CSE 599: Interplay between Convex Optimization and Geometry **Winter 2018**

Lecture 1: Introduction

Lecturer: Yin Tat Lee

Disclaimer: *Please tell me any mistake you noticed.*

1.1 Course Information

Objective: Recently, there are lots of progress on getting faster algorithms for various theory optimization problems, especially for combinatorial optimization. For example, the fastest algorithms for maximum flow, matroid intersection, submodular minimization are improved for the first time in decades. For theory, we will cover some of the optimization techniques used in those results and, more importantly, some basic mathematical tools underlying them. As for application, many of these algorithms are very different from what people are using and have potential to be disruptive.

Course requirement: 2 problem sets (exercises in lecture notes $\neq$ problem sets) (50%), final project (50%)

Course project: It can be anything related to convexity. For example:

1. A research project related to convexity. It can be any open problem mentioned in the class, any problem you are interested in or any part of your own ongoing research.
2. A survey type report on something related to convexity.
3. An implementation project on some ideas learned from the class or more generally, geometric methods for convex optimization.
4. Others

(Optional) I encourage you to send me a 1 page project proposal by Feb 21, describing what you want to do and why you think it is interesting.

Prerequisites: I try to be as self-contained as possible. However, I assume you are familiar with linear algebra, multivariate calculus, probability, inequalities and algorithms. Most importantly, I expect you love mathematics!

1.2 Course Overview

Convexity is a main source of easy problems. To exaggerate, some(?) even think that optimization problem is easy if and only if it can be rewritten as a convex problem. See Figure 1.1 for a list of convex problems with increasing difficulty. In this course, we will first cover the cutting plane methods. These methods show that all convex optimization problems are in polynomial time (with some oracle assumptions). Next, we will cover the first order method. Although it is not appeared in the Figure 1.1, it is the method of choice in many machine learning applications. We will cover some which have more impact on theory. Then, we will cover different methods for solving linear systems, including the Cholesky decomposition for Laplacian matrices. Finally, we will talk about interior point methods.

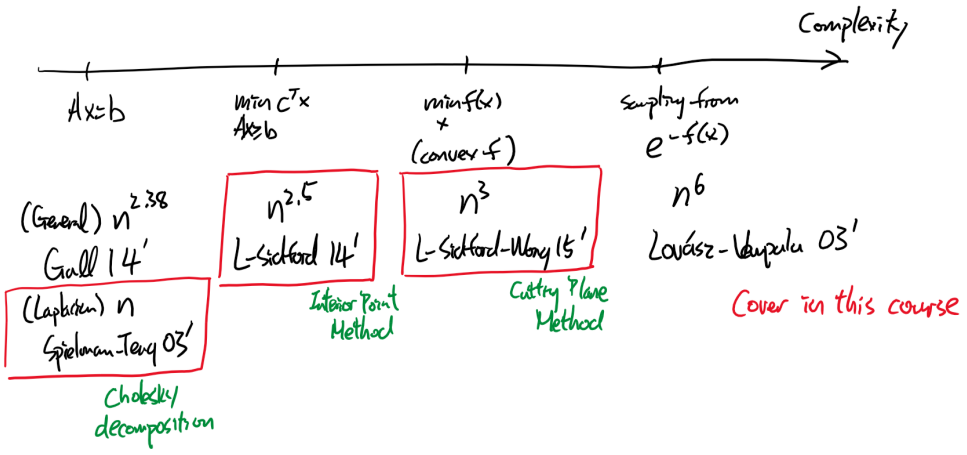

Figure 1.1: A list of convex problems with increasing difficulty

Although convex optimization is studied for more than a half century, there are still many basic open problems. This is an example:

Problem. Can we solve a random sparse linear system with $O(n)$ non-zeros faster than n^2 time? (One can do it $O(n^2 \log(n/\varepsilon))$ time by Chebyshev polynomial. Not 100% sure.)

In every lecture, I will give some open problem(s). Some are famous and some are not. If you are interested in the problem, feel free to talk to me. I have some thought on many of the problems. If you make enough progress on any "problem" (not exercise) in the lecture note, then you will get 100% score for the course and do not need to any problem set or project.[1]

1.3 Convex sets

Definition 1.3.1. A set $K \subset \mathbb{R}^n$ is convex if for any $x, y \in K$ and any $0 \leq \theta \leq 1$, we have $\theta x + (1-\theta) y \in K$. A set K is a polyhedron if $K = \{x \in \mathbb{R}^n \ : \ Ax \leq b\}$ for some matrix A and some vector b.

The main reason convex set is so nice is that we can also use a hyperplane to separate a point outside the set and the set itself. Such hyperplane allows us to do binary search to find a point in a convex set. As we will see in the next lecture, such hyperplane can be computed efficiently.

Theorem 1.3.2. *Let K be a convex set in $\mathbb{R}^n$ and $y \notin K$. Then, we can find a $\theta \in \mathbb{R}^n$ such that*

$$\langle \theta, y \rangle \geq \max_{x \in K} \langle \theta, x \rangle .$$

If K is closed, the inequality is strict.

Remark. In infinite dimension, this theorem is called Hahn–Banach theorem and is useful for functional analysis.

Proof. Let x^* be a point in K closest to y, namely $x^* \in \arg\min_{x \in K} \|x - y\|^2$. Using convexity of K, for any $x \in K$ and any $0 \leq t \leq 1$, we have that

$$\|(1-t)x^* + tx - y\|^2 \geq \|x^* - y\|^2 .$$

[1] I took a theory course in undergrad with a similar policy. I solved a problem and got a STOC paper.

Expand the left-hand side, we have

$$\begin{aligned}\|(1-t)x^* + tx - y\|^2 &= \|x^* - y + t(x - x^*)\|^2 \\ &= \|x^* - y\|^2 + 2t\langle x^* - y, x - x^*\rangle + O(t^2).\end{aligned}$$

Taking $t \to 0^+$, we have that

$$\langle x^* - y, x - x^*\rangle \geq 0 \text{ for all } x \in K.$$

□

Another application of this theorem is that it shows polyhedrons is essentially as general as convex sets. Throughout this course, many theorems about polyhedrons can be generalized to convex sets for this reason.

Corollary 1.3.3. *Any closed convex set K can be written as the intersection of half space as follows*

$$K = \bigcap_{\theta \in \mathbb{R}^n} \left\{ x : \ \langle \theta, x\rangle \leq \max_{y \in K} \langle \theta, y\rangle \right\}.$$

Namely, any convex set "is"[2] a polyhedron with possibly infinitely many constraints.

There are many important convex sets and here I only list some that appears in this course.

Example. Convex sets: Polytope $\{Ax \leq b\}$, Ellipsoid $\{x^\top Ax \leq 1\}$, positive definite cone $\{A \in \mathbb{R}^{n\times n} \text{ such that } A \succeq 0\}$, norm ball $\{x : \ \|x\|_p \leq 1\}$ for all $p \geq 1$.

1.4 Convex functions

Definition 1.4.1. A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex if for all $x, y \in \mathbb{R}^n$ and $0 \leq t \leq 1$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Certainly, the most important class of convex functions is convex set :).

Definition 1.4.2. Given a convex set K, we define

$$\ell_K = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{elses} \end{cases}.$$

If we let $\mathrm{dom} f \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \ f(x) < +\infty\}$, then the definition shows that $\mathrm{dom} f$ is a convex set. By looking on the set of points above the graph, we obtain a convex set called epigraph.

Definition 1.4.3. The epigraph of f is $\mathrm{epi} f \stackrel{\text{def}}{=} \{(x,t) \subset \mathbb{R}^n \times \mathbb{R} : \ t \geq f(x)\}$.

Exercise 1.4.4. A function f is convex if and only if $\mathrm{epi} f$ is a convex set.

This characterization shows that $\min_x f(x)$ is same as $\min_{(t,x)\in \mathrm{epi} f} t$. Therefore, convex optimization is same as finding an extreme point of a convex set. Another important feature of convex set is the following:

Exercise 1.4.5. The sublevel set $\{x \in \mathbb{R}^n : \ f(x) \leq t\}$ is convex.

[2] Formally, polyhedron only has finitely many constraints.

In particular, this shows that the set of minimizer are connected. Therefore, local minimum is same as global minimum. We note that the converse is not true. We call a function is quasi-convex if every sublevel set is convex.

Similar to the convex set, we have a separation theorem similar to Theorem 1.3.2. This shows that binary search on convex function is also possible.

Theorem 1.4.6. *Let f be a continuously differentiable function. Then, f is convex if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ for all } x, y.$$

Proof. Fix any $x, y \in \mathbb{R}^n$. Let $g(t) = f(tx + (1-t)y)$. Suppose f is convex, so is g. Then, we have

$$g(t) \leq (1-t)g(0) + tg(1)$$

which implies that

$$g(1) \geq g(0) + \frac{g(t) - g(0)}{t}.$$

Taking $t \to 0^+$, we have that $g(1) \geq g(0) + g'(0)$. In other words, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle .$$

Since we will not use the converse, we left it as an exercise. □

Checking if a function is convex can be difficult in general. However, if the function is twice differentiable, it can be done by a simple calculus as follows.

Theorem 1.4.7. *Let f be a twice continuously differentiable function on $\mathbb{R}^n$. Then, f is convex if and only if*

$$\nabla^2 f(x) \succeq 0 \text{ for all } x \in \mathbb{R}^n.$$

Remark. The Hessian $\nabla^2 f$ of f is the $n \times n$ matrix defined by $(\nabla^2 f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$. We write $A \succeq 0$ if for any $\theta \in \mathbb{R}^n$, $\theta^\top A \theta \geq 0$.

Proof. Suppose that $\nabla^2 f(x) \succeq 0$. Fix any $x, y \in \mathbb{R}^n$. Let $g(t) = f(tx + (1-t)y)$. Then, we have

$$\begin{aligned} g(t) &= g(0) + \int_0^t g'(s) ds \\ &= g(0) + tg'(0) + \int_0^t \int_0^s g''(u) du ds \\ &= g(0) + tg'(0) + \int_0^t (t-s) g''(s) ds. \end{aligned}$$

Similarly, we have

$$g(1) = g(0) + g'(0) + \int_0^1 (1-s) g''(s) ds$$

Therefore, we have that

$$g(t) - (1-t)g(0) - tg(1) = \int_0^t (t-s) g''(s) ds - t \int_0^1 (1-s) g''(s) ds \leq 0$$

where we used that $g''(s) = (x-y)^\top \nabla^2 f(tx + (1-t)y)(x-y) \geq 0$.

Since we will not use the converse, we left it as an exercise. □

Here are some convex functions. For a longer list, please take a look on convex function in Wikipedia.

Example. Convex functions: x, x^+, e^x, x^a for $a \geq 1$, $-\log(x)$, $x\log x$, $\|x\|_p$ for $p \geq 1$, $(x,y) \to \frac{x^2}{y}$, $A \to \log\det A$, $(x,Y) \to x^\top Y^{-1}x$, $\log\sum_i e^{x_i}$, $(\prod_i x_i)^{\frac{1}{n}}$.

Here is an example on how to check a function is convex. Sometimes, checking if the Hessian is positive definite requires some inequalities.

Exercise 1.4.8. Prove that the function $f(x) = \log\sum_i e^{x_i}$ is convex.

Proof. Note that $\frac{\partial f}{\partial x_i} = \frac{e^{x_i}}{\sum_i e^{x_i}}$. Hence, we have

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{e^{x_i}}{\sum_i e^{x_i}} 1_{i=j} - \frac{e^{x_i+x_j}}{(\sum_i e^{x_i})^2}.$$

For any $\theta \in \mathbb{R}^n$, we have that

$$\theta^\top \nabla^2 f(x)\theta = \sum_i \frac{e^{x_i}\theta_i^2}{\sum_i e^{x_i}} - \sum_{i,j} \frac{e^{x_i+x_j}\theta_i\theta_j}{(\sum_i e^{x_i})^2}.$$

The right-hand side is non-negative because

$$\sum_{i,j} e^{x_i+x_j}\theta_i\theta_j = (\sum_i e^{x_i}\theta_i)^2 \leq (\sum_i e^{x_i}) \cdot (\sum_i e^{x_i}\theta_i^2).$$

□

1.5 Convex Optimization

In this course, we focus on problems of the form

$$\min_{x\in\mathbb{R}^n} f(x)$$

where f is convex. This includes $\min_{g(x)\leq c} f(x)$ because the function

$$f(x) + \ell_{g(x)\leq c}$$

is convex.

Example. Convex problems: linear regression $\min_x \|Ax-b\|_p$ for $p \geq 1$, linear program $\min_{Ax\leq b} c^\top x$, semi-definite programming $\min_{A_i\bullet X = b_i, X\succeq 0} C\bullet X$, logistic regression $\min_x \sum \log(1+e^{-y_i(a_i^T x + b_i)})$, shortest path problem, maximum flow problem, matroid intersection, $\cdots$.

For convex problems, it is easy to check if a given point is optimal or not. This is not just useful for computation purpose, but even for constructing objects and proves all kind of theorems. I will give an illustration of this in next section.

Theorem 1.5.1 (Optimality condition)**.** *Let f be a continuously differentiable convex function. Then, x is the minimizer of f if and only if $\nabla f(x) = 0$.*

Proof. If $\nabla f(x) \neq 0$, then $f(x - \varepsilon\nabla f(x)) = f(x) - (\varepsilon + o_\varepsilon(1))\|\nabla f(x)\|^2 < f(x)$ for small enough ε. Hence, such point cannot be the minimizer.

On the other hand, if $\nabla f(x) = 0$, Theorem 1.4.6 shows that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle = f(x) \text{ for all } y.$$

□

The study of algebraic properties of convex functions and convex sets is a huge subject that we can only touch its surface today. For example, many results here can be generalized to settings that is not differentiable. If you are interested, please refer to [3]. In a later lecture, we will cover some geometric properties of convex functions that is again another huge subject.

1.5.1 Example: Matrix Balancing

We are given a positive matrix A. The goal is to find a diagonal matrix D such that the sum of i^{th} row of DAD^{-1} equals to the sum of i^{th} column of DAD^{-1}. This is one of the basic operation in numerical linear algebra to normalize a matrix. It turns out the existence of such rescaling can be found by considering the convex function

$$f(x) = \sum_{i,j=1}^{n} A_{ij} e^{x_i - x_j}.$$

If we fix the first coordinate $x_1 = 0$, it is easy to see the function blows up when $\|x\|_\infty$ is large, in particular $f(x) \geq e^{\|x\|_\infty / n}(\min_{ij} A_{ij})$. Therefore, the minimizer must exists. By the optimality condition, we have that $\frac{\partial f}{\partial x_i}(x^*) = 0$ for all $i \neq 1$, which implies

$$\sum_{j=1}^{n} A_{ij} e^{x_i - x_j} - \sum_{j=1}^{n} A_{ji} e^{x_j - x_i} = 0.$$

(Although we fixed $x_1 = 0$, it is easy to recover the condition for $i = 1$ by using $i \neq 1$.) Therefore, each row sum of $A_{ij} e^{x_i - x_j}$ equals to each column sum $A_{ij} e^{x_i - x_j}$. The rescaling we want is simply $D_{ii} = e^{x_i}$. One extra benefit of this proof is that it also gives an algorithm for finding such rescaling, which eventually leads to an nearly linear time algorithm [1]. This is a hot topic in theory community recently because its relation to bipartite matching, polynomial identity testing, communication complexity, $\cdots$. It is even used recently to solve an decades-long open problem in operator theory [2].

References

[1] Michael B Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained newton's method and interior point methods. *arXiv preprint arXiv:1704.02310*, 2017.

[2] Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, and Akshay Ramachandran. The paulsen problem, continuous operator scaling, and smoothed analysis. *arXiv preprint arXiv:1710.02587*, 2017.

[3] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.