A GRADUATE COURSE

IN

OPTIMIZATION

A GRADUATE COURSE

IN

OPTIMIZATION

CIE6010 Notebook

Prof. Yin Zhang

The Chinese University of Hongkong, Shenzhen

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Contents

Acknowledgments

This book is from the CIE6010 in fall semester, 2018.

CUHK(SZ)

Notations and Conventions

X — Set

$\inf X \subseteq \mathbb{R}$ — Infimum over the set X

$\mathbb{R}^{m \times n}$ — set of all $m \times n$ real-valued matrices

$\mathbb{C}^{m \times n}$ — set of all $m \times n$ complex-valued matrices

x_i — ith entry of column vector $\boldsymbol{x}$

a_{ij} — (i,j)th entry of matrix $\boldsymbol{A}$

$\boldsymbol{a}_i$ — ith column of matrix $\boldsymbol{A}$

$\boldsymbol{a}_i^{\mathrm{T}}$ — ith row of matrix $\boldsymbol{A}$

$\mathbb{S}^n$ — set of all $n \times n$ real symmetric matrices, i.e., $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $a_{ij} = a_{ji}$ for all i,j

$\mathbb{H}^n$ — set of all $n \times n$ complex Hermitian matrices, i.e., $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ and $\bar{a}_{ij} = a_{ji}$ for all i,j

$\boldsymbol{A}^{\mathrm{T}}$ — transpose of $\boldsymbol{A}$, i.e, $\boldsymbol{B} = \boldsymbol{A}^{\mathrm{T}}$ means $b_{ji} = a_{ij}$ for all i,j

$\boldsymbol{A}^{\mathrm{H}}$ — Hermitian transpose of $\boldsymbol{A}$, i.e, $\boldsymbol{B} = \boldsymbol{A}^{\mathrm{H}}$ means $b_{ji} = \bar{a}_{ij}$ for all i,j

$\mathrm{trace}(\boldsymbol{A})$ — sum of diagonal entries of square matrix $\boldsymbol{A}$

$\mathbf{1}$ — A vector with all 1 entries

$\mathbf{0}$ — either a vector of all zeros, or a matrix of all zeros

$\boldsymbol{e}_i$ — a unit vector with the nonzero element at the ith entry

$\mathcal{C}(\boldsymbol{A})$ — the column space of $\boldsymbol{A}$

$\mathcal{R}(\boldsymbol{A})$ — the row space of $\boldsymbol{A}$

$\mathcal{N}(\boldsymbol{A})$ — the null space of $\boldsymbol{A}$

$\mathrm{Proj}_{\mathcal{M}}(\boldsymbol{A})$ — the projection of $\boldsymbol{A}$ onto the set $\mathcal{M}$

Chapter 1

Week1

1.1. Monday

1.1.1. Introduction to Optimizaiton

The usual optimization formulation is given by:

$$\min f(\boldsymbol{x}), \quad \text{where } f : \mathbb{R}^n \mapsto \mathbb{R}$$
$$\text{such that} \quad \boldsymbol{x} \in X \subseteq \mathbb{R}^n$$

One example of the set X is given by:

$$X = \left\{ \boldsymbol{x} \in \mathbb{R}^n \,\middle|\, \begin{array}{l} C_i(\boldsymbol{x}) = \boldsymbol{0}, i = 1,2,\ldots,m \leq n \\ h_i(\boldsymbol{x}) \geq \boldsymbol{0}, i = 1,2,\ldots,p \end{array} \right\}$$

Linear programming can be easily solved, but Integer linear programming is much harder. The equivalent LP formulation is given by:

$$\min \quad \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{Ax} = \boldsymbol{b}$$
$$\boldsymbol{c} \leq \boldsymbol{Bx} \leq \boldsymbol{c}'$$

1.2. Wednesday

1.2.1. Reviewing for Linear Algebra

Questions:

- What is the necessary and sufficient condition for the linear system $Ax = b$ to have a solution x?

  Answer: $b \in \mathcal{C}(A)$.

- For $A \in \mathbb{S}^n$, what is the necessary and sufficient condition for $A \succeq 0$?

  Answer: $x^{\mathrm{T}} A x \geq 0$ for $\forall x \in \mathbb{R}^n$; or $\lambda_i(A) \geq 0$ for all i.

1.2.2. Reviewing for Calculus

For function $f : \mathbb{R}^n \mapsto \mathbb{R}$:

- We use notation $f \in \mathcal{C}^n$ to denote f is **continuously differentiable to nth order**. This course will basically deal with such functions.

- We use notation $\nabla f(x)$ to denote the **Gradient** of f at x; and $\nabla^2 f(x)$ denotes the second order derivative of f at x. Note that $\nabla^2 f(x) \in \mathbb{S}^n$ for $f \in \mathcal{C}^1$.

- We use notation $\mathbb{S}^n$ to denote the set of all symmetric $n \times n$ matrices, i.e.,

$$\mathbb{S}^n = \{X \in \mathbb{R}^{n \times n} \mid X^{\mathrm{T}} = X\}$$

  Moreover, $\mathbb{S}^n_+$ denotes the set of all symmetric $n \times n$ matrices with all eigenvalues non-negative:

$$\mathbb{S}^n_+ = \{X \in \mathbb{R}^{n \times n} \mid X^{\mathrm{T}} = X \succeq 0\}$$

1.2.3. Introduction to Optimization

The usual optimization formulation is given by:

$$\min f(\boldsymbol{x}), \quad \text{where } f : \mathbb{R}^n \mapsto \mathbb{R}$$
$$\text{such that} \quad \boldsymbol{x} \in X \subseteq \mathbb{R}^n$$

- The simplest case for the constraint is $X = \mathbb{R}^n$, which leads to **unconstrainted** optimization problem.

- Or $X = P$ is a **polyhedron**, i.e., the boundaries for the region are all lines.

Definition 1.1 [Constraint Regions] In space $\mathbb{R}^n$,

- the hyper-plane is defined as:
$$\left\{ \boldsymbol{x} \middle| \boldsymbol{a}^{\mathrm{T}} \boldsymbol{x} = \boldsymbol{\beta} \right\}$$
with constants $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}$

- the half-space is defined as
$$\left\{ \boldsymbol{x} \middle| \boldsymbol{a}^{\mathrm{T}} \boldsymbol{x} \le \boldsymbol{\beta} \right\}$$

- the polyhedron is defined as the **intersection** of a **finite** number of hyperplanes or half-spaces

Next, we give the definition for the basic optimization problem:

Definition 1.2 [Linear Programming] The Linear Programming is given by:

$$\min \quad \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x},$$
$$\text{such that} \quad \boldsymbol{x} \in P(\text{polyhedron})$$

Or it can be reformulated as:

$$\min \quad c^{\mathrm{T}} x,$$
$$\text{such that} \quad A_I x \leq b_I$$
$$A_E x = b_E \in \mathbb{R}^m, \quad m < n.$$

■

Definition 1.3   [Optimality] x^* is said to be :

- the **local minimum** of $f(x)$ if there exists small ϵ such that

$$f(x^*) \leq f(x), \quad \forall x \in \mathcal{B}(x^*, \epsilon) \bigcap X := \{ x \mid \|x - x^*\| \leq \epsilon \} \bigcap X$$

- the **global minimum** if

$$f(x^*) \leq f(x), \quad \forall x \in X$$

■

Ⓡ   Unless specified, when we want to minimize a non-convex function, it usually means we only find its **local minimum**. This is because usually the local minimum is good enough.

The optimization task is essentially find x^* such that

$$x^* = \arg\min_{x \in X} f(x) \in \mathbb{R}^n.$$

philosophy (optimization sufficient and necessity). philosophy of relaxation (convex nulls)

The Optimality conditions are the **most important** theoretical tools for optimization.

> **Theorem 1.1** — **Optimality condition.** The optimality condition contains
>
> 1. Necessary Condition (exclude non-optimal points):
>
> $$n = 1 \text{ special case:} \begin{cases} \text{1st order: } f'(x) = 0 \\ \text{2nd order: } f''(x) \geq 0 \end{cases} \implies \begin{cases} \text{1st order: } \nabla f(x) = 0 \\ \text{2nd order: } \nabla f^2(x) \succeq 0 \end{cases}$$
>
> 2. Sufficient Condition (may identify optimal solutions)
>
> $$n = 1 \text{ special case:} \begin{cases} \text{1st order: } f'(x) = 0 \\ \text{2nd order: } f''(x) > 0 \end{cases} \implies \begin{cases} \text{1st order: } \nabla f(x) = 0 \\ \text{2nd order: } \nabla f^2(x) \succ 0 \end{cases}$$

Proof. The $n = 1$ special case can imply the general case for optimality condition. For multivariate f, we set $\boldsymbol{x} = \boldsymbol{x}^* + td$ with t to be the stepsize and d to be the direction. For fixed t and d, we define $h(t) = f(\boldsymbol{x}) = f(\boldsymbol{x}^* + td)$. It follows that

$$h'(t) = \nabla^{\mathrm{T}} f(\boldsymbol{x}^* + td)d$$

We find $h'(0) = \nabla^{\mathrm{T}} f(\boldsymbol{x}^*)d$ for $\forall$ d, which implies $\nabla f(\boldsymbol{x}^*) = 0$. ∎

Note that there is a gap between necessary and sufficient conditions, which puts us in an embarrassing position. However, the convex condition can save us:

> **Theorem 1.2** If f is convex in $\mathcal{C}^1$, then $\nabla f(\boldsymbol{x}) = 0$ is the **necessary** and **sufficient** condition.

Chapter 2

Week2

2.1. Tuesday

2.1.1. Reviewing and Announments

Tutorial: Thursday 7:00pm -9:00pm, ChengDao 208

Homework is due every Monday.

The first homework has been uploaded.

To proof the optimality condition in $\mathbb{R}^n$, we set $h(t) = f(x^* + td)$ for fixed x^* and d. It follows that

$$h'(t) = \nabla^{\mathrm{T}} f(x^* + td)d$$

and

$$h''(t) = d^{\mathrm{T}} \nabla^2 f(x^* + td)d$$

By the optimality condition for $\mathbb{R}$, we derive the necessary condition:

$$\begin{cases} h'(0) = \nabla^{\mathrm{T}} f(x^*)d = 0 \text{ for } \forall d \implies \nabla f(x^*) = 0; \\ h''(0) = d^{\mathrm{T}} \nabla^2 f(x^*)d = 0 \text{ for } \forall d \implies \nabla^2 f(x^*) \succeq 0 \end{cases}$$

together with the sufficient condition:

$$\begin{cases} \nabla f(x^*) = 0; \\ \nabla^2 f(x^*) \succ 0 \end{cases}$$

2.1.2. Quadratic Function Case Study

Given a quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$$

w.l.o.g., assume the matrix $\boldsymbol{Q}$ is symmetric (recall the quadratic section studied in linear algebra).

> **Definition 2.1** [Stationarity] A point $\boldsymbol{x}^*$ is said to be the stationary point of $f(\boldsymbol{x})$ if $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. ∎

To minimize such a function without constraint, we apply the optimality condition:

1. The first order optimality condition is given by:

$$\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0}$$

   The stationary point of the quadratic function $f(\boldsymbol{x})$ exists iff $\boldsymbol{b} \in \mathcal{C}(\boldsymbol{Q})$.

2. The second order necessary condition should be:

$$\nabla^2 f(\boldsymbol{x}) = \boldsymbol{Q} \succeq 0$$

For this special case, if $\boldsymbol{Q} \succeq 0$, then $f(\boldsymbol{x})$ is convex, the solutions to $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ are local minimum points. Furthermore, they are global minimum points (prove by Taylor Expansion). However, for general functions, we cannot obtain such good results.

Least Squares Problem. Such a problem has been well-studied in statistics given by:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$$

The first order derivative of the minimizer should satisfy:

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$$

Note that $A^{\mathsf{T}}b \in \mathcal{C}(A^{\mathsf{T}}A)$, thus the least squares problem always has a solution. However, such a solution is not unique unless A is full rank.

A Non-trival Quadratic Function. To minimize the function

$$f(x,y) = \frac{1}{2}(\alpha x^2 + \beta y^2) - x$$

We take the first order derivative to be zero:

$$\nabla f(x,y) = \begin{bmatrix} \alpha x - 1 \\ \beta y \end{bmatrix} = 0$$

The second order derivative is given by:

$$\nabla^2 f(x,y) = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$$

The optimal solutions depend on the value of α and β: (although we haven't introduce the definition for convex formally)

- If $\alpha, \beta > 0$, then this problem is **strongly convex**. By the necessary and sufficient optimality condition for convex problem, we find that $(\frac{1}{\alpha}, 0)$ is the unique local minimum (It is also the global minimum by plotting the figure).

- If $\alpha = 0$, this problem has no solution. The objective value $f(x,y) \to -\infty$ as $x \to \infty$.

- If $\beta = 0, \alpha > 0$, this problem is convex. By the necessary and sufficient optimality condition for convex problem, $\{(\frac{1}{\alpha}, \xi) \mid \xi \in \mathbb{R}\}$ is the set of local minimum. (By plotting the graph, we find that such set is the set of global minimum points)

- For $\alpha > 0, \beta < 0$ case, this problem is non-convex. Actually, $f(x,y) \to -\infty$ as $y \to \infty$. Hence, this problem has no global minimum point.

9

A Non-trival Function Study. To minimize the function

$$\min \quad f(\boldsymbol{y}) = e^{y_1} + \cdots + e^{y_n}$$
$$\text{such that} \quad y_1 + \cdots + y_n = S$$

We can transform such a constrainted optimization problem into unconstrainted. Let $y_n = S - y_1 - \cdots - y_{n-1}$ and substitute it into the objective function, it suffices to solve

$$\min e^{y_1} + \cdots + e^{y_{n-1}} + e^{S - y_1 - \cdots - y_{n-1}}$$

The stationary point should satisfy:

$$e^{y_i} = e^{S - y_1 - \cdots - y_{n-1}}, \qquad i = 1, 2, \ldots, n-1$$

Or equivalently, $y_1 = y_2 = \cdots = y_{n-1} = y_n$. Hence we derive the unique stationary point:

$$y_1^* = y_2^* = \cdots = y_n^* = \frac{S}{n}$$

The value on the stationary point is $f(y^*) = ne^{S/n}$. By checking the second order sufficient optimality condition,

$$\frac{f}{\partial y_i \partial y_j} = \begin{cases} e^{y_i} + e^{S - y_1 - \cdots - y_{n-1}} & i = j \\ e^{S - y_1 - \cdots - y_{n-1}} & i \neq j \end{cases} \implies \nabla^2 f = e^{S - y_1 - \cdots - y_{i-1} - y_i - y_{n-1}} \boldsymbol{E} + \text{diag}(e^{y_1}, \ldots, e^{y_{n-1}})$$

where $\boldsymbol{E}$ is a matrix with entries all ones. Thus $\nabla^2 f \succ 0$ for any stationary point. By the second order sufficient optimality condition, this stationary point is local minimum. Actually, for this special problem, this unique local minimum point is the global minimum.

(R) In this problem, we find that this stationary point is the unique local minimum point, but the unique local minimum point is not necessarily the global minimum point, unless the function is **coercive** or the feasible region is compact. Here is the counter-example: $f(x) = x^2 - x^4$. We will discuss the definition for coercive in the future.