Université Gustave Eiffel

GRADIENT DESCENT ON MANIFOLDS

Data Science Project

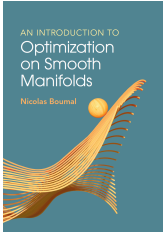
Students: CHAU Dang Minh
LAM Nhat Quan
Alhassane BAH

Outline

Introduction

Introduction

- Many optimization problems involve constraints that can be naturally modeled as manifolds.
- Gradient descent on manifolds extends traditional gradient descent methods to handle these constraints effectively.
- Presentation goals: build up the concept of gradient on manifolds and obtain basic convergence results.
- Main reference: An Introduction to Optimization on Smooth Manifolds by Nicolas Boumal.



Manifolds

Manifolds

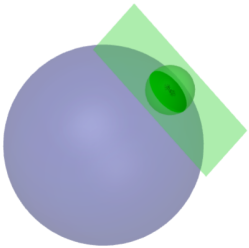
Embedded Submanifolds of a Linear Space

Motivation

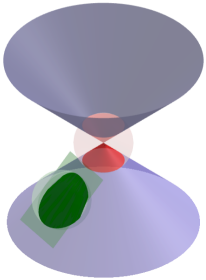
A subset $\mathcal{M}$ of $\mathbb{R}^d$ is an embedded manifold of dimension n if for each point $x \in \mathcal{M}$, there exists a neighborhood in $\mathcal{M}$ of x (i.e. $\mathcal{M} \cap U$ for some open set $U \subset \mathbb{R}^d$ containing x) that is approximate to an open subset of $\mathbb{R}^n$.

Motivation

A subset $\mathcal{M}$ of $\mathbb{R}^d$ is an embedded manifold of dimension n if for each point $x \in \mathcal{M}$, there exists a neighborhood in $\mathcal{M}$ of x (i.e. $\mathcal{M} \cap U$ for some open set $U \subset \mathbb{R}^d$ containing x) that is approximate to an open subset of $\mathbb{R}^n$.



A sphere is a manifold



A cone is not a manifold because every neighborhood of the tip (in red) cannot be approximated by a plane

Embedded Submanifolds of a Linear Space

We consider smooth spaces: for each $x \in \mathcal{M}$, there exists an open set $U \subset \mathbb{R}^d$ containing x and a smooth map $h : U \to \mathbb{R}^{d-n}$ such that $M \cap U = h^{-1}(\{0\})$.

By being approximate to $\mathbb{R}^n$, we mean that for any direction $v \in \mathbb{R}^d$ that is a tangent vector to $\mathcal{M}$ at x, we have

$$h(x + tv) = o(t).$$

We rely on curves for the definition of tangent vectors (and also later definitions).

Definition 1 (Tangent space)

The tangent space $T_x\mathcal{M}$ at a point $x \in \mathcal{M}$ is the set of all tangent vectors to $\mathcal{M}$ at x i.e.

$$T_x\mathcal{M} = \{\gamma'(0) \,|\, \gamma : (-\epsilon, \epsilon) \to \mathcal{M} \in \mathcal{C}^\infty(-\epsilon, \epsilon), \gamma(0) = x\}.$$

Embedded Submanifolds of a Linear Space

We can use Taylor expansion to write

$$h(x+tv) = h(x) + t\mathrm{D}h(x)[v] + o(t) = t\mathrm{D}h(x)[v] + o(t).$$

Proposition 2

For every $x \in \mathcal{M}$, we have

$$T_x\mathcal{M} \subseteq \ker(\mathrm{D}h(x)).$$

Proof.
Let $v \in T_x\mathcal{M}$, then there is a smooth γ such that $\gamma(0) = x$ and $\gamma'(0) = v$. Consider $g(t) = h(\gamma(t))$. Since $\gamma(t) \in \mathcal{M}$ for all t, we have $g(t) = 0$ for all t. Thus, $g'(0) = 0$. By chain rule,

$$g'(0) = \mathrm{D}h(\gamma(0))[\gamma'(0)] = \mathrm{D}h(x)[v] = 0.$$

Hence, $v \in \ker(\mathrm{D}h(x))$. □

Embedded Submanifolds of a Linear Space

Consider two possible issues:

1. If $T_x\mathcal{M} \subsetneq \ker(\mathrm{D}h(x))$. Since then there are vectors in $\ker(\mathrm{D}h(x))$ that are not tangent to $\mathcal{M}$ at x but can be used to approximate $\mathcal{M}$ near x.
   - For example, define cone shown previously by $h(x, y, z) = z^2 - x^2 - y^2$. At the tip $(0, 0, 0)$, we have $\mathrm{D}h(0, 0, 0) = [0\ 0\ 0]$ and $\ker \mathrm{D}h(0, 0, 0) = \mathbb{R}^3$. So any vector in $\mathbb{R}^3$ can be used to approximate the cone near the tip. But $v = (0, 0, 1)$ is not a tangent vector.

   Hence, we want $T_x\mathcal{M} = \ker(\mathrm{D}h(x))$.
2. Now $T_x\mathcal{M}$ is a linear space. So we also want dim $T_x\mathcal{M}$ to be constant for all $x \in \mathcal{M}$.

Suppose that dim $T_x\mathcal{M} = n$ for all $x \in \mathcal{M}$. Then, from the previous proposition, we have

$$n = \dim T_x\mathcal{M} = \dim \ker(\mathrm{D}h(x)) = d - \mathrm{rank}\mathrm{D}h(x).$$

Or $\mathrm{rank}\mathrm{D}h(x) = d - n$.

Conversely, we can show that if $\mathrm{rank}\mathrm{D}h(x) = d - n$, then the two conditions above are satisfied.

Embedded Submanifolds of a Linear Space

Definition 3

A subset $\mathcal{M}$ of $\mathbb{R}^d$ is an embedded submanifold of dimension n if for each $x \in \mathcal{M}$, there exists an open set $U \subset \mathbb{R}^d$ containing x and a smooth map $h : U \to \mathbb{R}^{d-n}$ such that

$$M \cap U = h^{-1}(\{0\}) \text{ and } \forall x \in \mathcal{M} \cap U, \operatorname{rank} \mathrm{D}h(x) = d - n.$$

Proposition 4

Using the convention that $\mathbb{R}^0 = \{0\}$, every open subset of $\mathbb{R}^d$ is a d-dimensional embedded submanifold of $\mathbb{R}^d$.

Proposition 5

Let $\mathcal{M} \subset \mathbb{R}^m$ and $\mathcal{N} \subset \mathbb{R}^n$ be embedded submanifolds of dimensions k and ℓ respectively. Then, $\mathcal{M} \times \mathcal{N} \subset \mathbb{R}^{m+n}$ is an embedded submanifold of dimension $k + \ell$.

Manifolds

Examples of Optimization on Manifolds

Eigenvalues and Singular Values

The problem of finding the largest eigenvalue of a symmetric matrix $A \in \mathbb{R}^{d \times d}$ is stated as

$$\max_{x \in \mathbb{R}^d} \langle x, Ax \rangle \quad \text{s.t. } \|x\|_2 = 1.$$

The constraint $\|x\|_2 = 1$ defines the unit sphere $S^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2^2 - 1 = 0\}$.

The function $h(x) = \|x\|_2^2 - 1$ is smooth and $\mathrm{D}h(x) = 2x^\top$, which has rank 1 for all $x \in S^{d-1}$. Hence, S^{d-1} is an embedded submanifold of $\mathbb{R}^d$ of dimension $d - 1$.

Eigenvalues and Singular Values

The problem of finding the largest eigenvalue of a symmetric matrix $A \in \mathbb{R}^{d \times d}$ is stated as

$$\max_{x \in \mathbb{R}^d} \langle x, Ax \rangle \quad \text{s.t. } \|x\|_2 = 1.$$

The constraint $\|x\|_2 = 1$ defines the unit sphere $S^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2^2 - 1 = 0\}$.

The function $h(x) = \|x\|_2^2 - 1$ is smooth and $\mathrm{D}h(x) = 2x^\top$, which has rank 1 for all $x \in S^{d-1}$. Hence, S^{d-1} is an embedded submanifold of $\mathbb{R}^d$ of dimension $d - 1$.

Similarly, the problem of finding the largest singular value of $M \in \mathbb{R}^{m \times n}$ can be written as

$$\max_{x \in S^{m-1}, y \in S^{n-1}} \langle x, My \rangle,$$

which is an optimization problem on the manifold $S^{m-1} \times S^{n-1}$.

Principal Component Analysis

Let $X \in \mathbb{R}^{d \times n}$ be the data matrix with n samples of dimension d. The goal is to find an orthogonal basis of dimension r that maximizes the weighted variance of the data when projected onto the basis.

$$\max_{U \in \mathbb{R}^{d \times r}} \operatorname{tr}(U^\top X W X^\top U) \quad \text{s.t. } U^\top U = I_r.$$

The set $\mathrm{St}(r, d) = \{U \in \mathbb{R}^{d \times r} \mid U^\top U = I_r\}$ is called the Stiefel manifold.

Gradient on Manifolds

Motivation

If $f : \mathbb{R}^d \to \mathbb{R}$, then $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$. We want to generalize in such a way that if $f : \mathcal{M} \to \mathbb{R}$, then $\mathrm{grad} f : \mathcal{M} \to \mathcal{M}$.

Recall that to define the gradient in $\mathbb{R}^d$, we need the differential and an inner product, which in turn needs a linear structure.

We already have $T_x\mathcal{M}$.

Differential on Manifolds

For $f : \mathbb{R}^d \to \mathbb{R}^q$, the differential $\mathrm{D}f(x) : \mathbb{R}^d \to \mathbb{R}^q$ defined by

$$\mathrm{D}f(x)(v) = \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t} = \frac{\mathrm{d}}{\mathrm{d}t} f(x+tv)\Big|_{t=0} = (f \circ \gamma)'(0)$$

This means how f changes when we move from x in the straight direction v.

The problem with manifolds is that the line $x + tv$ (for t in some interval) may not lie in $\mathcal{M}$.

But we can use a curve in $\mathcal{M}$.

Definition 6 (Differential)

Let $f : \mathcal{M} \to \mathcal{M}'$. The differential of f at $x \in \mathcal{M}$ is the linear map $\mathrm{D}f(x) : T_x\mathcal{M} \to T_{f(x)}\mathcal{M}'$ defined by

$$\mathrm{D}f(x)[v] = \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\bigg|_{t=0} = (f \circ \gamma)'(0).$$

Here, $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ is any smooth curve passing through x with velocity v i.e. $\gamma(0) = x$ and $\gamma'(0) = v$.

Here we need to check that $Df(x)[v]$ does not depend on the choice of γ. Details are given in the appendix.

Gradient on Manifolds

We need conditions under which $\mathrm{grad} f(x)$ is well-defined by the usual

$$\langle \mathrm{grad} f(x), v \rangle_x = \mathrm{D} f(x)[v], \quad \forall v \in T_x\mathcal{M}.$$

Definition 7 (Riemannian metric)

A Riemannian metric is an inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ that varies smoothly with x i.e. for any smooth vector fields X, Y on $\mathcal{M}$, the function $x \mapsto \langle X(x), Y(x) \rangle_x$ is smooth.

Gradient on Manifolds

Considering every $x \in \mathcal{M}$, we need the notion of tangent bundle

$$T\mathcal{M} = \{(x, v) \,|\, x \in \mathcal{M} \text{ and } v \in T_x\mathcal{M}\}.$$

Definition 8 (Gradient)

Let $f : \mathcal{M} \to \mathbb{R}$ be smooth on a Riemannian manifold $\mathcal{M}$. The gradient $\mathrm{grad} f$ is a vector field such that

$$\forall (x, v) \in T\mathcal{M}, \langle \mathrm{grad} f(x), v \rangle_x = \mathrm{D}f(x)[v].$$

Computation of Gradient - Retraction

Definition 9 (Retraction)

A retraction is a smooth map $R : T\mathcal{M} \to \mathcal{M} : (x, v) \mapsto R_x(v)$ such that each curve $c(t) = R_x(tv)$ satisfies $c(0) = x$ and $c'(0) = v$.

Proposition 10

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

$$\text{grad} f(x) = \nabla(f \circ R_x)(0)$$

Gradient Descent on Manifolds

First-order Optimality Conditions

We need the notion of critical points such that the Fermat theorem is reserved. Again, curves help: a point $x \in \mathcal{M}$ is a critical point of $f : \mathcal{M} \to \mathbb{R}$ if the velocity of any curve passing through x is 0.

Definition 11 (Critical point)

A point $x \in \mathcal{M}$ is a critical point of $f : \mathcal{M} \to \mathbb{R}$ if for any smooth curve $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ with $\gamma(0) = x$, we have

$$(f \circ \gamma)'(0) \geq 0.$$

The definition uses $(f \circ \gamma)'(0) \geq 0$ and that is equivalent: we can consider $t \mapsto c(t)$ and $t \mapsto c(-t)$.

Proposition 12

A point $x \in \mathcal{M}$ is a critical point of $f : \mathcal{M} \to \mathbb{R}$ if and only if $\operatorname{grad} f(x) = 0$.

Riemannian Gradient Descent

The framework is the iteration

$$x^{(k+1)} = R_{x^{(k)}}\left(-\alpha^{(k)}\text{grad}f(x^{(k)})\right),$$

where $x^{(0)}$ is initialized in $\mathcal{M}$ and $\alpha^{(k)} > 0$ is the step size.

Theorem 13

Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth, bounded-below function on a Riemannian manifold $\mathcal{M}$. Choose a retraction R such that grad$f$ is L-Lipschitz. If $\alpha^{(k)} \in [0, 2/L]$ for every step, then the sequence $\{x^{(k)}\}$ generated by Riemannian gradient descent satisfies

$$\lim_{k\to\infty} \|\text{grad}f(x^{(k)})\| = 0.$$

Thank you for listening !

Appendix