Negative curvature obstructs acceleration for g-convex optimization,
even with exact first-order oracles

COLT 2022

Chris Criscitiello
Nicolas Boumal
OPTIM, Chair of Continuous Optimization
Institute of Mathematics, EPFL

Question

Is there a fully accelerated first-order algorithm for geodesically convex optimization with exact oracles?

Question

Is there a fully accelerated first-order algorithm for geodesically convex optimization with exact oracles?

Short answer: No.

Question

Is there a fully accelerated first-order algorithm for geodesically convex optimization with exact oracles?

Short answer: No.

Slightly longer answer: We show there are Riemannian manifolds and regimes where gradient descent is optimal (worst-case complexity).

Question

Is there a fully accelerated first-order algorithm for geodesically convex optimization with exact oracles?

Short answer: No.

Slightly longer answer: We show there are Riemannian manifolds and regimes where gradient descent is optimal (worst-case complexity).

Builds on work of Hamilton and Moitra (2021), who show the answer is no when algorithms receive noisy information.

Hamilton and Moitra: "A No-Go Theorem for Acceleration in the Hyperbolic Plane" (2021)

Geodesically convex optimization

$$\min_{x \in D} f(x)$$

Search space D is a g-convex subset of a Riemannian manifold $\mathcal{M}$:

Cost f is μ-strongly g-convex:

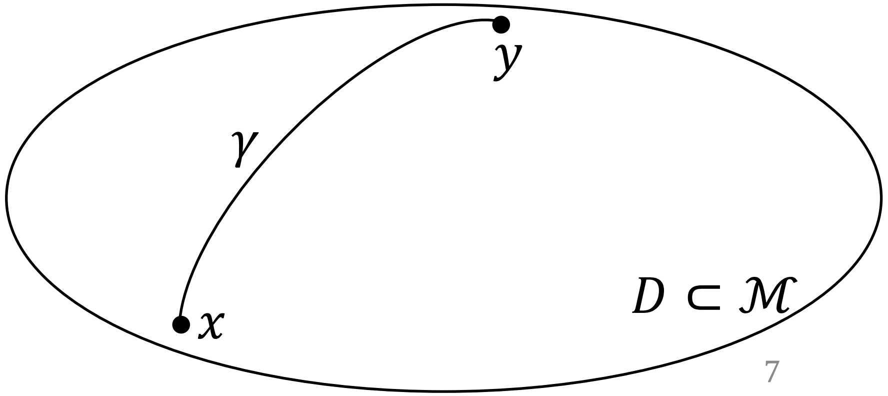
Geodesically convex optimization

$$\min_{x \in D} f(x)$$

Search space D is a g-convex subset of a Riemannian manifold $\mathcal{M}$:

For each $x, y \in D$, there is a unique minimizing geodesic $t \mapsto \gamma(t)$ contained in D, connecting x, y.

Cost f is μ-strongly g-convex:

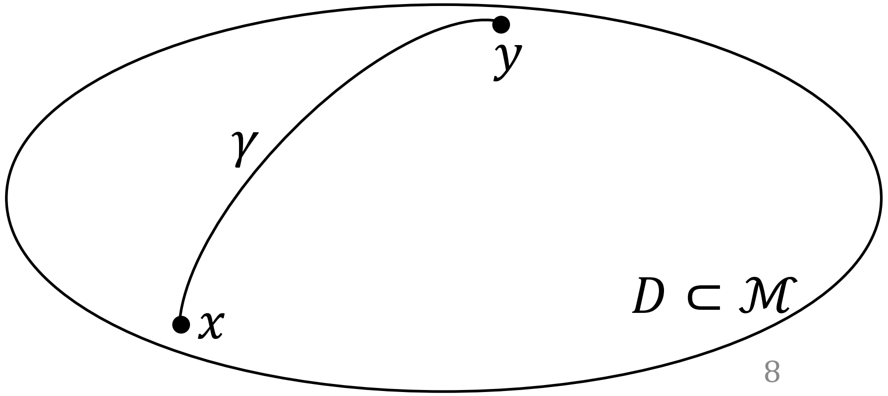


Geodesically convex optimization

$$\min_{x \in D} f(x)$$

Search space D is a g-convex subset of a Riemannian manifold $\mathcal{M}$:

For each $x, y \in D$, there is a unique minimizing geodesic $t \mapsto \gamma(t)$ contained in D, connecting x, y.



Cost f is μ-strongly g-convex:

$$t \mapsto f\big(\gamma(t)\big)$$

is μ-strongly convex for any geodesic γ in D.

Hadamard manifolds

Complete, simply connected, with non-positive (intrinsic) curvature.

Hadamard manifolds

Complete, simply connected, with non-positive (intrinsic) curvature.

Euclidean space: $\mathcal{M} = \mathbb{R}^d$

Hyperbolic space

Hadamard manifolds

Complete, simply connected, with non-positive (intrinsic) curvature.

Euclidean space: $\mathcal{M} = \mathbb{R}^d$

Hyperbolic space

Positive definite matrices: $\mathcal{M} = \{P \in \mathbf{R}^{n \times n} : P = P^\top \text{ and } P \succ 0\}$
with affine-invariant metric $\langle X, Y \rangle_P = \mathrm{Tr}(P^{-1} X P^{-1} Y)$.

Hadamard manifolds

Complete, simply connected, with non-positive (intrinsic) curvature.

Euclidean space: $\mathcal{M} = \mathbb{R}^d$

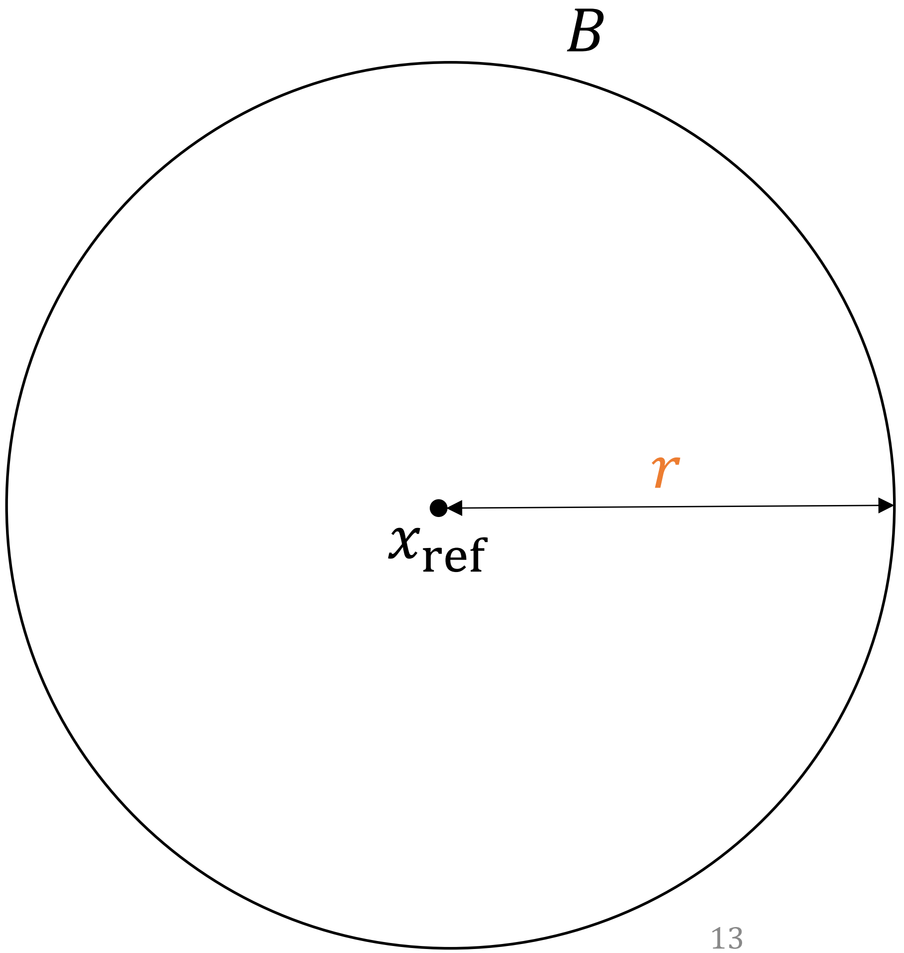
Hyperbolic space

Positive definite matrices: $\mathcal{M} = \{P \in \mathbf{R}^{n\times n}: P = P^\top \text{ and } P \succ 0\}$
with affine-invariant metric $\langle X, Y\rangle_P = \mathrm{Tr}(P^{-1}XP^{-1}Y)$.

Non-example: Sphere

Computational task

Geodesic ball $B = B(x_{\mathrm{ref}}, r)$ of radius r in Hadamard space $\mathcal{M}$.

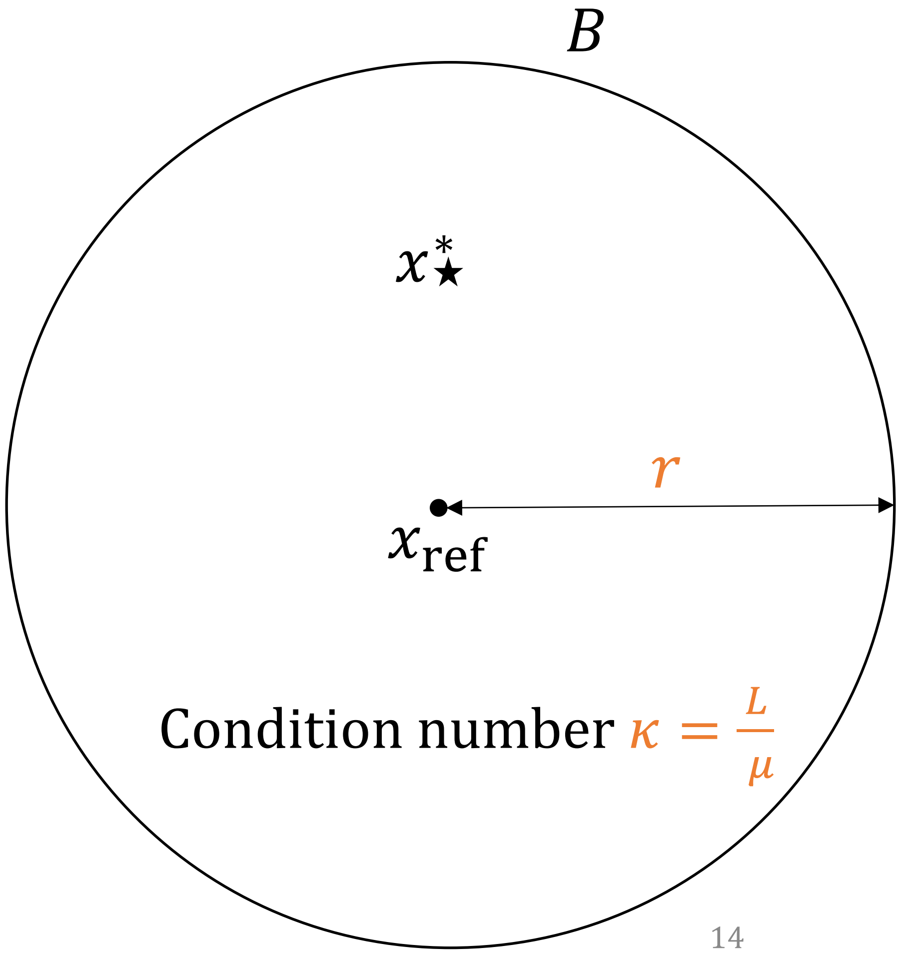


Computational task

Geodesic ball $B = B(x_{\mathrm{ref}}, r)$ of radius r in Hadamard space $\mathcal{M}$.

You know:

- f is L-smooth in B and μ-strongly g-convex in $\mathcal{M}$;
- f has a unique minimizer x^* in B.



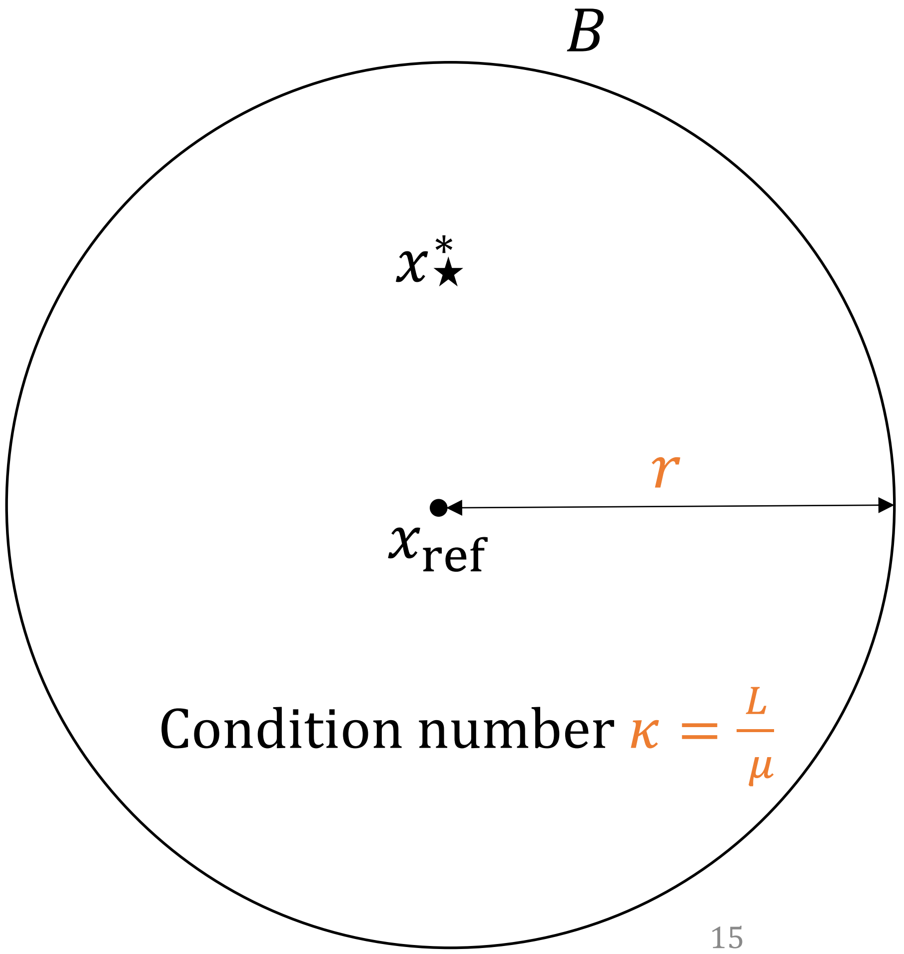
Computational task

Geodesic ball $B = B(x_{\mathrm{ref}}, r)$ of radius r in Hadamard space $\mathcal{M}$.

You know:

- f is L-smooth in B and μ-strongly g-convex in $\mathcal{M}$;
- f has a unique minimizer x^* in B.

You can query an oracle at x to get $f(x), \nabla f(x)$
(exact info, no noise).



Computational task

Geodesic ball $B = B(x_{\mathrm{ref}}, r)$ of radius r in Hadamard space $\mathcal{M}$.

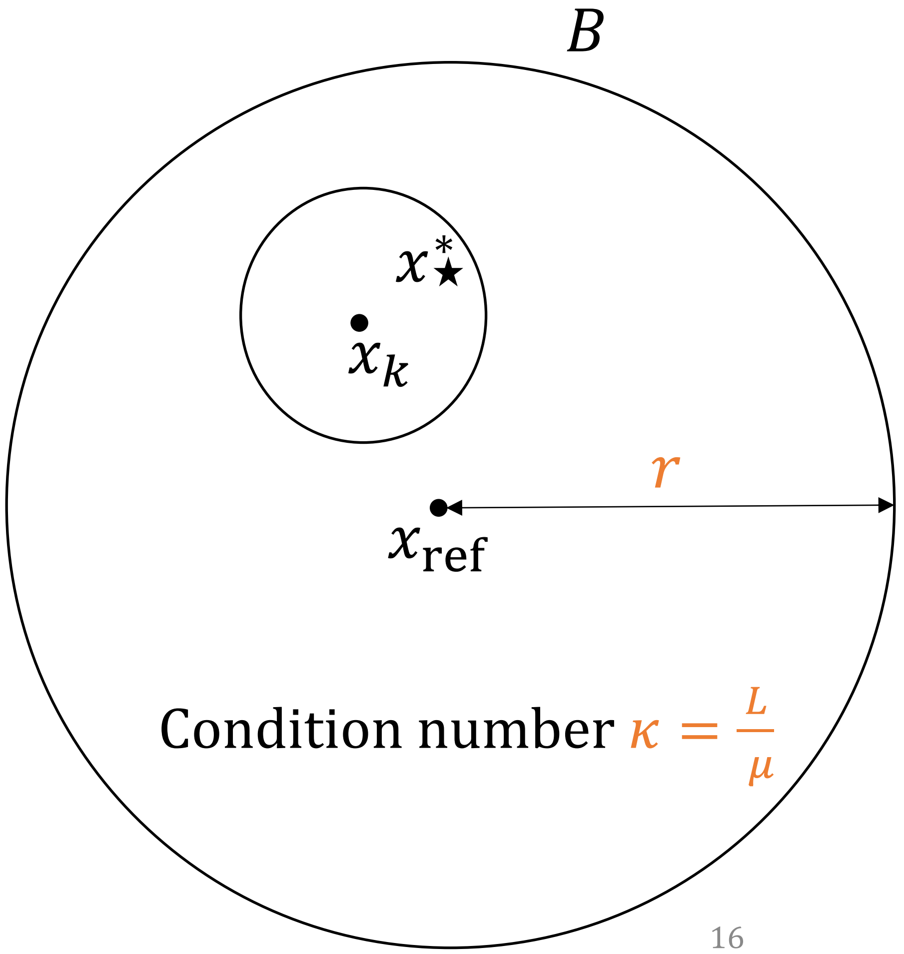
You know:

- f is L-smooth in B and μ-strongly g-convex in $\mathcal{M}$;
- f has a unique minimizer x^* in B.

You can query an oracle at x to get $f(x), \nabla f(x)$ (exact info, no noise).

Task: find a ball of radius $r/5$ containing x^*.

Least number of oracle queries necessary?



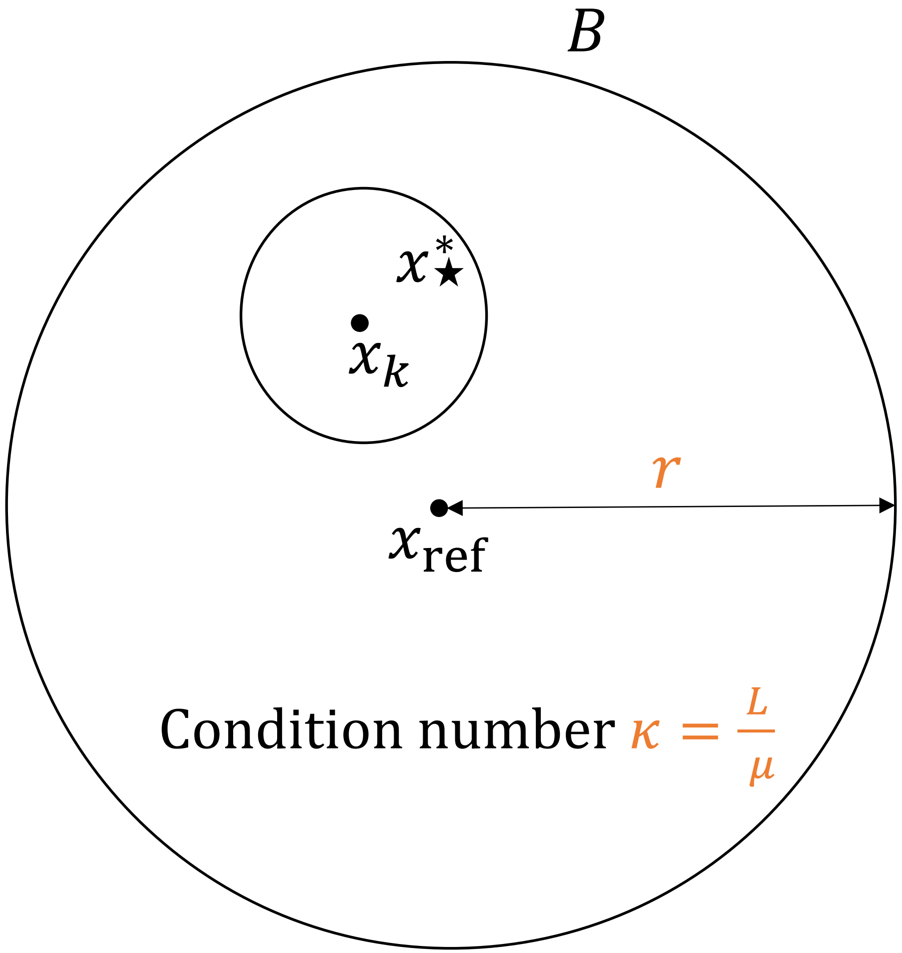
What happens in $\mathbb{R}^d$?

If $\mathcal{M} = \mathbb{R}^d$:

Gradient Descent (GD)

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

$O(\kappa)$ oracle queries.



What happens in $\mathbb{R}^d$?

If $\mathcal{M} = \mathbb{R}^d$:

Gradient Descent (GD)

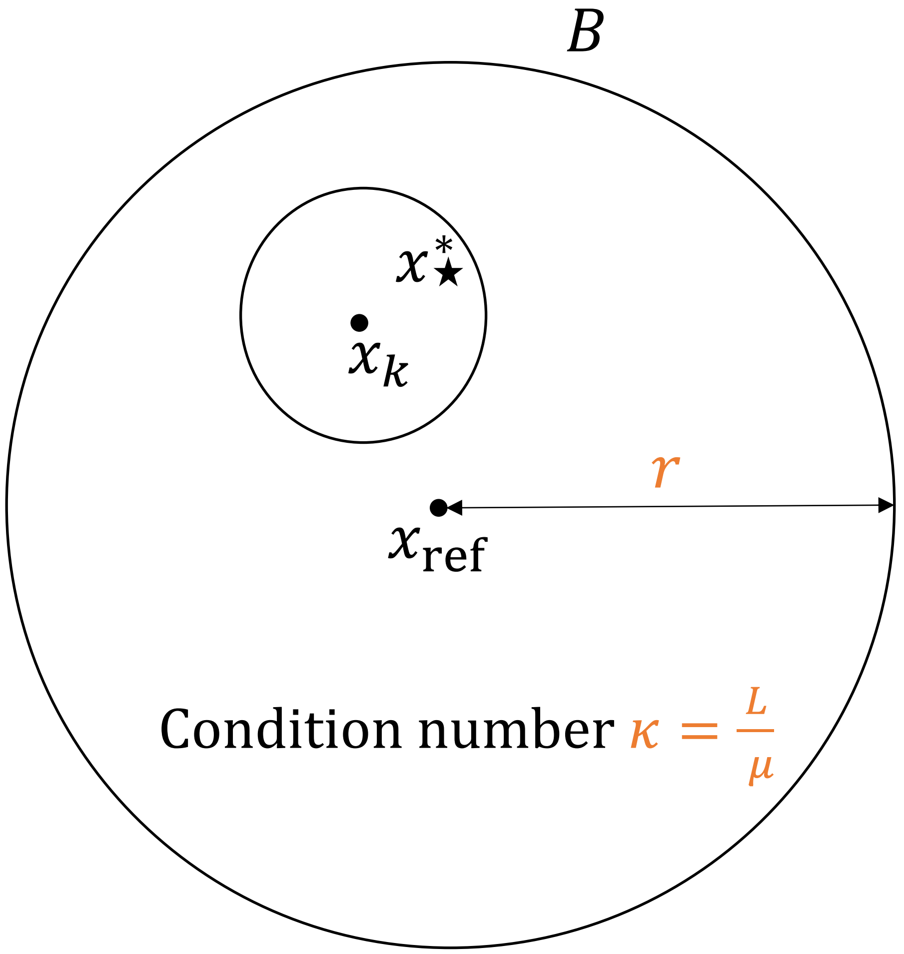
$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

$O(\kappa)$ oracle queries.

Nesterov's Accelerated Gradient method (NAG)

$$y_k = x_k + (1-\theta)v_k$$
$$x_{k+1} = y_k - \eta \nabla f(y_k)$$
$$v_{k+1} = x_{k+1} - x_k$$

$\tilde{O}(\sqrt{\kappa})$ oracle queries.



What happens in $\mathbb{R}^d$?

If $\mathcal{M} = \mathbb{R}^d$:

Gradient Descent (GD)

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

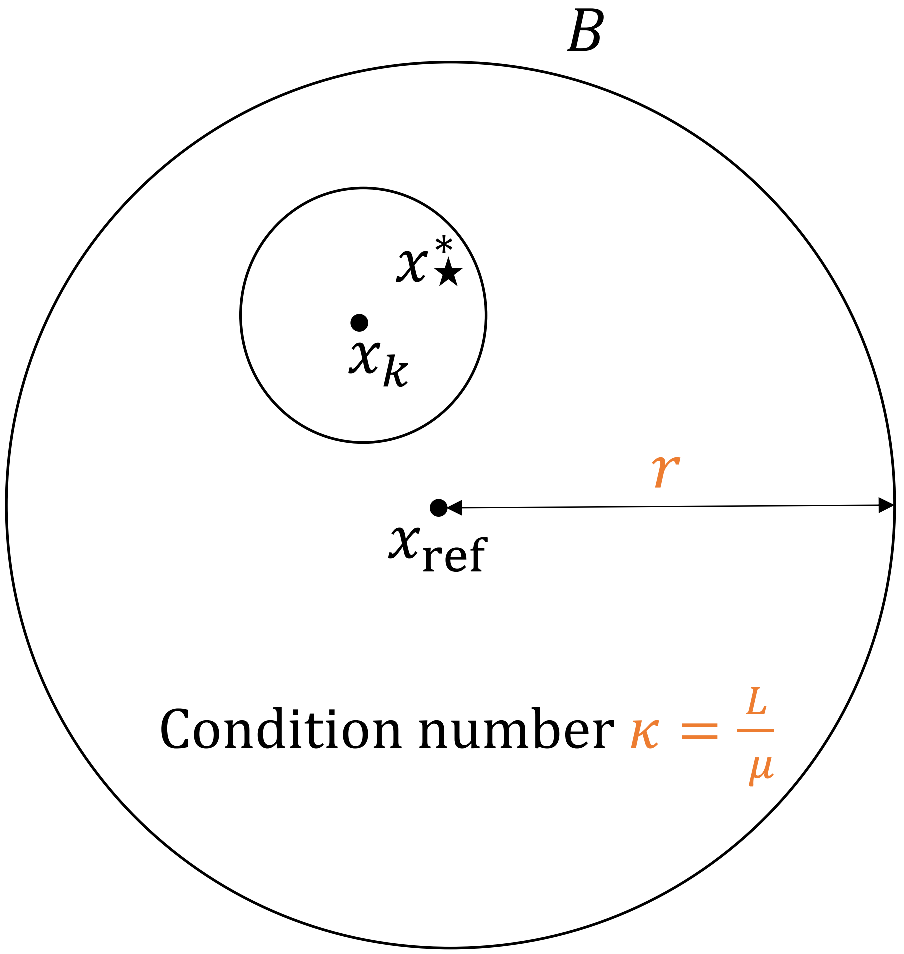
$O(\kappa)$ oracle queries.

Nesterov's Accelerated Gradient method (NAG)

$$y_k = x_k + (1-\theta) v_k$$
$$x_{k+1} = y_k - \eta \nabla f(y_k)$$
$$v_{k+1} = x_{k+1} - x_k$$

$\tilde{O}(\sqrt{\kappa})$ oracle queries.

NAG has optimal oracle complexity; GD does not.



Optimal methods

What about on Riemannian manifolds?

Riemannian GD (RGD) requires $O(\kappa)$ oracle queries (when for example $\mathcal{M}$ is a hyperbolic space).

$$x_{k+1} = \exp_{x_k}(-\eta \operatorname{grad} f(x_k))$$

Optimal methods

What about on Riemannian manifolds?

Riemannian GD (RGD) requires $O(\kappa)$ oracle queries (when for example $\mathcal{M}$ is a hyperbolic space).

$$x_{k+1} = \exp_{x_k}(-\eta \operatorname{grad} f(x_k))$$

Is there an algorithm using only $\tilde{O}(\sqrt{\kappa})$ queries in general?

Optimal methods

What about on Riemannian manifolds?

Riemannian GD (RGD) requires $O(\kappa)$ oracle queries (when for example $\mathcal{M}$ is a hyperbolic space).

$$x_{k+1} = \exp_{x_k}(-\eta \operatorname{grad} f(x_k))$$

Is there an algorithm using only $\tilde{O}(\sqrt{\kappa})$ queries in general?

Partial positive result (Zhang, Ahn, Sra, Martinez-Rubio, Alimisis, et al.): you can accelerate in some cases (e.g., r small).

Main results

Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ whose sectional curvatures are in the interval $[\mathrm{K}_{lo}, \mathrm{K}_{up}]$ with $\mathrm{K}_{up} < 0$.

Let $r = c_2\, \kappa \,/\sqrt{-K_{lo}}$.

For hyperbolic spaces, $K_{lo} = K_{up} = K < 0$

Main results

Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ whose sectional curvatures are in the interval $[\mathrm{K}_{lo}, \mathrm{K}_{up}]$ with $\mathrm{K}_{up} < 0$.

Let $r = c_2 \kappa / \sqrt{-K_{lo}}$.

For every deterministic algorithm $\mathcal{A}$, there is a C^∞ function f which is

- 1-strongly g-convex in all of $\mathcal{M}$;
- κ-smooth in the geodesic ball $B(x_{\text{origin}}, r)$;
- and has (unique) minimizer in $B(x_{\text{origin}}, r)$;

Main results

Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ whose sectional curvatures are in the interval $[\mathrm{K}_{lo}, \mathrm{K}_{up}]$ with $\mathrm{K}_{up} < 0$.
Let $r = c_2\, \kappa / \sqrt{-K_{lo}}$.

For every deterministic algorithm $\mathcal{A}$, there is a C^∞ function f which is

- 1-strongly g-convex in all of $\mathcal{M}$;
- κ-smooth in the geodesic ball $B(x_{\text{origin}}, r)$;
- and has (unique) minimizer in $B(x_{\text{origin}}, r)$;

such that algorithm $\mathcal{A}$ requires at least

$$\Omega\left(\sqrt{\frac{K_{up}}{K_{lo}}}\frac{\kappa}{\log \kappa}\right)$$

queries in order to find a point $x \in \mathcal{M}$ within $r/5$ of the minimizer of f.

Main results

Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ whose sectional curvatures are in the interval $[\mathrm{K}_{lo}, \mathrm{K}_{up}]$ with $\mathrm{K}_{up} < 0$.
Let $r = c_2\, \kappa / \sqrt{-K_{lo}}$.

For every deterministic algorithm $\mathcal{A}$, there is a C^{∞} function f which is

- 1-strongly g-convex in all of $\mathcal{M}$;
- κ-smooth in the geodesic ball $B(x_{\text{origin}}, r)$;
- and has (unique) minimizer in $B(x_{\text{origin}}, r)$;

such that algorithm $\mathcal{A}$ requires at least

$$\Omega\left(\sqrt{\frac{K_{up}}{K_{lo}}\frac{\kappa}{\log \kappa}}\right) \Longrightarrow$$

$O(\sqrt{\kappa})$ rate is impossible;
RGD is optimal (up to log).

queries in order to find a point $x \in \mathcal{M}$ within $r/5$ of the minimizer of f.

Other settings

$n \times n$ positive definite matrices with affine-invariant metric.

Smooth nonstrongly g-convex optimization ($\mu = 0$).

Nonsmooth g-convex optimization.

Negative curvature

Geodesic ball volumes grow much faster in negatively curved spaces than flat spaces.

Negative curvature

Geodesic ball volumes grow much faster in negatively curved spaces than flat spaces.

It’s harder to find a point in a ball just because there's so much more space to explore.

Negative curvature

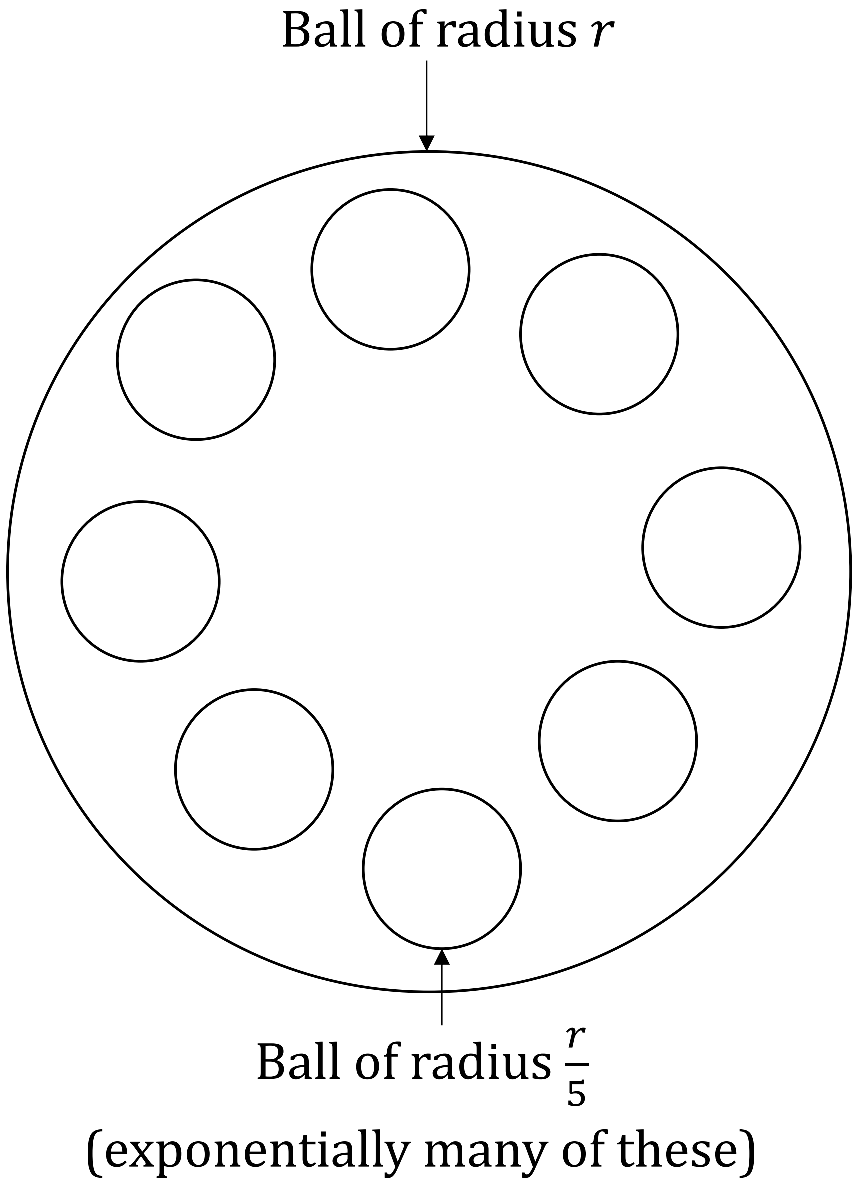
Geodesic ball volumes grow much faster in negatively curved spaces than flat spaces.

It’s harder to find a point in a ball just because there's so much more space to explore.

How many disjoint balls of radius $r/5$ contained in every ball of radius r?

$e^{\Theta(rd)}$ in hyperbolic space

$e^{\Theta(d)}$ in Euclidean space



Future directions

Tighter upper/lower bounds, e.g., Kim and Yang (2022)

Randomized algorithms which receive exact information?

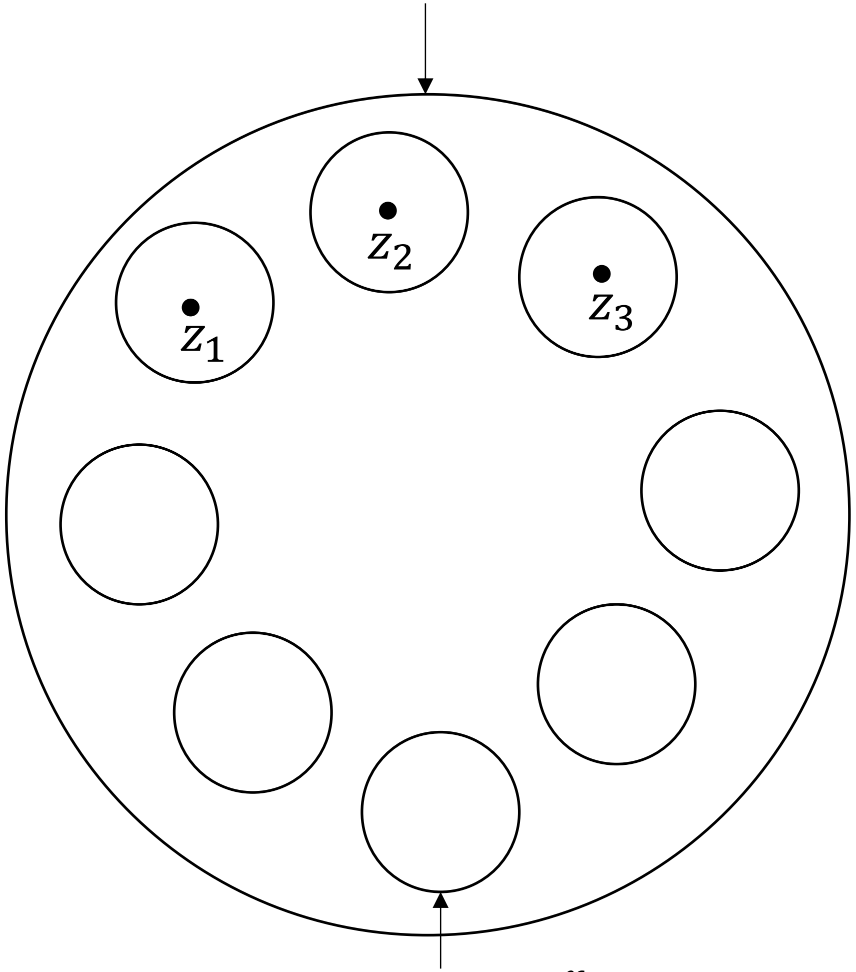
Ellipsoid method?

Interior-point methods?

Proof technique

Proof technique

Ball of radius r



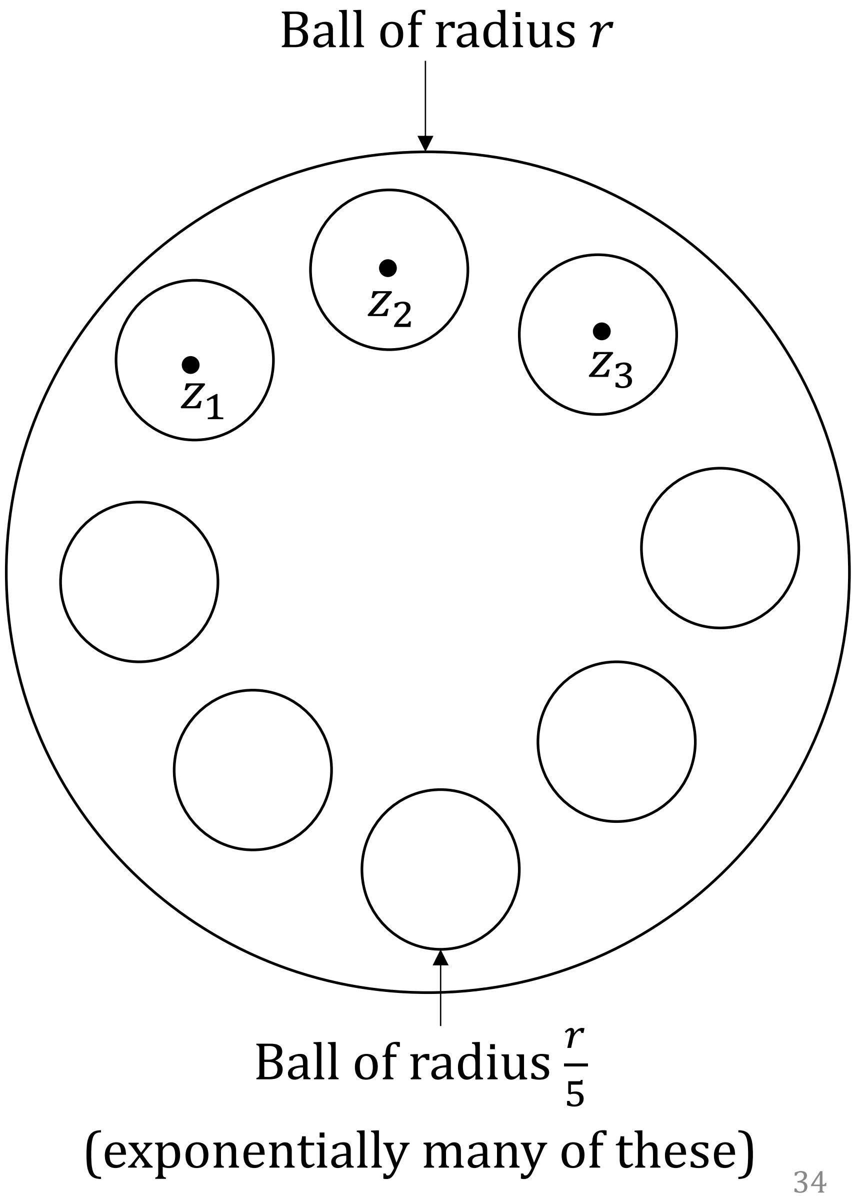
Ball of radius $\frac{r}{5}$
(exponentially many of these)

Proof technique

Hamilton and Moitra consider the functions

$$x \mapsto \frac{1}{2}\,\mathrm{dist}\left(x, z_j\right)^2, j = 1, \ldots, N$$

Show that in expectation (over noisiness of queries), any algorithm makes at most limited progress per query.



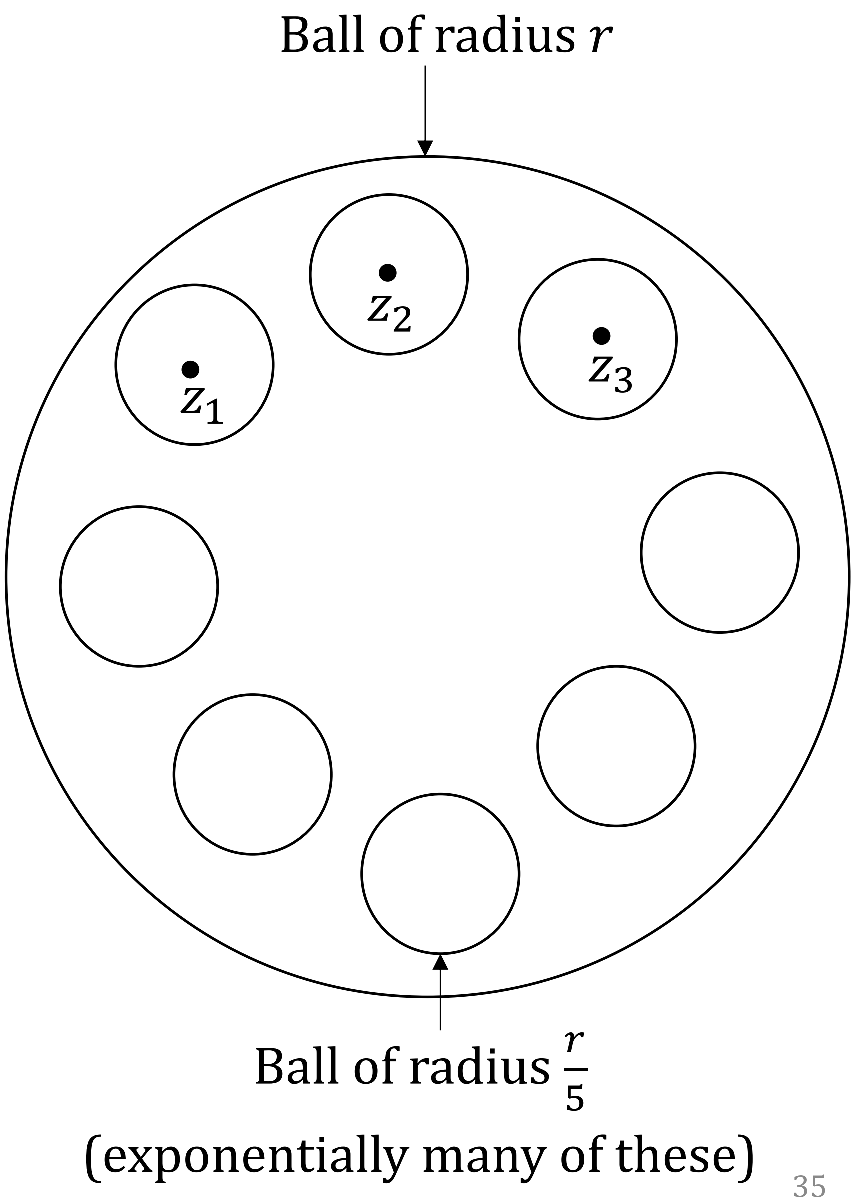
Proof technique

Hamilton and Moitra consider the functions

$$x \mapsto \frac{1}{2}\mathrm{dist}(x, z_j)^2, j = 1, \ldots, N$$

Gradients of these functions point directly towards the minimizer

- Ok if there is noise
- A problem if queries are exact



Proof technique

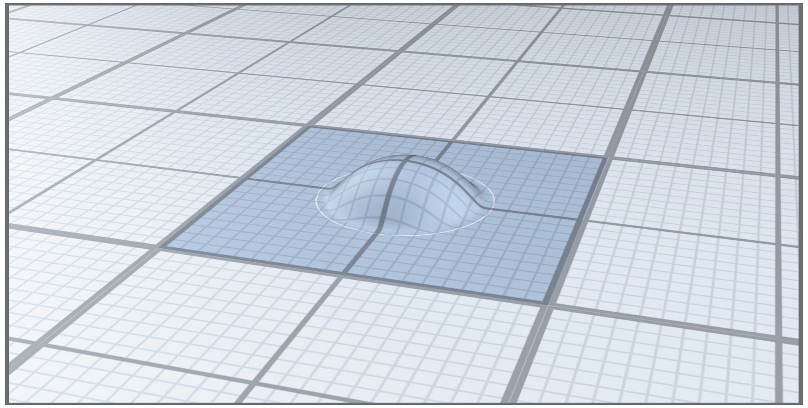
Our solution:

The hard functions we consider are squared distance functions plus a perturbation

$$x \mapsto \frac{1}{2}\operatorname{dist}(x, z_j)^2 + H_{j,k}(x), \qquad \|\operatorname{Hess} H_{j,k}(x)\| \leq \frac{1}{2}.$$

For any algorithm, the perturbation $H_{j,k}$ is constructed adversarially using a resisting oracle.

Proof technique



Our solution:

Perturbation is a sum of bump functions

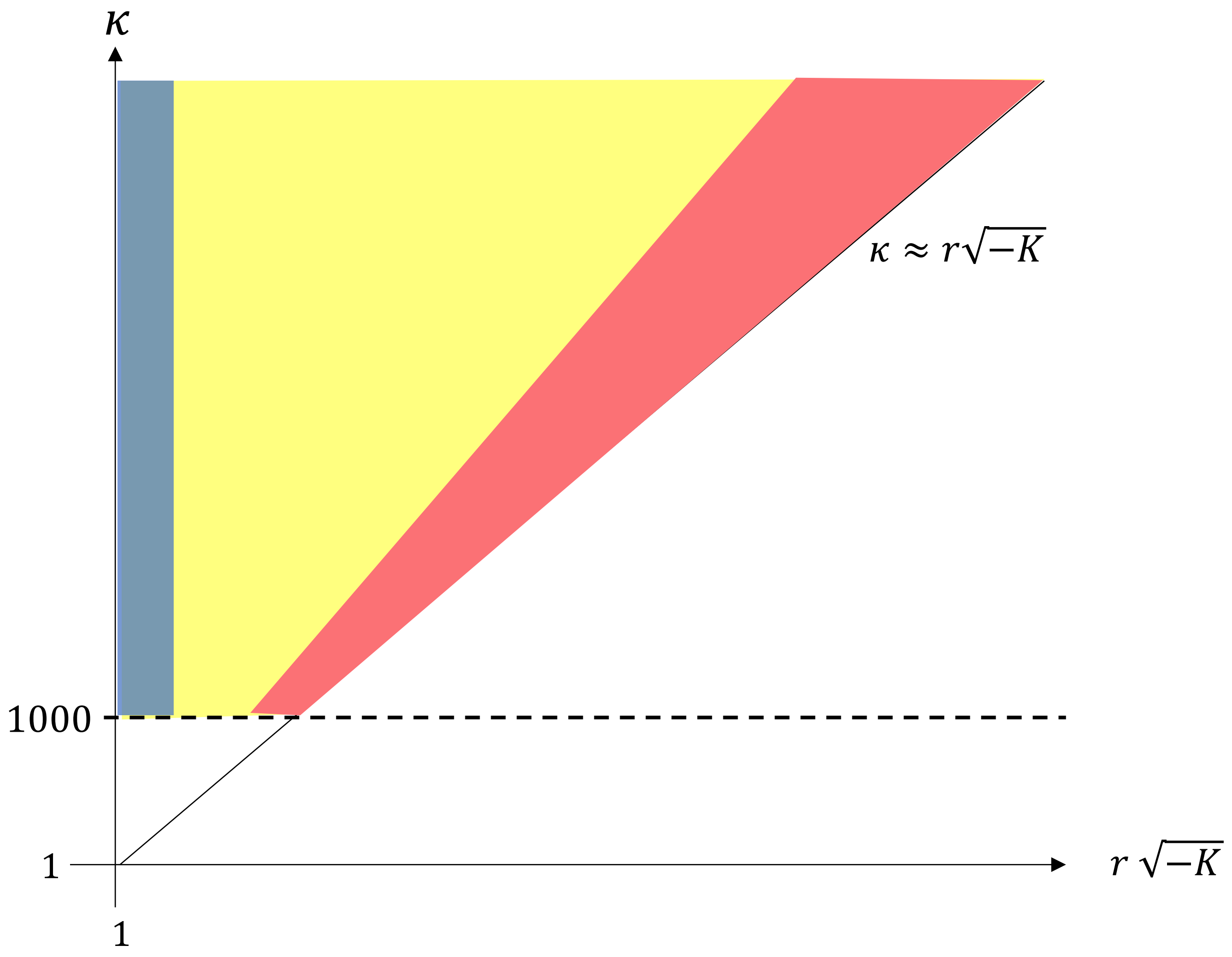
$$\mathrm{H}_{j,k}(x) = \sum_{m=1}^{k} h_{j,m}$$

One bump function $h_{j,m}$ is added for each query made by the algorithm.

Support of the bump $h_{j,m}$ is centered at the the query x_m.

Appendix

What we know (for hyperbolic spaces)



Main results

$n \times n$ positive definite matrices with affine-invariant metric

Main results

$n{\times}n$ positive definite matrices with affine-invariant metric

It is Hadamard, but does not satisfy assumptions of previous theorem: sectional curvature can be zero.

Main results

$n \times n$ positive definite matrices with affine-invariant metric

It is Hadamard, but does not satisfy assumptions of previous theorem: sectional curvature can be zero.

Still, can prove the lower bound $\Omega\left(\frac{1}{n}\frac{\kappa}{\log \kappa}\right)$.

Nonstrongly g-convex case

Can also show that acceleration is impossible in the nonstrongly g-convex case ($\mu = 0$).

Nonstrongly g-convex case

Can also show that acceleration is impossible in the nonstrongly g-convex case ($\mu = 0$).

Have the lower bound $\Omega\left(\frac{1}{\epsilon} \cdot \frac{1}{\log^3(\epsilon^{-1})}\right)$ for finding a point x with $f(x) - f(x^*) \leq \epsilon$.

Means a version of RGD is optimal.

Nonstrongly g-convex case

Can also show that acceleration is impossible in the nonstrongly g-convex case ($\mu = 0$).

Have the lower bound $\Omega\left(\frac{1}{\epsilon} \cdot \frac{1}{\log^3(\epsilon^{-1})}\right)$ for finding a point x with $f(x) - f(x^*) \le \epsilon$.

Means a version of RGD is optimal.

Compare with NAG, which uses at most $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ queries in Euclidean spaces.

Applications

- Fréchet mean (intrinsic averaging on Hadamard spaces) (e.g., Karcher)
- Gaussian mixture models (Hosseini + Sra)
- Optimistic likelihoods for Gaussians (Nguyen et al.)
- Robust Covariance estimation (Weisel + Zhang, Franks + Moitra)
- Metric learning (Zadeh et al.)
- Variants on PCA (Tang + Allen) [MLEs for matrix normal models]
- Operator/tensor scaling (Allen Zhu et al., Burgisser et al.)
    - Brascamp-Lieb constants, computational complexity, polynomial identity testing, hardness of robust subspace recovery, etc.
- Tree-like embeddings (Bacak)
- Sampling on Riemannian manifolds (Goyal + Shetty)
- Landscape analysis (e.g., Ahn + Suarez)

Application: robust covariance estimation

IID samples $x_i \in \mathbb{R}^p, i = 1, \ldots, n$, coming from an elliptical distribution:

$$x \sim u\, \Sigma^{1/2} v$$

where $\Sigma \succ 0$ is fixed (the shape matrix), u is a scalar r.v., and $v \sim \mathbb{S}^{p-1}$.

Application: robust covariance estimation

IID samples $x_i \in \mathbb{R}^p, i = 1, \ldots, n$, coming from an elliptical distribution:

$$x \sim u\,\Sigma^{1/2} v$$

where $\Sigma \succ 0$ is fixed (the shape matrix), u is a scalar r.v., and $v \sim \mathbb{S}^{p-1}$.

Tyler's M-estimator for the shape matrix:

$$\hat{\Sigma} = \underset{\Sigma \succ 0,\ \mathrm{Tr}(\Sigma) = p}{\operatorname{argmin}} \frac{p}{n} \sum_{i=1}^{n} \log\left(x_i^\top \Sigma^{-1} x_i\right) + \log\det(\Sigma)$$

Can also be derived as an MLE.

Application: robust covariance estimation

IID samples $x_i \in \mathbb{R}^p, i = 1, \ldots, n$, coming from an elliptical distribution:

$$x \sim u\, \Sigma^{1/2} v$$

where $\Sigma \succ 0$ is fixed (the shape matrix), u is a scalar r.v., and $v \sim \mathbb{S}^{p-1}$.

Tyler's M-estimator for the shape matrix:

$$\hat{\Sigma} = \operatorname*{argmin}_{\Sigma \succ 0,\ \mathrm{Tr}(\Sigma) = p} \frac{p}{n} \sum_{i=1}^{n} \log\left(x_i^\top \Sigma^{-1} x_i\right) + \log\det(\Sigma)$$

Is g-convex for PD matrices (with affine-invariant metric).

→ new algorithms/analysis + analysis for Tyler's iterative procedure

Application: robust covariance estimation

IID samples $x_i \in \mathbb{R}^p, i = 1, \dots, n$, coming from an elliptical distribution:

$$x \sim u\, \Sigma^{1/2} v$$

where $\Sigma \succ 0$ is fixed (the shape matrix), u is a scalar r.v., and $v \sim \mathbb{S}^{p-1}$.

Tyler's M-estimator for the shape matrix:

$$\hat{\Sigma} = \underset{\Sigma \succ 0,\ \mathrm{Tr}(\Sigma)=p}{\operatorname{argmin}} \frac{p}{n} \sum_{i=1}^{n} \log\left(x_i^\top \Sigma^{-1} x_i\right) + \log\det(\Sigma)$$

Is a specific instance of the operator scaling problem.

Sources: Weisel + Zhang, Franks + Moitra