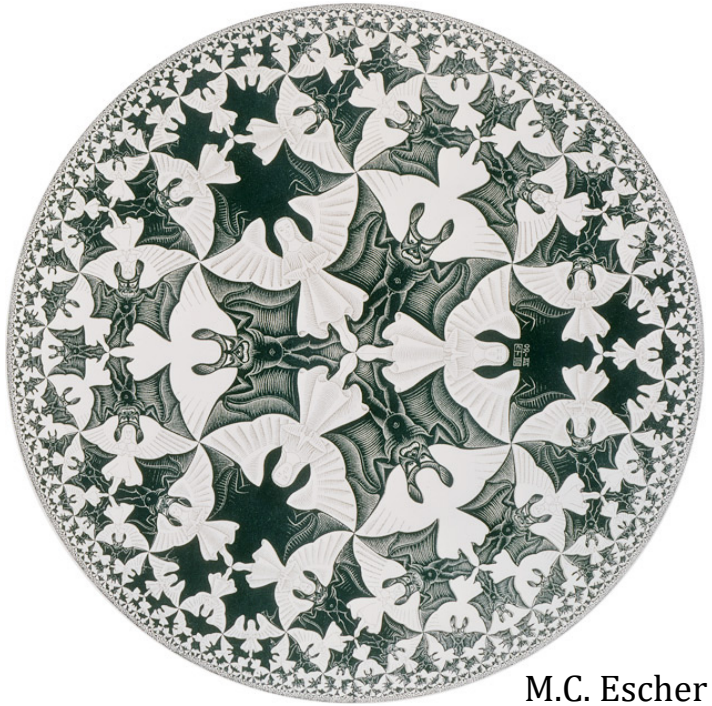
A bumpy start

Why we can't *quite* accelerate gradient methods on negatively curved spaces

October 18, 2022

Nicolas Boumal, with Chris Criscitiello

OPTIM, Institute of Mathematics, EPFL



M.C. Escher

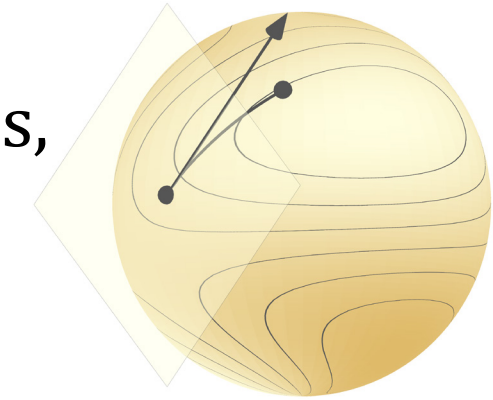
Optimization from Euclid to Riemann: Fifty years of mostly smooth sailing

$$\min_x f(x)$$

If f is smooth on a Euclidean space $\mathcal{E}$, we can design algorithms to (try to) minimize f using $\nabla f, \nabla^2 f, \ldots$

If f is instead defined on a Riemannian manifold $\mathcal{M}$, we still have gradients and Hessians. Plenty of applications.

Many classical Euclidean algorithms generalize to manifolds, often with essentially the same guarantees and limitations.



From Euclid to Riemann: non-convex f

Gradient descent finds ε-critical points in $\sim \frac{1}{\varepsilon^2}$ iterations assuming Lipschitz ∇f.

Euclid: folklore. Riemann: Boumal, Absil & Cartis '18. It's tight: Carmon, Duchi, Hinder & Sidford '19.

Accelerated gradient does so in $\sim \frac{1}{\varepsilon^{1.75}}$ iterations assuming also Lipschitz $\nabla^2 f$.

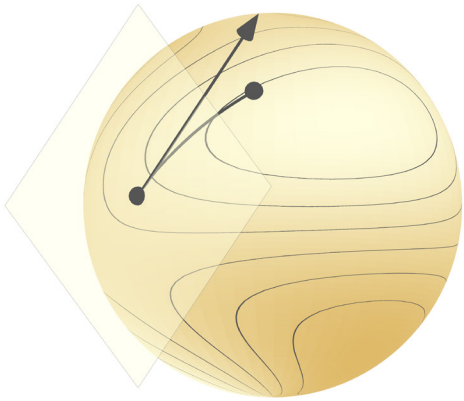
Euclid: Carmon++ '17; Jin, Netrapalli & Jordan '18. Riemann: Criscitiello & Boumal '22. Almost tight: Carmon++ '19.

Cubic regularization does so in $\sim \frac{1}{\varepsilon^{1.5}}$ by querying $\nabla^2 f$ (all of it).

Euclid: Nesterov & Polyak '08. Riemann: Agarwal, Boumal, Bullins & Cartis '20. It's tight: Carmon++ '19.

And more: trust regions, SGD, BFGS, nonlinear CG, …

Curvature affects constants, but only modest qualitative changes.



From Euclid to Riemann: convex f

Gradient descent finds ε-critical points in $\sim \frac{1}{\varepsilon}$ iterations assuming Lipschitz ∇f.

Euclid: folklore. Riemann: Zhang & Sra '16. It's tight: folklore.

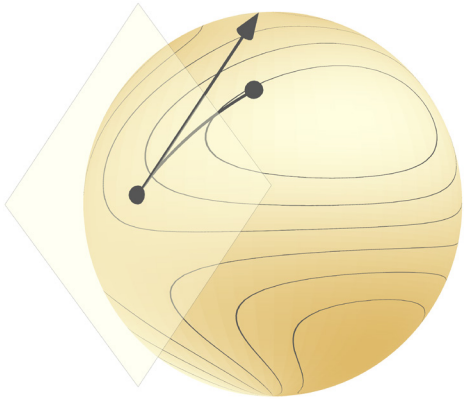
Gradient descent does so in $\sim \kappa \log \frac{1}{\varepsilon}$ iterations if f is also strongly convex.

Euclid: folklore. Riemann: Zhang & Sra '16 + details in Criscitiello & B. '22, App. K. It's the right rate *for GD*: folklore.

On Euclidean space, this can be accelerated to $\sim \sqrt{\kappa} \log \frac{1}{\varepsilon}$ ...

Euclid: Nesterov. It's tight: Nemirovski & Yudin.

... But on Riemannian manifolds: no result *quite* the same.



Template for what we might have wanted

We consider some Riemannian manifold $\mathcal{M}$.

Let $\mathcal{F}_\kappa$ be the class of functions on $\mathcal{M}$ with condition number κ, that is, $f \in C^\infty(\mathcal{M})$ has L-Lipschitz ∇f, is μ-strongly convex and $\kappa = L/\mu$.

An algorithm queries f and ∇f at $x_0, x_1, x_2, \ldots$ adaptively.

Aspirational Theorem. There exists an algorithm such that, for all $f \in \mathcal{F}_\kappa$, given $x_0 \in \mathcal{M}$, there is $k \lessapprox \sqrt{\kappa}$ s.t. $\operatorname{dist}(x_k, x^\star) \leq \frac{\operatorname{dist}(x_0, x^\star)}{5}$.

Mathematics > Optimization and Control

[Submitted on 14 Jan 2021 (v1), last revised 17 Jan 2021 (this version, v2)]

No-go Theorem for Acceleration in the Hyperbolic Plane

Linus Hamilton, Ankur Moitra

In recent years there has been significant effort to adapt the key tools and ideas in convex optimization to the Riemannian setting. One key challenge has remained: Is there a Nesterov-like accelerated gradient method for geodesically convex functions on a Riemannian manifold? Recent work has given partial answers and the hope was that this ought to be possible. Here we dash these hopes. We prove that in a noisy setting, there is no analogue of accelerated gradient descent for geodesically convex functions on the hyperbolic plane. Our results apply even when the noise is exponentially small. The key intuition behind our proof is short and simple: In negatively curved spaces, the volume of a ball grows so fast that information about the past gradients is not useful in the future.

Comments: 12 pages

Mathematics > Optimization and Control

[Submitted on 14 Jan 2021 (v1), last revised 17 Jan 2021 (this version, v2)]

No-go Theorem for Acceleration in the Hyperbolic Plane

Linus Hamilton, Ankur Moitra

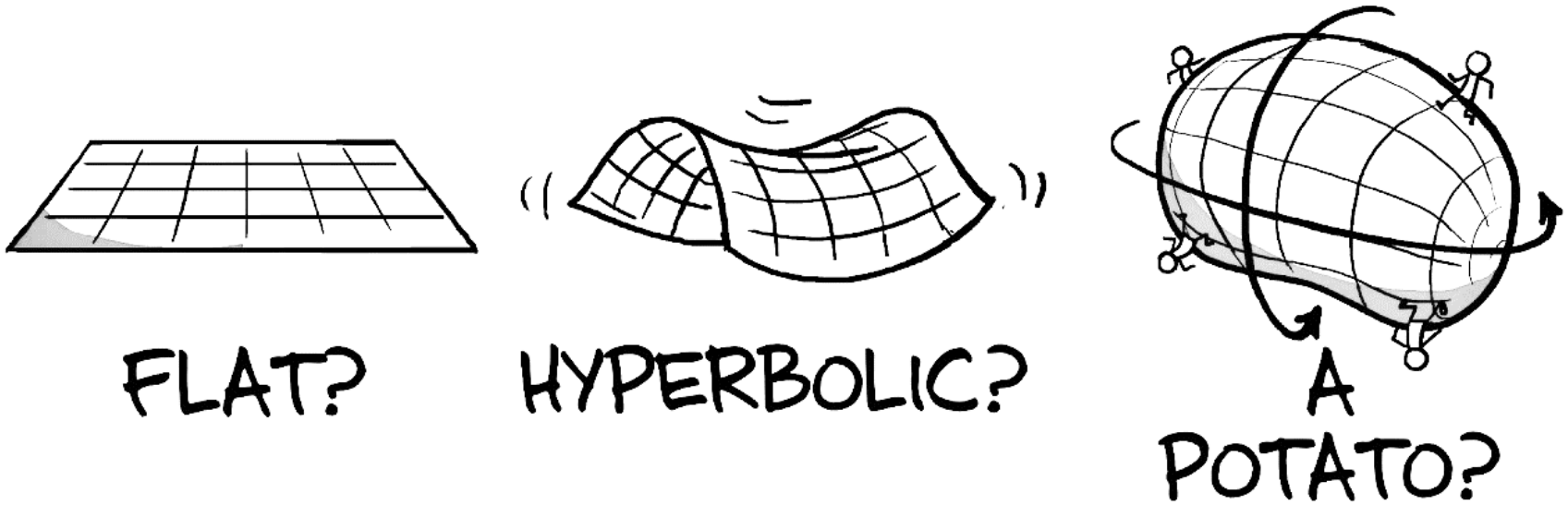
In recent years there has been significant effort to adapt the key tools and ideas in convex optimization to the Riemannian setting. One key challenge has remained: Is there a Nesterov-like accelerated gradient method for geodesically convex functions on a Riemannian manifold? Recent work has given partial answers and the hope was that this ought to be possible. Here we dash these hopes. We prove that in a noisy setting, there is no analogue of accelerated gradient descent for geodesically convex functions on the hyperbolic plane. Our results apply even when the noise is exponentially small. The key intuition behind our proof is short and simple: In negatively curved spaces, the volume of a ball grows so fast that information about the past gradients is not useful in the future.

Comments: 12 pages

* and the algorithm can only query in some bounded domain

Does your space support geodesically convex f?

f is g-convex if $f \circ \gamma$ is convex for all geodesic curves γ.



Thus, we focus on Hadamard manifolds:

Curvature ≤ 0, complete, simply connected. E.g.: hyperbolic; SPD.

PhDcomics.com

Smooth & strongly g-convex $f \in C^\infty(\mathcal{M})$

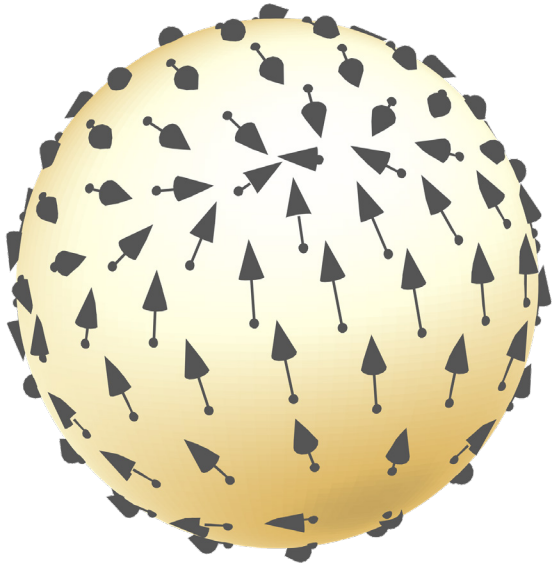
∇f is a vector field on $\mathcal{M}$ such that $\nabla f(x) \in \mathrm{T}_x\mathcal{M}$ is steepest ascent direction.

$\nabla^2 f(x) \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$ is the (Riemannian) derivative of ∇f at x. It's symmetric.

∇f is L-Lipschitz continuous and f is μ-strongly g-convex if, for all x,

$$\mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq L I$$

The catch: this cannot hold on *all* of $\mathcal{M}$ if curvature is negative.



Effects of negative curvature: take one

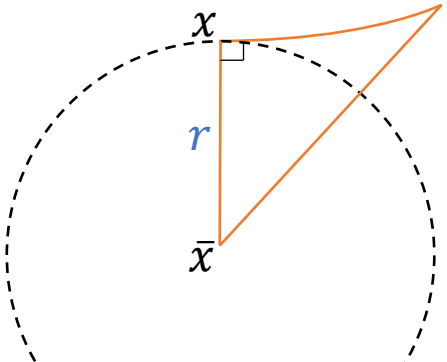
Pythagoras is only an inequality: $c^2 \geq a^2 + b^2$.

Consider $f(x) = \frac{1}{2}\text{dist}(x, \bar{x})^2$.

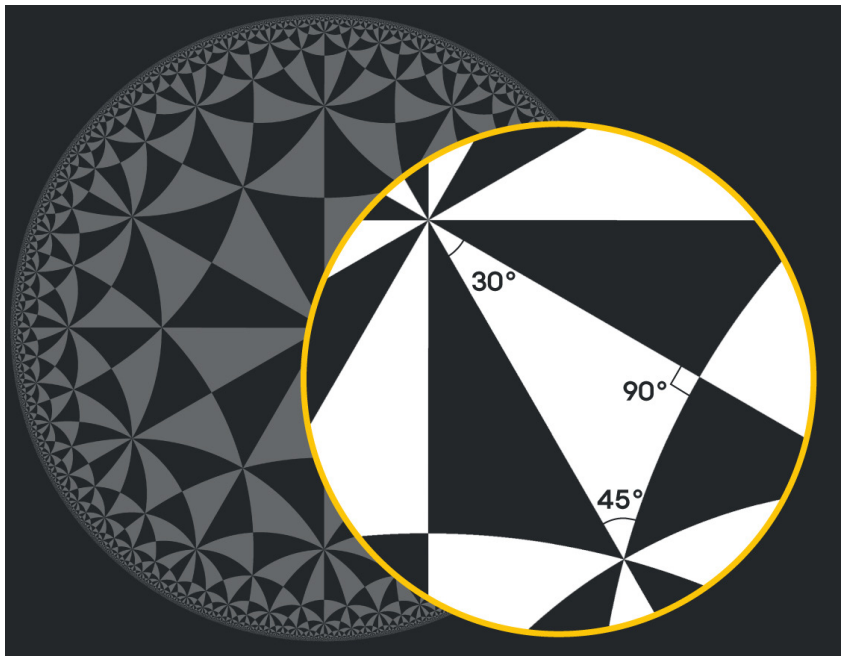
Euclidean space: $\nabla^2 f(x) = I$ so $\mu = L = 1$ and $\kappa = 1$.

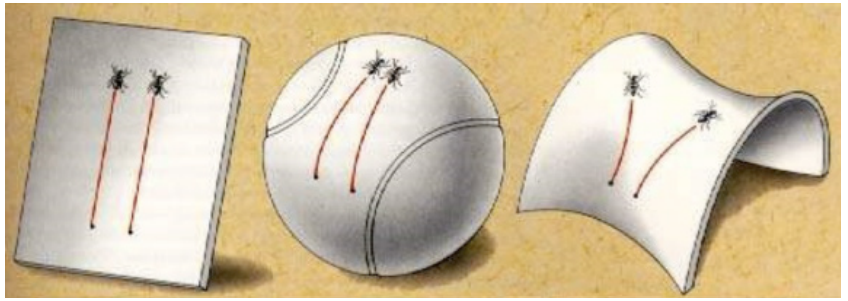
Hyperbolic space: $\mu = 1$ but $L = \frac{r}{\tanh r} \approx r$ so $\kappa \approx r$.

Fact: If f has L-Lipschitz gradient and is μ strongly g-convex on a ball of radius r (large), then $\kappa \geq \frac{1}{10} r$.



Quanta





Fact: If f has L-Lipschitz gradient and is μ strongly g-convex on a ball of radius r (large), then $\kappa \geq \frac{1}{10} r$.

Main theorem

Let $\mathcal{M}$ be hyperbolic space (curvature -1)—for this talk.

$f \in C^\infty(\mathcal{M})$ is in function class $\mathcal{F}_{\kappa,r,\bar{x}}$ if it is μ-strongly convex on $\mathcal{M}$, its gradient is L-Lipschitz on $B(\bar{x}, r)$, $\kappa = L/\mu$ and $\operatorname{dist}(\bar{x}, x^\star) \leq \frac{3}{4} r$.

Consider any algorithm that queries f and ∇f at $x_0, x_1, x_2, \ldots$

Theorem. For any $\kappa \geq 1000$, set r such that $\kappa = 12r + 9$.

There exists $f \in \mathcal{F}_{\kappa,r,\bar{x}}$ s.t. $\operatorname{dist}(x_k, x^\star) > \frac{r}{5}$ for all $k < \dfrac{\kappa}{1000 \log(10\kappa)}$.

Fact: If f has L-Lipschitz gradient and is μ strongly g-convex on a ball of radius r (large), then $\kappa \geq \frac{1}{10} r$.

Main theorem: a few comments

We can't pick $r \gg \kappa$ as otherwise $\mathcal{F}$ is empty.

We could pick $r \ll \kappa$ – but not *too* small, otherwise curvature fades.

Some algorithms achieve *eventual or local acceleration* [Zhang & Sra '18; Ahn & Sra '20; Martinez-Rubio '21; Alimisis, Orvieto, Bécigneul & Lucchi '21].

Promising recent work targets $1 \ll r \ll \kappa$ (say, $r \approx \sqrt{\kappa}$).

Leads: [Kim & Yang '22], and a recent follow-up by others on OpenReview.

Theorem. For any $\kappa \geq 1000$, set r such that $\kappa = 12r + 9$.
There exists $f \in \mathcal{F}_{\kappa, r, \bar{x}}$ s.t. $\operatorname{dist}(x_k, x^\star) > \frac{r}{5}$ for all $k < \frac{\kappa}{1000 \log(10\kappa)}$.

Proof technique: a resisting oracle

The algorithm queries on oracle for $f(x_k)$ and $\nabla f(x_k)$, $k = 0,1,2 \ldots$

Its aim: find x_k such that $\mathrm{dist}(x_k, x^\star) \leq \frac{r}{5}$, as fast as possible.

The oracle replies in such a way that there exists a compatible $f \in \mathcal{F}$.

Its aim: for as long as possible, ensure existence of *two* compatible functions in $\mathcal{F}$ whose minimizers are more than $2\frac{r}{5}$ apart.

Starting point: pick many $f_i \in \mathcal{F}$ whose minimizers are $\geq \frac{r}{2}$ apart.

Starting point: pick many $f_i \in \mathcal{F}$ whose minimizers are $\geq \frac{r}{2}$ apart.

Effects of negative curvature: take two

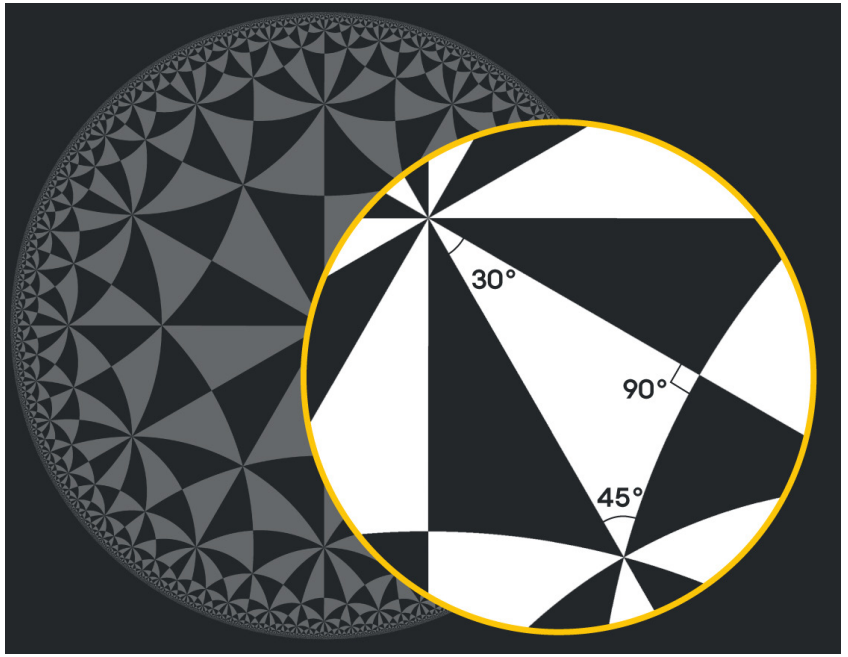
Balls have volume exponential in their radius.

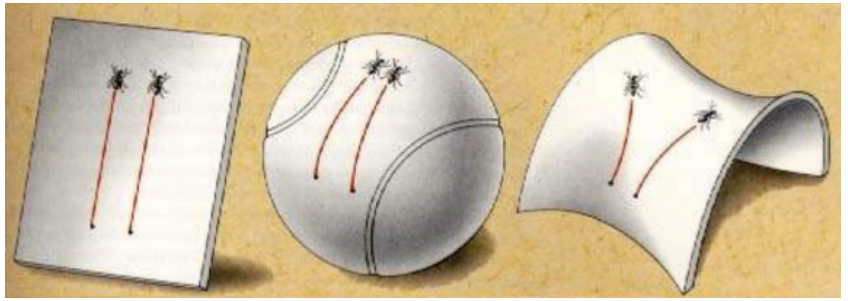
Intuitively, this is because geodesics diverge exponentially.

How many balls of radius $\frac{r}{4}$ fit in a ball of radius r?

Euclidean space $\mathbf{R}^d$: $\sim \frac{r^d}{\left(\frac{r}{4}\right)^d} = 4^d$ independent of r

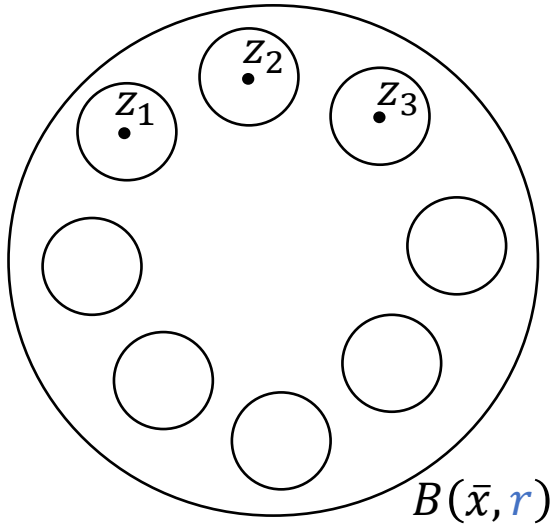
Hyperbolic plane: $\sim \frac{e^r}{e^{r/4}} = e^{\frac{3}{4}r}$ exponential in r





Starting point: pick many $f_i \in \mathcal{F}$ whose minimizers are $\geq \frac{r}{2}$ apart.

(Part of) Hamilton & Moitra's insight



We can pack $N \approx e^r$ balls of radius $\frac{r}{4}$ in the ball $B(\bar{x}, r)$.

Their centers $z_1, \ldots, z_N$ satisfy $\text{dist}(z_i, z_j) \geq \frac{r}{2} > 2\frac{r}{5}$ for all $i \neq j$.

For each z_i, let $f_i(x) = \frac{1}{2}\text{dist}(x, z_i)^2$. Note: $f_i \in \mathcal{F}_{\kappa, r, \bar{x}}$ with $\kappa \approx r$.

The oracle answers each query such that a fixed fraction remain compatible.

Since we start with exponentially many, at least 2 survive after $\sim r \approx \kappa$ queries.

How can the oracle keep many f_i compatible?

For each z_i, let $f_i(x) = \frac{1}{2}\mathrm{dist}(x, z_i)^2$. Note: $f_i \in \mathcal{F}_{\kappa,r,\bar{x}}$ with $\kappa \approx r$.

The oracle answers query x_k such that a fixed fraction remain compatible.

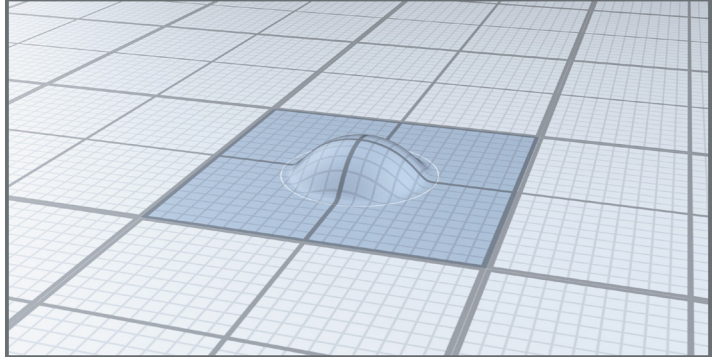
Hiccup: $-\nabla f_i(x_k) \in \mathrm{T}_{x_k}\mathcal{M}$ points directly at z_i (the minimizer!)

These gradients are all different… but many are similar.

H&M's approach: allow the oracle to return a *noisy* gradient $g \approx \nabla f_i(x_k)$.

Ours: *modify* many of the f_i so they have the *same* gradient at x_k: return that.

Bump functions instead of noise



At query x_k, we add bumps to many f_i.

Bumps have (small) compact support.

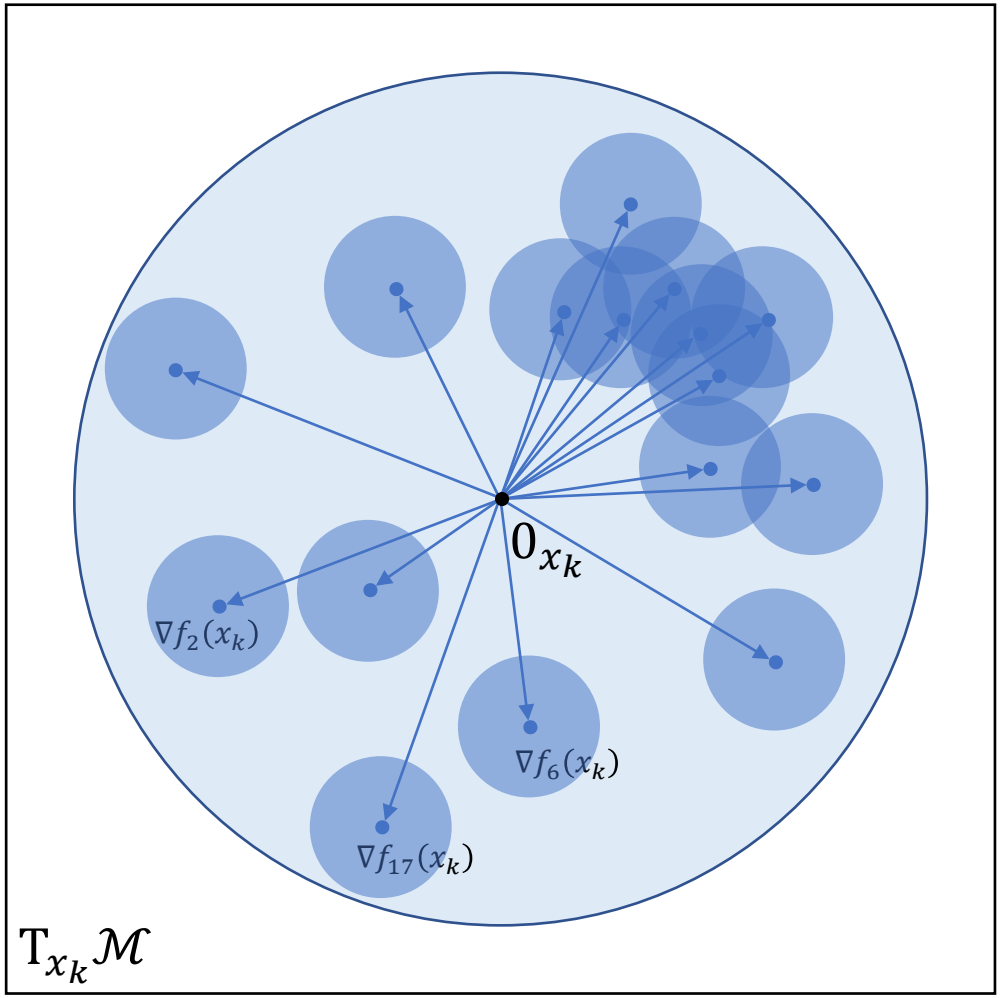
Allow us to steer gradients at x_k.

Bumps have bounded ∇^2 so f_i stay in $\mathcal{F}$.

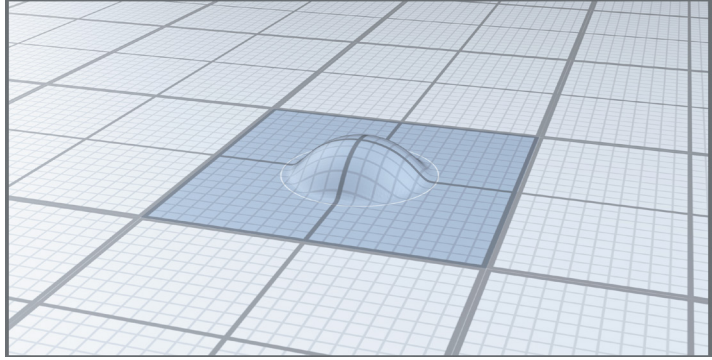
If x_k is far from $x_0, \ldots, x_{k-1}$, the f_i are pristine near x_k: can add sizable bump.

Technicality: assume $x_k \in B(\bar{x}, r)$ so we can bound $\|\nabla f_i(x_k)\|$.

Then pigeonhole. Then reduction.



Bump functions instead of noise



At query x_k, we add bumps to many f_i.

Bumps have (small) compact support.

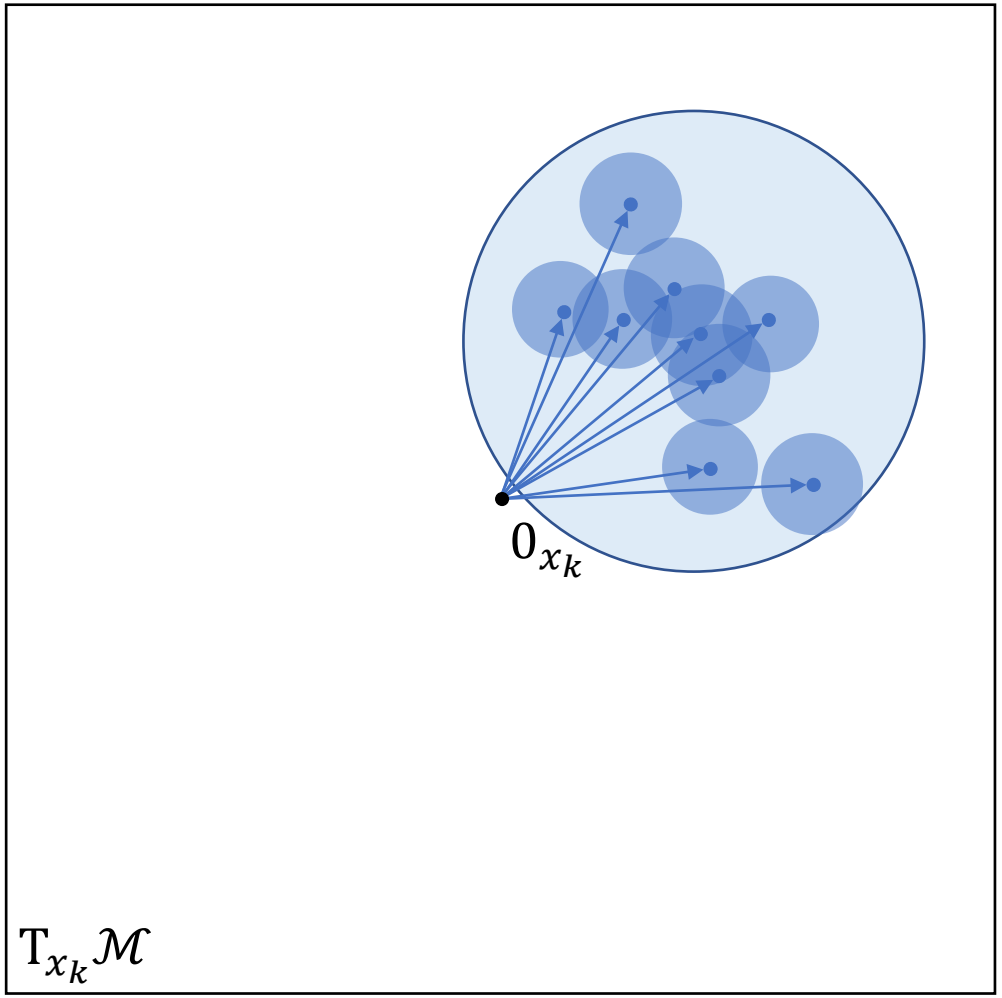
Allow us to steer gradients at x_k.

Bumps have bounded ∇^2 so f_i stay in $\mathcal{F}$.

If x_k is close to some x_ℓ with $\ell < k$, then be careful not to break past work.

Still fine because the f_i already have the same gradient at x_ℓ.

Smaller bumps suffice to align them at x_k.

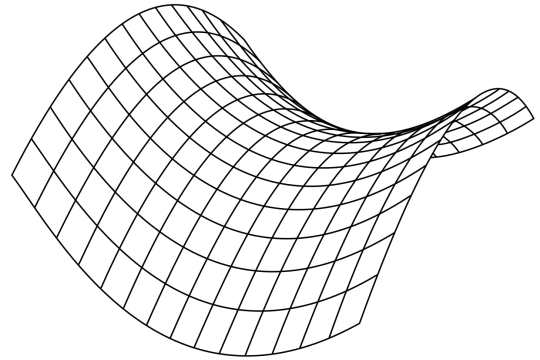


A meta comment

Resisting oracles require one to build many f's in the function class.
We tried and failed a lot.

E.g., not generally clear how to interpolate gradients with g-convex f.

Adding bumps to a nice starter f is the only thing that worked for us.

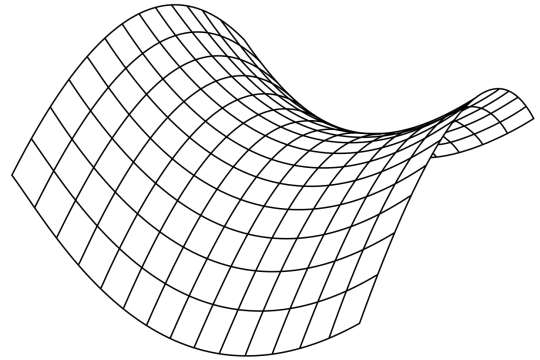


Conclusions

Results extend to (not strong) g-convexity; SPD cone; but randomized?

When it comes to accelerated gradients on Hadamard spaces,
we can’t have the thing we might have thought we should have.

That’s not to say there are no meaningful other types of acceleration.



Examples of Hadamard manifolds

Euclidean space: $\mathbf{R}^n$ with $\langle u, v \rangle = u^\top v$. Constant curvature: 0

Hyperbolic space: $\{x \in \mathbf{R}^{n+1} : x_0^2 = 1 + x_1^2 + \cdots + x_n^2\}$ Constant curvature: -1

with Minkowski metric $\langle u, v \rangle = -u_0 v_0 + u_1 v_1 + \cdots + u_n v_n$.

Positive definite matrices $\{X \in \mathbf{R}^{n \times n} : X = X^\top \text{ and } X \succ 0\}$ Variable curvature, ≤ 0

with affine invariant metric $\langle U, V \rangle_X = \mathrm{Tr}(X^{-1} U X^{-1} V)$.