Convex Optimization

Jaden Wang

January 20, 2021

Contents

Chapter 1

Theoretical Foundation

1.1 Introduction

An optimization problem looks like

$$\min_{x \in C} f(x)$$

where $f(x)$ is the **objective function** and $C \subseteq \mathbb{R}^n$ is the **constraint set**. C might look like

$$C = \{x : g_i(x) \leq 0 \ \forall \ i = 1, \ldots, m\}.$$

Remark. We can always turn a maximization problem into a minimization problem as the following:

$$\min_x f(x) = -\max_x -f(x).$$

Therefore, WLOG, we will stick with minimization.

Example. An assistant professor earns \$100 per day, and they enjoy both ice cream and cake. The optimization problem aims to maximize the utility (*e.g.* happiness) of ice cream $f_1(x_1)$ and of cake $f_2(x_2)$. The constraints we have is that $x_1 \geq 0, x_2 \geq 0$, and $x_1 + x_2 \leq 100$.

To maximize both utility, it might be natural to define

$$F(\operatorname{vec} x) = \begin{pmatrix} f_1(x_1) \\ f_2(x_2) \end{pmatrix}, \operatorname{vec} x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

and maximize F. However, this isn't a well-defined problem, because *there is no total order on* $\mathbb{R}^n$! That is, we don't have a good way to compare whether a vector is bigger than another vector, except in the cases when the same direction of inequality can be achieved for all components of two vectors and a partial order can be established. For this kind of **multi-objective** optimization problem, we can look for Pareto-optimal points in these special cases. We can also try to convert the output into a scalar as the following:

$$\min_x f_1(x) + \lambda \cdot f_2(x_2)$$

for some $\lambda > 0$ that reflects our preference for cake vs ice cream. But this can be subjective.

Thus, For the remainder of this class, we are only going to assume $f : \mathbb{R}^n \to \mathbb{R}$.

Moreover, for $f : \mathbb{R} \to \mathbb{R}$, it's very easy to solve by using root finding algorithms or grid search. So since interesting problems occur with vector inputs, we will simply use x to represent vectors.

Notation. min asks for the minimum value, whereas arg min asks for the minimizer that yields the minimum value.

1.1.1 Lipschitz continuity

Example. Let's consider a variant of the Dirichlet function, $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \begin{cases} x & \text{if } x \in \mathbb{Q} \\ 1 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

Then the solution to the problem

$$\min_{x \in [0,1]} f(x) = 0$$

is $x = 0$ by observation. However, the function is not smooth and a small perturbation can yield wildly different values. Thus, it is not tractable to solve this numerically.

This requires us to add a smoothness assumption:

Definition: Lipschitz continuity

$f : \mathbb{R}^n \to \mathbb{R}$ is ***L*-Lipschitz continuous** with respect to a norm $\|\cdot\|$ if for all $x, y \in \mathbb{R}^n$,

$$|f(x) - f(y)| \leq L \cdot \|x - y\|.$$

Note. Lipschitz continuity implies continuity and uniform continuity. It is a stronger statement because it tells us *how* the function is (uniformly) continuous. However, it doesn't require differentiability.

Definition: l_p norms

For $1 \leq p < \infty$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

For $p = \infty$,

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|.$$

Remark. $\|x\|_1$ and $\|x\|_2^2$ have separable terms as they are sums of their components. $\|x\|_2^2$ is also differentiable which makes it the nicest norm to optimize.

Example. Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. Let $C = [0,1]^n$, *i.e.* in $\mathbb{R}^2$, C is a square. To solve the problem

$$\min_{x \in C} f(x),$$

since we have few assumption, there is no better method (in the worst case sense) than the **uniform grid method**. The idea is that we pick $p+1$ points in each dimension, *i.e.* $\{0, \frac{1}{p}, \frac{2}{p}, \dots, 1\}$, so we would have $(p+1)^n$ points in total.

Let x^* be a global optimal point, then there exists a grid point $\widetilde{x}$ s.t.

$$\|x^* - \widetilde{x}\|_\infty \le \frac{1}{2} \cdot \frac{1}{p}.$$

Thus by Lipschitz continuity,

$$\begin{aligned} |f(x^*) - f(\widetilde{x})| &\le L \cdot \|x^* - \widetilde{x}\|_\infty \\ &\le \frac{1}{2}\frac{L}{p} \end{aligned}$$

So we can find $\widetilde{x}$ by taking the discrete minimum of all $(p+1)^n$ grid points.

In (non-discrete) optimization, we usually can't exactly find the minimizer, but rather find something very close.

Definition: epsilon-optimal solution

x is a **ε-optimal solution** to $\min_{x \in C} f(x)$ if $x \in C$ and

$$f(x) - f^* \le \varepsilon$$

where $f^* = \min_{x \in C} f(x)$.

Our uniform grid method gives us an ε-optimal solution with $\varepsilon = \frac{L}{2p}$, and requires $(p+1)^n$ function evaluations. Writing p in terms of ε, we have $p = \frac{L}{2\varepsilon}$

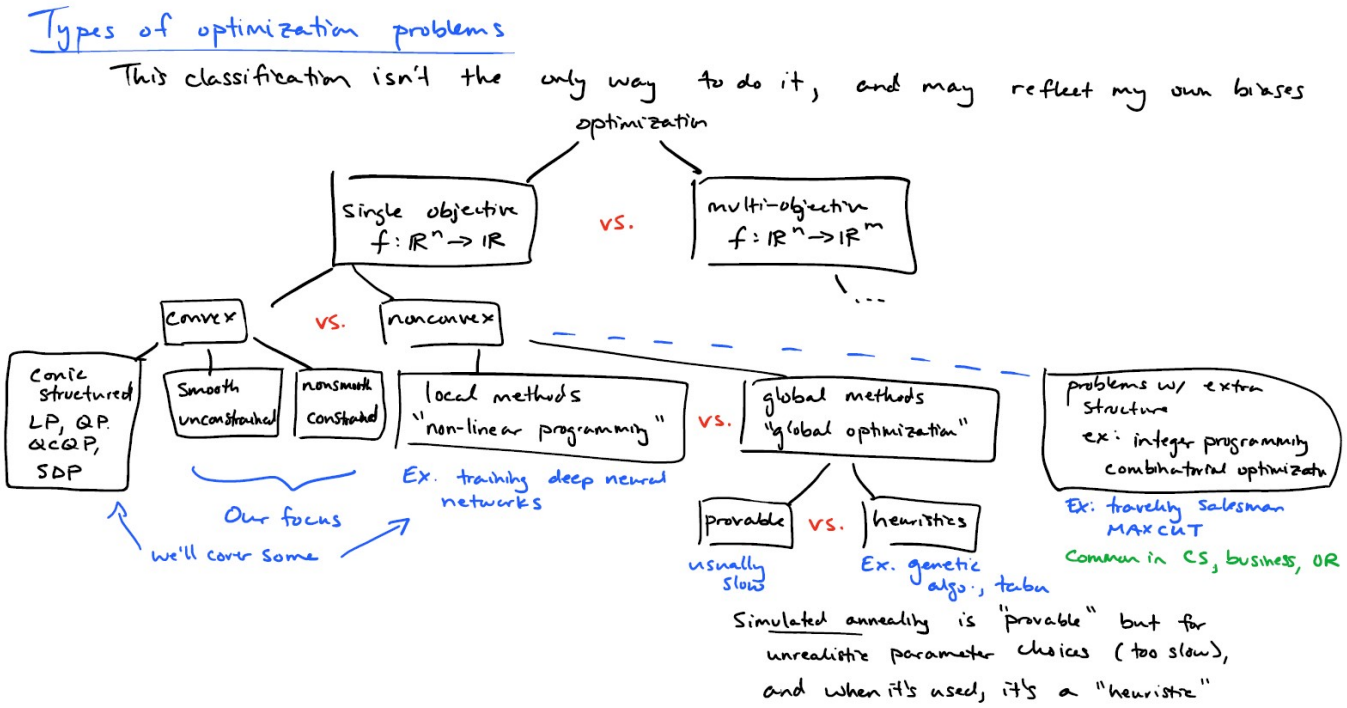
so equivalently it requires $\left(\frac{2L}{\varepsilon}+1\right)^n$ function evaluations, which approximately is ε^{-n}.

For $\varepsilon = 10^{-6}$, $n = 100$, it requires 10^{600} function evaluations. This is really bad!

Take-aways from this example:

- curse-of-dimensionality: there can be trillions of variables in a Google Neural Network. It would be intractable using the grid method.
- we need more assumptions to allow us to use more powerful methods.

1.1.2 categorization



1.2 Minimizers

We are given a generic problem $\min_{x\in C} f(x), C \subseteq \mathbb{R}^n$. Then a **feasible point** x means $x \in C$. A **solution** or **minimizer** or **global minimizer** x^* means

1) $x^* \in C$

2) $\forall\ y \in C, f(x^*) \leq f(y)$

It might not be unique, *i.e.* $x^* \in \arg\min_{x\in C} f(x)$.

Example.

$$\min_{x\in\mathbb{R}} f(x) \text{ where } f(x) = 0\ \forall\ x.$$

Sometimes the solution may not exist (even for convex problems).

Example.

$$\min_{x\in(0,1)} x^2.$$

x^* is a **local minimizer** if x^* is feasible and there exists an $\varepsilon > 0$ s.t. $f(x^*) \leq f(y)\ \ \forall\ y \in C \cap B_\varepsilon(x^*) \coloneqq \{y : \|y - x^*\|\} < \varepsilon$. A **strict local minimizer** simply doesn't achieve equality. x^* is an **isolated local minimum** if it is a local minimum and no other local minimum are nearby.

Example (isolated but not strict)**.**

$$f(x) = \begin{cases} x^4 \cos\left(\frac{1}{x}\right) + 2x^4 & x \neq 0 \\ 0 & x = 0 \end{cases}$$

$x^* = 0$ is strict but not isolated due to the rapid oscillation near $x = 0$.

Notation. $f \in \mathcal{C}^3$ means f, f', f'', f''' all exist and are continuous. $f \in \mathcal{C}^3(\mathbb{R}^n)$ means $f, \nabla f, \nabla^2 f, \nabla^3 f$ all exists and are continuous.

1.2.1 connections with Calculus 1

Recall that in Cal 1, we first find the stationary/critical points in the domain. Then we add the boundary points and minimize over the small (finite) set of candidates.

In high-dimension optimization, we cannot check critical points and the boundary separately because the set of points in the boundary becomes infinite. More-

over, there can be infinite critical points too.

Necessary condition: if x^* is a local or global minimizer and $C = \mathbb{R}^n$, then x^* is a **critical point**. But the converse is false.

Notation. The boundary of C is denoted as $\partial C \coloneqq \overline{C} \setminus \operatorname{int} C$.

If x^* is a critical point but is not a local or global minimizer, then it's a **saddle point**.

Theorem: Weierstrass Theorem

If f is continuous and C is compact, then f achieves its infimum over C. That is,

$$\inf_{x \in C} f(x) = \min_{x \in C} f(x).$$

Note. This is the same as the Extreme Value Theorem.

Proof

First let's prove a claim.

Claim. Every compact set K is closed and bounded.

Closed: suppose not, the compact set K doesn't contain all its limit points. That is, there exists a limit point $x \notin K$ s.t. a sequence $(x_n) \subseteq K$ converges to x. But that also means that all subsequences of (x_n) converges to $x \notin K$ as well, contracting with the definition of compactness that for every sequence in K there exists a subsequence that converges inside K.

Bounded: suppose not, for all $M > 0$, there exists a $x \in K$ s.t. $\|x\| > M$. This allows us to find a sequence $(x_n) \subseteq K$ s.t. $x_n > n$. This way every subsequnece is also unbounded and cannot converge, contradicting with the definition of sequential compactness.

Now let's begin proof proper. Since C is compact and f is continuous, the image of C under f, $f(C)$, is also compact (this follows from sequential definition of continuity). By the claim $f(C)$ is bounded and closed, meaning that it has an infimum (completeness axiom) and contains the infimum (closed). Thus, f achieves its infimum over C. □

Remark. It would be nice if our constraints C are compact. But at the very least, we want our constraint sets to be closed. For example, $\|Ax - b\| \leq \varepsilon$ instead of $\|Ax - b\| < \varepsilon$.

Several things to note about the feasible set C:

If $C = \emptyset$, the problem is infeasible. This is not always easy to spot.

In this class, C will usually be convex and not integral, *i.e.* $\mathbb{Z}^n$.

Integral constraint is problematic because the optimal integer solution might not be at all close to the optimal real solution, so we cannot obtain it by solving for the real solution first and then round it.