Emory University

# MATH 347 Non Linear Optimization

Learning Notes

Jiuru Lyu

January 24, 2024

## Contents

**1 Math Preliminaries** **2**

1.1 Introduction to Optimization . . . . . . . . 2

1.2 Linear Algebra Review . . . . . . . . 4

**2 Unconstrained Optimization** **11**

**3 Least Square** **12**

**4 Constrained Optimization** **13**

# 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., $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{m \times n}$, $\dots$), 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 $S = \mathbb{R}^n$, $S = \mathbb{R}^{m \times n}$, $\dots$ $(m, n$ are fixed).
  - *Constrained problems*: where $S \subsetneq \mathbb{R}^n$, $S \subsetneq \mathbb{R}^{m \times n}$, $\dots$.

    *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: $S = \{(x_1, x_2) : x_1, x_2 \geq 0\} \subsetneq \mathbb{R}62$. $\longrightarrow$ 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 $\min\limits_{x \in \mathbb{R}} f(x)$, 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 $\lim\limits_{x \to x_0} \varepsilon(x - x_0)$.

**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 $v_1 \dots, v_d \in \mathbb{R}^d$ is a *basis* in $\mathbb{R}^d$ if $\forall\, x \in \mathbb{R}^d$, $\exists!\ \alpha_1, \dots, \alpha_d \in \mathbb{R}$ *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 $\ell_2$ norm in $\mathbb{R}^d$.

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

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

- **Definition 1.2.15 ($\infty$-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:** $\left|x^T y\right| \leq \|x\|_2 \|y\|_2$.

**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 $\lambda$ 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.* $Av_i = \lambda v_i \quad \forall 1 \leq i \leq d$.

**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 $\lambda$'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*

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

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

# 2 Unconstrained Optimization

# 3 Least Square

# 4 Constrained Optimization