

CO 367

272284444

September 2023

Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Lecture 1-Preliminaries . . . . .	2
1.2	Lecture 2 . . . . .	5
<b>2</b>	<b>Unconstrained Optimization</b>	<b>6</b>
2.1	Lecture 2 . . . . .	6
2.2	Lecture 3 . . . . .	8
2.3	Lecture 4 . . . . .	9
<b>3</b>	<b>Linear Least Squares &amp; Solving Linear Systems</b>	<b>12</b>
3.1	Lecture 5 . . . . .	12
3.2	Lecture 6 . . . . .	14
3.3	Lecture 7 . . . . .	16

# 1 Introduction

## 1.1 Lecture 1-Preliminaries

### Definition 1.1 – 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 1.1 – 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 1.2 – Classification of Quadratic Forms

Let  $Q$  be a quadratic form of a matrix  $A$ . Then  $Q$  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.
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.
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.
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.
5. indefinite if  $Q(x)$  can be positive or negative. Or there are positive and negative eigenvalues for  $A$ .

### Definition 1.3 – 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 1.4 – 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 1.5 – 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 1.6 – Chain Rule

WIP

### Definition 1.7 – 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  $\cos(\theta) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$ .

With corresponding norm to be the Euclidean Norm

### Definition 1.8 – Open ball

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

### Definition 1.9 – map

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

### Definition 1.10 – open set

Let  $D \subset \mathbb{R}^n$ ,  $D$  open set.  $\forall x \in D, \exists \delta > 0$ , s.t  $B_\delta(x) \subset D$

**Definition 1.11 – 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 1.12 – 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  $f, g_i, h_i$  are all **affine** function and  $D = \mathbb{R}^2$ , then we have an LP

**Definition 1.13 – affine**

$$f(x) = Ax + b \tag{1}$$

where  $b \neq 0$

**Definition 1.14 – 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 1.15 – 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 1.16 – 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  $x \in I$ .

### Definition 1.17 – 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

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

## 1.2 Lecture 2

### Definition 1.18 – 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 1.2 – 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 1.19 – Contour

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

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

### Definition 1.20 – 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 1.21 – 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 1.3 – 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

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

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.

## 2 Unconstrained Optimization

### 2.1 Lecture 2

#### Problem 2.1 – 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

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

#### Theorem 2.1 – Taylor's Theorem on the real line

Let  $f : (a, b) \rightarrow \mathbb{R}$ , and  $\bar{x}, x \in (a, b)$ , then there exists  $z$  strictly between  $x, \bar{x}$  such that

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

or equivalently

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

### Lemma 2.1 – Directional Derivative

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

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

Then the **directional derivative**, denoted  $f'(x; d)$  of  $f$  at  $x$  at the direction  $d$  is

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

### Problem 2.2

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

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

### Corollary 2.1

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

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

### Definition 2.1 – Hessian

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

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

### Theorem 2.2 – 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 there exists a point  $z \in [\bar{x}, x]$  such that

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

**Lemma 2.2**

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

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

**2.2 Lecture 3****Definition 2.2 – Matrix Norm**

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

**Definition 2.3**

Define  $f, D, \bar{x}$ ,  $D$  is an open set Then:

1. Nec: If  $\bar{x}$  is a local minimum for  $f$  on  $D$ , then  $\nabla f(\bar{x}) = 0$  and  $\nabla^2 f(\bar{x}) \succeq 0$  is positive semidefinite.
2. Suff: If  $\nabla f(\bar{x}) = 0$ ,  $\nabla^2 f(\bar{x}) \succ 0$  is positive definite then  $\bar{x}$  is a strict local minimum of  $f$  on  $D$ .

*Proof.* 1. updated later after I confirmed some details with the professor □

**Definition 2.4 – 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 2.3 – 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 2.4**

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 2.5 – Cauchy's method of steepest descent**

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

$$\|\nabla f(x_k)\| \approx 0? \text{ IF yes Stop}$$

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

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

repeat



**Problem 2.5 – 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.

**2.3 Lecture 4****Definition 2.6 – 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 2.7 – 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 2.1 – 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 **leading** principal minors  $\det(A[I])$ .

**Definition 2.8 – 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 2.3 – 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 2.4 – 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 2.2**

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 2.2**

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.

**Problem 2.6 – 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 2.3 – Weierstrass Extreme Value Theorem**

If  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuous, and if  $D \subset \mathbb{R}^n$  is a closed and bounded set, then  $f$  is bounded below and the minimum value is attained on  $D$ .

**Definition 2.9 – 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 2.4 – Coercive Functions and Minimizers**

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

**Definition 2.10 – 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 2.11 – 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  $f(x) = \alpha$ .

### 3 Linear Least Squares & Solving Linear Systems

#### 3.1 Lecture 5

##### Problem 3.1 – 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.

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

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

##### Definition 3.1 – 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 3.1

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 3.2 – $\ell_1$ 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 3.2 – 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, using the definition of linear transformation definition of the gradient (**WTF is this**), 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.

### Definition 3.3 – 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 3.4 – 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 3.5 – 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

## 3.2 Lecture 6

Goal: Solving normal equation/non linear case

### Definition 3.6 – 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 3.7 – Orthogonal Matrix**

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

**Definition 3.8 – 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).

**Definition 3.9 – QR Factorization**

For any  $m \times n$  matrix  $A$ , there exists an  $m \times m$  orthogonal matrix  $Q$  ( $Q Q^T = I$ ) 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 3.2 – 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  $A$ :

$$\begin{aligned} (A^T A)x &= A^T b \\ ((Q_1 R_1)^T Q_1 R_1)x &= (Q_1 R_1)^T b \\ (R_1^T Q_1^T Q_1 R_1)x &= R_1^T Q_1^T b \\ R_1^T R_1 x &= R_1^T Q_1^T b \\ R_1 x &= Q_1^T b \end{aligned}$$

Since  $Q_1$  is orthogonal

Since  $R_1$  is invertible

**Definition 3.10 – 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 3.3 – 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](#)

**Theorem 3.3 – 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**. 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**).

**3.3 Lecture 7****Definition 3.11 – 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 3.12 – Jacobian Matrix**

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

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

**Problem 3.4 – 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  $\alpha$ .