Université Gustave Eiffel

GRADIENT DESCENT ON MANIFOLDS

Data Science Project

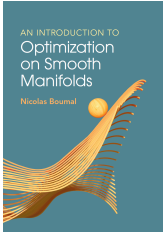
Students: CHAU Dang Minh
LAM Nhat Quan
Alhassane BAH

Outline

Introduction

Introduction

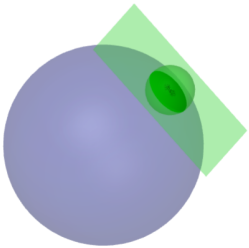
- Many optimization problems involve constraints that can be naturally modeled as manifolds.
- Gradient descent on manifolds extends traditional gradient descent methods to handle these constraints effectively.
- Presentation goals: build up the concept of gradient on manifolds and obtain basic convergence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



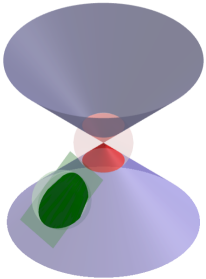
Embedded Submanifolds of a Linear Space

Embedded Submanifolds of a Linear Space

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 submanifolds: 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 to define tangent vectors (and also later definitions).

Definition (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

Now 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

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

By the rank-nullity theorem, we have $\operatorname{rank}(\mathrm{D}h(x)) \leq d - n$. Thus,

$$\dim(\ker(\mathrm{D}h(x))) = d - \operatorname{rank}(\mathrm{D}h(x)) \geq n.$$

On the other hand, $\dim(T_x\mathcal{M}) \leq n$.

Therefore, if there is $x \in \mathcal{M}$ such that $\operatorname{rank}(\mathrm{D}h(x)) < d - n$, then

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

That means 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 (we want to avoid this situation).

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.

Embedded Submanifolds of a Linear Space

Definition

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

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

We may add the theorem that this definition is equivalent to the diffeomorphism definition.

Examples of Optimization on Manifolds

I will take two from the book.

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 $\mathcal{M}$.

Differential on Manifolds

Definition (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 $\mathrm{grad} f$ 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 (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 $X, Y : \mathcal{M} \to T\mathcal{M}$, the function $x \mapsto \langle X(x), Y(x) \rangle_x$ is smooth.

Definition (Gradient)

Computation of Gradient - Retraction

Let $T\mathcal{M} = \{(x, v) \,|\, x \in \mathcal{M} \text{ and } v \in T_x\mathcal{M}\}$, called the tangent bundle.

Definition (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

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 $x \in \mathcal{M}$,

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

Should we add exponential map here? We will have to introduce geodesics first.

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 (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

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

Gradient Descent

The framework is the iteration

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \mathrm{grad} f(x^{(k)}),$$

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

I want to compare with projection method.

Convergence Results

Thank you for listening !

Appendix

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unproces
data that should have been added to the final page this extra page has been added to receiv

If you rerun the document (without altering it) this surplus page will go away, because LaTeX
now knows how many pages to expect for this document.