Escaping Saddle Points on Manifolds

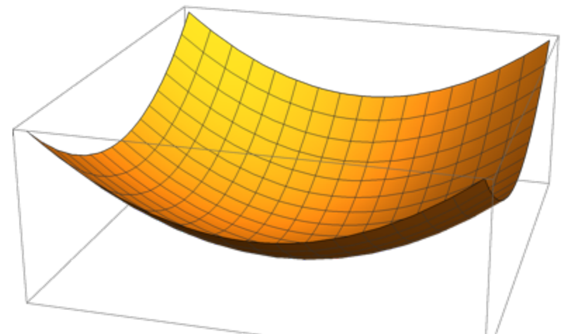
Chris Criscitiello and Nicolas Boumal (Princeton University)

The problem:

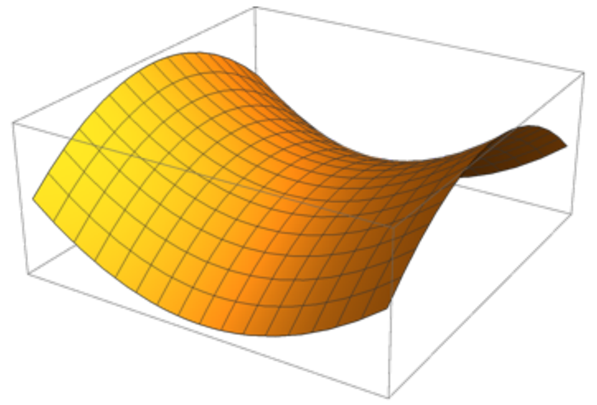
- M is a d-dimensional smooth manifold with Riemannian metric
- Riemannian metric induces gradient $\mathrm{grad}\, f(x)$, Hessian $\mathrm{Hess}\, f(x)$
- Problem:

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

with f nonconvex.

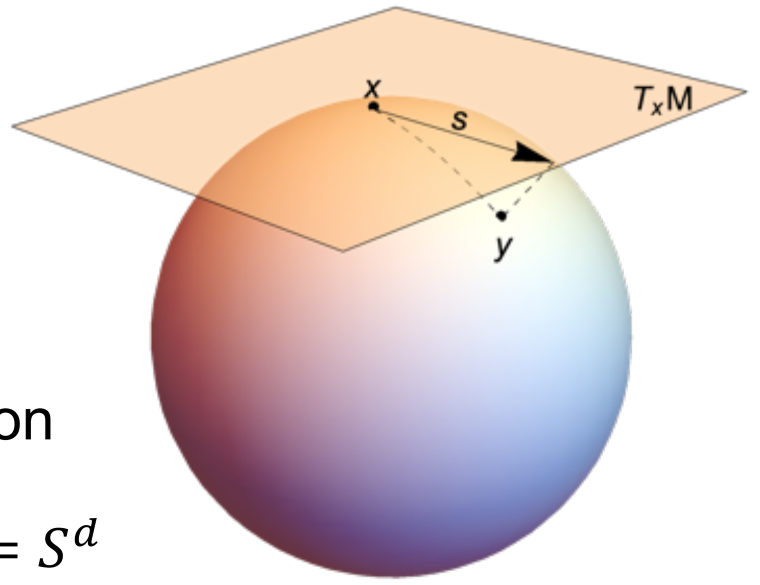


- Global minimum is hard to find. Instead seek:
- ϵ-FOCP: $\|\mathrm{grad}\, f(x)\| \le \epsilon$
- ϵ-SOCP: $\|\mathrm{grad}\, f(x)\| \le \epsilon,\ \lambda_{min}(\mathrm{Hess}\, f(x)) \ge -\sqrt{\rho\epsilon}$



- **Objective: Find SOCP *without* Hessian queries.**
- Applications:
  - **numerical linear algebra -** spectral decompositions, low-rank Lyapunov equations
  - **signal and image processing -** shape analysis, diffusion tensor imaging, community detection on graphs, rotational video stabilization
  - **statistics and machine learning -** matrix/tensor completion, metric learning, Gaussian mixtures, activity recognition, independent component analysis
  - **robotics and computer vision** - simultaneous localization and mapping, structure from motion, pose estimation

Optimization on manifolds:



- To move on the manifold, use retractions:

$$y = \mathrm{Retr}_x(s), \quad s \in T_xM$$

- Tangent space T_xM gives possible directions
- E.g., follow geodesics, or use metric projection

$$\mathrm{Retr}_x(s) = \frac{x+s}{\|x+s\|} \text{ for } M = S^d$$

- Riemannian gradient descent (RGD): $x_{t+1} = \mathrm{Retr}_{x_t}(-\eta\, \mathrm{grad}\, f(x_t))$
- RGD visits an ϵ-FOCP in $O(\epsilon^{-2})$ iterations.
- **Pullback** $\hat{f}_x : T_xM \to \mathbb{R}$: $\hat{f}_x(s) = f(\mathrm{Retr}_x(s))$

Euclidean case (Jin, Netrapalli, Ge, Kakade, Jordan 2019):

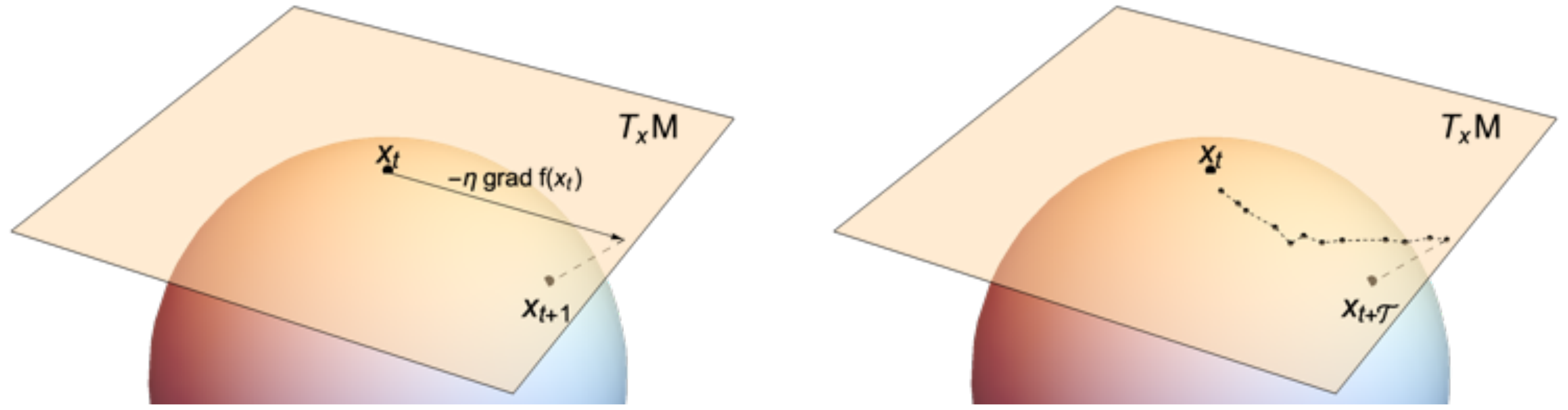
- Jin et al.'s setting:

$$\min f(x) \quad \text{subject to } x \in \mathbb{R}^d$$

- Perturbed Gradient Descent:
  - If $\|\nabla f(x_t)\| \ge \epsilon$, perform a GD step $x_{t+1} = x_t - \eta \nabla f(x_t)$.
  - If $\|\nabla f(x_t)\| < \epsilon$, **perturb** then perform $\mathcal{T}$ GD steps.
- Visits an ϵ-SOCP in $O(\epsilon^{-2} \log^4(d))$ iterations with high probability.
- Intuition: Saddle points are unstable.
- **Proof relies heavily on vector spaces. How to overcome this?**

Our extension to smooth manifolds:

- **Make batches of steps in a single tangent space.**
- Perturbed Riemannian Gradient Descent (PRGD):
  - (a) If $\|\mathrm{grad}\, f(x_t)\| \ge \epsilon$, perform an RGD step $x_{t+1} = \mathrm{Retr}_{x_t}(-\eta\, \mathrm{grad}\, f(x_t))$.
  - (b) If $\|\mathrm{grad}\, f(x_t)\| < \epsilon$, enter tangent space $T_{x_t}M$, then perturb and perform $\mathcal{T}$ GD steps on the pullback $\hat{f}_{x_t}$ *in that tangent space*. Retract back to manifold.



(a) (b)

- Visits an ϵ-SOCP in $O(\epsilon^{-2} \log^4(d))$ iterations with high probability.
- **Extends Jin et al.'s analysis (almost) seamlessly.**

Competing Extension (Sun, Flammarion, Fazel 2019):

- Sun et al. *perform* all steps on the manifold and *analyze* them in a common tangent space.
- More natural but also more technical.
- Similar but different regularity assumptions on f.
- $\mathrm{Retr} = \mathrm{Exp}$: move along geodesics.
- Iteration complexity: same dependence in ϵ and d; also curvature?

Details:

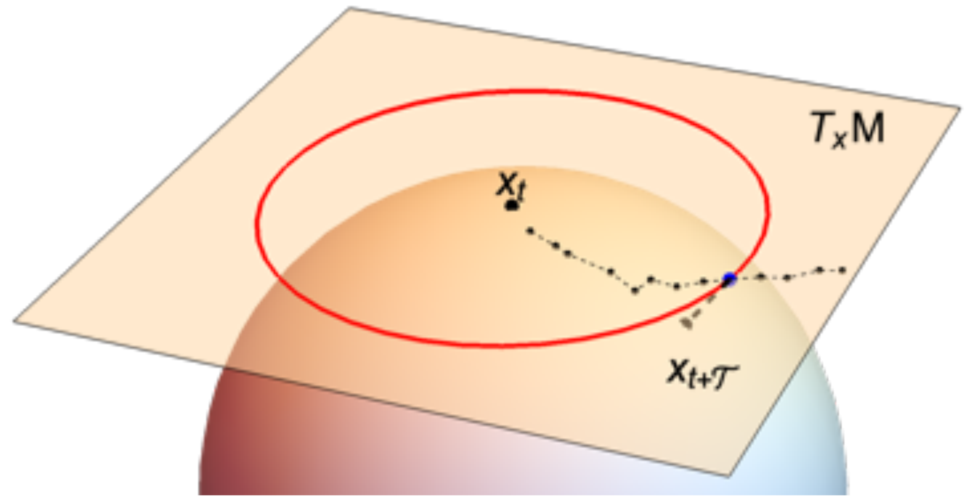
- Assumptions:
  - (A1) f is lower-bounded.
  - (A2) Gradient of the pullback is "Lipschitz" in a ball:

$$\|\nabla \hat{f}_x(s) - \nabla \hat{f}_x(0)\| \le L\|s\| \ \ \forall s \in T_xM \text{ with } \|s\| \le b.$$

  - (A3) Hessian of the pullback is "Lipschitz" in a ball:

$$\|\nabla^2 \hat{f}_x(s) - \nabla^2 \hat{f}_x(0)\| \le \rho\|s\| \ \ \forall s \in T_xM \text{ with } \|s\| \le b.$$

  - (A4) Second-order retraction.



- Issue: What if tangent space iterates escape the ball of radius b?
  - Handle with Jin et al.'s improve-or-localize lemma.
  - Require $\epsilon \le b^2\rho$.
- So, more precisely, PRGD visits an ϵ-SOCP in $O(\max\{\epsilon^{-2}, b^{-4}\}\log^4(d))$ iterations with high probability.
- PCA: $\max \frac{1}{2} x^T A x$ subject to $x \in S^{d-1}$, $L = \frac{5}{2}\|A\|, \rho = 9\|A\|, b = \infty$.

Future Directions:

- Role of curvature of M?
- Adaptive scheme that doesn't need to know smoothness parameters?
- Perturbed Stochastic Gradient Descent (PSGD, Jin et al. 2019)?
- Running many steps in a single tangent space before retracting means more classical methods can be adapted. In particular, it may be easier to generalize:
  - Parallelized schemes
  - Coordinate descent algorithms
  - Accelerated schemes
- See also *trivializations* paper by M. Lezcano Casado.