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, Lausanne, Switzerland

November 8, 2019
Manifold constrained optimization

We consider the problem

\[
\begin{align*}
\text{minimize} & \quad f(x), \quad \text{subject to} \quad x \in M
\end{align*}
\]

Same as Euclidean space, generally we cannot find global optimum in polynomial time, so we want to find an approximate local minimum.

Plot of saddle point in Euclidean space.

Contour of function value on sphere.
Example of manifolds

1. Sphere. \( \{ x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i^2 = r^2 \} \).

2. Stiefel manifold. \( \{ X \in \mathbb{R}^{m \times n} : X^T X = I \} \).

3. Grassmannian manifold. \( \text{Grass}(p, n) \) is set of \( p \) dimensional subspaces in \( \mathbb{R}^n \).

4. Bruer-Monteiro relaxation. \( \{ X \in \mathbb{R}^{m \times n} : \text{diag}(X^T X) = 1 \} \).
A curve in a continuous map \( \gamma : t \rightarrow \mathcal{M} \).
Usually \( t \in [0, 1] \), where \( \gamma(0) \) and \( \gamma(1) \) are start and end points of the curve.
Tangent vector and tangent space

We use

\[ \dot{\gamma}(t) = \lim_{\tau \to 0} \frac{\gamma(t + \tau) - \gamma(t)}{\tau} \]

as the velocity of the curve, \( \dot{\gamma}(t) \) is a tangent vector at \( \gamma(t) \in \mathcal{M} \).

\( x \in \mathcal{M} \) can be start point of many curves, and a tangent space \( T_x\mathcal{M} \) is the set of tangent vectors at \( x \).

Tangent space is a metric space.
Gradient of a function

Let $f : \mathcal{M} \to \mathbb{R}$ be a function defined on $\mathcal{M}$, and $\gamma$ be a curve on $\mathcal{M}$.

The directional derivative of $f$ in direction $\dot{\gamma}(0)$ is\(^1\)

$$\dot{\gamma}(0)f = \left. \frac{d(f(\gamma(t)))}{dt} \right|_{t=0} = \lim_{\tau \to 0} \frac{f(\gamma(0 + \tau)) - f(\gamma(0))}{\tau}$$

Then we can define $\text{grad}f(x) \in T_x\mathcal{M}$, which satisfies

$$\langle \text{grad}f, y \rangle = yf$$

for all $y \in T_x\mathcal{M}$.

---

\(^1\)Usually $\dot{\gamma}$ denotes differential operator and $\gamma'$ denotes the tangent vector. They are closely related. To avoid confusion, we always use $\dot{\gamma}$.
Vector field

The gradient of a function is a special case of vector field of a manifold.
A vector field is a function from a point in $\mathcal{M}$ to tangent vector at that point.
Connection

Denote a smooth vector field on $\mathcal{M}$ by $\mathfrak{X}(\mathcal{M})$. A connection defines directional derivative

$$\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$$

satisfying

$$\nabla_{fx+gy}u = f\nabla_x u + g\nabla_y u,$$
$$\nabla_x (au + bv) = a\nabla_x u + b\nabla_x v,$$
$$\nabla_x (fu) = (xf)u + f(\nabla_x u).$$

Note that

$$\nabla_{e_i} u = \sum_j (\partial_i u^j e_j + u^j \nabla_{e_i} e_j).$$

A special connection is Riemannian (Levi-Civita) connection.
A directional Hessian is defined as

\[ H(x)[u] = \nabla_u \text{grad} f(x) \]

for \( u \in \mathcal{T}_x \mathcal{M}^2 \).

Similar as gradient, we can define a Hessian from directional version.

\[ \langle H(x)u, v \rangle = \langle \nabla_u \text{grad} f(x), v \rangle, \; \forall u, v \in \mathcal{T}_x \mathcal{M}. \]

It is a symmetric operator.

\[ ^2 \text{In Riemannian geometry, one writes } u_x \text{ to indicate } u \in \mathcal{T}_x \mathcal{M} \text{ and the directional Hessian is } \nabla_{u_x} \text{grad} f. \]
A geodesic is a special class of curves on the manifold, which satisfies zero acceleration condition

\[ \nabla_{\dot{\gamma}(t)} \dot{\gamma} = 0. \]
For any $x \in \mathcal{M}$, $y \in T_x \mathcal{M}$ and the geodesic $\gamma$ defined by $y$, 

$$\gamma(0) = x, \quad \dot{\gamma}(0) = y$$

we call the mapping $\operatorname{Exp}_x(y) : T_x \mathcal{M} \to \mathcal{M}$ such that $\operatorname{Exp}_x(y) = \gamma(1)$ as exponential map.

There is a neighborhood with radius $\mathcal{I}$ in $T_x \mathcal{M}$, such that for all $y \in T_x \mathcal{M}$, $\|y\| \leq \mathcal{I}$, exponential map is a bijection/diffeomorphism.
Parallel transport

The parallel transport $\Gamma$ transports a tangent vector $w$ along a curve $\gamma$, satisfying the zero acceleration condition

$$\nabla_{\gamma t} w_t = 0, \quad w_t = \Gamma_{\gamma(0)}^{\gamma(t)} w.$$
Curvature tensor

The curvature tensor describes how curved the manifold is. It relates to the second order structure of the manifold.
A definition by connection is

\[ R(x, y)w = \nabla_x \nabla_y w - \nabla_y \nabla_x w - \nabla_{[x,y]} w. \]

where \( x, y, w \) are in tangent space of the same point\(^3\).

\(^3\)[\(x, y\)] is the Lie bracket defined by \([x, y]f = xyf - yxf\)
Curvature tensor

\[ R(x, y)w = \lim_{t, \tau \to 0} \frac{\Gamma^{0,0}_{0,\tau y} \Gamma^{0,\tau y}_{tx,\tau y} \Gamma^{tx,\tau y}_{tx,0} \Gamma^{tx,0}_{0,0} w - w}{t \tau} \]
Smooth function on Riemannian manifold

We consider the manifold constrained optimization problem

\[ \min_{x} f(x), \text{ subject to } x \in \mathcal{M} \]

assuming the function and manifold satisfying

1. There is a finite constant \( \beta \) such that
   \[ \| \text{grad} f(y) - \Gamma_{x}^{y} \text{grad} f(x) \| \leq \beta d(x, y) \quad \text{for all } x, y \in \mathcal{M}. \]

2. There is a finite constant \( \rho \) such that
   \[ \| H(y) - \Gamma_{x}^{y} H(x) \Gamma_{y}^{x} \|_{2} \leq \rho d(x, y) \quad \text{for all } x, y \in \mathcal{M}. \]

3. There is a finite constant \( K \) such that
   \[ |R(x)[u, v]| \leq K \quad \text{for all } x \in \mathcal{M} \text{ and } u, v \in T_{x}\mathcal{M}. \]

\( f \) may not be convex.
Taylor expansion of a smooth function

For $x, y \in \mathcal{E}$ Euclidean space,

\[
f(y) - f(x) = \langle y - x, \nabla f(x) + \nabla^2 f(x)(y - x) \rangle + \int_0^1 (\nabla^2 f(x + \tau(y - x)) - \nabla^2 f(x))(y - x) d\tau \rangle.
\]

For $x, y \in \mathcal{M}$, let $\gamma$ denote the geodesic where $\gamma(0) = x$, $\gamma(1) = y$, then

\[
f(y) - f(x) = \langle \dot{\gamma}(0), \text{grad} f(x) + \frac{1}{2} \nabla \dot{\gamma}(0) \text{grad} f + \Delta \rangle
\]

where $\Delta = \int_0^1 \Delta(\gamma(\tau)) d\tau$,

\[
\Delta(\gamma(\tau)) = \int_0^1 (\Gamma^x_{\gamma(\tau)} \nabla \dot{\gamma}(\tau) \text{grad} f - \nabla \dot{\gamma}(0) \text{grad} f) d\tau.
\]
Riemannian gradient descent

On a smooth manifold, there exists a $\eta$ such that, if

$$x_{t+1} = \text{Exp}_{x_t}(-\text{grad}f(x_t)),$$

then

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\text{grad}f(x_t)\|^2.$$

Converge to first order stationary.
Proposed algorithm for escaping saddle

Hope to escape from saddle point and converge to an approximate local minimum.

1. At iterate $x$, check the norm of gradient.

2. If large: do $x^+ = \text{Exp}_x(-\eta \nabla f(x))$ 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).
Difficulty of second order analysis

1. We use linearization in first order analysis, for second order, manifold has a second order structure as well.

2. Consider power method in Euclidean space. We need to prove that the biggest eigenvector direction of $x$ grows exponentially.

3. If it’s iteration of variable, we have to consider gradient in different tangent spaces.

4. Some recent work require strong assumptions such as flat manifold, product manifold.

5. Other recent work assume smoothness parameters of composition of function and manifold operator, which are hard to check.
Useful lemmas

Let \( x \in \mathcal{M} \) and \( y, a \in T_x \mathcal{M} \). Let us denote by \( z = \text{Exp}_x(a) \) then

\[
d(\text{Exp}_x(y + a), \text{Exp}_z(\Gamma_x^z y)) \leq c(K) \min\{\|a\|, \|y\|\}(\|a\| + \|y\|)^2.
\]
Useful lemmas

Holonomy.

$$\|\Gamma^x_z\Gamma^z_y\Gamma^y_x w - w\| \leq c(K)d(x, y)d(y, z)\|w\|.$$ 

Similar to definition of curvature tensor, if a vector is parallel transported around a closed curve, then the change is bounded by the area whose boundary is the curve.
Useful lemmas

Euclidean: $f(x) = x^T H x \Rightarrow x^+ = (I - \eta H)x$.

Exponential growth in a vector space.

If function $f$ is $\beta$ gradient Lipschitz, $\rho$ Hessian Lipschitz, curvature constant is bounded by $K$, $x$ is a $(\epsilon, -\sqrt{\hat{\rho}} \epsilon)$ saddle point, and define $u^+ = \text{Exp}_u(-\eta \text{grad} f(u))$ and $w^+ = \text{Exp}_w(-\eta \text{grad} f(w))$. If a small enough neighborhood$^4$,

$$\|\text{Exp}_x^{-1}(w^+) - \text{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\text{Exp}_x^{-1}(w) - \text{Exp}_x^{-1}(u))\| \leq C(K, \rho, \beta) d(u, w) (d(u, w) + d(u, x) + d(w, x)),$$

for some explicit constant $C(K, \rho, \beta)$.

$^4$Quantified in paper.
Theorem

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.
Burer-Monteiro factorization.

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

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

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

can be factorized as

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

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

Iteration versus function value.