Optimization

Lecture 2: Some mathematical overheads

Parin Chaipunya
KMUTT
└Mathematics @ Faculty of Science
└The Joint Graduate School of Energy and Environments

Areas of research:

- Multi-agent optimization: Bilevel programs, Game theory
- Optimization modeling: mainly focused on energy and environmental applications

parin.cha@kmutt.ac.th
parinchaipunya.com
github.com/parinchaipunya

Last update: Januaray 2026

Table of contents

Section 1

Vector-matrix calculus

Subsection 1

Vectors and matrices

Vector notation

Consider the set $\mathbb{R}^n$.

- Any vector $x \in \mathbb{R}^n$ is represented either by a **tuple** or a **column vector**, *i.e.*

$$x = (x_1, \ldots, x_n) = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} .$$

- A vector $x \in \mathbb{R}^n$ is interpreted either as a **point** or a **direction** in the space $\mathbb{R}^n$.
- A vector x above is not the same as its transpose, which is

$$x^t = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} .$$

  It is not uncommon to write $(\mathbb{R}^n)^*$ as the set of these row vectors.
- We write

$$0 = (0, \ldots, 0)$$

  to denote the **zero vector**.
- The dimension of the zero vector 0 can be ambigiuous.

Norms and distances

The norms and distances here are considered in the Euclidean sense.

- The **norm** of a vector $x \in \mathbb{R}^n$ is defined by

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{x^t x}\,.$$

- The **distance** between two vectors $x, y \in \mathbb{R}^n$ is defined by

$$\begin{aligned} d(x,y) &= \|x - y\| \\ &= \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} \\ &= \sqrt{(x-y)^t (x-y)}\,. \end{aligned}$$

We would like to note that both $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ and $d(\cdot, *) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ are continuous functions.

Norms and distances

The following theorem concludes some useful properties of the norm and distance functions defined previously.

Theorem

For any $x, y, z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, the following holds.

- $\|x\| = 0 \iff x = 0$.
- $\|\lambda x\| = |\lambda| \|x\|$.
- $\|x + y\| \leq \|x\| + \|y\|$.
- $d(x, y) = 0 \iff x = y$.
- $d(x, y) = d(y, x)$.
- $d(x, z) \leq d(x, y) + d(y, z)$.

Matrix notation

- Consider the set $\mathbb{R}^{m \times n}$ of $m \times n$ matrices.
- We also write $0_{m \times n}$ or just 0 to denote the $m \times n$ matrix whose elements are all zero.
- We write $I_{n \times n}$ or just I to denote the $n \times n$ identity matrix.

Let $M \in \mathbb{R}^{n \times n}$ be a symmetric matrix.

- We say that M is **positive semidefinite** (or **PSD**) if all eigenvalues of M are ≥ 0.
- We say that M is **positive definite** (or **PD**) if all eigenvalues of M are > 0.
- We say that M is **negative semidefinite** (or **NSD**) if all eigenvalues of M are ≤ 0.
- We say that M is **negative definite** (or **ND**) if all eigenvalues of M are < 0.
- We say that M is **indefinite** if it has both positive and negative eigenvalues.

It is useful to note that $M = A^t A$ is always PSD.

Subsection 2

Gradients and Hessians

Gradients

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ (a function of n real variables).

- The **gradient** of f at $x^0 \in \mathbb{R}^n$ is a vector in $\mathbb{R}^n$ defined by

$$\nabla f(x^0) = \left(\frac{\partial f}{\partial x_1}(x^0), \ldots, \frac{\partial f}{\partial x_n}(x^0) \right) = \left(f_{x_1}(x^0), \ldots, f_{x_n}(x^0) \right).$$

- The gradient ∇f is viewed as a function $\mathbb{R}^n \to \mathbb{R}^n$.
- A function f is said to be **continuously differentiable** if ∇f exists and is a continuous function.
- The vector $\nabla f(x^0)$ is the direction at which f increases the fastest. This is called the **steepest ascent** property.

Differentials of vector functions

Consider a vector function $F : \mathbb{R}^n \to \mathbb{R}^m$.

- We may regard it as a vector of scalar functions $F_1, \ldots, F_m : \mathbb{R}^n \to \mathbb{R}$ for which

$$F(x) = (F_1(x), \ldots, F_m(x)) .$$

- We then define the differential of F at $x^0 \in \mathbb{R}^n$ as the **Jacobian matrix** as follows

$$JF(x^0) = \begin{bmatrix} (F_1)_{x_1}(x^0) & \ldots & (F_1)_{x_n}(x^0) \\ \vdots & \ddots & \vdots \\ (F_m)_{x_1}(x^0) & \ldots & (F_m)_{x_n}(x^0) \end{bmatrix}_{m \times n} = \begin{bmatrix} \nabla F_1(x^0)^t \\ \ldots \\ \nabla F_m(x^0)^t \end{bmatrix}$$

- For $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$Jf(x) = \nabla f(x)^t .$$

The chain rule

Theorem 1 (The chain rule.)
Let $F : \mathbb{R}^n \to \mathbb{R}^m$ *and* $G : \mathbb{R}^k \to \mathbb{R}^n$. *Then*

$$J(F \circ G)(x) = JF(G(x))JG(x) \qquad (x \in \mathbb{R}^k)$$

provided that all the Jacobians exist.

Hessians

- The **Hessian** of f at $x^0 \in \mathbb{R}^n$ is a $n \times n$ matrix defined by

$$\nabla^2 f(x^0) = J(\nabla f)(x^0) = \begin{bmatrix} f_{x_1x_1}(x^0) & \dots & f_{x_1x_n}(x^0) \\ f_{x_2x_1}(x^0) & \dots & f_{x_2x_n}(x^0) \\ \vdots & \ddots & \vdots \\ f_{x_nx_1}(x^0) & \dots & f_{x_nx_n}(x^0) \end{bmatrix}.$$

- The Hessian $\nabla^2 f$ is viewed as a function $\mathbb{R}^n \to \mathbb{R}^{n \times n}$.
- A function f is said to be **twice continuously differentiable** if $\nabla^2 f$ exists and is a continuous function.
- The Hessian $\nabla^2 f$ is the **local curvature matrix**. Along the direction $d \in \mathbb{R}^n$, the curvature of the graph of f at x^0 is measured by $d^t \nabla^2 f(x^0) d$.

Subsection 3

Calculus of affine and quadratic functions

Calculus of an affine function

- An **affine function** is a function $f : \mathbb{R}^n \to \mathbb{R}$ of the form

$$f(x) = f(x_1, \ldots, x_n) = a_0 + a_1 x_1 + \cdots + a_n x_n = a_0 + a^t x \,. \tag{1}$$

- If $a_0 = 0$, that is

$$f(x) = a^t x \,,$$

  then we say that f is **linear**.
- A linear function f always has the property that $f(0) = 0$.
- An affine function is a translation of a linear function.
- If f is an affine function defined in (1), then

$$\nabla f(x) = a \quad \text{and} \quad \nabla^2 f(x) = 0_{n \times n} \,.$$

Calculus of an affine operator

An **affine operator** $T : \mathbb{R}^n \to \mathbb{R}^m$ is defined as

$$T(x) = Mx + q\,,$$

for some fixed $M \in \mathbb{R}^{m\times n}$ and $q \in \mathbb{R}^m$.

One obtains that

$$JT(x) = A\,.$$

Calculus of a quadratic function

- A **quadratic function** is a function $f : \mathbb{R}^n \to \mathbb{R}$ of the form

$$f(x) = \sum_{i=1}^{n} \sum_{j=i}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} a_i x_i + a_0 = \frac{1}{2} x^t Q x + a^t x + a_0 \,,$$

$$Q = \begin{bmatrix} 2a_{11} & a_{12} & \dots & a_{1n} \\ a_{12} & 2a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & 2a_{nn} \end{bmatrix} .$$

- This quadratic function has the following gradient and Hessian

$$\nabla f(x) = Qx + a \quad \text{and} \quad \nabla^2 f(x) = Q \,.$$

Calculus of a quadratic function

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) = \|Ax - b\|^2 ,$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then f could be expressed as

$$\begin{aligned} f(x) &= \|Ax - b\|^2 \\ &= (Ax - b)^t (Ax - b) \\ &= x^t A^t A x - 2(A^t b)^t x + b^t b . \end{aligned}$$

This means f is actually a quadratic function, and we may derive that

$$\nabla f(x) = 2A^t A x - 2A^t b \quad \text{and} \quad \nabla^2 f(x) = 2A^t A .$$

Calculus of a quadratic function

Alternatively, the function $f(x) = \|Ax - b\|^2$ can be viewed as a composite function $g \circ h$, where

$$\begin{aligned} g &: \mathbb{R}^m \to \mathbb{R} \quad & g(y) &= \|y\|^2 = y^t I y \\ h &: \mathbb{R}^n \to \mathbb{R}^m \quad & h(x) &= Ax - b\,. \end{aligned}$$

Here, the chain rule applies and one gets

$$\begin{aligned} Jg(y) &= \nabla g(y)^t = 2y^t \\ Jh(x) &= A \\ J(g \circ h)(x) &= Jg(h(x))Jh(x) = 2(Ax - b)^t A \\ &= 2(Ax)^t A - 2b^t A \\ \nabla f(x) &= \nabla (g \circ h)(x) \\ &= \Big(J(g \circ h)(x)\Big)^t \\ &= 2A^t Ax - 2A^t b\,. \end{aligned}$$

Section 2

The calculus of extreme value operators

inf, sup, min, max

For $A \subset \mathbb{R}$, define

$$\mathscr{L}(A) = \Big\{ \ell \in \mathbb{R} \mid \ell \leq a \quad (\forall a \in A) \Big\}$$
$$\mathscr{U}(A) = \Big\{ u \in \mathbb{R} \mid u \geq a \quad (\forall a \in A) \Big\}.$$

inf, sup, min, max

We define the following **extreme value operators**

- The **infimum** operator:

$$\ell = \inf A \iff \Big[\ell \in \mathscr{L}(A)\Big] \wedge \Big[\forall \ell' \in \mathscr{L}(A) : \ \ell \geq \ell'\Big] .$$

- The **supremum** operator:

$$u = \sup A \iff \Big[u \in \mathscr{U}(A)\Big] \wedge \Big[\forall u' \in \mathscr{U}(A) : \ u \leq u'\Big] .$$

- The **minimum** operator:

$$\ell = \min A \iff \Big[\ell = \inf A\Big] \wedge \Big[\ell \in A\Big] .$$

- The **maximum** operator:

$$u = \max A \iff \Big[u = \sup A\Big] \wedge \Big[u \in A\Big] .$$

arg min and arg max

It is usually the case that we consider

$$\min_{x \in \mathcal{X}^{\text{feas}}} J(x) := \min \left\{ J(x) \mid x \in \mathcal{X}^{\text{feas}} \right\} \quad \text{or} \quad \max_{x \in \mathcal{X}^{\text{feas}}} J(x) := \max \left\{ J(x) \mid x \in \mathcal{X}^{\text{feas}} \right\},$$

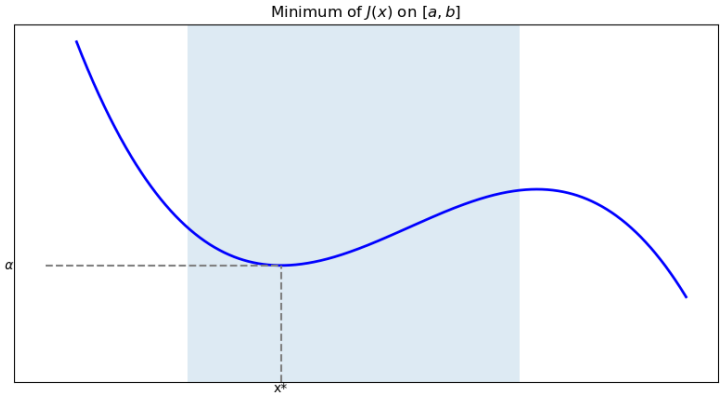
where we actually want to find the optimal decision x that is feasible and minimizes or maximizes the objective value. Hence we are interested in the following **solution set**:

$$\begin{aligned} \arg\min_{x \in \mathcal{X}^{\text{feas}}} J(x) : &= \arg\min \left\{ J(x) \mid x \in \mathcal{X}^{\text{feas}} \right\} \\ &= \left\{ \bar{x} \in \mathcal{X}^{\text{feas}} \mid J(\bar{x}) = \min J(x) \right\} \end{aligned}$$

or

$$\begin{aligned} \arg\max_{x \in \mathcal{X}^{\text{feas}}} J(x) : &= \arg\max \left\{ J(x) \mid x \in \mathcal{X}^{\text{feas}} \right\} \\ &= \left\{ \bar{x} \in \mathcal{X}^{\text{feas}} \mid J(\bar{x}) = \max J(x) \right\}. \end{aligned}$$

min and arg min



In the above figure, we have

$$\min_{x \in \mathcal{X}^{\text{feas}}} J(x) = \alpha \quad \text{and} \quad \arg\min_{x \in \mathcal{X}^{\text{feas}}} J(x) = \{x^*\}\,.$$

Extreme value operator calculus

Theorem

(a) $\min J(x) = -\max[-J(x)]$
(b) $\max J(x) = -\min[-J(x)]$
(c) $\arg\min J(x) = \arg\max[-J(x)]$ (arg min is obtained from arg max of negative objective.)
(d) $\arg\max J(x) = \arg\min[-J(x)]$ (arg max is obtained from arg min of negative objective.)

Theorem

For $\lambda > 0$ and $\beta \in \mathbb{R}$, we have the following.

(e) $\min[\lambda J(x)] = \lambda \min J(x)$
(f) $\min[J(x) + \beta] = [\min J(x)] + \beta$
(g) $\arg\min[\lambda J(x)] = \arg\min J(x)$ (The positive scaling does not affect arg min.)
(h) $\arg\min[J(x) + \beta] = \arg\min J(x)$ (Shifting up or down does not affect arg min.)

Theorem

(i) If J is non-negative, then

$$\min[J(x)]^2 = [\min J(x)]^2 \quad \text{and} \quad \arg\min[J(x)]^2 = \arg\min J(x)\,.$$

Community's choice

Since one could obtain max and max (and also arg max and arg min) from one another, we only need to study one of them.

For more than half a decade, the optimization community has decided to study **minimization**.

(However, the economists' community has decided otherwise to study **maximization**.)

That's it!

Key concept takeaways.

- Positive and negative (semi)definiteness
- Gradients, Hessians, and Jacobians
- Calculus of affine and quadratic functions
- Calculus of extreme value operators

Thank you.