

# CO 367

272284444

September 2023

## Contents

<b>1</b>	<b>Review</b>	<b>2</b>
1.1	Calculus . . . . .	2
1.2	Linear Algebra . . . . .	7
<b>2</b>	<b>Introduction</b>	<b>10</b>
2.1	Lecture 1-Preliminaries . . . . .	10
2.2	Lecture 2 . . . . .	12
<b>3</b>	<b>Unconstrained Optimization</b>	<b>14</b>
3.1	Lecture 2 . . . . .	14
3.2	Lecture 3 . . . . .	15
3.3	Lecture 4 . . . . .	18
<b>4</b>	<b>Linear Least Squares &amp; Solving Linear Systems</b>	<b>22</b>
4.1	Lecture 5 . . . . .	22
4.1.1	Linear Least Squares . . . . .	22
4.1.2	Best Linear Least Squares . . . . .	24
4.2	Lecture 6 . . . . .	24
4.3	Lecture 7 . . . . .	29
<b>5</b>	<b>Iterative Methods for Unconstrained Optimization</b>	<b>30</b>
5.1	Lecture 8-10 . . . . .	30
5.1.1	Line Search Strategy . . . . .	31
5.1.1.1	Finding Descent Direction . . . . .	31
5.1.2	Second order model: Newton Type . . . . .	32
5.1.2.1	Finding Step Size . . . . .	33
5.1.2.2	Steepest Descent Method . . . . .	34
5.1.2.3	Backtracking Line Search . . . . .	35
5.1.2.4	Newton's Method . . . . .	35
5.1.2.5	Quasi-Newton Methods . . . . .	36
5.2	Lecture 11 . . . . .	36
5.2.1	Convergence of Line Search Methods . . . . .	36
5.2.2	Convergence Rate of Steepest Descent . . . . .	37
5.2.3	Convergence Rate of Newton's Method . . . . .	37
5.2.4	Trust Region Strategy . . . . .	38
<b>6</b>	<b>Convex set and function</b>	<b>39</b>
6.1	Lecture 11-13 . . . . .	39

# 1 Review

## 1.1 Calculus

### Definition 1.1 – Norm

$\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  is a mapping from  $\mathbb{R}^n$  to  $\mathbb{R}$  such that

- For every  $\mathbf{x} \in \mathbb{R}^n$ ,  $\|\mathbf{x}\| \geq 0$
- $\|\mathbf{x}\| = 0$  if and only if  $\mathbf{x} = \mathbf{0}$ .
- For every  $\alpha \in \mathbb{R}$  and for every  $\mathbf{x} \in \mathbb{R}^n$ ,  $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
- (Triangle inequality) For every  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ ,

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

Two common norms,  $\ell_1$  and  $\ell_2$  norms:

1.  $\ell_1$ :  $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| = |x_1| + |x_2| + \cdots + |x_n|$
2.  $\ell_2$ :  $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

### Proposition 1.1 – Cauchy-Schwarz Inequality

For every  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$|\mathbf{x}^T \mathbf{y}| = \langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

with equality if and only if  $\mathbf{x} = \alpha \mathbf{y}$  for some  $\alpha \in \mathbb{R}$ .

### Definition 1.2 – Neighborhood/Open ball

Given  $\delta > 0$ ,  $\bar{\mathbf{x}} \in \mathbb{R}^n$ , the open ball  $B_\delta(\bar{\mathbf{x}}) = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \bar{\mathbf{x}}\|_2 < \delta\}$

### Definition 1.3 – Sequence Convergence in $\mathbb{R}^n$

We say that a sequence  $\{\mathbf{x}_k\} \subseteq \mathbb{R}^n$  converges to  $\mathbf{x}^* \in \mathbb{R}^n$  and write

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}^*$$

if for every  $\epsilon > 0$  there exists  $N \in \mathbb{N}$  such that for every  $k \geq N$ ,  $\|\mathbf{x}_k - \mathbf{x}^*\| < \epsilon$ .

### Definition 1.4 – Limit Point

If  $\{\mathbf{x}_k\}$  has a subsequence that converges to  $\mathbf{x}^*$ , then  $\mathbf{x}^*$  is called a limit point of  $\{\mathbf{x}_k\}$ .

Given a set  $E \subseteq \mathbb{R}^n$ , if there exists a sequence  $\{\mathbf{x}_k\} \subseteq E$  that converges to  $\mathbf{x}^*$ , then  $\mathbf{x}^*$  is called a limit point of  $E$ .

### Definition 1.5 – Closed Set

A set  $E \subseteq \mathbb{R}^n$  is closed if it contains all of its limit points.

That is, for every sequence  $\{\mathbf{x}_k\} \subseteq E$  with

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}^*$$

if  $\mathbf{x}^* \in E$ , then  $E$  is closed.

### Definition 1.6 – Interior Point

A point  $\mathbf{x} \in E$  is an interior point of  $E$  if there is a neighborhood/open ball of  $\mathbf{x}$  that is contained in  $E$ .

### Definition 1.7 – Open Set

A set  $E$  is open if all of its elements are interior points.

So, for any point you pick in  $E$ , you can find a small neighborhood around that point which is entirely contained in  $E$  (there are no boundary points in  $E$ ).

### Theorem 1.1 – Properties of Open/Closed Sets

The followings hold:

- A set is closed (open) if its complement,  $\mathbb{R}^n \setminus E$ , is open (closed).
- Union of finitely many closed sets is closed
- Intersection of (finitely or infinitely many) closed sets is closed.
- Intersection of finitely many open sets is open
- Union of (finitely or infinitely many) open sets is open.

### Definition 1.8 – Bounded

A set  $E \subset \mathbb{R}^n$  is bounded if it can be contained in a ball of finite radius. That is, there exists a neighborhood,  $B_\delta(\mathbf{x})$ , such that  $E \subseteq B_\delta(\mathbf{x})$ .

### Definition 1.9 – Compact

A set  $E \subset \mathbb{R}^n$  is compact if it is closed and bounded.

### Definition 1.10 – Lipschitz Continuous/Contraction

Let  $E \subseteq \mathbb{R}^n$ . Let  $f : E \rightarrow \mathbb{R}^m$  be a function. We say  $f$  is Lipschitz continuous on  $E$  if there exists a constant  $L > 0$  such that for all  $\mathbf{x}, \mathbf{y} \in E$ ,

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

### Theorem 1.2 – Continuity of Functions

Let  $E \subseteq \mathbb{R}^n$ ,  $f, g : E \rightarrow \mathbb{R}^m$  and  $\alpha \in \mathbb{R}$ . If  $f$  and  $g$  are continuous at  $\mathbf{x}_0$ , then

1.  $f + g, fg, \alpha f$  are continuous at  $\mathbf{x}_0$
2.  $\frac{f}{g}$  is continuous at  $\mathbf{x}_0$  provided that  $g(\mathbf{x}_0) \neq 0$ .

### Definition 1.11 – Formal Definition of Derivative

The **derivative** of  $f$  at  $a$  is defined as

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

if the limit exists.

An alternate definition is

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

**Theorem 1.3 – Extreme Value Theorem (EVT)**

If  $E \subset \mathbb{R}^n$  is a compact set (closed and bounded) and  $f : E \rightarrow \mathbb{R}$  is a continuous function, then  $f$  attains a maximum and minimum on  $E$ . That is, there exists points  $\mathbf{x}_{\min}, \mathbf{x}_{\max} \in E$  such that for all  $\mathbf{x} \in E$ ,

$$f(x) \leq f(\mathbf{x}_{\max}) \quad \text{and} \quad f(\mathbf{x}_{\min}) \leq f(\mathbf{x})$$

**Theorem 1.4 – Continuously Differentiable**

$f$  is continuously differentiable at  $\mathbf{x}_0$  if all partial derivatives exist and are continuous in a neighborhood of  $\mathbf{x}_0$ . We say  $f$  is continuously differentiable,  $f \in C^1$ , if its partial derivatives are continuous everywhere.

$f$  is twice differentiable on  $E$  if  $\nabla^2 f(x)$  exists for all  $\mathbf{x} \in E$ . If each entry of the Hessian  $\nabla^2 f(\mathbf{x})$  is continuous, we say  $f$  is twice differentiable on  $E$ ,  $f \in C^2$ .

**Theorem 1.5**

If  $f : E \rightarrow \mathbb{R}$  is twice differentiable, then the Hessian is symmetric.

**Theorem 1.6 – Mean Value Theorem (MVT)**

Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be continuous on the closed interval  $[a, b]$  and differentiable on the open interval  $(a, b)$ . Then there exists  $c \in (a, b)$  such that

$$f(b) - f(a) = f'(c)(b - a)$$

**Lemma 1.1 – Directional Derivative**

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $\bar{x}, d \in \mathbb{R}^n$  where  $d$  is the direction. We define

$$\phi(\epsilon) = f(\bar{x} + \epsilon d) : \mathbb{R} \rightarrow \mathbb{R}$$

the value of the function  $f$  at a point that is displaced from  $\bar{x}$  by a distance of  $\epsilon$  in the direction  $d$ . Then the **directional derivative**, denoted  $f'(x; d)$  of  $f$  at  $x$  at the direction  $d$  is

$$f'(x; d) = \phi'(0) = \lim_{\epsilon \rightarrow 0} \frac{f(x + \epsilon d) - f(x)}{\epsilon} = \nabla f(x)^T d$$

**Definition 1.12 – Directional Derivative (Different notation)**

The directional derivative of  $f$  at a point  $\bar{x} \in \mathbb{R}^n$  in the direction  $d$  is

$$f'(\bar{x}; d) = \left. \frac{d}{ds} f(\bar{x} + sd) \right|_{s=0}$$

**Theorem 1.7**

If  $f$  is differentiable at  $\bar{x}$ , then

$$f'(\bar{x}; d) = \nabla f(\bar{x})^T d$$

*Proof.* We only prove in the case where  $\bar{x} = (a, b) \in \mathbb{R}^2$ .

$$\begin{aligned}
 f'(\bar{x}; d) &= \left. \frac{d}{ds} f(\bar{x} + sd) \right|_{s=0} \\
 &= \left. \frac{d}{ds} f(\underbrace{a + sd_1}_x, \underbrace{b + sd_2}_y) \right|_{s=0} \\
 &= \left[ \frac{\partial f}{\partial x} \frac{dx}{ds} + \frac{\partial f}{\partial y} \frac{dy}{ds} \right]_{s=0} && \text{Chain rule} \\
 &= \left[ \frac{\partial}{\partial x} f(a + sd_1, b + sd_2) \cdot d_1 + \frac{\partial}{\partial y} f(a + sd_1, b + sd_2) \cdot d_2 \right]_{s=0} \\
 &= \frac{\partial f}{\partial x}(a, b) \cdot d_1 + \frac{\partial f}{\partial y}(a, b) \cdot d_2 \\
 &= \left( \frac{\partial f}{\partial x}(a, b), \frac{\partial f}{\partial y}(a, b) \right) \cdot (d_1, d_2) \\
 &= \nabla f(a, b) \cdot d
 \end{aligned}$$

□

### Problem 1.1

Let  $f(x, y, z) = x^2z + y^3z^2 - xyz$  with  $d = \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix}$ . Then the **directional derivative** in the direction  $d$  is

$$\nabla f(x, y, z)^T d = \begin{pmatrix} 2xz - yz \\ 3y^2z^2 - xz \\ x^2 + 2y^3z - xy \end{pmatrix}^T \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix} = -2xz + yz + 3x^2 + 6y^3z - 3xy$$

### Theorem 1.8 – Taylor’s Theorem on the real line

Let  $f : (a, b) \rightarrow \mathbb{R}$ , and  $\bar{x}, x \in (a, b)$ . Then the Taylor’s series centered at  $\bar{x}$  (approximation near  $\bar{x}$ ) is

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{f''(z)}{2}(x - \bar{x})^2$$

where  $z$  is between  $x$  and  $\bar{x}$  that gives the largest value of  $f''(z)$ . The term  $\frac{f''(z)}{2}(x - \bar{x})^2$  is the **error term**

Equivalently,

$$f(\bar{x} + \Delta x) = \underbrace{f(x) + f'(x)\Delta x}_{\text{Linear approximation}} + o(|\Delta x|) \text{ (little O)}$$

This formula emphasizes its use in approximating changes in  $f$  for small changes in  $x$ , denoted  $\Delta x$ .  $o(|\Delta x|)$ , the error term, means that the error goes to 0 faster than  $|\Delta x|$  as  $\Delta x$  goes to 0. Therefore, this is saying that the linear approximation becomes more and more accurate for smaller  $\Delta x$ .

### Theorem 1.9 – Multivariate Taylor

Consider a  $C^2$ -smooth function  $f : U \rightarrow \mathbb{R}$  on an open set  $U \subset \mathbb{R}^n$ . If  $\bar{x}$  and  $x$  are such that the segment  $[\bar{x}, x] := \{\bar{x} + t(x - \bar{x}) : t \in [0, 1]\}$  is contained in  $U$ , then the Taylor series expansion of  $f$  centered

around  $\bar{x}$  is

$$f(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle \nabla^2 f(z)(x - \bar{x}), (x - \bar{x}) \rangle$$

where  $z$  is between  $x$  and  $\bar{x}$ .

### Theorem 1.10 – Taylor’s Theorem

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be continuously differentiable and let  $\mathbf{d} \in \mathbb{R}^n$  be a direction vector. Then

$$f(\mathbf{x} + \mathbf{d}) = f(\mathbf{x}) + \underbrace{\nabla f(\mathbf{x} + t\mathbf{d})^T \mathbf{d}}_{\text{directional derivative}}$$

for some  $t \in (0, 1)$ . Moreover, if  $f$  is twice differentiable, then

$$f(\mathbf{x} + \mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{d} + \frac{1}{2} \underbrace{\mathbf{d}^T \nabla^2 f(\mathbf{x} + t\mathbf{d}) \mathbf{d}}_{\text{second order directional derivative}}$$

for some  $t \in (0, 1)$ .

This theorem provides a way to approximate the value of the function  $f$  at a point  $\mathbf{x} + \mathbf{d}$  based on the value and derivatives of  $f$  at or near the point  $\mathbf{x}$ .

### Theorem 1.11 – Taylor’s Theorem (alternative)

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be continuously differentiable at  $\mathbf{x}^*$ . Then  $\forall \mathbf{x} \in \mathbb{R}^n$

$$f(\mathbf{x}) = \underbrace{f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*)}_{\text{linear approximation of } f \text{ at } \mathbf{x}^*} + o(\|\mathbf{x} - \mathbf{x}^*\|)$$

where  $o(\|\mathbf{x} - \mathbf{x}^*\|)$  is the error term that goes to 0 faster than  $\|\mathbf{x} - \mathbf{x}^*\|$  as  $\mathbf{x} \rightarrow \mathbf{x}^*$ .

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be twice continuously differentiable at  $\mathbf{x}^*$ . Then  $\forall \mathbf{x} \in \mathbb{R}^n$

$$f(\mathbf{x}) = \underbrace{f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \nabla^2 f(\mathbf{x}^*) (\mathbf{x} - \mathbf{x}^*)}_{\text{quadratic approximation of } f \text{ at } \mathbf{x}^*} + o(\|\mathbf{x} - \mathbf{x}^*\|^2)$$

where  $o(\|\mathbf{x} - \mathbf{x}^*\|^2)$  is the error term that goes to 0 faster than  $\|\mathbf{x} - \mathbf{x}^*\|^2$  as  $\mathbf{x} \rightarrow \mathbf{x}^*$ .

### Problem 1.2 – Lagrange Multiplier Example

Maximize  $f(x, y) = x^2 y$  subject to  $g(x, y) = x^2 + y^2 = 1$ . By using Lagrange multipliers, we know that the maximizer,  $(x^*, y^*)$  satisfies

$$\nabla f(x^*, y^*) = \lambda \nabla g(x^*, y^*)$$

We have

$$\begin{aligned} \nabla g(x, y) &= \begin{bmatrix} 2x \\ 2y \end{bmatrix} \\ \nabla f(x, y) &= \begin{bmatrix} 2xy \\ x^2 \end{bmatrix} \end{aligned}$$

Then, using Lagrange multipliers, we solve

$$\begin{bmatrix} 2xy \\ x^2 \end{bmatrix} = \lambda \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

Since we also have the constraint  $x^2 + y^2 = 1$ , we solve the system of equations

$$\begin{aligned} 2xy &= 2\lambda x \\ x^2 &= 2\lambda y \\ x^2 + y^2 &= 1 \end{aligned}$$

Solving it, we get

$$(x, y) = \left( \pm\sqrt{\frac{2}{3}}, \pm\sqrt{\frac{1}{3}} \right), \quad \lambda = y$$

Testing each point, we get that  $\left( \sqrt{\frac{2}{3}}, \sqrt{\frac{1}{3}} \right)$  is the maximizer of  $f$ .

## 1.2 Linear Algebra

### Definition 1.13 – Characteristic Polynomial

The characteristic polynomial of  $A \in \mathbb{R}^{n \times n}$  is

$$p(\lambda) = \det(A - \lambda I)$$

The degree of  $p(\lambda)$  is  $n$ , and the leading term is  $(-1)^n \lambda^n$ .

The eigenvalues of  $A$  are the roots of the characteristic polynomial.

### Definition 1.14 – Eigenvectors and Eigenvalues

$0 \neq v \in \mathbb{R}^n$  is an **eigenvector** of  $A$  if there exists  $\lambda \in \mathbb{R}$  such that  $Av = \lambda v$ . The number  $\lambda$  is called an **eigenvalue** of  $A$ .

### Theorem 1.12 – Finding Eigenvectors and Eigenvalues

Let  $A$  be a matrix.

1. Set up the characteristic equation. We find

$$\det(A - \lambda I) = 0$$

2. Solve for  $\lambda$ . These are the eigenvalues.
3. Plug eigenvalues  $\lambda_1, \dots, \lambda_n$  into  $(A - \lambda I)v = 0$  and solve for  $v$ . These are the eigenvectors.

### Definition 1.15 – Quadratic Form

Let  $A$  be a symmetric matrix and  $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ . The **quadratic form**  $Q$  of the matrix  $A$  is defined as

$$Q = x^T A x$$

### Problem 1.3 – Example

Consider the matrix  $A = \begin{bmatrix} 5 & -5 \\ -5 & 1 \end{bmatrix}$ . The quadratic form of  $A$  is

$$Q(x) = 5x_1^2 - 10x_1x_2 + x_2^2$$

**Definition 1.16 – Classification of Quadratic Forms**

Let  $Q = x^T Ax$  be a quadratic form of a matrix  $A$ . Then  $A$  is

1. positive definite if  $Q(x) > 0$  for all non-zero vectors  $x$ , and  $Q(x) = 0$  if and only if  $x = 0$ . Or all eigenvalues of  $A$  are positive. Denoted by  $A \succ 0$ .
2. positive semidefinite if  $Q(x) \geq 0$  for all vectors  $x$ , with  $Q(x) = 0$  occurring for some non-zero vectors  $x$ . Or all eigenvalues of  $A$  are non-negative. Denoted by  $A \succeq 0$ .
3. negative definite if  $Q(x) < 0$  for all non-zero vectors  $x$ , and  $Q(x) = 0$  if and only if  $x = 0$ . Or all eigenvalues of  $A$  are negative. Denoted by  $A \prec 0$ .
4. negative semidefinite if  $Q(x) \leq 0$  for all vectors  $x$ , with  $Q(x) = 0$  occurring for some non-zero vectors  $x$ . Or all eigenvalues of  $A$  are non-positive. Denoted by  $A \preceq 0$ .
5. indefinite if  $Q(x)$  can be positive or negative. Or there are positive and negative eigenvalues for  $A$ .

**Theorem 1.13 – Orthogonal Spectral Decomposition**

Let  $A \in \mathbb{S}^n$ . Then  $A$  has an **orthogonal spectral decomposition**

$$A = \sum_i \lambda_i u_i u_i^T = U D U^T$$

where  $U$  is orthogonal with the orthogonal eigenvectors  $u_i$  as columns and  $D$  is a diagonal matrix with real eigenvalues on the diagonal.

**Proposition 1.2**

Let  $A \in \mathbb{S}^n$ . The following are equivalent (Positive definite):

1.  $A \succ 0$ .
2. All the eigenvalues of  $A$  are in  $\mathbb{R}_{++}^n$ , the interior of the nonnegative orthant.
3.  $A$  has a real symmetric positive definite square root,  $A = SS$ ,  $S \in \mathbb{S}_{++}^n$ .
4.  $A$  has a lower triangular factorization, a Cholesky factorization,  $A = LL^T$  and  $L$  has positive diagonal elements.
5. All principal minors of  $A$  are positive.
6. All leading principal minors of  $A$  are positive.

And the following are equivalent (Positive semidefinite):

1.  $A \succeq 0$ .
2. All the eigenvalues of  $A$  are in  $\mathbb{R}_+^n$ , the nonnegative orthant.
3.  $A$  has a real symmetric square root,  $A = SS$ ,  $S \in \mathbb{S}^n$ .



4.  $A$  has a lower triangular factorization, a Cholesky factorization,  $A = LL^T$ .
5. All principal minors of  $A$  are nonnegative.

#### Definition 1.17 – Principal Submatrices

Let

$$A = \begin{bmatrix} 1 & 1 & 2 & 7 \\ 1 & 1 & 4 & 6 \\ 2 & 4 & 7 & 8 \\ 7 & 6 & 8 & 1 \end{bmatrix}, \quad I = \{1, 3\}, \quad A[I] = \begin{bmatrix} 1 & 2 \\ 2 & 7 \end{bmatrix}$$

Then  $A[I]$  is a **principal submatrix** of  $A$ .

#### Definition 1.18 – Principal Minors

Let  $A \in \mathbb{S}^n$ , where  $\mathbb{S}^n$  is the set of all symmetric  $n \times n$  matrices.

1.  $\det(A[I])$  is called the **principal minor** of  $A$ .
2. If  $I = \{1, \dots, k\}$  then  $\det(A[I])$  is called the **leading principal minor** of  $A$ .

#### Proposition 1.3 – Characterizing Positive Definiteness with Principal Minors

Let  $A \in \mathbb{S}^n$ . Then

1.  $A \succeq 0 \iff \det(A[I]) \geq 0$  for all principal minors  $\det(A[I])$ .
2.  $A \succ 0 \iff \det(A[I]) > 0$  for all **leading** principal minors  $\det(A[I])$ .

#### Theorem 1.14

Let  $A \in \mathbb{R}^{n \times n}$  be symmetric. Then the following are equivalent:

1.  $A$  is positive semidefinite (definite).
2. All eigenvalues of  $A$  are nonnegative (positive).
3.  $A$  can be factored as  $A = BB^T$  where  $B$  is an  $n \times p$  matrix for some  $p$ . (Cholesky factorization)

#### Definition 1.19 – Diagonally Dominant

A matrix  $A$  is diagonally dominant if for every  $i$ ,

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$$

So each diagonal element is greater than or equal to the sum of the absolute values of the other elements in the same row.

It is called strictly diagonally dominant if for every  $i$ ,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

**Proposition 1.4**

Let  $A \in \mathbb{R}^{n \times n}$  be a symmetric, diagonally dominant matrix whose diagonal entries are nonnegative, then  $A$  is positive semidefinite.

Let  $A \in \mathbb{R}^{n \times n}$  be a symmetric, strictly diagonally dominant matrix whose diagonal entries are positive, then  $A$  is positive definite.

Note that the converse is not true.

**Definition 1.20 – Four Fundamental Subspaces**

Let  $A$  be a  $m \times n$  matrix.

- The range space of  $A$  is defined as  $\text{Range}(A) = \{Ax : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$ .
- The range space of  $A^T$  is defined as  $\text{Range}(A^T) = \{A^T y : y \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$ .
- The null space of  $A$  is defined as  $\text{Null}(A) = \{x \in \mathbb{R}^n : Ax = 0\} \subseteq \mathbb{R}^n$ .
- The null space of  $A^T$  is defined as  $\text{Null}(A^T) = \{y \in \mathbb{R}^m : A^T y = 0\} \subseteq \mathbb{R}^m$ .

**Theorem 1.15 – Rank Nullity Theorem**

Let  $A$  be an  $m \times n$  matrix. Then

$$\text{rank}(A) + \dim(\text{Null}(A)) = n$$

Recall that rank of a matrix is the number of pivots in the reduced row echelon form of the matrix. The dimension of a subspace is the number of linearly independent vectors that span the subspace.

## 2 Introduction

### 2.1 Lecture 1-Preliminaries

**Definition 2.1 – Big O and little o**

Big O is basically the rate of growth of that function. A function  $f(n)$  is of order 1, or  $O(1)$  if there exists some non zero constant  $c$  such that

$$\frac{f(n)}{c} \rightarrow 1$$

as  $n \rightarrow \infty$ .

Little o is the upper bound of the rate of growth of that function. Therefore, a function  $f(n)$  is of order 1, or  $o(1)$  if for all constants  $c > 0$ ,

$$\frac{f(n)}{c} \rightarrow 0$$

as  $n \rightarrow \infty$ .

**Definition 2.2 – Differentiability Based on Big o and Little o**

If  $f$  is differentiable at  $x = a$ , then

$$f(a + h) = f(a) + f'(a)h + o(h)$$

Conversely, if there exists constants  $A$  and  $B$  such that

$$f(a + h) = A + Bh + o(h)$$

then  $f$  is differentiable at  $x = a$ . Moreover,  $A = f(a)$  and  $B = f'(a)$ .

**Definition 2.3 – Product Rule**

If  $f, g$  are differentiable at  $x = a$ , then

$$f(a + h) = f(a) + f'(a)h + o(h), \quad g(a + h) = g(a) + g'(a)h + o(h)$$

Then

$$\begin{aligned} p(a + h) &= f(a + h)g(a + h) \\ &= f(a)g(a) + [f(a)g'(a) + g(a)f'(a)]h + o(h) \end{aligned}$$

Then by above theorem,  $p = fg$  is differentiable at  $x = a$ , and  $p'(a) = f(a)g'(a) + g(a)f'(a)$ .

**Definition 2.4 – Chain Rule**

WIP

**Definition 2.5 – Inner Product Space**

Let  $x \in \mathbb{R}^n$ , represented as:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The inner product space is defined as:

$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i \quad (\text{dot product})$$

The angle between vectors  $x$  and  $y$  is given by  $\cos(\theta) = \frac{\langle x, y \rangle}{\|x\|}$ .

With corresponding norm to be the Euclidean Norm

**Definition 2.6 – map**

Suppose the map  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ .

**Definition 2.7 – differ**

We define  $f$  to be in  $C^1, C^2$  on an open set  $D \subseteq \mathbb{R}^n$ , denoted  $f \in C^1(D), C^2(D)$ , respectively, if the partial first  $\frac{\partial f(x)}{\partial x_i}$  and second  $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$  derivatives exist and are continuous for all  $i, j$ , respectively. We then get the gradient vector in  $\mathbb{R}^n$  and the  $n \times n$  symmetric Hessian matrix, respectively denoted as:

$$\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_i} \right) \in \mathbb{R}^n, \quad \nabla^2 f(x) = \left[ \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right] \in \mathbb{S}^n.$$

Here,  $\mathbb{S}^n$  is the vector space of  $n \times n$  symmetric matrices.

**Definition 2.8 – General Nonlinear opt. function NLO**

The general problem of nonlinear optimization, denoted NLO, is defined as follows: Given  $C^2$ -smooth functions  $f, g_i, h_j : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  for  $i = 1, \dots, m$  and  $j = 1, \dots, p$ , where  $D$  is an open subset of  $\mathbb{R}^n$ , the objective is to find the optimal value  $p^*$  and an optimum  $x^*$  of NLO, represented as:

$$p^* := \min f(x) \text{ s.t. } g_i(x) \leq 0, \quad \forall i = 1, \dots, m, h_j(x) = 0, \quad \forall j = 1, \dots, p, x \in D$$

If  $f, g_i, h_j$  are all **affine** function and  $D = \mathbb{R}^2$ , then we have an LP

**Definition 2.9 – affine**

$$f(x) = Ax + b \quad (1)$$

where  $b \neq 0$

**Definition 2.10 – Types of Minimality**

Consider  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $D \subset \mathbb{R}^n$ . Then  $\bar{x} \in D$  is:

- a *global minimizer* for  $f$  on  $D$  if  $f(\bar{x}) \leq f(x)$  for all  $x \in D$ .
- a *strict global minimizer* for  $f$  on  $D$  if  $f(\bar{x}) < f(x)$  for all  $x \in D$  where  $x \neq \bar{x}$ .
- a *local minimizer* for  $f$  on  $D$  if there exists  $\delta > 0$  such that  $f(\bar{x}) \leq f(x)$  for all  $x \in D \cap B_\delta(\bar{x})$ .
- a *strict local minimizer* for  $f$  on  $D$  if there exists  $\delta > 0$  such that  $f(\bar{x}) < f(x)$  for all  $x \in D \cap B_\delta(\bar{x})$  where  $x \neq \bar{x}$ .

**Definition 2.11 – Linear Approximation**

Suppose  $f$  is a function that is differentiable on an interval  $I$  containing the point  $a$ . The **linear approximation** to  $f$  at  $a$  is the linear function

$$L(x) = f(a) + f'(a)(x - a)$$

for  $x \in I$ .

**Definition 2.12 – Quadratic Approximation**

Similar as above, the **quadratic approximation** to  $f$  at  $a$  is the quadratic function

$$Q(x) = f(a) + f'(a)(x - a) + \frac{1}{2}f''(a)(x - a)^2$$

for  $x \in I$ .

**2.2 Lecture 2****Definition 2.13 – General NLO/NLP**

A **Non-linear Optimization Problem** (NLP) is of the following form:

$$\underbrace{p^*}_{\text{Optimal Value}} = \min \underbrace{f(x)}_{\text{Objective function}}$$

s.t.

$$\begin{aligned} g(x) = (g_i(x)) &\leq 0 \in \mathbb{R}^m \\ h(x) = (h_j(x)) &= 0 \in \mathbb{R}^p \end{aligned}$$

### Problem 2.1 – Example

$$\min (x_1 - 2)^2 + (x_2 - 1)^2$$

s.t.

$$\begin{aligned} x_1^2 - x_2 &\leq 0 & (g_1(x) \leq 0) \\ x_1 + x_2 - 2 &\leq 0 & (g_2(x) \leq 0) \end{aligned}$$

### Definition 2.14 – Contour

For  $\alpha \in \mathbb{R}$ , the **contour** of a function  $f$  is

$$C_\alpha = \{x \in \mathbb{R}^n : f(x) = \alpha\}$$

### Definition 2.15 – Feasible Set

The **feasible set** is

$$F = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0, x \in D\}$$

(Is  $D$  the domain??)

### Definition 2.16 – Gradient

The **gradient** of  $f$  is

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

For the optimal solution  $x^*$ , we have

$$\alpha \nabla f(x^*) = \lambda_1 \nabla g_1(x^*) + \lambda_2 \nabla g_2(x^*)$$

for some  $\alpha, \lambda_1, \lambda_2 \in \mathbb{R}$ .

We will see later that we can choose  $\alpha = 1$  and we need  $\lambda_1 \geq 0, \lambda_2 \geq 0$ .

### Problem 2.2 – Max-cut Problem

Given a weighted graph  $G = (\underbrace{V}_{\text{vertices}}, \underbrace{E}_{\text{edges}}, \underbrace{w}_{\text{weight}})$ , a **cut** is  $U \subseteq V, U \neq \emptyset$ . The objective function

$$\max \frac{1}{2} \sum_{\substack{i \in U, j \notin U \\ (i,j) \in E}} w_{i,j}$$

maximizes the sum of edges in a cut.

Formulating as an NLP, we introduce variables  $x_i \in \{\pm 1\}, i = 1, \dots, n$ . Then the Max-cut problem (MC) is as

follows:

$$\max \quad \frac{1}{2} \sum_{ij \in E} w_{ij}(1 - x_i x_j)$$

Why 1/2 s.t.

$$x_i \in \{\pm 1\} \quad (\text{equivalent to } x_i^2 = 1) \quad \forall i = 1, \dots, n$$

This works because

$$1 - x_i x_j = \begin{cases} 0 & \text{if } x_i = x_j \quad (i, j \text{ in the same set, } U \text{ or } U^c) \\ 2 & \text{otherwise} \end{cases}$$

MC is a **quadratically constrained quadratic program** (QOP) since each constraint  $x_i \in \{-1, 1\}$  is equivalent to the quadratic constraint  $x_i^2 = 1$ . Note that MC is an NP-hard problem.

### 3 Unconstrained Optimization

#### 3.1 Lecture 2

##### Problem 3.1 – Simplest Case - No Constraints

Let  $\Omega \subseteq \mathbb{R}^n$  be an open set. Assume  $f$  is sufficiently smooth (differentiable) then the NLP with no constraints is

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

##### Definition 3.1 – Secant Line

A secant line is a line that connects two points on a function.

##### Theorem 3.1 – Chain Rule (2 dimensions)

Let  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ , and two other functions  $x(t) : \mathbb{R} \rightarrow \mathbb{R}$  and  $y(t) : \mathbb{R} \rightarrow \mathbb{R}$ . Let  $\phi(t) = f(x(t), y(t))$ . The chain rule then states

$$\frac{d\phi}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt}$$

##### Problem 3.2 – Example of Chain Rule in 2 Dimensions

Let  $f(x, y) = x^2 + y^2$ . We want to find the rate of change of  $f$  along a curve defined by  $x(t) = t$  and  $y(t) = 2t$ . The partial derivatives of  $f$  are:

$$\frac{\partial f}{\partial x} = 2x, \quad \frac{\partial f}{\partial y} = 2y$$

The derivatives of  $x(t)$  and  $y(t)$  are

$$\frac{dx}{dt} = 1, \quad \frac{dy}{dt} = 2$$

Then we get

$$\begin{aligned} \frac{d}{dt} f(x(t), y(t)) &= \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} \\ &= 2x \cdot 1 + 2y \cdot 2 \\ &= 2(t) \cdot 1 + 2(2t) \cdot 2 \\ &= 10t \end{aligned}$$

**Corollary 3.1**

Let  $f : (a, b) \rightarrow \mathbb{R}$

1. (Necessity) If  $\bar{x}$  is a **local minimizer** of  $f$  on  $(a, b)$ , then  $f'(\bar{x}) = 0$  and  $f''(\bar{x}) \geq 0$ .
2. (sufficient) If  $f(\bar{x}) = 0$ ,  $f''(\bar{x}) > 0$  then  $\bar{x}$  is a **strict local minimizer** of  $f$ .

*Proof.*

Since  $\bar{x}$  is the local minimizer of  $f$ , given small  $|\Delta x|$ ,

$$\begin{aligned} f(\bar{x} + \Delta x) - f(\bar{x}) &\geq 0 \\ f(\bar{x} + \Delta x) &= f(\bar{x}) + f'(\bar{x}) * \Delta x + o(|\Delta x|) \\ f'(\bar{x}) * \Delta x + o(|\Delta x|) &\geq 0 \\ \frac{f'(\bar{x})}{|\Delta x|} * \Delta x + \frac{o(|\Delta x|)}{|\Delta x|} &\geq 0 \end{aligned}$$

Since this holds for all positive and negative  $\Delta x$ . Thus  $f'(\bar{x}) = 0$ . For the second order proof, we used Taylor thm to construct the contradiction  $\square$

**Definition 3.2 – Hessian**

The **Hessian** of  $f$  at  $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$  is the matrix

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

**Lemma 3.1**

Let  $v \in \mathbb{R}^n$ . Then

$$v = 0 \iff \langle v, d \rangle = 0, \quad \forall d \in \mathbb{R}^n$$

**3.2 Lecture 3****Definition 3.3 – Matrix Norm**

$$\|Q\| = \max_{\|x\|=1} \|Qx\| = \text{Largest singular value of } A$$

**Theorem 3.2 – Second Order Optimality Conditions (Min)**

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be twice continuously differentiable at an open set  $D$ . Then

1. Necessary conditions: If  $\bar{x}$  is a local minimizer for  $f$  on  $D$ , then

$$\nabla f(\bar{x}) = 0 \quad \text{and} \quad \nabla^2 f(\bar{x}) \succeq 0$$

2. Sufficient conditions: If  $\nabla f(\bar{x}) = 0$ ,  $\nabla^2 f(\bar{x}) \succ 0$  is positive definite then  $\bar{x}$  is a strict local minimizer

of  $f$  on  $D$ .

*Proof of Necessary conditions.* Assume  $\bar{x}$  is a local minimizer of  $f$ , we show that  $\nabla f(\bar{x}) = 0$  and  $\nabla^2 f(\bar{x}) \succeq 0$ . By multivariate Taylor Theorem, we have

$$f(\bar{x} + d) = f(\bar{x}) + \nabla f(\bar{x})^T d + o(\|d\|)$$

Rearranging, then dividing both sides by  $\|d\|$  gives

$$\frac{f(\bar{x} + d) - f(\bar{x})}{\|d\|} = \nabla f(\bar{x})^T \frac{d}{\|d\|} + \frac{o(\|d\|)}{\|d\|}$$

Since  $\bar{x}$  is a local minimizer,  $f(\bar{x} + d) \geq f(\bar{x})$ , thus  $\frac{f(\bar{x} + d) - f(\bar{x})}{\|d\|} \geq 0$ , so  $\nabla f(\bar{x})^T \frac{d}{\|d\|} \geq 0$  as well. Now,  $\frac{d}{\|d\|}$  is a unit vector with arbitrary  $d$ , so  $\nabla f(\bar{x})^T v \geq 0$  for all unit vectors  $v$ .

Consider  $+v$  and  $-v$ , we have  $\nabla f(\bar{x})^T v \geq 0$  and  $\nabla f(\bar{x})^T (-v) \geq 0$ , so  $\nabla f(\bar{x})^T v = 0$  for all  $v$ . Thus,  $\nabla f(\bar{x}) = 0$ .

We now show that  $\nabla^2 f(\bar{x}) \succeq 0$ . By multivariate Taylor Theorem, we have

$$f(\bar{x} + d) = f(\bar{x}) + \nabla f(\bar{x})^T d + \frac{1}{2} d^T \nabla^2 f(\bar{x}) d + o(\|d\|^2)$$

Rearranging and dividing both sides by  $\|d\|^2$  gives

$$\frac{f(\bar{x} + d) - f(\bar{x}) - \nabla f(\bar{x})^T d}{\|d\|^2} = \frac{1}{2} \frac{d^T \nabla^2 f(\bar{x}) d}{\|d\|^2} + \frac{o(\|d\|^2)}{\|d\|^2}$$

Since  $\nabla f(\bar{x}) = 0$  from first part, we get

$$\frac{f(\bar{x} + d) - f(\bar{x})}{\|d\|^2} = \frac{1}{2} \frac{d^T \nabla^2 f(\bar{x}) d}{\|d\|^2} + \frac{o(\|d\|^2)}{\|d\|^2}$$

Similar to first part,  $\frac{d^T \nabla^2 f(\bar{x}) d}{\|d\|^2} \geq 0$ . Since  $\|d\|^2 \geq 0$ , we have  $d^T \nabla^2 f(\bar{x}) d \geq 0$ . The Hessian is symmetric, so  $\nabla^2 f(\bar{x}) \succeq 0$ .  $\square$

### Theorem 3.3 – Second Order Optimality Conditions (Max)

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be twice continuously differentiable at an open set  $D$ . Then

1. Necessary conditions: If  $\bar{x}$  is a local maximizer for  $f$  on  $D$ , then

$$\nabla f(\bar{x}) = 0 \quad \text{and} \quad \nabla^2 f(\bar{x}) \preceq 0$$

2. Sufficient conditions: If  $\nabla f(\bar{x}) = 0$ ,  $\nabla^2 f(\bar{x}) \prec 0$  is positive definite then  $\bar{x}$  is a strict local maximizer of  $f$  on  $D$ .

*Proof.* 1. updated later after I confirmed some details with the professor  $\square$

### Theorem 3.4

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be twice continuously differentiable and let  $\bar{x}$  be a critical point of  $f$ . Then

- $\bar{x}$  is a global minimizer if  $\nabla^2 f(\mathbf{x}) \succeq 0$  for all  $\mathbf{x} \in \mathbb{R}^n$ .
- $\bar{x}$  is a strict global minimizer if  $\nabla^2 f(\mathbf{x}) \succ 0$  for all  $\mathbf{x} \in \mathbb{R}^n$ .
- $\bar{x}$  is a global maximizer if  $\nabla^2 f(\mathbf{x}) \preceq 0$  for all  $\mathbf{x} \in \mathbb{R}^n$ .
- $\bar{x}$  is a strict global maximizer if  $\nabla^2 f(\mathbf{x}) \prec 0$  for all  $\mathbf{x} \in \mathbb{R}^n$ .



**Definition 3.4 – Critical/Stationary Points**

A point  $\bar{x} \in U$  is a critical point of a function  $f : U \rightarrow \mathbb{R}$  if  $\nabla f(\bar{x})$  exists and satisfies  $\nabla f(\bar{x}) = 0$ .

**Problem 3.3 – Algorithm to Find Local Minimizer**

Given  $f : \mathbb{R} \rightarrow \mathbb{R}$  and  $f'(\bar{x}) \neq 0$ , then  $x_{new} = \bar{x} - (\text{step}) * f'(\bar{x})$ .

The idea is that if  $f'(\bar{x}) > 0$ , then we know that the function is increasing at  $\bar{x}$ , so we want to move to the left to obtain the minimum. Similarly, if  $f'(\bar{x}) < 0$ , then we know that the function is decreasing at  $\bar{x}$ , so we want to move to the right to obtain the minimum.

**Problem 3.4**

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $\phi(\epsilon) = f(\bar{x} + \epsilon d)$

using Taylor expansion  $f(\bar{x} + \epsilon d) = f(\bar{x}) + \epsilon \nabla f(\bar{x})^T d + o(\|\epsilon\|)$  **shouldnt be  $\epsilon d$ ? or  $d$  is the unit vector**

let  $d = -\nabla f(\bar{x}) / \|\nabla f(\bar{x})\|$ ,  $f(\bar{x}) - \epsilon \|\nabla f(\bar{x})\|^2 + o(\epsilon) < f(\bar{x})$  (if  $\nabla f(\bar{x}) \neq 0$ )

i.e test nec condition

If  $\nabla f(\bar{x}) \neq 0$ , then  $x_{new} = \bar{x} + \epsilon(-\nabla f(\bar{x}))$  Move to the deepest direction

**Definition 3.5 – Cauchy's method of steepest descent**

<https://www.math.usm.edu/lambers/mat419/lecture10.pdf>  $x_0 \in \mathbb{R}^n$ .

Is  $\nabla f(x_k) \approx 0$ ? If yes Stop

O.W, find a  $\alpha > 0$

$x_{k+1} = x_k - \alpha \nabla f(x_k)$

repeat

**Problem 3.5 – Example of finding global and local minimizers**

Find global and local minimizers of  $f(x, y) = x^3 - 12xy + 8y^3$ .

We first find the gradient and the Hessian:

$$\nabla f(x, y) = \begin{pmatrix} 3x^2 - 12y \\ -12x + 24y^2 \end{pmatrix}$$

$$\nabla^2 f(x, y) = \begin{bmatrix} -6x & -12 \\ -12 & 48y \end{bmatrix}$$

We can find the critical points when we solve for  $\nabla f(x, y) = 0$ . Solving it, we get solutions  $(0, 0)$  or  $(2, 1)$ .

The Hessian at  $(0, 0)$  is

$$\nabla^2 f(0, 0) = \begin{bmatrix} 0 & -12 \\ -12 & 0 \end{bmatrix}$$

The eigenvalues of  $\nabla^2 f(0, 0)$  are  $-12, 12$ . Therefore it is indefinite. So  $(0, 0)$  is a saddle point.

The Hessian at  $(2, 1)$  is

$$\nabla^2 f(2, 1) = \begin{bmatrix} -12 & -12 \\ -12 & 48 \end{bmatrix}$$

Checking all leading principal minors, we see that they are all positive. So  $\nabla^2 f(2, 1)$  is positive definite. So  $(2, 1)$  is a local minimizer.

### 3.3 Lecture 4

#### Definition 3.6 – Principal Submatrices

Let

$$A = \begin{bmatrix} 1 & 1 & 2 & 7 \\ 1 & 1 & 4 & 6 \\ 2 & 4 & 7 & 8 \\ 7 & 6 & 8 & 1 \end{bmatrix}, \quad I = \{1, 3\}, \quad A[I] = \begin{bmatrix} 1 & 2 \\ 2 & 7 \end{bmatrix}$$

Then  $A[I]$  is a **principal submatrix** of  $A$ .

#### Definition 3.7 – Principal Minors

Let  $A \in \mathbb{S}^n$ , where  $\mathbb{S}^n$  is the set of all symmetric  $n \times n$  matrices.

1.  $\det(A[I])$  is called the **principal minor** of  $A$ .
2. If  $I = \{1, \dots, k\}$  then  $\det(A[I])$  is called the **leading principal minor** of  $A$ .

#### Proposition 3.1 – Characterizing Positive Definiteness with Principal Minors

Let  $A \in \mathbb{S}^n$ . Then

1.  $A \succeq 0 \iff \det(A[I]) \geq 0$  for all principal minors  $\det(A[I])$ .
2.  $A \succ 0 \iff \det(A[I]) > 0$  for all principal minors  $\det(A[I])$ .
3.  $A \succ 0 \iff \det(A[I]) > 0$  for all **leading** principal minors  $\det(A[I])$ .

#### Definition 3.8 – Eigenvectors and Eigenvalues

$0 \neq v \in \mathbb{R}^n$  is an **eigenvector** of  $A$  if there exists  $\lambda \in \mathbb{R}$  such that  $Av = \lambda v$ . The number  $\lambda$  is called an **eigenvalue** of  $A$ .

#### Theorem 3.5 – Finding Eigenvectors and Eigenvalues

Let  $A$  be a matrix.

1. Set up the characteristic equation. We find

$$\det(A - \lambda I) = 0$$

2. Solve for  $\lambda$ . These are the eigenvalues.
3. Plug eigenvalues  $\lambda_1, \dots, \lambda_n$  into  $(A - \lambda I)v = 0$  and solve for  $v$ . These are the eigenvectors.

#### Theorem 3.6 – Orthogonal Spectral Decomposition

Let  $A \in \mathbb{S}^n$ . Then  $A$  has an **orthogonal spectral decomposition**

$$A = \sum_i \lambda_i u_i u_i^T = U D U^T$$

where  $U$  is orthogonal with the orthogonal eigenvectors  $u_i$  as columns and  $D$  is a diagonal matrix with

real eigenvalues on the diagonal.

**Corollary 3.2**

Let  $A \in \mathbb{S}^n$ . Then

1.  $A \succeq 0$  (positive semidefinite) iff all eigenvalues of  $A$  are nonnegative.
2.  $A \succ 0$  (positive definite) iff all eigenvalues of  $A$  are positive.

**Proposition 3.2**

Let  $A \in \mathbb{S}^n$ . The following are equivalent (Positive definite):

1.  $A \succ 0$ .
2. All the eigenvalues of  $A$  are in  $\mathbb{R}_{++}^n$ , the interior of the nonnegative orthant.
3.  $A$  has a real symmetric positive definite square root,  $A = SS$ ,  $S \in \mathbb{S}_{++}^n$ .
4.  $A$  has a lower triangular factorization, a Cholesky factorization,  $A = LL^T$  and  $L$  has positive diagonal elements.
5. All principal minors of  $A$  are positive.
6. All leading principal minors of  $A$  are positive.

And the following are equivalent (Positive semidefinite):

1.  $A \succeq 0$ .
2. All the eigenvalues of  $A$  are in  $\mathbb{R}_+^n$ , the nonnegative orthant.
3.  $A$  has a real symmetric square root,  $A = SS$ ,  $S \in \mathbb{S}^n$ .
4.  $A$  has a lower triangular factorization, a Cholesky factorization,  $A = LL^T$ .
5. All principal minors of  $A$  are nonnegative.

**Theorem 3.7**

$A \in \mathbb{S}^n$  is positive definite if and only if it can be factored as  $A = LL^T$ , where  $L$  is a lower triangular matrix with positive diagonal entries.

*Proof of  $\Leftarrow$ .* Let  $A = LL^T$  where  $L$  is a lower triangular matrix with positive diagonal entries. Then

$$\begin{aligned} x^T Ax &= x^T LL^T x \\ &= y^T y && \text{where } y = L^T x \\ &= \|y\|^2 \geq 0 \end{aligned}$$

$\|y\|^2 = 0$  iff  $y = 0$ . Since  $L$  is lower triangular with positive diagonal entries, it is invertible, thus  $y = Lx = 0$  iff  $x = 0$ . Thus  $x^T Ax > 0$  for all  $x \neq 0$ , so  $A$  is positive definite.  $\square$

**Problem 3.6 – Motivation**

When can we guarantee that global minimizers of  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  exist?

For example, the real valued function on  $\mathbb{R}$   $f(x) = e^x$  is bounded below by 0 but has no minimizers. The

minimum value is 0 but is not attained.

**Proposition 3.3 – Weierstrass Extreme Value Theorem**

If  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuous, and if  $D \subset \mathbb{R}^n$  is a compact set (closed and bounded), then  $f$  attains its global maximum and minimum on  $D$ .

**Problem 3.7 – Example of a continuous function that does not attain its minimum**

Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be defined by  $f(x) = e^x$ . Then  $f$  is continuous, but  $D = \mathbb{R}$  is not compact (closed but not bounded). We can see that  $f$  does not have any global minimizer.

**Definition 3.9 – Coercive function**

A continuous function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is **coercive** if for any sequence  $x_i$  with  $\|x_i\| \rightarrow \infty$ , it must be the case that  $f(x_i) \rightarrow +\infty$ . In other words,

$$\lim_{\|x\| \rightarrow \infty} f(x) = +\infty$$

Here are some examples:

1.  $f_1(x) = x^2$  is coercive.
2.  $g(x) = x$  is not coercive (because as  $x \rightarrow -\infty$ ,  $g(x) \rightarrow -\infty \neq \infty$ ).
3.  $h(x) = e^x$  is not coercive.

**Proposition 3.4 – Coercive Functions and Minimizers**

A coercive function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  has a global minimizer.

**Definition 3.10 – Level Sets**

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function and let  $\alpha \in \mathbb{R}$ . An  $\alpha$ -level set of  $f$  is defined by

$$L_\alpha = \{x \in \mathbb{R}^n : f(x) = \alpha\}$$

That is, all points  $x$  such that  $f(x) = \alpha$ .

- When  $n = 2$ , we call this a level curve.
- When  $n = 3$ , we call this a level surface.
- When  $n > 3$ , we call this a level hypersurface.

**Definition 3.11 – Sub-level set**

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function and let  $\alpha \in \mathbb{R}$ . An  $\alpha$ -sublevel set of  $f$  is defined by

$$S_\alpha(f) = \{x \in \mathbb{R}^n : f(x) \leq \alpha\}$$

That is, all points  $x$  below the line  $f(x) = \alpha$ .

**Problem 3.8**

Can we find minimizer for  $q : \mathbb{R}^n \rightarrow \mathbb{R}$ .  $q(x) = \frac{1}{2}x^T Qx + b^T x + \alpha$  ( $Q = Q^T$ )

Let  $p^* = \inf q(x)$

- When is it finite?
- When is it attained?

**Proposition 3.5 – bdd blow Need to confirm**

$q(x)$  is bdd below iff  $Q \succeq 0$  and  $b \in \text{Range}(Q)$  and  $0 = \nabla q(x) = Qx - b$  so  $x^* = Q^{-1}b$  is Orthogonal (why) so always attained

*Proof.* Since  $Q$  is positive semidef, we can use spectral decomposition,  $Q = UDU^T$ ,  $U^T U = I$ .  $x^T Qx = x^T UDU^T x$  which is also a quadratic form, let  $y = U^T x$ ,  $x = Uy$

Thus for  $q(x)$ , we sub  $x = Uy$ , then we get

$$\begin{aligned} q(x) &= \frac{1}{2}(Uy)^T Q(Uy) + b^T Uy + \alpha \\ &= \frac{1}{2}y^T Dy + (U^T b)^T y + \alpha \quad \text{Let } \bar{b} = U^T b \\ &= \sum \lambda_i y_i^2 + \bar{b}_i y_i + \alpha \end{aligned}$$

Which is a separable problem.  $p_i^* = \min \frac{1}{2}\lambda_i y_i^2 + \bar{b}_i y_i$   
 $p_i^*$  is finite iff  $\lambda_i \geq 0$ ,  $\lambda_i = 0 \implies \bar{b}_i = 0$  Why.  $\bar{b}_i = u_i^T b$  which is the  $i$ th eigenvector. □

**Proposition 3.6**

If  $f$  is coercive,  $f$  maps  $\mathbb{R}^n$  to  $\mathbb{R}$  and cts, then there exists  $\bar{x} \in \text{argmin } f(x)$

*Proof.* Omit □

**Theorem 3.8 –  $q(x)$  is coercive iff  $Q$  is positive definite**

*Proof.* Omit □

## 4 Linear Least Squares & Solving Linear Systems

### 4.1 Lecture 5

#### 4.1.1 Linear Least Squares

##### Problem 4.1 – Motivation For Least Squares

Suppose we have a series of observed values from an experiment:

$$\{(t_1, s_1), (t_2, s_2), \dots, (t_m, s_m)\}$$

where  $t_i$  is the time and  $s_i$  is the observed value at time  $t_i$ . We want to find a polynomial function

$$p(t) = x_0 + x_1 t + \dots + x_n t^n$$

that fits the data. So we want to find coefficients  $x_0, \dots, x_n$  such that  $p(t_i) \approx s_i$  for all  $i$ . More formally, we want to minimize the absolute value of the error of each term. The error ( $\ell_1$  norm) is defined as

$$|e_i| = |p(t_i) - s_i|$$

This can be formulated into a  $\ell_1$  norm minimization problem:

$$\min \left\{ \sum_{i=1}^m |p(t_i) - s_i| : (x_0, \dots, x_n) \in \mathbb{R}^{n+1} \right\}$$

This is a non-differentiable optimization problem since we have absolute values which make it not smooth. So we can reformulate it as a linear program:

$$\min \sum_{i=1}^m \lambda_i$$

s.t.

$$s_i - p(t_i) \leq \lambda_i \quad \text{for all } i = 1, \dots, m$$

$$p(t_i) - s_i \leq \lambda_i \quad \text{for all } i = 1, \dots, m$$

This minimization problem is called **compressive sensing**.

##### Definition 4.1 – Vandermonde Matrix

Let

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \dots & t_1^n \\ 1 & t_2 & t_2^2 & \dots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \dots & t_m^n \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad b = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$A$  is called a **Vandermonde matrix**.

##### Theorem 4.1

The Vandermonde Matrix

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad b = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

is full column rank if  $n + 1 \leq m$  and the points  $t_i$  are distinct.

#### Definition 4.2 – $\ell_1$ and $\ell_2$ Norm

The  $\ell_1$  norm of a vector  $x$  is defined to be

$$\|x\| = \sum |x_i|$$

The  $\ell_2$  norm of a vector  $x$  is defined to be

$$\|x\| = \sqrt{\sum x_i^2}$$

#### Problem 4.2 – Linear Least Squares Problem

Recall our  $\ell_1$  norm minimization problem:

$$\min \left\{ \sum_{i=1}^m |p(t_i) - s_i| : (x_0, \dots, x_n) \in \mathbb{R}^{n+1} \right\}$$

We can instead use  $\ell_2$  norm defined as  $\|e\|_2 = \sqrt{\sum e_i^2}$ . So our  $\ell_2$  minimization problem is

$$\min \left\{ \sum_{i=1}^m (p(t_i) - s_i)^2 : (x_0, \dots, x_n) \in \mathbb{R}^{n+1} \right\}$$

where  $p(t) = x_0 + x_1 t + \cdots + x_n t^n$ . Using the Vandermonde matrix, we can rewrite our problem to be

$$\min \frac{1}{2} \|Ax - b\|^2$$

where

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad b = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The objective function is  $g(x) = \frac{1}{2} \|Ax - b\|^2$ . Let's first expand  $g(x)$ :

$$\begin{aligned} g(x) &= \frac{1}{2} \|Ax - b\|^2 \\ &= \frac{1}{2} (Ax - b)^T (Ax - b) \\ &= \frac{1}{2} (Ax)^T Ax - (Ax)^T b + \frac{1}{2} \|b\|^2 \\ &= \frac{1}{2} x^T A^T Ax - x^T A^T b + \frac{1}{2} \|b\|^2 \end{aligned}$$

Then, applying the gradient on both sides, we have

$$\nabla g(x) = A^T Ax - A^T b$$

To find the critical points, we solve for  $\nabla g(x) = 0$ . So the critical points are  $x^*$  that satisfy the equation

$$A^T Ax = A^T b$$

This is also called a **normal equation**.

also something about the condition number, i dont really understand.

#### Theorem 4.2 – Matrix Calculus

$$\nabla(Ax) = A^T$$

$$\nabla(x^T A) = A$$

$$\nabla(x^T x) = 2x$$

$$\nabla(x^T Ax) = (A + A^T)x$$

#### Definition 4.3 – Frechet Derivative

Let  $h : U \rightarrow W$ , where  $U$  is an open subset of  $V$ , and  $V, W$  are finite dimensional vector spaces. The function  $h$  is **Frechet differentiable** at  $x \in U$  if there exists a linear transformation  $A : V \rightarrow W$  such that

$$\lim_{d \rightarrow 0} \frac{\|h(x+d) - h(x) - Ad\|}{\|d\|} = 0$$

idk

#### 4.1.2 Best Linear Least Squares

Suppose  $\bar{x}$  is a solution to the normal equation

$$A^T Ax = A^T b$$

When  $A$  is not full column rank, the solution  $\bar{x}$  is not unique. Therefore, the solution set to  $A^T Ax = A^T b$  is

$$S = \{\bar{x} + y : y \in \text{null}(A)\}$$

Then, an algorithm might choose  $x \in S$  with very large norm. This is not the best solution.

#### Definition 4.4 – Best Linear Least Square

The best linear least square is the least squares solution of minimal norm. It is the solution to the problem

$$\min_x \{\|x\| : x \in \text{argmin}\|Ax - b\|\}$$

## 4.2 Lecture 6

Goal: Solving normal equation/non linear case

#### Definition 4.5 – Singular Values of a Matrix

The singular values of a matrix  $A$  are the square roots of the eigenvalues of the matrix  $A^T A$ . They are always non-negative real numbers.

The number of non-zero singular values of a matrix equals the rank of that matrix.



**Definition 4.6 – Condition Number of a Matrix**

Suppose  $A \in \mathbb{R}^{m \times n}$ ,  $m > n$  is full column rank. The condition number of the matrix  $A$ ,  $\text{cond}(A)$ , is the ratio of the largest to smallest nonzero singular values of  $A$ . Let  $\sigma_{\max}$  be the largest singular value and  $\sigma_{\min}$  be the smallest singular value. Then

$$\text{cond}(A) = \frac{\sigma_{\max}}{\sigma_{\min}}$$

**Definition 4.7 – SVD Decomposition**

Let  $A$  be an  $m \times n$  matrix. Then  $A$  can be factored into

$$\underbrace{A}_{m \times n} = \underbrace{U}_{m \times m} \underbrace{\Sigma}_{m \times n} \underbrace{V^T}_{n \times n}$$

where

- $U$  is an  $m \times m$  orthogonal matrix consisting of eigenvectors of  $AA^T$
- $V^T$  is the transpose of an  $n \times n$  matrix containing the eigenvectors of  $A^T A$
- $\Sigma$  is a diagonal matrix with  $r = \text{rank}(A)$  positive eigenvalues of  $AA^T$  (Singular values of  $A$ ) on the diagonal.

there is a section on piazza posted lecture notes that shows why using SVD decomposition to solve normal equation is a bad idea. Not sure if i should include here

**Definition 4.8 – Orthogonal Matrix**

A matrix  $Q$  is orthogonal if  $Q^T Q = I$ .

**Definition 4.9 – Orthonormal Columns**

A matrix  $Q$  has orthonormal columns if each column vector is a unit vector (norm is 1), and any two distinct columns are orthogonal (inner product is 0).

**Problem 4.3****Applicable Model for "best" data fitting**

$$(s_i, p_i), \quad i = 1, \dots, m$$

We got linear model and we need to solve  $A^T Ax = A^T b$  normal equation

Today: Solve normal equation (whenever case)

i.e., this is from optimality condition for

$$\begin{aligned} \min \frac{1}{2} \|Ax - b\|_2^2 &= f(x) \\ \text{s.t. } A &\in M_{m,n} \\ x &\in \mathbb{R}^n \end{aligned}$$

$$\text{stationary: } \nabla f(\bar{x}) = A^T(A\bar{x} - b) = 0$$

An alternate view is  $\min \|y - b\|$

, s.t.  $y \in \text{Range}(A) \Rightarrow$  a projection problem

i.e., project  $b$  onto  $\text{Range of } A$

$$\text{Aside: use SVD of } A = U\Sigma V^T : \Sigma = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \\ & 0 & \end{bmatrix}, \quad r = \text{rank}(A) \sigma_i > 0$$

SVD can be used for a "rank revealing" decomposition.

$$U = [U_1 \quad U_2]$$

Where:

$U_1$  :  $m \times r$  orthogonal columns span  $\mathcal{R}(A)$

$U_2$  :  $m \times (m-r)$  spanning  $\mathcal{N}(A^T)$

Properties:

$$U^T U = I \quad \text{and} \quad I = U_1^T U_1 = (U_1^T U_1)^T = \mathcal{C}(U_1 U_1^T) = U_1 U_1^T$$

This means  $U_1 U_1^T$  is symmetric idempotent, i.e., orthogonal projection.

To solve the projection minimizing  $\|y - b\|$  where  $y \in \text{Range}(\mathcal{R}(A))$ :

$$\text{optimal } y^* = U_1 U_1^T b$$

Defining the Moore-Penrose generalized inverse:

$$A^+ = V \Sigma^+ U^T$$

where

$$\Sigma^+ = \begin{bmatrix} \frac{1}{\sigma_1} & \cdots & 0 \\ 0 & \cdots & 0 \end{bmatrix}$$

and

$$\gamma_i = \frac{b_i}{\sigma_i} \quad \text{if } \sigma_i \neq 0$$

Finally, to find  $x$  such that  $Ax = y^* = U_1 U_1^T b = AA^+ b$ : If  $A$  has full column rank (i.e.,  $\mathcal{R}(A^T) = \mathbb{R}^n$ ), then  $x$  is unique. Otherwise  $\bar{x} + v, v \in \mathcal{N}(A)$  is a solution for all  $v$  in  $\mathcal{N}(A)$  Choosing a large solution. This solution of min is called "Best least square solution"

we can write it as bilevel problem

The solution of minimum norm is called the "best least square solution". This can be represented as a "bilevel problem":

$$\min_{\hat{x}} \|\hat{x}\|_2 \quad \text{for values when different}$$

subject to:

$$x \in \arg \min_z \frac{1}{2} \|Ax - b\|_2^2$$

The bLSS (best least square solution) is the unique LSS (least square solution) in  $\mathcal{R}(A^T)$ .

#### Definition 4.10 – Condition number

Given:

$$A = U\Sigma V^T \quad (\text{SVD of } A)$$

Where  $\Sigma$  is a diagonal matrix with:

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_r \end{pmatrix} \quad \text{and} \quad \sigma_r \geq 0$$

The condition number of matrix  $A$  is given by:

$$\text{Cond}(A) = \frac{\sigma_1}{\sigma_r}$$

If  $A$  is not of full rank, then:

$$\text{Cond}(A) = \infty \implies \text{Cond}(A^T) = \text{Cond}(A)$$

We also have:

$$A^T A = V\Sigma^T U^T U \Sigma V^T = V\Sigma^2 V^T$$

With:

$$\Sigma^2 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_r^2 \end{pmatrix}$$

Thus:

$$\text{Cond}(A^T A) = \frac{\sigma_1^2}{\sigma_r^2} = \text{Cond}(A)^2$$

For the solution  $B$ :

$$Bx^*(\text{True sol}) = b, B\hat{x}(\text{Numerical}) = b + \delta b$$

Where:

$$\frac{\|x^* - \hat{x}\|}{\|x^*\|} \leq \text{Cond}(B) \frac{\delta \|b\|}{\|b\|}$$

To avoid solving the normal equation with squares condition number, we should use QR

#### Definition 4.11 – QR Factorization

For any  $m \times n$  matrix  $A$ , there exists an  $m \times m$  orthogonal matrix  $Q$  ( $QQ^T = I$ ) and an  $m \times n$  upper triangular matrix  $R$  ( $R_{i,j} = 0, \forall i < j$ ) satisfying  $A = QR$ . Moreover, if the columns of  $A$  are linearly independent then we can get

$$\begin{aligned} A &= QR \\ &= Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \\ &= [Q_1 \quad Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \\ &= Q_1 R_1 \end{aligned}$$

where

- $R_1$  is an invertible  $n \times n$  upper triangular matrix
- $0$  is an  $(m - n) \times n$  zero matrix
- $Q_1$  is an  $m \times n$  matrix with orthonormal columns
- $Q_2$  is an  $m \times (m - n)$  matrix with orthonormal columns

**Theorem 4.3 – QR Factorization on Normal Equation**

Assuming that the columns of  $A$  are linearly independent, then the normal equation  $A^T A x = A^T b$  can be solved by applying QR factorization to  $A$ :

$$\begin{aligned}
 (A^T A)x &= A^T b \\
 ((Q_1 R_1)^T Q_1 R_1)x &= (Q_1 R_1)^T b \\
 (R_1^T Q_1^T Q_1 R_1)x &= R_1^T Q_1^T b \\
 R_1^T R_1 x &= R_1^T Q_1^T b && \text{Since } Q_1 \text{ is orthogonal} \\
 R_1 x &= Q_1^T b && \text{Since } R_1 \text{ is invertible}
 \end{aligned}$$

**Definition 4.12 – Methods of Solving General Linear Systems**

Suppose we are given a linear system  $Bx = b$ , and we know that this system has a solution, i.e.  $b \in \text{range}(B)$ . There are 3 important algorithms/factorizations used to find  $x$ :

- Gaussian Elimination (LU factorization) ( $PB = LU$ )
- QR factorization
- SVD, singular value decomposition

**Problem 4.4 – Solving Large Positive Definite Systems**

Suppose we have a linear system,  $Ax = b$ , with  $A$  positive definite. If  $x^*$  is a solution, then  $Ax^* - b = 0$ . Then this is equivalent to minimizing the function

$$f(x) = \frac{1}{2} \|Ax - b\|^2, \nabla f(x) = Ax - b = 0$$

Dont understand this and the part after as well. You will have to add more notes here. Link to notes [HERE](#)

**Definition 4.13 – Steepest Descent Direction in First order model**

The steepest descent direction in the first order model is  $d = -\frac{\nabla f(x_c)}{\|\nabla f(x_c)\|}$

**Theorem 4.4 – Conjugate Gradient Method**

The first search direction is the negative gradient,

$$v_0 = -\nabla q(x_0)$$

with  $q = f$ . At the  $k$ th iteration:

$$v_{k+1} = -\nabla q(x_k) + \beta_k v_k$$

where  $\beta_k$  is chosen to ensure  $\langle Av_{k+1}, v_k \rangle = 0$ . This guarantees that the directions are  $A$ -conjugate **wtf is**

A conjugate  $u, v$  conjugate, if  $u^T Av = 0$ . We then set

$$x_{k+1} = x_k + \alpha_{k+1} v_{k+1}$$

where  $\alpha_{k+1}$  is chosen from an exact line search (what is line search).

#### Problem 4.5

$$\min_x f(x) \\ x \in D \subset \mathbb{R}$$

- 1<sup>st</sup> & 2<sup>nd</sup> order optimality condition both necessary and sufficient
- 1<sup>st</sup>, 2<sup>nd</sup> order model

### 4.3 Lecture 7

#### Definition 4.14 – Nonlinear Least Square

Suppose we have  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  where

$$F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix}$$

Then the nonlinear least squares problem is

$$\min \{h(x)\}$$

where

$$h(x) = \frac{1}{2} \|F(x)\|^2 = \frac{1}{2} \langle F(x), F(x) \rangle = \frac{1}{2} \sum_{i=1}^m f_i^2(x)$$

#### Definition 4.15 – Jacobian Matrix

Let  $F$  be defined as above, then the Jacobian matrix is  $J(x) = F'(x)$  where

$$F'(x) = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix}$$

#### Problem 4.6 – Solving Nonlinear Least Squares

For the nonlinear least squares problem defined above, we consider the special case  $m = n$ . Then we can consider the problem as solving the square system of nonlinear equations  $F(x) = 0$ . Recall that for current approximation  $x_c$ ,

$$0 = F(x_c + d) \approx F(x_c) + \underbrace{F'(x_c)}_{\text{Jacobian}} \underbrace{d}_{\text{search direction}}$$

So we solve

$$F'(x_c)d = -F(x_c)$$

which is called the Newton equation. Then we can take a step in the search/Newton direction  $d$  to get a new approximation  $x_{c+1} = x_c + \alpha d$  for appropriate step length  $\alpha$ .  
 In general,  $\nabla h(x_c) = \sum_{i=1}^m f_i(x_c) \nabla f_i(x_c) = J(x_c)^T F(x_c)$

**Definition 4.16 – Second Order Model for Nonlinear Problem**

$h(x + \nabla x) \sim h(x)$  [upload later](#)

**Definition 4.17 – Argmin and argmax**

argmin  $f(x)$  is the set of all minimizers of  $f(x)$ , similarly argmax  $f(x)$  is the set of all maximizers of  $f(x)$ .

**Theorem 4.5 – Lagrange Multipliers for Equality Constraints**

Suppose that  $f : C \rightarrow \mathbb{R}$  and  $h : C \rightarrow \mathbb{R}^m$  are sufficiently smooth functions on the open set  $C \subseteq \mathbb{R}^n$ . Let  $x \in C$  be a local minimum of  $f$  subject to the constraints  $h(x) = 0, x \in C$ . In addition, assume the following regularity condition at  $x$  (constraint qualification at  $x$ )

$$h'(x) \text{ is onto}$$

Then there exists  $\lambda \in \mathbb{R}^m$  (a Lagrange multiplier vector) such that

$$0 = \nabla f(x) + \langle \lambda, h'(x) \rangle$$

Equivalently, the gradient  $\nabla f(x)$  is in the range of  $h'(x)^T$ .

**Problem 4.7 – DNE of multiplier**

Recall:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } f_i(x) = 0, \quad i = 1, \dots, m \\ &L(x, \lambda) = f(x) - \sum_{i=1}^m \lambda_i f_i(x) \quad (\text{Lagrangian multiplier}) \end{aligned}$$

If  $x^*$  is a local minimum and

$$\{\nabla f_i(x^*)\} \text{ is a linearly independent set, then } \nabla L(x^*, \lambda) = \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla f_i(x^*) = 0 \text{ for some } \lambda \in \mathbb{R}$$

**Example:**

$$\begin{aligned} f(x) &= x \\ h(x) &= x^2 = 0 \\ L(x, \lambda) &= x + \lambda x^2, \quad \text{with constraint } x^2 = 0 \\ 0 &= \nabla L(x, \lambda) = 1 + 2\lambda x = 1 + 0. \text{ No } \lambda \text{ exist} \end{aligned}$$

## 5 Iterative Methods for Unconstrained Optimization

### 5.1 Lecture 8-10

Before, we have iterative methods to solve linear and linear least squares system.

**Goal:** Solve more general problems of minimization.

An iterative algorithm is a procedure that produces an (infinite) sequence of points  $\{x_k\}$ , in  $\mathbb{R}^n$  such that our sequence converges to a critical point of  $f$  or to a point that satisfies the second order necessary conditions of optimality.

### 5.1.1 Line Search Strategy

Most Line Search Algorithms have the following procedure:

Choose a starting point  $x_0$ . For  $k = 0, 1, 2, \dots$

1. Choose a search direction  $d_k$ .
2. Choose a step length  $\alpha_k > 0$  (that satisfies  $f(x_k + \alpha_k d_k) < f(x_k)$ )
3. Update  $x_{k+1} = x_k + \alpha_k d_k$
4. Stop if a stopping criterion is satisfied.

#### Definition 5.1 – Line Search Strategy

At each iteration  $k$ , we choose a vector (search direction)  $d_k \neq 0 \in \mathbb{R}^n$ , then choose  $\alpha_k > 0$  (step length) that approximately solves

$$\alpha_k \in \operatorname{argmin} f(x_k + \alpha d_k)$$

Then we update  $x_{k+1} = x_k + \alpha_k d_k$ .

#### 5.1.1.1 Finding Descent Direction

#### Definition 5.2 – Descent Direction

Let  $x \in U \subseteq \mathbb{R}^n$  with  $U$  being an open set. Then  $d \in \mathbb{R}^n$  is a descent direction for  $f$  at  $x$  if there exists  $\bar{\alpha} > 0$  such that

$$x + \alpha d \in U, f(x + \alpha d) < f(x)$$

for all  $0 < \alpha < \bar{\alpha}$ .

So  $f(x_{k+1}) < f(x_k)$ .

#### Lemma 5.1

Let  $f$  be sufficiently smooth on an open set  $U$  and let  $\bar{x} \in U$ . Let  $d \in \mathbb{R}^n$  satisfy

$$\langle d, \nabla f(\bar{x}) \rangle = \nabla f(\bar{x})^T d < 0$$

Then  $d$  is a descent direction for  $f$  at  $\bar{x}$ .

(i.e., the directional derivative of  $f$  at  $\bar{x}$  in the direction of  $d$  is negative, so we know that it is the descent direction.)

*Proof.* From the hypothesis we have

$$f(x + td) = f(x) + t \nabla f(x) \cdot d + o(t).$$

Therefore we deduce

$$\frac{f(x + td) - f(x)}{t} = \langle d, \nabla f(x) \rangle + \frac{o(t)}{t} < 0, \text{ for all sufficiently small } t > 0.$$

Moreover, since  $U$  is open, we have  $x + td \in U$  for all sufficiently small  $t > 0$ . □

**Problem 5.1 – Example(direction of steepest descent)**

We saw

$$\min \nabla f(x)^T \cdot d \quad \text{s.t.} \quad \|d\| = 1$$

get

$$d = -\frac{\nabla f(x)^T}{\|\nabla f(x)\|}$$

**Cauchy direction of steepest descent**

Alternatively,

$$-\nabla f(x)^T \cdot d \leq \|d\| \cdot \|\nabla f(x)^T\| = \|d\| \cdot \|\nabla f(x)\|$$

but

$$d = -\nabla f(x)^T \text{ gives equality.}$$

**So C.S yields s.t direction**

If  $B \succ 0$ , then

$$d = -B^{-1}\nabla f(x)^T \text{ in a "deflated gradient" direction.}$$

**Proposition 5.1 – Existence of B**

Suppose  $\nabla f(\bar{x})^T d < 0$  Then there exists  $B \succ 0$  that satisfies  $d = -B^{-1}\nabla f(\bar{x})$

*Proof.* Ignore it for now text

□

**5.1.2 Second order model: Newton Type**

Model  $m_q(x) = f(x_c) + \nabla f(x_c)^T \Delta x + \frac{1}{2} \Delta x^T B_c \Delta x$  where  $B_c \simeq \nabla^2 f(x_c)$

**Lemma 5.2 – TFAE**

1.  $m_q(\Delta x)$  is bdd below
2.  $B_c \succeq 0$  and  $\nabla f(x_c) \in \text{range}(B_c)$
3.  $\Delta \bar{x} = -B_c^{-1} \nabla f(x_c)$  is a global minimum of  $m_q$

*Proof.* 2 to 1:

Suppose 2 holds, we have  $B_c \succeq 0$  and  $\exists \bar{\Delta x}, B_c \bar{\Delta x} = -\nabla f(x_c)$ . This is also the stationary condition. Again, since  $B_c$  is psd, there exists a psd square root such that  $ss = B_c$ .

Then let's complete the square

$$m_q(\Delta x) =$$

confirm

□

The main difference between line search methods is the choice of the descent direction

**Problem 5.2 – Example**

If  $\nabla f(x) \neq 0$  then  $d = -\nabla f(x)$  is a descent direction since

$$d^T \nabla f(x) = -\nabla f(x)^T \nabla f(x) = -\|\nabla f(x)\|_2^2 < 0$$

Another option is letting  $d = -H^{-1}\nabla f(x)$  where  $H \succ 0$  and  $\nabla f(x) \neq 0$  since

$$d^T \nabla f(x) = -\nabla f(x)^T H^{-1} \nabla f(x) < 0$$

since  $H \succ 0$ , we have  $H^{-1} \succ 0$ , so  $\nabla f(x)^T H^{-1} \nabla f(x) > 0$ .



**Theorem 5.1 – Existence of a Descent Direction**

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be continuously differentiable and let  $\bar{x} \in \mathbb{R}^n$ . Assume that  $d$  is a descent direction of  $f$  at  $\bar{x}$ , so

$$\nabla f(\bar{x})^T d < 0$$

Then there exists  $\delta > 0$  such that for every  $\alpha \in (0, \delta]$ ,

$$f(\bar{x} + \alpha d) < f(\bar{x})$$

*Proof.* By Taylor's theorem, for all  $x \in \mathbb{R}^n$ ,

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + o(\|x - \bar{x}\|)$$

Then

$$f(\bar{x} + \alpha d) = f(\bar{x}) + \nabla f(\bar{x})^T (\bar{x} + \alpha d - \bar{x}) + o(\|\bar{x} + \alpha d - \bar{x}\|)$$

Simplifying gives

$$f(\bar{x} + \alpha d) = f(\bar{x}) + \alpha \underbrace{\nabla f(\bar{x})^T d}_{<0} + o(\|\alpha d\|)$$

Rearranging and dividing both sides by  $\alpha\|d\|$ , we get

$$\frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha\|d\|} = \frac{\nabla f(\bar{x})^T d}{\|d\|} + \frac{o(\|\alpha d\|)}{\alpha\|d\|}$$

Taking the limit as  $\alpha \rightarrow 0$ , we get

$$\lim_{\alpha \rightarrow 0} \frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha\|d\|} = \frac{\nabla f(\bar{x})^T d}{\|d\|}$$

Since  $\nabla f(\bar{x})^T d < 0$ , we know that the limit is negative. Then there exists  $\delta > 0$  such that for all  $\alpha \in (0, \delta]$ , we have

$$\frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha\|d\|} < 0$$

□

**5.1.2.1 Finding Step Size**

The process of finding the step size  $\alpha_k$  is called line search. Some common choices for step size are as follows:

1.  $\alpha_k = \alpha$  (a constant) for every  $k$
2. Exact line search:  $\alpha_k = \operatorname{argmin}_{\alpha \geq 0} f(x_k + \alpha d_k)$
3. Inexact line search: A step size  $\alpha_k$  that achieves a sufficient decrease in  $f$ . One popular inexact line search conditions are called the Wolfe conditions.

**Problem 5.3 – Exact Line Search**

For exact line search, we define a new function

$$\phi(\alpha) = f(x_k + \alpha d_k)$$

We can differentiate w.r.t.  $\alpha$  using the chain rule to obtain

$$\phi'(\alpha) = \nabla f(x_k + \alpha d_k)^T d_k$$

Then we set  $\phi'(\alpha) = 0$  to obtain the critical point of  $\phi$ . Then we can use the second derivative test to determine if the critical point is a local minimizer. If it is, then we have found the optimal step size  $\alpha_k$ .

**Definition 5.3 – Wolfe Conditions**

Let  $0 < c_1 < c_2 < 1$ . The Wolfe conditions are

1. Sufficient decrease condition:

$$f(x_k + \alpha d_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T d_k$$

or equivalently

$$\phi(\alpha_k) \leq \phi(0) + c_1 \alpha \phi'(0)$$

2. Curvature condition:

$$\nabla f(x_k + \alpha d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k$$

or equivalently

$$\phi'(\alpha_k) \geq c_2 \phi'(0)$$

**Theorem 5.2 – Existence of Step Lengths that Satisfy Wolfe Conditions**

Assume  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable. Let  $d_k$  be a descent direction at  $x_k$  and assume  $f$  is bounded below along the ray  $\{x_k + \alpha d_k : \alpha \geq 0\}$ . Then if  $0 < c_1 < c_2 < 1$  there exist intervals of step lengths satisfying the Wolfe conditions.

*Proof.* Omitted □

**5.1.2.2 Steepest Descent Method**

The steepest descent method is easy to implement as it requires the calculation of the gradient but not second derivatives. The descent direction at each iteration is chosen as

$$d_k = -\nabla f(x_k)$$

**Lemma 5.3**

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be continuously differentiable and let  $\bar{x} \in \mathbb{R}^n$  such that  $\nabla f(\bar{x}) \neq 0$ . Then the optimal solution of the following minimization problem

$$\min_{d \in \mathbb{R}^n} \{ \nabla f(\bar{x})^T d : \|d\|_2 = 1 \}$$

is

$$d^* = \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$$

*Proof.* By Cauchy-Schwarz inequality, we have

$$\nabla f(\bar{x})^T d \geq -\|\nabla f(\bar{x})\|_2 \|d\|_2 = -\|\nabla f(\bar{x})\|_2$$

The smallest value of  $d$  is  $d = \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$  since if we plug in  $d$  into the inequality, we get

$$\nabla f(\bar{x})^T \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2} = \frac{-\|\nabla f(\bar{x})\|_2^2}{\|\nabla f(\bar{x})\|_2} = -\|\nabla f(\bar{x})\|_2$$

So  $d^* = \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$  is a minimizer of this problem. □

**Theorem 5.3**

Let  $\{x_k\}$  be the sequence generated by the method of steepest descent method where the step sizes are chosen by the exact line search. Then for every  $k \geq 0, k \in \mathbb{N}$

$$(x_{k+2} - x_{k+1})^T (x_{k+1} - x_k) = 0$$

**5.1.2.3 Backtracking Line Search****Definition 5.4 – Backtrack Inexact Line Search**

We make use of the Wolfe Conditions.

1. Initialize  $\alpha > 0$ , for example, choose  $\alpha = 1, c \in (0, 1)$
2. if  $(\phi(\alpha) > \phi(0) + c\alpha\phi'(0))$ 
  - (a) while  $(\phi(\alpha) > \phi(0) + c\alpha\phi'(0))$ 
    - i.  $\alpha = \alpha/2$
3. else if  $(\phi(2\alpha) \leq \phi(0) + 2c\alpha\phi'(0))$ 
  - (a) while  $(\phi(2\alpha) \leq \phi(0) + 2c\alpha\phi'(0))$ 
    - i.  $\alpha = 2\alpha$
4. Output =  $\alpha$ .

**5.1.2.4 Newton's Method**

In Newton's Method, the search direction is given by

$$d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

The Newton step is defined only when  $\nabla^2 f(x_k)$  is positive definite.

**Definition 5.5 – Newton's Method**

Input: Initial point:  $x_0$ , Tolerance:  $\epsilon > 0$

1. Solve the linear system of equations  $\nabla^2 f(x_k)d_k = -\nabla f(x_k)$  for  $d_k$ , so  $d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$
2. Update  $x_{k+1} = x_k + \alpha_k d_k$  where  $\alpha_k = 1$  for Newton's method
3. If  $\|\nabla f(x_{k+1})\|_2 \leq \epsilon$ , stop. Otherwise, go to step 1.

**Definition 5.6 – Operator Norm**

The operator norm of a matrix  $Q \in \mathbb{R}^{m \times n}$  is defined by

$$\|Q\|_2 = \max\{\|Qx\|_2 : x \in \mathbb{R}^n, \|x\|_2 = 1\}$$

This definition satisfies the properties of the norm. For every  $A \in \mathbb{R}^{m \times n}$ ,

- $\|A\|_2 = 0 \iff A = 0$
- $\|\alpha A\|_2 = |\alpha| \|A\|_2$  for every  $\alpha \in \mathbb{R}$
- $\|A + B\|_2 \leq \|A\|_2 + \|B\|_2$

**Theorem 5.4**

For every  $A \in \mathbb{R}^{m \times n}$ ,  $x \in \mathbb{R}^n$ , we have

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2$$

**5.1.2.5 Quasi-Newton Methods**

Since computing the Hessian matrix is expensive, we can use an approximation of the Hessian matrix instead. We now set the search direction as

$$d_k = -B_k^{-1} \nabla f(x_k)$$

where the symmetric and positive definite matrix  $B_k$  is updated at every iteration by a quasi-Newton updating formula.

There are many quasi-Newton methods, but we will only discuss the BFGS method. Consider the quadratic model of the objective function:

$$m_k(d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d$$

where  $B_k$  is a positive definite matrix and will be updated at every iteration. The minimizer of this quadratic model is

$$d_k = -B_k^{-1} \nabla f(x_k)$$

It is used as the search direction and the new iterate is given by

$$x_{k+1} = x_k + \alpha_k d_k$$

where  $\alpha_k$  is the step length, chosen so that it satisfies the Wolfe conditions. In this method, we match the gradient of  $m_{k+1}$  to the gradient of  $f$  at  $x_k$  and  $x_{k+1}$ . That is, we require

$$\nabla m_{k+1}(0) = \nabla f(x_{k+1})$$

$$\nabla m_{k+1}(-\alpha_k d_k) = \nabla f(x_k)$$

From the second equation above, we get

$$\nabla_{k+1}(-\alpha_k d_k) = \nabla f(x_{k+1}) - \alpha_k B_{k+1} d_k = \nabla f(x_k)$$

Which implies

$$\nabla(x_{k+1}) - \nabla f(x_k) = B_{k+1}(x_{k+1} - x_k)$$

This is the secant equation. We can use this equation to update  $B_{k+1}$ .

**5.2 Lecture 11****5.2.1 Convergence of Line Search Methods****Definition 5.7 – Key Quantity**

The key quantity is the angle between the search direction and the gradient. It is

$$\cos \theta_k = \frac{\langle \nabla f(x_k), v_k \rangle}{\|\nabla f(x_k)\| \|v_k\|}$$

We will use this together with Wolfe conditions to prove convergence.

**Theorem 5.5 – Convergence of Line Search Methods**

Let  $x_0 \in \mathbb{R}^n \rightarrow \mathbb{R}$  be  $C^1$ -smooth, and  $f$  be bounded below. Let  $0 < c_1 < c_2 < 1$  and suppose that there exists a Lipschitz constant  $L > 0$  such that

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

on an open set containing the level set  $\{z : f(z) \leq f(x_0)\}$ .

Let  $\{x_k\}$  be generated using descent directions  $v_k$  and step lengths that satisfy the Wolfe conditions. Then we have

$$\sum_{k=1}^{\infty} \cos^2 \theta_k \|\nabla f(x_k)\|^2 < \infty$$

*Proof.* In course notes. □

**Corollary 5.1**

If there exists  $\delta > 0$  satisfying  $\cos^2 \theta_k \geq \delta > 0$  for all  $k$ , then

$$\lim_{k \rightarrow \infty} \nabla f(x_k) = 0$$

**5.2.2 Convergence Rate of Steepest Descent**

For exact line search, we have

$$x_{k+1} = x_k - \left( \frac{\|\nabla f(x_k)\|^2}{\langle A \nabla f(x_k), \nabla f(x_k) \rangle} \right) \nabla f(x_k)$$

If we plot out our iterations, we can see that there is a zig-zagging behaviour and the directions  $x_{k+1} - x_k$  and  $x_k - x_{k-1}$  are orthogonal.

**Theorem 5.6 – Zig Zagging of Steepest Descent**

For the general steepest descent method with an exact step-size, we have

$$(x_{k+1} - x_k)^T (x_k - x_{k-1}) = 0$$

So they are orthogonal.

**Theorem 5.7 – Rate of Convergence of Steepest Descent**

Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $C^2$ -smooth and the iterates  $\{x_k\}$  generated by steepest descent with exact line search converge to  $x^*$ , that is,  $\lim_{k \rightarrow \infty} x_k = x^*$ . Suppose that  $\nabla^2 f(x^*)$  is positive definite. Then for all large  $k$ , we have

$$f(x_{k+1}) - f(x^*) \leq K(f(x_k) - f(x^*))$$

with

$$0 \leq K = \left( \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^2 < 1$$

where  $\lambda_1$  and  $\lambda_n$  are the largest and smallest eigenvalues of  $\nabla^2 f(x^*)$ .

**5.2.3 Convergence Rate of Newton's Method****Definition 5.8 – Comparing Algorithms**

upload later

### Definition 5.9 – $p$ -th Order Convergence

If the sequence of residuals/errors  $r_k \rightarrow 0$  and

$$\lim_{k \rightarrow \infty} \frac{\|r_{k+1}\|}{\|r_k\|^p} \rightarrow r > 0$$

then the order of convergence is called  $p$ -th order.

If  $p = 1, r < 1$ , then this is called linear convergence. (For s.d, linear,  $r \cong r(\text{cond}(\text{Hessian at } x^*))$ )

If  $p = 2$ , then this is called quadratic convergence.

If

$$\lim_{k \rightarrow \infty} \frac{\|r_{k+1}\|}{\|r_k\|} \rightarrow 0$$

then this is called superlinear convergence.

### Theorem 5.8 – Convergence of Newton's Method

Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $C^2$ -smooth,  $\nabla^2 f(x^*)$  is positive definite,  $\nabla f(x^*) = 0$  (also minor condition: Hessian is locally Lipschitz continuous, which in particular holds if  $f$  is  $C^3$ -smooth). Consider the iterates

$$\begin{aligned} x_{k+1} &= x_k + t_k v_k \\ &= x_k + t_k (\nabla^2 f(x_k))^{-1} (-\nabla f(x_k)) \end{aligned}$$

where  $t_k$  is the step size chosen to satisfy the Wolfe conditions (with  $c_1 < \frac{1}{2}$ ). Then if the starting point  $x_0$  is sufficiently close to  $x^*$ , then we have:

1.  $t_k = 1$  satisfies the Wolfe conditions (pure/free Newton step).
2.  $x_k$  converges to  $x^*$ .
3. If we choose  $t_k = 1$ , then we have quadratic convergence locally of the iterates, that is for some  $r \geq 0$  and some  $x_0$  close enough to the strict local minimum  $x^*$ , we have

$$\|x_{k+1} - x^*\| \leq r \|x_k - x^*\|^2$$

$$\|\nabla f(x_{k+1})\| \leq r \|\nabla f(x_k)\|^2$$

### Problem 5.4

$$d_{SC} = -\nabla f(x_k), d_N = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

levenberg marquardt algorithm  $-(\nabla^2 f(x_k) + \lambda I)^{-1} \nabla f(x_k) \lambda > 0$

### Problem 5.5

This leads to trust region method, follows by 5.8.1 Trust Region Methods Outline. note: Find  $d_k$  is trust region subproblem

### 5.2.4 Trust Region Strategy

### Problem 5.6 – Trust Region Strategy

Construct a model for  $f$  at  $x_k$ ,  $f(x_k + \Delta x) \sim m_q(\Delta x) = f(x_k) + \Delta x^T \nabla f(x_k) + \frac{1}{2} \Delta x^T B_k \Delta x$ , where

$B_k \simeq \nabla^2 f(x_k)$ . Construct region where we trust the model. Region can be a ball, eg.  $\Omega_k = \{\Delta x \mid \|\Delta x\| \leq \delta_k^2\}$

The difference in trust region strategy and line search strategy is that in line search strategy, we first choose the direction, then choose step size. In trust region, we first choose the step size, then we choose the direction.

#### Definition 5.10 – Trust Region Strategy

In each iteration, we construct a model of  $f$ . That is, in each step we consider  $m_k : \mathbb{R}^n \rightarrow \mathbb{R}$  that is a simple function that approximates  $f$  well on some simple set  $\Omega_k$  (the trust region) around our current approximation  $x_k$ . Then we find the new approximation

$$\hat{x} = \operatorname{argmin}_{x \in \Omega_k} m_k(x)$$

A common model is the quadratic model

$$m_k(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle x - x_k, B_k(x - x_k) \rangle$$

where  $B_k \approx \nabla^2 f(x_k)$  approximates the Hessian. If the values  $f(\hat{x})$  and  $m_k(\hat{x})$  are close, then we declare  $x_{k+1} = \hat{x}$ . Otherwise, we shrink the size of the trust region  $\Omega_k$  and repeat the process.

Usually  $\Omega_k$  is a ball, ellipsoid, or a box around  $x_k$ .

The main points are how to choose a model function  $m_k$ , and how to choose a trust region  $\Omega_k$ .

#### Theorem 5.9 – Trust Region Methods Outline

Outline:

1. Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , an iterate  $x_k$ , a local model function

$$m_k(d) = f(x_k) + \langle \nabla f(x_k), d \rangle + \frac{1}{2} \langle B_k d, d \rangle$$

and a trust region radius  $\nabla_k > 0$

2. Find  $d_k \in \operatorname{argmin}\{m_k(d) : \|d\| \leq \nabla_k\}$
3. Use the relative decrease (actual/predicted decrease)

$$\rho_k = \frac{f(x_k) - f(x_k + d_k)}{m_k(0) - m_k(d_k)}$$

to gauge the strength of the model in order to obtain the new trust region radius  $\Delta_{k+1}$  and to decide whether to accept the new point as  $x_{k+1} = x_k + d_k$  or use  $x_{k+1} = x_k$ .

## 6 Convex set and function

### 6.1 Lecture 11-13

#### Definition 6.1 – Line Segment

Let  $x, y \in \mathbb{R}^n$ . The line segment joining  $x$  and  $y$  is the set

$$[x, y] = \{\lambda x + (1 - \lambda)y : \lambda \in [0, 1]\}$$

#### Definition 6.2 – Convex set

Set  $S$  in  $\mathbb{R}^n$  is a convex set if for every  $x, y \in S$  and for every  $0 \leq \lambda \leq 1$

$$\lambda x + (1 - \lambda)y \in S$$

That is, the line segment  $[x, y]$  is contained in  $S$ .

### Problem 6.1 – Example of Convex Sets

The following are convex sets:

1. The empty set  $\emptyset$ , any single point  $\{x_0\}$ , and  $\mathbb{R}^n$
2. Lines in  $\mathbb{R}^n$ . For example, a line through the point  $x_0$  in the direction  $d$

$$L = \{x_0 + \alpha d : \alpha \in \mathbb{R}\}$$

3. A halfspace in  $\mathbb{R}^n$ :  $H = \{x : a^T x \leq c\}$
4. An open halfspace in  $\mathbb{R}^n$ :  $H^o = \{x : a^T x < c\}$
5. A polyhedron in  $\mathbb{R}^n$ :  $P = \{x : Ax \leq b\}$
6. An open and closed ball in  $\mathbb{R}^n$
7. The set of  $n \times n$  positive semi definite and positive definite matrices:

$$S_+^n = \{A \in \mathbb{R}^{n \times n} : A \succeq 0\}, \quad S_{++}^n = \{A \in \mathbb{R}^{n \times n} : A \succ 0\}$$

### Definition 6.3 – Convex combination

Let  $S$  be a convex set and let  $x_1, \dots, x_n \in S$ . If  $\lambda_1, \lambda_2, \dots, \lambda_n$  are nonnegative numbers such that

$$\sum_{i=1}^n \lambda_i = 1$$

then

$$\sum_{i=1}^n \lambda_i x_i$$

is called a convex combination and is contained in  $S$ .

### Lemma 6.1

$S$  is a convex set iff  $S$  contains all its convex combination

### Definition 6.4

Given  $S \in \mathbb{R}^n$ , the convex hull of  $S$ ,  $\text{conv } S$  is the smallest convex set containing  $S$  and is equal to the set of all convex combinations of points in  $S$ .

### Lemma 6.2

Suppose  $S$  is a compact set, then  $\text{conv } S$  is a compact set.



**Definition 6.5 – Affine Set**

Let  $S \subseteq \mathbb{R}^n$  be a subspace, and  $a \in \mathbb{R}^n$ . Then  $L = a + S$  is an affine set (i.e. it is a translation of a subspace).

Moreover,  $\dim L = \dim S$ . In addition, let  $T \subseteq \mathbb{R}^n$ . Then the affine hull,  $\text{aff } T$ , is the smallest affine set containing  $T$ . If  $T$  is a convex set, then  $\dim(T) = \dim(\text{aff } T)$ .

**Definition 6.6 – Convex Cone**

$K \subseteq \mathbb{R}^n$  is a convex cone if the Minkowski sums satisfy

1.  $\lambda K \subseteq K$  for all  $\lambda \geq 0$
2.  $K + K \subseteq K$

**Definition 6.7 – Hyperplane**

For  $a \in \mathbb{R}^n, b \in \mathbb{R}$ , their hyperplane is defined as

$$H = \{x \in \mathbb{R}^n : a^T x = b\}$$

**Definition 6.8 – Halfspace**

For  $a \in \mathbb{R}^n, b \in \mathbb{R}$ , their halfspace is defined as

$$H = \{x \in \mathbb{R}^n : a^T x \leq b\}$$

**Definition 6.9 – Polytope**

A polytope (polyhedron set) is a finite intersection of halfspaces.

**Definition 6.10 – Face**

Let  $C$  be a convex set. The convex set  $F \subseteq C$  is a face of  $C$ , denoted as  $F \trianglelefteq C$  if

$$x, y \in C, \quad z \in (x, y) \cap F \implies x, y \in F$$

That is, for any two points  $x, y \in C$ , if a point  $z$  lies in the open line segment between  $x$  and  $y$  (denoted  $(x, y)$ ) also lies in  $F$ , then  $x, y \in F$ .

If this is true, then  $F$  is a face of  $C$ .

**Theorem 6.1**

If  $C$  is a convex cone, then the faces of  $C$  are convex cones.

*Proof.* Exercise. □

**Definition 6.11 – Non-negative Polar (Dual) Cone**

The non-negative polar (dual) cone of a set  $C$  is

$$C^+ = C^* = \{\phi : \langle \phi, c \rangle \geq 0, \forall c \in C\}$$

**Definition 6.12 – Conjugate Face**

Let  $F$  be a face of the convex cone  $K$  denoted  $F \trianglelefteq K$ . The conjugate face of  $F$  is

$$F^* = F^\perp \cap K^*$$

**Definition 6.13 – Exposed Face**

Let  $F$  be a face of the convex cone  $K$ ,  $F \trianglelefteq K$ . Then  $F$  is an exposed face if there exists  $\phi \in K^*$  such that

$$F = K \cap \phi^\perp$$

where  $\phi^\perp = \{\phi\}^\perp$ .

**Definition 6.14 – Extreme Point**

Let  $C \subseteq \mathbb{R}^n$ . The point  $x \in C$  is an extreme point of  $C$  if

$$x \in [y, z], \quad y, z \in C \implies x = y = z$$

That is, there is no nontrivial convex combination with points in  $C$ .

**Definition 6.15 – Convex Functions and Strictly Convex Functions**

Let  $f : C \rightarrow \mathbb{R}$  where  $C \subseteq \mathbb{R}^n$  is a convex set. We say that  $f$  is a convex function if for every  $x, y \in C$  and for every  $\lambda \in [0, 1]$ ,

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

We say that  $f$  is strictly convex if for every  $x, y \in C$ ,  $x \neq y$  and for every  $\lambda \in [0, 1]$ ,

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

**Theorem 6.2 – First Order Condition**

Assume that  $f(x)$  is continuously differentiable on a convex set  $C \subseteq \mathbb{R}^n$ . Then the function  $f(x)$  is

1. convex if and only if for all  $x, y \in C$ ,

$$f(x) + \nabla f(x)^T(y - x) \leq f(y)$$

2. strictly convex if and only if for all  $x, y \in C$  with  $x \neq y$ ,

$$f(x) + \nabla f(x)^T(y - x) < f(y)$$

*Proof.* See past course notes

□

**Corollary 6.1**

Let  $f(x)$  be a continuously differentiable convex function on an open convex set  $C \subseteq \mathbb{R}^n$ . If  $x^* \in C$  is a critical point of  $f$ , then  $x^*$  is a global minimizer of  $f$  in  $C$ .

**Theorem 6.3 – Second Order Condition**

Assume that  $f(x)$  is twice continuously differentiable on an open convex set  $C \subseteq \mathbb{R}^n$ . Then the function  $f(x)$  is convex if and only if  $\nabla^2 f(x)$  is positive semidefinite for every  $x \in C$ .

---

*Proof.* See past course notes

□

**Theorem 6.4**

Let  $C \subseteq \mathbb{R}^n$  be a convex set. Then any local minimizer (in  $C$ ) of a convex function  $f : C \rightarrow \mathbb{R}$  is also a global minimizer of  $f$  on  $C$ . Furthermore, any local minimizer of a strictly convex function  $f : C \rightarrow \mathbb{R}$  is the unique global minimizer of  $f$  on  $C$ .