COMP7180: Quantitative Methods for Data Analytics and Artificial Intelligence

Lecture 11: Matrix Calculus

Jun Qi

Research Assistant Professor in Computer Science @ Hong Kong Baptist University

Affiliated Associate Professor in Electronic Engineering @ Fudan University

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

29/10/2024

Hints on Problem 7

- 7 (a) Using the definition of expectation

$$E_Y[E_X[X \mid Y]] = \sum_y E_X(X \mid Y = y)p(y) = \sum_y \left(\sum_x x p(x \mid y) \right) p(y)$$

- 7 (b) Conditional Variance Component:

$$E_Y[\mathrm{Var}[X \mid Y]] = E_Y[E[X^2 \mid Y] - (E[X \mid Y])^2]$$

Variance of Expectation Component:

$$\mathrm{Var}_Y[E[X \mid Y]] = E_Y[(E[X \mid Y])^2] - (E[E[X \mid Y]])^2$$

Combine the above two terms…

Outline

- The Derivatives of Vector Functions
- The Chain Rule for Vector Functions
- Maximum Likelihood Estimation

The Derivatives of Vector Functions

Let $\mathbf{x}$ and $\mathbf{y}$ be vectors of orders n and m respectively:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix},$$

where each component y_i may be a function of all the x_j, a fact represented by saying that $\mathbf{y}$ is a function of $\mathbf{x}$, or

$$\mathbf{y} = \mathbf{y}(\mathbf{x}).$$

If $n = 1$, $\mathbf{x}$ reduces to a scalar, which we call x. If $m = 1$, $\mathbf{y}$ reduces to a scalar, which we call y. Various applications are studied in the following subsections.

Derivative of Vector with Respect to Vector

The derivative of the vector $\mathbf{y}$ with respect to vector $\mathbf{x}$ is the $n \times m$ matrix

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Derivative of a Scalar with Respect to Vector

If y is a scalar

$$\frac{\partial y}{\partial \mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}.$$

It is also called the gradient of y for a vector variable x, denoted by ∇y.

1.3 Derivative of Vector concerning Scalar

$$\frac{\partial \mathbf{y}}{\partial x} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}$$

Example 1

Given $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

and

$$y_1 = x_1^2 - x_2$$
$$y_2 = x_3^2 + 3x_2$$

the partial derivative matrix $\partial \mathbf{y}/\partial \mathbf{x}$ is computed as follows:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} \\ \frac{\partial y_1}{\partial x_3} & \frac{\partial y_2}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 & 0 \\ -1 & 3 \\ 0 & 2x_3 \end{bmatrix}$$

Some useful vector derivative formulas

$$\frac{\partial \ Cx}{\partial \mathbf{x}} = C^T$$

$$\frac{\partial \ x^T C}{\partial \mathbf{x}} = C$$

$$\frac{\partial \ x^T x}{\partial \mathbf{x}} = 2x$$

Homework

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} x_t C_{1t} \\ \sum_{t=1}^{n} x_t C_{2t} \\ \vdots \\ \sum_{t=1}^{n} x_t C_{nt} \end{bmatrix}$$

$$\frac{\partial \ \mathrm{Cx}}{\partial \mathrm{x}} = \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{n1} \\ c_{12} & c_{22} & \cdots & c_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1n} & c_{2n} & \cdots & c_{nn} \end{pmatrix} = C^T$$

Important Property of Quadratic Form x^TCx

$$\frac{\partial\ (\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}{\partial \mathbf{x}} = \left(\mathbf{C}+\mathbf{C}^{\mathrm{T}}\right)\mathbf{x}$$

Proof:

$$\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} = \sum_{i=1}^{n}\left[x_i\sum_{j=1}^{n}\left(x_j C_{ij}\right)\right]$$

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} x_t C_{1t} \\ \sum_{t=1}^{n} x_t C_{2t} \\ \vdots \\ \sum_{t=1}^{n} x_t C_{nt} \end{bmatrix}$$

$$\Rightarrow \frac{\partial(\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}{\partial x_k} = \frac{\partial\left\{\sum_{i=1}^{n}\left[x_i\sum_{j=1}^{n}\left(x_j C_{ij}\right)\right]\right\}}{\partial x_k} = \frac{\partial\left\{x_k\sum_{j=1}^{n}\left(x_j C_{kj}\right)\right\}}{\partial x_k} + \frac{\partial\left\{\sum_{i=1}^{n}\left[x_i x_k C_{ik}\right]\right\}}{\partial x_k}$$

$$= \sum_{j=1}^{n} x_j C_{kj} + \sum_{i=1}^{n} x_i C_{ik}$$

$$\Rightarrow \frac{\partial(\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}{\partial x} = \mathbf{C}\mathbf{x} + \mathbf{C}^{\mathrm{T}}\mathbf{x} = \left(\mathbf{C}+\mathbf{C}^{\mathrm{T}}\right)\mathbf{x}$$

If C is symmetric, $\dfrac{\partial\ (\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}{\partial \mathbf{x}} = 2C\,\mathbf{x}$

The Chain Rule for Vector Functions

Let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{bmatrix} \quad \text{and} \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}$$

where z is a function of y, which is in turn a function of x, we can write

$$\left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right)^T = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} & \cdots & \frac{\partial z_1}{\partial x_n} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial z_m}{\partial x_1} & \frac{\partial z_m}{\partial x_2} & \cdots & \frac{\partial z_m}{\partial x_n} \end{bmatrix}$$

Each entry of this matrix may be expanded as

$$\frac{\partial z_i}{\partial x_j} = \sum_{q=1}^{r} \frac{\partial z_i}{\partial y_q} \frac{\partial y_q}{\partial x_j} \qquad \begin{cases} i = 1, 2, \ldots, m \\ j = 1, 2, \ldots, n. \end{cases}$$

The Chain Rule for Vector Functions (Cont.)

Then

$$\left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right)^T = \begin{bmatrix} \sum \frac{\partial z_1}{y_q}\frac{\partial y_q}{\partial x_1} & \sum \frac{\partial z_1}{\partial y_q}\frac{\partial y_q}{\partial x_2} & \cdots & \sum \frac{\partial z_2}{\partial y_q}\frac{\partial y_q}{\partial x_n} \\ \sum \frac{\partial z_2}{\partial y_q}\frac{\partial y_q}{\partial x_1} & \sum \frac{\partial z_2}{\partial y_q}\frac{\partial y_q}{\partial x_2} & \cdots & \sum \frac{\partial z_2}{\partial y_q}\frac{\partial y_q}{\partial x_n} \\ \vdots & & & \\ \sum \frac{\partial z_m}{\partial y_q}\frac{\partial y_q}{\partial x_1} & \sum \frac{\partial z_m}{\partial y_q}\frac{\partial y_q}{\partial x_2} & \cdots & \sum \frac{\partial z_m}{\partial y_q}\frac{\partial y_q}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_1}{\partial y_2} & \cdots & \frac{\partial z_1}{\partial y_r} \\ \frac{\partial z_2}{\partial y_1} & \frac{\partial z_2}{\partial y_2} & \cdots & \frac{\partial z_2}{\partial y_r} \\ \vdots & & & \\ \frac{\partial z_m}{\partial y_1} & \frac{\partial z_m}{\partial y_2} & \cdots & \frac{\partial z_m}{\partial y_r} \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial y_r}{\partial x_1} & \frac{\partial y_r}{\partial x_2} & \cdots & \frac{\partial y_r}{\partial x_n} \end{bmatrix}$$

$$= \left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}}\right)^T \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^T = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\frac{\partial \mathbf{z}}{\partial \mathbf{y}}\right)^T.$$

On transposing both sides, we finally obtain $\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}}\frac{\partial \mathbf{z}}{\partial \mathbf{y}}$,

This is the chain rule for vectors (different from the conventional chain rule of calculus, the chain of matrices builds toward the left)

Example 2

x, y are as in Example 1 and z is a function of y defined as

$$z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix}, \text{ and } \begin{cases} z_1 = y_1^2 - 2y_2 \\ z_2 = y_2^2 - y_1 \\ z_3 = y_1^2 + y_2^2 \\ z_4 = 2y_1 + y_2 \end{cases}, \text{ we have}$$

$$\frac{\partial z}{\partial y} = \begin{pmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_2}{\partial y_1} & \frac{\partial z_3}{\partial y_1} & \frac{\partial z_4}{\partial y_1} \\ \frac{\partial z_1}{\partial y_2} & \frac{\partial z_2}{\partial y_2} & \frac{\partial z_3}{\partial y_2} & \frac{\partial z_4}{\partial y_2} \end{pmatrix} = \begin{pmatrix} 2y_1 & -1 & 2y_1 & 2 \\ -2 & 2y_2 & 2y_2 & 1 \end{pmatrix}.$$

Therefore,

$$\frac{\partial z}{\partial x} = \frac{\partial y}{\partial x}\frac{\partial z}{\partial y} = \begin{pmatrix} 2x_1 & 0 \\ -1 & 3 \\ 0 & 2x_3 \end{pmatrix} \begin{pmatrix} 2y_1 & -1 & 2y_1 & 2 \\ -2 & 2y_2 & 2y_2 & 1 \end{pmatrix} = \begin{pmatrix} 4x_1y_1 & -2x_1 & 4x_1y_1 & 4x_1 \\ -2y_1 - 6 & 1 + 6y_2 & -2y_2 + 6y_2 & 1 \\ -4x_3 & 4x_3y_2 & 4x_3y_2 & 2x_3 \end{pmatrix}$$

Maximum Likelihood (ML) Estimation

- To start our new topic, we introduce a simple question.

Tossing a coin. If the possibility to appear the head is μ, then fliping a coin is a Bernoulli Distribution with parameter μ.

Bernoulli(μ)

$P(X=1) = \mu$

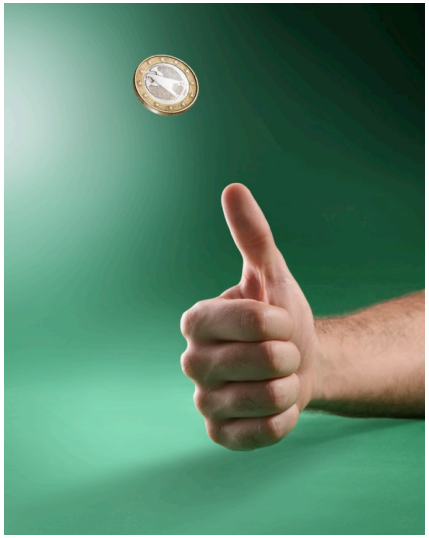
$P(X=0) = 1-\mu$

X is the random variable:

X=1 means the head appears; X=0 means the tail appears.

Maximum Likelihood (ML) Estimation

Suppose that $x_1, x_2, \ldots, x_n$ (i.i.d) represent the outcomes of n independent Bernoulli trials (for example, coin flipping), each with success probability μ.



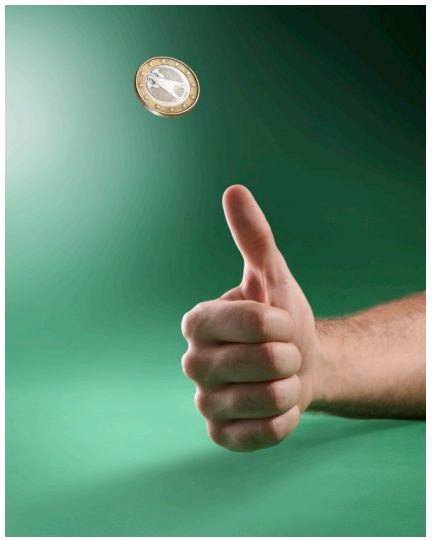
Question: assume μ is unknown, can we estimate μ by given data $x_1, x_2, \ldots, x_n$? **For which μ is $x_1, x_2, \ldots x_n$ most likely?**

Maximum Likelihood (ML) Estimation

Question: assume μ is unknown, can we estimate μ by given data $x_1, x_2, \ldots, x_n$? **For which μ is $x_1, x_2, \ldots x_n$ most likely?**

Why do we need to estimate the parameters?

Because in machine learning, to estimate sutiable and effective parameters is vital to build an effective model.



To address this issue, we introduce

Maximum Likelihood (ML) Estimation

Maximum Likelihood (ML) Estimation

- In many artificial intelligence and machine learning applications, the objective is **to estimate the model parameters from the given data**.

- For example, given a distribution class $P(X;\alpha)$, where α is a parameter from a parameter space. Now given data (x1,x2,...,xn) which are drawn from an unknown distribution $P(X;\alpha_0)$, we want to ask that how to select a suitable parameter α_0 by given data (x1,x2,...,xn)?

- The **Maximum Likelihood Estimation (MLE)** is one of the most widely used methods of estimating the parameters of a model.

Maximum Likelihood (ML) Estimation

- The method of Maximum Likelihood Estimation (MLE) selects the set of values of the model parameters that maximizes the likelihood function.
- In other words, the basic principle of MLE is to choose values that "explain" the data best by maximizing the probability of the data we've seen as a function of the parameters.
- The **Maximum Likelihood Estimation (MLE)** is one of the most widely used methods of estimating the parameters of a model.
- It answers the question: What values of parameters would make the observations most probable ?

Maximum Likelihood (ML) Estimation

- A distribution class $P(X;\alpha)$, where α is from a parameter space Δ.
- For each α from the space Δ, $P(X;\alpha)$ corresponds to a distribution.
- We have data $S=(x1,x2,\ldots,xn)$, which are drawn from an unknown distribution P(X), Independent and identically distributed (i.i.d.).

Problem: what is the optimal parameter α^* selected from the parameter space Δ, such that the seleced distribution $P(X;\alpha^*)$ is the most possible distribution sampling data S?

Maximum Likelihood (ML) Estimation: Example

- See the question introduced at the beginning.
- Suppose that $x_1, x_2, \ldots, x_n$ (i.i.d) represent the outcomes of n independent Bernoulli trials, each with success probability μ.

- The parameter α is μ.
- The parameter space Δ is $\{\mu: 0<\mu<1\}$.
- Data $S=(x_1, x_2, \ldots, x_n)$
- Distribution class $P(X;\mu)$ is: to each μ,
  $P(X=1;\mu) = \mu$; $P(X=0;\mu) = 1-\mu$
  So $P(X=x_i;\mu) = \mu^{x_i}(1-\mu)^{1-x_i}$

Maximum Likelihood (ML) Estimation: Example

- See the question introduced at the beginning.
- Suppose that $x_1, x_2, \ldots, x_n$ (i.i.d) represent the outcomes of n independent Bernoulli trials, each with success probability μ.

- The parameter α is μ.
- The parameter space Δ is $\{\mu: 0<\mu<1\}$.
- Data $S=(x_1, x_2, \ldots, x_n)$
- Distribution class $P(X;\mu)$ is: to each μ,
  $P(X=1;\mu) = \mu$; $P(X=0;\mu) = 1-\mu$
  So $P(X=x_i;\mu) = \mu^{x_i}(1-\mu)^{1-x_i}$

Maximum Likelihood (ML) Estimation

The selected distribution $P(X;\alpha^*)$ is the most possible distribution sampling data S=(x1,x2,...,xn), i.i.d..

Understanding above sentence, we can formulate it as follows:

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} P(x1, x2, \ldots, xn; \alpha)$$

here we assume $P(X; \alpha)$ is a discrete distribution.

- $\max_{\alpha \in \Delta} P(x1, x2, \ldots, xn; \alpha)$

  means the largest probability for $P(X; \alpha)$ that S is observed.

Maximum Likelihood (ML) Estimation

Because (x1,...,xn), are Independent and identically distributed,

$$\underset{\alpha \in \Delta}{\operatorname{argmax}}\, P(x1, x2, \ldots, xn; \alpha)$$

is equal to

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

How to solve this problem?

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha)$$

A commonly used method is to derivate $\prod_{i=1}^{n} P(X = xi; \alpha)$.

- However, it is difficult to derivate $\prod_{i=1}^{n} P(X = xi; \alpha)$ directly, because of the Multiplier operator $\prod_{i=1}^{n}$.

Maximum Likelihood (ML) Estimation

To reduce the affects of Multiplier operator $\prod_{i=1}^{n}$, we take a small trick (we use the property of log function to help us):

$$\log \prod_{i=1}^{n} ai = \sum_{i=1}^{n} \log ai$$

Step 1. We take log function.

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \log \prod_{i=1}^{n} P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \log \prod_{i=1}^{n} P(X = xi; \alpha)$$

Step 2. Using the property of log function:

$$\log \prod_{i=1}^{n} P(X = xi; \alpha) = \sum_{i=1}^{n} \log P(X = xi; \alpha)$$

Maximum Likelihood (ML) Estimation

Therefore,

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \prod_{i=1}^{n} P(X = xi; \alpha) = \underset{\alpha \in \Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = xi; \alpha)$$

Step 3. We need to optimize

$$\underset{\alpha \in \Delta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X = xi; \alpha) \quad (1)$$

and obtain the optimal solution.

The solution of Eq. (1) is called Maximum Likelihood Estimation.

Thank You!