Optimality on Riemannian Manifolds

Xiao Wang

Institute for Theoretical Computer Science

School of Computing and Artificial Intelligence

SUFE

Definition

The Riemannian gradient of a smooth function $f : M \to \mathbb{R}$ on a Riemannian manifold M is a vector field on M denoted by $\mathrm{grad} f$ such that, for all $x \in M$, $\mathrm{grad} f(x)$ is the unique tangent vector in $T_x M$ satisfying:

$$df(x)v = \langle v, \mathrm{grad} f(x) \rangle_x$$

for all $v \in T_x M$, where $df(x)$ is 1-form defined on $T_x M$.

In local coordinate system, gradient can be written

$$\mathrm{grad} f(x) = \sum_{i,j} g^{ij} \frac{\partial f}{\partial x_j} \frac{\partial}{\partial x_i}.$$

Proposition

Let $f : M \to \mathbb{R}$ be a smooth function on Riemannian manifold M equipped with a retraction Retr*. Then for any $x \in M$,*

$$\mathrm{grad} f(x) = \mathrm{grad}(f \circ \mathrm{Retr}_x)(0),$$

where $f \circ \mathrm{Retr}_x : T_x M \to \mathbb{R}$ is defined on a Euclidean space.

Proof.

By the chain rule, for any tangent vector $v \in T_x M$,

$$D(f \circ \mathrm{Retr}_x)(0)[v] = Df(\mathrm{Retr}_x(0))[D\mathrm{Retr}_x(0)[v]] = Df(x)[v],$$

since $\mathrm{Retr}_x(0) = x$ and $D\mathrm{Retr}_x(0)$ is the identity map. The definition of gradient gives

$$\langle \mathrm{grad}(f \circ \mathrm{Retr}_x)(0), v \rangle_x = \langle \mathrm{grad} f(x), v \rangle_x.$$

□

First-order optimality conditions

Proposition

Let $f : M \to \mathbb{R}$ be a smooth function on a Riemannian manifold. If x is a local minimizer of f, then $\operatorname{grad} f(x) = 0$.

Given initial point $x_0 \in M$ and a retraction on M, iterate

$$x_{k+1} = \operatorname{Retr}_x(-\eta_k \operatorname{grad} f(x_k))$$

is called *Riemannian gradient descent.*

Assumptions

- There exists $f^* \in \mathbb{R}$ such that $f(x) \geq f^*$ for all $x \in M$.
- At each iteration, the algorithm achieves sufficient decrease, in that there exists a constant $c > 0$ such that for all k

$$f(x_k) - f(x_{k+1}) \geq c \left\| \operatorname{grad} f(x_k) \right\|^2.$$

Proposition

Let f be a smooth function satisfying above assumptions. Let $x_0, x_1, ...$ be iterates generated by RGD with constant c. Then

$$\lim_{k\to\infty} \|\mathrm{grad} f(x_k)\| = 0.$$

Furthermore, for any $K \geq 1$, there exists k in $0, ..., K-1$ such that

$$\mathrm{grad} f(x_k) \leq \sqrt{\frac{f(x_0) - f^*}{c}} \frac{1}{\sqrt{K}}.$$

Definition

Let M be a Riemannian manifold, equipped with the Riemannian connection ∇. The Riemannian Hessian of f at x is a linear operator $\mathrm{Hess} f(x) : T_x M \to T_x M$ defined as follows:

$$\mathrm{Hess} f(x) u = \nabla_u \mathrm{grad} f$$

In general, if we consider ∇X as an operator that maps u to $\nabla_u X$, $\nabla \mathrm{grad} f$ can be computed in local coordinate system, following an alternative definition of connection.

Let E be a real vector bundle on M, and $\Gamma(E)$ is the set of smooth sections of E on M.

Definition (Connection on general vector bundle)

A connection on a vector bundle E is a map

$$\nabla : \Gamma(E) \to \Gamma(T^*M \otimes E),$$

which satisfies the following conditions:

1) For any $s_1, s_2 \in \Gamma(E)$,
$$\nabla(s_1 + s_2) = Ds_1 + Ds_2$$

2) For $s \in \Gamma(E)$ and any $\alpha \in C^\infty M$
$$\nabla(\alpha s) = d\alpha \otimes s + \alpha Ds$$

Suppose X is a smooth tangent vector field on M and $s \in \Gamma(E)$. $\nabla_X s = \langle X, \nabla s \rangle$ returns a value by pairing between TM and $T * M$ (evaluation of vector and its dual).

If ∇ is defined on tangent bundle, choose any local coordinate system (U, x_i) of M, then the natural basis $\{\partial_i\}$ forms a local frame field of the tangent bundle TM on U. Using Christoffel symbol, we have the following expansion

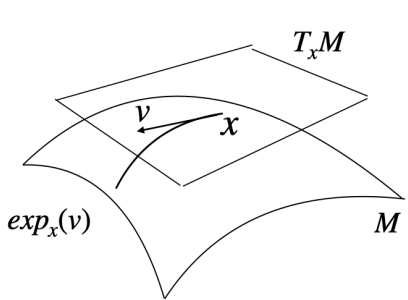
$$\nabla \partial_i = \sum_{k,j} \Gamma_{ik}^j dx_k \otimes \partial_j.$$

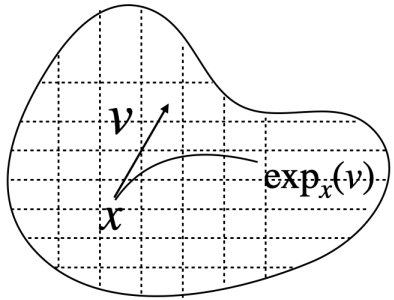
The calculation rule of connection gives $\nabla \mathrm{grad} f$

$$\begin{aligned}
\nabla \mathrm{grad} f &= \nabla \left(g^{ij} \frac{\partial f}{\partial x_i} \partial_i \right) \\
&= d \left(g^{ij} \frac{\partial f}{\partial x_j} \right) \otimes \partial_i + g^{ij} \frac{\partial f}{\partial x_j} \nabla \partial_i \\
&= \frac{\partial}{\partial x_k} \left(g^{ij} \frac{\partial f}{\partial x_j} \right) dx_k \otimes \partial_i + g^{ij} \frac{\partial f}{\partial x_j} \Gamma_{ik}^m dx_k \otimes \partial_m
\end{aligned}$$

Riemannian gradient descent with stepsize α_t

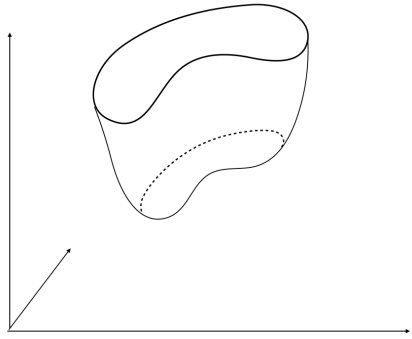
$$x_{t+1} = \mathrm{Exp}_{x_t}(-\alpha_t \mathrm{grad} f(x_t))$$



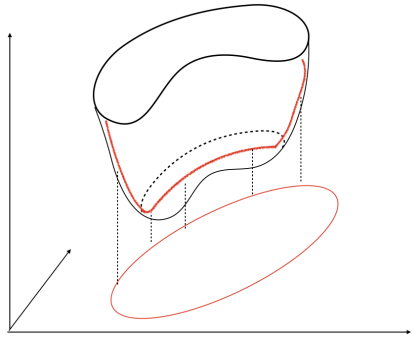


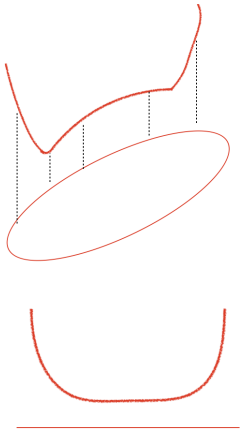
Geodesic Convexity

A non convex function $f : \mathbb{R}^2 \to \mathbb{R}$.



When constrained on an arc of a circle, the function can be convex along the arc.





Geodesically convex set

A set $U \subset M$ is geodesically convex if for any $\mathbf{x}, \mathbf{y} \in U$, there is a geodesic γ with $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$ and $\gamma(t) \in U$ for all $t \in [0, 1]$.

Geodesically convex function

f is call ed geodesically convex on a g.c. set U if for any $\mathbf{x}, \mathbf{y} \in M$ and any geodesic γ connecting $\mathbf{x}$ and $\mathbf{y}$, it holds that

$$f(\gamma(t)) \leq (1 - t)f(\mathbf{x}) + tf(\mathbf{y}).$$

Definition

A function $f : S \to \mathbb{R}$ is geodesically μ-strongly convex for some $\mu > 0$ if the set S is geodesically convex and for any geodesic segment $c : [0, 1] \to M$ whose image is in S we have

$$f(c(t)) \leq (1 - t)f(c(0)) + tf(c(1)) - \frac{t(1 - t)\mu}{2} L(c)^2,$$

where $L(c) = \|c'(0)\|$ is the length of the geodesic segment. The latter is equivalent to the requirement that $f \circ c : [0, 1] \to \mathbb{R}$ be $\mu L(c)^2$-strongly convex in the usual sense.

Theorem

If $f : S \to \mathbb{R}$ is geodesically strictly convex, then it admits at most one local minimizer, which is necessarily the global minimizer.

When a geodesically convex function admits a maximizer, this maximizer typically occurs on the "boundary" of the geodesically convex domain.

- If M is a connected, compact Riemannian manifold and $f : M \to \mathbb{R}$ is continuous and geodesically convex, then f is constant.
- If S_1 is a geodesically convex set in a Riemannian manifold M_1, and similarly for S_2 in M_2, then $S_1 \times S_2$ is geodesically convex in $M_1 \times M_2$.
- If $S \to \mathbb{R}$ is geodesically convex, then f is continuous on the interior of S.

Differentiable geodesically convex functions

Theorem

Let S be a geodesically convex set on a Riemannian manifold M and f be a real function, differentiable in a neighborhood of S. Then, $f : S \to \mathbb{R}$ is geodesically convex if and only if for any geodesic segment $c : [0,1] \to M$ contained in S we have:

$$t \in [0,1], \quad f(c(t)) \geq f(x) + t\langle \mathrm{grad} f(x), c'(0)\rangle_x.$$

Moreover, f is geodesically μ-strongly convex if and only if, whenever $c'(0) \neq 0$,

$$t \in [0,1], \quad f(c(t)) \geq f(x) + t\langle \mathrm{grad} f(x), c'(0)\rangle_x + t^2 \frac{\mu}{2} L(c)^2.$$

Finally, f is geodesically strictly convex if and only if, whenever $c'(0) \neq 0$,

$$t \in (0,1], \quad f(c(t)) > f(x) + t\langle \mathrm{grad} f(x), c'(0)\rangle_x.$$

Logarithmic map

For $\mathbf{x} \in M$, let $\mathrm{Log}_{\mathbf{x}}$ denote the logarithmic map at $\mathbf{x}$,

$$\mathrm{Log}_{\mathbf{x}}(\mathbf{y}) = \mathrm{argmin}_{\mathbf{u} \in T_{\mathbf{x}} M} \mathrm{Exp}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}$$

with domain such that this is uniquely defined.

Riemannian Accelerated Gradient Descent (Zhang & Sra)

$$\begin{aligned}
\mathbf{y}_t &= \mathrm{Exp}_{\mathbf{x}_t}\left(s_1 \mathrm{Log}_{\mathbf{x}_t}(\mathbf{v}_t)\right) \\
\mathbf{x}_{t+1} &= \mathrm{Exp}_{\mathbf{y}_t}(-\alpha \mathrm{grad} f(\mathbf{y}_t)) \\
\mathbf{v}_{t+1} &= \mathrm{Exp}_{\mathbf{y}_t}\left(s_2 \mathrm{Log}_{\mathbf{y}_t}(\mathbf{v}_t) - s_3 \mathrm{grad} f(\mathbf{y}_t)\right)
\end{aligned}$$

where s_1, s_2 and s_3 are parameters related to Lipschitz constant, geodesic convexity and step-size.

Second-order optimality

Proposition

Consider a smooth function $f : M \to \mathbb{R}$. If x is a local or global minimizer of f, then $\operatorname{grad} f(x) = 0$ *and* $\operatorname{Hess} f(x) \geq 0$.

Proof.

We can assume $\operatorname{Hess} f(x)$ is not positive semidefinite. Then there exists a tangent vector $v \in T_x M$ such that $\langle \operatorname{Hess} f(x)[v], v \rangle_x = -2a < 0$, for some positive a. Let $c : I \to M$ be a smooth curve passing through x with velocity v at $t = 0$. Then, using the Taylor expansion and the fact that $\operatorname{grad} f(x) = 0$, we have

$$f(c(t)) = f(x) + \frac{t^2}{2} \langle \operatorname{Hess} f(x)[v], v \rangle_x + O(t^3) = f(x) - at^2 + O(t^3).$$

Hence, there exists t' such that

$$f(c(t)) < f(x) \quad \text{for all } t \in (0, t'],$$

contradicting the fact that x is a local minimum.

Riemannian Newton's method

For second-order retractions, the Taylor expansion gives

$$f(\mathrm{Retr}_x(s)) \approx f(x) + \langle \mathrm{grad} f(x), s\rangle_x + \frac{1}{2}\langle \mathrm{Hess} f(x)[s], s\rangle_x := m_x(s).$$

A minimizer of m_x must be a critical point of m_x.

$$\mathrm{grad} m_x(s) = \mathrm{grad} f(x) + \mathrm{Hess} f(x)[s],$$

and s is a critical point if and only if

$$\mathrm{Hess} f(x)[s] = -\mathrm{grad} f(x).$$

As long as $\mathrm{Hess} f(x)$ is invertible, there exists a unique solution, and this can be used to define the Newton method.

$$\text{Solve } \mathrm{Hess} f(x_k)[s_k] = -\mathrm{grad} f(x_k), \quad \text{for } s_k \in T_{x_k} M \tag{1}$$

$$x_{k+1} = \mathrm{Retr}_{x_k}(s_k) \tag{2}$$

Further reading: An Introduction to Optimization on Smooth Manifolds, Nicolas Boumal