What is the role of curvature in the complexity of optimization on manifolds?

Nicolas Boumal

Princeton University

with P.-A. Absil, N. Agarwal, B. Bullins, C. Cartis and C. Criscitiello

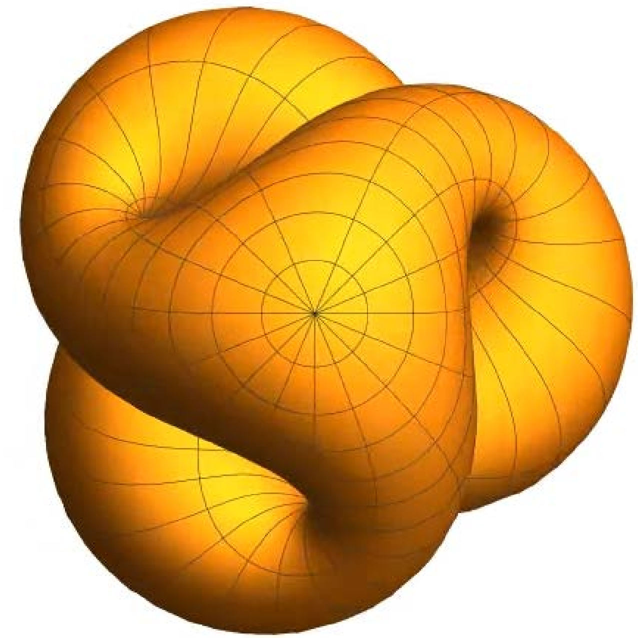
Nicolas Boumal, Oaxaca 2019
Curvature and optimization

https://nypost.com/2019/02/19/youtube-is-helping-the-flat-earth-conspiracy-movement-grow/

https://www.brookeandjubalradio.com/featured/brooke-and-jubal/content/2018-03-09-second-date-tommy-and-hope-why-on-earth/

“Apparently, some people believe the Earth is shaped like a donut”

—Vice.com, Nov. 2018

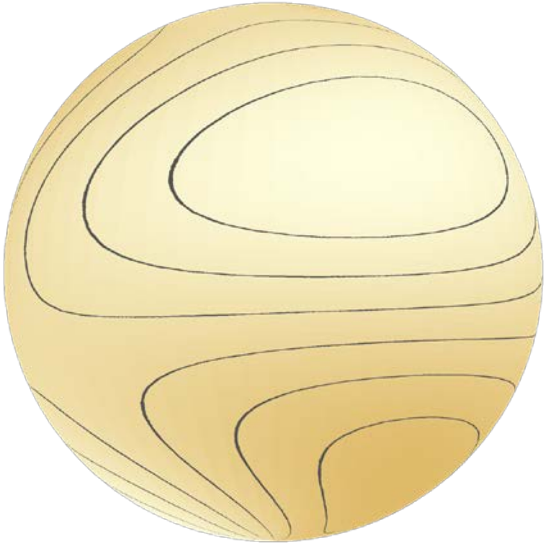


“How does curvature affect optimization?”

Picture: http://homepages.math.uic.edu/~ddumas/teaching/2017/fall/math549/boy/

Optimization on smooth manifolds

$$\min_x f(x) \ \text{ subject to } x \in \mathcal{M}$$



Linear spaces

Unconstrained; linear equality constraints

Low rank (matrices, tensors)

Recommender systems, large-scale Lyapunov equations, …

Orthonormality (Grassmann, Stiefel, rotations)

Dictionary learning, structure from motion, SLAM, PCA, ICA, SBM,…

Positivity (positive definiteness, positive orthant)

Metric learning, Gaussian mixtures, diffusion tensor imaging, …

Symmetry (quotient manifolds)

Invariance under group actions

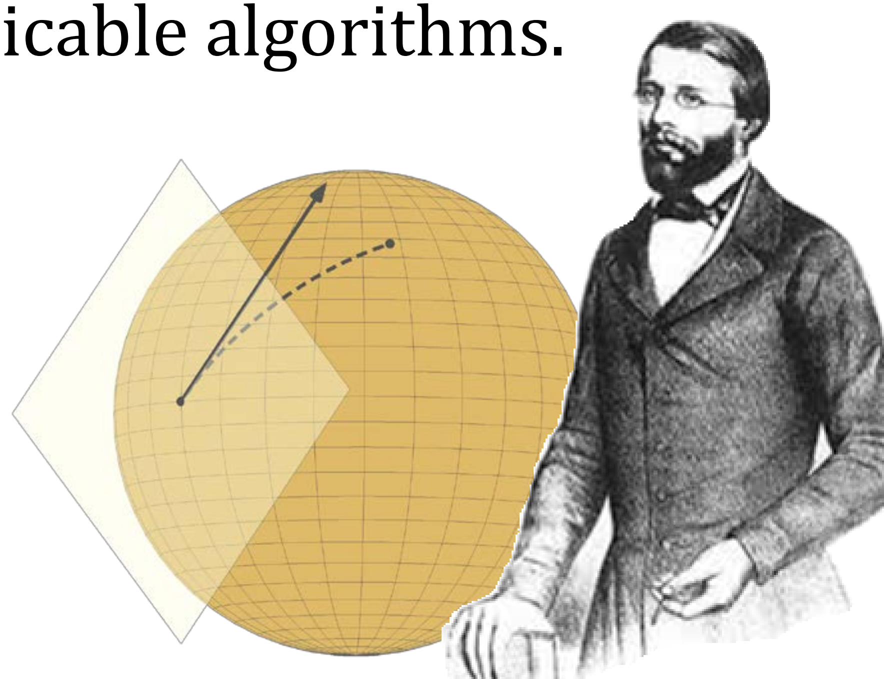
A Riemannian structure gives us gradients and Hessians

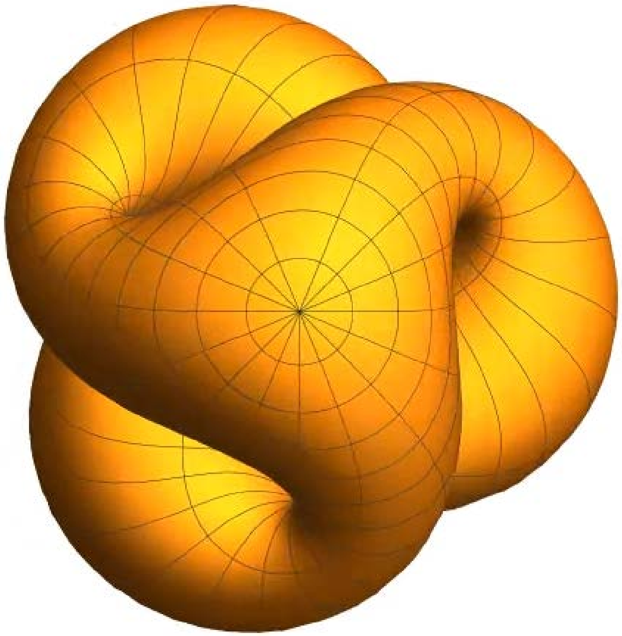
The essential tools of smooth optimization are defined generally on Riemannian manifolds.

Unified theory, broadly applicable algorithms.

First ideas from the '70s.

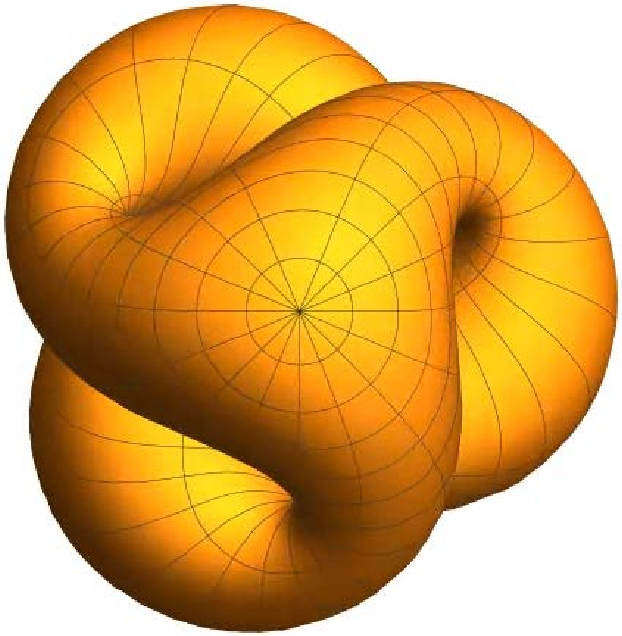
First practical in the '90s.





"All other things being equal, is it harder to optimize if the space is more curved?"

Picture: http://homepages.math.uic.edu/~ddumas/teaching/2017/fall/math549/boy/



Whatever that means…

"All other things being equal, is it harder to optimize if the space is more curved?"

Picture: http://homepages.math.uic.edu/~ddumas/teaching/2017/fall/math549/boy/

Does curvature impede optimization?

Message 1

Under a natural Lipschitz assumption: no.

Message 2

Unclear if curvature affects Lipschitz constant.

Target: approximate critical points

$$\|\mathrm{grad} f(x)\| \leq \varepsilon$$

Iteration complexity of gradient descent?

1. Classical analysis in $\mathbf{R}^n$.

2. Extended to manifolds ~2016.
   Zhang & Sra; B., Absil & Cartis; Bento, Ferreira & Melo

Gradient descent in $\mathbf{R}^n$

A1 $f(x) \geq f_{\text{low}}$ for all $x \in \mathbf{R}^n$

A2 ∇f is L-Lipschitz: $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$

Algorithm: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

Complexity: $\|\nabla f(x_K)\| \leq \varepsilon$ for some $K \leq 2L(f(x_0) - f_{\text{low}})\frac{1}{\varepsilon^2}$

$$\textbf{A2} \Rightarrow f(y) - f(x) - \langle y - x, \nabla f(x)\rangle \leq \frac{L}{2}\|y - x\|^2$$

$$\Rightarrow f(x_{k+1}) - f(x_k) + \frac{1}{L}\langle \nabla f(x_k), \nabla f(x_k)\rangle \leq \frac{1}{2L}\|\nabla f(x_k)\|^2$$

$$\Rightarrow f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2$$

$$\textbf{A1} \Rightarrow f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{K} f(x_k) - f(x_{k+1}) > \frac{\varepsilon^2}{2L}(K + 1)$$

Lipschitz gradients on complete manifolds

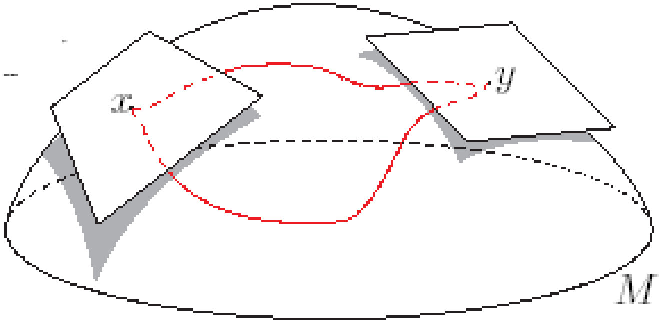


Image: D.K. Wise, IOP Science

Using parallel transport and exponential map:

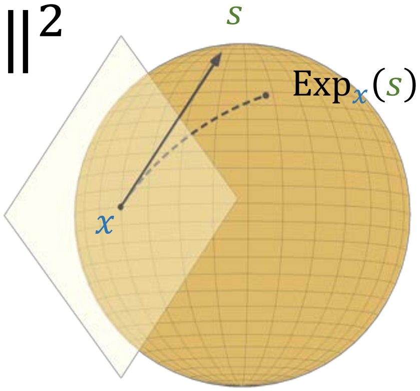
$$\|\mathrm{grad} f(y) - \mathrm{P}_{y \leftarrow x} \mathrm{grad} f(x)\| \leq L \cdot \mathrm{dist}(x, y),$$

$\mathrm{P}_{y \leftarrow x}$ is parallel transport along $\gamma(t) = \mathrm{Exp}_x(ts)$ from $x = \gamma(0)$ to $y = \gamma(1) = \mathrm{Exp}_x(s)$.

Implies the key quadratic bound:

$$f\big(\mathrm{Exp}_x(s)\big) - f(x) - \langle s, \mathrm{grad} f(x)\rangle \leq \frac{L}{2}\|s\|^2$$

RGD: $x_{k+1} = \mathrm{Exp}_{x_k}\left(-\frac{1}{L}\mathrm{grad} f(x_k)\right)$



Gradient descent on $\mathcal{M}$

A1 $f(x) \geq f_{\text{low}}$ for all $x \in \mathcal{M}$

A2 $\|\text{grad}f(y) - \text{P}_{y \leftarrow x}\text{grad}f(x)\| \leq L \cdot \text{dist}(x, y)$

Algorithm: $x_{k+1} = \text{Exp}_{x_k}\left(-\frac{1}{L}\text{grad}f(x_k)\right)$

$\Rightarrow \|\text{grad}f(x_K)\| \leq \varepsilon$ with $K \leq 2L(f(x_0) - f_{\text{low}})\frac{1}{\varepsilon^2}$

Curvature-free complexity!

Second-order target

$$\|\mathrm{grad} f(x)\| \leq \varepsilon \qquad \mathrm{Hess} f(x) \succcurlyeq -\sqrt{\varepsilon}$$

Assume Lipschitz continuous Riemannian Hessian.

1. Riemannian trust regions: $O\left(\varepsilon^{-2.5}\right)$
   With Absil and Cartis, arXiv:1605.08101
2. Riemannian cubic regularization: $O\left(\varepsilon^{-1.5}\right)$
   With Agarwal, Bullins and Cartis, arXiv:1806.00065
   See also Zhang and Zhang, arXiv:1805.05565

Both curvature free. But this one is tricky:

3. Riemannian perturbed GD: $O\left(\log(d)^4 \varepsilon^{-2}\right)$
   With Criscitiello, arXiv:1906:04321
   See also Sun, Flammarion and Fazel, arXiv:1906:07355

Riemannian Lipschitz, **with** Riemannian curvature in bounds

Geodesically convex optimization (Zhang & Sra 2016)
RSVRG (Zhang, Reddi & Sra 2016)
SGD with averaging (Tripuraneni, Flammarion, Bach & Jordan 2018)
Perturbed gradient descent (Sun, Flammarion & Fazel 2019)

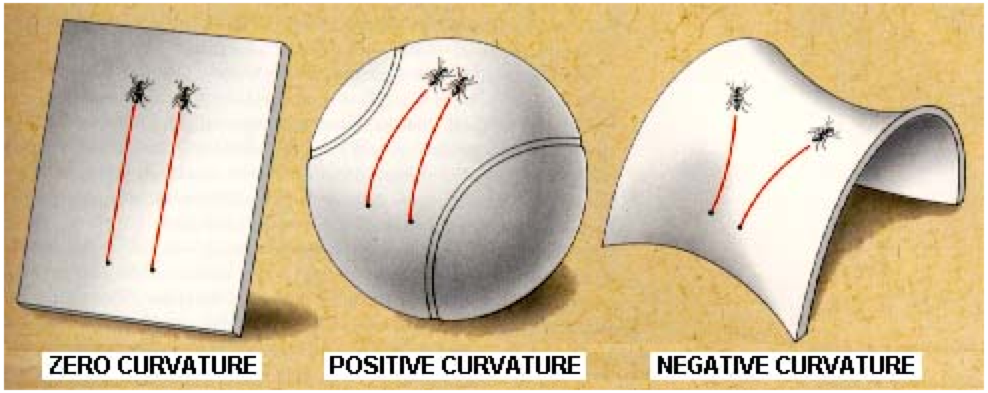
Riemannian Lipschitz, **no** Riemannian curvature in bounds

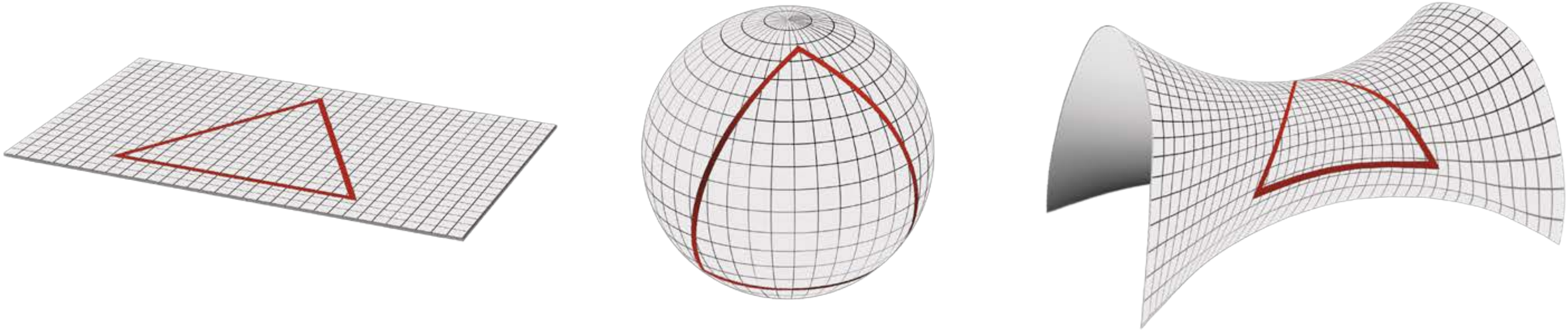
Gradient descent (Bento, Ferreira & Melo 2017)
Trust-regions (B., Absil & Cartis 2018)
Adaptive regularization with cubics (Agarwal, B., Bullins & Cartis 2019)
R-Spider (stochastic) (Zhang, Zhang & Sra 2018)
Frank-Wolfe (Weber & Sra 2017)

Pullback Lipschitz, no curvature in bounds, but maybe hidden

Gradient descent (B., Absil & Cartis 2018)
Trust-regions
Adaptive regulization with cubics
Perturbed gradient descent (Criscitiello & Boumal, 2019)

How does curvature come up in complexity bounds?





Geodesics: ?
Triangles: https://januscosmologicalmodel.com/negativemass

Does curvature affect Lipschitz cnsts?

Here are two possible ways to address this.
Consider $f: \mathbf{R}^n \to \mathbf{R}$ with Lipschitz gradient:

1. Restrict to a Riemannian submanifold $\mathcal{M} \subset \mathbf{R}^n$.
   Constant L is affected by *extrinsic* curvature.

2. Deform $\mathbf{R}^n$ into a Riemannian manifold.
   Derivative of metric does affect L,
   but link with curvature is indirect.

Case in point: one-dimensional manifolds have *no* intrinsic curvature, yet see these effects.