

A GRADUATE COURSE
IN
OPTIMIZATION

A GRADUATE COURSE
IN
OPTIMIZATION
CIE6010 Notebook

Prof. Yin Zhang

The Chinese University of Hong Kong, Shenzhen



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

Contents

Acknowledgments	vii
Notations	ix
1 Week1	1
1.1 Monday	1
1.1.1 Introduction to Optimizaiton	1
1.2 Wednesday	2
1.2.1 Reviewing for Linear Algebra	2
1.2.2 Reviewing for Calculus	2
1.2.3 Introduction to Optimization	3
2 Week2	7
2.1 Monday	7
2.1.1 Reviewing and Announments	7
2.1.2 Quadratic Function Case Study	8
2.2 Wednesday	11
2.2.1 Convex Analysis	11
3 Week3	17
3.1 Wednesday	17
3.1.1 Convex Analysis	17
3.1.2 Iterative Method	18
3.2 Thursday	22
3.2.1 Announcement	22
3.2.2 Sparse Large Scale Optimization	22

4	Week4	27
4.1	Wednesday	27
4.1.1	Comments for MATLAB Project	27
4.1.2	Local Convergence Rate	28
4.1.3	Newton's Method	29
4.1.4	Tutorial: Introduction to Convexity	30
5	Week5	33
5.1	Monday	33
5.1.1	Review	33
5.1.2	Existence of solution to Quadratic Programming	36
5.2	Wednesday	39
5.2.1	Comments about Newton's Method	39
5.2.2	Constant Step-Size Analysis	40

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	i th entry of column vector \mathbf{x}
a_{ij}	(i, j) th entry of matrix \mathbf{A}
\mathbf{a}_i	i th column of matrix \mathbf{A}
\mathbf{a}_i^T	i th row of matrix \mathbf{A}
\mathbb{S}^n	set of all $n \times n$ real symmetric matrices, i.e., $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $a_{ij} = a_{ji}$ for all i, j
\mathbb{H}^n	set of all $n \times n$ complex Hermitian matrices, i.e., $\mathbf{A} \in \mathbb{C}^{n \times n}$ and $\bar{a}_{ij} = a_{ji}$ for all i, j
\mathbf{A}^T	transpose of \mathbf{A} , i.e, $\mathbf{B} = \mathbf{A}^T$ means $b_{ji} = a_{ij}$ for all i, j
\mathbf{A}^H	Hermitian transpose of \mathbf{A} , i.e, $\mathbf{B} = \mathbf{A}^H$ means $b_{ji} = \bar{a}_{ij}$ for all i, j
$\text{trace}(\mathbf{A})$	sum of diagonal entries of square matrix \mathbf{A}
$\mathbf{1}$	A vector with all 1 entries
$\mathbf{0}$	either a vector of all zeros, or a matrix of all zeros
\mathbf{e}_i	a unit vector with the nonzero element at the i th entry
$\mathcal{C}(\mathbf{A})$	the column space of \mathbf{A}
$\mathcal{R}(\mathbf{A})$	the row space of \mathbf{A}
$\mathcal{N}(\mathbf{A})$	the null space of \mathbf{A}
$\text{Proj}_{\mathcal{M}}(\mathbf{A})$	the projection of \mathbf{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:

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

One example of the set X is given by:

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

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

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{Ax} = \mathbf{b} \\ & \mathbf{c} \leq \mathbf{Bx} \leq \mathbf{c}' \end{aligned}$$

1.2. Wednesday

1.2.1. Reviewing for Linear Algebra

Questions:

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

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

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

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

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

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

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

1.2.3. Introduction to Optimization

The usual optimization formulation is given by:

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

- The simplest case for the constraint is $X = \mathbb{R}^n$, which leads to **unconstrained** 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:

$$\{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} = \beta\}$$

with constants $\mathbf{a} \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$

- the half-space is defined as

$$\{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} \leq \beta\}$$

- 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:

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

Or it can be reformulated as:

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

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

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

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

- the **global minimum** if

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

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 \mathbf{x}^* such that

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

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):

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

2. Sufficient Condition (may identify optimal solutions)

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

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

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

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

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 $\nabla f(\mathbf{x}) = 0$ 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

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

and

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

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

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

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

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

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

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

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

1. The first order optimality condition is given by:

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

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

2. The second order necessary condition should be:

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

For this special case, if $\mathbf{Q} \succeq 0$, then $f(\mathbf{x})$ is convex, the solutions to $\nabla f(\mathbf{x}) = \mathbf{0}$ 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:

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

The first order derivative of the minimizer should satisfy:

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

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

A Non-trivial 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 $f(x, y) \rightarrow -\infty$ as $x \rightarrow \infty$.
- 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, $f(x, y) \rightarrow -\infty$ as $y \rightarrow \infty$. Hence, this problem has no global minimum point.

A Non-trivial Function Study. To minimize the function

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

We can transform such a constrained optimization problem into unconstrained. 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:

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


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 $f(\mathbf{y}^*) = ne^{S/n}$. By checking the second order sufficient optimality condition,

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

where \mathbf{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.

-  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 lecture will study the convex analysis.

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

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

i.e., the line segment between arbitrarily two elements lies in \mathcal{C} ■

R 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 $\text{dom } f$ is convex and

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

for $\forall x, y \in \text{dom } f$ 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). ■

R

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

Proof. 1. From the definition for convexity,

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

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

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

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

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla^T f(\mathbf{z})(\mathbf{x} - \mathbf{z}) \quad (2.1)$$

$$f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla^T f(\mathbf{z})(\mathbf{y} - \mathbf{z}) \quad (2.2)$$

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. $\nabla^2 f(\mathbf{x}) \succeq 0$ for $\forall \mathbf{x} \in \text{dom } f$.

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

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

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

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

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

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

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

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

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

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

From Theorem(2.1) we imply f is convex. ■

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

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

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

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

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

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 $\text{dom } f$ is convex and

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

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

R Strict convex implies the uniqueness 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 $\text{dom } f$ is convex and $\exists \alpha > 0$ such that $f(\mathbf{x}) - \alpha \mathbf{x}^T \mathbf{x}$ is convex; or equivalently,

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

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 $\nabla f(\mathbf{x}) = 0$ 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

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

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

■

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

Last time we have shown that for a unconstrained problem, $\nabla f(\mathbf{x}) = 0$ is the necessary and sufficient condition for global minimum ensurance. However, the case for constrained problem will be different.

Proposition 3.1 For the **constrained** problem

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

\mathbf{x} is a global minimum iff

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0$$

for $\forall \mathbf{y} \in X$.

Proof. Since f is convex, the inequality below holds:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{x}, \mathbf{y} \in X$$

Note that \mathbf{x} is a global minimum iff $f(\mathbf{y}) \geq f(\mathbf{x})$, $\forall \mathbf{y} \in X$. Combining the inequality above, the proof is complete. ■

R

- Such a condition is not so useful unless \mathbf{y} lies in the whole space, at that time we have no choice but $\nabla f(\mathbf{x}) = \mathbf{0}$. (otherwise we can construct a \mathbf{y} to let the inner product negative.)
- An equivalent version of the condition is that every **feasible** direction is **ascending**.

Definition 3.1 [Descending Direction] The vector $\mathbf{d} \in \mathbb{R}^n$ is said to be a **descending direction** of f at \mathbf{x} if

$$\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle < 0.$$

This definition is the motivation of descent method.

3.1.2. Iterative Method

Definition 3.2 [Descent Method] At any non-stationary \mathbf{x} , i.e., $\nabla f(\mathbf{x}) \neq \mathbf{0}$, we find the descending direction \mathbf{d} , i.e., $\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle < 0$. We update our old \mathbf{x} as:

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

The key is how to choose \mathbf{d} and α . We have a general formula for \mathbf{d} :

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

where $\mathbf{D} \in \mathbb{S}^n$ and $\mathbf{D} \succ 0$. (Verify by yourself that \mathbf{d} satisfies the descending direction definition)

1. $D = I$ implies gradient method (Steepest Descent).
2. $D = (\nabla^2 f(\mathbf{x}))^{-1}$ implies the Newton's method.

Nonlinear LS. The optimization problem is

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

The gradient function is

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

where $J(\mathbf{x}) \in \mathbb{R}^{m \times n}$ is said to be the Jacobian matrix of $g(\mathbf{x})$.

The second order derivative function is given as: (complete the calculation process by yourself)

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

the second term in RHS is complicated and hard to compute. To solve this LS problem, the Gauss-Newton method directly ignore it, which leads to the descent direction

$$\mathbf{d} = -(J^T J)^{-1} J^T g(\mathbf{x})$$

Choice of Step Length α . We apply the Limited Minimization Rule to find α , i.e., for fixed $s > 0$, choose α^r such that

$$\min_{\alpha^r \in (0, s]} f(\mathbf{x}^r + \alpha^r \mathbf{d}^r).$$

Usually this rule is too computationally expensive. The alternative ways are:

- Choose α just as a constant
- Choose $\alpha^r \rightarrow 0$ as $r \rightarrow \infty$ but also satisfies the infinite travel condition

$$\sum_{r=0}^{\infty} \alpha^r = \infty$$

Adding Lipschitz condition will make the choice of step-length easier:

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

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

for all \mathbf{x}, \mathbf{y} . ■



- It is useful to note that convexity places a lower bound on the growth of the function at every point; whereas Lipschitzness places an upper bound on the growth of the function that is linear in the perturbation i.e., $\|\mathbf{x} - \mathbf{y}\|_2$. Also note that Lipschitz functions need not be differentiable. However, differentiable functions with bounded gradients are always Lipschitz.
- The Lipschitz condition induces that for iterative method we have

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

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

Convergence Rate Analysis. We apply the Lipschitzness to analysis the rate of convergence first. Setting $h(t) = f(\mathbf{x} + t\alpha\mathbf{d})$, we find that

$$\begin{aligned}
f(\mathbf{x} + \alpha\mathbf{d}) - f(\mathbf{x}) &= h(1) - h(0) = \int_0^1 h'(t) dt \\
&= \int_0^1 \langle \nabla f(\mathbf{x} + t \cdot \alpha\mathbf{d}), \alpha\mathbf{d} \rangle dt \\
&= \int_0^1 [\langle \nabla f(\mathbf{x} + t \cdot \alpha\mathbf{d}), \alpha\mathbf{d} \rangle - \langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle + \langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle] dt \\
&= \langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t \cdot \alpha\mathbf{d}) - \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle dt \\
&\leq \langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle + \int_0^1 \|\nabla f(\mathbf{x} + t \cdot \alpha\mathbf{d}) - \nabla f(\mathbf{x})\| \cdot \|\alpha\mathbf{d}\| dt \\
&\leq \langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle + L \int_0^1 t\alpha^2 \|\mathbf{d}\|^2 dt \\
&= \underbrace{\langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle}_{\text{negative}} + \frac{L\alpha^2 \|\mathbf{d}\|^2}{2}
\end{aligned}$$

Choice of Step Length. To get the optimal step length α , differentiating the RHS w.r.t. α leads to

$$\langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle + L\alpha \|\mathbf{d}\|^2 = 0 \implies \alpha = -\frac{\langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle}{L\|\mathbf{d}\|^2} > 0,$$

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

$$\alpha = \frac{1}{L}.$$

3.2. Thursday

3.2.1. Announcement

The assignment 2 requires to do a MATLAB project. The grade usually depends on your understanding of the reading materials and the time spent on experimentation.

3.2.2. Sparse Large Scale Optimization

Given an underlying signal $\mathbf{x} \in \mathbb{R}^n$ satisfying the undermined system $\mathbf{Ax} = \mathbf{b}$, we aim to recover the desired $\hat{\mathbf{x}}$. It suffices to solve the optimization problem

$$\begin{aligned} \min \quad & \|\mathbf{D}\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{Ax} = \mathbf{b} \\ & \mathbf{A} \in \mathbb{R}^{m \times n}, m < n \end{aligned}$$

with \mathbf{D} to be the difference matrix and $\min \|\mathbf{D}\mathbf{x}\|_1$ is sparsity promoting. Here we list two basic but effective ways to solve such a problem.

Linear Programming Approach. One way is to reformulate the problem into LP.

1. Define new variables $t_i = |(\mathbf{D}\mathbf{x})_i|$, we can reformulate the origin problem as:

$$\begin{aligned} \min \quad & \sum_{i=1}^n t_i \\ \text{s.t.} \quad & \mathbf{Ax} = \mathbf{b} \\ & -t_i \leq \sum_{k=1}^n d_{ik}x_k \leq t_i \\ & \mathbf{A} \in \mathbb{R}^{m \times n}, m < n \end{aligned}$$

2. Alternatively, recall what we have learnt in MAT3007. Define slack variables

$(\mathbf{D}\mathbf{x})_i = u_i - v_i$, where $u_i := (\mathbf{D}\mathbf{x})_i^+$, $v_i = (\mathbf{D}\mathbf{x})_i^-$. It suffices to solve

$$\begin{aligned} \min \quad & \sum_{i=1}^n (u_i + v_i) \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & -t_i \leq \sum_{k=1}^n d_{ik}x_k \leq t_i \\ & \mathbf{A} \in \mathbb{R}^{m \times n}, m < n \\ & u_i, v_i \geq 0 \end{aligned}$$

However, linear programming is not the optimal way to solve large-scale problem.

Gredient-Based Approach. We can also transform it into the unconstraint minimization problem, i.e., we add the penalty for the constraint $\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}$:

$$\min \|\mathbf{D}\mathbf{x}\|_1 + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$$

You may see that this reformulation is not exactly equivalent to the origin problem. However, it is not meaningful to stress $\mathbf{A}\mathbf{x}$ should exactly equal to \mathbf{b} , as there exists some noise perturbing the equality in nature.

Another problem is that this objective function is not differentiable once there is at least zero entry from $\mathbf{D}\mathbf{x}$. Thus we do the approximation

$$|t| \approx \sqrt{t^2 + \sigma}, \text{ for small } \sigma > 0.$$

Hence, it suffices to solve

$$\min f(x) := \Theta_\sigma(\mathbf{D}\mathbf{x}) + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \quad (3.1)$$

where

$$\Theta_\sigma(\mathbf{y}) = \sum_{i=1}^n \sqrt{y_i^2 + \sigma}$$

Descent Direction. Since problem(3.1) is convex, taking the derivative leads to minimum point. Hence we use the gredient descent method, i.e., $\mathbf{d} = -\nabla f(\mathbf{x})$.

Although this direction is not optimal (trying another direction may be faster after

several iterations), let's assume we are short-sighted such that we just want to take the steepest direction.

Hence the iterative algorithm to solve this problem can be summarized into one formula: Take a initial guess \mathbf{x}^0 , then for $r = 0, 1, 2 \dots$

$$\mathbf{x}^{r+1} = \mathbf{x}^r - \alpha^r \nabla f(\mathbf{x}^r)$$

Stopping Criteria. The stopping criteria has two conditions, either one is satisfied is ok. Always keep mind of scaling for stopping criteria, i.e., how large of an objective should depend on the scale of the problem.

- First is $\|\nabla f(\mathbf{x}^k)\| \leq 10^{-2} \|\nabla f(\mathbf{x}^0)\|$, i.e., the iterative method converge to the near stationary point
- Another is $|f(\mathbf{x}^k) - f(\mathbf{x}^{k+1})| \leq 10^{-8} |f(\mathbf{x}^k)|$, i.e., the function does not change too much.

The next questions turn out that how to choose initial guess? How to choose step-length? Is steepest descent usually effective?

1. For large-scale optimization, the steepest descent is usually one of the **best** way among iterative methods.
2. To choose the initial guess, sometimes we choose the LS solution, i.e., enter the matlab command $A' A / A' b$.
3. Last lecture we tell that we can choose step-length to be the Lipschitz constant, but the disadvantage is that the constant L for large scale optimization is too small. We have a better alternative.

Armijo Condition and BB Step. The motivation is that we aim to let

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) \leq f(\mathbf{x}^k) + C_1 \alpha \langle \nabla f(\mathbf{x}), \mathbf{d}^k \rangle, (0 < C_1 < 1) \quad (3.2)$$

i.e., our updated function value should be at least less than the old function minus the descent decrease gain, i.e., it should sufficiently decrease faster than a constant times

the steepest gradient descent.

This method has a geometrically meaning:

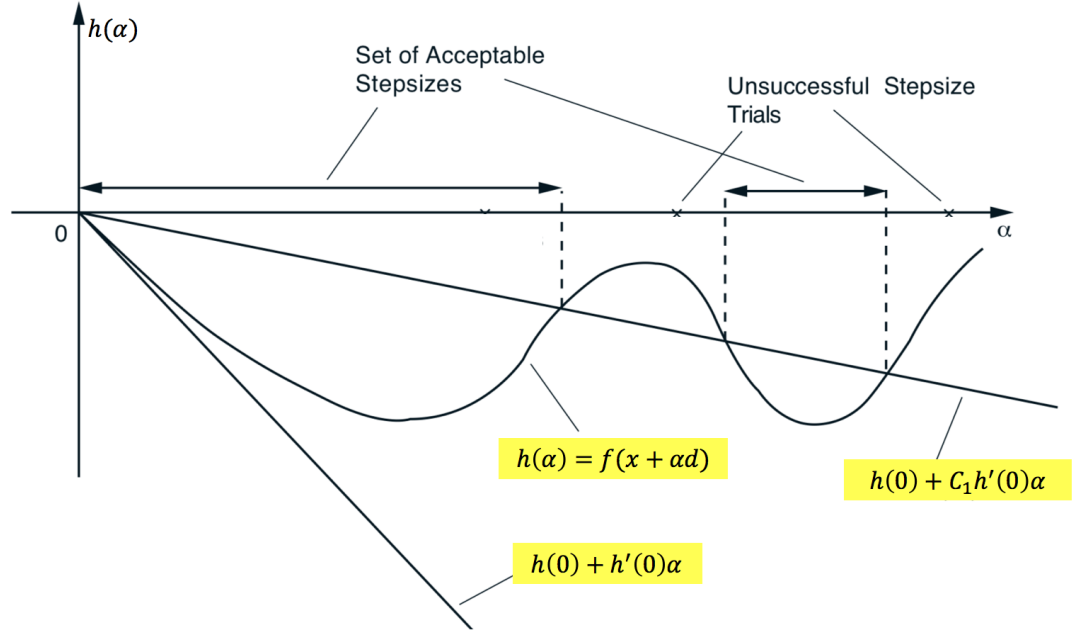


Figure 3.1: Geometric Interpretation of Armijo Condition

We set $h(\alpha) := f(\mathbf{x} + \alpha \mathbf{d})$, then $h'(0) = \langle \nabla f(\mathbf{x} + \alpha \mathbf{d}), \mathbf{d} \rangle$, thus the tangent line at $h(0)$ is given by:

$$h(0) + \alpha h'(0) := f(\mathbf{x}) + \alpha \langle \nabla f(\mathbf{x} + \alpha \mathbf{d}), \mathbf{d} \rangle$$

Geometrically we can see that no α can be chosen such that the updated function value $h(\alpha)$ is less than this tangent line. Hence we make the tangent line more flat, i.e., we want to find α such that the updated function value $h(\alpha)$ is below the line $h(0) + C_1 h'(0) \alpha$, $0 < C_1 < 1$:

$$h(\alpha) \leq h(0) + C_1 h'(0) \alpha$$

How to choose such α ? Take a initial long step-length $\bar{\alpha}$ first, if condition(3.2) is not satisfied, try step length $\beta \bar{\alpha}, \beta^2 \bar{\alpha}, \dots$ respectively. (Take a big step, if not satisfied, shorten the step.)

R However, it is not suggested to do that. Although it is mathematically true,

during the computer run, the step-length will decrease exponentially.

How to choose C_1 ? Empirically, $C_1 = 10^{-3}$ or 10^{-4} , i.e., it is very flat.

How to choose initial $\bar{\alpha}$? It depends on the scale of functionm which requires for your reading of materials.

How to choose the value of β ? Do the experiment. (0.5, 0.8 for example).

Chapter 4

Week4

4.1. Wednesday

4.1.1. Comments for MATLAB Project

You need to take care of several things during your assignment:

Do Not Repeat Computation. For example, enter

$$|f(x^{k+1}) - f(x^k)| \leq 10^{-\varepsilon} |f(x^k)|,$$

is very bad, since you evaluated this function three times.

Arrange Computation Properly. For example, compute

$$\mathbf{x}\mathbf{x}^T\mathbf{A}$$

is very expensive, but $\mathbf{x}(\mathbf{x}^T\mathbf{A})$ is not. Be aware of the size of matrices or vectors.

Appreciate sparsity. For example, when faced with high-dimensional matrices, compute $\mathbf{A}\mathbf{D}\mathbf{A}^T$ is bad for using $\mathbf{D} = \text{diag}(\mathbf{d})$, while using $\mathbf{D} = \text{sparse}(\mathbf{1} : \mathbf{n}, \mathbf{1} : \mathbf{n}, \mathbf{d})$ is better.

Grading Criteria. Your code should be at least faster than the testing script, while the error should be smaller than the testing script.

4.1.2. Local Convergence Rate

The study of the rate of convergence is often the dominant criteria for selecting appropriate algorithms. In this lecture we only focus on the local behaviour of the method in a neighborhood of an optimal solution.

Definition 4.1 [Q_1 Factor] Restrict the attention to a convergent sequence $\{\mathbf{x}^k\}$ with limit \mathbf{x}^* . Define an error function $e_k = \|\mathbf{x}^k - \mathbf{x}^*\| \rightarrow 0$. The Q_1 factor of $\{\mathbf{x}^k\}$ is given as:

$$Q_1 = \limsup_{k \rightarrow \infty} \frac{e_{k+1}}{e_k}$$

Here we want to study the performance of e_k . In our case, we compare $\{e_k\}$ with the geometric progression

$$\beta^k, \quad k = 0, 1, \dots$$

- If there exists $\beta \in (0, 1)$ such that

$$Q_1 \leq \beta,$$

then we can show $e_k \leq q\beta^k$ for some $q > 0$. In this case $\{e_k\}$ is said to be **Q-linear convergent**.

- If $Q_1 = 0$, then we say $\{e_k\}$ is **Q-super-linear convergent**
- If $Q_1 = 1$, then we say $\{e_k\}$ is **Q-sub-linear convergent**



1. Q-linear convergence is not always so good, e.g., $\beta = .999$ may require 20000 iterations to meet satisfaction, while $\beta = .1$ may only require 20.
2. Linear is always better than sublinear; super linear is always better than linear.

We have a faster type of convergence:

Definition 4.2 [Q_2 Factor]

$$Q_2 = \limsup_{k \rightarrow \infty} \frac{e^{k+1}}{(e^k)^2}$$

If $Q_2 = M < +\infty$, i.e., $e^{k+1} = O((e^k)^2)$, then $\{x^k\} \rightarrow x^*$ Q -quadratically. ■

Newton's method generally gives us the quadratic convergence.

4.1.3. Newton's Method

The newton's method requires to solve a non-linear system of equations $\nabla f(x) = 0$.

We don't know how to solve non-linear system in general. Fortunately, Newton gives us the remediation: in order to search for \mathbf{d} to make $\nabla f(\mathbf{x} + \mathbf{d}) = 0$, do the linearization first.

$$\nabla f(\mathbf{x} + \mathbf{d}) \approx \nabla f(\mathbf{x}) + \langle \nabla^2 f(\mathbf{x}), \mathbf{d} \rangle = 0 \quad (4.1)$$

To get the optimal solution, it suffices to solve (4.1) for \mathbf{d} :

$$\mathbf{d} = -(\nabla^2 f(\mathbf{x}))^{-1} \nabla f(\mathbf{x}),$$

and hence update the solution to be $\mathbf{x} \leftarrow \mathbf{x} + \alpha \mathbf{d}$.

Interpretation. In order to minimize a strictly convex function $f(x)$, we find

$$f(\mathbf{x} + \mathbf{d}) \approx f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle + \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{x}) \mathbf{d} := q(\mathbf{d})$$

It suffices to minimize $q(\mathbf{d})$:

$$\nabla q(\mathbf{d}) = 0 \iff \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x}) \mathbf{d} = 0$$

How to guarantee \mathbf{d} is the descent direction? Not necessarily we are able to do that. It becomes an art.

Convergence of Rate.

Proposition 4.1 Newton's method guarantees the Q-quadratic convergence.

Proof. Given a nonlinear system $F(x) = 0$, suppose the sequence $\{\mathbf{x}^k\}$ is generated by Newton with limit \mathbf{x}^* and $F(\mathbf{x}^*) = 0$. By Newton's iteration,

$$\mathbf{x}^{k+1} = \mathbf{x}^k - [F'(\mathbf{x}^k)]^{-1}F(\mathbf{x}^k), \quad (4.2)$$

which follows that

$$\begin{aligned} \mathbf{x}^{k+1} - \mathbf{x}^* &= \mathbf{x}^k - \mathbf{x}^* - [F'(\mathbf{x}^k)]^{-1} \left(F(\mathbf{x}^k) - F(\mathbf{x}^*) \right) \\ &= [F'(\mathbf{x}^k)]^{-1} \left(F(\mathbf{x}^*) - F(\mathbf{x}^k) - F'(\mathbf{x}^k)(\mathbf{x}^* - \mathbf{x}^k) \right) \end{aligned} \quad (4.3)$$

Note that $F(\mathbf{x}^*) = F(\mathbf{x}^k) + F'(\mathbf{x}^k)(\mathbf{x}^* - \mathbf{x}^k) + O(\|\mathbf{x}^k - \mathbf{x}^*\|^2)$, which implies

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq \|[F'(\mathbf{x}^k)]^{-1}\| O(\|\mathbf{x}^k - \mathbf{x}^*\|^2) = O(\|\mathbf{x}^k - \mathbf{x}^*\|^2).$$

■

R During the proof we assume two things:

1. Step-size is 1!
 2. The limit exists.
- Hence, in practice, to implement Newton's method, **1** is the first choice; but gradient descent method is not.
 - Newton's method is good for nice problems, i.e., the function is convex, and the inverse of gradient is easy to solve.
 - In machine learning most time we implement the gradient descent method, since Newton's method is expensive for computing the inverse.

4.1.4. Tutorial: Introduction to Convexity

First we discuss some exercises:

1. Given a sequence of convex functions f_i , the maximum over all functions

$$\max\{f_1(x), f_2(x), \dots, f_n(x)\} := f(x)$$

is also convex. (Proof using Epi-Graph)

2. Given a sequence of convex functions f_i , the combination $\sum_i \lambda_i f_i$ is also convex
3. Composition function may not be convex, .e.g., if $h = x^2$ and $g = -x$, we find $g \circ h$ is concave.
4. However, if g is convex and h is affine, then $g \circ h$ is convex; if g, h is convex, and g is non-decreasing, then $g \circ h$ is convex.

Convexity Examples. Given the problem

$$\begin{array}{ll} \min & \|\mathbf{x}\|_0 \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b}, \end{array}$$

which is difficult to solve. Alternatively, we relax it and solve

$$\begin{array}{ll} \min & \|\mathbf{x}\|_1 \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{array}$$

Also, solving the problem

$$\begin{array}{ll} \min & f(X) \\ \text{s.t.} & \text{rank}(X) = k \end{array}$$

is hard, we relax it and solve

$$\begin{array}{ll} \min & f(X) \\ \text{s.t.} & \|X\|_* \leq k \end{array}$$

Announcement. Learn and implement these things by yourself:

- Luo's note: Lecture #4, P7, Nesterov's optimal 1st-order (also called acceleration) method

- Textbook P67-P72.

Chapter 5

Week5

5.1. Monday

5.1.1. Review

Optimality Condition. Given a general problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X \subseteq \mathbb{R}^n \\ & f \text{ is } \mathcal{C}^1 \text{ or } \mathcal{C}^2 \end{aligned}$$

One of the most important thing is the optimality condition.

- Unconstrained: $X = \mathbb{R}^n$. (First order and second order)
- Constrained:
 - 1st order necessary condition: Let \mathbf{x}^* be a local minimum, then

$$\langle \nabla f(\mathbf{x}^*), (\mathbf{x} - \mathbf{x}^*) \rangle \geq 0, \forall \mathbf{x} \in X$$

- For convex function f , the above is also the sufficient condition, since

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), (\mathbf{x} - \mathbf{x}^*) \rangle, \forall \mathbf{x}, \mathbf{x}^* \in X.$$

The optimality condition for constrained problem is difficult to check. We want a more efficient way, which will be discussed later.

Iterative descent methods. For unconstrained problem, we consider the iterative descent methods:

$$\mathbf{x} \leftarrow \mathbf{x} - \alpha \cdot \mathbf{D} \cdot \nabla f(\mathbf{x}) \quad (\mathbf{D} \succ 0)$$

- If $\mathbf{D} = \mathbf{I}$, it is the first order (gradient) method.
- If $\mathbf{D} = (\nabla^2 f(\mathbf{x}))^{-1}$, it is the second order (Newton's) method
- Sometimes it is difficult to compute $\nabla^2 f(\mathbf{x})$. We can apply finite difference method to accurately approximate the Hessian matrix.
- If $\mathbf{D} = (\mathbf{J}^T \mathbf{J})^{-1}$ with Jacobian matrix for nonlinear least squares problem, it is the Gauss-Newton method.
- Sometimes we apply rough method to approximate the Hessian matrix inaccurately, which is called the **Quasi-Newton** method. The most famous one is BFGS (L-BFGS).

There are more generalized iterative descent methods, such as the accelerated descent method tried in Assignment 3.

Reading materials. CG-conjugate gradient methods; and Nestorov's method (**optimal** accelerated method in **worse** case).

There is a method which is much faster than Nestorov's method in most cases:

$$\mathbf{D}^r = \frac{1}{L} \mathbf{I} + \frac{\mathbf{S}^r (\mathbf{S}^r)^T}{(\mathbf{y}^r)^T (\mathbf{y}^r)} \succ 0$$

$$\alpha = 1$$

$$\mathbf{S}^r = \mathbf{x}^{r+1} - \mathbf{x}^r$$

$$\mathbf{y}^r = \nabla f(\mathbf{x}^{r+1}) - \nabla f(\mathbf{x}^r)$$

Step-size.

- Back-tracking with Armijo condition:

$$\text{Armijo condition} \quad f(\mathbf{x} + \alpha \mathbf{d}) \leq f(\mathbf{x}) + C_1 \alpha \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle, 0 < C_1 < 1$$


- Wolfe condition for line search:

$$\text{Wolfe condition} \quad \begin{cases} f(\mathbf{x} + \alpha \mathbf{d}) \leq f(\mathbf{x}) + C_1 \alpha \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle, & 0 < C_1 < 1 \\ \langle \nabla f(\mathbf{x} + \alpha \mathbf{d}), \mathbf{d} \rangle \geq C_2 \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle, & 0 < C_2 < 1 \end{cases}$$

Define $h(\alpha) = f(\mathbf{x} + \alpha \mathbf{d})$, then the Wolfe condition is essentially

$$\begin{cases} h(\alpha) \leq h(0) + C_1 h'(0) \\ h'(\alpha) \geq C_2 h'(0) \end{cases}$$

- Constant step-size: $\alpha^r \equiv \frac{1}{L}$ with L be the Lipschitz constant of $\nabla f(\mathbf{x})$.
- $\alpha^r \rightarrow 0$ with $\sum \alpha^r = +\infty$.

 Amijo condition guarantees that $f(\mathbf{x}^r) - f(\mathbf{x}^{r+1}) \geq -C_1 \alpha^r \langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle$. Assume $f(\mathbf{x}) > -\infty$, then

$$\alpha^r \langle \nabla f(\mathbf{x}^r), \mathbf{d}^r \rangle \rightarrow 0$$

We want $\nabla f(\mathbf{x}^r) \rightarrow 0$, which means your direction \mathbf{d}^r should not be perpendicular to $\nabla f(\mathbf{x}^r)$ after some iterations. If choosing $\mathbf{d}^r = -\nabla f(\mathbf{x}^r)$, then $\alpha^r \|\nabla f(\mathbf{x}^r)\|^2 \rightarrow 0$, which implies $\nabla f(\mathbf{x}^r) \rightarrow 0$.

Under reasonable conditions, applying first order condition we expect $\nabla f(\mathbf{x}^r) \rightarrow 0$. Is \mathbf{x}^r always convergent? not necessarily.

Local convergence rate. The first order method has linear or sub-linear convergence rate; while the second order method has quadratic convergence rate.

Finite difference Method. Given $F(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}^n$, its Jacobian is given by:

$$F'(\mathbf{x}) = \begin{bmatrix} \nabla^T F_1(\mathbf{x}) \\ \vdots \\ \nabla^T F_n(\mathbf{x}) \end{bmatrix}$$

Its j th column is given by:

$$\begin{aligned} F'(\mathbf{x})\mathbf{e}_j &:= \lim_{h \rightarrow 0} \frac{F(\mathbf{x} + h\mathbf{e}_j) - F(\mathbf{x})}{h} \\ &\approx \frac{F(\mathbf{x} + h\mathbf{e}_j) - F(\mathbf{x})}{h} \text{ for small } h \end{aligned}$$

where for $\varepsilon = 10^{-8}$,

$$h = \varepsilon \max\{1, |x_j|\} \text{sign}(x_j),$$

more-multiplying the term $\text{sign}(x_j)$ means we avoid subtract between \mathbf{x} and $h\mathbf{e}_j$.

5.1.2. Existence of solution to Quadratic Programming

Theorem 5.1 Let $\{S^k\}$ be a sequence of non-empty closed nested sets. Suppose that all **asymptotic** sequences corresponding to asymptotic directions of $\{S^k\}$ are **retractive**, then $\bigcap_{k=0}^{\infty} S^k$ is **non-empty**.

Definition 5.1 Let $\{S^k\}$ be a sequence of non-empty closed nested sets. We say that a vector $\mathbf{d} \neq \mathbf{0}$ is an **asymptotic direction** of $\{S^k\}$ if there exists a sequence $\{\mathbf{x}^k\}$ such that

$$\mathbf{x}^k \in S^k, \quad \mathbf{x}^k \neq \mathbf{0}, \quad k = 0, 1, 2, \dots$$

and

$$\|\mathbf{x}^k\| \rightarrow \infty, \quad \frac{\mathbf{x}^k}{\|\mathbf{x}^k\|} \rightarrow \frac{\mathbf{d}}{\|\mathbf{d}\|}$$

- $(\{\mathbf{x}^k\}, \mathbf{d})$ is said to be the asymptotic pair of $\{S^k\}$
- $\{\mathbf{x}^k\}$ is said to be **retractive** if there exists a bounded sequence of positive numbers $\{\alpha^k\}$ and \bar{k} such that

$$\mathbf{x}^k - \alpha^k \mathbf{d} \in S^k, \quad \forall k \geq \bar{k}$$

■

Theorem 5.2 Let \mathbf{Q} be a positive semi-definite symmetric $n \times n$ matrix, let \mathbf{c} and $\mathbf{a}_1, \dots, \mathbf{a}_r$ be vectors in \mathbb{R}^n , and let b_1, \dots, b_r be scalars. Assume that the optimal value of the problem

$$\begin{aligned} \min \quad & \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x} \\ \text{such that} \quad & \mathbf{a}_j^T \mathbf{x} + b_j \leq 0, \quad j = 1, 2, \dots, r, \end{aligned} \quad (5.1)$$

is finite. Then the problem has **at least one optimal** solution.

Proof. Suppose f^* is the optimal solution. The feasible region is denoted by:

$$F = \{\mathbf{x} \mid \mathbf{a}_j^T \mathbf{x} + b_j \leq 0, j = 1, \dots, r\}.$$

Set a decreasing sequence $\{\gamma^k\}$ with limit f^* , and set

$$S^k := \{\mathbf{x} \in F \mid \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x} \leq \gamma^k\}$$

Thus the set of optimal solutions is $\bigcap_{k=0}^{\infty} S^k$. It suffices to show that all asymptotic sequences corresponding to asymptotic directions are **retractive**.

Asymptotic Directions are essentially Boundary Directions. For fixed asymptotic pair $(\{\mathbf{x}^k\}, \mathbf{d})$, we claim that

$$\mathbf{Q} \mathbf{d} = 0, \langle \mathbf{c}, \mathbf{d} \rangle \leq 0 \quad (5.2)$$

$$\langle \mathbf{a}_j, \mathbf{d} \rangle \leq 0, j = 1, 2, \dots, r \quad (5.3)$$

For first equality, define $\mathbf{d}^k = \frac{\mathbf{x}^k}{\|\mathbf{x}^k\|}$. Since $\mathbf{x}^k \in S^k$, we have

$$(\mathbf{d}^k)^T \mathbf{Q} \mathbf{d}^k + \frac{\langle \mathbf{c}, \mathbf{d}^k \rangle}{\|\mathbf{x}^k\|} \leq \frac{\gamma^k}{\|\mathbf{x}^k\|^2} \quad (5.4)$$

Taking $k \rightarrow \infty$, we imply $(\mathbf{d})^T \mathbf{Q} \mathbf{d} \leq 0$, and therefore $\mathbf{Q} \mathbf{d} = 0$ as $\mathbf{Q} \succeq 0$.

Due to (5.4) and the semi-definiteness of \mathbf{Q} , we have

$$\langle \mathbf{c}, \mathbf{d}^k \rangle \leq \|\mathbf{x}^k\| (\mathbf{d}^k)^T \mathbf{Q} \mathbf{d}^k + \langle \mathbf{c}, \mathbf{d}^k \rangle \leq \frac{\gamma^k}{\|\mathbf{x}^k\|}$$

Taking $k \rightarrow \infty$, we imply $\langle \mathbf{c}, \mathbf{d}^k \rangle \leq 0$. Similarly, $\langle \mathbf{a}_j, \mathbf{d} \rangle \leq 0$ since $\langle \mathbf{a}_j, \mathbf{d}^k \rangle \leq -\frac{b_j}{\|\mathbf{x}^k\|}$.

Finiteness of optimal value. Next we show $\langle \mathbf{c}, \mathbf{d} \rangle = 0$. For a feasible vector $\bar{\mathbf{x}}$, consider $\tilde{\mathbf{x}} := \bar{\mathbf{x}} + m\mathbf{d}$ for any positive m , which is also feasible as $\langle \mathbf{a}_j, \mathbf{d} \rangle \leq 0$. Then checking the function evaluated at $\tilde{\mathbf{x}}$:

$$f^* \leq (\tilde{\mathbf{x}})^T \mathbf{Q}(\tilde{\mathbf{x}}) + \langle \mathbf{c}, \tilde{\mathbf{x}} \rangle = (\bar{\mathbf{x}})^T \mathbf{Q}\bar{\mathbf{x}} + \langle \mathbf{c}, \bar{\mathbf{x}} \rangle + m\langle \mathbf{c}, \mathbf{d} \rangle$$

As f^* is finite, $\langle \mathbf{c}, \mathbf{d} \rangle \geq 0$, i.e., $\langle \mathbf{c}, \mathbf{d} \rangle = 0$.

As a result, for fixed \mathbf{x}^k , the function evaluated at $\mathbf{x}^k - \alpha\mathbf{d}$ satisfies

$$(\mathbf{x}^k - \alpha\mathbf{d})^T \mathbf{Q}(\mathbf{x}^k - \alpha\mathbf{d}) + \langle \mathbf{c}, \mathbf{x}^k - \alpha\mathbf{d} \rangle = (\mathbf{x}^k)^T \mathbf{Q}\mathbf{x}^k + \langle \mathbf{c}, \mathbf{x}^k \rangle \leq \gamma^k, \forall \alpha, k,$$

Feasiblness of $\mathbf{x}^k - \alpha\mathbf{d}$. It suffices to choose $\alpha > 0$ to let $\mathbf{x}^k - \alpha\mathbf{d} \in F$ for sufficiently large k , i.e.,

$$\langle \mathbf{a}_j, \mathbf{x}^k - \alpha\mathbf{d} \rangle + b_j \leq 0, \quad j = 1, \dots, r$$

- This is true for $\forall \alpha > 0$ if $\langle \mathbf{a}_j, \mathbf{d} \rangle = 0$
- Otherwise, suppose $\langle \mathbf{a}_j, \mathbf{d} \rangle < -\varepsilon$. Thus $\langle \mathbf{a}_j, \mathbf{d}^k \rangle < -\varepsilon$ for sufficiently large k , i.e., $\langle \mathbf{a}_j, \mathbf{x}^k \rangle \leq -\varepsilon\|\mathbf{x}^k\|$. Combining the unboundness of $\{\mathbf{x}^k\}$, we imply

$$\langle \mathbf{a}_j, \mathbf{x}^k - \alpha\mathbf{d} \rangle + b_j \leq -\varepsilon\|\mathbf{x}^k\| - \alpha\langle \mathbf{a}_j, \mathbf{d} \rangle + b_j < 0$$

■

5.2. Wednesday

Announcement. You need to study the textbook by yourself. Those materials will be tested in the mid-term.

5.2.1. Comments about Newton's Method

Newton's Method may not necessarily have quadratic rate of convergence. One counter-example is the optimization on the function x^2 , where we have the iteration

$$x^{k+1} = x^k - \frac{1}{2x^k}(x^k)^2 = \frac{x^k}{2},$$

which has linear rate of convergence.

Theorem 5.3 — Sufficient Condition for quadratic convergence of Newton's method.

To minimize the function $f(x)$ with gradient function $F(x) = \nabla f(x)$, the Newton's method iteration gives

$$x^{k+1} = x^k - [F'(x^k)]^{-1} F(x^k),$$

the quadratic (local) rate of convergence is guaranteed if the following conditions hold:

1. there exists x^* such that $F(x^*) = 0$
2. $[F'(x^*)]^{-1}$ exists
3. F is **Lipschitz continuous** near x^* .

For example, to minimize the function $x^2 - x$, the iteration gives the quadratic convergence:

$$x^{k+1} - 1 = \frac{1}{2}(x^k - 1)\left(1 - \frac{1}{x^k}\right) = O((x^k - 1)^2)x^{k+1} = x^k - (2x^k)$$

5.2.2. Constant Step-Size Analysis

For the unconstrained minimization

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

with convex function $f \in \mathcal{C}^1$, our iteration for constant step-size is given by

$$x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k).$$

Now we are interested in the convergence rate of this choice of step-size.

Theorem 5.4 — Convergence Rate for invariant step-size. Given a convex function $f \in \mathcal{C}^2$ with Lipschitz gradient, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$

the iteration $x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k)$ with local minimum point x^* gives super-linear convergence rate, i.e.,

$$f(x^k) - f(x^*) \leq \frac{L\|x^0 - x^*\|^2}{k+1}$$

First prove a simple proposition:

Proposition 5.1 A convex function $f \in \mathcal{C}^2$ with Lipschitz gradient constant L has a bounded Hessian matrix:

$$\nabla^2 f(x) \preceq L\mathbf{I}$$

Proof for proposition(5.1). Otherwise $\exists x_0, v$ such that

$$\|\nabla^2 f(x_0)v\| > L\|v\|.$$

Thus we apply Taylor expansion near x_0 for $x = x_0 + v$:

$$\nabla f(x) = \nabla f(x_0) + \nabla^2 f(x_0)(x - x_0) + o(1)(x - x_0)$$

It follows that

$$\|\nabla f(x) - \nabla f(x_0)\| = \|\nabla^2 f(x_0)(x - x_0) + o(1)(x - x_0)\| \leq L\|x - x_0\|$$

Thus for sufficiently small v , we have $\|\nabla f(x) - \nabla f(x_0)\| \leq L\|x - x_0\|$, which is a contradiction. ■

Step 1: Apply Lipschitz condition. We do the Taylor expansion of x^k for the point x^{k+1} :

$$\begin{aligned} f(x^{k+1}) &= f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2}(x^{k+1} - x^k)^T \nabla^2 f(x^k + \tau(x^{k+1} - x^k))(x^{k+1} - x^k) \\ &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \end{aligned}$$

implying that

$$\|\nabla f(x^k)\|^2 \leq 2L[f(x^k) - f(x^{k+1})]$$

and therefore

$$\sum_{k=0}^r \|\nabla f(x^k)\|^2 \leq \sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 \leq 2L[f(x^0) - f(x^*)] \leq L^2\|x^0 - x^*\|^2 \quad (5.5)$$

Step 2: Applying Convexity of f . By the convexity of f and the bound on its gradient, we can estimate the total error $\sum_{k=0}^r (f(x^k) - f(x^*))$:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - \frac{1}{L}\nabla f(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - \frac{2}{L}\langle \nabla f(x^k), x^k - x^* \rangle + \frac{1}{L^2}\|\nabla f(x^k)\|^2 \\ &\leq \|x^k - x^*\|^2 - \frac{2}{L}(f(x^k) - f(x^*)) + \frac{1}{L^2}\|\nabla f(x^k)\|^2 \end{aligned}$$

which implies

$$\begin{aligned}
\sum_{k=0}^r f(x^k) - f(x^*) &\leq \frac{L}{2} \sum_{k=0}^r \left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \right] + \frac{1}{2L} \sum_{k=0}^r \|\nabla f(x^k)\|^2 \\
&\leq \frac{L}{2} \left[\|x^0 - x^*\|^2 - \|x^{r+1} - x^*\|^2 \right] + \frac{L}{2} \|x^0 - x^*\|^2 \\
&\leq L \|x^0 - x^*\|^2
\end{aligned}$$

Step 3: applying monotonicity of $f(x^k) - f(x^*)$. By the monotonicity of $f(x^k) - f(x^*)$,

$$\begin{aligned}
f(x^r) - f(x^*) &\leq \frac{1}{r+1} \sum_{k=0}^r f(x^k) - f(x^*) \\
&\leq \frac{L \|x^0 - x^*\|^2}{r+1}
\end{aligned}$$

R The convergence rate for this method is **super-linear**. This upper bound is order-tight, i.e., we can find one example satisfying the equality.

However, the constant step-size method can be faster if the given function f is **strongly convex**.

Proposition 5.2 The iteration $x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k)$ for a strongly convex function $f \in \mathcal{C}^2$ with Lipschitz gradient gives linear convergence rate, i.e.,

$$f(x^k) - f(x^*) = O(\rho^k),$$

with $\rho = 1 - \frac{\sigma}{L}$.

Step 1: Lipschitz gradient gives upper bound on $f(x^{k+1}) - f(x^k)$. Applying Taylor Expansion,

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &= \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2} (x^{k+1} - x^k)^T \nabla^2 f(x^k + \tau(x^{k+1} - x^k)) (x^{k+1} - x^k) \\
&\leq \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2
\end{aligned}$$

Substituting $x^{k+1} - x^k = -\frac{1}{L}\nabla f(x^k)$, we derive:

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2L}\|\nabla f(x^k)\|^2$$

Step 2: Strongly convex gives upper bound on $f(x^*) - f(x)$. For $\forall x, y$, we have

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2}\|y - x\|^2 \end{aligned}$$

where $\min_i \lambda_i(\nabla^2 f(x)) \geq \sigma > 0, \forall$. Minimizing both sides in terms of y gives

$$f(x^*) \geq f(x) - \frac{1}{2\sigma}\|\nabla f(x)\|^2$$

Step 3: Re-arrange bounds on $f(x^{k+1}) - f(x^*)$. Applying those bounds, we derive

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq f(x^k) - f(x^*) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - f(x^*) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - f(x^*) - \frac{\sigma}{L}[f(x^k) - f(x^*)] \\ &= (1 - \frac{\sigma}{L})[f(x^k) - f(x^*)] \end{aligned}$$

Applying this trick recursively,

$$f(x^k) - f(x^*) \leq (1 - \frac{\sigma}{L})^k [f(x^0) - f(x^*)]$$

Thus we have shown the linear convergence rate $f(x^k) - f(x^*) = O(\rho^k)$ with $\rho = 1 - \frac{\sigma}{L}$.

Reading Assignment: .

1. Nesterov's Introductory Courses on Convex Programming
2. Prof. Luo's note #4.

