Escaping from saddle points on Riemannian manifolds

Yue Sun†, Nicolas Flammarion‡, Maryam Fazel†

† Department of Electrical and Computer Engineering, University of Washington, Seattle
‡ Department of Electrical Engineering and Computer Science, University of California, Berkeley

October 12, 2019
**Exponential map and parallel transport.**

Exponential map is a projection-like operation mapping from tangent space to manifold, where the curve from $x \rightarrow \text{Exp}_x(y)$ is a geodesic with initial velocity $y$.

Parallel transport is an operation that translates a tangent vector from $T_x \mathcal{M}$ to $T_y \mathcal{M}$ along a geodesic.
Manifold constrained optimization

We consider the manifold constrained optimization problem

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

assuming the function and manifold satisfying

1. There is a finite constant $\beta$ such that
   $$\|\nabla f(y) - \Gamma^y_x \nabla f(x)\| \leq \beta d(x, y)$$
   for all $x, y \in \mathcal{M}$.

2. There is a finite constant $\rho$ such that
   $$\|H(y) - \Gamma^y_x H(x)\Gamma^x_y\|_2 \leq \rho d(x, y)$$
   for all $x, y \in \mathcal{M}$.

3. There is a finite constant $K$ such that
   $$|K(x)[u, v]| \leq K$$
   for all $x \in \mathcal{M}$ and $u, v \in T_x \mathcal{M}^1$.

$f$ may not be convex.

---

$^1K(x)[u, v]$ denotes the curvature constant of $\mathcal{M}$ at $x$ in direction $u, v$. 
Algorithm

1. At iterate $x$, check the norm of gradient.
2. If large: do $x^+ = \exp_x(-\eta \nabla f(x))$ to decrease function value.
3. If small: near either a saddle point or a local min. Perturb iterate by adding appropriate noise, run a few iterations.
   3.1 if $f$ decreases, iterates escape saddle point (and alg continues).
   3.2 if $f$ doesn’t decrease: at approximate local min (alg terminates).
Theorem (Jin et al., Euclidean space)

Perturbed GD converges to a \((\epsilon, -\sqrt{\rho \epsilon})\)-stationary point of \(f\) in

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

iterations.

We replace Hessian Lipschitz \(\rho\) by \(\hat{\rho}\) as a function of \(\rho\) and \(K\) and we quantify it in the paper.

Theorem (manifold)

Perturbed RGD converges to a \((\epsilon, -\sqrt{\hat{\rho}(\rho, K) \epsilon})\)-stationary point of \(f\) in

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

iterations.
Experiment

Burer-Monteiro factorization.

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

$$\max_{X \in S^{d \times d}} \text{trace}(AX),$$

s.t. $\text{diag}(X) = 1, X \succeq 0, \text{rank}(X) \leq r$.

can be factorized as

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

when $r(r + 1)/2 \leq d, p(p + 1)/2 \geq d$.