Note: This is a small overview of what we discussed during the tutorial session. To get the full picture, please attend the session and ask questions.

1 Linear System of Equations

We talked about how to solve $Ax = b$ during the class and for Task 1, we did $A = X^\top X$ and $b = X^\top y$. (x is θ in that case).

We solved it by taking the inverse of A. The other solutions would be elimination (RREF - Reduced row-echelon form), QR Decomposition, and so on. These methods are used because of numerical stability, computational efficiency, and memory consideration.

You can watch QR decomposition from the link Click here .

Example:

Suppose $A = \begin{pmatrix} 14 & 6 \\ 6 & 3 \end{pmatrix}$ and b $= \begin{pmatrix} 29 \\ 13 \end{pmatrix}$

$$\begin{pmatrix} 14 & 6 \\ 6 & 3 \end{pmatrix} \theta = \begin{pmatrix} 29 \\ 13 \end{pmatrix}$$

We create an augmented matrix and we do operations such as multiplication and addition. We should get an identity matrix where pivots are 1 and below and upper pivots all are zero. We do the same operations to the right side to get theta values.

$$\left[\begin{array}{cc|c} 14 & 6 & 29 \\ 6 & 3 & 13 \end{array}\right] \rightsquigarrow^{\frac{1}{14}R_1 \to R_1} \left[\begin{array}{cc|c} 1 & \frac{3}{7} & \frac{29}{14} \\ 6 & 3 & 13 \end{array}\right] \rightsquigarrow^{-6R_1 + R_2 \to R_2} \left[\begin{array}{cc|c} 1 & \frac{3}{7} & \frac{29}{14} \\ 0 & \frac{3}{7} & \frac{4}{7} \end{array}\right] \rightsquigarrow^{-R_2 + R_1 \to R_1} \left[\begin{array}{cc|c} 1 & 0 & \frac{3}{2} \\ 0 & \frac{3}{7} & \frac{4}{7} \end{array}\right]$$

$$\rightsquigarrow^{\frac{7}{3}R_2 \to R_2} \left[\begin{array}{cc|c} 1 & 0 & \frac{3}{2} \\ 0 & 1 & \frac{4}{3} \end{array}\right]$$

2 Matrices and Vectors

The linear combinations of matrix columns when we multiply matrix with a vector. It should come on the right side. (Dimensions should always match $3 * 3 \times 3 * 1 = 3 * 1$).

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 0 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 7 \end{bmatrix} = 2 * \text{col-1} + 3 * \text{col-2} + 7 * \text{col-3} = 2 * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + 3 * \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix} + 7 * \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix} = \begin{bmatrix} 15 \\ 11 \\ 36 \end{bmatrix}$$

The linear combinations of matrix rows when we multiply vector with matrix. It should come on the left side. (Dimensions should always match $1 * 3 \times 3 * 3 = 1 * 3$).

$$\begin{bmatrix} 2 & 3 & 7 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 0 \\ 1 & 2 & 4 \end{bmatrix} = 2 * \text{row-1} + 3 * \text{row-2} + 7 * \text{row-3} = 2 * \begin{bmatrix} 1 & 2 & 1 \end{bmatrix} + 3 * \begin{bmatrix} 1 & 3 & 0 \end{bmatrix} + 7 * \begin{bmatrix} 1 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 12 & 27 & 30 \end{bmatrix}$$

Matrix Multiplication:

First way:

When we multiply A matrix (3×3 matrix) by B matrix (3×4 matrix), we get C matrix (3×4) where $C = AB$.

To get value in the third row and fourth column of C:

C_{34} = (row 3 of A) × (col 4 of B) = $\sum_{k=1}^{3} a_{3k}b_{k4}$

General formula: $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$

Second way:

Columns of C matrix are combinations of columns of A matrix when we multiply matrix A by matrix B. (Dimensions should always match $3*3 \times 3*4 = 3*4$).

$$\begin{bmatrix}1&2&1\\1&3&0\\1&2&4\end{bmatrix}\begin{bmatrix}1&2&1&0\\1&3&0&1\\1&2&4&2\end{bmatrix} = \left[\begin{bmatrix}1&2&1\\1&3&0\\1&2&4\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}1&2&1\\1&3&0\\1&2&4\end{bmatrix}\begin{bmatrix}2\\3\\2\end{bmatrix}, \begin{bmatrix}1&2&1\\1&3&0\\1&2&4\end{bmatrix}\begin{bmatrix}1\\0\\4\end{bmatrix}, \begin{bmatrix}1&2&1\\1&3&0\\1&2&4\end{bmatrix}\begin{bmatrix}0\\1\\2\end{bmatrix}\right] = \begin{bmatrix}4&10&5&4\\4&11&1&3\\7&16&17&10\end{bmatrix}$$

Third way:

Rows of C matrix are combinations of rows of B matrix when we multiply matrix A by matrix B. (Dimensions should always match $3*3 \times 3*4 = 3*4$).

In the same way above, we take rows of A matrix and multiply with B. As an example, first row of A multiplied by B matrix will give first row of C matrix.

3 Basic Properties and Notations

Some properties and notations about matrices:

$$\begin{aligned}(A+B)^\top &= A^\top + B^\top\\ (A+B)^{-1} &\neq A^{-1} + B^{-1}\\ (AB)^\top &= B^\top A^\top\\ (AB)^{-1} &= B^{-1}A^{-1}\\ (A^\top)^\top &= A\end{aligned}$$

Some notations about norms:

$$\begin{aligned}||x||_p &= (\sum_{i=1}^{N} |x_i|^p)^{\frac{1}{p}}\text{- general formula with respect to p}\\ ||x||_1 &= |x_1| + |x_2| + \ldots + |x_i|\text{- called Taxi Cab norm, also L1 norm (p = 1)}\\ ||x||_2 &= \sqrt{x_1^2 + x_2^2 + \ldots + x_i^2}\text{- called Euclidean norm (p = 2)}\\ ||x||_2^2 &= x_1^2 + x_2^2 + \ldots + x_i^2\text{- called squared Euclidean norm}\\ ||x||_\infty &= max(x_1, x_2, \ldots, x_i)\text{- infinity max norm (p = }\infty)\end{aligned}$$

4 Differentiation and Gradients

First way:
It is from book with Chapter 5.2 - 5.4 Click here.

Here we use numerator layout. We should describe gradient as a row vector, not a column vector.

Def: For a function $f : \mathcal{R}^n \to \mathcal{R}$, $x \mapsto f(x), x \in \mathcal{R}^n$ of n variables $x_1, \ldots, x_n$, we define partial derivatives and collect in the row vector as:

$$gradf = \frac{df}{dx} = \begin{bmatrix}\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \ldots & \frac{\partial f}{\partial x_n}\end{bmatrix} \in \mathcal{R}^{1\times n}$$

Def: (Jacobian) The collection of all first-order partial derivatives of a vector-valued function $\mathbf{f} : \mathcal{R}^n \to \mathcal{R}^m$ is called the Jacobian. The Jacobian $\mathbf{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

$$\mathbf{J} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \dots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix} \in \mathcal{R}^{m \times n} \text{ where } \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathcal{R}^n$$

Example:

Consider a linear model $y = X\theta$ where $X \in \mathcal{R}^{N \times D}, \theta \in \mathcal{R}^D,$ and $y \in \mathcal{R}^N$.

We define function: $L(e) = ||e||_2^2$ where $e(\theta) = y - X\theta$. L is called a least-squares loss function.

$\frac{\partial L}{\partial \theta} \in \mathcal{R}^{1 \times D}$ because L is in R and θ is in $\mathcal{R}^D$.

Chain rule: $\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial e} \frac{\partial e}{\partial \theta}$

We know that $||e||^2 = e^\top e$ and we get:

$\frac{\partial L}{\partial e} = 2e^\top \in R^{1 \times N}$ because because L is in R and e is in $\mathcal{R}^N$. That's why we get transpose to match dimension.

We obtain:

$\frac{\partial e}{\partial \theta} = -X \in \mathcal{R}^{N \times D}$ because e is in $\mathcal{R}^N$ and θ is in $\mathcal{R}^D$. That's why it is X, not transpose of X to match dimension.

Combining all:

$\frac{\partial L}{\partial \theta} = -2e^\top X = -2(y - X\theta)^\top X \in \mathcal{R}^{1 \times D}$

We want to have now column vector as $\mathcal{R}^{D \times 1}$, so take transpose by using notation $(AB)^\top = B^\top A^\top$:

$\frac{\partial L}{\partial \theta} = -2(e^\top X)^\top = -2X^\top e = -2X^\top (y - X\theta)$.

You can also read Chapters 5.45.5 to understand gradients of matrices.

Second way:

It is from the tutorial session of Randolf Scholz.

Riesz representation theorem and tensor representations are explained in the class. We also showed the transpose of a linear map. Here you can find an example solution:

Example:

- Represent the k-th left accumulated derivative as an inner product.
- Express next derivative as a linear map x.
- Combine it with the previous term to get the next accumulated
- Use the properties of the transpose to isolate the delta-term in the ket-competent (right side $|v>$). Then the resulting bra-component (left side $< u|$) is the gradient!

$L = \frac{1}{N} \sum_{n=1}^{N} (y - f_\theta(x))^2$

$\frac{dL}{d\theta} = \frac{dL}{df_\theta(x)} \frac{df_\theta(x)}{\theta} = [\Delta f_\theta(x) \mapsto \left\langle \frac{2}{N}(f_\theta(x) - y) \middle| \Delta f_\theta(x) \right\rangle] \odot [\Delta\theta \mapsto X\Delta\theta] = [\Delta\theta \mapsto \left\langle \frac{2}{N}(f_\theta(x) - y) \middle| X\Delta\theta \right\rangle] = [\Delta\theta \mapsto \left\langle X^\top \frac{2}{N}(f_\theta(x) - y) \middle| \Delta\theta \right\rangle]$ which gives us gradient as $\frac{2}{N} X^\top (X\theta - y)$.

5 Positive Semidefinite Matrices

Def: A square symmetric matrix $H \in \mathcal{R}^{n \times n}$ is positive semidefinite if $v^\top Hv \geq 0$ and positive definite if $v^\top Hv > 0$ for all $v \neq 0$.

The following conditions are equivalent:

i) $v^\top Hv \geq 0$ for all v

ii) All eigenvalues of H are non-negative.

It implies that the matrix behaves like a positive scalar value under transformations, indicating a kind of stretching or contracting rather than rotation in certain directions. (All changes are in the same direction (let's say in the first quadrant)).

We need it in image processing, and computer graphics to know how matrices transfer shapes; we also need it in convexity and other ML problems.