Optimization Methods

Instructor: Dr. Liang Zheng

School of Sciences

Harbin Institute of Technology (Shenzhen)

Today's Topics

- About this course
  - Course Outline
  - Brief Intro. of the Instructor
  - Reference Books
- Chapter I: Mathematical Modeling
- Chapter II: Fundamental Math

About This Course - Outline

- **Prerequisites:**

  Advanced Mathematics and Linear Algebra

- **Grading System:**

  Homework Assignments: 30%

  Final Exam: 70%

- **Closed books and notes during exam.**

  ONE formula A4 sheet is allowed during exam.

Brief Introduction of the Instructor

- Liang Zheng
  - Ph.D., Mechanical Engineering

    University of Wisconsin-Madison
  - Assistant Dean & Associate Professor, School of Sciences, HIT Shenzhen.
  - Three years experience of teaching at University of Wisconsin-Madison. Courses taught include Statics, Dynamics, Mechanics of Materials.
  - Main Research Fields: Microlithography, Fatigue of innovative materials, Finite Element Analysis.

Contact Information

- Dr. Liang Zheng

  Room 624, Building G

  Email: icon_lzheng@hit.edu.cn

- Teaching Assistants:

  Moring Section:

  Xiao-Bing Wu

  Tel: 15235112179

  Email: 3086899634@qq.com

  Rang-Fei Ren

  Tel: 18833974197

  Email: 2441184616@qq.com

  Afternoon Section:

  Yi-Tian Qiu

  Tel: 15940069067

  Email: q15940069067@163.com

  Ke-Ze Rong

  Tel: 18268677358

  Email: 690803667@qq.com

Reference books

- R. Fletcher, “Practical Method of Optimization”, 2nd Edition, Wiley, 1987.
- J. Nocedal and S. J. Wright, “Numerical Optimization”, 1999.
- J. Frederic Bonnans, “Numerical Optimization: Theoretical and Practical Aspects”, Springer, 2002.

Chapter 1: Intro. to Optimization

- **Examples of Optimization**:
  - Nature optimizes:.
    - Potential energy
    - Rays of light
    - Fish swimming
    - ?
  - People optimize:.
    - Airline companies: cost
    - Investors: profit
    - Manufacturers: efficiency
    - ?

Chapter 1: Introduction (Cont'd)

- **Objective: a quantitative measurement of the performance of the system under study.**
  - **The objective could be any quantity such as profit, time, or combination of quantities that can be represented by a single number.**
- **The objective (function) depends on certain characteristics of the system, called variables.**
- **In most cases, constraints are often existed and applied to the variables, in some ways.**

Chapter 1: Introduction (Cont'd)

- **Mathematical modeling: the process of identifying the objective, variables, and constraints for a given problem is known as mathematical modeling.**
- **Construction of an appropriate model is the first step - in many cases, the most important step - in the optimization process.**
- **Once the model has been formulated, an optimization algorithm can be used to find its solution.**

1.1 Mathematical Formulation of Optimization

- **Optimization is the process of minimizing or maximizing of a function subject to constraints on its variables.**
- **The following notations are used in this class:**

***X*: is the vector of variables, also called unknowns;**

***f*: is the objective function, a function of X that needs to be maximized or minimized;**

***C*: is the vector of constraints that the variables must satisfy. It is a vector function of variables X.**

1.1 Optimization (Cont'd)

Then the Optimization problem can be written as:

$$\min_{x \in R^n} f(X)$$

s. t. (subject to):

$$c(X) = \begin{cases} g_u(x_1, x_2, \cdots x_n) \le 0 & u = 1,2,\cdots p \\ h_v(x_1, x_2, \cdots x_n) = 0 & v = 1,2,\cdots m \end{cases}$$

$X = (x_1, x_2, \cdots, x_n)^T$ **: the vector of variables;**

R^n **: n-dimensional real Euclidean space;**

g_u , h_v **: constraints of** X**.**

1.2 Examples

Example 1-1:

A no-lid box is made by cutting a small square piece with same size at each corner of a 3 m × 3 m square wood lamella. Determine the dimensions of this small square piece to maximize the volume of the box.

1.2 Examples (Cont'd)

Example 1-2:

A factory is developing two products, named A and B. The Table lists the supplies and the demands of materials, man hours, electricity, as well as the profits of making each product. Determine the production plan of these two products to maximize the daily profit.

Product	Material/kg	Hour/h	Electricity kwh	Profit /$
A	9	3	4	60
B	4	10	5	120
Supply	360	300	200	?

1.3 Further Discussion

- **Optimization problems can be divided into linear problems and nonlinear problems.**
- **When the objective function and constraint functions are all the linear functions of the variables, the optimization problems are called “linear”; otherwise, the problems are considered as “nonlinear”.**
- **In general, optimization problems in the fields of economy, management, or production planning are linear; problems in engineering fields are nonlinear.**

1.4 Mathematical Model

- In general, the mathematical model is composed of the objective function, variables and constraint functions.
- Variables can be classified as independent and dependent variables. They can be also classified as continuous and discrete variables. During the formulation of the model, only independent variables are considered.
- For complex problems with many variables, the variables of less importance are treated as constants first. After the simplified model is formulated and the problem is solved, those variables can be treated back as variables to improve the accuracy of final solutions.

1.4 Mathematical Model (II)

- **Most of the optimization methods and algorithms are applied for only continuous variables. For problems involved discrete variables, the general way is to assume these variables as continuous, then solve the problems with optimization methods. At last, the optimal solutions can be discretized to acquire the final solutions of the problems.**

1.4 Mathematical Model (III)

- **Domain**: the set of values assigned to the independent variables of a function.
- Assuming there are n variables $x_1, x_2, \ldots, x_n$, then a n-dimensional real space can be formulated, called **Euclidean Space**, denoting as R^n.
- **Constraints**: also called constraint functions. For the variable vector $X=[x_1, x_2, \ldots, x_n]^T$, the following inequalities or equalities:

$$g_u(X) \leq 0 \qquad (u = 1, 2, \cdots, p)$$
$$h_v(X) = 0 \qquad (v = 1, 2, \cdots, m)$$

are called the constraint functions.

1.4 Mathematical Model (IV)

- **Feasible Region**: An region enclosed by multiple constraint boundaries. All points in this region satisfy all constraint functions. It can be expressed as a set denoting as:

$$\& = \{X \mid g_u(X) \le 0,\ h_v(X) = 0 \quad (u=0, 1, \ldots, p;\ v=0, 1, \ldots, m)\}$$

- **Objective function**: is the quantitative criterion for assessing the optimization process. In this course, only optimization problems with single objective are discussed.
- **Contours** of the function: a set of points for which the function has a constant value. The contours explicitly show the variation of function values and can be used to determine the optimal solutions for the problems.

1.5 The Graphic Method

- **The Graphic Method is used to solve the optimization problems via the visualization of the mathematical model and data. It is mainly used to solve simple 2-D problems.**
- **The general procedures are listed as follows:**

1) Determine the domain of the variables;

2) Identify the feasible region of the solutions;

3) Plot a few contours of the objective function to search the descent direction of the function.

4) Find the optimal solutions of the problem。

1.5 The Graphic Method (II)

- **Example 1-3: Solve the following problem using the graphic method:**

$$\min f(X) = x_1^2 + x_2^2 - 4x_1 + 4$$

$$s.t. \quad g_1(X) = -x_1 + x_2 - 2 \le 0$$

$$g_2(X) = x_1^2 - x_2 + 1 \le 0$$

$$g_3(X) = -x_1 \le 0$$

1.6 Descent Methods / Algorithms

- **In optimization problems, the solutions are usually obtained via the numerical methods, not the analytical methods. The numerical methods in optimization theories are called the iteration methods (algorithms).**
- **Since the objective functions are defined to be minimized in the mathematical models, these algorithms are often called descent algorithms.**

1.6 Descent Methods/Algorithms (II)

- An algorithm is referred to as a **descent method** if it generates a sequence of points X_0, $X_1, \cdots, X_k, X_{k+1}, \cdots$, such that:

  $f(X_0) > f(X_1) > \cdots > f(X_k) > f(X_{k+1}) > \cdots$ for all k.

  and the limit of this sequence is the minimal value of the objective function, i.e.:

$$\lim_{k \to \infty} X_k = X^*$$

1.6.1 General Formulation of Descent Algorithms

- **In the optimization methods, the iteration points are generated by the following iteration formula:**

$$X = X_k + \alpha S_k$$

where, X_k is the current iteration point, S_k is the search direction and α is the step length.

- **Often, the new iteration point is chosen as the point having the minimal function value along the direction S_k, i.e.:**

$$X_{k+1} = X_k + \alpha_k S_k$$

where, α_k is the optimal step length.

1.6.2 General Procedures for Descent Algorithms

1. Given an initial estimate X_0 and a sufficiently small convergence number $\varepsilon > 0$. Set $k = 0$.
2. Select the search direction S_k.
3. Determine the optimal step length α_k to minimize the function value $f(X_k+\alpha S_k)$. Obtain the new iteration point X_{k+1} via $X_{k+1} = X_k+\alpha_k S_k$.
4. If X_{k+1} satisfies the convergence criterion, i.e., the termination criterion, then X_{k+1} is the optimal solution; otherwise set k = k+1, i.e., choose X_{k+1} as the new iteration point, go to Step 2.

1.6.3 Convergence and Termination Criteria of Descent Algorithms

- **When the function values of the iteration points in the sequence generated by the algorithm strictly reduce, and ultimately reach the minimal value of the optimization problem, then it is said that this algorithm is of convergence. The speed that the point sequence is approaching to the minimal value is called the speed of convergence of the algorithm.**
- **A good optimization algorithm is expected to have not only good convergence but a fast speed of convergence.**

1.6.4 Rate of Convergence

- **Rate of Convergence of an algorithm:**

  For a constant $\sigma \in (0,1)$ which is unrelated with the iteration, there exists an integer $\beta \geq 1$ such that:

$$\lim_{k\to\infty} \frac{\left\| X_{k+1} - X^* \right\|}{\left\| X_k - X^* \right\|^{\beta}} = \sigma$$

  $\beta=1$: Linear rate of convergence;

  $1<\beta<2$: Superlinear rate of convergence;

  $\beta=2$: Quadratic rate of convergence.

- **Quadratic convergence is faster than superlinear convergence; Superlinear convergence is faster than linear convergence.**

1.6.5 Termination Criteria

1. Criterion based on Point Distance:

$$\left\| X_{k+1} - X_k \right\| \le \varepsilon$$

2. Criterion based on Function Value Difference:

$$\left| f(X_{k+1}) - f(X_k) \right| \le \varepsilon \qquad \left| \frac{f(X_{k+1}) - f(X_k)}{f(X_k)} \right| \le \varepsilon$$

3. Criterion based on Gradient:

$$\left\| \nabla f(X_{k+1}) \right\| \le \varepsilon$$

1.6.6 Classification of Optimization Algorithms

Problem	Characteristics	Features of Algorithm	Algorithms / Methods
Linear	linear functions of variables	Vortex Conversion	Simplex Method
Nonlinear	Unconstrained Optimization	One-dimensional Search	Golden Section Search Method, Fibonacci Search Method, Quadratic Interpolation Method, Cubic Interpolation Method.
		Use information of derivatives	Steepest Descent Method, Newton Method, Quasi-Newton Method, Conjugate Gradient Method.
		Don’t use info.	Powell Method
	Constrained Optimization	Solve directly	Feasible Direction Method
		Solve indirectly	Penalty Function Method, SQP Method, etc.

Chapter 2: Fundamentals of Optimization

- ***n*-dimensional Vector**

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad X = [x_1, x_2, \cdots, x_n]^T$$

- ***m* × *n* Matrix**

$$A = A_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

2.2 Directional Derivative and Gradient

- **Partial derivative**: **the rate of variation at the point X_k along the coordinate axis x_j, for a function $f(X)$ with multiple variables $x_1, x_2, \ldots$; denoted as $\partial f(X_k)/\partial x_j$.**
- **Directional derivative**: **the rate of variation at the point X_k along any direction S; denoted as $\partial f(X_k)/\partial S$.**
- **For an arbitrary function with n variables $x_1, x_2, \ldots, x_n$, the directional derivative can be expressed as:**

$$\frac{\partial f(X_k)}{\partial S}=\frac{\partial f(X_k)}{\partial x_1}\cos\alpha_1+\frac{\partial f(X_k)}{\partial x_2}\cos\alpha_2+\cdots\frac{\partial f(X_k)}{\partial x_n}\cos\alpha_n$$

$$=\begin{bmatrix}\frac{\partial f(X_k)}{\partial x_1}, & \frac{\partial f(X_k)}{\partial x_2}, & \cdots & \frac{\partial f(X_k)}{\partial x_n}\end{bmatrix}\begin{bmatrix}\cos\alpha_1\\ \cos\alpha_2\\ \vdots\\ \cos\alpha_n\end{bmatrix}=[\nabla f(X_k)]^T S_0 \qquad \textbf{(2-1)}$$

2.2 Derivative and Gradient (II)

where: $S_0 = \left[\cos\alpha_1, \cos\alpha_2, \cdots, \cos\alpha_n\right]^T$

is the unit vector of the direction *S*;

$$\nabla f(X_k) = gradf(X_k) = \left[\frac{\partial f(X_k)}{\partial x_1}, \quad \frac{\partial f(X_k)}{\partial x_2}, \quad \cdots \quad \frac{\partial f(X_k)}{\partial x_n}\right]^T$$

is called the gradient of *f*(x) at point X*k*;

From (2-1), it can be shown that:

$$\frac{\partial f(X_k)}{\partial S} = \left[\nabla f(X_k)\right]^T S_0 = \left\|\nabla f(X_k)\right\| \cdot \left\|S_0\right\| \cdot \cos\left\langle \nabla f(X_k), S_0 \right\rangle$$

$$= \left\|\nabla f(X_k)\right\| \cdot 1 \cdot \cos\left\langle \nabla f(X_k), S_0 \right\rangle = \left\|\nabla f(X_k)\right\| \cos\left\langle \nabla f(X_k), S_0 \right\rangle$$

2.2 Derivative and Gradient (III)

Characteristics of the gradient:

1) The gradient of a function at one point is a vector comprised of all the first-order partial derivatives at this point. It is the comprehensive description of the variation of the function at this point.

2) The direction of the gradient is the direction the function value increases most rapidly; the direction of the negative gradient is the direction the function value decreases most rapidly.

2.2 Derivative and Gradient (IV)

锐角

Characteristics of the gradient (Cont'd):

钝角

3) Directions having acute angles with the gradient are the directions that function values get increased; Directions having obtuse angles with the gradient are the directions that function values get decreased.

4) The gradient describes the local variation of the function values at one point. At areas outside the neighborhood of this point, the variation of the function values can NOT be determined or described by this gradient.

2.2 Example

Ex.2-1: Determine and plot the gradients of the function $f(X) = (x_1-2)^2 + (x_2-1)^2$ at points $X_1=[5, 5]^T$ and $X_2 = [6, 4]^T$.

2.3 Taylor Series / Expansion

- **Taylor Expansion is a representation of a function as an infinite sum of terms calculated from the values of its derivatives at a single point.**
- **Taylor series of a function with one variable is:**

$$f(x) = f(x_k) + f'(x_k)\cdot(x - x_k) + \frac{1}{2} f''(x_k)\cdot(x - x_k)^2 + \cdots + R_n$$

where R_n is the remainder.

- **Taylor series of a function with multiple variables:**

$$f(X) \approx f(X_k) + \left[\nabla f(X_k)\right]^T (X - X_k) + \frac{1}{2}(X - X_k)^T \nabla^2 f(X_k)(X - X_k)$$

This formula is called the quadratic approximation of the function f(X): 二阶近似

2.3 Taylor Expansion (II)

where, $\nabla^2 f(X_k)$ is a matrix composed of the second-order partial derivatives of the function at point X_k, called the Hessian (matrix) of $f(X)$ at X_k and denoted as $H(X_k)$:

$$H(X_k) = \nabla^2 f(X_k) = \begin{bmatrix} \dfrac{\partial^2 f(X_k)}{\partial x_1^2} & \dfrac{\partial^2 f(X_k)}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(X_k)}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(X_k)}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(X_k)}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(X_k)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 f(X_k)}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(X_k)}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(X_k)}{\partial x_n^2} \end{bmatrix}$$

2.4 Positive Definite Quadratic Function (PDQF)

- **Quadratic functions are the simplest nonlinear functions. They are of significance in the theory of optimization.**
- **Using Taylor expansion, quadratic functions can be generally written in the form of vectors as:**

$$f(X) = \frac{1}{2} X^T H X + B^T X + C \qquad \textbf{(2-2)}$$

where:

B is a constant vector, as the gradient of the function;

H is a *n*×*n* constant matrix, as the Hessian of the function;

$X^T HX$ is called the quadratic form and H is called the quadratic form matrix. 二次型

2.4 PDQF (II)

- Matrices can be classified as **positive definite**, **negative definite**, and **indefinite**.
- For an arbitrary non-zero vector p:

  if $p^T \mathrm{H} p > 0$, then H is **positive definite**;

  if $p^T \mathrm{H} p \geq 0$, then H is **positive semi-definite**;

  if $p^T \mathrm{H} p < 0$, then H is **negative definite**;

  if $p^T \mathrm{H} p \leq 0$, then H is **negative semi-definite**;

  if $p^T \mathrm{H} p$ depends on p, then H is **indefinite**.
- If the Hessian $H(X)$ in (2-2) is positive definite, then $f(\mathrm{X})$ is a positive definite quadratic function.

2.4 PDQF (III)

- **PDQFs have the following characteristics:**

1. The contours of PDQFs are a set of concentric ellipses (ellipsoids); The center of this set is the minimum point of the PDQF.

2. For the non-positive-definite QFs, the contours near the minimum point are approximately as ellipses (ellipsoids); the contours become irregular when they are away from the minimum point; illustrated as follows:

2.5 Convexity

- **The term "Convex" can be applied to both sets and functions.**

1. Convex Set

- **A non-null set $S \in R^n$ is a convex set if the straight line segment connecting any two points in S lies entirely inside S. That is, for any two points $x, y \in S$, there has $\alpha x+(1-\alpha)y \in S$ for arbitrary $\alpha \in [0, 1]$.**

2. Convex Function

- **f is a convex function if its domain S is a convex set and for any two points $x, y \in S$, the straight line connecting these two points lies above the graph of the function. That is:**

$$f[\alpha x+(1-\alpha)y] \le \alpha f(x)+(1-\alpha)f(y), \text{ for all } \alpha \in [0, 1].$$

2.6 Minimizer / Maximizer

- **As we have known, the concepts of Minimization and Maximization can be converted into each other by changing the sign in the inequalities.**

1. Global Minimizer: A point x^* is a global minimizer if $f(x^*) \leq f(x)$ for all $x \in S$.

2. Local Minimizer: A point x^* is a local minimizer if there is a neighborhood N of x^* such that $f(x^*) \leq f(x)$ for all $x \in N$.

3. Strong Minimizer: A point x^* is a strong minimizer if $f(x^*) < f(x)$ instead of $f(x^*) \leq f(x)$.

2.7 Descent Direction

Definition:

- **f(X) is differentiable at and p is a known non-zero vector. If there is $[\nabla f(X_k)]^T p < 0$, then p is called the Descent Direction of f(X) at X_k.**
- **From the definition, one can see:**

$$\nabla f(X_k)^T p = \|\nabla f(X_k)\| \cdot \|p\| \cos\theta < 0 \quad \Rightarrow \quad \theta > \pi / 2$$

- **It is easy to observe that $p = -\nabla f(X_k)$ is the descent direction of f(X) at X_k. In fact, $p = -\nabla f(X_k)$ is not only a descent direction, but the Steepest Descent Direction of f(X) at X_k.**

2.8 Extremum Conditions (I)

For Functions with only one variable:

- **The necessary condition for $f(x)$ with one variable to reach the extremum at x_k is that the first-order derivative at this point equals zero; the sufficient condition is that the corresponding second-order derivative is not zero, i.e.:**

$$f'(x_k) = 0 \qquad \& \qquad f''(x_k) \neq 0$$

when $f''(x_k) > 0$, the function reaches the minimum value; when $f''(x_k) < 0$, the function achieves the maximum value.

2.8 Extremum Conditions (II)

For Functions with multiple variables:

Taylor's Theorem:

1) If *f* is once continuously differentiable in an open neighborhood of X* and *p* is a non-zero vector, then:

$$f(X^*+p)=f(X^*)+\nabla f(X^*+\lambda p)^T p \qquad for\ \ \lambda\in(0,1)$$

2) If *f* is twice continuously differentiable, then:

$$f(X^*+p)=f(X^*)+\nabla f(X^*)^T p+\frac{1}{2}p^T\nabla^2 f(X^*+\lambda p)^T p$$

Lemma:

- ***f* is differentiable at X_k and *p* is a descent direction of *f*(X) at X_k, then there is $\lambda > 0$, s. t. $f(X_k+\lambda p) < f(X_k)$.**

2.8 Extremum Conditions (III)

First-order Necessary Condition:

- **If X^* is a local minimizer and f is continuously differentiable in an open neighborhood of X^*, then there is $\nabla f(X^*) = 0$.**

X^* is called a stationary point if $\nabla f(X^*) = 0$. Hence, any local minimizer must be a stationary point.

Second-order Necessary Condition:

- **If X^* is a local minimizer and $\nabla^2 f$ is continuous in an open neighborhood of X^*, then there has $\nabla f(X^*) = 0$ and $\nabla^2 f(X^*)$ is positive semi-definite.**

2.8 Extremum Conditions (IV)

Second-order Sufficient Condition:

- **Suppose that $\nabla^2 f$ is continuous in an open neighborhood of X^* and there has $\nabla f(X^*) = 0$ and $\nabla^2 f(X^*)$ is positive definite, then X^* is a strict local minimizer of f.**

Theorem:

- **When f is convex, any local minimizer X^* is a global minimizer of f; If in addition f is differentiable, then any stationary point X^* is a global minimizer of f.**

2.8 Summary on ECs (V)

- For a function $f(X)$ with multiple variables, the necessary condition for $f(X)$ to achieve the extremum at X^* is that the gradient of the function at this point equals to zero, i.e. $\nabla f(X^*) = 0$; the sufficient condition is that the Hessian $\nabla^2 f(X^*)$ is either positive or negative definite. That is:

i) $\nabla f(X^*)=0$ and $\nabla^2 f(X^*)$ positive definite $\Rightarrow$ Minimum Point

ii) $\nabla f(X^*)=0$ and $\nabla^2 f(X^*)$ negative definite $\Rightarrow$ Maximum Point

iii) $\nabla f(X^*)=0$ and $\nabla^2 f(X^*)$ indefinite $\Rightarrow$ Not a Extremum Point

2.8 Example

Ex. 2-2: Use Taylor's Expansion to convert the following function (at $X_1 = [0, 2]^T$) to a linear function and a quadratic function, respectively:

$$f(X) = x_1^3 - x_2^3 + 3x_1^2 + 2x_2^2 - 8x_1$$