Emory University
MATH 347 Non Linear Optimization
Learning Notes

Jiuru Lyu

February 8, 2024

Contents

1 Math Preliminaries

1.1 Introduction to Optimization

Definition 1.1.1 (Optimization Problem). The main optimization problem can be stated as follows

$$\min_{x \in S} f(x), \tag{1}$$

where

- x is the *optimization variable*,
- S is the *feasible set*, and
- f is the *objective function*.

Remark 1.1 $\max_{x \in S} f(x) = -\min_{x \in S} -f(x)$. *Hence, we will only study minimization problems.*

> **Theorem 1.1.2 Solving an Optimization Problem**
>
> - Theoretical Analysis: analytic solution
> - Numerical solution/optimization

Definition 1.1.3 (Solution Methods depend on the type of x, S, and f).

- When x is continuous (e.g., $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{m \times n}$, $\ldots$), then the optimization problem stated in Eq. (1) is a *continuous optimization problem*. *It will also be the focus of this class.*

  Opposite to continuous optimization problems, we have *discrete optimization problem* if x is discrete.

  If x has both types of components, then we call the problem *mixed*.

- Depending on S, we can have
  - *Unconstrained problems*: where $S = \mathbb{R}^n$, $S = \mathbb{R}^{m \times n}$, $\ldots$ $(m, n$ are fixed).
  - *Constrained problems*: where $S \subsetneq \mathbb{R}^n$, $S \subsetneq \mathbb{R}^{m \times n}$, $\ldots$.

    Both types of problems will be studied.

- Depending on f, we have
  - *Smooth optimization problems*: f has first and/or second order derivatives.

    Only smooth optimization problems will be studied.
  - *Non-smooth optimization problems*: f is not differentiable.

Definition 1.1.4 (Linear Optimization/Program). If f is linear and S consists of linear constrains, then the optimization problem is called a *linear problem/program.*

Example 1.1.5 Classification of Optimization Problems

1. Consider the following problem

$$\min_{x_1,x_2,x_3} x_1^2 - 4x_1x_2 + 3x_2x_3 + \sin x_3$$

Solution 1.

- Optimization variable: $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. $\longrightarrow$ continuous.
- Feasible set: $S = \mathbb{R}^3$. $\longrightarrow$ unconstrained.
- Objective function: $f(x_1, x_2, x_3) = x_1^2 - 4x_1x_2 + 3x_2x_3 + \sin x_3$. $\longrightarrow$ smooth but non-linear.

$\square$

2. Consider the following problem

$$\max_{\substack{4x_1+7x_2+3x_3\leq 1\\ x_1,x_2,x_3\geq 0}} x_1 + 2x_2 + 3x_3$$

Solution 2.

- Optimization variable: $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. $\longrightarrow$ continuous.
- Feasible set: $S = \{(x_1, x_2, x_3) : x_1, x_2, x_3 \geq 0, 4x_1 + 7x_2 + 3x_3 \leq 1\} \subsetneq \mathbb{R}^3$. $\longrightarrow$ constrained.
- Objective function: $f(x_1, x_2, x_3) = x_1 + 2x_2 + 3x_3$. $\longrightarrow$ smooth and linear.

$\square$

Remark 1.2 *This problem can be considered as the budget constrained optimization problem in Economics.*

3. Consider the following problem

$$\min_{x_1,x_2\geq 0} 4x_1 - 3|x_2| + \sin\left(x_1^2 - 2x_2\right)$$

Solution 3.

- Optimization variable: $x = (x_1, x_2) \in \mathbb{R}^2$. $\longrightarrow$ continuous.

- Feasible set: $S = \{(x_1, x_2) : x_1, x_2 \geq 0\} \subsetneq \mathbb{R}^2$. $\longrightarrow$ constrained.
- Objective function: $f(x_1, x_2) = 4x_1 - 3|x_2| + \sin(x_1^2 - 2x_2)$. $\longrightarrow$ non-smooth and non-linear.

$\square$

Remark 1.3 *In this particular problem,* $x_2 \geq 0$*, and so* $f(x_1, x_2) = 4x_1 - 3x_2 + \sin(x_1^2 - 2x_2)$ *on the feasible set. Hence, this problem can be equivalently written as*

$$\min_{x_1, x_2 \geq 0} 4x_1 - 3x_2 + \sin\left(x_1^2 - 2x_2\right),$$

which is a smooth optimization problem.

1.2 Linear Algebra Review

Example 1.2.1 Why linear algebra for optimization?

Consider $\min_{x \in \mathbb{R}} f(x)$, where $f(x) = c + bx + ax^2,\ a, b, c \in \mathbb{R}$.

- $a > 0$: $x^* = -\dfrac{b}{2a}$ is a global minimum and $f(x^*) = c - \dfrac{b^2}{4a}$.
- $a < 0$: no minimum exists.
- $a = 0$: $f(x) = c + bx$.
  - $b \neq 0$: no minimum exists.
  - $b = 0$: $f(x) = c$, and every x is a minimum point.

We can approximate any smoothing function using Taylor's approximation and make them simple into the case discussed above.

Theorem 1.2.2 Taylor's Approximation

$$f(x) = \underbrace{f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2}_{q(x)} + \underbrace{\varepsilon(x - x_0)(x - x_0)^2}_{\text{error}},$$

where $\lim_{x \to x_0} \varepsilon(x - x_0)$.

Remark 1.4 *The hope is that the quadratic approximation will inform us on the behavior of* f *near* x_0 *and be useful for instance in referring* x_0 *on the subject of optimality.*

Definition 1.2.3 (Quadratic Approximation in Higher Dimensions). When $d > 1$, we consider $\min_{x \in \mathbb{R}^d} f(x)$. Then, the *quadratic approximation* of f is defined as

$$q(x) \coloneqq c + \langle b, x \rangle + \langle x, Ax \rangle,$$

where $c \in \mathbb{R}$, $b \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$.

Remark 1.5 *Then, to know if a minimum exists, we need information on the matrix A and the vector b.*

Definition 1.2.4 (Vector, $\mathbb{R}^d$). We define a *vector* in $\mathbb{R}^d$ as a column vector.

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d,\ x_i \in \mathbb{R}.$$

On $\mathbb{R}^d$, we also have the following operations defined

- Addition:
$$\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_d + y_d \end{pmatrix},\ x_i, y_i \in \mathbb{R}$$

- Scalar multiplication:
$$\alpha \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_d \end{pmatrix}, \alpha, x_i \in \mathbb{R}$$

Definition 1.2.5 (Basis of $\mathbb{R}^d$). A collection of vectors $v_1 \ldots, v_d \in \mathbb{R}^d$ is a *basis* in $\mathbb{R}^d$ if $\forall\, x \in \mathbb{R}^d$, $\exists!\ \alpha_1, \ldots, \alpha_d \in \mathbb{R}$ *s.t.* $x = \alpha_1 v_1 + \cdots + \alpha_d v_d$.

Example 1.2.6 The Standard Basis

The *standard basis* is defines as

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix},$$

where 1 is at the i-th position for $1 \leq i \leq d$. Note that $\forall x \in \mathbb{R}^d,\ x = x_1 e_1 + \cdots + x_d e_d$.

Notation 1.7.

$$0_d = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Definition 1.2.8 (Inner Product). $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is an *inner product* if

- (symmetry) $\langle x, y \rangle = \langle y, x \rangle \quad \forall x, y \in \mathbb{R}^d$
- (additivity) $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle \quad \forall x, y, z \in \mathbb{R}^d$
- (homogeneity) $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle \quad \forall x, y \in \mathbb{R}^d, \ \lambda \in \mathbb{R}$
- (positive definiteness) $\langle x, x \rangle \geq \quad \forall x \in \mathbb{R}^d$ and $\langle x, x \rangle = 0 \iff x = 0$

Example 1.2.9 Examples of Inner Products

1. **Definition 1.2.10 (Dot Product).** The *dot product* of $x, y \in \mathbb{R}^d$ is defined as

$$\langle x, y \rangle = x_1 y_1 + \cdots + x_d y_d = \sum_{i=1}^{d} x_i y_i \quad \forall x, y \in \mathbb{R}^d.$$

   It is also referred as the *standard inner product*, and we often use the notation $x \cdot y$ to denote it.

2. **Definition 1.2.11 (Weighted Dot Product).** The *weighted dot product* of $x, y \in \mathbb{R}^d$ with some weight w is defined as

$$\langle x, y \rangle_w = \sum_{i=1}^{d} w_i x_i y_i,$$

   where $w_1, \ldots, w_d > 0$ are called *weights*.

Remark 1.6 *When $d = 2$, then $\langle x, y \rangle = |x||y| \cos \angle(x, y)$. Dot product measure how correlated are two vectors (with respect to their directions).*

Definition 1.2.12 (Vector Norm). $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ is a *norm* if

- (non-negativity) $\|x\| \geq 0 \quad \forall x \in \mathbb{R}^d$ and $\|x\| = 0 \iff x = 0$
- (positive homogeneity) $\|\lambda x\| = |\lambda| \|x\| \quad \forall \lambda \in \mathbb{R}, \ x \in \mathbb{R}^d$
- (triangular inequality) $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^d$.

Remark 1.7 *Vector norm introduces the notion of length of vectors in $\mathbb{R}^d$.*

Example 1.2.13 Examples of Vector Norms

- If $\langle \cdot, \cdot \rangle$ is an inner product on $\mathbb{R}^d$, then

$$\|x\| = \sqrt{\langle x, x \rangle} \quad \forall x \in \mathbb{R}^d$$

is a norm. For instance,

$$\|x\|_2 = \sqrt{x \cdot x} = \left(\sum_{i=1}^{d} x_i^2 \right)^{\frac{1}{2}}.$$

This norm is called the *standard (Euclidean)* or ℓ_2 norm in $\mathbb{R}^d$.

- **Definition 1.2.14 (ℓ_p Norms).** Suppose $p \geq 1$, then

$$\|x\|_p := \left(\sum_{i=1}^{d} x_i^p \right)^{\frac{1}{p}}.$$

- **Definition 1.2.15 (∞-Norms).**

$$\|x\|_\infty := \max_{1 \leq i \leq d} |x_i| \quad \forall x \in \mathbb{R}^d.$$

Remark 1.8 $\lim_{p \to \infty} \|x\|_p = \|x\|_\infty$.

Theorem 1.2.16 Cauchy-Schwarz Inequality
Assume that $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is an inner product, then

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle \quad \forall x, y \in \mathbb{R}^d.$$

In particular, if $\|x\| = \sqrt{\langle x, x \rangle}$, then

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\| \quad \forall x, y \in \mathbb{R}^d.$$

For the standard inner product, we have

$$\left| \sum_{i=1}^{n} x_i y_i \right| \leq \|x\|_2 \cdot \|y\|_2 \quad \forall x, y \in \mathbb{R}^d.$$

The equality holds when x and y are linearly dependent.

Definition 1.2.17 (Matrix). Let $d, m \in \mathbb{N}$. We say that $A \in \mathbb{R}^{d\times m}$ is a $d \times m$ *matrix* if

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d1} & a_{d2} & \cdots & a_{dm} \end{pmatrix} = \Big(a_{ij}\Big)_{i=1,j=1}^{d,m}$$

Definition 1.2.18 (Operations with Matrices).

- Let $A, B \in \mathbb{R}^{d\times m}$, then $\Big(A + B\Big)_{i,j} = a_{ij} + b_{ij} \quad \forall i, j$.
- Let $A \in \mathbb{R}^{d\times m}$ and $\alpha \in \mathbb{R}$, then $\Big(\alpha A\Big)_{ij} = \alpha a_{ij} \quad \forall i, j$.
- Let $A \in \mathbb{R}^{d\times m}$ and $B \in \mathbb{R}^{m,n}$, then $AB \in \mathbb{R}^{d\times n}$, and $\Big(AB\Big)_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj} \quad \forall i, j$.

Remark 1.9 *Matrix multiplication is not commutative. In fact, if $A \in \mathbb{R}^{d\times m}$ and $B \in \mathbb{R}^{m\times n}$, then BA is defined if and only if $n = d$. In that case, $AB \in \mathbb{R}^{d\times d}$ and $BA \in \mathbb{R}^{m\times m}$, and so if $m \neq d$, AB and BA have different sizes. Finally, even if $m = d = n$, $AB \neq BA$ in general.*

Definition 1.2.19 (Linear Transformation). The mapping $\mathcal{L} : \mathbb{R}^m \to \mathbb{R}^d$ is called *linear* if $\mathcal{L}(\alpha x_1 + \beta x_2) = \alpha\mathcal{L}(x_1) + \beta\mathcal{L}(x_2)$.

Theorem 1.2.20 Matrices and Linear Transformation
$\forall A \in \mathbb{R}^{d\times m}$, $\mathcal{L}_A(x) = Ax$ is a linear mapping from $\mathbb{R}^m$ to $\mathbb{R}^d$. Moreover, $\forall\mathcal{L} : \mathbb{R}^m \to \mathbb{R}^d$ linear, $\exists! A \in \mathbb{R}^{d\times m}$ *s.t.* $\mathcal{L} = \mathcal{L}_A$.

Proof 1. Here, we offer an intuition on why this is true. Suppose $A \in \mathbb{R}^{d\times m}$ and $x \in \mathbb{R}^m$ *s.t.*

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dm} \end{pmatrix} \quad \text{and} \quad x \in \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \in \mathbb{R}^{m\times 1}.$$

Then, $Ax \in \mathbb{R}^{d\times 1}$ is the following

$$Ax = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1m}x_m \\ \vdots \\ a_{d1}x_1 + \cdots + a_{dm}x_m \end{pmatrix} \in \mathbb{R}^{d\times 1}.$$

So, if $\mathcal{L}_A(x) = Ax$ for $x \in \mathbb{R}^m$, then $\mathcal{L}_A : \mathbb{R}^m \to \mathbb{R}^d$ is linear. ■

Theorem 1.2.21 Matrix Multiplication as Composite Linear Transformations
Suppose $\mathcal{L}_A : \mathbb{R}^m \to \mathbb{R}^d$ and $\mathcal{L}_B : \mathbb{R}^n \to \mathbb{R}^m$, where $A \in \mathbb{R}^{d\times m}$ and $B \in \mathbb{R}^{m\times n}$. Define $\mathcal{L}(x) = \mathcal{L}_A \circ \mathcal{L}_B(x) = \mathcal{L}_A(\mathcal{L}_B(x)) \quad \forall x \in \mathbb{R}^n$. Then, $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^d$. Since $\mathcal{L}_A$ and $\mathcal{L}_B$ are linear, we found that $\mathcal{L}$ is also linear. Hence, $\mathcal{L} = \mathcal{L}_C$ *f.s.* $C \in \mathbb{R}^{d\times n}$. It turns out that $C = AB$.

Definition 1.2.22 (Transpose of Matrix). Let $A \in \mathbb{R}^{d\times m}$, then its transpose $A^T \in \mathbb{R}^{m\times d}$, and

$$\left(A^T\right)_{ij} = a_{ji}.$$

Corollary 1.2.23 : If $x, y \in \mathbb{R}^d$, then $\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i = x^T y = xy^T$.

Proof 2. Suppose $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$, then $x^T = \begin{pmatrix} x_1 & \cdots & x_d \end{pmatrix}$.

$$x^T y = \begin{pmatrix} x_1 & \cdots & x_d \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = x_1 y_1 + \cdots + x_d y_d.$$

■

Corollary 1.2.24 Cauchy-Schwarz: $\left|x^T y\right| \le \|x\|_2 \|y\|_2$.

Definition 1.2.25 (Trace of a Matrix). Assume that $A \in \mathbb{R}^{d\times d}$, the *trace* of A, denoted as $\mathrm{Tr}(A)$, is defined as

$$\mathrm{Tr}(A) = \sum_{i=1}^{d} a_{ii}.$$

Definition 1.2.26 (Determinant of a Matrix). Assume that $A \in \mathbb{R}^{d\times d}$, the *determinant* of A, denoted as $\det(A)$, is defined as

$$\det(A) = \sum_{\sigma\in S_d} (-1)^{i(\sigma)} a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{d\sigma(d)},$$

where S_d is the set of all possible permutation of size d and $i(\sigma)$ denotes the sign of the permutation.

Definition 1.2.27 (Eigenvalue and Eigenvector). Assume that $A \in \mathbb{R}^{d\times d}$. We say that λ is an *eigenvalue* for A if $\exists x \in \mathbb{R}^d \backslash \{0\}$ *s.t.* $Ax = \lambda x$. In this case, x is called an *eigenvector*.

Definition 1.2.28 (Diagonalizability). A matrix $A \in \mathbb{R}^{d\times d}$ is called *diagonalizable* if $\exists$ basis $v_1, \ldots, v_d$ ***s.t.*** $Av_i = \lambda v_i \quad \forall 1 \le i \le d$.

Theorem 1.2.29 Diagonalization, Singular Value Decomposition (SVD) of Squared Matrices
Assume that A is diagonalizable and

$$V = \begin{pmatrix} v_1 & v_2 & \cdots & v_d \end{pmatrix}.$$

Then, $A = VDV^{-1}$, where D is a diagonal matrix such that

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}.$$

Example 1.2.30 Application of Diagonalization

$$A^2 = \left(VDV^{-1}\right)\left(VDV^{-1}\right) = VD\underbrace{V^{-1}V}_{I}DV^{-1} = VD^2V^{-1}.$$

Generally,

$$A^n = VD^nV^{-1} = V\begin{pmatrix} \lambda_1^n & & 0 \\ & \ddots & \\ 0 & & v_d^n \end{pmatrix}V^{-1}.$$

Remark 1.10 *Remarks on Diagonalization*

- *There might be repeating eigenvalues. Typically, we enumerate λ's* s.t. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.
- *In general, it is hard to decide whether A is diagonalizable.* For example, rotation matrices have no eigenvectors nor eigenvalues.
- *If A is symmetric; that is $A = A^T$, then A is diagonalizable. Moreover, we can choose basis* $v_1, \ldots, v_d$ s.t.

  $$v_i^T v_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}.$$

  Such bases are called orthonormal. *In matrix form, if $V = \begin{pmatrix} v_1 & v_2 & \cdots & v_d \end{pmatrix}$, then*

  $$V^TV = \begin{pmatrix} v_1^T \\ \vdots \\ v_d^T \end{pmatrix}\begin{pmatrix} v_1 & \cdots & v_d \end{pmatrix} = I.$$

  That is, $V^T = V^{-1}$, and hence $A = VDV^{-1} = VDV^T$.

1.3 Basic Topology

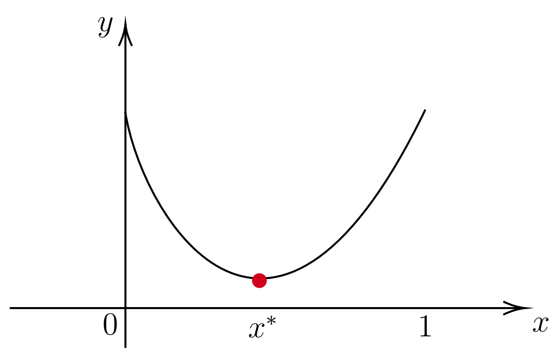
Example 1.3.1 Introduction

Consider the optimization problem $\min_{x \in [0,1]} f(x)$. Suppose that $x^* \in [0,1]$ is a solution for this problem, then we have

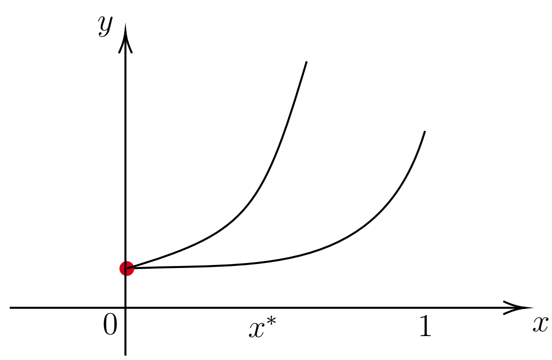
$$f(x) \geq f(x^*) \quad \forall x \in [0,1].$$

Then, we can conduct a case study on the necessary condition we need to have on $f'(x)$.

1. $x^* \in (0,1) \implies f'(x^*) = 0$.



2. $x^* = 0 \implies f'(x^*) \geq 0$



3. $x^* = 1 \implies f'(x^*) \leq 0$.

Definition 1.3.2 (Open/Closed Ball). The *open ball* with center $c \in \mathbb{R}^n$ and radius $r > 0$ is the set

$$B(c,r) := \{x \in \mathbb{R}^n : \|x - c\| < r\}.$$

The *closed ball* with center $c \in \mathbb{R}^n$ and radius $r > 0$ is the set

$$B[c,r] := \{x \in \mathbb{R}^n : \|x - c\| \leq r\}.$$

Remark 1.11 *The boundary is not included in an open ball.*

Definition 1.3.3 (Interior Point). Assume that $U \in \mathbb{R}^n$. We say that $x \in U$ is an *interior point* if $\exists\, r > 0$ *s.t.* $B(x, r) \subseteq U$. The set of all interior points of U is denoted by $\text{int}(U)$

Example 1.3.4 Interior Point Example

Suppose $U = [0, 1]$. Prove that $\text{int}(U) = (0, 1)$.

Proof 1. To prove this, we have to show $\text{int}(U) \subseteq (0, 1)$ and $(0, 1) \subseteq \text{int}(U)$.

($\supseteq$): Let $x \in (0, 1)$. *WTS:* $x \in \text{int}(U)$. Take $r = \min\{x, 1 - x\}$, then the open ball $B(x, r) \subseteq U$. *proof omitted.* So, $x \in \text{int}(U)$, and thus $(0, 1) \subseteq \text{int}(U)$. □

($\subseteq$): Let $x \in \text{int}(U)$. *WTS:* $x \in (0, 1)$. *omitted.* ■

Definition 1.3.5 (Open Set). A set $U \subseteq \mathbb{R}^n$ is called *open* if $\text{int}(U) = U$.

Example 1.3.6 Open Set Counterexample

$U = [0, 1]$ in Example 1.3.4 is not an open set.

Remark 1.12 *When f is defined over an open set U, then we can define differentiability on f on U.*

Definition 1.3.7 (Closed Set). A set $F \subseteq \mathbb{R}^n$ is called a *closed set* if $\forall\ (x_n)_{n=1}^{\infty} \subseteq F$ such that $\lim\limits_{n\to\infty} x_n = x \implies x \in F$.

Example 1.3.8 Closed Set

- Take $F = \mathbb{R}^n$, then F is a closed set because we have taken everything into the set.
- $F = [0, 1]$ is closed.

  Proof 2. Take $x_1, x_2, \ldots, x_n \cdots \in [0, 1]$. That is, $0 \leq x_n \leq 1, \quad \forall n \geq 1$. Then, set $x = \lim\limits_{n\to\infty} x_n$. It must be that $0 \leq x \leq 1$, or $x \in [0, 1]$. ■

- $F = (0, 1]$ is not closed.

  Proof 3. Take $x_1, \ldots, x_n, \cdots \in (0, 1]$, where $x_n = \dfrac{1}{n} \quad \forall n \geq 1$. Then, $0 \leq x_n \leq 1$. However, notice that $x = \lim\limits_{n\to\infty} x_n = \lim\limits_{n\to\infty} \dfrac{1}{n} = 0 \notin (0, 1]$. Hence, F is not closed. ■

Remark 1.13 *In general, optimization problems are set on closed sets for otherwise, we cannot guarantee, in general, existence of optimal solutions.*

Example 1.3.9 Optimization Problem on a Set that is not Cloased
Suppose $f(x) = x$ and consider the optimization problem

$$\min_{0<x\leq 1} f(x) = \min_{0<x\leq 1} x.$$

Then we know that this problem does not have a solution.

Remark 1.14 *A set can be neither open nor closed.*

Definition 1.3.10 (Boundary Points). A point x is a *boundary point* for U if $\forall\ r > 0,\ B(x,r)$ contains points from both U and its complement. The set of all boundary points of U is denoted by $\mathrm{bd}(U)$.

Example 1.3.11 Boundary Pooints

- $U = [0,1] \implies \mathrm{bd}(U) = \{0,1\}$.
- $U = (0,1] \implies \mathrm{bd}(U) = \{0,1\}$.

Definition 1.3.12 (Compact Set). A set $C \in \mathbb{R}^n$ is called *compact* if it is **closed** and ***bounded***. The latter means that $\exists\ M > 0$ *s.t.* $\|x\| \leq M \quad \forall x \in C$.

1.4 Continuity and Differentiability

Definition 1.4.1 (Continuity). Let $S \subseteq \mathbb{R}^n,\ f : S \to \mathbb{R}, x \in S$. We say that f is *continuous at* x if

$$\lim_{\substack{z\to x \\ z\in S}} f(z) = f(x).$$

If f is continuous at all points $x \in S$, we simply say f is *continuous* on S. We also use the notation $f \in \mathcal{C}(S)$.

Theorem 1.4.2 Weierstrass Theorem
Assume that $S \subseteq \mathbb{R}^n$ is a compact set, and $f : S \to \mathbb{R}$ is a continuous function. Then $\exists\ x_{\min}, x_{\max} \in S$ *s.t.*

$$f(x) \geq f(x_{\min}) \quad \forall x \in S \quad \text{and} \quad f(x) \leq f(x_{\max}) \quad \forall x \in S.$$

In other words, $\min_{x\in S} f(x)$ and $\max_{x\in S} f(x)$ problems are guaranteed to have solutions.

Example 1.4.3 Classes of Continuous Functions

1. Polynomials.
2. $\sin(x)$ and $\cos(x)$; $\tan(x)$ and $\cot(x)$ at certain domain.
3. Exponents: $e^{ax},\ a \in \mathbb{R}$.
4. Logarithm: $\ln x,\ x > 0$.
5.

Theorem 1.4.4 Building Continuous Functions

- If f and g are continuous, then $f \cdot g$, $f + g$, and af are continuous $\forall a \in \mathbb{R}$.
- If f and g are continuous, then $\frac{f}{g}$ is continuous for x *s.t.* $g(x) \neq 0$.
- If f, g are continuous and $h = f \circ g$ makes sense, then h is continuous.

Definition 1.4.5 (Differentiability). Let $S \subseteq \mathbb{R}^n$, $x \in int(S)$, and $f : S \to \mathbb{R}$. Then, the *i-th partial derivative of f at x* is the limit (if it exists)

$$\frac{\partial f(x)}{\partial x_i} = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}, \quad \text{where } e_i \text{ is the standard basis.}$$

If all partial derivatives exist, then we assemble them in a column vector called *gradient*.

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{pmatrix}^T$$

We say that f is *continuously differentiable* on S if $\exists\, U$ open set *s.t.* $S \subseteq U$ and $\nabla f(x)$ exists $\forall x \in U$ and is continuous. In this case, we write $f \in \mathcal{C}^1(S)$.

Example 1.4.6 Continuous Function that is not Continuously Differentiable

Consider $f(x) = |x|$. Then we know its derivative

$$f'(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ \text{undefined} & x = 0. \end{cases}$$

So, $f \in \mathcal{C}(\mathbb{R})$ but $f \notin \mathcal{C}^1(\mathbb{R})$.

Definition 1.4.7 (Directional Derivative). Let $f \in \mathbb{R}^n \backslash \{0\}$. Then, the *directional derivative of* f *at* x is the limit (if it exists)

$$f'(x; d) = \lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t}.$$

Remark 1.15 *If* $f \in \mathcal{C}^1(S)$*, then*

$$f'(x; d) = \boldsymbol{\nabla} f(x)^T \cdot d.$$

However, the converse is not true in general. Indeed, for $f(x) = |x|$*, we have that* $f'(0; 1) = 1$ *(the positive direction), and* $f'(0; -1) = -1$ *(the negative direction). But* $f'(0)$ *does not exist.*

Definition 1.4.8 (Second-Order Differentiability). The (i, j)*-th partial derivative of* f *at* x is

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f(x)}{\partial x_j}\right).$$

If all second order partial derivatives exist and are continuous on S, we say that f is *twice continuously differentiable* on S and write $f \in \mathcal{C}^2(S)$.

If $f \in \mathcal{C}^2(S)$, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \quad \forall i, j.$$

If f has all second-order partial derivatives at x, then we denote the *Hessian of* f at x by the matrix

$$\boldsymbol{\nabla}^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j=1}^n$$

If $f \in \mathcal{C}^2(S)$, then $\boldsymbol{\nabla}^2 f(x)$ is *symmetric* for all $x \in S$.

Definition 1.4.9 (Small-O Notation). $o(r)$ is the *small-o notation* and means that this quantity is much smaller than r. For example, $o(\|y - x\|)$ is any quantity *s.t.*

$$\lim_{y \to x} \frac{o(\|y - x\|)}{\|y - x\|} = 0.$$

Theorem 1.4.10 Taylor Approximation I
If f is differentiable at x, then

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

Theorem 1.4.11 Taylor Approximation II
If f is twice differentiable at x, then

$$f(y) = f(x) + \boldsymbol{\nabla} f(x)^T (y - x) + \frac{1}{2}(y - x)^T \boldsymbol{\nabla}^2 f(x)(y - x) + \underbrace{o\big(\|y - x\|^2\big)}_{\text{small error}}.$$

Theorem 1.4.12 Taylor Approximation III
If f is twice differentiable at x, then

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(c)(y - x) \text{ for some } c(x, y, f \ldots),$$

where the point c is dependent on x, y, and f, but we do not know exactly what c is.

Remark 1.16 *From Taylor Approximation II to III, we improve our approximation from an expression with a small error, to an exact equation. However, the trade-off here is that we have to introduce a new constant c, which we do not have any information about.*

2 Unconstrained Optimization

2.1 Global and Local Optima

Definition 2.1.1 (Global Minimum and Maximum). Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{R}^n$. Then, $x^* \in S$ is called a

- *global minimum point* of f over S if $f(x) \geq f(x^*)$ for any $x \in S$.
- *strict global minimum point* of f over S if $f(x) > f(x^*)$ for any $x^* \neq x \in S$.
- *global maximum point* of f over S if $f(x) \leq f(x^*)$ for any $x \in S$.
- *strict global maximum point* of f over S if $f(x) < f(x^*)$ for any $x^* \neq x \in S$.

Definition 2.1.2 (Feasible Set and Feasible Solution). The set S on which the optimization of f is performed is called the *feasible set*, and any only $x \in S$ is called a *feasible solution.*
Definition 2.1.3 (Minimizer and Maximizer). We refer to a global minimum point as a *minimizer* or a *global minimizer*, and a global maximum point as a *maximizer* or a *global maximizer.* A vector $x^* \in S$ is called a *global optimum* of f over S if it is either a global minimum or a global maximum.
Definition 2.1.4 (Maximal and Minimal Value). The *maximal value* of f over S is defined as the supremum of f over S:

$$\max\{f(x) : x \in S\} = \sup\{f(x) : x \in S\}.$$

If $x^* \in S$ is a global maximum of f over S, then the maximum value of f over S is $f(x^*)$. The *minimal value* of f over S is the infimum of f over S,

$$\min\{f(x) : x \in S\} = \inf\{f(x) : x \in S\},$$

and is equal to $f(x^*)$ when x^* is a global minimizer of f over S.

Remark 2.1 (Difference between $\min$ and $\inf$) *For $A \subseteq \mathbb{R}$, $\min A = y^*$ if $y^* \in A$, and $y^* \leq y \quad \forall y \in A$. On the other hand, $\inf A = y^*$ if $y^* \leq y \quad \forall y \in A$, and any $y' > y^*$ is NOT a lower bound for A.*

Remark 2.2 *There could be several global minimum points, but there could be only one minimal value.*

Definition 2.1.5 (Set of Global Minimizers and Global Maximizers). The set of *all global minimizers* of f over S is denoted by

$$\arg\min\{f(x) : x \in S\}$$

and the set of all global maximizers of f over S is denoted by

$$\arg\max\{f(x) : x \in S\}.$$

Definition 2.1.6 (Local Minima and Maxima). Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{R}^n$. Then, $x^* \in S$ is called a

- *local minimum point* of f over S if there exists $r > 0$ for which $f(x^*) \leq f(x)$ for any $x \in S \cap B(x^*, r)$.
- *strict local minimum point* of f over S if there exists $r > 0$ for which $f(x^*) < f(x)$ for any $x^* \neq x \in S \cap B(x^*, r)$.
- *local maximum point* of f over S if there exists $r > 0$ for which $f(x^*) \geq f(x)$ for any $x \in S \cap B(x^*, r)$.
- *strict local maximum point* of f over S if there exists $r > 0$ for which $f(x^*) > f(x)$ for any $x^* \neq x \in S \cap B(x^*, r)$.

Lemma 2.1.7 Fermat's Theorem: For a one-dimensional function f defined and differentiable over an interval (a, b), if a point $x^* \in (a, b)$ is a local maximum or minimum, then $f'(x^*) = 0$.

Remark 2.3 *Moving into multidimensional extension of this Lemma, the result states that the gradient is zero at local optimum points. We refer to such an optimality condition as a* first order optimality condition .

Theorem 2.1.8 First Order Optimality Condition for Local Optima Points
Let $f : U \to \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $x^* \in \operatorname{int}(U)$ is a local optimum point and that all the partial derivatives of f exist at x^*. Then, $\nabla f(x^*) = 0$.

Proof 1. Let $i \in \{1, 2, \ldots, n\}$ and consider the one-dimensional function $g(t) = f(x^* + te_i)$, where e_i is the standard basis. Note that g is differentiable at $t = 0$ and that $g'(0) = \dfrac{\partial f}{\partial x_i}(x^*)$. Since x^* is a local optimum point of f, it follows that $t = 0$ is a local optimum of g, which immediately implies that $g'(0) = 0$. The latter equality is exactly the same as $\dfrac{\partial f}{\partial x_i}(x^*) = 0$. Since this is true for any $i \in \{1, 2, \ldots, n\}$, the result $\nabla f(x^*) = 0$ follows. ■

Remark 2.4 *Our proof of the multidimensional First Order Condition relies on the first order optimality condition for one-dimensional functions.*

Remark 2.5 *Theorem 2.1.8 presents a* necessary *optimality condition: the gradient vanishes at all local optimum points, which are interior points of the domain of the function; however, the reverse claim is not true since there could be points which are not local optimum points whose gradient is zero.*

Definition 2.1.9 (Stationary Points). Let $f : U \to \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $x^* \in \operatorname{int}(U)$ and that f is differentiable over some neighborhood of x^*. Then, x^* is called a *stationary point* of f if $\nabla f(x^*) = 0$.

Remark 2.6 *Theorem 2.1.8 essentially states that local optimum points are necessarily stationary points. However, again, stationary points are not necessarily local optimum points.*

2.2 Classification of Matrices

Definition 2.2.1 (Positive Definiteness, Negative Definiteness). A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called

- *positive semidefinite*, denoted by $A \succeq 0$, if $x^T Ax \geq 0$ for every $x \in \mathbb{R}^n$.
- *positive definite*, denoted by $A \succ 0$, if $x^T Ax > 0$ for every $x \neq 0 \in \mathbb{R}^n$.
- *negative semidefinite*, denoted by $A \preceq 0$, if $x^T Ax \leq 0$ for every $x \in \mathbb{R}^n$.
- *negative definite*, denoted by $A \prec 0$, if $x^T Ax < 0$ for every $x \neq 0 \in \mathbb{R}^n$.
- *indefinite* if there exist x and $y \in \mathbb{R}^n$ such that $x^T Ax > 0$ and $y^T Ay < 0$.

Remark 2.7 *A matrix is negative (semi)definite if and only if $-A$ is positive (semi)definite.*

Lemma 2.2.2 Necessary Condition for Definiteness of Matrices: If $A \in \mathbb{R}^{n\times n}$ is a positive definite matrix, then its diagonal elements are positive. If $A \in \mathbb{R}^{n\times n}$ is a semidefinite matrix, then its diagonal elements are nonnegative. Similarly, if A is a negative definite matrix, then its diagonal elements are negative. If A is a negative semidefinite matrix, then tis diagonal elements are nonpositive.

Remark 2.8 *Note that Lemma 2.2.2 gives a necessary condition for a positive definite matrix. It is not sufficient. That is, one can easily generate a matrix with positive diagonal entries that is not positive definite.*

Lemma 2.2.3 : Let A be a symmetric $n \times n$ matrix. If there exists positive and negative elements in the diagonal of A, then A is indefinite.

Theorem 2.2.4 Eigenvalue Characterization Theorem
Let A be a symmetric $n \times n$ matrix. Then,

- A is positive definite if and only if all its eigenvalues are positive.
- A is positive semidefinite if and only if all its eigenvalues are nonnegative.
- A is negative definite if and only if all its eigenvalues are negative.
- A is negative semidefinite if and only if all its eigenvalues are nonpositive.
- A is indefinite if and only if it has both positive and negative eigenvalues.

Corollary 2.2.5 : Let A be a positive semidefinite (definite) matrix. Then, $\operatorname{tr}(A)$ and $\det(A)$ are nonnegative (positive).

Lemma 2.2.6 L:et $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$. Then, D is

- positive definite if and only if $d_i > 0 \quad \forall i$.
- positive semidefinite if and only if $d_i \geq 0 \quad \forall i$.
- negative definite if and only if $d_i < 0 \quad \forall i$.
- negative semidefinite if and only if $d_i \leq 0 \quad \forall i$.
- indefinite if and only if $\exists\, i, j$ *s.t.* $d_i > 0, d_j < 0$.

Proposition 2.2.7 : Let A be a symmetric 2×2 matrix. Then, A is positive semidefinite (definite) if and only if both $\operatorname{tr}(A) \geq 0$ and $\det(A) \geq 0$ ($\operatorname{tr}(A) > 0$ and $\det(A) > 0$).

Example 2.2.8 Square Root of Matrices

For any positive semidefinite matrix A, we can define the square root of matrix $A^{1/2}$. Let $A = UDU^T$ by the spectral decomposition. Then, $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$, where d_i's are eigenvalues of A. Since A is positive semidefinite, we have $d_1, \ldots, d_n \geq 0$. Now, define

$$A^{1/2} = UEU^T,$$

where $E = \operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$. Then

$$A^{1/2}A^{1/2} = UEU^TUEU^T = UEEU^T = UDU^T = A.$$

The matrix $A^{1/2}$ is also known as the *positive semidefinite square root.*

Definition 2.2.9 (Principal Minor). Given an $n \times n$ matrix, the determinant of the upper left $k \times k$ sub-matrix is called the k*-th principal minor* and is denoted by $D_k(A)$.

Example 2.2.10 Principal Minor

Consider a 3×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Then the principal minors are

$$D_1(A) = a_{11},\ D_2(A) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\ D_3(A) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

Theorem 2.2.11 Principal Minors Criterion
Let A be an $n \times n$ symmetric matrix. Then, A is positive definite if and only if all its principal minors are positive. That is, $D_1(A) > 0, \ldots, D_n(A) > 0$.

Remark 2.9 *When the matrix becomes large, computing its determinant will be hard. So, other method to determine the definiteness of a matrix shall be introduced.*

Definition 2.2.12 (Diagonally Dominant Matrices). Let A be a symmetric $n \times n$ matrix. Then,

- A is *diagonally dominant* if

$$|A_{ii}| \geq \sum_{j \neq i} |A_{ij}|, \quad \text{for } i = 1, 2, \ldots, n.$$

- A is called *strictly diagonally dominant* if

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|, \quad \text{for } i = 1, 2, \ldots, n.$$

Theorem 2.2.13 Positive (Semi)Definiteness of Diagonally Dominant Matrices

- Let A be a symmetric $n \times n$ diagonally dominant matrix whose diagonal elements are nonnegative. Then, A is positive semidefinite.
- Let A be a symmetric $n \times n$ strictly diagonally dominant matrix whose diagonal elements are positive. Then, A is positive definite.

3 Least Square

4 Constrained Optimization