Math Revision Session

3: Matrix Algebra (3)

Jukina HATAKEYAMA

The University of Osaka, Department of Economics

April 24, 2025

Solving a System of Linear Equations with Matrices

Step 1: Represent the system as a matrix equation

- Consider the system of linear equations:

$$\begin{cases} 2x + y & = 5 \\ 3x + 4y & = 6 \end{cases}$$

- This system can be written in matrix form as:

$$A\mathbf{x} = \mathbf{b}$$

where

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}$$

Solving the System using the Inverse Matrix

Step 2: Solve for $\mathbf{x}$ using the inverse of A

- To find $\mathbf{x}$, use the formula:

$$\mathbf{x} = A^{-1}\mathbf{b}$$

- First, calculate the inverse of matrix A:

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} 4 & -1 \\ -3 & 2 \end{pmatrix}$$

  where $\det(A) = (2)(4) - (1)(3) = 5$.

- So,

$$A^{-1} = \frac{1}{5} \begin{pmatrix} 4 & -1 \\ -3 & 2 \end{pmatrix}$$

Computing the Solution

Step 3: Multiply A^{-1} with $\mathbf{b}$

- Now, multiply A^{-1} with $\mathbf{b}$:

$$\mathbf{x} = \frac{1}{5}\begin{pmatrix} 4 & -1 \\ -3 & 2 \end{pmatrix}\begin{pmatrix} 5 \\ 6 \end{pmatrix}$$

- Perform the matrix multiplication:

$$\mathbf{x} = \frac{1}{5}\begin{pmatrix} (4)(5) + (-1)(6) \\ (-3)(5) + (2)(6) \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 20 - 6 \\ -15 + 12 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 14 \\ -3 \end{pmatrix}$$

- Finally, we get the solution:

$$\mathbf{x} = \begin{pmatrix} 14/5 \\ -3/5 \end{pmatrix} = \begin{pmatrix} 2.8 \\ -0.6 \end{pmatrix}$$

Solving Systems of Linear Equations using Cramer's Rule

Overview of Cramer's Rule:

- A method for solving systems of linear equations in matrix form.
- The solutions for each variable are found using the determinants of matrices and cofactor matrices.

The system of equations:

$$A\mathbf{x} = \mathbf{b}$$

where A is the coefficient matrix, $\mathbf{x}$ is the vector of variables, and $\mathbf{b}$ is the constant vector.

Definition of Cramer's Rule

The solution for each variable x_i is given by the following formula:

$$x_i = \frac{\mathsf{det}(A_i)}{\mathsf{det}(A)}$$

where A_i is the matrix obtained by replacing the i-th column of matrix A with the constant vector $\mathbf{b}$.

Example: System of Three Variables

The system of equations:

$$\begin{cases} x + y + z &= 6 \\ 2x + 3y + z &= 14 \\ 3x + y + 2z &= 10 \end{cases}$$

In matrix form:

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 6 \\ 14 \\ 10 \end{pmatrix}$$

Calculating the Determinant

The determinant of matrix A is:

$$\begin{aligned}
\mathsf{det}(A) &= 1(3 \times 2 - 1 \times 1) - 1(2 \times 2 - 1 \times 3) + 1(2 \times 1 - 3 \times 3) \\
&= 1(6 - 1) - 1(4 - 3) + 1(2 - 9) \\
&= 5 - 1 - 7 = -3
\end{aligned}$$

Creating the Cofactor Matrices

- **Matrix** A_x: The matrix obtained by replacing the first column with the constant vector $\mathbf{b}$:

$$A_x = \begin{pmatrix} 6 & 1 & 1 \\ 14 & 3 & 1 \\ 10 & 1 & 2 \end{pmatrix}$$

- **Matrix** A_y: The matrix obtained by replacing the second column with $\mathbf{b}$:

$$A_y = \begin{pmatrix} 1 & 6 & 1 \\ 2 & 14 & 1 \\ 3 & 10 & 2 \end{pmatrix}$$

- **Matrix** A_z: The matrix obtained by replacing the third column with $\mathbf{b}$:

$$A_z = \begin{pmatrix} 1 & 1 & 6 \\ 2 & 3 & 14 \\ 3 & 1 & 10 \end{pmatrix}$$

Calculating the Determinants of the Cofactor Matrices

The determinants of the cofactor matrices are:

$$\mathsf{det}(A_x) = -4, \quad \mathsf{det}(A_y) = -10, \quad \mathsf{det}(A_z) = -38$$

Solving for the Variables

Using Cramer's Rule, we solve for each variable:

$$x = \frac{\mathsf{det}(A_x)}{\mathsf{det}(A)} = \frac{-4}{-3} = \frac{4}{3}$$

$$y = \frac{\mathsf{det}(A_y)}{\mathsf{det}(A)} = \frac{-10}{-3} = \frac{10}{3}$$

$$z = \frac{\mathsf{det}(A_z)}{\mathsf{det}(A)} = \frac{-38}{-3} = \frac{38}{3}$$

Final Solution

Therefore, the solution to the system of equations is:

$$x = \frac{4}{3}, \quad y = \frac{10}{3}, \quad z = \frac{38}{3}$$

Matrix Differentiation

Matrix differentiation involves computing the derivative of a matrix expression with respect to another matrix or vector. Some useful rules and examples are shown below.

1. Derivative of a Scalar with Respect to a Vector

If $\mathbf{x}$ is a column vector of size $n \times 1$ and $f(\mathbf{x})$ is a scalar function, then:

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \quad \text{is the gradient vector} \quad \nabla_{\mathbf{x}} f(\mathbf{x})$$

The gradient is a column vector containing the partial derivatives of the scalar function with respect to each component of $\mathbf{x}$.

2. Derivative of a Scalar with Respect to a Matrix

If A is an $m \times n$ matrix and $f(A)$ is a scalar function of A, the derivative is:

$$\frac{\partial f(A)}{\partial A}$$

is the matrix of partial derivatives of each element of $f(A)$ with respect to each element of A.

3. Common Matrix Derivatives

- Derivative of a quadratic form:

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T A \mathbf{x}) = (A + A^T)\mathbf{x}$$

- Derivative of the trace of a matrix:

$$\frac{\partial}{\partial A}\mathsf{tr}(A) = I$$

  where I is the identity matrix.

- Derivative of the determinant:

$$\frac{\partial}{\partial A}\det(A) = \det(A)(A^{-1})^T$$

4. Gradient of a Vector with Respect to a Matrix

For a vector $\mathbf{y} = A\mathbf{x} + \mathbf{b}$, the gradient with respect to the matrix A is:

$$\frac{\partial \mathbf{y}}{\partial A} = \mathbf{x}^T$$

These rules are fundamental for optimization problems involving matrix expressions.

Consider a scalar function $f(x)$ where $x \in \mathbb{R}^n$. The gradient of $f(x)$ with respect to x is the vector of partial derivatives of $f(x)$ with respect to each component of x:

$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

1. Example: Gradient of a Quadratic Form For $f(x) = x^T A x$, where A is a symmetric matrix, the gradient is:

$$\nabla_x f(x) = 2Ax$$

2. Example: Gradient of a Linear Function For $f(x) = \mathbf{b}^T x$, where $\mathbf{b}$ is a vector, the gradient is:

$$\nabla_x f(x) = \mathbf{b}$$

3. Example: Gradient of a Norm For $f(x) = \|x\|_2^2$, the gradient is:

$$\nabla_x f(x) = 2x$$

The gradient provides the direction of the steepest ascent of the function.

Let A be an $m \times n$ matrix, and $f(A)$ be a scalar function of A. The derivative of $f(A)$ with respect to A is a matrix where each element is the partial derivative of $f(A)$ with respect to the corresponding element of A:

$$\frac{\partial f(A)}{\partial A} = \begin{pmatrix} \frac{\partial f(A)}{\partial a_{11}} & \frac{\partial f(A)}{\partial a_{12}} & \cdots & \frac{\partial f(A)}{\partial a_{1n}} \\ \frac{\partial f(A)}{\partial a_{21}} & \frac{\partial f(A)}{\partial a_{22}} & \cdots & \frac{\partial f(A)}{\partial a_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial a_{m1}} & \frac{\partial f(A)}{\partial a_{m2}} & \cdots & \frac{\partial f(A)}{\partial a_{mn}} \end{pmatrix}$$

Example: Gradient of the Frobenius Norm For $f(A) = \mathrm{tr}(A^T A) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2$, the derivative is:

$$\frac{\partial f(A)}{\partial A} = 2A$$

In this case, each element a_{ij} is differentiated to give $2a_{ij}$, so the derivative of $f(A)$ is $2A$.

Vector Version of the Chain Rule

The chain rule in vector form allows us to compute the derivative of a composition of functions. Let $\mathbf{f}(\mathbf{x})$ be a vector-valued function, and let $\mathbf{g}(\mathbf{x})$ be a vector of functions. The chain rule for vector functions is expressed as:

$$\frac{d}{dx}\mathbf{f}(\mathbf{g}(x)) = \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \cdot \frac{d}{dx}\mathbf{g}(x)$$

In matrix form, the chain rule is written as:

$$\frac{d}{dx}\mathbf{f}(\mathbf{g}(x)) = \mathbf{J}_f(\mathbf{g}(x)) \cdot \mathbf{J}_g(x)$$

Where $\mathbf{J}_f(\mathbf{g}(x))$ is the Jacobian matrix of $\mathbf{f}$ with respect to $\mathbf{g}$, and $\mathbf{J}_g(x)$ is the Jacobian matrix of $\mathbf{g}$ with respect to x.

Example of the Chain Rule in Vectors

Consider the following composition of functions:

$$\mathbf{f}(\mathbf{g}(x)) = \begin{pmatrix} f_1(g_1(x), g_2(x)) \\ f_2(g_1(x), g_2(x)) \end{pmatrix} \quad \text{and} \quad \mathbf{g}(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \end{pmatrix}$$

The Jacobian matrices for $\mathbf{f}$ and $\mathbf{g}$ are:

$$\mathbf{J}_f(\mathbf{g}(x)) = \begin{pmatrix} \frac{\partial f_1}{\partial g_1} & \frac{\partial f_1}{\partial g_2} \\ \frac{\partial f_2}{\partial g_1} & \frac{\partial f_2}{\partial g_2} \end{pmatrix}$$

$$\mathbf{J}_g(x) = \begin{pmatrix} \frac{dg_1}{dx} \\ \frac{dg_2}{dx} \end{pmatrix}$$

Thus, the derivative of $\mathbf{f}(\mathbf{g}(x))$ is:

$$\frac{d}{dx}\mathbf{f}(\mathbf{g}(x)) = \mathbf{J}_f(\mathbf{g}(x)) \cdot \mathbf{J}_g(x)$$

OLS

In the ordinary least squares (: OLS) method, we aim to estimate the parameters β that minimise the sum of squared residuals.
The linear model for observation i is:

$$y_i = x_i^T \beta + \epsilon_i$$

where:

- y_i is the dependent variable for observation i,
- $x_i \in \mathbb{R}^k$ is the vector of independent variables (including the intercept) for observation i,
- $\beta \in \mathbb{R}^k$ is the vector of coefficients to be estimated,
- ϵ_i is the error term for observation i.

The goal is to estimate β by minimizing the residual sum of squares:

$$\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 = \min_{\beta} \sum_{i=1}^{n} (y_i - X_i \beta)^2$$

OLS Model in Stacked Form

In the stacked form of the OLS model, we combine all observations into a single system of equations. The model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^n$$

where:

- $\mathbf{y}$ is the vector of all dependent variables, $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$,
- $\mathbf{X}$ is the matrix of all independent variables, $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$,
- $\boldsymbol{\beta}$ is the vector of coefficients, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)^T$,
- $\boldsymbol{\epsilon}$ is the vector of all error terms, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$.

The residual sum of squares (RSS) is minimized:

$$\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 = \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 = \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

Let $\hat{\beta}$ denote the unique solution of the optimisation problem, then the F.O.C is:

$$\nabla_{\beta}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$

Then, we have:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

by using the chain rule.

Differences Between the Jacobian, Hessian, and Gradient

The Jacobian, Hessian, and Gradient are important concepts related to derivatives in multivariable calculus. Each plays a different role depending on the context and type of function.

- **Gradient**:
  - The gradient is a vector that points in the direction of the steepest increase of a scalar-valued function.
  - It corresponds to the derivative in one-dimensional functions, and for multivariable functions, it is represented as a vector.
  - For a scalar function $\mathbf{f}(\mathbf{x})$, the gradient is expressed as:

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

- **Jacobian**:
  - The Jacobian is a matrix of all first-order partial derivatives of a vector-valued function.
  - For a vector function $\mathbf{f}(\mathbf{x})$, the Jacobian matrix is defined as:

$$\mathbf{J}_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

- **Hessian**:
  - The Hessian is a square matrix of second-order partial derivatives of a scalar-valued function.
  - It provides information about the curvature of a function and is useful for optimization problems, especially in finding critical points.
  - For a scalar function $\mathbf{f}(\mathbf{x})$, the Hessian matrix is defined as:

$$H(\mathbf{f}) = \begin{pmatrix} \frac{\partial^2 \mathbf{f}}{\partial x_1^2} & \frac{\partial^2 \mathbf{f}}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \mathbf{f}}{\partial x_1 \partial x_n} \\ \frac{\partial^2 \mathbf{f}}{\partial x_2 \partial x_1} & \frac{\partial^2 \mathbf{f}}{\partial x_2^2} & \cdots & \frac{\partial^2 \mathbf{f}}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathbf{f}}{\partial x_n \partial x_1} & \frac{\partial^2 \mathbf{f}}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 \mathbf{f}}{\partial x_n^2} \end{pmatrix}$$