Escaping from saddle points on Riemannian manifolds

Yue Sun†, Nicolas Flammarion‡, Maryam Fazel†

† Department of Electrical and Computer Engineering, University of Washington, Seattle ‡ School of Computer and Communication Sciences, EPFL

Abstract

We consider minimizing a nonconvex, smooth function $f(x)$ on a smooth manifold $x \in \mathcal{M}$. We show that a perturbed Riemannian gradient algorithm converges to a *second-order* stationary point in a number of iterations that is polynomial in appropriate smoothness parameters of f and $\mathcal{M}$, and polylog in dimension. This matches the best known rate for unconstrained smooth minimization.

Background and motivation

Consider the optimization problem

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad f(x), \\ &\text{subject to} \quad x \in \mathcal{M}, \end{aligned}$$

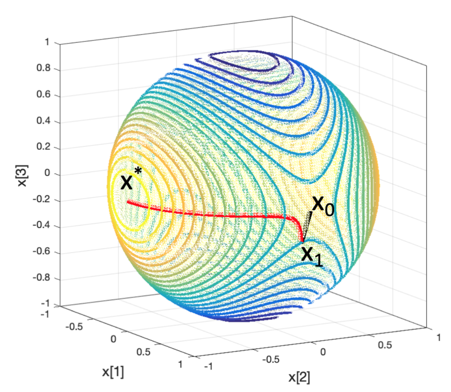


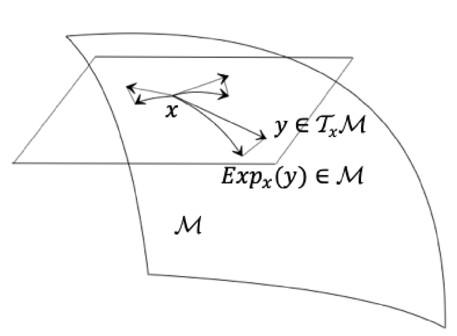
Figure: Escaping from saddle trajectory.

where $\mathcal{M}$ is a manifold of dimension d, optimization variable is $x \in \mathcal{M}$, and (nonconvex) function $f(x)$ is twice differentiable. Finding global optimum is generally not possible. We seek an approximate *second order stationary point* on the manifold (defined in main theorem) using *first-order algorithms*.

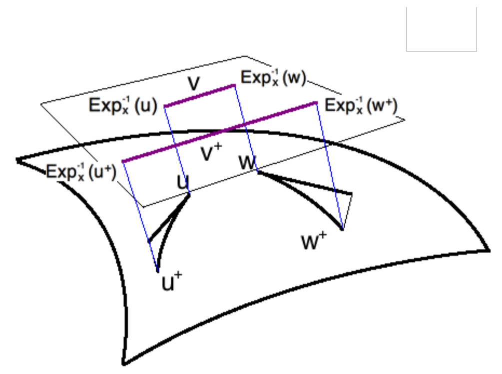
Related work:

- Unconstrained case: convergence rate of perturbed GD is polynomial in smoothness parameters and d [1] .
- Equality-constrained case (with explicit constraints): convergence rate of noisy GD is polynomial in smoothness parameters and polylog in d [2].

Here we study perturbed Riemannian GD and show convergence rate is polylog in d and polynomial in smoothness parameters. This extends best known unconstrained rates to the case of non-Euclidean, manifold constrained problems (e.g., optimization on matrix manifolds).



(a)



(b)

Figure: (a) Exponential map on manifold; (b) Progress of two iterate sequences.

Taylor series and smoothness assumptions

Notation: $\mathrm{Exp}_x(y)$ denotes the exponential map, $\mathrm{grad} f(x)$ and $H(x)$ are Riemannian gradient and Hessian of $f(x)$; x is a saddle point, and $\breve{u} = \mathrm{Exp}_x^{-1}(u)$.

- **Riemannian gradient descent.** Let f have a β-Lipschitz gradient. There exists $\eta = \Theta(1/\beta)$ such that Riemannian gradient descent step $u^+ = \mathrm{Exp}_u(-\eta \mathrm{grad} f(u))$ (c.f. Euclidean case: $u^+ = u - \eta \nabla f(u)$) monotonically decreases f by $\frac{\eta}{2}\|\mathrm{grad} f(u)\|^2$.
- **ρ-Lipschitz Hessian.** Let $\hat{f}_x = f \circ \mathrm{Exp}_x$ have a ρ-Lipschitz Hessian, then
$$\hat{f}_x(\breve{u}) = f(u) \le f(x) + \langle \mathrm{grad} f(x), \breve{u} \rangle + \frac{1}{2} H(x)[\breve{u}, \breve{u}] + \frac{\rho}{6}\|\breve{u}\|^3.$$
- **Two perturbed iterates; negative curvature direction.** Let u, w be perturbations of x, then
$$\left\| (\breve{w^+} - \breve{u^+}) - (\breve{w} - \breve{u}) + \eta H(x)[\breve{w} - \breve{u}] \right\| \le \eta \hat{\rho} \|\breve{w} - \breve{u}\| (\|\breve{w} - x\| + \|\breve{u} - x\|).$$

$\hat{\rho}$ is a function of (1) Hessian Lipschitz constant of $f(\cdot)$, (2) Hessian Lipschitz constant of $\mathrm{Exp}.(\cdot)$, (3) spectral norm of Riemannian curvature tensor, (4) injectivity radius.

Main theorem

Let smoothness assumptions above hold. With probability δ, perturbed Riemannian GD takes

$$O\left(\frac{\beta(f(x_0) - f(x^*))}{\epsilon^2} \log^4\left(\frac{\beta d (f(x_0) - f(x^*))}{\epsilon^2 \delta}\right)\right)$$

iterations to reach an $(\epsilon, -\sqrt{\hat{\rho}\epsilon})$-stationary point, where $\|\mathrm{grad} f(x)\| \le \epsilon$ and $\lambda_{\min} H(x) \ge -\sqrt{\hat{\rho}\epsilon}$.

Algorithm (informal)

- At iterate x, check the norm of gradient
- If large: do $x^+ = \mathrm{Exp}_x(-\eta \mathrm{grad} f(x))$ to decrease function value
- If small: near either a saddle point or a local min. Perturb iterate by adding appropriate noise, run a few iterations
  - if f decreases, iterates escape saddle point (and alg continues)
  - if f doesn't decrease: at approximate local min (alg terminates).

Example – Burer-Monteiro factorization.

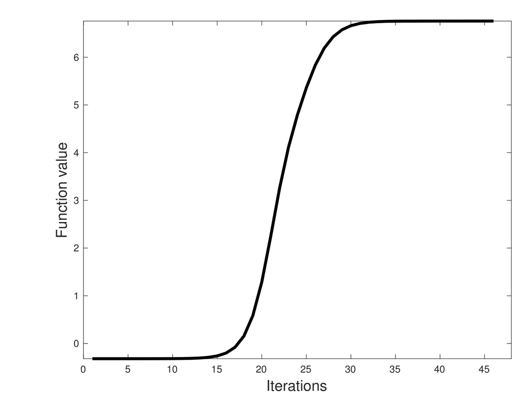
Let $A \in \mathbb{S}^{d \times d}$, the problem

$$\max_{X \in \mathbb{S}^{d \times d}} \mathrm{trace}(AX),$$
$$s.t.\ \mathrm{diag}(X) = 1, X \succeq 0, \mathrm{rank}(X) \le r.$$

can be factorized as

$$\max_{Y \in \mathbb{R}^{d \times p}} \mathrm{trace}(AYY^T),\ s.t.\ \mathrm{diag}(YY^T) = 1.$$

when $r(r+1)/2 \le d$, $p(p+1)/2 \ge d$.



Iteration versus function value. The iterations start from a saddle point, is perturbed by the noise, and converges to a local minimum (that is proven global as well).

Contributions

For (nonconvex) optimization on Riemannian manifold, perturbed Riemannian GD has a rate

- Polylog in dimension (improving 'polynomial in dimension' rate in [2])
- Comparable polynomial dependence on ϵ and β as in unconstrained case [1].
- Explicit polynomial dependence of curvature constant, which is implicit in [3].

Future work

It is known that accelerated method works in escaping saddle framework [4], it's also of interest whether we can run accelerated algorithm on manifolds.

Another recent trend is to consider optimization problem with equality and inequality constraints [5, 6]. They require solution or approximation oracle for NP-hard problems in general (including copositivity test).

Bibliography

[1] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1724–1732.

[2] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points – online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, 2015, pp. 797–842.

[3] C. Criscitiello and N. Boumal, "Efficiently escaping saddle points on manifolds," *arXiv preprint arXiv:1906.04321*, 2019.

[4] C. Jin, P. Netrapalli, and M. I. Jordan, "Accelerated gradient descent escapes saddle points faster than gradient descent," *arXiv preprint arXiv:1711.10456*, 2017.

[5] A. Mokhtari, A. Ozdaglar, and A. Jadbabaie, "Escaping saddle points in constrained optimization," *arXiv preprint arXiv:1809.02162*, 2018.

[6] M. Nouiehed, J. D. Lee, and M. Razaviyayn, "Convergence to second-order stationarity for constrained non-convex optimization," *arXiv preprint arXiv:1810.02024*, 2018.