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:

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

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:

$$\begin{aligned} \min \quad & \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{c} \leq \boldsymbol{B}\boldsymbol{x} \leq \boldsymbol{c}' \end{aligned}$$

1.2. Wednesday

1.2.1. Reviewing for Linear Algebra

Questions:

- What is the necessary and sufficient condition for the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ to have a solution $\boldsymbol{x}$?

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

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

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

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

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

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

1.2.3. Introduction to Optimization

The usual optimization formulation is given by:

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

- 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

$$\left\{ \boldsymbol{x} \middle| \boldsymbol{a}^{\mathrm{T}} \boldsymbol{x} \leq \boldsymbol{\beta} \right\}$$

- 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:

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

Or it can be reformulated as:

$$\begin{aligned} \min \quad & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}, \\ \text{such that} \quad & \boldsymbol{A}_I\boldsymbol{x} \leq \boldsymbol{b}_I \\ & \boldsymbol{A}_E\boldsymbol{x} = \boldsymbol{b}_E \in \mathbb{R}^m, \quad m < n. \end{aligned}$$

■

Definition 1.3 [Optimality] $\boldsymbol{x}^*$ is said to be :

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

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

- the **global minimum** if

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

■

R 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 $\boldsymbol{x}^*$ such that

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

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):

$$n=1 \text{ special case: } \begin{cases} \text{1st order: } f'(x)=0 \\ \text{2nd order: } f''(x)\geq 0 \end{cases} \Longrightarrow \begin{cases} \text{1st order: } \nabla f(x)=0 \\ \text{2nd order: } \nabla f^2(x)\succeq 0 \end{cases}$$

2. Sufficient Condition (may identify optimal solutions)

$$n=1 \text{ special case: } \begin{cases} \text{1st order: } f'(x)=0 \\ \text{2nd order: } f''(x)> 0 \end{cases} \Longrightarrow \begin{cases} \text{1st order: } \nabla f(x)=0 \\ \text{2nd order: } \nabla f^2(x)\succ 0 \end{cases}$$

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. Monday

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 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})$, thus the least squares problem always has a solution. However, such a solution is not unique unless $\boldsymbol{A}$ 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:

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

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.

A Non-trival Function Study. To minimize the function

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

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,

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

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.

2.2. Wednesday

2.2.1. Convex Analysis

This lectue will study the convex analysis.

Definition 2.2 [Convex] The subset $\mathcal{C} \subseteq \mathbb{R}^n$ is convex if

$$x, y \in \mathcal{C} \implies \{\lambda x + (1-\lambda)y \mid \alpha \in [0,1]\} \subset \mathcal{C},$$

i.e., the line segment between arbitrarily two elements lines in $\mathcal{C}$ ∎

Ⓡ Intersections of convex sets are convex. Empty set is assumed to be convex.

Definition 2.3 [Convex] The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if dom f is convex and

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

for $\forall x, y \in$ dom f and $\forall \lambda \in [0,1]$, i.e., the function evaluated in the line segment is lower than secant line between x and y (f lies below secant line). ∎

Ⓡ

- f is convex iff $-f$ is concave. (The concave definition simply changes the inequality direction in Def.(2.3))
- Affines are both convex and concave.
- The convexity depends on the domain of the function.

For a second order differentiable function, we have a much easier way to determine its convexity.

Theorem 2.1 If $f \in \mathcal{C}^1$, then the followings are equivalent:

1. f is convex

2. $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla^{\mathrm{T}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$ for $\forall x, y \in \text{dom } f$, i.e., f lines above the tangent line.

Proof. 1. From the definition for convexity,

$$f(y) - f(x) \geq \frac{f(\lambda x + (1-\lambda)y) - f(x)}{1-\lambda}$$

Letting $\lambda \to 1$, the RHS becomes a direction derivative:

$$f(y) - f(x) \geq \nabla^{\mathrm{T}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

2. To show the converse, we let $z = \lambda x + (1-\lambda)y$. By applying the inequality in (2.1) twice, we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \nabla^{\mathrm{T}} f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z}) \tag{2.1}$$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \nabla^{\mathrm{T}} f(\boldsymbol{z})(\boldsymbol{y} - \boldsymbol{z}) \tag{2.2}$$

Letting Eq.(2.1) times λ add Eq.(2.2) times $(1-\lambda)$, we derive that f is convex.

■

Theorem 2.2 If $f \in \mathcal{C}^2$, then the followings are equivalent:

1. f is convex
2. $\nabla^2 f(\boldsymbol{x}) \succeq 0$ for $\forall x \in \text{dom } f$.

Proof. We rewrite $f(\boldsymbol{y})$ by applying Taylor expansion:

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla^{\mathrm{T}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}), \tag{2.3}$$

for some $t \in [0,1]$.

1. If f is convex, from Theorem(2.1) and Eq.(2.3), we derive

$$(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}) \geq 0 \implies \frac{(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}}}{\|\boldsymbol{y} - \boldsymbol{x}\|} \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \frac{(\boldsymbol{y} - \boldsymbol{x})}{\|\boldsymbol{y} - \boldsymbol{x}\|} \geq 0$$

Set $d := \frac{(\boldsymbol{y}-\boldsymbol{x})}{\|\boldsymbol{y}-\boldsymbol{x}\|}$ and let $\boldsymbol{y} \to \boldsymbol{x}$, we derive

$$\boldsymbol{d}^{\mathrm{T}} \nabla^2 f(\boldsymbol{x}) \boldsymbol{d} \geq 0,$$

which implies $\nabla^2 f(\boldsymbol{x}) \succeq 0$ since $\boldsymbol{d}$ could have an arbitrary direction.

2. To show the converse, due to the semidefiniteness of $\nabla^2 f(\boldsymbol{x})$, we obtain a new inequality from Eq.(2.3):

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla^{\mathrm{T}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

From Theorem(2.1) we imply f is convex.

■

Definition 2.4 [Epigraph] The Epigraph of f is given by:

$$\mathrm{Epi}(f) := \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \mathrm{dom}\ f, t \geq f(x)\} \subseteq \mathbb{R}^{n+1}$$

■

Theorem 2.3 f is convex iff $\mathrm{Epi}(f)$ is convex.

Proof. 1. Suppose f is convex. For any $(x,t),(y,s) \in \mathrm{Epi}(f)$, it suffices to show

$$(\lambda x + (1-\lambda)y, \lambda t + (1-\lambda)s) \in \mathrm{Epi}(f) \Longleftrightarrow \lambda t + (1-\lambda)s \geq f(\lambda x + (1-\lambda)y).$$

The convexity of f implies that

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &\leq \lambda f(x) + (1-\lambda) f(y) \\ &\leq \lambda t + (1-\lambda)s. \end{aligned}$$

2. The reverse direction is obvious by applying definitions.

■

Definition 2.5 [Strict Convex] The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is strict convex if dom f is convex and

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

for $\forall x \neq y, x, y \in$ dom f and $\forall \lambda \in (0,1)$ ■

R Strict convex implies the uniquesness of minimum

However, for function $f(x) = \frac{1}{x}$, the curvature becomes more and more flat. We want to exclude such kind of functions.

Definition 2.6 The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be strongly convex if dom f is convex and $\exists \alpha > 0$ such that $f(\boldsymbol{x}) - \alpha \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}$ is convex; or equivalently,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla^{\mathrm{T}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$$

■

R The strong convexity places a quadratic lower bound in the curvature of the function, i.e., the function must rise up at least as fast as a quadratic function. How fast it rises depends on the parameter α.

The convexity properties are extremely useful in forcing optimization algorithms to rapidly converge to optima. However, most functions are not convex. The most important result that requires convexity is given below:

Theorem 2.4 If f is convex in $\mathcal{C}^1$, then $\nabla f(\boldsymbol{x}) = 0$ is the **necessary** and **sufficient** condition for **global** minimum.

Note that convex function does not have a local minimum that is not global minimum.

Proof. If $f \in \mathcal{C}^1$ is convex, recall the Theorem(2.1) that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla^{\mathrm{T}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{2.4}$$

1. If $\nabla f(\boldsymbol{x}) = 0$, then Eq.(2.3) implies $f(\boldsymbol{y}) \geq f(\boldsymbol{x})$ for $\forall y$.
2. If $\boldsymbol{x}$ is the global minimum, recall the optimality conditon, $\nabla f(\boldsymbol{x}) = \mathbf{0}$.

■

In practice, we cannot solve all convex optimization problems. So we need to carefully study the structure of every problem we have faced.

Chapter 3

Week3

3.1. Wednesday

Assignment 2 posted.

CIE6010: Exercise 1.2.9 and 1.3.9; together with MATLAB project.

3.1.1. Convex Analysis

For the **constrained** problem

$$\begin{aligned} \min \quad & f(x) \\ & \boldsymbol{x} \in X \subseteq \mathbb{R} \\ & f \text{ is convex in } \mathcal{C}^1 \end{aligned}$$

$\boldsymbol{x}$ is a global minimum iff $\nabla^{\mathrm{T}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \geq 0$ for $\forall y \in X$

Such a condition is not useful unless $\boldsymbol{y}$ lies in the whole space, at that time we have no choice but $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$.

An equivalent version of the condition is that every **feasible** direction is **ascending**.

Definition 3.1 [Descending Direction] $\boldsymbol{d} \in \mathbb{R}^n$ is a **descending direction** of f at $\boldsymbol{x}$ if

$$\nabla^{\mathrm{T}} f(\boldsymbol{x}) \boldsymbol{d} < 0.$$

This definition is the motivation of descent method.

3.1.2. Iterative Method

Definition 3.2 [Descent Method] At ant non-stationary $\boldsymbol{x}$, i.e., $\nabla f(x) \neq \mathbf{0}$, we find descent d, i.e., $\nabla^{\mathrm{T}} f(\boldsymbol{x})\boldsymbol{d} < 0$. We update our old $\boldsymbol{x}$ as:

$$\boldsymbol{x}^{r+1} \leftarrow \boldsymbol{x}^r + \alpha^r \boldsymbol{d}^r, \quad \alpha > 0.$$

The key is how to chooose $\boldsymbol{d}$ and α. We have a general formula for $\boldsymbol{d}$, which is the **descent direction**:

$$\boldsymbol{d} = -\boldsymbol{D} \cdot \nabla f(\boldsymbol{x}),$$

where $\boldsymbol{D} \in \mathbb{S}^n$ and $\boldsymbol{D} \succ 0$.

e.g., $\boldsymbol{D} = \boldsymbol{I}$ implies gradient method (Steepest Descent).

$\boldsymbol{D} = \left(\nabla^2 f(\boldsymbol{x})\right)^{-1}$ implies the Newton's method.

Nonlinear LS. The optimization problem is

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) = \tfrac{1}{2}\textstyle\sum_{i=1}^m g_i^2(\boldsymbol{x}) := \tfrac{1}{2}\|g(\boldsymbol{x})\|_2^2 \\ & g(\boldsymbol{x}) = \begin{pmatrix} g_1(\boldsymbol{x}) & g_2(\boldsymbol{x}) & \cdots & g_m(\boldsymbol{x}) \end{pmatrix}^{\mathrm{T}} \end{aligned}$$

The gredient function is

$$\nabla f(\boldsymbol{x}) = \sum_{i=1}^m g_i(\boldsymbol{x}) \nabla g_i(\boldsymbol{x}) = \underbrace{\begin{bmatrix} \nabla g_1(\boldsymbol{x}) & \cdots & \nabla g_m(\boldsymbol{x}) \end{bmatrix}}_{\nabla g(x) \in \mathbb{R}^{n \times m}} \begin{bmatrix} g_1(\boldsymbol{x}) \\ \vdots \\ g_m(\boldsymbol{x}) \end{bmatrix} = \nabla g(\boldsymbol{x}) \cdot g(\boldsymbol{x}) = \boldsymbol{J}^{\mathrm{T}}(\boldsymbol{x}) g(\boldsymbol{x}),$$

where $J(\boldsymbol{x}) \in \mathbb{R}^{m \times n}$ is the Jacobian of g.

The second order derivative is given as:

$$\nabla^2 f(\boldsymbol{x}) = \boldsymbol{J}^{\mathrm{T}}(\boldsymbol{x}) \boldsymbol{J}(\boldsymbol{x}) + \sum_{i=1}^m g_i() \nabla^2 g_i(\boldsymbol{x})$$

the second term in RHS is complicated and hard to compute. The Gauss-Newton

method directly ignore it.

Choice of Step Length α. Exact line search

$$\min_{\alpha\in(0,M]} f(\boldsymbol{x}^r + \alpha^r \boldsymbol{d}^r)$$

usually it is too expensive. Sometimes we choose α as a constant.

Definition 3.3 [Lipschitz Continuous] ∇f is **Lipschitz continuous** with Lipschitz constant L if

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

for all $\boldsymbol{x}, \boldsymbol{y}$. ■

R This condition guarantees that

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^{r+1}) \geq \frac{L}{2}\|\nabla f(\boldsymbol{x}^r)\|^2$$

From this inequality, we imply that the result of convergence is $\nabla f(\boldsymbol{x}^r) \to 0$, but the minimum point is still un-guaranteed. In Deep Learing people often train the data using this way, which is not so rigorous.

We set $h(t) = f(\boldsymbol{x} + t\alpha\boldsymbol{d})$, then we can use Lipschitz continuous to analysis the rate of convergence and the choice of step-length:

$$\begin{aligned}
f(\boldsymbol{x} + \alpha\boldsymbol{d}) - f(\boldsymbol{x}) &= \int_0^1 \nabla^{\mathrm{T}} f(\boldsymbol{x} + t\cdot\alpha\boldsymbol{d})\alpha d\,\mathrm{d}t \\
&= \int_0^1 \left[\nabla^{\mathrm{T}} f(\boldsymbol{x} + t\cdot\alpha\boldsymbol{d})\alpha\boldsymbol{d} - \nabla^{\mathrm{T}} f(\boldsymbol{x})\alpha\boldsymbol{d} + \nabla^{\mathrm{T}} f(\boldsymbol{x})\alpha\boldsymbol{d}\right] \\
&\leq \nabla^{\mathrm{T}} f(\boldsymbol{x})\alpha\boldsymbol{d} + \int_0^1 \|\nabla f(\boldsymbol{x} + \alpha\boldsymbol{d}) - \nabla f(\boldsymbol{x})\| \cdot \|\alpha\boldsymbol{d}\|\,\mathrm{d}t \\
&\leq \nabla^{\mathrm{T}} f(\boldsymbol{x})\alpha\boldsymbol{d} + L\int_0^1 t\alpha^2\|d\|^2\,\mathrm{d}t \\
&= \underbrace{\nabla^{\mathrm{T}} f(\boldsymbol{x})\alpha\boldsymbol{d}}_{\text{negative}} + \frac{L\alpha^2\|d\|^2}{2}
\end{aligned}$$

Differentiate the RHS w.r.t. α leads to

$$\nabla^{\mathrm{T}} f(\boldsymbol{x})\boldsymbol{d} + L\alpha\|\boldsymbol{d}\|^2 = 0 \implies \alpha = -\frac{\nabla^{\mathrm{T}} f(\boldsymbol{x})\boldsymbol{d}}{L\|\boldsymbol{d}\|^2} > 0,$$

which seems a reasonable choice. If $\boldsymbol{d}$ is the steepest descent direction, the step-length becomes:

$$\alpha = \frac{1}{L} \text{ (constant)}$$