Iteration complexity of optimization on smooth manifolds

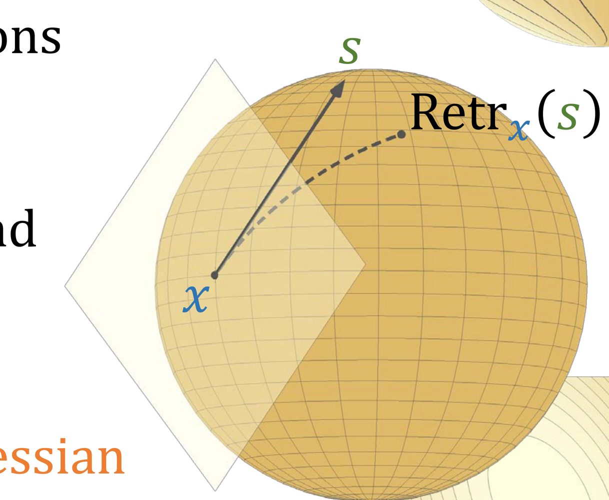
Nicolas Boumal (Princeton University), joint work with P.-A. Absil (UCLouvain), Naman Agarwal (Google), Brian Bullins (Princeton), Coralia Cartis (Oxford)

Mostly harmless

Optimization on smooth manifolds is not that different from unconstrained optimization on a linear space:

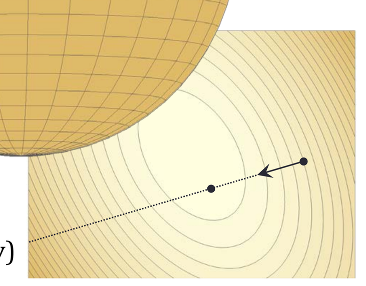
$$\min_x f(x) \quad \text{subject to} \quad x \in \mathcal{M}$$

Tangent spaces: allowed directions
E.g.: $\mathrm{T}_x\mathcal{M} = \{s \in \mathbf{R}^n : x^\mathrm{T} s = 0\}$



Retractions: tools to move around
E.g.: $\mathrm{Retr}_x(s) = \frac{x+s}{\|x+s\|}$

Riemannian metric: gradient, Hessian
E.g.: $\langle s_1, s_2\rangle_x = s_1^\mathrm{T} s_2$



These ideas have been around since the 70s (Luenberger, Gabay)

Algorithms we know and love

$$x_{k+1} = \mathrm{Retr}_{x_k}(s_k)$$

Riemannian gradient descent, trust-regions and cubic regularization:

- $s_k = -\alpha_k \,\mathrm{grad}\, f(x_k)$
- $s_k \approx \operatorname{argmin}_{s \in \mathrm{T}_{x_k}\mathcal{M}, \|s\| \le \Delta_k} f(x_k) + \langle s, \mathrm{grad}\, f(x_k)\rangle + \frac{1}{2}\langle s, \mathrm{Hess}\, f(x_k)[s]\rangle$
- $s_k \approx \operatorname{argmin}_{s \in \mathrm{T}_{x_k}\mathcal{M}} f(x_k) + \langle s, \mathrm{grad}\, f(x_k)\rangle + \frac{1}{2}\langle s, \mathrm{Hess}\, f(x_k)[s]\rangle + \frac{\sigma_k}{3}\|s\|^3$

Manifolds that matter

Any Cartesians products of all of these:

- Unit norm vectors (spheres)
- Matrices with orthonormal columns (Stiefel manifold)
- Subspaces of $\mathbf{R}^n$ of dimension k (Grassmann manifold)
- Fixed-rank matrices (general, symmetric, psd…)
- Low-rank tensors (Tucker, tensor train)
- Euclidean distance matrices
- Rotation matrices
- Positive probability distributions
- Positive definite matrices
- Many quotients by group actions
- …

Familiar looking bounds

For any $x_0 \in \mathcal{M}$, worst-case iteration complexity:

Riemannian gradient descent

$O(\varepsilon^{-2})$ for $\|\mathrm{grad} f(x)\| \le \varepsilon$

Riemannian trust regions

$O(\varepsilon^{-2})$ for $\|\mathrm{grad} f(x)\| \le \varepsilon$
$O(\varepsilon^{-3})$ for $\lambda_{\min}(\mathrm{Hess} f(x)) \ge -\varepsilon$ too

Riemannian adaptive regularization with cubics

$O(\varepsilon^{-1.5})$ for small gradient (optimal)
$O(\varepsilon^{-3})$ for second-order too

RGD and RTR: arXiv:1605.08101, IMA JNA 2019
ARC: arXiv:1806.00065 (see also Zhang & Zhang, arXiv:1805.05565)

Proof for gradient descent

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

A2 $f(\mathrm{Retr}_x(s)) - f(x) - \langle s, \mathrm{grad} f(x)\rangle \le \frac{L}{2}\|s\|^2$

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

Complexity: $\|\mathrm{grad} f(x_K)\| \le \varepsilon$ with $K \le 2L(f(x_0) - f_{\mathrm{low}})\frac{1}{\varepsilon^2}$

$$\textbf{A2} \Rightarrow f(x_{k+1}) - f(x_k) + \frac{1}{L}\|\mathrm{grad} f(x_k)\|^2 \le \frac{1}{2L}\|\mathrm{grad} f(x_k)\|^2$$

$$\Rightarrow f(x_k) - f(x_{k+1}) \ge \frac{1}{2L}\|\mathrm{grad} f(x_k)\|^2$$

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

Main assumptions

Mostly the same as in linear spaces:

- f lower-bounded on $\mathcal{M}$
- Sufficient decrease per iteration, either in the actual cost function (RGD) or in the model (RTR, ARC)
- Regularity assumptions for f; this is key!

Standard assumptions would be, e.g., Lipschitz gradient.
However, this is uncomfortable on manifolds:

$$\mathrm{dist}(\mathrm{grad} f(x), \mathrm{grad} f(y)) \le L\,\mathrm{dist}(x, y)$$

Far easier to compare only scalars:

$$f(\mathrm{Retr}_x(s)) \le f(x) + \langle s, \mathrm{grad} f(x)\rangle + \frac{L}{2}\|s\|^2$$

Same thing for Lipschitz Hessian: go up one order.

When things are different

To get optimal rates for ARC, we need more work.

Pullback: $\hat{f}_x = f \circ \mathrm{Retr}_x : \mathrm{T}_x\mathcal{M} \to \mathbf{R}$

Solving the subproblem, we make $\mathrm{grad}\hat{f}_x(s)$ small.

For complexity bound, we need $\mathrm{grad} f(\mathrm{Retr}_x(s))$ small.

On linear spaces, they are the same ($\mathrm{Retr}_x(s) = x + s$).

On manifolds,

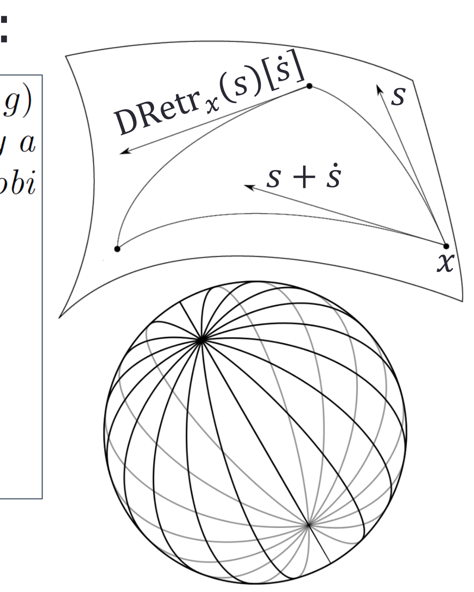
$$\mathrm{grad}\hat{f}_x(s) = \left(\mathrm{DRetr}_x(s)\right)^{\mathrm{adj}}[\mathrm{grad} f(\mathrm{Retr}_x(s))]$$

Need to control minimum singular value; two tools:

Jacobi field comparison theorem (Lee 1997):

Theorem 11.2. (Jacobi Field Comparison Theorem) *Suppose (M, g) is a Riemannian manifold with all sectional curvatures bounded above by a constant C. If γ is a unit speed geodesic in M, and J is any normal Jacobi field along γ such that $J(0) = 0$, then*

$$|J(t)| \ge \begin{cases} t\,|D_tJ(0)| & \text{for } 0 \le t, & \text{if } C = 0; \\ R\sin\frac{t}{R}\,|D_tJ(0)| & \text{for } 0 \le t \le \pi R, & \text{if } C = \frac{1}{R^2} > 0; \\ R\sinh\frac{t}{R}\,|D_tJ(0)| & \text{for } 0 \le t, & \text{if } C = -\frac{1}{R^2} < 0. \end{cases}$$



Maximum theorem (Bergé 1963):

Theorem 2. *If ϕ is an upper semi-continuous numerical function in $X \times Y$ and Γ is a u.s.c. mapping of X into Y such that, for each x, $\Gamma x \neq \varnothing$, the numerical function M defined by*

$$M(x) = \max\{\phi(x, y) \,/\, y \in \Gamma x\}$$

is upper semi-continuous.