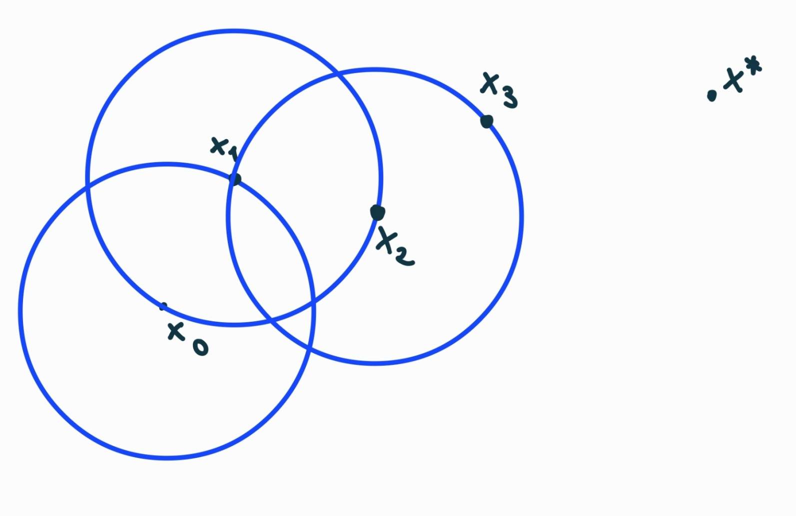
5. Ball optimization oracle (Ref. 1), (Ref. 2)

Sequentially optimize with linear rates in a ball of radius $O(1)$.
If done $O(\zeta_{R_0})$ times, you optimize globally. Initial distance: $R_0 = d(x_0, x^*)$.



5. Ball optimization oracle (Ref. 1), (Ref. 2)

Sequentially optimize with linear rates in a ball of radius $O(1)$.
If done $O(\zeta_{R_0})$ times, you optimize globally. Initial distance: $R_0 = d(x_0, x^*)$.



6. Mapping to Euclidean space (II) (Ref.)

Manifold: Locally symmetric space (all applications satisfy this). Actually it works slightly more broadly.
For f L-smooth and μ-strongly convex in a ball of center x_0, and diameter $\approx \min\{\sqrt{\frac{\mu}{L}}, \frac{\mu}{G}\}$, pulling back:

$$\hat{f} : \mathbb{R}^d \to \mathbb{R}, \quad \hat{f}(\hat{x}) = f(\mathrm{Exp}_{x_0}(\hat{x})),$$

results in $\Theta(L)$-smooth, $\Theta(\mu)$-strongly convex Euclidean function.
This technique is not ours, it is from (CB20), but we use it with the proximal method for an L-smooth function with $\lambda = 1/L$:

$$\min\left\{f(x) + \frac{L}{2}d(x, x_0)^2\right\}$$

Condition number: ζ_D. Thus, we just need diameter $D \leq \zeta_D$ if $x^* \in$ the ball. Holds for a $D = O(1)$.
This relaxes the required diameter from $O(\sqrt{\mu/L})$ to $O(1)$.

7. Showing naturally-ocurring iterate boundedness (Ref.)

1. Monotonous methods stay in the level set. But this is too bad.
2. Subproblems of proximal methods have much smaller level sets.
3. Mirror descent approaches can give us natural boundedness.
   - Euclidean step-size: we stay in a bigger ball of diameter $O(R_0\zeta_{R_0})$.
   - Smaller step size by a $\frac{1}{\zeta_{R_0}}$ factor: We stay in a ball of diameter $O(R_0)$.
   - In the hyperbolic space we can do much better. Can this be generalized?

Projected Riemannian Gradient Descent & Prox Subproblems

$D \stackrel{\text{def}}{=} \mathrm{diam}(\mathcal{X})$, $R \stackrel{\text{def}}{=} \mathrm{Lips}(F, \mathcal{X})/L$, $\lambda \stackrel{\text{def}}{=} 1/L$.

- Metric projection. Efficient steps.

$$x_{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}\left(\mathrm{Exp}_{x_t}\left(-\frac{1}{L+\zeta/\lambda}\nabla F(x_t)\right)\right).$$

Rates: $\widetilde{O}(\zeta_R\zeta_D)$, where $F(x) = f(x) + \frac{1}{2\lambda}d(x,\hat{x})^2$.

Projected Riemannian Gradient Descent & Prox Subproblems

$D \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{X})$, $R \stackrel{\text{def}}{=} \operatorname{Lips}(F, \mathcal{X})/L$, $\lambda \stackrel{\text{def}}{=} 1/L$.

- Metric projection. Efficient steps.

$$x_{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}\left(\operatorname{Exp}_{x_t}\left(-\frac{1}{L+\zeta/\lambda}\nabla F(x_t)\right)\right).$$

Rates: $\widetilde{O}(\zeta_R \zeta_D)$, where $F(x) = f(x) + \frac{1}{2\lambda}d(x, \hat{x})^2$.

- Quadratic upper model in the **tangent space**. ¿Efficient steps?

$$x_{t+1} \leftarrow \operatorname{argmin}_{y\in\mathcal{X}}\{\langle\nabla F(x_t), \operatorname{Exp}_{x_t}^{-1}(y)\rangle_{x_t} + \frac{L+\zeta/\lambda}{2}d(x_t, y)^2\}.$$

Rates: $\widetilde{O}(\zeta_D)$, where $F(x) = f(x) + \frac{1}{2\lambda}d(x, \hat{x})^2$.

Projected Riemannian Gradient Descent & Prox Subproblems

$D \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{X})$, $R \stackrel{\text{def}}{=} \operatorname{Lips}(F, \mathcal{X})/L$, $\lambda \stackrel{\text{def}}{=} 1/L$.

- Metric projection. Efficient steps.

$$x_{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}\left(\operatorname{Exp}_{x_t}\left(-\frac{1}{L+\zeta/\lambda}\nabla F(x_t)\right)\right).$$

Rates: $\widetilde{O}(\zeta_R \zeta_D)$, where $F(x) = f(x) + \frac{1}{2\lambda}d(x,\hat{x})^2$.

- Quadratic upper model in the **tangent space**. ¿Efficient steps?

$$x_{t+1} \leftarrow \operatorname{argmin}_{y\in\mathcal{X}}\{\langle\nabla F(x_t), \operatorname{Exp}_{x_t}^{-1}(y)\rangle_{x_t} + \frac{L+\zeta/\lambda}{2}d(x_t,y)^2\}.$$

Rates: $\widetilde{O}(\zeta_D)$, where $F(x) = f(x) + \frac{1}{2\lambda}d(x,\hat{x})^2$.

- Composite quadratic upper model in the **tangent space**. ¿Efficient steps?

$$x_{t+1} \leftarrow \operatorname{argmin}_{y\in\mathcal{X}}\{\langle\nabla F(x_t), \operatorname{Exp}_{x_t}^{-1}(y)\rangle_{x_t} + \frac{L}{2}d(x_t,y)^2 + g(y)\}.$$

Rates: $\widetilde{O}(1)$, where $F(x) = f(x)$ and $g(x) = \frac{1}{2\lambda}d(x,\hat{x})^2$.

Different Results and Trade-Offs in Smooth G-Convex Riem. Optimization

$R \stackrel{\text{def}}{=} d(x_0, x^*)$, $\zeta_D = \Theta(D\sqrt{|K_{\min}|}+1)$ if $K_{\min} < 0$ else 1. $K_{\min} \stackrel{\text{def}}{=} \min\{\text{sectional curv.}\}$, $\kappa = L/\mu$.

	Result	g-convex	μ-st. g-cvx	K?	C/NC?	D?	Needs R?
0	(Nes05)	$O(\sqrt{\frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\sqrt{\kappa})$	0	NC	$O(R)$	No | No
1	**(Mar22)**	$\widetilde{O}(\zeta^{\frac{3}{2}}\sqrt{\zeta+\frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\zeta^{\frac{3}{2}}\sqrt{\kappa})$	ctant.$\neq 0$	C	$O(R)$	Yes | Yes
2	(CB22)	-	$\widetilde{\Omega}(\zeta)$	$\leq c < 0$	-	-	-
3	**(MP23)**	$\widetilde{O}(\zeta^2\sqrt{\zeta+\frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\zeta^2\sqrt{\kappa})$	Hadamard*	C & NC	$O(R)$	Yes | No
4	**(MRCP23)**	$\widetilde{O}(\zeta\sqrt{\zeta+\frac{LR^2}{\varepsilon}})$	$\widetilde{O}(\sqrt{\zeta\kappa}+\zeta)$	Hadamard	C & NC	$O(R)$	Yes | No
5	(CB23)	$\widetilde{\Omega}(\zeta+\frac{LR^2}{\zeta\sqrt{\varepsilon}})$	$\widetilde{\Omega}(\sqrt{\kappa}+\zeta)$	ctant < 0	-	-	-
6	**(MRP24).1**	$O(\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\kappa)$	ctant < 0	NC	$O(R)$	No | No
7	**(MRP24).2**	$O(\zeta\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\kappa)$	bounded	NC	$O(R\zeta_R)$	No | No
8	**(MRP24).3**	$O(\zeta\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta\kappa)$	bounded	NC	$O(R)$	Yes | Yes
9	**(MRP24).4**	$O(\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\kappa)$	Hadamard	C	$O(R)$	Yes | Yes