Numerical Optimization - Summary

Andy Tràn

October 10, 2022

This will be a very personalised summary for me to use to study for the course Numerical Optimization (AIST 3010). It might be complete, it might not be, it will probably not be. For questions you can refer to andtran@ethz.ch. This summary is based of the lecture notes and should be used as a supplement to the lectures.

Contents

1 Lecture 1 - 05/09/2022: Introduction to Optimization

- **Main lecture**: Monday 12:30 - 2:15, Wednesday 5:30 -6:15 (only ESTR3112)
- **Tutorial lecture**: Wednesday 4:30-5:15
- **Prereq**: Multivariable calculus, linear algebra
- **Course materials:** Homework and exam questions solely based on lecture notes.
- **Email**: farnia@cse.cuhk.edu.hk
- **Office Hour**: Tuesday 2-3 pm, SHB Building, Office 918
- **Learning goals:**
  1. Formulating optimization problems belonging to standard optimization categories for engineering and AI tasks
  2. Applying standard optimization algorithms to solve linear and convex programming tasks
  3. Implementing standard optimization algorithms over Python
- **Grading**: Homework 0.20, midterm 0.30, final .50, participation additionally 0.05

2 Tutorial 1 - 07/09/2022: Introduction to Optimization

Example 1 (Transportation problem) *we want to minimize the total cost of transporting a commodity from m factories to n stores. We have to following constraints:*

- *factory i can supply at most a_i units of the commodity*
- *store j needs at least b_j units of commodity*
- *the cost of shipping from factory i to store j is $c_{i,j}$ per unit*

We get the following system

- *Optimization variables: $x_{i,j}$, the amount of units from fac i to store j*
- *Objective function: $\sum_{i,j} c_{i,j}x_{i,j}$*
- *Constraint function: $\sum_i x_{i,j} \leq a_i$, $\sum_j x_{i,j} \geq b_j$,, $x_{i,j} \geq 0$*

For above problem, there is no analytical solution, only an interative solution.

Example 2 (Manufacturing task) *we want to maximize the profit of producing n products from m raw materials, given that*

- *We have a profit of c_i per unit of Product i*
- *We have b_j available units of raw material j*
- *We need $a_{i,j}$ units of raw material j for manufacturing one unit of i*

We get the following system

- *Optimization variable:* x_i*, amount of units per product i*
- *Objective function:* $\sum_i c_i x_i$
- *Constraint function:* $\sum_i x_i a_{i,j} \leq b_j$ *for all* j

Obviously you have to define what the allowed values are for i and j, which is left as an exercise to the reader.

Example 3 (Sorting task) *given real numbers* $c_1, ..., c_n \in \mathbb{R}$ *we want to find the k smallest numbers*

- *Optimization variable: For every* $1 \leq i \leq n, x_i = \begin{cases} 1 & \text{if } c_i \text{ is among the smallest } k \\ 0 & \text{otherwise} \end{cases}$
- *Objective function:* $\sum_i^n = x_i c_i$
- *Constraint function:* $\sum_i^n x_i = k$*,* $x_i(1 - x_i) = 0$ *for all* i

3 Tutorial 2 - 14/09/2022: Vectors

Vectors

A vector $x = [x_1, ..., x_n]$ is a collection of numbers, arranged in a column or a row, which can be thought of as the coordinates of a point in a n-dimensional space.

- Addition: is defined elementwise, given the dimensions are the same
- Scalar product: element wise multiplication with a scalar.

Note: we by default assume a vector follows a column-representation, with real values. For row format we can just transpose.

Definition 4 (Linearly independent) *Given we have n vectors, we call them l.i. when* $c_1 x_1 + ... + c_n x_n = 0$ *only has one solution, namely all* $c_1, ..., c_n = 0$

Definition 5 (Basis) *For a subspace* $S \in \mathbb{R}^d$*, is a set of l.i. vectors* $B = [x_1, ..., x_m]$ *such that every vector* $x \in S$ *is a linear combination of the vectors in B.*
Standardbasis: defined where $0 \leq i \leq m$*, where i'th coordinate of* $e_i \in S_b$ *is 1 and all the others 0.*

Definition 6 (Euclidean length) $x := \sqrt{x_1^2 + ... + x_n^2}$

Definition 7 (Norm function) *Properties of a norm function*

1. $||x|| \geq 0$*, equal to 0 only when* $x = 0$
2. *For every* $c \in \mathbb{R}, ||cx|| = |c|||x||$
3. *For every* $x, y \in \mathbb{R}^d, ||x + y|| \leq ||x|| + ||y||$

Definition 8 (l_p**-norm)** *For every* $p \geq 1$ *we define* l_p *norm* $|| \cdot ||_p : \mathbb{R}^d \to \mathbb{R}$ *as* $||x||_p = (|x_1|^p + ...|x_n|^p)^{\frac{1}{p}}$

Note: $l_\infty = \max\limits_{1\leq i\leq d} x_i$
Theorem: $l_p - norms$ are decreasing in $p \geq 1$, so

$$1 \leq p \leq q \leq \infty \Rightarrow ||x||_p \geq ||x||_q.$$

Definition 9 (Inner Product) *For every $x = [x_1, ..., x_n], y = [y_1, ..., y_n]$ we define their inner product $< x, y >= \sum_i^n x_i y_i = x^t y$*

4 Lecture 2, Part 2 - 19/09/2022: Vectors and Matrices

Satz 10 *Cauchy-Schwarz Inequality For two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, the following inequality holds:*

$$| < \boldsymbol{x}, \boldsymbol{y} > | \leq ||\boldsymbol{x}||||\boldsymbol{y}||.$$

Satz 11 *Angle between two vectors We can calculate the angle between to vectors namely,*

$$\cos\theta = \frac{< \boldsymbol{x}, \boldsymbol{y} >}{||\boldsymbol{x}||||\boldsymbol{y}||}.$$

They are orthogonal when the scalar product is 0, they are aligned the same or opposite position when the angle is 0 or 180 degrees.

Definition 12 (Mutual Orthogonal Vectors) *a group of vectos $x^{(1)}, ..., x^{(k)}$ For all $i \neq j :< x^{(i)}, x^{(j)>} = 0$*

Proposition 13 (Mutual Orthogonality implies linear independence) *A set of mutually orthogonal vectors are linearly independent*

A collection of vectors are called orthogonal if they have unit euclidean length and are mutually orthogonal...

Vectors in Optimization problems

From the future on we write an optimization problem as follows:

$$\underset{\mathbf{x}}{min} f(\mathbf{x}\text{subject to} g_1(\mathbf{x}) \leq 0 \ldots g_n(\mathbf{x}) \leq 0.$$

Matrix

A matrix $A \in \mathbb{R}^{m\times n}$ is a two-dimensional array of numbers $\begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{m,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{m,2} \\ \cdot & \cdot & \cdot & \\ a_{1,n} & a_{2,n} & \cdots & a_{m,n} \end{bmatrix}$

The product of two matrix $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times t}$ is defined as

$$[AB]_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}.$$

- **The identity matrix** I_n is defined as an $n \times n$ matrix with the diagonal being 1 and all other elements being 0.

- **Matrix vector product** is the same as matrix multiplication, but t being 1
- **Range of Matrix**: $\mathcal{R}(A)$ the subspace of all vectors following from linear combinations of A's combinations
- **Rank of matrix**: $Rank(A)$ the dimension of subspace $\mathcal{R}(A)$
- **Full-rank matrix**: $Rank(A_{m\times n}) \leq min(m,n)$ and a matrix is full rank if $Rank(A_{m\times n}) = min(m,n)$
- **Null-space**: $\mathcal{N}(A)$: the subspace of all vectors which A maps to 0. $\{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$
- **Determinant**: $det(A) = \sum_j^n (-1)^{i+j} a_{i,j} det(A_{(i,j)})$

Rules about determinant for square matrices: $det(A) = \begin{cases} n, & \text{when regular} \\ 0, & \text{otherwise} \end{cases}$

Definition 14 (Invertible matrices) *we call a square matrix $A_{n\times n}$ invertible when a matrix exists such that*

$$A^{-1}A = I_n or AA^{-1}.$$

holds

Proposition 15 (Non-singular matrices) *Following propositions are equivalent.*

1. *A is invertible*
2. *A is non-singular, i.e $det(A) \neq 0$*
3. *A is full rank, i.e $Rank(A) = n$*
4. *A has linearly independent rows or columns*
5. *A has zero null subspace*
6. *A has*

Basic identities

- $(A^T)^{-1} = (A^{-1})^T$
- $(AB)^T = B^T A^T$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $det(A^T) = det(A)$
- $det(A^{-1}) = \frac{1}{det(A)}$
- empty

Eigenvalues and Eigenvectors

Definition 16 (Eigenvector and eigenvalues) *We call a non-zero vector $\boldsymbol{v} \neq \boldsymbol{0}$ an eigenvector of a matrix $A_{n\times n}$ if for a coefficient λ*

$$A\boldsymbol{v} = \lambda\boldsymbol{v}.$$

1. Eigenvalues can be characterized as the solutions for $det(A - \lambda I) = 0$
2. Every $A \in \mathbb{R}^{n\times n}$ has n eigenvalues $\lambda_1, ..., \lambda_n$ counting multiplicities
3. All eigenvalues of a symmetric matrix are real numbers and can be sorted as $\lambda_1 \geq ... \geq \lambda_n$

Proposition 17 (Determinant and eigenvalues) *$det(A) = \lambda_1\lambda_2...\lambda_n$ and so A is invertible $\Leftrightarrow$ all its eigenvalues are non-zero.*

Proposition 18 (Eigenvectors of symmetric matrices) *Supposed v_1 and v_2 are two eigenvectors of a symmetric matri xA for different eigenvalues $\lambda_1 \neq \lambda_2$ then*

$$v_1 \perp v_2 \text{ i.e. } v_1^T v_2 = 0.$$

Satz 19 (Spectral theoren) *For a symmetric matrix A, the set of mutual orthogonal eigenvectors $v_1, ..., v_n$ corresponding to A's eigenvalues $\lambda_1, ..., \lambda_n$. In addition if $v_1, ..., v_n$ are normalized and have unit norm, then*

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T = V\Lambda V^T.$$

, where $V = [v_1|v_2|...|v_n]$, $\Lambda = diag(\lambda_1, ..., \lambda_n)$

Example 20 *Spectral decomposition for the following matrix $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$, calculating $det(A - \lambda I) \overset{!}{=} \lambda_1 = -1, \lambda_2 = 3$ and the eigenvectors result in $v_1 = [\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}]$, $v_2 = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$*

Definition 21 (Positive semi-definite matrix) *A symmetric matrix A is called positive semi-definite (PSD) if for every vector $\boldsymbol{x} \in \mathbb{R}^n$ we have*

$$x^t A x \geq 0.$$

positive definite when the above equality holds strictly for $\boldsymbol{x} \neq 0$

Satz 22 (Eigenvalues and PSD matrix) *Following propositions are equivalent*

- *A is positive semi-definite*
- *A's eigenvalues are all non-negative*
- *For a matrix H we have $A = HH^T$*

Ordering positive smei-definite matrices

Definition 23 (Partial order for Matrices) *for two square matrices A, B we say*

- *$A \succeq B$ if A - B is a positive semidefinite matrix*
- *$A \succ B$ if A - B is a positive definite matrix*
- *$A \preceq B$ if A - B is a negative definite matrix*
- *$A \prec B$ if A - B is a negative definite matrix*

Note what happens when $B = 0$ and note that $A \not\succeq B$ does not imply $A \prec B$

5 Lecture 3 - 26/09/2022: Multivariable Functions and Calculus

Graphs and Epigraphs

- Consider a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$,
- Graph: The set of all points $(x, f(x)) \in \mathbb{R}^{n+1}$, graph $f := \{(x, f(x)) : x \in \mathbb{R}^n\}$
- Epigraph: Set of points on top of f's Graph, $epif := \{(x,t) : x \in \mathbb{R}^n, t \geq f(x)\}$
- Example $f([x_1, x_2]) = x_1^2 + x_2^2$

Contour set

Definition 24 (Contour Set) *For a real value $t \in \mathbb{R}$ the* ***contour*** *of f is the set of vectors mapped to value t.*

$$C_f(t) = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = t\}.$$

Linear and affine functions

$f : \mathbb{R}^n \to \mathbb{R}$ is called a linear function if

1. For every $x, y \in \mathbb{R}^n : f(x+y) = f(x) + f(y)$
2. For every $x \in \mathbb{R}^n, c \in \mathbb{R} : f(cx) = cf(x)$
3. $f(0) = 0$, which follows from point (2).
4. We call f an affine function if $g(x) = f(x) - f(0)$ is a linear function

Satz 25 (Linear functons and inner products) *Suppose that f is a linear function, then there exists a vector $a \in \mathbb{R}^n$ such that:*

$$f(\mathbf{x}) = \mathbf{a^T x} = a_1x_1 + ... + a_nx_n.$$

Hyperplane and Halfspaces

- **Hyperplane:** a contour of a linear function, that is for a vector $\mathbf{a} \in \mathbb{R}^n$ and real number $b \in \mathbb{R}$ we define a hyperplane H as:

$$H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{x} = b\}.$$

- **Halfspace:** the space on top or bottom of a hyperplane, that is for a vector $a \in \mathbb{R}^n$ and real number $b \in \mathbb{R}$ we define halfspace H_- and H_+ as

$$H_+ = \{x \in \mathbb{R}^n : a^t x \leq b\}.$$

, conversely for H_+

Quadratic functions

- **Quadratic function:** a polynomial function in which the highest-degree term is of degree 2

Definition 26 (General quadratic function) *The following provides a general form of a quadratic function in terms of $A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$, and $c \in \mathbb{R}$*

$$f(x) = \frac{1}{2} x^t A x + b^t x + c.$$

Gradients

Definition 27 (Gradient) *if f is differentiable at a point $x \in \mathbb{R}^n$, we define the gradient as*

$$\nabla f(x) = [\frac{df(x)}{dx_1}, ..., \frac{df(x)}{dx_2}].$$

Gradient used for local function approximation

- In many optimization problems, we need to locally approximate the objective and constraint functions with an affine function
- At every point x_0 the gradient $\nabla f(x_0)$ results in an affine function approximating function f around x_0
- The approximation error is defined as $f(x) = f_{app}(x) + \epsilon(x)$

$$f(x) \approx f(x_0) + \nabla(f(x_0))^T (x - x_0).$$

Satz 28 (Error of first-order taylor series expansion) *Suppose that f is differentiable at x_0, then error $\epsilon(x)$ is vanishing near x_0*

$$lim_{x \to x_0} \frac{\epsilon(x)}{||x - x_0||} = 0.$$

Geometric interpretation of gradients

For a unit vector $\mathbf{v}$ and real $\epsilon > 0$

$$f(x_0 + ev) \approx f(x_0) + \epsilon \nabla f(x_0)^T v.$$

- $\nabla f(x_0)^T v > 0$: increases
- $\nabla f(x_0)^T v < 0$: decreases
- $\nabla f(x_0)^T v = 0$: doesn't change

Therefore we have a maximal rate of local variation along the gradient. In contrast there is a zero rate of variation along any direction orthogonal to the gradient. This shows gradients are orthogonal to contour sets.

Hessian

6 Lecture 5 - 03/10/2022: ..

Convex sets

- Linear combination of vectors $x_1, ..., x_k$ with scalar coefficients θ_1, θ_k: $\sum_{i=1}^{k} \theta_i x_i$
- Convex combination is a linear combination with coefficients $\theta_1, ..., \theta_k$ that satisfy: $\theta_i \forall i$ and $\sum_{i=1} k\theta_i = 1$

Definition 29 (Convex set) *$S \subseteq \mathbb{R}^d$ is called a convex set if for every $x, y \in S$, S includes the line segment between x and y*

$$For\ all x, y \in S, \theta \in [0, 1] : \theta x(1 - \theta)y \in S.$$

Proposition 30 *A convex set S is closed to any convex combination of its points, i.e. for every $k \in \mathbb{N}$, non-negative $\theta_1, ..., \theta_k \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$*

- Examples of convex sets
  1. Hyperplanes $\{\mathbf{x} : \mathbf{a^T x} = b\}$
  2. Halfspaces $\{\mathbf{x} : \mathbf{a^T x} \leq b\}$
  3. Norm balls $\{\mathbf{x} : ||\mathbf{x}|| \leq \epsilon\}$
- Convexity preserving operations: convex sets are closed to the following operations:
  - Intersection
  - Affine and inverse-Affine mappings: $f : \mathbb{R}^d \to \mathbb{R}^m, f(x) = Ax + b$
  - Linear fractional functions: $f : \mathbb{R}^d \to \mathbb{R}^m, f(x) = \frac{Ax+b}{c^T x+d}$, where $dom(f) =$

Convex functions:

Satz 31 (Affine functions and convex combinations) *A multivariable function $f : \mathbb{R}^d \to \mathbb{R}$ is an affine function if and only if for all*

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \theta \in [0, 1] : f(\theta x + (1 - \theta)y = \theta f(x) + (1 - \theta)f(y)).$$

Definition 32 (Convex functions) *We call f a convex function if $dom(f)$ is a convex set and $\forall x, y \in dom(f), \theta \in [0, 1 :] f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$. Also we call f concafe if $-f$ is convex.*

Observation: an affine function f is both convex and concafe. If a function is both convex and concave then it's also affine.

Examples of scalar functions:

- Convex functions
  1. powers: $f(x) = x^p$ for $dom(f) = \mathbb{R}_+, p \geq_1$ or $p \leq 0$
  2. Absolute powers: $f(x) = |x|^p$ for $dom(f) = \mathbb{R}, p \geq 1$
  3. Exponential: $f(x) = e^{ax+b}$ for $dom(f)\mathbb{R}$, $a, b \in \mathbb{R}$

- Concave functions:
  1. Powers: $f(x) = x^p$ for $dom(f) = \mathbb{R}_+, 0 \leq p \leq 1$
  2. Logarithm: $f(x) = log(x)$ for $dom(f) = \mathbb{R}_+$

Examples of multivariable functions:

- Convex functions:
  1. Norms: $f(x) = ||x||$ for $dom(x) = \mathbb{R}^d$
  2. Norm powers: $f(x) = ||x||^p$ for $dom(f) = \mathbb{R}^d, p \geq 1$
  3. Log-sum-exp: $f(x) = log($ $sum_{i=1}^d e^{x_i})$ for $dom(f) = \mathbb{R}^d$
- Concave functions:
  1. Entropy: $\sum_{i=1}^d x_i log(\frac{1}{x_i})$ for $dom(f) = \mathbb{R}_+^d$
  2. Geometric mean: $f(x) = (x_1 x_2 ... x_d)^{\frac{1}{d}}$ for $dom(f) = R_+^d$

Methods of identifying convex functions:

1. Verifying the defining inequality (hard!!)
2. Epigraph and sub-level sets
3. Gradients and hessians
4. Operations that preserve convexity

Epigraph and sub-level sets

Proposition 33 (Epigraph of convex functions) *A function f is convex if and only if its epigraph is a convex set*

- We defined the contour (level) set of $C_f(t)$ as the set of inputs mapped to t:
$$C_f(t) = \{x \in dom(f) : f(x) = t\}.$$
- We defined the sublevel set $S_f(t)$ as the set of inputs mapped to below t:
$$S_f(t) = \{x \in dom(f) : f(x) \leq t\}.$$

Proposition 34 *If f is a convex function, $S_f(t)$ will be a convex set for every t.*

Gradient and Hessian of convex functions

Satz 35 (First-order convexity condition) *A differentiable f is convex if and only if its domain is convex and*
$$\forall x, y \in dom(f) : f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

We can see the RHS is the first-order taylor series expanion around x, if f is twice differentiable we can use the second-order Taylor expansion

Satz 36 (Second-order convexity condition) *A twice-differentiable f is convex if and only if its domain is convex and*
$$\forall x \in dom(f) : H_f(x) is\ PSD.$$

Convex quadratic functions

- Consider a quadratic function characteerized by symmetric A,b,c:

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c.$$

- We earlier derived the gradient and Hessian as $\nabla f(x) = Ax + b, H_f(x) = A$
- Second-order condition: f is convex if and only if A is PSD
- Application to the least-squares objective $f(x) = ||Ax + b||^2$

Convexity preserving operations

:

- Positive scalar multiplication
- Addition of convex functions
- Composition with affine functions, i.e. $f(Ax + b)$, if f is convex
- Pointwise maximum: $max(f_1(x), ..., f_k(x))$ is convex if each f_i is convex
- Composition $f(g(x))$ is convex if f,g are convex and f is non-decreasing in every entry

7 Lecture 6: ...