MAT5010 Introduction to Nonlinear Optimization

Zhi-Quan (Tom) Luo

School of Science and Engineering

The Chinese University of Hong Kong, Shenzhen

luozq@cuhk.edu.cn

Organization

Instructor:
Dr. Tom Luo (`luozq@cuhk.edu.cn`), Daoyuan Building, Room 505

Textbook:
D. Bertsekas, *Nonlinear Programming*. Athena Scientific;
Course notes to be distributed in class

Prerequisites:
Calculus, Linear Algebra, Statistics, and Matlab (or C)

Lecture Time:
Monday, 10:00 - 13:00, Sept 5
Wednesdays 8:30 -11:30, Sept 21, 28, Oct 19, 26, Nov 23, 30, Dec 14
Thursdays 12:30 -15:30 , Sept 22, 29, Oct 20, 27 (midterm), Nov 24, Dec 1, 15

Location: Zhixin Building room 201A

Course Objectives and Grading

- to present the basic theory and important algorithms, concentrating on results that are useful in large scale computation and contemporary applications
- to give students the background required to use the methods in their own research or engineering work
- to give students a thorough understanding of how optimization problems are solved, and some experience in solving them

Intended Audience:
Anyone who uses or will use scientific computing or optimization in engineering or related work.

Course Requirement and Grading:
In-class midterm (35%), Final course project (40%),
Problem Sets and Programming Assignments (almost weekly) (25%)

Lecture Topics

- Introduction. Mathematical Review.
- Unconstrained Optimization - Optimality Conditions
- Gradient Methods: Convergence Analysis. Rate of Convergence
- Newton and Gauss–Newton Methods
- Optimization Over a Convex Set; Optimality Conditions
- Feasible Direction Methods
- Constrained Optimization; Lagrange Multipliers. Introduction to Duality
- Penalty Methods. Augmented Lagrangian Methods.
- Alternating Directions Method of Multipliers.
- A General Approximate Gradient Projection (AGP) Framework
- Rate of Convergence Analysis of AGP, Error Bounds
- AGP for Nonconvex, Nonsmooth Problems. Lasso, Group Lasso
- Proximity Operator, Bregman Distance, Proximal Gradient (PG) Method
- Accelerated PG Methods (Nesterov's Method), Incremental Gradient Methods
- Stochastic Approximation Methods

Additional Topics

If time permits, we will also go over

- Conic Programming (SDP, SOCP)
- Conic Duality
- Interior Point Methods for Conic Programming
- SDP Relaxation for Quadratic Optimization, Signal Processing Applications
- Approximation Quality of SDP Relaxation

What is Optimization?

$$\boxed{\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in X \end{array}}$$

where

- $\boldsymbol{x}$ is the decision variable (discrete, continuous)
- f is the objective function or cost function (Differentiable, convex, linear,...)
- X is the feasible region (convex, nonempty, ...)

Key questions:

- When is the problem feasible? Does optimal solution exist?
- How do we determine if a candidate $\boldsymbol{x}$ is an optimal solution?
- How to find an optimal solution numerically? (should do better than exhaustive search)

The Importance and the Process of Optimization

Importance of Optimization:

- fundamental to engineering design and analysis, with ubiquitous applications in many disciplines
- a mathematical foundation for engineering and science, just like probability theory

The Process of Optimization:

- **pre-optimization analysis:** number of variables, number of constraints, continuous/discrete, convex/nonconvex, smooth/nonsmooth, practical engineering requirement on solution accuracy, solution time, current state of the art
- **choose a suitable optimization strategy:** decide on whether to code your own algorithm or use off-the-shelf implementation; decide on initial point and termination criterion
- **post-optimization analysis:** sensitivity, Lagrangian multipliers, feasibility/infeasibility, duality gap
- **troubleshoot and iterate**: repeat the above process

How Do You Use Optimization?

Questions:

- How do I know that the answer from my computer run is the global minimum?
- Which algorithm should I use?
- Is my problem convex?
- Why doesn't my algorithm terminate?
- My cost function is nondifferentiable, should I smooth it?
- Why does my algorithm run into numerical difficulty? It's so slow.
- How do I initialize my algorithm? Which stepsize should I use?

Answer: take this course.

Mathematical Review

- Notations: Sets, Inf, Sup, functions, derivatives, gradients
- Vectors, matrices
- Norms, sequences, limits, continuity
- Mean value theorems
- Implicit function theorem
- Contraction mappings

Notations

- **Sets**:

$$X, \ x \in X, \ X_1 \cap X_2, \ X_1 \cup X_2$$

The set of real (complex) numbers is denoted by $\mathbb{R}$ ($\mathbb{C}$).

- **Inf and Sup**:

The supremum of a nonempty set $X \subset \mathbb{R}$ is the smallest scalar y such that

$$y \geq x, \quad \text{for all } x \in X.$$

Similarly, the infimum of a set $X \subset \mathbb{R}$ is the largest scalar y such that

$$y \leq x, \quad \text{for all } x \in X.$$

If $\sup X \in X$ ($\inf X \in X$), then we say $\sup X = \max X$ ($\inf X = \min X$).

$$\sup\{1/n : n \geq 1\} =? \qquad \inf\{\sin n : n \geq 1\} =?$$

- **Function:**

$$f : X \mapsto Y, \ X \text{ is called the domain, } Y \text{ is called the range}$$

Monotonicity. Inverse function: f^{-1}

Vectors and Subspaces

- **Linear combination**: if $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, then

$$\alpha \boldsymbol{x} + \beta \boldsymbol{y} = (\alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, \ldots, \alpha x_n + \beta y_n)$$

- **Subspaces and linear independence:**

  $S \in \mathbb{R}^n$ is called a subspace if it is closed under linear combination.

  A set of vectors $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^r\}$ are linearly independent if there does not exist a $(\alpha_1, \ldots, \alpha_r) \neq 0$ s.t.

$$\alpha_1 \boldsymbol{x}^1 + \alpha_2 \boldsymbol{x}^2 + \cdots + \alpha_r \boldsymbol{x}^r = \mathbf{0}.$$

  Basis and dimension of a subspace;

  Inner product: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_i x_i y_i$;

  Orthogonality: $\boldsymbol{x} \perp \boldsymbol{y}$ iff $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

  Orthogonal complement of a subspace S:

$$S^\perp := \{\boldsymbol{x} \mid \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0, \ \forall \ \boldsymbol{y} \in S\}.$$

- **Cauchy-Schwartz inequality:**

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$$

Matrices

- For any matrix $\boldsymbol{A}$, we use a_{ij} (or A_{ij}) to denote its (i,j)th entry.
- Matrix addition, multiplication, transpose, symmetric matrices $\boldsymbol{A} = \boldsymbol{A}'$.

$$[\boldsymbol{AB}]' = \boldsymbol{B}'\boldsymbol{A}', \ \ \boldsymbol{AB} \neq \boldsymbol{BA}.$$

- Let $\boldsymbol{A}$ be a matrix of size $m \times n$. Range of $\boldsymbol{A}$: $R(\boldsymbol{A})$, null space of $\boldsymbol{A}$: $N(\boldsymbol{A})$.

$$R(\boldsymbol{A}) = N^{\perp}(\boldsymbol{A})$$

  Rank of $\boldsymbol{A}$ $\operatorname{rank}(\boldsymbol{A})$. Full rank matrix $\boldsymbol{A}$: $\operatorname{rank}(\boldsymbol{A}) = \min\{m, n\}$.

- Square matrix $(m = n)$, identity matrix $\boldsymbol{I}$, determinant $\det(\boldsymbol{A})$, inverse $\boldsymbol{A}^{-1}$.

$$\boldsymbol{A}^{-1} \text{ exists} \quad \text{iff} \quad \det(\boldsymbol{A}) \neq 0$$

- Let $X \subset \mathbb{R}^n$, $\boldsymbol{A}$ be a matrix of size $m \times n$, then the image of X under $\boldsymbol{A}$ is

$$\boldsymbol{A}X = \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in X\}$$

- **Inner product:**

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{Tr}(\boldsymbol{AB}') = \sum_{i,j} A_{ij} B_{ij}$$

Square Matrices

- **Useful identities:** $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$, $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
- **Orthonormal matrices:** $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$.
- **(Complex) Eigenvalue λ:** $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ for some $\boldsymbol{x} \neq \mathbf{0}$.
- **Spectral radius:** $\rho(\boldsymbol{A}) = \max_i\{|\lambda_i| : \lambda_i$ is an eigenvalue of $\boldsymbol{A}\}$.
- **Eigen-decomposition of a symmetric matrix:**

$$\boldsymbol{A} = \boldsymbol{P}'\boldsymbol{\Lambda}\boldsymbol{P},$$

  where $\boldsymbol{P}$ is orthonormal, $\boldsymbol{\Lambda}$ is diagonal and real.
- **Positive (semi-) definite matrix:** $\boldsymbol{A} \succeq \mathbf{0}$.

$$\boldsymbol{A} \succeq \mathbf{0},\ \boldsymbol{B} \succeq \mathbf{0} \Rightarrow \boldsymbol{A} + \boldsymbol{B} \succeq \mathbf{0}.$$

- **Square root $\boldsymbol{A}^{1/2}$:** $\boldsymbol{A}^{1/2} := \boldsymbol{P}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{P}'$, where $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ is the eigen-decomposition of $\boldsymbol{A} \succeq \mathbf{0}$.
- **A useful property:** $\boldsymbol{A} \succeq \mathbf{0} \iff \langle \boldsymbol{A}, \boldsymbol{B} \rangle \geq 0,$ for all $\boldsymbol{B} \succeq \mathbf{0}$.

Matrices

- **Singular values** $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$**:** σ_i^2 is an eigenvalue of $\boldsymbol{A}\boldsymbol{A}'$.
- **Condition number:** $\kappa(\boldsymbol{A}) = \sigma_1/\sigma_n$.
- **Singular values decomposition:** $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$, with $\boldsymbol{U}$ and $\boldsymbol{V}$ orthonormal, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \succeq \boldsymbol{0}$ diagonal.
- **Norms:**

  Frobenious norm: $\|\boldsymbol{A}\|_F = \left(\sum_{i,j} |A_{ij}|^2\right)^{1/2} = \left(\sum_i \sigma_i^2\right)^{1/2}$

  Nuclear norm: $\|\boldsymbol{A}\|_* = \sum_i \sigma_i$

  Matrix 2-norm: $\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_i \sigma_i$

  and $\|\boldsymbol{A}\|^2 = \|\boldsymbol{A}\boldsymbol{A}'\| = \|\boldsymbol{A}'\boldsymbol{A}\|$.
- **Useful inequalities:**

$$\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|, \quad \|\boldsymbol{A}\|_* \geq \|\boldsymbol{A}\|_F \geq \|\boldsymbol{A}\|_2 \geq \rho(\boldsymbol{A}).$$

- **Cauchy-Schwartz inequality:**

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle \leq \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F$$

Derivatives and Mean Value Theorems

Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a continuously twice differentiable.

- **Derivative:**

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} := \lim_{t \to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{e}_i) - f(\boldsymbol{x})}{t}$$

  where $\boldsymbol{e}_i$ is the ith unit vector of $\mathbb{R}^n$.

- **Gradient vector:**

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_n} \right)^T$$

- **Hessian matrix:**

$$\nabla^2 f = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_i \partial x_j} \right]$$

- **Taylor expansion:**

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})'(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{y})'\nabla^2 f(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{y}) + o(\|\boldsymbol{x} - \boldsymbol{y}\|^2)$$

- **Mean value theorem:** there exist $\boldsymbol{\xi}, \boldsymbol{\eta}$ in the line segment connecting $\boldsymbol{x}$ and $\boldsymbol{y}$ such that

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \nabla f(\boldsymbol{\xi})'(\boldsymbol{y} - \boldsymbol{x})$$

and

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})'(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{y})'\nabla^2 f(\boldsymbol{\eta})(\boldsymbol{x} - \boldsymbol{y})$$

- **Jacobian matrix:** For a vector valued continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, define the Jacobian matrix

$$\nabla f(\boldsymbol{x}) = [\nabla f_1(\boldsymbol{x}), \nabla f_2(\boldsymbol{x}), \ldots, \nabla f_m(\boldsymbol{x})]\,.$$

- **Chain rule:** for $f : \mathbb{R}^k \mapsto \mathbb{R}^m$ and $g : \mathbb{R}^m \mapsto \mathbb{R}^n$, let $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$. Then

$$\nabla h(\boldsymbol{x}) = \nabla f(\boldsymbol{x})\nabla g(f(\boldsymbol{x})).$$

For example, we have

$$\nabla(f(\boldsymbol{A}\boldsymbol{x})) = \boldsymbol{A}'\nabla f(\boldsymbol{A}\boldsymbol{x}), \;\; \nabla^2 f(\boldsymbol{A}\boldsymbol{x})) = \boldsymbol{A}'\nabla^2 f(\boldsymbol{A}\boldsymbol{x})\boldsymbol{A}.$$

Implicit Function Theorem

- Let $G(x,y) : \mathbb{R}^2 \mapsto \mathbb{R}$ be a continuously differentiable function and $P = (x^0, y^0)$ is a point such that
$$\frac{\partial G}{\partial y}(x^0, y^0) \neq 0$$
If $G(x,y) = G(x^0, y^0)$, then y may be expressed as a function of x in a neighborhood containing P; i.e., there exists a differentiable function $y = g(x)$ such that
$$y^0 = g(x^0), \quad \text{and} \quad G(x, g(x)) = 0 \text{ for all } x \text{ close to } x^0.$$

- Let $G(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{R}^n \mapsto \mathbb{R}^m$ be a continuously differentiable mapping where $\boldsymbol{y} \in \mathbb{R}^m$ and
$$\nabla_{\boldsymbol{y}} G = \left[\frac{\partial G_j}{\partial y_i}\right]_{m \times m} \quad \text{is nonsingular at } P = (\boldsymbol{x}^0, \boldsymbol{y}^0)$$
If $G(\boldsymbol{x}, \boldsymbol{y}) = G(\boldsymbol{x}^0, \boldsymbol{y}^0)$, then $\boldsymbol{y}$ may be expressed as a function of $\boldsymbol{x}$ locally around P; i.e., there exists a differentiable mapping over $\boldsymbol{y} = g(\boldsymbol{x})$ such that
$$\boldsymbol{y}^0 = g(\boldsymbol{x}^0), \quad \text{and} \quad G(\boldsymbol{x}, g(\boldsymbol{x})) = \boldsymbol{0} \text{ for all } \boldsymbol{x} \text{ close to } \boldsymbol{x}^0.$$

Contraction Mappings

- **Lipschitzian Property:** $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ satisfies

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq \gamma \|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall\ \boldsymbol{x}, \boldsymbol{y}$$

  γ is called the Lipschitz constant.

- If $\gamma \leq 1$, then f is called a *non-expansive mapping*.

- If $\gamma < 1$, then f is called a *contraction mapping*

- **Fixed point theorem:** If f is a contraction, then the iterated function sequence

$$\boldsymbol{x},\ f(\boldsymbol{x}),\ f(f(\boldsymbol{x})),\ f(f(f(\boldsymbol{x}))), \ldots$$

  converges to a unique fixed point $\boldsymbol{x}^*$ (independent of $\boldsymbol{x}$) satisfying

$$\boldsymbol{x}^* = f(\boldsymbol{x}^*).$$