# CO367, Course Notes

Transcribed by Louis Castricato

October 19, 2018

# Introduction

Mathematical Optimization or Mathematical Programming
Informally: Find a best solution to the model of a problem
*best* according to a given objective/criterion
Applications include

1. Operations research
   (a) Scheduling + Planning
   (b) Supply Chain Management
   (c) Vehicular Routing
   (d) Power Grid Optimization
2. Statistics and Machine Learning
   (a) Curve FItting
   (b) Classification, Clustering, SVM,...
   (c) Deep Learning
3. Finance
4. Optimal Control
5. Biology

(OPT) min f(x) s.t.

$$g_i(x) \leq 0, \text{ for } i \in \{1, 2, 3, \dots, m\}$$

Remarks

- a. $\max f(x) = -(\min - f(x))$
- b. $\{x \in \mathbb{R}^n : g(x) \geq 0\} = \{x \in \mathbb{R}^n : -g(x) \leq 0\}$
- c. $\{x \in \mathbb{R}^n : g(x) \geq b\} = \{x \in \mathbb{R}^n : -g(x) - b \leq 0\}$

## Classification of Problems - 1

- if $f(x) = 0, \forall x \in \mathbb{R}^n \implies$ (OPT) is a feasibility problem
- if we have $m = 0$ constraints $\implies$ (OPT) is an unconstrained optimization problem

## Classification of Problems - 2

Q: Why do we need f and g? A: In abscence of hyp. on f and g, (OPT) is unsolvable.

## Note: "Black box" optimization framework

All that is given is an oracle function that can compute values of $f(x)$ $\forall x$ in the domain of $f$

Example: consider

$$\begin{aligned} &\min f(x) \\ &\text{s.t.} g(x) \leq 0, \text{ for } i \in [1, m] \cap \mathbb{N} \\ &h(x) \leq 0 \\ h(x), &\text{ when } x \in \mathbb{Z}^n, \text{ do: } 0 \\ &h(x), \text{ do: } 1 \end{aligned}$$

in other words, we only want integral solutions.

**Definition 0.1. Discrete Optimization:** When the constraint of OPT restrict to a lattice, we have a discrete optimization problem

**Definition 0.2. Continuous:** A function $f : D \mapsto \mathbb{R}$ is continuous over $D$ ($f \in C^k(D)$) if all its $k^{\text{th}}$ derivatives are continuous over $D$.

Consider the following examples

$$\begin{aligned} f(x) \text{ when } x \geq 2, \text{ do: } 1 \\ f(x), \text{ do: } -1 \end{aligned}$$

Then $f(x)$ is not continuous.

In another example we have $g(x)$, do: $\text{abs}(x-2)$. Then $g(x) \in C^0$.

**Definition 0.3. Gradient:** Let $f \in C^1(D)$ for $D \subseteq \mathbb{R}^n$. The gradient is $\nabla f : D \mapsto \mathbb{R}^n$ if it satifies $\nabla f \in C^0(D)$ and is given by $\nabla f(x) = \begin{bmatrix} \frac{\delta f}{\delta x_1(x)} \cdots \frac{\delta f}{\delta x_n(x)} \end{bmatrix}$.

**Definition 0.4. Hessian:** Let $f \in C^2(D)$ for $D \subseteq \mathbb{R}^n$. Its Hessian is $\nabla^2 f : D \mapsto \mathbb{R}^n$. It satisfies $\nabla^2 f \in C^0(D)$ and is given by

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

**Definition 0.5. Linear:** A function $f : D \mapsto \mathbb{R}$, $D \subseteq \mathbb{R}^n$ is linear if $\exists c \in \mathbb{R}^n$ where $f(x) = c^T x, \forall x \in D$. Then $\nabla f(x) = c$ and $\nabla^2 f(x) = 0$.

**remark:** if $f, g_i$ are linear, then OPT is a linear programming function.

# 1 Linear Algebra

A vector and matrix norm.

**Definition 1.1. Norm:** A norm $\|\cdot\|$ on $\mathbb{R}^n$ assigns a scalar $\|x\|$ to every $x \in \mathbb{R}^n$ *s.t.*

1. $\|x\| \geq 0, \forall x \in \mathbb{R}^n$
2. $\|cx\| = |c|\|x\|, \forall x \in \mathbb{R}^n \forall c \in \mathbb{R}$
3. $\|x\| = 0 \iff x = 0$
4. $\|x + y\| \leq \|x\| + \|y\| \ \forall x, y \in \mathbb{R}^n$

$L^k$ norm $\|x\| = \left(\sum (x_i)^k\right)^{\frac{1}{k}}$ in particular,

1. Manhattan Norm $= L_1$
2. Euclidean Norm $= L_2$
3. Infinite Norm $= L_\infty = \max(|x_i|)$

Schwartz inequaloty: $\forall x, y \in \mathbb{R}^n$
$|x^T y| \leq \|x\|_2 \cdot \|y\|_2$

**Theorem 1.1.** ***Pythagorean Theorem:*** *If $x, y \in \mathbb{R}$ are orthogonal then $\|x + y\|^2 = \|x\|^2 + \|y\|^2$ under $L_2$.*

**Definition 1.2. Matrix Norm:** Given a vector norm $\|\cdot\|$, the induced magtrix norm associates a scalar $\|x\|$ to all $A \in \mathbb{R}^{n \times n}$
$\|A\| = \max\|Ax\|$ where $\|x\| = 1$

**Property of Matrix norm:**
$\|A\|_2 = \max\|Ax\|_2 = \max|y^T Ax|$, where $\|x\|_2 = 1$ and $\|y\|_2 = 1$.
Proof is trivial by schwartz inequality.
$\|A\| = \|A^T\|_2$
**TFAE:** [1]

1. $A$ is nonsingular
2. $A^T$ is nonsingular
3. $\forall x \in \mathbb{R}^n$ if $x \neq 0$ then $Ax \neq 0$.
4. $\forall b \in \mathbb{R}^n$, $\exists x \in \mathbb{R}^n$ *s.t.* $Ax = b$ and $x$ is unique
5. The columns of $A$ are linearly independent
6. The rows of $A$ are linearly independent

[1] The following are equivalent

7. A unique inverse of $A$ exists

8. If $B$ is a matrix *s.t.* an inverse of $B$ exists, then $(AB)^{-1} = A^{-1}B^{-1}$

**Definition 1.3. Eigenvalue:** The characteristic polynomial $\Phi : \mathbb{R} \mapsto \mathbb{R}$ of $A \in \mathbb{R}^{n\times n}$ is $\Phi(\lambda) = \det(A - \lambda I)$. It has $n$ complex roots, the eigenvalues of $A$. GIven an eigenvalue $\lambda$ of $A$, $x \in \mathbb{R}^n$ is its corresponding eigenvector of $A$ if $Ax = \lambda x$.

**Properties :** Given $A \in \mathbb{R}^{n\times n}$

1. $\lambda$ is an eigenvalue $\iff$ $\exists$ a corresponding eigenvector $x$.

2. $A$ is simuglar $\iff$ it has a zero eigenvalue

3. If $A$ is triangular, then its eigenvalues are its diagonal elements

4. If $S \in \mathbb{R}^{n\times n}$ is nonsingular and $B = SAS^{-1}$ then $A$ and $B$ have the same eigenvalues.

5. If the eigenvalues of $A$ are $\{\lambda_1, \dots, \lambda_n\}$ then
   (a) the eigenvalues of $A + cI$ are $c + \lambda_1, \dots, c + \lambda_n$.
   (b) the eigenvalues of $A^k$ are $\lambda_1^k, \dots, \lambda_n^k$. This also holds for $k = -1$.
   (c) the eigenvalues of $A^T$ are the same as the eigenvalues of $A$.

**Definition 1.4. Spectral Radius:** The spectral radius $\rho(A)$ of $A \in \mathbb{R}^{n\times n}$ is the maximum magnitude of its eigenvalues.

**Property:**

**Lemma 1.2.** *For any induced norm,* $\|\cdot\|$, $\rho(A) \le \|A^k\|^{\frac{1}{k}}$ $\forall k \in \mathbb{N}$

**Proof:** By defn, $\|A^k\| = \max\|A^k y\| = \max\frac{\|A^k y\|}{\|y\|}$, where $\|y\| = 1$.
Let $\lambda$ be an eigenvalue of $A$, and $x$ its eigenvector. Then

$$\|A^k\| \ge \frac{\|A^k x\|}{\|x\|} = \frac{\|A^{k-1}Ax\|}{\|x\|} = \frac{A^{k-1}\lambda x}{\|x\|} = \dots = \frac{\|\lambda^k x\|}{\|x\|} = \frac{(|\lambda^k|\|x\|)}{\|x\|} = \|\lambda^k\|$$

So for any eigenvalue, $\|A^k\| \ge |\lambda^k| \implies \|A^k\|^{\frac{1}{k}} \ge \lambda \implies \rho(A) \le \|A^k\|^{\frac{1}{k}}$.

**Lemma 1.3.** *For any induced norm,* $\|\cdot\|$, $\lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} = \rho(A)$. *Furthermore,* $\lim_{k\to\infty} A^k = A$ *iff* $\rho(A) \le 1$.

**Proof:** Exercise!

**Symmetrix Matricies:**
**Property:** Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix.

1. its eigenvalues are real

2. its eigenvectors are mutually orthongal
3. assume its eigenvectors are normalized. Let $(\lambda_i, v_i)$ refer to an eigenpair. Then $A = \sum \lambda_i x_i x_i^T$.

**Proof:** Exercise!

**Lemma 1.4.** *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix, then $\|A\|_2 = p(A)$.*

**Proof:** from before, $\rho(A) \leq \|A^k\|^{\frac{1}{k}}$ and in particular we have that $p(A) \leq \|A\|_2$. Now all we need to do is show that $p(A) \geq \|A\|_2$.
As the eigenvectors $x_i$ $i = 1, \ldots, n$ of C are mutually orthogonal we can write any $y \in \mathbb{R}^n$ as $y = \sum \beta_i x_i$ for some $\beta \in \mathbb{R}^n$.

By pythagoras' theorem, $\|y\|_2 = \sum \beta_i^2 \cdot \|x\|_2^2$. Hence $Ay = A \sum \beta_i^2 \cdot \|x\|_2^2 = \sum \beta_i \lambda_i x_i$. Again we can apply pythagoras'

$$\begin{aligned}
\|Ay\|_2^2 &= \| \sum \beta_i \lambda_i x_i \|_2^2 \\
&= \sum \beta_i \lambda_i^2 \|x\|_2^2 \\
&= \sum |\lambda_i|^2 \cdot |\beta_i|^2 \cdot \|x\|_2^2 \\
&\leq \sum \rho(A)^2 |\beta_i|^2 \|x\|_2^2 \\
&= \rho(A)^2 \sum |\beta_i|^2 \|x\|_2^2 \\
&= \rho(A)^2 \|y\|_2^2
\end{aligned}$$

This then implies that

$$\begin{aligned}
&\|A\|_2 \leq \rho(A) \|y\|_2 \\
\implies &A = \max \frac{\|Ay\|_2}{\|y\|_2} \leq \frac{(\rho(A)\|y\|_2)}{\|y\|_2}, \text{ where } y \neq 0 \\
\implies &\|A\|_2 \leq \rho(A)
\end{aligned}$$

Therefore $\|A\|_2 = \rho(A)$.

**Lemma 1.5.** *Let $A \in \mathbb{R}^{n \times n}$ be symmetric, with eigen values $\lambda_1 \leq \ldots \leq \lambda_n \in \mathbb{R}$. Then $\forall y \in \mathbb{R}^n$ we have that $\lambda_1 \|y\|_2^2 \leq y^T A y \leq \lambda_n \|y\|_2^2$.*

**Proof:** Express y as $\sum \beta_i x_i$, $i = 1, \ldots, n$ where $\beta_i \in \mathbb{R}$, $x_i$ are orthongal eigenvectors of $A$. Firstly:

$$y^T A y = (\sum \beta_i x_i)^T (\sum \beta_i \lambda_i x_i) = \sum \beta_i^2 \lambda_i \|x_i\|_2^2$$

WLOG, assume that $\|x_i\|_2 = 1$ by normalization. So $y^T Ay = \sum \lambda\beta_i^2$. Secondly: $\|y\|_2^2 = \sum \beta_i^2$

$$\sum \lambda_1\beta_1^2 \leq \sum \lambda_i\beta_i^2 \leq \sum \lambda_n\beta_n^2 \implies \lambda_1\|y\|_2^2 \leq y^T Ay \leq \lambda_n\|y\|_2^2.$$

**Lemma 1.6.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then $\|A^k\|_2 = \|A\|_2^k$.*

**Proof:**

Since $A$ is symmetric, we have that $(A^k)^T = A^k$ and $\|A^k\|_2 = \rho(A^k)$. So $\rho(A^k) = \rho(A)^k$. Therefore $\|A\|_2^k = \|A^k\|_2$.