1 Definitions For All

(a) **Partial Derivative**: Referred to when we take a derivative of a multi-variable function with respect to one of its input. Example: $f(x_1, x_2) = x_1^3 + e^{x_2}x_1$. If we take derivative with respect to x_1, this is denoted as

$$\frac{\partial f}{\partial x_1}$$

We treat other variables as constant so,

$$= \frac{\partial(x_1^3 + e^{x_2}x_1)}{\partial x_1} = 3x_1^2 + e^{x_2}$$

(b) **Vector Derivative:** We can express the above as a column vector where the i-th entry is the partial derivative of the function w.r.t i-th input entry in the input vector. So, $f : \mathbb{R}^2 \to \mathbb{R}$.

$$\nabla f(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 3x_1^2 + e^{x_2} \\ x_1 e^{x_2} \end{bmatrix}$$

Note: On LHS, $\mathbf{x}$ is in bold as this indicates a vector, NOT a scalar.

(c) **Matrix Derivative:** Above, we discussed derivative of a vector w.r.t to a scalar. But, we can also take derivative of a vector w.r.t another vector.
Consider a vector-valued function $\mathbf{y} = f(\mathbf{x})$ where $f : \mathbb{R}^M \to \mathbb{R}^N$. Then, the derivative $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$is defined as

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} \cdots \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} \cdots \frac{\partial y_n}{\partial x_2} \\ \vdots & \ddots \\ \frac{\partial y_1}{\partial x_m} & \frac{\partial y_2}{\partial x_m} \cdots \frac{\partial y_n}{\partial x_m} \end{bmatrix}$$

(d) **Numerator vs. Denominator Layout:** There are two different layouts to express vector/matrix derivatives, namely the numerator and the denominator layout. In this course, we will always specify which layout to use. These layouts are mostly the same and can easily be switched using transpose operations. To demonstrate this better, some examples are shown below:

	Numerator Layout	**Denominator Layout**
$\frac{\partial y}{\partial \mathbf{x}}$	1-D row vector	1-D column vector
$\frac{\partial \mathbf{y}}{\partial x}$	1-D column vector	1-D row vector
$\frac{\partial \mathbf{u}^T\mathbf{v}}{\partial \mathbf{y}}$	$\mathbf{u^T}$	$\mathbf{u}$
$\frac{\partial A\mathbf{v}}{\partial \mathbf{v}}$	A	A^T

Note: If the variables are in bold, they are vectors. A capital variable denotes a matrix. Otherwise, they are scalars. Knowing $A \in \mathbb{R}^{M \times N}$ and $\mathbf{v} \in \mathbb{R}^{N \times 1}$, verify that the dimensions make sense for both the layouts.

A handy way to distinguish numerator vs. denominator layout is to remember that **the layout type corresponds to the number of rows in the output matrix**. In numerator layout, the output matrix has number of rows equal to the size of the numerator, while in denominator layout, the output matrix has number of rows equal to the size of the denominator.

(e) **Shortcuts**: You're in some luck as we've compiled some shortcuts here that you can refer to whenever you're stuck. This will come in handy a lot in ML.

	Numerator layout	Denominator layout	Notes
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}$	I_N	I_N	$\mathbf{v} \in \mathbb{R}^N$
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T$	I_N	I_N	$\mathbf{v} \in \mathbb{R}^N$
$\frac{\partial}{\partial \mathbf{v}}\ t\mathbf{v}$	tI_N	tI_N	$\mathbf{v} \in \mathbb{R}^N$
$\frac{\partial}{\partial \mathbf{u}}\ \mathbf{u}^T\mathbf{v}$	$\mathbf{v}^T$	$\mathbf{v}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{u}^T\mathbf{v}$	$\mathbf{u}^T$	$\mathbf{u}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T\mathbf{v}$	$2\mathbf{v}^T$	$2\mathbf{v}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T A\mathbf{v}$	$\mathbf{v}^T(A + A^T)$	$(A + A^T)\mathbf{v}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T A\mathbf{v}$	$2\mathbf{v}^T A$	$2A\mathbf{v}$	If $A = A^T$
$\frac{\partial}{\partial \mathbf{v}}\ A\mathbf{v}$	A	A^T	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T A$	A^T	A	

Super Important: Always verify that your dimensions match on both sides of the equality!

2 Deriving Vector Derivatives

1. Let $J(\mathbf{z}) = ||\mathbf{y} - \mathbf{z}||_2^2$ where $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$. Find $\nabla J(\mathbf{z})$.

$$
\begin{aligned}
J(\mathbf{z}) &= \|\mathbf{y} - \mathbf{z}\|_2^2 \\
&= \left\| \begin{bmatrix} y_1 - z_1 \\ y_2 - z_2 \end{bmatrix} \right\|_2^2 \\
&= (y_1 - z_1)^2 + (y_2 - z_2)^2 \\
\frac{\delta J}{\delta z_1} &= -2(y_1 - z_1) \\
\frac{\delta J}{\delta z_2} &= -2(y_2 - z_2) \\
\nabla J(\mathbf{z}) &= -2(\mathbf{y} - \mathbf{z})
\end{aligned}
$$

2. Suppose we have a function that takes in a vector $\mathbf{x} \in \mathbb{R}^{3\times 1}$ and squares each element individually, returning another vector, $\mathbf{y} = f(\mathbf{x})$.

$$
f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{bmatrix}
$$

What is $\frac{\delta \mathbf{y}}{\delta \mathbf{x}}$? Use numerator layout.

Intuitively, one might think $\frac{\delta \mathbf{y}}{\delta \mathbf{x}} = 2\mathbf{x}$. However, we can quickly see that this is not the case because the dimensions wouldn't make sense. $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ has shape 3×3 whereas $\mathbf{x}$ has shape 3×1

$$
\begin{aligned}
\frac{\delta \mathbf{y}}{\delta \mathbf{x}} &= \begin{bmatrix} \frac{\delta y_1}{\delta x_1} & \frac{\delta y_1}{\delta x_2} & \frac{\delta y_1}{\delta x_3} \\ \frac{\delta y_2}{\delta x_1} & \frac{\delta y_2}{\delta x_2} & \frac{\delta y_2}{\delta x_3} \\ \frac{\delta y_3}{\delta x_1} & \frac{\delta y_3}{\delta x_2} & \frac{\delta y_3}{\delta x_3} \end{bmatrix} \\
&= \begin{bmatrix} 2x_1 & 0 & 0 \\ 0 & 2x_2 & 0 \\ 0 & 0 & 2x_3 \end{bmatrix}
\end{aligned}
$$

3 How About a Proof?

1. Assuming denominator layout, prove $\frac{\delta}{\delta \mathbf{v}} \mathbf{v}^T A \mathbf{v} = (A^T + A)\mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^2$ and $A \in \mathbb{R}^{2\times 2}$.

   Hint: Start by expanding $A\mathbf{v}$ and then expanding $\mathbf{v}^T A \mathbf{v}$.

   Hint: Write out the scalar partial derivatives.

   Hint: Work both top-down and bottom-up, i.e. also expand the right side so you can see where you are going.

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$
$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$
$$A^T = \begin{bmatrix} A_{1,1} & A_{2,1} \\ A_{1,2} & A_{2,2} \end{bmatrix}$$

$$\begin{aligned} \mathbf{v}^T A \mathbf{v} &= \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &= \begin{bmatrix} A_{1,1}v_1 + A_{2,1}v_2 & A_{1,2}v_1 + A_{2,2}v_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &= A_{1,1}v_1^2 + A_{2,1}v_1v_2 + A_{1,2}v_1v_2 + A_{2,2}v_2^2 \end{aligned}$$

$$\begin{aligned} \frac{\delta}{\delta v_1} \mathbf{v}^T A \mathbf{v} &= 2A_{1,1}v_1 + A_{2,1}v_2 + A_{1,2}v_2 \\ \frac{\delta}{\delta v_2} \mathbf{v}^T A \mathbf{v} &= A_{2,1}v_1 + A_{1,2}v_1 + 2A_{2,2}v_2 \\ \frac{\delta}{\delta \mathbf{v}} \mathbf{v}^T A \mathbf{v} &= \begin{bmatrix} 2A_{1,1}v_1 + A_{2,1}v_2 + A_{1,2}v_2 \\ A_{2,1}v_1 + A_{1,2}v_1 + 2A_{2,2}v_2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} (A^T + A)\mathbf{v} &= \begin{bmatrix} 2A_{1,1} & A_{1,2} + A_{2,1} \\ A_{1,2} + A_{2,1} & 2A_{2,2} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &= \begin{bmatrix} 2A_{1,1}v_1 + A_{1,2}v_1 + A_{2,1}v_2 \\ A_{1,2}v_1 + A_{2,1}v_1 + 2A_{2,2}v_2 \end{bmatrix} \end{aligned}$$

It turns out this statement holds for vectors and matrices of any dimension. Can you prove the same statement for the general case?

4 Take it Back Now, Y'all

How does this all relate to linear regression? As you can recall from lecture, the objective function for linear regression is

$$J(\boldsymbol{\theta}) = \frac{1}{N} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \tag{1}$$

After expanding, this is equivalent to

$$J(\boldsymbol{\theta}) = \frac{1}{N} \left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T X^T\mathbf{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right) \tag{2}$$

Our goal is find a closed-form solution to our objective function. In other words, we can find the best $\boldsymbol{\theta}$ that minimizes the equation above. Let's do this step by step.

1. Thanks to your awesome proving skills, we now know

$$\frac{\delta}{\delta \mathbf{v}} \mathbf{v}^T A\mathbf{v} = (A^T + A)\mathbf{v} \tag{3}$$

We'll also assume

$$\frac{\delta \mathbf{v}^T\mathbf{u}}{\delta \mathbf{v}} = \mathbf{u} \tag{4}$$

(try proving this on your own). Using these facts, find $\nabla J(\boldsymbol{\theta})$, the derivative of the objective function with respect to $\boldsymbol{\theta}$.

$$\begin{aligned}
\nabla J(\boldsymbol{\theta}) &= \frac{\delta}{\delta\boldsymbol{\theta}} \left(\frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T X^T\mathbf{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right)\right) \\
&= \frac{1}{N}\left(0 - 2X^T\mathbf{y} + \frac{\delta}{\delta\boldsymbol{\theta}}\boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right) && \text{by (4)} \\
&= \frac{1}{N}\left(0 - 2X^T\mathbf{y} + (X^T X + \left(X^T X\right)^T)\boldsymbol{\theta}\right) && \text{by (3)} \\
&= \frac{1}{N}\left(0 - 2X^T\mathbf{y} + 2X^T X\boldsymbol{\theta}\right) \\
&= \frac{2}{N}\left(-X^T\mathbf{y} + X^T X\boldsymbol{\theta}\right)
\end{aligned}$$

2. What is the closed-form solution of the objective function?

$$\begin{aligned}
0 &= \frac{2}{N}\left(-X^T\mathbf{y} + X^T X\boldsymbol{\theta}\right) \\
&= -X^T\mathbf{y} + X^T X\boldsymbol{\theta} \\
X^T X\boldsymbol{\theta} &= X^T\mathbf{y} \\
\boldsymbol{\theta} &= \left(X^T X\right)^{-1} X^T\mathbf{y}
\end{aligned}$$