

# Explanatory Notes for 6.390

Shauntclair Ruiz (Current TA)

Fall 2022

## X. Matrix Derivatives

In general, we want to be able to combine the powers of matrices and calculus:

- **Matrices:** the ability to store lots of **data**, and do fast linear operations on all that data at the **same time**.

**Example:** Consider

$$\mathbf{w}^T \mathbf{x} = \begin{bmatrix} w_1 & w_2 & \cdots & w_m \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \sum_{i=1}^m x_i w_i \quad (1)$$

In this case, we're able to do  $m$  different **multiplications** at the same time! This is what we like about matrices.

In this case, we're thinking about vectors as  $(m \times 1)$  matrices.

- **Calculus:** analyzing the way different variables are **related**: how does changing  $x$  affect  $y$ ?

**Example:** Suppose we have

$$\frac{\partial f}{\partial x_1} = 10 \quad \frac{\partial f}{\partial x_2} = -5 \quad (2)$$

Now we know that, if we increase  $x_1$ , we increase  $f$ . This **understanding** of variables is what we like about derivatives.

### Concept 1

**Matrix derivatives** allow us to find **relationships** between large volumes of **data**.

- These "relationships" are **derivatives**: consider  $dy/dx$ . How does  $y$  change if we modify  $x$ ? Currently, we only have **scalar derivatives**.
- This "data" is stored as **matrices**: blocks of data, that we can do linear operations (matrix multiplication) on.

Our goal is to work with many scalar derivatives at the **same time**.

In order to do that, we can apply some **derivative** rules, but we have to do it in a way that **agrees** with **matrix** math.

Our work is a careful balancing act between getting the **derivatives** we want, without violating the **rules** of matrices (and losing what makes them useful!)

**Example:** When we multiply two matrices, their inner shape has to match: in the below case, they need to share a dimension  $b$ .

$$\underbrace{(a \times b)}_X \underbrace{(b \times c)}_Y \quad (3)$$

We can't do anything that would **violate** this rule: otherwise, our **equations** don't make sense, and we get stuck. This means we need to build our math carefully.

First, we'll look at the **properties** of derivatives. Then figure out how to usefully apply them to **vectors**, and then **matrices**.

## X.1 Review: Partial Derivatives

One more comment, though - we may have many different variables floating around. This means we **have** to use the multivariable **partial derivative**.

### Definition 2

The **partial derivative**

$$\frac{\partial B}{\partial A}$$

Is used when there may be **multiple variables** in our functions.

The rule of the partial derivative is that we keep every **independent** variable other than A and B **fixed**.

**Example:** Consider  $f(x, y) = 2x^2y$ .

$$\frac{\partial f}{\partial x} = 2(2x)y \quad (4)$$

Here, we kept  $y$  *fixed* - we treat it as if it were an unchanging **constant**.

Using the partial derivative lets us keep our work tidy: if **many** variables were allowed to **change** at the same time, it could get very confusing.

If this is too complicated, we can change those variables *one at a time*. We get a partial derivative for each of them, holding the others **constant**.

Our **total** derivative is the result of all of those different variables, **added** together. This is how we get the **multi-variable chain rule**.

Imagine keeping track of  $k$  different variables  $x_i$  with  $k$  different changes  $\Delta x_i$  at the same time! That's a headache.

**Definition 3**

The **multi-variable chain rule** in 3-D  $\{(x, y, z)\}$  is given as

$$\frac{df}{ds} = \overbrace{\frac{\partial f}{\partial x} \frac{\partial x}{\partial s}}^{\text{only modify } x} + \overbrace{\frac{\partial f}{\partial y} \frac{\partial y}{\partial s}}^{\text{only modify } y} + \overbrace{\frac{\partial f}{\partial z} \frac{\partial z}{\partial s}}^{\text{only modify } z}$$

If we have  $k$  variables  $\{x_1, x_2, \dots, x_k\}$  we can generalize this as:

$$\frac{df}{ds} = \sum_{i=1}^k \overbrace{\frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial s}}^{x_i \text{ component}}$$

## X.2 Thinking about derivatives

The typical definition of derivatives

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (5)$$

Gives an *idea* of what sort of things we're looking for. It reminds us of one piece of information we need:

- Our derivative **depends** on the **current position**  $x$  we are taking the derivative at.

We need this because derivative are **local**: the relationship between our variables might change if we move to a different **position**.

But, the problem with vectors is that each component can act **separately**: if we have a vector, we can change in many different "directions".

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad (6)$$

**Example:** Suppose we want a derivative  $\partial B / \partial A$ :  $\Delta a_1, \Delta a_2$ , and  $\Delta a_3$  could each, separately, have an effect on  $\Delta b_1$  and/or  $\Delta b_2$ . That requires 6 different derivatives,  $\partial b_i / \partial a_j$ .

Every component of the input  $A$  can potentially modify **every** component of the output  $B$ .

3 dimensions of  $A$  times  
2 dimensions of  $B$ : 6  
combinations.

One solution we could try is to just collect all of these derivatives into a **vector** or **matrix**.

**Concept 4**

For the **derivative** between two objects (scalars, vectors, matrices) **A** and **B**

$$\frac{\partial \mathbf{B}}{\partial \mathbf{A}}$$

We need to get the **derivatives**

$$\frac{\partial \mathbf{b}_j}{\partial \mathbf{a}_i}$$

between every **pair** of elements  $\mathbf{a}_i, \mathbf{b}_j$ : each pair of elements could have a **relationship**.

The total number of elements (or "size") is...

$$\text{Size}\left(\frac{\partial \mathbf{B}}{\partial \mathbf{A}}\right) = \text{Size}(\mathbf{B}) * \text{Size}(\mathbf{A})$$

Collecting these values into a **matrix** will gives us all the information we need.

But, how do we gather them? What should the **shape** look like? Should we **transpose** our matrix or not?

### X.3 Derivatives: Approximation

To answer this, we need to ask ourselves *why* we care about these derivatives: their **structure** will be based on what we need them for.

- We care about the **direction of greatest decrease**, the gradient. For example, we might want to adjust weight vector  $\mathbf{w}$  to reduce  $\mathcal{L}$ .
- We also want other derivatives that have the **same** behavior, so we can combine them using the **chain rule**.

Let's focus on the first point: we want to **minimize**  $\mathcal{L}$ . Our focus is the **change** in  $\mathcal{L}$ ,  $\Delta \mathcal{L}$ .

We want to take steps that reduce our loss  $\mathcal{L}$ .

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} \approx \frac{\text{Change in } \mathcal{L}}{\text{Change in } \mathbf{w}} = \frac{\Delta \mathcal{L}}{\Delta \mathbf{w}} \quad (7)$$

Thus, we **solve** for  $\Delta \mathcal{L}$ :

All we do is multiply both sides by  $\Delta \mathbf{w}$ .

$$\Delta \mathcal{L} \approx \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \Delta \mathbf{w} \quad (8)$$

Since this derivation was gotten using scalars, we might need a **different** type of multiplication for our **vector** and **matrix** derivatives.

**Concept 5**

We can use derivatives to **approximate** the change in our output based on our input:

$$\Delta \mathcal{L} \approx \frac{\partial \mathcal{L}}{\partial w} \star \Delta w$$

Where the  $\star$  symbol represents some type of **multiplication**.

We can think of this as a **function** that takes in change in  $\Delta w$ , and returns an **approximation** of the loss.

We already understand **scalar** derivatives, so let's move on to the **gradient**.