Lecture 1: Introduction to optimization

DSA5103 Optimization algorithms for data modelling

Yangjing Zhang

12-Jan-2023

NUS

Contacts

- Lecturer: Yangjing Zhang
- Email: yj.zhang [AT] nus.edu.sg/zhangyangjing [AT] u.nus.edu
- Office: S17-05-16
- Feel free to contact me through emails with proper salutation:
  - ▷ Dr. Zhang (✓formal)
  - ▷ Yangjing (✓informal)
  - ▷ Hi/Hi Professor (✗)
- Programming tools: I will use Matlab for demonstration. Feel free to use Python/R/Julia

About

Real data and problems

Optimization model

e.g.,
linear/logistic regression
support vector machine
principal componet analysis
Gaussian graphical models
... ...

Optimization algorithms

e.g.,
(stochastic) gradient descent methods
block coordinate descent method
ADMM
... ...

Probably will not cover deep learning, reinforcement learning ...

Assessment

Final grade based on:

- Final exam: 40%
  - ▷ Interpretation, computation
  - ▷ Implementation of some algorithms
  - ▷ No proof required
- Homework (#4 sets): 60%

Today's content[1]

1. Motivating examples
2. Graphical illustration
3. General nonlinear programming
4. Basic calculus and linear algebra
5. Necessary/Sufficient optimality conditions
6. Convex sets/functions

[1] Most of the contents are adopted from [1]

Motivating examples

Simple examples

Example 1

If K units of capital and L units of labor are used, a company can produce KL units of a product. Capital can be purchased at $\$4$ per unit and labor can be purchased at $\$1$ per unit. A total of $\$8000$ is available to purchase capital and labor. How can the firm maximize the quantity of the product manufactured?

Simple examples

Example 1

If K units of capital and L units of labor are used, a company can produce KL units of a product. Capital can be purchased at $\$4$ per unit and labor can be purchased at $\$1$ per unit. A total of $\$8000$ is available to purchase capital and labor. How can the firm maximize the quantity of the product manufactured?

Solution. Let $K =$ unites of capital purchased, and $L =$ units of labor purchased. The problem to solve is

$$\begin{aligned} &\text{maximize} && KL \\ &\text{subject to} && 4K + L \leq 8000, \quad K \geq 0, \quad L \geq 0. \end{aligned}$$

Simple examples

Example 2

It costs a company $\$c$ to produce a unit of a product. If the company charges $\$p$ per unit, and the customers demand $D(p)$ units, what price should the company charge to maximize its profit?

Simple examples

Example 2

It costs a company $\$c$ to produce a unit of a product. If the company charges $\$p$ per unit, and the customers demand $D(p)$ units, what price should the company charge to maximize its profit?

Solution. The firm's decision variable is p, and the profit is $(p-c)D(p)$. The problem to solve is

$$\begin{aligned} \text{maximize} \quad & (p-c)D(p) \\ \text{subject to} \quad & p \geq 0. \end{aligned}$$

Linear regression

- Predict the price for house with area 3500

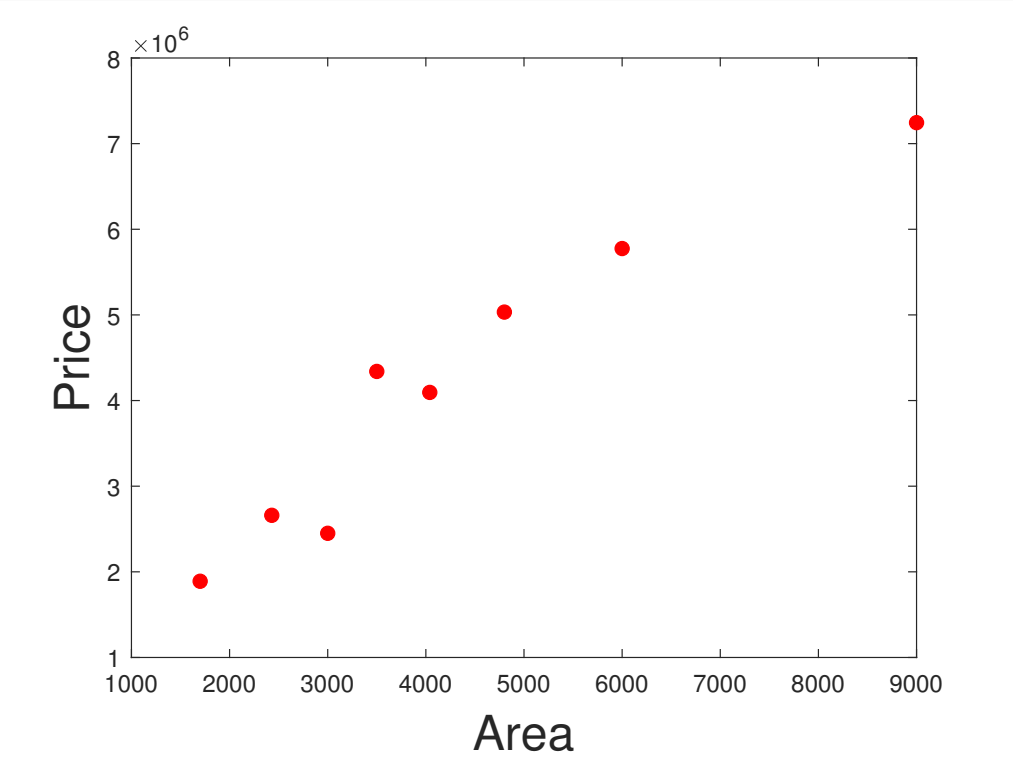


Figure 1: Housing prices data[2]

[2] https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Linear regression

area	#bedrooms	#bathrooms	stories	$\cdots$	price
7420	4	2	2	$\cdots$	13300000
8960	4	4	4	$\cdots$	12250000
		$\vdots$			

- Predictor/feature vector $a_1 = (7420, 4, 2, 2, \dots)^T \in \mathbb{R}^p$
- Response/target $b_1 = 13300000$
- p: #features. n: #samples
- Fit a linear model $b = x^T a + \alpha$

Linear regression

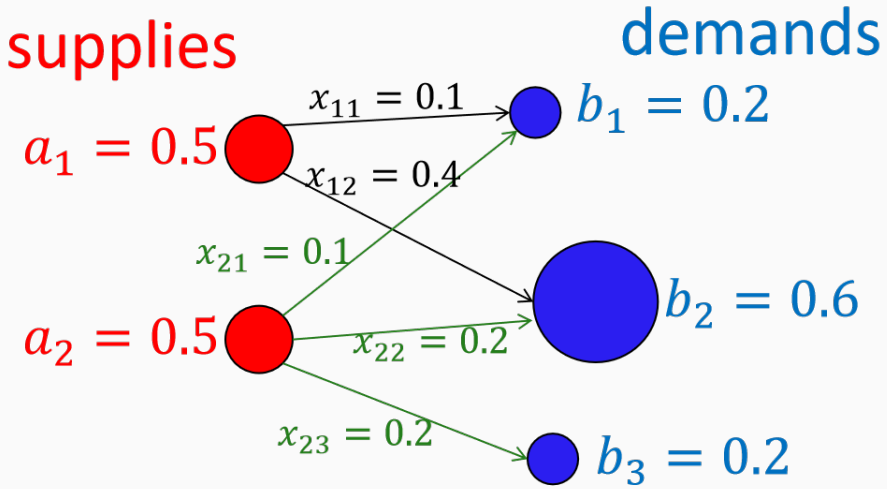
- Given data (a_i, b_i), $i = 1, \ldots, n$ where $a_i \in \mathbb{R}^p$ is the predictor vector and $b_i \in \mathbb{R}$ is the response.
- Fit a linear model $b = x^T a + \alpha$.
- The goal is to find $x \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$

$$\text{minimize}_{x,\alpha} \quad \frac{1}{2} \sum_{i=1}^{n} (b_i - x^T a_i - \alpha)^2$$

Optimal transportation (OT)

Example: OT

s supplies $a_1, \ldots, a_s$ of goods must be transport to meet t demands $b_1, \ldots, b_t$ of customers. The cost of transporting one unit of ith good to jth customer is c_{ij}. What is the optimal transportation plan (x_{ij} unit of ith good transported to jth customer) to minimize the cost?



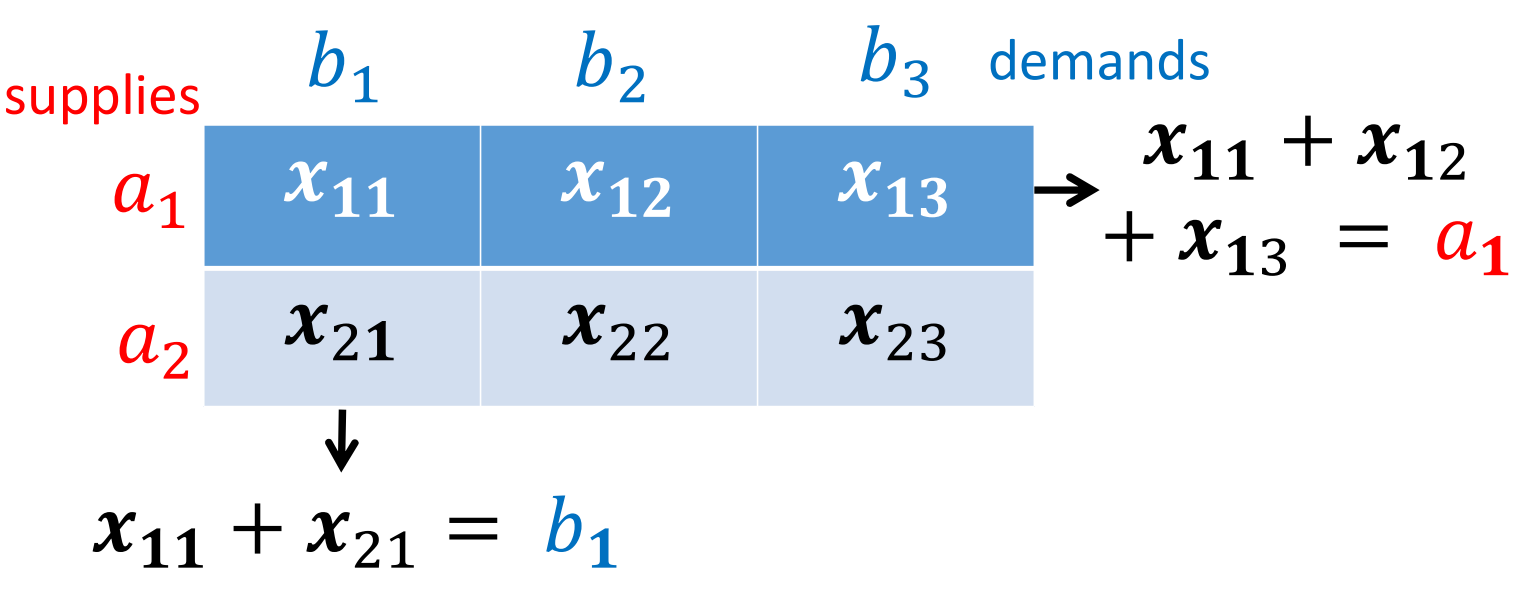
- For example, the cost c_{ij} may be determined by the distance
- A feasible transportation plan $\begin{pmatrix} 0.1 & 0.4 & 0 \\ 0.1 & 0.2 & 0.2 \end{pmatrix}$
- [Exercise] Construct another feasible transportation plan

Optimal transportation (OT)

Cost:

$$\sum c_{ij} x_{ij}$$

Constraints:



Optimal transportation(OT)

Solution. Suppose x_{ij} unit of ith good is transported to jth customer. The cost is $\sum_{ij} c_{ij}x_{ij}$. The problem to solve is

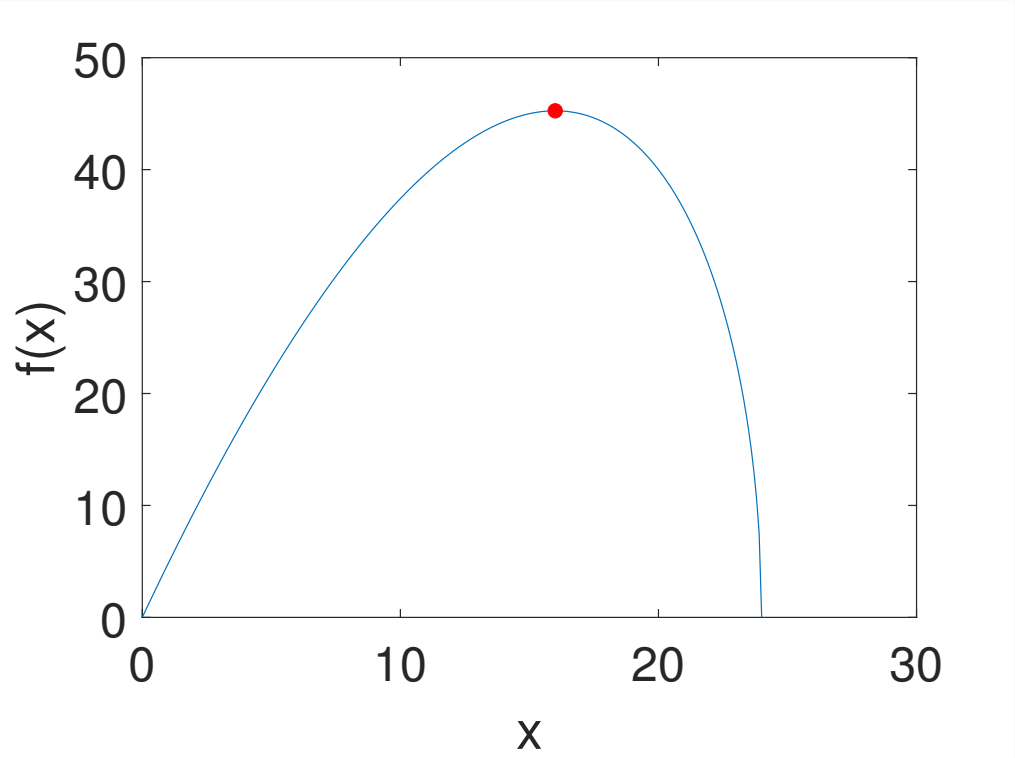
$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{s}\sum_{j=1}^{t} c_{ij}x_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{t} x_{ij} = a_i,\ i \in [s] = \{1, \ldots, s\} \\ & \sum_{i=1}^{s} x_{ij} = b_j,\ j \in [t] \\ & x_{ij} \geq 0,\ i \in [s],\ j \in [t]. \end{aligned}$$

- Such a problem arises in computing the Wasserstein distance (a hot topic in machine learning and statistics) to measure the similarity of two discrete probability distributions.
- In this case, $\sum_{i=1}^{s} a_i = 1$ and $\sum_{i=1}^{s} b_j = 1$ correspond to two discrete probability distributions.

Graphical illustration

Graphical illustration in 1-dimension

$$\text{maximize} \quad f(x) = x\sqrt{24 - x}$$



$x = 16$ achieves the maximal value $16\sqrt{24 - 16}$

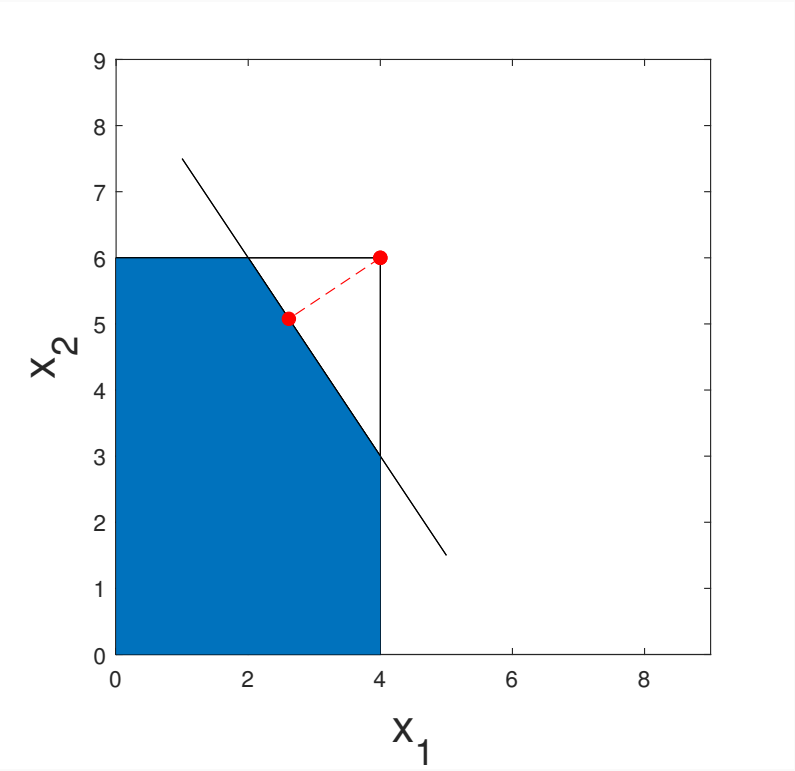
Graphical illustration in 2-dimension

$$
\begin{aligned}
\text{minimize} \quad & f(x) = (x_1 - 4)^2 + (x_2 - 6)^2 \\
\text{subject to} \quad & x_1 \leq 4 \\
& x_2 \leq 6 \\
& 3x_1 + 2x_2 \leq 18 \\
& x_1 \geq 0 \\
& x_2 \geq 0
\end{aligned}
$$

- Note that $(x_1 - 4)^2 + (x_2 - 6)^2$ measures the distance between two points $(x_1, x_2)^T$ and $(4, 6)^T$
- Graphically, we want to find the shortest distance of $(x_1, x_2)^T$ from $(4, 6)^T$

Graphical illustration in 2-dimension

$$
\begin{aligned}
\text{minimize} \quad & (x_1 - 4)^2 + (x_2 - 6)^2 \\
\text{subject to} \quad & x_1 \leq 4 \\
& x_2 \leq 6 \\
& 3x_1 + 2x_2 \leq 18 \\
& x_1 \geq 0 \\
& x_2 \geq 0
\end{aligned}
$$



$(x_1, x_2) = (\frac{34}{13}, \frac{66}{13})^T$ achieves the minimal value $(\frac{34}{13} - 4)^2 + (\frac{66}{13} - 6)^2$

Graphical illustration

- Good for intuition
- Give solutions visually
- Only work in 1-dimension or 2-dimension
- For higher dimensional problems, one has to rely on algebraic conditions to find (as well as to verify) the solutions

General nonlinear programming

Nonlinear programming

A general nonlinear programming problem (NLP) is

$$\begin{array}{ll} \text{minimize (or maximize)} & f(x) \\ \text{subject to} & g_i(x) = 0,\ i \in [m] \\ & h_j(x) \leq 0,\ j \in [p] \end{array}$$

- f, g_i, and h_j are (possibly nonlinear) functions of variable $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$
- f objective function, $g_i(x) = 0$ equality constraint, $h_j(x) \leq 0$ inequality constraint
- It suffices to discuss minimization problems since

$$\text{minimize}\, f(x) \iff \text{maximize}\, -f(x)$$

Terminology and notation

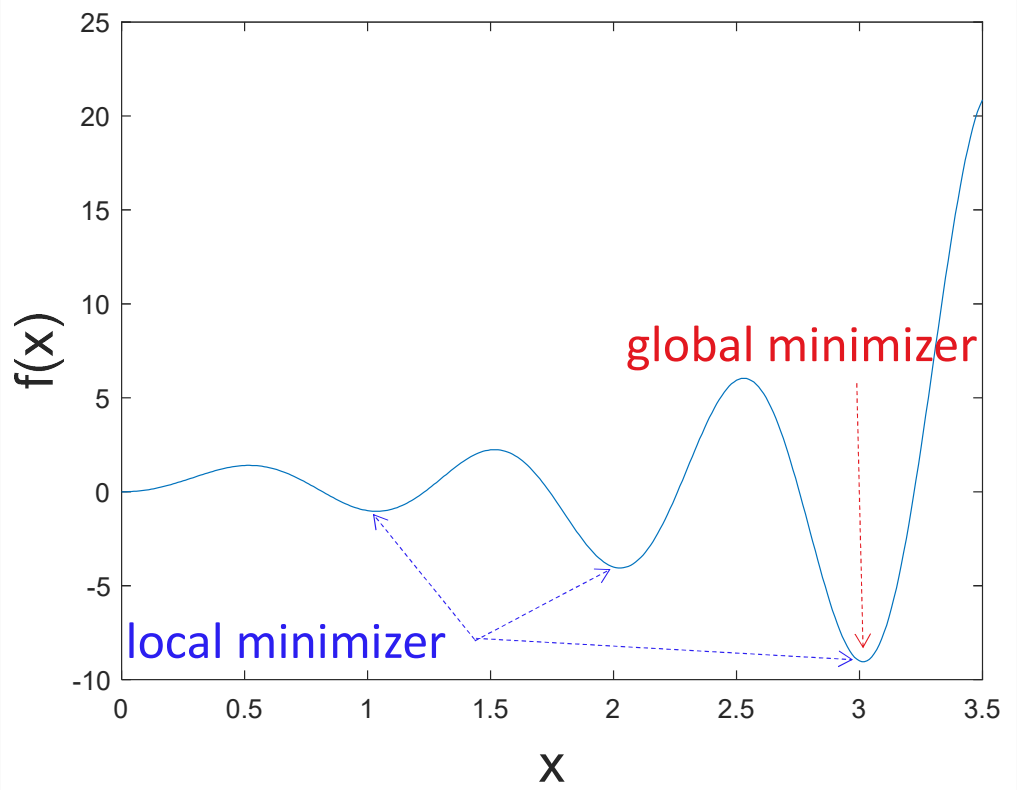
- Vectors in $\mathbb{R}^n$ will always be column vectors, unless otherwise specified.
  - ▷ row vector (x_1, x_2, x_3)
  - ▷ column vector $(x_1; x_2; x_3)$ or $(x_1, x_2, x_3)^T$
- The feasible set for an NLP is the set

$$S = \{x \in \mathbb{R}^n \mid g_1(x) = 0, \ldots, g_m(x) = 0,\ h_1(x) \leq 0, \ldots, h_p(x) \leq 0\}$$

- A point in the feasible set is a feasible solution or a feasible point; otherwise, it is an infeasible solution or infeasible point
- When there is no constraint, $S = \mathbb{R}^n$, we say the NLP is unconstrained

Local/global minimizer

$$\begin{aligned} \text{minimize} \quad & f(x) = (\sin(\pi x))^2 e^x - x^2 \\ \text{subject to} \quad & 0 \le x \le 3.5 \end{aligned}$$



- Local/global minimizer may not be unique
- A global minimizer is a local minimizer. However, the converse is not true in general

Local/global minimizer

Definition (Local minimizer and global minimizer)

Let S be the feasible set. Define $B_\epsilon(y) = \{x \in \mathbb{R}^n \mid \|x - y\| < \epsilon\}$ to be the open ball with center y and radius ϵ. ($\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$)

1. A point $x^* \in S$ is said to be a local minimizer of f if there exists $\epsilon > 0$ such that

$$f(x^*) \leq f(x) \quad \forall\, x \in S \cap B_\epsilon(x^*)$$

2. A point $x^* \in S$ is said to be a global minimizer of f if there exists $\epsilon > 0$ such that

$$f(x^*) \leq f(x) \quad \forall\, x \in S$$

Basic calculus and linear algebra

Interior point

Definition (Interior point)

Let $S \subseteq \mathbb{R}^n$ be a nonempty set. A point $x \in S$ is called an interior point of S if there exists $\epsilon > 0$ such that

$$B_\epsilon(x) \subseteq S$$

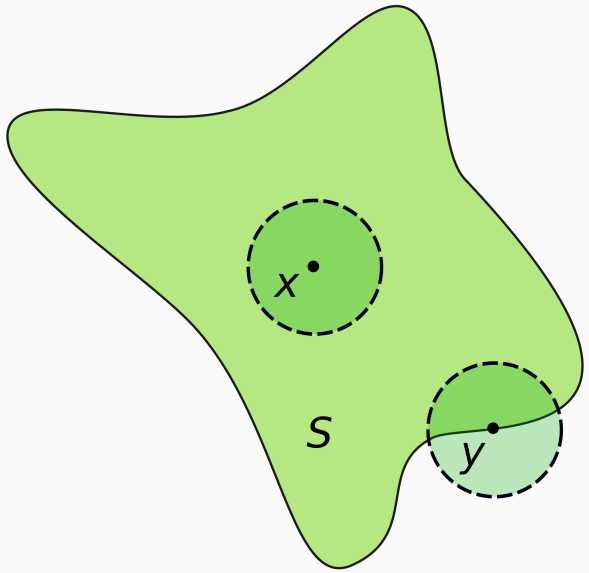


Figure 2: The point x is an interior point of S. The point y is on the boundary of S. Image from internet

Gradient vector

Definition (Gradient vector)

Let $S \subseteq \mathbb{R}^n$ be a nonempty set. Suppose $f : S \to \mathbb{R}$, and x is an interior point of S such that f is differentiable at x. Then the gradient vector of f at x is

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

- At x^*, the value $f(x)$ decreases most rapidly along the direction $-\nabla f(x^*)$. On the other hand, it increases most rapidly along the direction $\nabla f(x^*)$.

Hessian matrix

Definition (Hessian matrix)

Let $S \subseteq \mathbb{R}^n$ be a nonempty set. Suppose $f : S \to \mathbb{R}$, and x is an interior point of S such that f has second order partial derivatives at x. Then the Hessian of f at x is the $n \times n$ matrix

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

- The ij-entry of $H_f(x)$ is $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$. In general, $H_f(x)$ is not symmetric. However, if f has continuous second order derivatives, then the Hessian matrix is symmetric.

Example

Compute the gradient and Hessian

(1) $f(x) = x^3 + 2x + e^x$, $x \in \mathbb{R}$.

(2) $f(x) = x_1^3 + 2x_1x_2 + x_2^2$, $x = (x_1; x_2) \in \mathbb{R}^2$.

Example

Compute the gradient and Hessian

(1) $f(x) = x^3 + 2x + e^x$, $x \in \mathbb{R}$.

(2) $f(x) = x_1^3 + 2x_1x_2 + x_2^2$, $x = (x_1; x_2) \in \mathbb{R}^2$.

Solution (1).

$$\nabla f(x) = f'(x) = 3x^2 + 2 + e^x, \quad H_f(x) = f''(x) = 6x + e^x.$$

Solution (2).

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 3x_1^2 + 2x_2 \\ 2x_1 + 2x_2 \end{bmatrix}$$

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 6x_1 & 2 \\ 2 & 2 \end{bmatrix}$$

Positive (semi)definite matrices

Definition (Positive (semi)definite)

Let A be a real symmetric $n \times n$ matrix.

(1) A is said to be positive semidefinite if $x^T Ax \geq 0, \forall\, x \in \mathbb{R}^n$.

(2) A is said to be positive definite if $x^T Ax > 0, \forall\, x \neq 0$.

To check whether a matrix is positive semidefinite by definition is not an easy task in general. The following eigenvalue test may be useful to check for definiteness of a matrix:

- A is positive semidefinite if and only if every eigenvalue of A is nonnegative.
- A is positive definite if and only if every eigenvalue of A is positive.

Necessary/Sufficient optimality conditions

Unconstrained NLP

Consider an unconstrained NLP

$$\text{minimize} \quad f(x)$$

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is nonlinear and differentiable. A point x^* is called a stationary point of f if $\nabla f(x^*) = 0$.

Next, we investigate some optimality conditions of a local minimizer

- Necessary (optimality) condition
- Sufficient (optimality) condition

Necessary/sufficient condition

Necessary condition If x^* is a local minimizer of f, then

(1) x^* is a stationary point, i.e., $\nabla f(x^*) = 0$

(2) $H_f(x^*)$ is positive semidefinite

Sufficient condition If

(1) x^* is a stationary point, i.e., $\nabla f(x^*) = 0$ and

(2) $H_f(x^*)$ is positive definite,

then x^* is a local minimizer of f.

- The necessary condition allows us to confine our search for global minimizers within the set of stationary points
- The sufficient condition can help us verify that a point is indeed a local minimizer

Example

Example

Let $f(x) = x_1^2 + x_2^2 - 2x_2$, $x = (x_1; x_2) \in \mathbb{R}^2$.

(1) Find a stationary point of f.

(2) Verify that the stationary point is a local minimizer (by checking the sufficient condition).

Example

Example

Let $f(x) = x_1^2 + x_2^2 - 2x_2$, $x = (x_1; x_2) \in \mathbb{R}^2$.

(1) Find a stationary point of f.

(2) Verify that the stationary point is a local minimizer (by checking the sufficient condition).

Solution. We let

$$\nabla f(x^*) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x^*) \\ \frac{\partial f}{\partial x_2}(x^*) \end{bmatrix} = \begin{bmatrix} 2x_1^* \\ 2x_2^* - 2 \end{bmatrix} = 0 \Rightarrow x^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Thus $x^* = (0, 1)^T$ is a stationary point of f. Next, we verify that

$$H_f(x^*) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x^*) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x^*) \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

is positive definite (since the eigenvalues of $H_f(x^*)$ are 2, 2, which are positive).

Example

Example

Let $f(x) = x_1^2 x_2 + x_1 x_2^3 - x_1 x_2$, $x = (x_1; x_2) \in \mathbb{R}^2$.

(1) Verify that $x = (0; 1)$ a stationary point of f.

(2) Is $x = (0; 1)$ a local minimizer?

Example

Example

Let $f(x) = x_1^2 x_2 + x_1 x_2^3 - x_1 x_2$, $x = (x_1; x_2) \in \mathbb{R}^2$.

(1) Verify that $x = (0; 1)$ a stationary point of f.

(2) Is $x = (0; 1)$ a local minimizer?

Solution.

$$\nabla f(x) = \begin{bmatrix} 2x_1 x_2 + x_2^3 - x_2 \\ x_1^2 + 3x_1 x_2^2 - x_1 \end{bmatrix} \Rightarrow \nabla f(0; 1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Thus $x = (1; 0)$ is a stationary point of f. Next,

$$H_f(x^*) = \begin{bmatrix} 2x_2 & 2x_1 + 3x_2^2 - 1 \\ 2x_1 + 3x_2^2 - 1 & 6x_1 x_2 \end{bmatrix} \Rightarrow H_f(0; 1) = \begin{bmatrix} 2 & 2 \\ 2 & 0 \end{bmatrix}$$

is not positive semidefinite (since the eigenvalues of H_f are $1 + \sqrt{5}$, $1 - \sqrt{5} < 0$). The necessary condition fails, thus $x = (0; 1)$ is not a local minimizer.

Convex sets/functions

Convex sets

Definition Convex set

A set $D \in \mathbb{R}^n$ is said to be a convex set if for any two points x and y in D, the line segment joining x and y also lies in D. That is,

$$x, y \in D \Rightarrow \lambda x + (1 - \lambda) y \in D \quad \forall\, \lambda \in [0, 1].$$

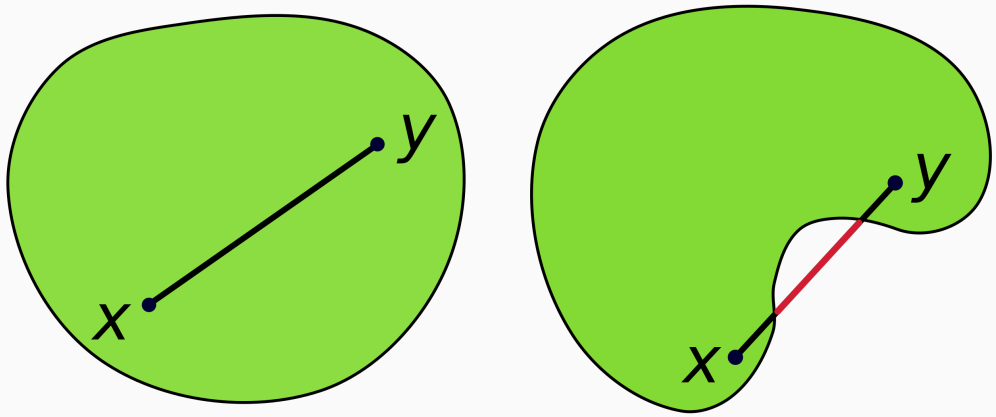


Figure 3: Left: convex. Right: non-convex. Image from internet

(Strictly) convex functions

Definition (Strictly) convex function

Let $D \subseteq \mathbb{R}^n$ be a convex set. Consider a function $f : D \to \mathbb{R}$.

(1) The function f is said to be convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \, \forall \, x, y \in D, \, \lambda \in [0,1].$$

(2) The function f is said to be strictly convex if

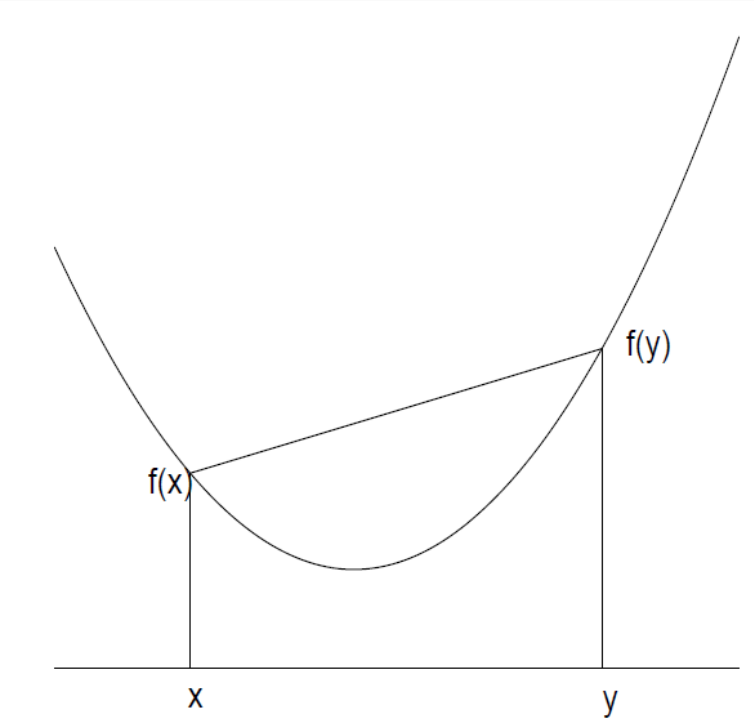
$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y),$$

for all distinct $x, y \in D$, $\lambda \in (0,1)$.

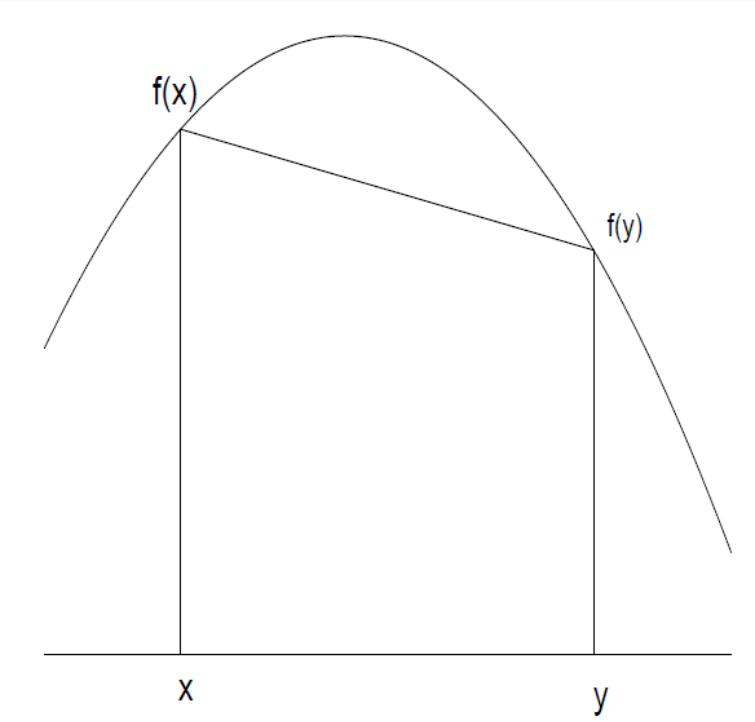
- Note that $-f$ is convex $\iff$ f is concave

Geometrical interpretation

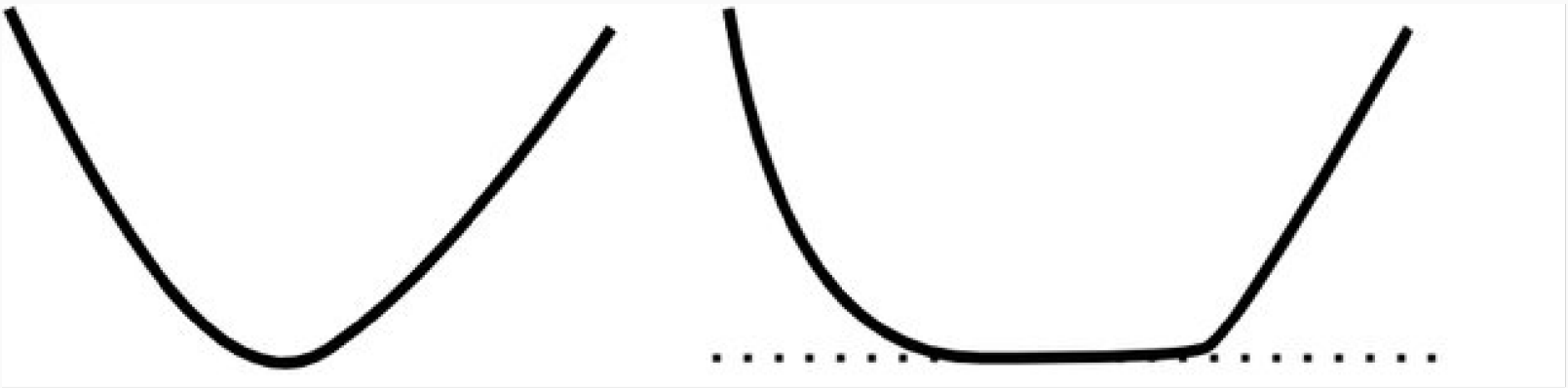
For a convex function f, the line segment joining $f(x)$ and $f(y)$ lies above the graph of f in $[x, y]$.



For a concave function f, the line segment joining $f(x)$ and $f(y)$ lies below the graph of f in $[x, y]$.

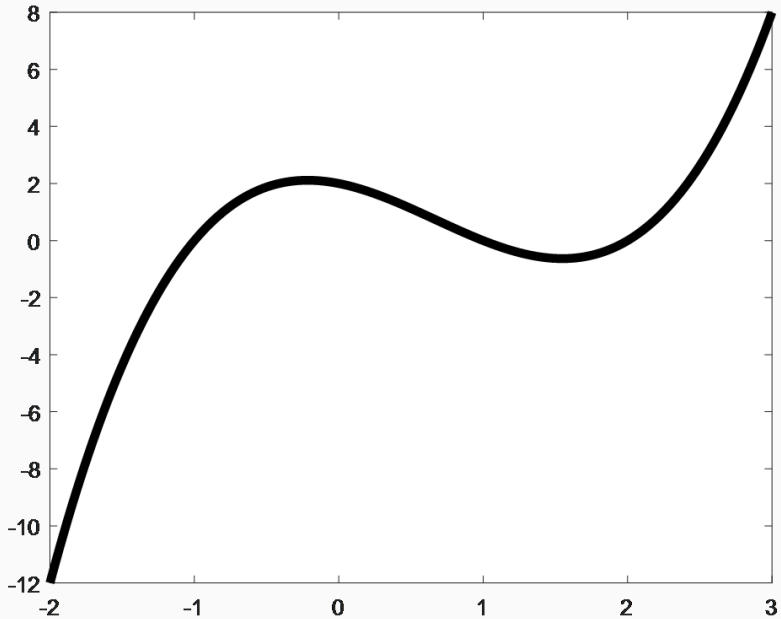


Geometrical interpretation



strictly convex

convex (but not strictly convex)



$f(x) = (x-1)(x-2)(x+1)$ is neither convex or concave on $[-2,3]$

Benefit of convexity

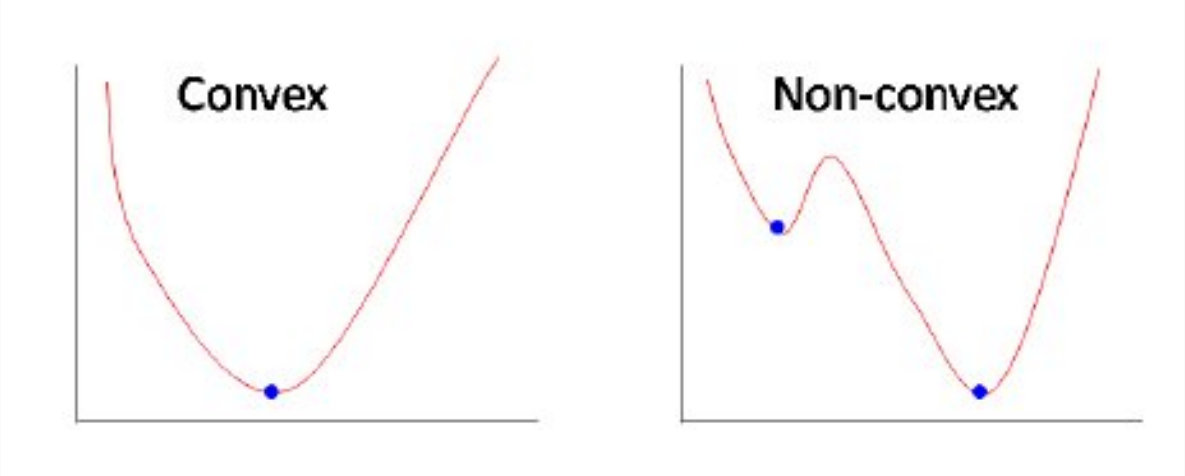
Consider an unconstrained convex NLP

$$\text{minimize} \quad f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex. It holds that

(1) any local minimizer is a global minimizer.

(2) if f is strictly convex, then the global minimizer is unique.



References i


K.-C. Toh.
MA5268: Theory and algorithms for nonlinear optimization.