Continuous Optimization: Algorithms and Complexity
ORIE 6365

Gradients and Optimality Condition

Nikita Doikov

Cornell University
School of Operations Research and Information Engineering (ORIE)

January 28, 2026

Outline

- Gradients
- Optimality condition

Differentiable Functions

Definition

Function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called **differentiable** at $x \in \mathbb{R}^n$ if there exists a linear operator $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(x+h) = f(x) + \mathcal{L}[h] + o(\|h\|)$$

$$\Leftrightarrow \quad \lim_{\|h\|\to 0} \frac{\|f(x+h)-f(x)-\mathcal{L}[h]\|}{\|h\|} = 0.$$

- In a finite-dimensional space $\mathbb{R}^n$, all norms are *topologically equivalent* (does not matter which norm to pick)
- Assume there are two operators $\mathcal{L}_1$ and $\mathcal{L}_2$ which satisfy the definition. Subtracting one from another gives $(\mathcal{L}_1 - \mathcal{L}_2)[h] = o(\|h\|)$, which implies $\mathcal{L}_1 \equiv \mathcal{L}_2$
- Hence, operator $\mathcal{L}$ is unique (if exists). It is called the **derivative** of f at x.

Notations: $\quad Df(x) \equiv df(x) \equiv f'(x) \equiv \mathcal{L}$

Derivatives

Derivative is the best local approximation of a function f at x by a linear function:

$$f(x+h) \quad \approx \quad f(x) + Df(x)[h]$$

- Df has two arguments: $x \in \mathbb{R}^n$ (the point) and $h \in \mathbb{R}^n$ (the shift)
- Note that $Df(x)[h]$ is linear in h, but <u>not in x</u>. For any $h, u \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$:

$$Df(x)[\alpha h + \beta u] \quad = \quad \alpha Df(x)[h] + \beta Df(x)[u].$$

Gradients

We fix an inner product $\langle \cdot, \cdot \rangle$ in our space.

- For vectors $x, y \in \mathbb{R}^n$, we will always use the standard dot product:

$$\langle x, y \rangle \quad := \quad \sum_{i=1}^{n} x^{(i)} y^{(i)}.$$

- For matrices $X, Y \in \mathbb{R}^{n \times m}$, this leads to $\langle X, Y \rangle := \operatorname{tr}(X^\top Y)$

In **optimization**, we mostly work with functions $f : \operatorname{dom} f \to \mathbb{R}$.

The gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at x is a vector $\nabla f(x) \in \mathbb{R}^n$ such that

$$Df(x)[h] \quad = \quad \langle \nabla f(x), h \rangle.$$

So, it holds: $\quad f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|)$

- Note that $Df(x)[h]$ does not depend on a coordinate system
- The gradient $\nabla f(x)$ depends on a coordinate system
- We can use $Df(x)$ and $\nabla f(x)$ interchangeably, if the coordinate system is fixed

Directional Derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}$

Directional derivative. For any $h \in \mathbb{R}^n$, the directional derivative of f at point x along direction h is the derivative of the univariate function $\varphi(t) = f(x + th)$ at zero:

$$\frac{\partial f(x)}{\partial h} \quad := \quad \varphi'(0) \quad = \quad \lim_{t \to 0} \frac{f(x+th) - f(x)}{t}.$$

Proposition. For a differentiable function f, its directional derivative can be computed as

$$\frac{\partial f(x)}{\partial h} \quad = \quad Df(x)[h] \quad = \quad \langle \nabla f(x), h \rangle \qquad \textbf{(check!)}$$

Computing Gradients: Two Ways

1. Construct $\nabla f(x)$ coordinate-wise **(hard way):**
   - Compute all partial dertivatives $\frac{\partial f(x)}{\partial x^{(i)}}$ for all coordinate directions $e_1, \dots, e_n \in \mathbb{R}^n$
   - Combine them into vector:

$$\nabla f(x) \quad = \quad \left(\frac{\partial f(x)}{\partial x^{(1)}}, \dots, \frac{\partial f(x)}{\partial x^{(n)}} \right)^\top \quad \in \quad \mathbb{R}^n.$$

   - Theoretically is completely fine. Computationally hard in practice!
2. Think of $\nabla f(x)$ as the derivative representation **(coordinate-free way):**
   - Compute $Df(x)[h]$ for an arbitrary $h \in \mathbb{R}^n$.
   - Then find the vector $\nabla f(x) \in \mathbb{R}^n$ such that $Df(x)[h] \equiv \langle \nabla f(x), h \rangle$.
   - Often is much easier. Especially useful for **matrix functions** and for **neural networks**.

Example

$$f(X) = \frac{1}{2}\|X\|_F^2 \quad \Rightarrow \quad Df(X)[H] = \mathrm{tr}(X^\top H) \quad \Rightarrow \quad \nabla f(X) = X$$

Exercise: consider $f(X) = \frac{1}{2}\|AX - B\|_F^2$

The Gradient: Summary

For a function $f : \operatorname{dom} f \to \mathbb{R}$, the gradient vector $\nabla f(x)$ is the representation of the derivative, which depends on the choice of the coordinate system and is defined by

$$f(x+h) \quad = \quad f(x) + \langle \nabla f(x), h \rangle + o(\|h\|)$$

- $\nabla f(x)$ has the same shape as x (a vector, a matrix, multiple tensors — layers in neural networks, ...)

Gradients are used

- as the main search direction in optimization algorithms
- as optimality conditions for solutions

First-Order Optimality Condition

Definition

Point $x^\star$ is called a **local minimum** of f if there exist a neighborhood $x^\star \in U$ (an open set) such that

$$f(x^\star) \quad \leq \quad f(x), \qquad \forall x \in U$$

- NB: if $U \equiv \operatorname{dom} f$, then $x^\star$ is a *global minimum* of f.

Theorem

Let $x^\star$ be a local minimum of a differentiable function f. Then,

$$\nabla f(x^\star) \quad = \quad 0.$$

Proof. Assume $\nabla f(x^\star) \neq 0$. Take $h := -\alpha \nabla f(x^\star)$ with $\alpha > 0$, and consider

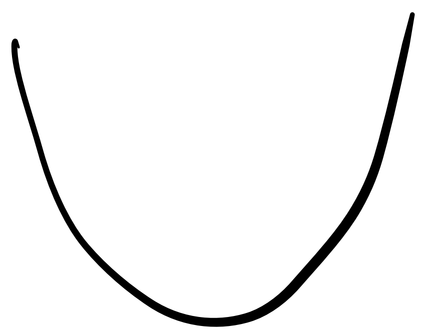
$$f(x^\star + h) \quad = \quad f(x^\star) - \alpha \|\nabla f(x^\star)\|^2 + o(\alpha).$$

For sufficiently small α, we get: $f(x^\star + h) \leq f(x^\star) - \frac{\alpha}{2}\|\nabla f(x^\star)\|^2 < f(x^\star)$. **(?!)** □

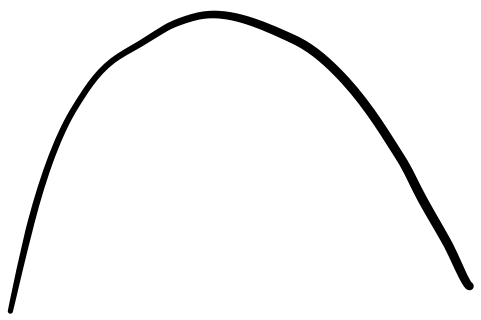
Stationary points

A point $x^\star$ such that $\nabla f(x^\star) = 0$ called a **stationary point**

local minimum



local maximum



saddle points