Linear Algebra and Calculus Review

Can Li

ChE 597: Computational Optimization
Purdue University

Disclaimer

This lecture is like a cheat sheet for linear algebra and calculus.
The theorems are given without proof.
It is impossible to memorize the theorems and formulas since they are not intuitive. I would strongly recommend watching the YouTube video by 3Blue1Brown listed in the references to get some geometric intuition.

Matrix and Vector Notation

- By $A \in \mathbb{R}^{m \times n}$, we denote a matrix with m rows and n columns, where the entries of A are real numbers.
- By $x \in \mathbb{R}^n$, we denote a vector with n entries. By convention, an n-dimensional vector is often thought of as a matrix with n rows and 1 column, known as a *column vector*. If we want to explicitly represent a *row vector*—a matrix with 1 row and n columns—we typically write $x^\top$
- The i-th element of a vector x is denoted x_i:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Matrix Notation

- We use the notation a_{ij} (or $A_{ij}, A_{i,j}$, etc.) to denote the entry of A in the i-th row and j-th column:

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

- We denote the j-th column of A by a_j or $A_{:,j}$:

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}.$$

- We denote the i-th row of A by $a_i^\top$ or $A_{i,:}$:

$$A = \begin{bmatrix} — & a_1^\top & — \\ — & a_2^\top & — \\ & \vdots & \\ — & a_m^\top & — \end{bmatrix}.$$

Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is the matrix

$$C = AB \in \mathbb{R}^{m \times p},$$

where

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Note that in order for the matrix product to exist, the number of

columns in A must equal the number of rows in B.

$$C = AB = \begin{bmatrix} - & a_1^\top & - \\ - & a_2^\top & - \\ \vdots & & \\ - & a_m^\top & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^\top b_1 & a_1^\top b_2 & \cdots & a_1^\top b_p \\ a_2^\top b_1 & a_2^\top b_2 & \cdots & a_2^\top b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^\top b_1 & a_m^\top b_2 & \cdots & a_m^\top b_p \end{bmatrix}$$

Examples: vector products

- Inner product (dot product) of two vectors. Given two vectors $x, y \in \mathbb{R}^n$

$$x^\top y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

- Outer product of two vectors. Given vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ (not necessarily of the same size), $xy^\top \in \mathbb{R}^{m \times n}$ is called the *outer product* of the vectors. It is a matrix whose entries are given by $(xy^\top)_{ij} = x_i y_j$, i.e.,

$$xy^\top \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

Matrix-Vector Multiplication

Given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $x \in \mathbb{R}^n$, their product is a vector $y = Ax \in \mathbb{R}^m$.

If we write A by rows, then we can express Ax as:

$$y = Ax = \begin{bmatrix} — & a_1^\top & — \\ — & a_2^\top & — \\ & \vdots & \\ — & a_m^\top & — \end{bmatrix} x = \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ \vdots \\ a_m^\top x \end{bmatrix}.$$

Alternatively, let's write A in column form. In this case, we see that:

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 \end{bmatrix} x_1 + \begin{bmatrix} a_2 \end{bmatrix} x_2 + \cdots + \begin{bmatrix} a_n \end{bmatrix} x_n.$$

In other words, y is a *linear combination* of the *columns* of A, where the coefficients of the linear combination are given by the entries of x.

Properties of Matrix Multiplication

- Matrix multiplication is **associative**:

$$(AB)C = A(BC).$$

- Matrix multiplication is **distributive**:

$$A(B + C) = AB + AC.$$

- Matrix multiplication is, in general, **not commutative**; that is, it can be the case that:

$$AB \neq BA.$$

  For example, if $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times q}$, the matrix product BA does not even exist if m and q are not equal!

The Transpose and Symmetric Matrices

The Transpose:

- The transpose of a matrix $A \in \mathbb{R}^{m \times n}$, written $A^\top \in \mathbb{R}^{n \times m}$, is defined as $(A^\top)_{ij} = A_{ji}$.
- Properties:

$$(A^\top)^\top = A, \quad (AB)^\top = B^\top A^\top, \quad (A+B)^\top = A^\top + B^\top.$$

Symmetric Matrices:

- A square matrix $A \in \mathbb{R}^{n \times n}$ is:
  - *Symmetric* if $A = A^\top$.
  - *Anti-symmetric* if $A = -A^\top$.
- Any square matrix can be decomposed as:

$$A = \frac{1}{2}(A + A^\top) + \frac{1}{2}(A - A^\top),$$

  where the first term is symmetric and the second is anti-symmetric.

The Trace of a Matrix

Definition: The *trace* of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}(A)$, is the sum of its diagonal elements:

$$\mathrm{tr} A = \sum_{i=1}^{n} A_{ii}.$$

Properties of the Trace:

- For $A \in \mathbb{R}^{n \times n}$, $\mathrm{tr} A = \mathrm{tr} A^\top$.
- For $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(A + B) = \mathrm{tr} A + \mathrm{tr} B$.
- For $A \in \mathbb{R}^{n \times n}$, $t \in \mathbb{R}$, $\mathrm{tr}(tA) = t\,\mathrm{tr} A$.
- For A, B such that AB is square, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
- For A, B, C such that ABC is square, $\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$, and so on for products of more matrices.

Norms of Vectors and Matrices

Definition of a Norm: A *norm* measures the "length" of a vector. Commonly used norms include:

- **Euclidean or ℓ_2 norm:**

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

  Note: $\|x\|_2^2 = x^\top x$.

- **ℓ_1 norm:**

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

- **ℓ_∞ norm:**

$$\|x\|_\infty = \max_i |x_i|.$$

- **ℓ_p norm:**

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad p \geq 1.$$

Norms of Vectors and Matrices

More formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity).
2. $f(x) = 0$ if and only if $x = 0$ (definiteness).
3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t|f(x)$ (homogeneity).
4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality).

Matrix Norm: The Frobenius norm is defined as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\operatorname{tr}(A^\top A)}.$$

The *2-norm* (spectral norm) of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as:

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$$

where $\lambda_{\max}(A^\top A)$ denotes the largest eigenvalue of the matrix $A^\top A$.

Linear Independence and Dependence

A set of vectors $\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^m$ is:

- **Linearly independent** if no vector can be represented as a linear combination of the remaining vectors.
- **Linearly dependent** if one vector can be expressed as a linear combination of the others:

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i,$$

for some scalars $\alpha_1, \dots, \alpha_{n-1} \in \mathbb{R}$.
zero vector is always linearly dependent on others.

Example: The vectors

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

are linearly dependent because $x_3 = -2x_1 + x_2$.

Rank of a Matrix

Definition:

- The *column rank* of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of linearly independent columns.
- The *row rank* is the size of the largest subset of linearly independent rows.
- For any matrix $A \in \mathbb{R}^{m \times n}$, the column rank equals the row rank, collectively referred to as the *rank* of A, denoted as $\text{rank}(A)$.

Properties of the Rank:

- $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is *full rank*.
- $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$:

$$\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B)).$$

- $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

The Inverse of a Matrix

Definition: The *inverse* of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted A^{-1}, is the unique matrix such that:

$$A^{-1}A = I = AA^{-1}.$$

Key Points:

- Not all matrices have inverses. Non-square matrices, for example, do not have inverses.
- A square matrix A is *invertible* or *non-singular* if A^{-1} exists; otherwise, it is *non-invertible* or *singular*.
- For A to have an inverse, it must be full rank.

Properties of the Inverse:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^T)^{-1} = (A^{-1})^T$

Application: For a linear system $Ax = b$, if A is invertible, the solution is $x = A^{-1}b$.

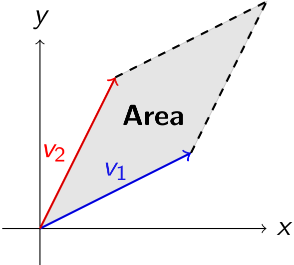
Matrix Determinant and Geometric Intuition

The determinant of a square matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$, denoted $\det(A)$, is a scalar value that encodes properties of the linear transformation represented by A.

- For a 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \det(A) = ad - bc$$

- The determinant represents the **signed area** of the parallelogram spanned by the column vectors of A.
- In higher dimensions ($n \times n$), it represents the **signed volume** of the parallelepiped spanned by the columns.

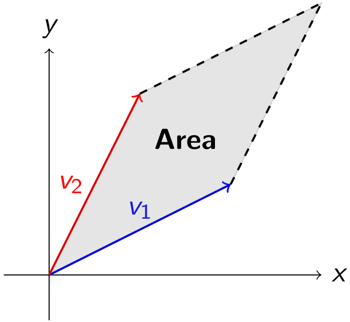


Matrix Determinant and Geometric Intuition

Key Properties:

- $\det(A) = 0$: The transformation collapses space (columns are linearly dependent).
- $\det(A) > 0$: Orientation is preserved.
- $\det(A) < 0$: Orientation is reversed.

Visual Example (2D):



Determinant Formula and Adjoint Method for Inverse

Determinant Formula for an $n \times n$ Matrix:

For $A = [a_{ij}] \in \mathbb{R}^{n \times n}$, the determinant is given by:

$$\det(A) = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{\setminus 1, \setminus j})$$

where $A_{\setminus 1, \setminus j}$ is the $(n-1) \times (n-1)$ matrix obtained by removing the 1st row and j-th column of A.

Adjoint Method to Calculate A^{-1}:

- The **adjoint matrix** of A, denoted $\operatorname{adj}(A)$, is the transpose of the cofactor matrix:

$$\operatorname{adj}(A)_{ij} = (-1)^{i+j} \det(A_{\setminus i, \setminus j}).$$

- If $\det(A) \neq 0$, the inverse of A is:

$$A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A)$$

Vector Spaces of Matrix $A \in \mathbb{R}^{m\times n}$

- **Column (Range) Space** $(\mathrm{Col}(A), \mathrm{Range}(A))$
  - Set of all linear combinations of the column vectors of A.
  - Subspace of $\mathbb{R}^m$. $\{Ax \mid \forall x \in \mathbb{R}^n\}$
  - Dimension equals the rank of A.
- **Row Space** $(\mathrm{Row}(A))$
  - Set of all linear combinations of the row vectors of A.
  - Subspace of $\mathbb{R}^n$. $\{A^T y \mid \forall y \in \mathbb{R}^m\}$
  - Dimension equals the rank of A.
  - For $A \in \mathcal{S}^n$ (symmetric matrix), $\mathrm{Col}(A) = \mathrm{Row}(A)$.
- **Null Space** $(\mathrm{Null}(A))$
  - Set of all vectors x such that $Ax = 0$.
  - Subspace of $\mathbb{R}^n$.
  - Dimension equals the $n-$rank(A).
- **Orthogonal Properties**
  - The row space and null space of A are orthogonal complements in $\mathbb{R}^n$.
    $\{w : w = u + v, u \in \mathrm{Row}\,(A)\,, v \in \mathrm{Null}(A)\} = \mathbb{R}^n$ and
    $\mathrm{Row}\,(A) \cap \mathrm{Null}(A) = \emptyset$
  - The column space and the left null space (null space of A^T) are orthogonal complements in $\mathbb{R}^m$.

Eigenvalues and Eigenvectors: Definition and Calculation

Definition: For a square matrix $A \in \mathbb{R}^{n \times n}$, a scalar $\lambda \in \mathbb{R}$ and a nonzero vector $v \in \mathbb{R}^n$ are called an **eigenvalue** and **eigenvector**, respectively, if:

$$Av = \lambda v$$

How to Calculate: 1. Solve the **characteristic equation**:

$$\det(A - \lambda I) = 0$$

This yields the eigenvalues λ.
2. For each eigenvalue λ, solve the system of linear equations:

$$(A - \lambda I)v = 0$$

to find the corresponding eigenvectors v.

Example: Let $A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$:

- Solve $\det(A - \lambda I) = 0$:

$$\det\left(\begin{bmatrix} 4 - \lambda & 1 \\ 2 & 3 - \lambda \end{bmatrix}\right) = 0$$

Properties of Eigenvalues and Eigenvectors

- For a matrix $A \in \mathbb{R}^{n \times n}$:
  - The sum of the eigenvalues equals the trace of A:

$$\sum_{i=1}^{n} \lambda_i = \mathrm{tr}(A)$$

  - The product of the eigenvalues equals the determinant of A:

$$\prod_{i=1}^{n} \lambda_i = \det(A)$$

- Eigenvectors corresponding to distinct eigenvalues are linearly independent.

PSD, PD, NSD, and ND Symmetric Matrices

Symmetric Matrix $A \in \mathbb{R}^{n \times n}$: $A = A^\top$**. Notation:** $A \in \mathbb{S}^n$

- **Positive Semidefinite (PSD)**: A is PSD if:

  $$x^\top A x \geq 0 \quad \text{for all } x \in \mathbb{R}^n$$

  All eigenvalues of A are $\lambda_i \geq 0$.
- **Positive Definite (PD)**: A is PD if:

  $$x^\top A x > 0 \quad \text{for all nonzero } x \in \mathbb{R}^n$$

  All eigenvalues of A are $\lambda_i > 0$.
- **Negative Semidefinite (NSD)**: A is NSD if:

  $$x^\top A x \leq 0 \quad \text{for all } x \in \mathbb{R}^n$$

  All eigenvalues of A are $\lambda_i \leq 0$.
- **Negative Definite (ND)**: A is ND if:

  $$x^\top A x < 0 \quad \text{for all nonzero } x \in \mathbb{R}^n$$

  All eigenvalues of A are $\lambda_i < 0$.
- A matrix is **indefinite** if it has both positive and negative eigenvalues.

Eigenvalue Decomposition (EVD)

Definition: For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, the **eigenvalue decomposition** is:

$$A = Q\Lambda Q^\top = \sum_{i=1}^{n} \lambda_i q_i q_i^\top$$

where:

- $Q = \begin{bmatrix} q_1 & q_2 & \dots & q_n \end{bmatrix}$ is an orthogonal matrix ($QQ^\top = I$) containing the eigenvectors of A.
- Λ is a diagonal matrix with the eigenvalues of A as its entries:

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$$

Key Properties:

- A is symmetric $\implies$ all eigenvalues are real.
- A is PSD $\implies$ all $\lambda_i \geq 0$.

Notation for Functions

$$f : A \to B$$

f is a function on the set dom $f \subseteq A$ into the set B; in particular, we can have dom f as a proper subset of the set A.

- $f : \mathbb{R}^n \to \mathbb{R}^m$ means that f maps (some) n-vectors into m-vectors; it does not mean that $f(x)$ is defined for all $x \in \mathbb{R}^n$.
- $f(x) = log(x)$. $f : \mathbb{R}^{++} \to \mathbb{R}$

Continuity of Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *continuous* at $x \in \text{dom } f$ if $\forall\ \epsilon > 0$, $\exists$ $\delta > 0$ such that

$$\forall y \in \text{dom } f, \quad \|y - x\|_2 \leq \delta \implies \|f(y) - f(x)\|_2 \leq \epsilon.$$

Continuity can be described in terms of limits: whenever the sequence $x_1, x_2, \ldots$ in dom f converges to a point $x \in \text{dom } f$, the sequence $f(x_1), f(x_2), \ldots$ converges to $f(x)$, i.e.,

$$\lim_{i \to \infty} f(x_i) = f\left(\lim_{i \to \infty} x_i\right).$$

A function f is *continuous* if it is continuous at every point in its domain.

Derivative and Differentiable Functions

Derivative: Let $f : \mathbb{R}^n \to \mathbb{R}^m$. We say f is **differentiable at** x if there exists a linear map $Df(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that:

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - Df(x)h\|}{\|h\|} = 0.$$

Differentiable Functions: A function f is differentiable on an open set U if it is differentiable at every point in U.
Differentiability $\implies$ continuity.

Gradient (for scalar functions): If $f : \mathbb{R}^n \to \mathbb{R}$, then the gradient of f at x is:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

It satisfies: $Df(x)h = \nabla f(x)^\top h$.

Gradient Interpretation and C^1 Functions

Interpretation of the Gradient: - The gradient points in the direction of maximum increase of f. - Its magnitude $\|\nabla f(x)\|$ represents the rate of change.

C^1 **Functions:** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is C^1 on an open set U if:

1. f is differentiable on U.
2. The map $x \mapsto Df(x)$ (or each partial derivative) is continuous on U.

For scalar functions $f : \mathbb{R}^n \to \mathbb{R}$, this means:

$$f \in C^1(U) \iff \frac{\partial f}{\partial x_i}(x) \text{ exists and is continuous for all } i.$$

C^1 functions are also called smooth functions.

Hessian Matrix

Definition: The **Hessian matrix** of a twice-differentiable scalar function $f : \mathbb{R}^n \to \mathbb{R}$ at a point $x \in \mathbb{R}^n$ is the $n \times n$ matrix of second-order partial derivatives:

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

Properties:

- $H_f(x)$ is symmetric if f is twice continuously differentiable ($f \in C^2$).
- The Hessian provides information about the curvature of f.

Taylor Series Expansion (Second Order)

Definition: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function. The **Taylor series expansion up to the second order** around a point $x \in \mathbb{R}^n$ is given by:

$$f(x+h) \approx f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top H_f(x) h,$$

where:

- $\nabla f(x)$ is the gradient of f at x,
- $H_f(x)$ is the Hessian matrix of f at x,
- $h = x' - x$ is the displacement vector.

Interpretation:

- The first term, $f(x)$, represents the function value at x.
- The second term, $\nabla f(x)^\top h$, is the linear approximation (gradient contribution).
- The third term, $\frac{1}{2} h^\top H_f(x) h$, accounts for the curvature (second-order effects).

Peano and Lagrange Remainders in Taylor Series

For a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the second-order Taylor expansion of f at x for a small increment h is:

$$f(x+h) = f(x) + \nabla f(x)^\top h + \tfrac{1}{2}\, h^\top H_f(x)\, h \; + \; R_2(x,h),$$

where $H_f(x)$ is the Hessian of f at x and $R_2(x,h)$ is the remainder.

1. Peano Form:

$$R_2(x,h) = o(\|h\|^2) \quad \text{as} \quad \|h\| \to 0.$$

This means $R_2(x,h)$ goes to zero faster than $\|h\|^2$.

2. Lagrange Form: There exists $\theta \in (0,1)$ such that

$$f(x+h) = f(x) + \nabla f(x)^\top h + \tfrac{1}{2}\, h^\top H_f\big(x + \theta\, h\big)\, h.$$

Equivalently,

$$R_2(x,h) = \tfrac{1}{2}\, h^\top \big[\, H_f\big(x + \theta\, h\big) - H_f\big(x\big) \big]\, h.$$

Key Insight: - The *Peano form* expresses how R_2 vanishes *asymptotically*. - The *Lagrange form* gives a pointwise representation of R_2 via the Hessian at an intermediate point.

Mean Value Theorem

Let $f\colon U \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function such that:

- f is **continuous** on the line segment joining x to y (i.e., on $\{\, x + t(y-x) : t \in [0,1] \}$),
- f is **differentiable** in the interior of that segment (i.e., for $t \in (0,1)$).

Then there exists some $\theta \in (0,1)$ such that

$$f(y) - f(x) \;=\; \nabla f\big(x + \theta\,(y-x)\big)^{\top}\big(y-x\big).$$

- The directional derivative of f at some point on the segment from x to y (in the direction $y-x$) matches the average rate of change $\frac{f(y)-f(x)}{\|y-x\|}$.
- Geometrically, $\nabla f\big(\dots\big)$ at this point captures how f changes most rapidly, and it aligns with the increment $y-x$.
- MVT can be seen as a corollary or special case of the Lagrange form of the first-order Taylor remainder.

Jacobian Matrix

Let $f: \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable vector-valued function:

$$f(x) = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) \end{bmatrix}.$$

The **Jacobian matrix** of f at $x \in \mathbb{R}^n$ is the $m \times n$ matrix of partial derivatives:

$$J_f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^\top f_1 \\ \vdots \\ \nabla^\top f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

- $J_f(x) \in \mathbb{R}^{m \times n}$, where m is the number of outputs, and n is the number of inputs.
- Each row corresponds to the gradient of a component function $f_i(x)$.
- If $m = n$, and $J_f(x)$ is invertible, f is locally invertible near x.

Chain Rule

Scalar form: If $y = f(g(x))$, where $g : \mathbb{R} \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$, then:

$$\frac{dy}{dx} = f'(g(x)) \cdot g'(x).$$

Matrix Form: If $f : \mathbb{R}^m \to \mathbb{R}^p$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, then for $F(x) = f(g(x))$,

$$J_F(x) = J_f(g(x))\, J_g(x)$$

where $J_g(x) \in \mathbb{R}^{m \times n}$ and $J_f(g(x)) \in \mathbb{R}^{p \times m}$.

Chain Rule for Second Derivative

A general chain rule for the second derivative is cumbersome in most cases, so we state it only for some special cases that we will need.

Composition with Scalar Function: Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(x) = g(f(x))$. By computing the partial derivatives, we get:

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T$$

Composition with Affine Function: Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times m}$, and $b \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(x) = f(Ax + b)$. Then:

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b)A.$$

References

- Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge university press. Appendix A.
- Linear algebra review notes by Zico Kolter. `https://www.cs.cmu.edu/~zkolter/course/linalg/index.html`
- YouTube videos review of linear algebra and calculus by 3Blue1Brown. `https://www.youtube.com/@3blue1brown/courses` **Strongly recommend!** Provides geometric intuition.