Lecture 1–2: Optimization Background

Yudong Chen

1 Introduction

Our standard optimization problem

$$\min_{x \in \mathcal{X}} f(x) \tag{P}$$

- x: a vector, optimization/decision variable
- $\mathcal{X}$: feasible set
- $f(x)$ objective function, real-valued
- $\max_x f(x) = -\min_x \{-f(x)\}$

The (optimal) value of (P):

$$\text{val}(\text{P}) = \inf_{x \in \mathcal{X}} f(x).$$

To fully specify (P), we need to specify

- vector space, feasible set, objective function;
- what it means to solve (P).

1.1 Can we even hope to solve an arbitrary optimization problem?

Example 1. Suppose we want to find positive integers x, y, z satisfying

$$x^3 + y^3 = z^3.$$

Can be formulated as a (continuous) optimization problem (P_F):

$$\begin{aligned} \min_{x,y,z,n} \quad & (x^n + y^n - z^n)^2 \\ \text{s.t.} \quad & x \ge 1, y \ge 1, z \ge 1, n \ge 3 \\ & \sin^2(\pi n) + \sin^2(\pi x) + \sin^2(\pi y) + \sin^2(\pi z) = 0. \end{aligned} \tag{P_F}$$

If we could certify whether $\text{val}(\text{P}_\text{F}) \neq 0$, we would have found a proof for Fermat's Last theorem (1637):

> For any $n \ge 3$, $x^n + y^n = z^n$ has no solutions over positive integers.

Proved by Andrew Wiles in 1994, a major mathematical breakthrough.

Example 2. In unconstrained optimization, there may exist many local minima like in the following picture. It is in general hard to find the global minima.

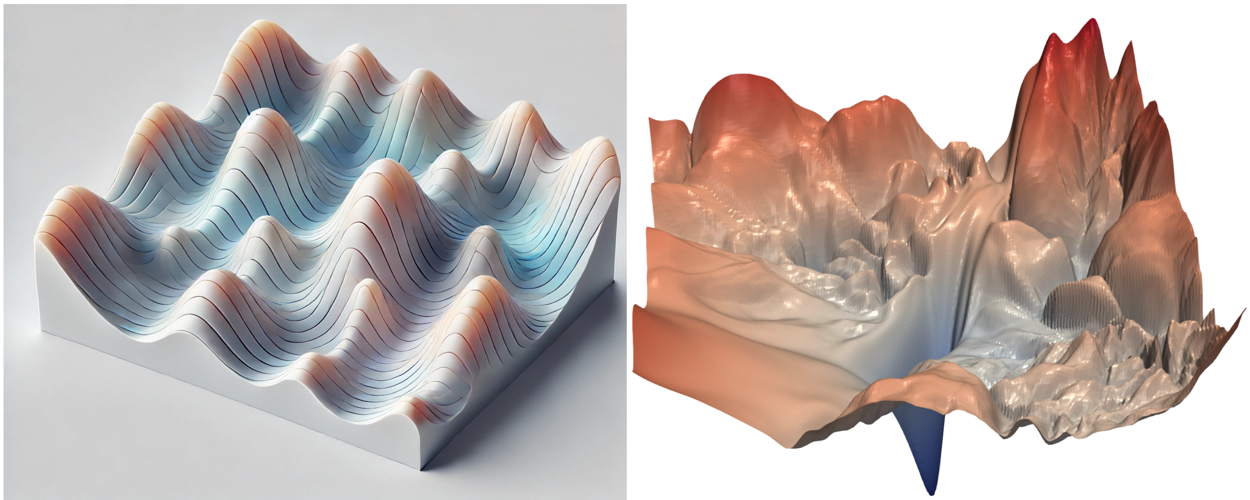


Figure 1: Left: an example of nonconvex function. Right: loss surfaces of ResNet-56 without skip connections (`https://arxiv.org/pdf/1712.09913.pdf`).

Therefore, we cannot hope for solving an arbitrary optimization problem.
We need some structure.

2 Specifying the optimization problem

2.1 Vector space

$(\mathbb{R}^d, \|\cdot\|)$: normed vector space, "primal space".
This is where the optimization variable and the feasible set live.

- The variable x is a (column) vector in $\mathbb{R}^d$.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}.$$

- The norm tells us how to measure distances in $\mathbb{R}^d$.

Most often, we will use the Euclidean norm $\|x\| = \|x\|_2 = \left(\sum_{i=1}^d x_i^2\right)^{1/2}$.

We sometimes also consider ℓ_p norm $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ for $p \geq 1$.

- $\|x\|_1 = \sum_i |x_i|$,
- $\|x\|_\infty = \max_{1 \leq i \leq d} |x_i|$.

(Try to plot the unit balls of $\ell_2, \ell_1, \ell_\infty$ norms.)

We will use $\langle \cdot, \cdot \rangle$ to denote the standard inner product

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{d} x_i y_i.$$

When we work with $\left(\mathbb{R}^d, \|\cdot\|_p\right)$, view $\langle y, x \rangle$ as the value of a linear function y at x. So, if we are measuring the length of x using the $\|\cdot\|_p$, we should measure the length of y using $\|\cdot\|_q$,where $\frac{1}{p} + \frac{1}{q} = 1$.

Definition 1 (Dual norm)**.** The dual norm of $\|\cdot\|$ is given by

$$\|z\|_* := \sup_{\|x\| \le 1} \langle z, x \rangle .$$

From the definition we immediately have the

Proposition 1 (*Holder's Inequality*)**.** *For all $z, y \in \mathbb{R}^d$:*

$$|\langle z, x \rangle| \le \|z\|_* \cdot \|x\| .$$

Proof. Fix any two vectors x, z. Assume $x \neq 0, z \neq 0$; otherwise the inequality trivially holds. Define $\hat{x} = \frac{x}{\|x\|}$. Then

$$\|z\|_* \ge \langle z, \hat{x} \rangle = \frac{\langle z, x \rangle}{\|x\|}$$

and hence $\langle z, x \rangle \le \|z\|_* \cdot \|x\|$. Applying same argument with x replaced by $-x$ proves $-\langle z, x \rangle \le \|z\|_* \cdot \|x\|$. □

Example 3. $\|\cdot\|_p$ and $\|\cdot\|_q$ are duals when $\frac{1}{p} + \frac{1}{q} = 1$. In particular, $\|\cdot\|_2$ is its own dual; $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are dual to each other.

In $\mathbb{R}^d$, all ℓ_p norms are equivalent. In particular,

$$\forall x \in \mathbb{R}^d, p \ge 1, r > p : \quad \|x\|_r \le \|x\|_p \le d^{\frac{1}{p} - \frac{1}{r}} \|x\|_r .$$

However, choice of norm affects how algorithm performance depends on dimension d.

2.2 Feasible set

The feasible set

$$\mathcal{X} \subseteq \mathbb{R}^d$$

specifies what solution points we are allowed to output.

If $\mathcal{X} = \mathbb{R}^d$, we say that (P) is *unconstrained*. Otherwise we say that (P) is *constrained*.

$\mathcal{X}$ can be specified in multiple ways:

- as an abstract geometric body (a ball, a box, a polyhedron, a convex set)
- via functional constraints:

$$\begin{aligned} g_i(x) &\le 0, i = 1, 2, \ldots, m, \\ h_i(x) &= 0, i = 1, \ldots, p \end{aligned}$$

Note that $f_i(x) \ge C$ is equivalent to taking $g_i(x) = C - f_i(x)$.

Example 4.

$$\mathcal{X} = \mathcal{B}_2(0,1) = \text{unit Euclidean ball}$$
$$\mathcal{X} = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i^2 \leq 1 \right\}.$$

In this class, we will always assume that $\mathcal{X}$ is *closed*.

Heine-Borel Theorem: $\mathcal{X} \subseteq \mathbb{R}^d$ is closed and bounded if and only if it is compact (if $\mathcal{X} \subset \bigcup_{\alpha \in A} U_\alpha$ for some family of open sets $\{U_\alpha\}$,then there there exists a finite subfamily $\{U_{\alpha_i}\}_{i=1}^n$ such that $\mathcal{X} \subseteq \bigcup_{1 \leq i \leq n} U_{\alpha_i}$.)

Weierstrass Extreme Value Theorem: If $\mathcal{X}$ is compact and f is a function that is defined and continuous on $\mathcal{X}$, then f attains its extreme values on $\mathcal{X}$.

What if $\mathcal{X}$ is not bounded? Consider $f(x) = e^x$. Then $\inf_{x \in \mathbb{R}} f(x) = 0$, but not attained.

When we work with unconstrained problems, we will normally assume that f is bounded from below.

Convex sets: Except for some special cases, we often assume that the feasible set $\mathcal{X}$ is convex (but we will consider both convex and nonconvex objective functions f).

Definition 2 (Convex set)**.** A set $\mathcal{X} \subseteq \mathbb{R}^d$ is *convex* if

$$\forall x, y \in \mathcal{X}, \forall \alpha \in (0,1) : (1-\alpha)x + \alpha y \in \mathcal{X}$$

Try to draw a picture.

We cannot hope to deal with arbitrary nonconvex constraints. E.g., $x_i(1-x_i) = 0 \iff x_i \in \{0,1\}$, integer programs.

2.3 Objective function

This represents the "cost" or "loss", which we want to minimize.

Extended real valued functions:

$$f : \mathcal{D} \to \mathbb{R} \cup \{-\infty, \infty\} \equiv \bar{\mathbb{R}}.$$

Here f is defined on $\mathcal{D} \subseteq \mathbb{R}^d$. Can extend the definition of f to all of $\mathbb{R}^d$ by assigning the value $+\infty$ at each point $x \in \mathbb{R}^d \setminus \mathcal{D}$.

Effective domain:

$$\text{dom}(f) = \left\{ x \in \mathbb{R}^d : f(x) < \infty \right\}$$

In the sequel, domain means effective domain.

"Linear and nonlinear optimization" $\approx$ "continuous optimization" (as contrast to discrete/combinatorial optimization)

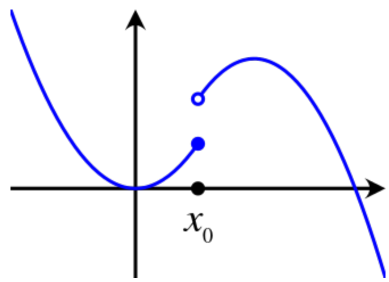
2.3.1 Lower semicontinuous functions

We mostly assume f to be continuous. This can be relaxed slightly as follows.

Definition 3. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is said to be *lower semicontinuous* (l.s.c) at $x_0 \in \mathbb{R}^d$ if

$$f(x_0) \le \liminf_{x \to x_0} f(x).$$

We say that f is l.s.c. on $\mathbb{R}^d$ if it is l.s.c. at every point $x \in \mathbb{R}^d$.



This definition is mainly useful for allowing indicator functions.

Example 5. Verify yourself: The indicator function of a closed set $\mathcal{X}$, defined as

$$I_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X}, \\ \infty, & x \notin \mathcal{X}, \end{cases}$$

is l.s.c. Using $I_{\mathcal{X}}$ we can write

$$\min_{x \in \mathcal{X}} f(x) \equiv \min_{x \in \mathbb{R}^d} \left\{ f(x) + I_{\mathcal{X}}(x) \right\},$$

thereby unifying constrained and unconstrained optimization.

2.3.2 Lipschitz-continuous and smooth functions

Unless we are abstracting away constraints like in Example 5, the least we will assume about f is that it is continuous.

Sometimes we consider stronger assumptions of f.

Definition 4. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is said to be

1. Lipschitz-continuous on $\mathcal{X} \subseteq \mathbb{R}^d$ (w.r.t. the norm $\|\cdot\|$) if there exists $M < \infty$ such that

$$\forall x, y \in \mathcal{X} : |f(x) - f(y)| \le M \|x - y\|.$$

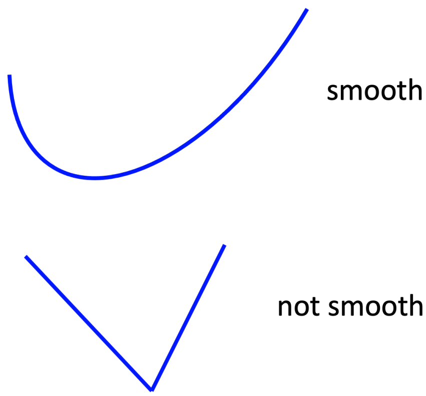
2. Smooth on $\mathcal{X} \subseteq \mathbb{R}^d$ (w.r.t. the norm $\|\cdot\|$) if f's gradient is Lipschitz-continuous, i.e., there exists $L < \infty$ such that[1]

$$\forall x, y \in \mathcal{X} : \|\nabla f(x) - \nabla f(y)\|_* \le L \|x - y\|.$$

(Gradient: $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$.)

[1] This definition can be viewed a quantitative version of C^1-smoothness.

- Picture:



In $\mathbb{R}^d$, Lipschitz-continuity in some norm implies the same for every other norm, but M may differ.

Example 6. $f(x) = \frac{1}{2}\|x\|_2^2$ is 1-smooth on $\mathbb{R}^2$ w.r.t. $\|\cdot\|_2$. The log-sum-exp (or softmax) function $f(x) = \log\left(\sum_{i=1}^d \exp(x_i)\right)$ is 1-smooth on $\mathbb{R}^d$ w.r.t. $\|\cdot\|_\infty$.

Example 7. Function that is continuously differentiable on its domain but not smooth in the above sense:

$$f(x) = \frac{1}{x}$$
$$\mathrm{dom}(f) = \mathbb{R}_{++}$$

2.3.3 Convex functions

Definition 5. $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is convex if $\forall x, y \in \mathbb{R}^d, \forall \alpha \in (0,1)$:

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y).$$

A picture.

Lemma 1. *$f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only its epigraph*

$$epi(f) := \left\{(x, a) : x \in \mathbb{R}^d, a \in \mathbb{R}, f(x) \leq a\right\} \subseteq \mathbb{R}^{d+1}$$

is convex.

Proof. Follows from definitions. Left as exercise. □

Definition 6. We say that a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper if $\exists x \in \mathbb{R}^d$ s.t. $f(x) \in \mathbb{R}$.

Lemma 2. *If $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper and convex, then $\mathrm{dom}(f)$ is convex.*