CO367, Course Notes

Transcribed by Louis Castricato

October 27, 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.

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

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^{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 $\nabla f(x) = \left[\begin{array}{ccc} \frac{\delta f}{\delta x_1(x)} & \cdots & \frac{\delta f}{\delta x_n(x)} \end{array} \right]$.

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 λ 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. λ 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,* $\|\cdot\|$, $\rho(A) \leq \|A^k\|^{\frac{1}{k}}$ $\forall k \in \mathbb{N}$

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

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

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

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* $\rho(A) \leq 1$.

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 (λ_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: $\|y\|_2^2 = \sum \beta_i^2$

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

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 $\|A\|_2^k = \|A^k\|_2$.

Lemma 1.7. *Let $A \in \mathbb{R}^{n\times n}$, then $\|A\|_2^2 = \|A^T A\|_2 = \|AA^T\|_2$.*

Proof: According to the schwartz inequality, $x^T y \le \|x\|_2 \cdot \|y\|_2$

$$\begin{aligned}\|Ax\|_2^2 = (Ax)^T(Ax) = (x^T A^T)(Ax) &= x^T \cdot A^T Ax \\ &\le \|x\|_2 \cdot \|A \cdot Ax\|_2 \\ &\le \|x\|_2 \cdot \|A^T A\|_2 \cdot \|x\|_2, \forall x \in \mathbb{R}^n\end{aligned}$$

Remark.

$$\begin{aligned}\|A\|_2^2 &= \max\frac{\|Ax\|_2^2}{\|X\|_2^2} \le \|A^T A\|_2 \\ \|A^T A\| &= \max\|y^T A^T\|_2 \cdot \|Ax\|_2 \\ &= (\max\|y^T A^T\|_2)(\max\|Ax\|_2) \\ &= \|A\|_2\end{aligned}$$

So we have that $\|A\|_2^2 = \|A^T A\|$. For $\|A\|_2^2 = \|AA^T\|$ repeat steps with A amd A^T swapped.

Lemma 1.8. *For any $A \in \mathbb{R}^{m\times n}$, $A^T A$ is psd and $A^T A$ is pd iff rank$(A) = n$*

Proof: The proof of this follows from the fact that a matrix with all positive eigenvalues is pd, and a matrix with all positive/zero eigenvalues is psd. Notice that $A^T A$ has all positive eigenvalues if rank$(A) = n$, and $A^T A$ has all positive/zero eigenvalues otherwise. If required, showing that $A^T A$ has all positive/zero eigenvalues can be done by multiplying their orthogonal decompositions.

Corollary. *If A is a square matrix, $A^T A$ is pd iff A is nonsingular.*

Properties:

1. A square symmetric matrix is psd iff all of its eigenvalues are ≥ 0
2. A square symmetric matrix is pd iff all of its eigenvalues are > 0

Proof (For statement 1): Let λ be an eigenvalue of a psd matrix A and let x be its corresponding nonzero eigenvector. Notice that

$$\begin{aligned}&x^T Ax \ge 0,\ \text{so } x^T\lambda x = \lambda\|x\|_2^2 \ge 0 \\ &\implies \lambda \ge 0\end{aligned}$$

Let $\{\lambda_i\}$ refer to the set of eigenvalues of A and let $\{x_i\}$ refer to its eigenvectors. As such, $\forall y \in \mathbb{R}^n$ y is a linear combination of $\{x_i\}$. Namely notice that we can write

$$\begin{aligned}
y &= \sum \beta_i x_i \\
y^T A y &= (\sum \beta_i x_i)^T \sum \beta_i A x_i \\
&= (\sum \beta_i x_i)^T \sum \beta_i \lambda_i x_i \\
&= \sum \beta_i^2 \lambda \|x_i\|_2^2 \geq 0
\end{aligned}$$

Statement 2 is left as an exercise to the reader.

Corollary. *The inverse of a pd matrix is also pd*

Proof: Trivial

2 Convexity

Definition 2.1. A set C is called convex if it is closed under convex combinations. Namely $\forall x, y \in C$, $\forall t \in [0,1]$ we have that $tx + (1-t)y \in C$.

Definition 2.2. A function f is said to be convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ $\forall x, y \in D$, $\forall \lambda \in [0,1]$. A function f is said to be strictly convex if $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$ $\forall x, y \in D$, $\forall \lambda \in [0,1]$.

Definition 2.3. A function f is said to be concave if $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)$ $\forall x, y \in D$, $\forall \lambda \in [0,1]$. A function f is said to be strictly convex if $f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda)f(y)$ $\forall x, y \in D$, $\forall \lambda \in [0,1]$.

Remark. Notice that convex sets are closed under intersections. The Minkowski sum of convex sets is convex. The image of a convex set under a linear transformation is convex. The proof of all three properties is left as an exercise to the reader.

Definition 2.4. Let f refer to a function with a convex domain C. The level sets of f are $\{x \in C : f(x) \leq \alpha\}$, $\forall \alpha \in R$.

Definition 2.5. Same f as above. The epigraph of f is a subset of $\mathbb{R}^{n+1}$ given by $\text{epi}(f) = \{(x, \alpha) : x \in C, \alpha \in R, f(x) \leq \alpha\}$.

Definition 2.6. Same f as above. The hypograph of f is a subset of $\mathbb{R}^{n+1}$ given by $\text{hypo}(f) = \{(x, \alpha) : x \in C, \alpha \in R, f(x) \geq \alpha\}$.

Remark. Some intuition. Notice that the intersection of the epi and hypo graph of a function is quite literally the graph of said function. Furthermore the epigraph of a function can be viewed as the region above the graph, inclusive, where as the hypograph of a function can be viewed as the region below the graph, inclusive.

Properties

1. If $f : C \to R$ is convex, then its level sets are also convex. The converse is not true
   (a) Consider the example of $f(x) = \sqrt{|x|}$.
2. $f : C \to R$ is convex iff its epigraph is a convex set.
3. $f : C \to R$ is concave iff its hypograph is a concave set.
4. $f : C \to R$ is linear iff its both concave and convex.
5. The sum of two convex functions is also convex
6. The sum of two concave functions is also concave
7. The max of two convex functions is a convex (piecewise) function
8. The max of two concave functions is a concave (piecewise) function
9. Any vector norm is convex

We'll prove the last statement and leave the rest as an exercise to the reader.

Proof: This proof relies on the fact that norms satisfy the triangle inequality. Let $f(x) = \|x\|$. Then notice that $\forall x, y \in D$, $\forall \lambda \in [0,1]$ we have that

$$\begin{aligned}&f(\lambda x + (1-\lambda)y)\\ &= \|\lambda x + (1-\lambda)y\|\\ &\leq \lambda\|x\| + (1-\lambda)\|y\|\\ &= \lambda f(x) + (1-\lambda)f(y)\end{aligned}$$

Theorem 2.1. *Taylor's theorem for univariate functions.*

Let $f : D \to R$.

$$f(x+h) = \sum \frac{h_i}{i!} f^i(x) + \Phi(h)$$

where Φ refers to the residual function. Namely

$$\Phi(h) = \frac{h^{k+1}}{(k+1)!} f^{k+1}(x + \lambda h)$$

for some $\lambda \in [0,1]$. Furthermore

$$\lim_{h \to 0} \frac{\Phi(h)}{h^k} = 0$$

Theorem 2.2. *Taylor's theorem for multivariate functions.*

Let $f : D^m \to R$*.*

$$f(x+h) = f(x) + h^T \nabla f(x) + \Phi(h)$$

where Φ *refers to the residual function. Namely*

$$\Phi(h) = \frac{1}{2} h^T \nabla^2 f(x + \lambda h) h$$

for some $\lambda \in [0, 1]$*. Furthermore*

$$\lim_{h \to 0} \frac{\Phi(h)}{\|h\|} = 0$$

Theorem 2.3. *2nd order*

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h + \Phi(h)$$

$$\lim_{h \to 0} \frac{\Phi(h)}{\|h\|} = 0$$

Notice that Taylor's theorem for univariate functions can be easily derived from Taylor's theorem for multivariate functions and vice versa.

Theorem 2.4. *Mean value Theorem*

Let $f : D \to R$ *be* C^1 *smooth. Then* $\forall x, y \in D\ \exists z \in [x, y]$ *suuch that* $f(y) = f(x) + \nabla f(z)(y - x)$*.*

Proof: follows from the zeroth order taylor expansion

Definition 2.7. The directional derivative of f in direction y is given by

$$\nabla_y f(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha y) - f(x)}{\alpha}$$

Definition 2.8. The gradient of f is given by

$$\nabla f = (\nabla_{e_1} f(x), \dots, \nabla_{e_k} f(x))$$

Corollary. *If* f *is* C^1 *smooth, the directional derivative of* f *in direction* y *can be computed as*

$$\nabla_y f = y \cdot \nabla f$$

Proof: Left as an exercise to the reader.

Lemma 2.5. *Let* C *be convex. and let* f *be differentiable over* C*.* f *is convex iff*

$$f \geq f(x) + (z - x)^T \nabla f(x),\ \forall x, y \in C$$

Remark. This is the most important theorem of this chapter!! Make sure you understand it.

Proof:

$(\implies)$

As C is convex, $x + (z-x)\alpha = \alpha z + (1-\alpha)x \in C$, $\forall \alpha \in [0,1]$

$$\lim_{\alpha \to 0} \frac{f(x + \alpha(z-x)) - f(x)}{\alpha} = (z-x)\nabla f(x)$$

But by convexity

$$f(x + \alpha(z-x)) - f(x)\alpha \leq f(z) - f(x)$$

Taking the limit of $\alpha \to 0$ of both sides gets us the desired result.

$(\impliedby)$

Assume that $f(z) \geq f(x) + (z-x)^T \nabla f(x)$. Let $a, b \in C$ be any points in the domain of f and let $c = \alpha a + (1-\alpha)b$. We can write

1. $f(a) \geq f(c) + (a-c)^T \nabla f(c)$
2. $f(b) \geq f(c) + (b-c)^T \nabla f(c)$

Multiply 1 by α and 2 by $(1-\alpha)$ and sum them

$$\alpha f(a) + (1-\alpha)f(b) \geq \alpha(f(c) + (a-c)^T \nabla f(c)) + (1-\alpha)(f(c) + (b-c)^T \nabla f(c))$$
$$\alpha f(a) + (1-\alpha)f(b) \geq f(c) + \alpha(a-c)^T \nabla f(c) + (1-\alpha)(b-c)^T \nabla f(c)$$
$$\alpha f(a) + (1-\alpha)f(b) \geq f(c) + (\alpha a - \alpha c + b - \alpha b - c + \alpha c)^T \nabla f(c)$$
$$\alpha f(a) + (1-\alpha)f(b) \geq f(c)$$
$$\alpha f(a) + (1-\alpha)f(b) \geq f(\alpha a + (1-\alpha)b)$$

Therefore, f is convex over C.

Remark. Drawing out what this theorem is describing aids in forming an intuition.

Properties: Assume that f is C^2 smooth.

1. If $\nabla^2 f$ is psd, then f is convex
2. If $\nabla^2 f$ is pd, then f is strictly convex.
3. If the domain of f is $\mathbb{R}^n$ and f is convex over D, then $\nabla^2 f(x)$ is psd $\forall x \in D$

The proof of these properties is left as an exercise to the reader.