Institute for Advanced Studies
in Basic Sciences
Gava Zang, Zanjan, Iran

Gradient Descent - Matrix Derivatives

Presented by: Elaheh Sheikhi (elahehsheikhi@iasbs.ac.ir)
Supervised by: Dr. Parvin Razzaghi (p.razzaghi@iasbs.ac.ir)

Winter 2024

What is a Loss Function?

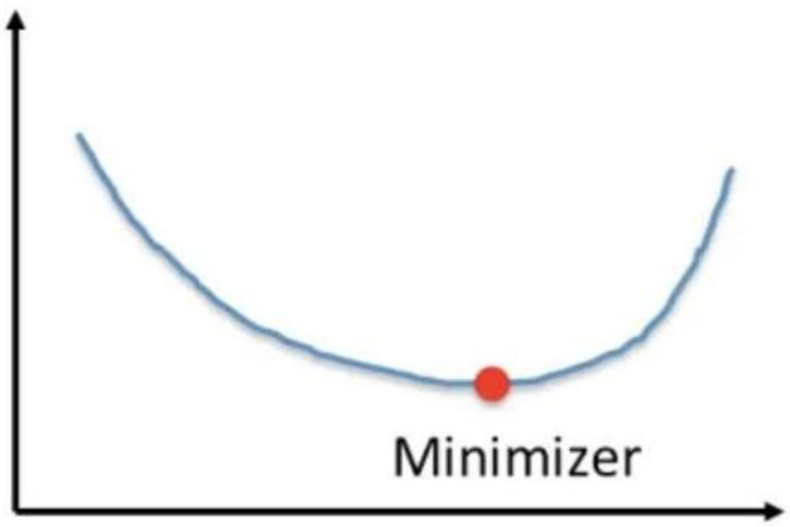
- **Definition:** A metric to measure the gap between model predictions and actual values.
- **Purpose:** Guides optimization algorithms to minimize error.
- **Example:** Predicting tomorrow's temperature and calculating the error.
- **Importance:**
  - Identifies model parameters needing adjustment.
  - Ensures consistent improvement in model performance.

$$\min_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\theta})$$

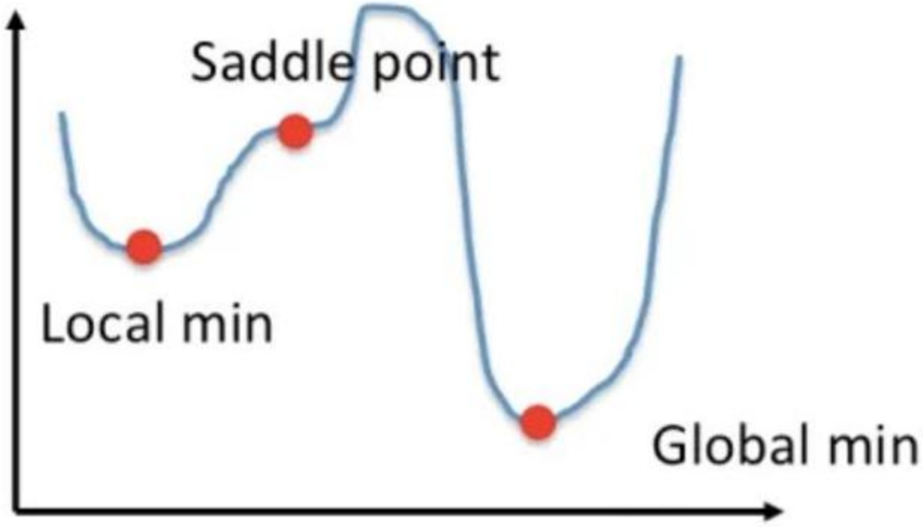
$$\boldsymbol{\theta}^* = \boldsymbol{argmin_{\theta} J}(\boldsymbol{\theta})$$

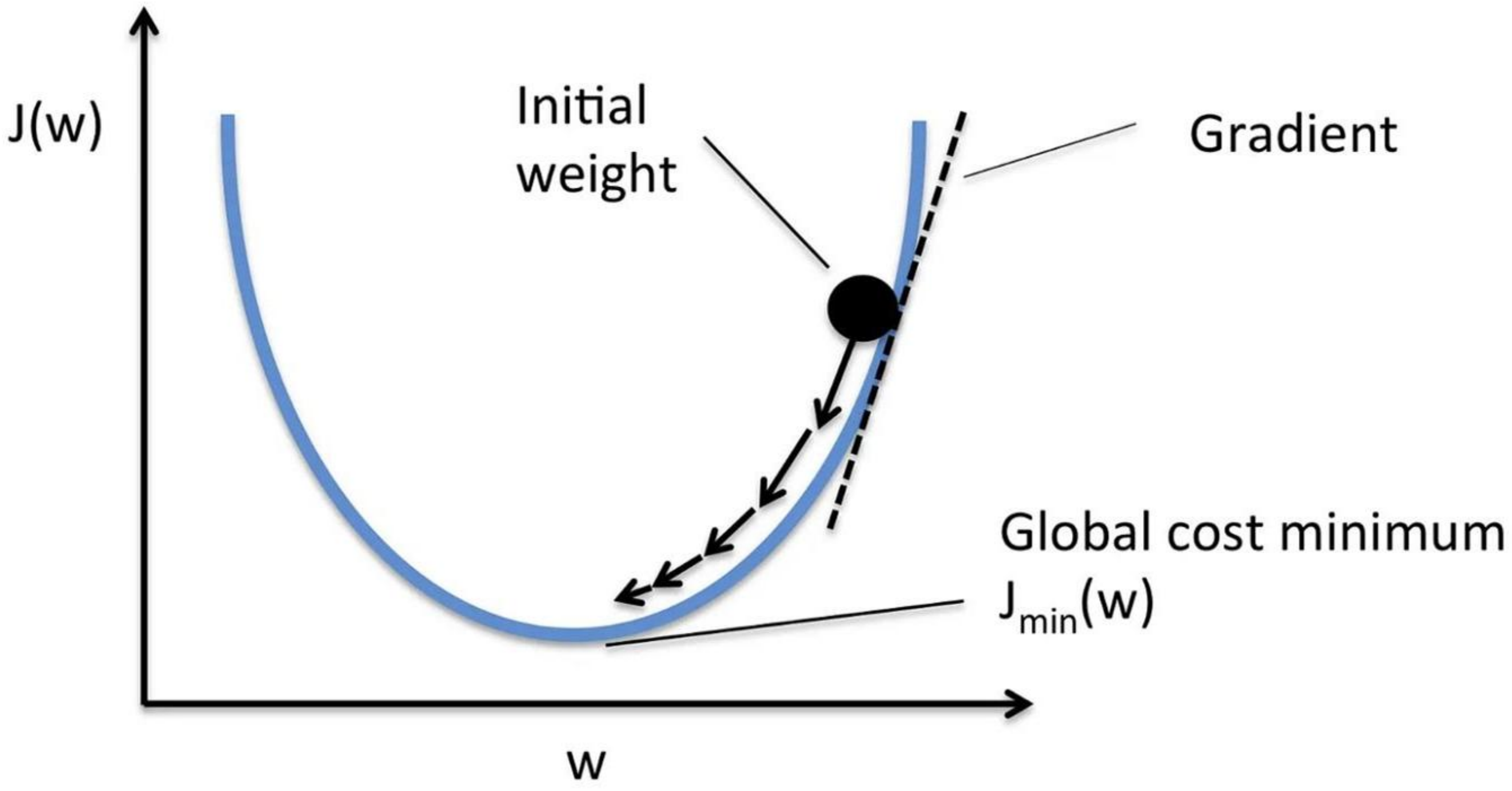
- **Simple Derivation:**
  - This represents the derivative of a function f(x) with respect to a single variable x
- **Gradient Calculation:**
  - This is the gradient of a function f(x1,x2,…,xn), which is a vector of partial derivatives with respect to each variable.

Convex



Non-Convex

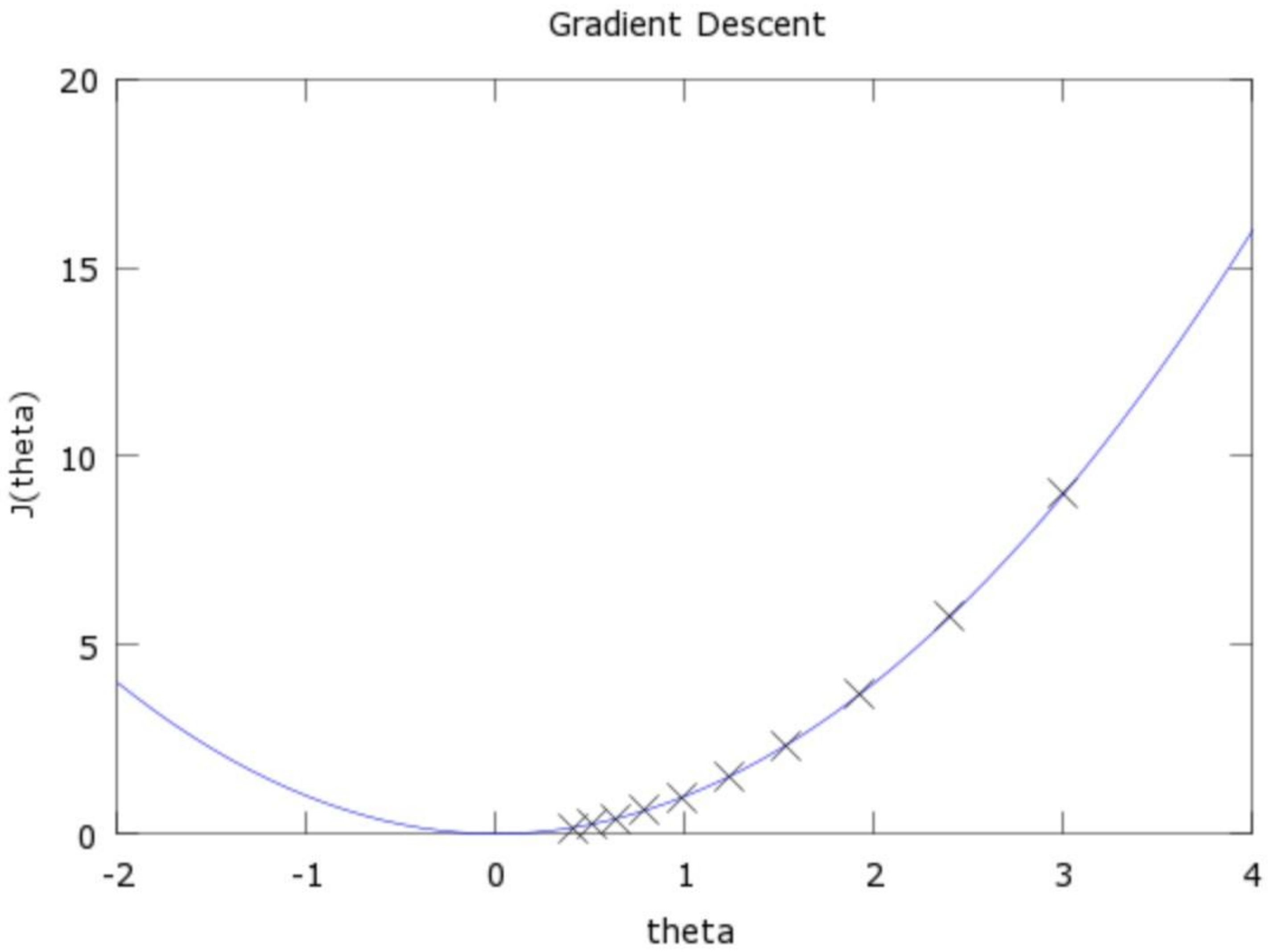




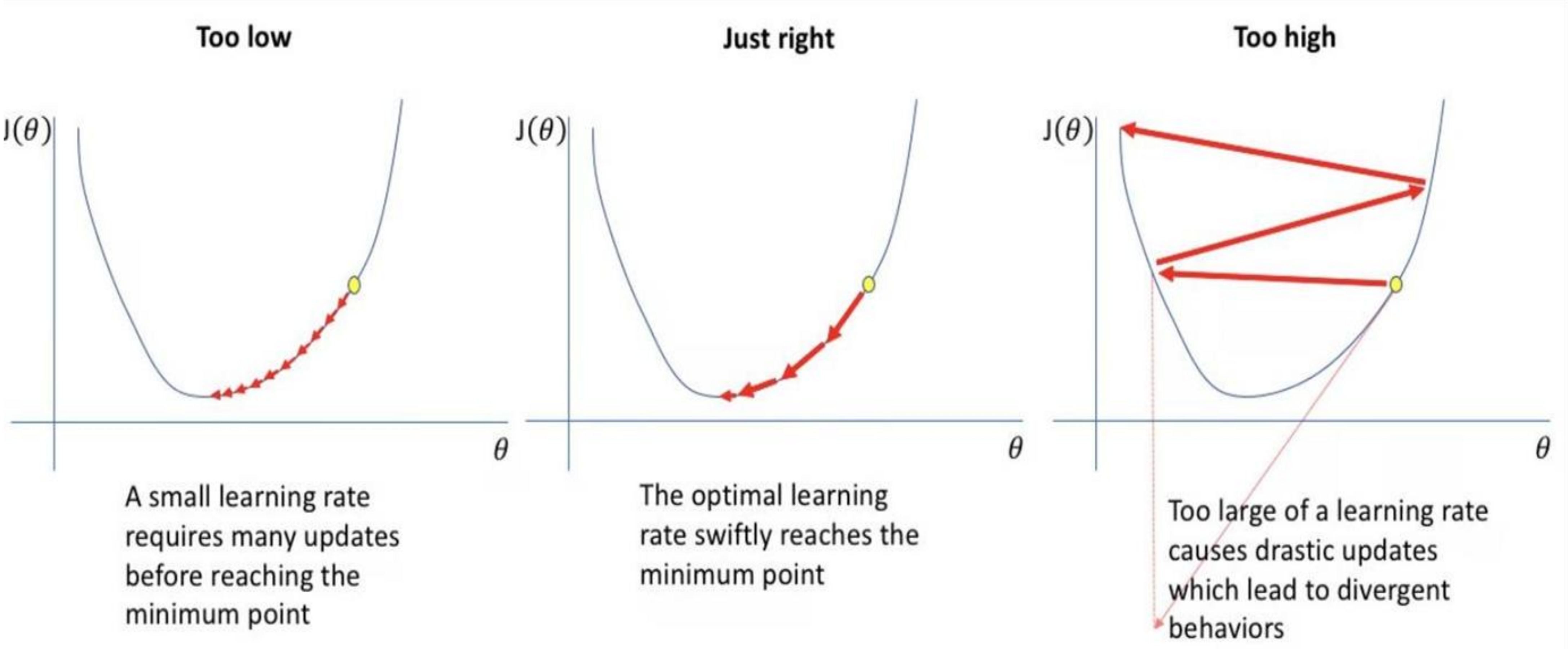
From: https://hackernoon.com/gradient-descent-aynk-7cbe95a778da

$$\theta = \theta - \eta * \nabla_\theta J(\theta)$$

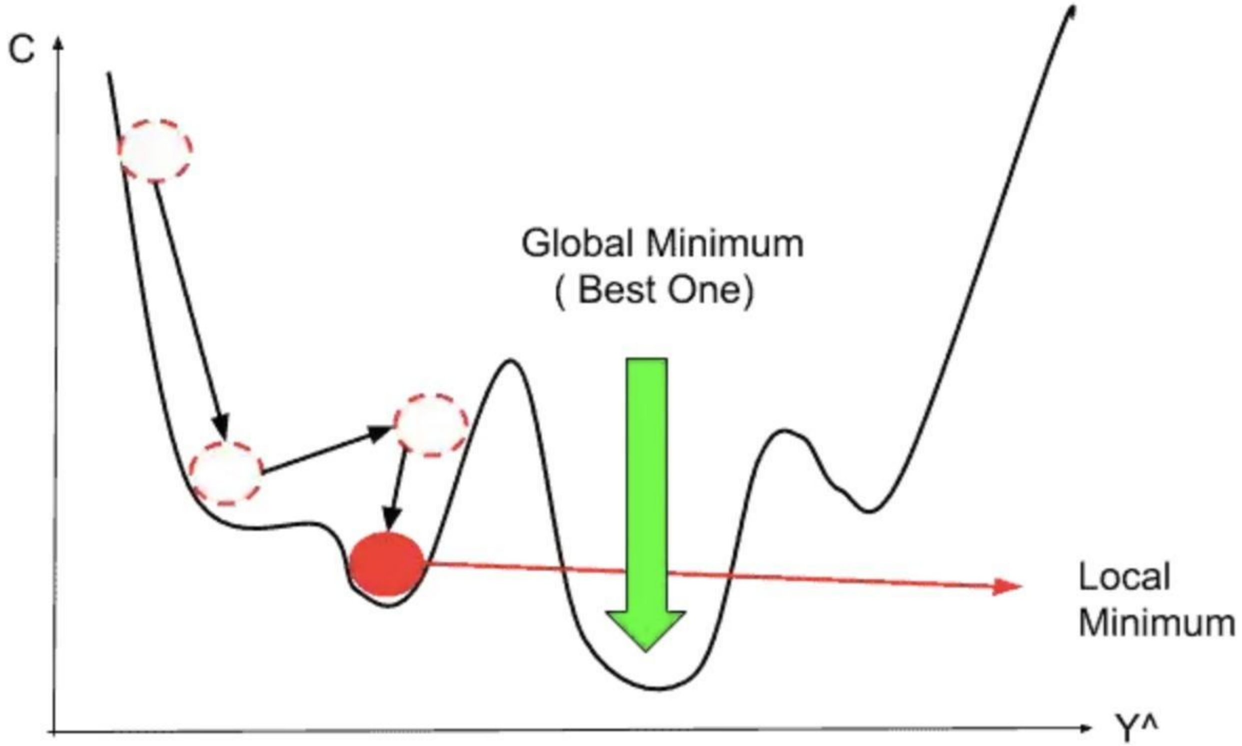
notes:



Itera-tion	θ	$\alpha \frac{d}{d\theta} J(\theta)$
1	3	0.6
2	2.4	0.48
3	1.92	0.384
4	1.536	0.307
5	1.229	0.246
6	0.983	0.197
7	0.786	0.157
8	0.629	0.126
9	0.503	0.101
10	0.403	0.081


Too low
Just right
Too high
$J(\theta)$
$J(\theta)$
$J(\theta)$
θ
θ
θ
A small learning rate requires many updates before reaching the minimum point
The optimal learning rate swiftly reaches the minimum point
Too large of a learning rate causes drastic updates which lead to divergent behaviors

Challenge in Gradient Descent:



6 Common Matrix derivatives:

Type	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

Derivatives by Scalar

Numerator Layout Notation

$$\frac{\partial y}{\partial x}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}$$

Denominator Layout Notation

$$\frac{\partial y}{\partial x}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix} \equiv \frac{\partial \mathbf{y}}{\partial x}^\top$$

Derivatives by Vector

Numerator Layout Notation

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \equiv \frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top}$$

Denominator Layout Notation

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \equiv \frac{\partial \mathbf{y}^\top}{\partial \mathbf{x}}$$

Derivative by Matrix

Numerator Layout Notation

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{m1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1n}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}$$

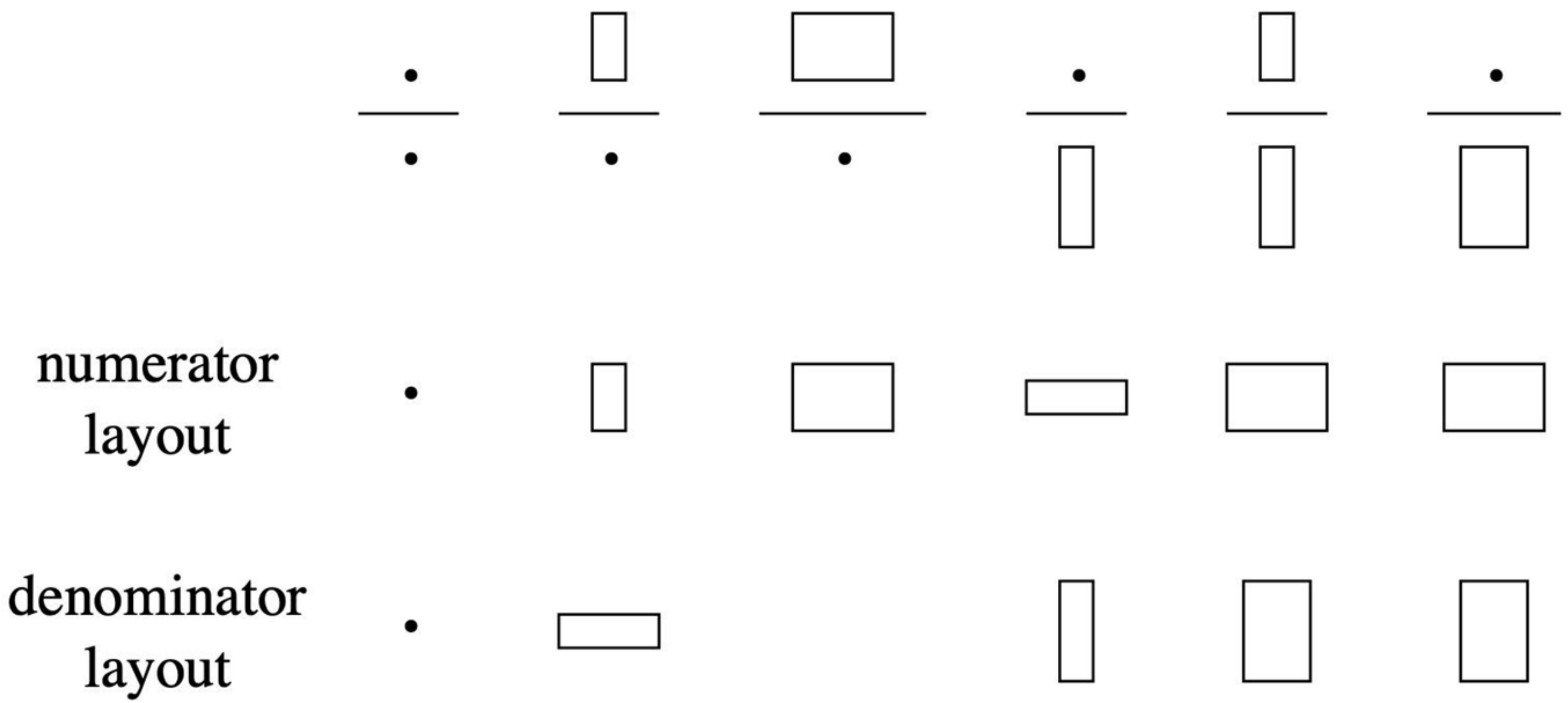
$$\equiv \frac{\partial y}{\partial \mathbf{X}^\top}$$

Denominator Layout Notation

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}$$

$$\equiv \frac{\partial y}{\partial \mathbf{X}}$$

Pictorial Representation



Here, scalar a, vector $\mathbf{a}$ and matrix $\mathbf{A}$ are not functions of x and $\mathbf{x}$.

$$\text{(C1)} \qquad \frac{d\mathbf{a}}{dx} = \mathbf{0} \quad \text{(column matrix)}$$

$$\text{(C2)} \qquad \frac{da}{d\mathbf{x}} = \mathbf{0}^\top \quad \text{(row matrix)}$$

$$\text{(C3)} \qquad \frac{da}{d\mathbf{X}} = \mathbf{0}^\top \quad \text{(matrix)}$$

$$\text{(C4)} \qquad \frac{d\mathbf{a}}{d\mathbf{x}} = \mathbf{0} \quad \text{(matrix)}$$

$$\text{(C5)} \qquad \frac{d\mathbf{x}}{d\mathbf{x}} = \mathbf{I}$$

$$\text{(C6)} \qquad \frac{d\,\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \frac{d\,\mathbf{x}^\top\mathbf{a}}{d\mathbf{x}} = \mathbf{a}^\top$$

$$\text{(C7)} \qquad \frac{d\mathbf{x}^\top\mathbf{x}}{d\mathbf{x}} = 2\,\mathbf{x}^\top$$

$$\text{(C8)} \qquad \frac{d(\mathbf{x}^\top\mathbf{a})^2}{d\mathbf{x}} = 2\,\mathbf{x}^\top\mathbf{a}\,\mathbf{a}^\top$$

$$\text{(C9)} \qquad \frac{d\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{A}$$

$$\text{(C10)} \qquad \frac{d\,\mathbf{x}^\top\mathbf{A}}{d\mathbf{x}} = \mathbf{A}^\top$$

$$\text{(C11)} \qquad \frac{d\,\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$$

$$\text{(SS1)} \qquad \frac{\partial(u+v)}{\partial x} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial x}$$

$$\text{(SS2)} \qquad \frac{\partial uv}{\partial x} = u\frac{\partial v}{\partial x} + v\frac{\partial u}{\partial x} \qquad \text{(product rule)}$$

$$\text{(SS3)} \qquad \frac{\partial g(u)}{\partial x} = \frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x} \qquad \text{(chain rule)}$$

$$\text{(SS4)} \qquad \frac{\partial f(g(u))}{\partial x} = \frac{\partial f(g)}{\partial g}\frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x} \qquad \text{(chain rule)}$$

(VS1) $$\frac{\partial a\mathbf{u}}{\partial x} = a\frac{\partial \mathbf{u}}{\partial x}$$

where a is not a function of x.

(VS2) $$\frac{\partial \mathbf{A}\mathbf{u}}{\partial x} = \mathbf{A}\frac{\partial \mathbf{u}}{\partial x}$$

where $\mathbf{A}$ is not a function of x.

(VS3) $$\frac{\partial \mathbf{u}^\top}{\partial x} = \left(\frac{\partial \mathbf{u}}{\partial x}\right)^\top$$

(VS4) $$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial x} = \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial x}$$

$$\text{(VS5)} \qquad \frac{\partial \mathbf{g}(\mathbf{u})}{\partial x} = \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial x} \qquad \text{(chain rule)}$$

with consistent matrix layout.

$$\text{(VS6)} \qquad \frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial x} = \frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}} \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial x} \qquad \text{(chain rule)}$$

with consistent matrix layout.

notes:

Any questions?