# MATRIX DERIVATIVE

MIN XU

The purpose of this guide is to show a simpler view of Matrix Derivative. Traditionally, *matrix derivative* is presented as a notation for organizing partial derivatives; however, I believe it is far easier on the mind and on the hand to think of Matrix Derivatives as Frechet Derivatives.

## 1. WHAT IS DERIVATIVE?

In one dimension, derivative is a number:

**Definition 1.** *(1D Derivative)*
*Let $f : \mathbb{R} \to \mathbb{R}$ be a function, we say $f$ is differentiable at $x_0$ if*

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = m$$

*where $m$ is a number we call $f'(x_0)$.*

We can do a little algebra and get an alternative form:

**Definition 2.** *(1D Derivative 2)*
*Define $f'(x_0)$ to be a number $m$ such that*

$$f(x_0 + h) - f(x_0) = mh + R_m(h) \tag{1.1}$$

***and** $R_m(h)$ is a function satisfying $\lim_{h \to 0} R_m(h) = 0$.*

Here we define the remainder term $R_m(h)$ to be $f(x_0 + h) - f(x_0) - mh$.

We can think of equation **??** this way: we are approximating $f(x_0 + h)$ by a linear function $f(x_0) + mh$ with the error term $R_m(h) = f(x_0 + h) - f(x_0) - mh$ going to 0 as $h$ gets smaller.

Indeed, in higher dimensions, we cannot represent the derivative as a single number but rather as a linear function.

**Definition 3.** *(higherD Derivative)*
*Define $f : \mathbb{R}^n \to \mathbb{R}^m$ as a function, let $x_0 \in \mathbb{R}^n$. We say that $f$ is differentiable at $x_0$ with derivative $M$ if $M : \mathbb{R}^n \to \mathbb{R}^m$ is **linear** and*

$$f(x_0 + h) - f(x_0) = M(h) + R_M(h)$$

*with $h \in \mathbb{R}^n$ and $\lim_{||h|| \to 0} \frac{||R_M(h)||}{||h||} = 0$.*

The *key* point here is that $M(h)$ and $R_M(h)$ are both functions $\mathbb{R}^n \to \mathbb{R}^m$. Note that $h$ is now a vector so in our limit, we must use $||h|| \to 0$.

Consider an example:

**Example 4.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function $f(x) = x^\mathsf{T} Ax$ where $x \in \mathbb{R}^n$ and $A$ is a $n \times n$ matrix. Then the derivative of $f$ at $x_0$ is a function $M$ where $M(h) = x_0^\mathsf{T}(A + A^\mathsf{T})h$.*

*Proof.*

$$f(x_0 + h) = (x_0 + h)^\mathsf{T} A(x_0 + h) \tag{1.2}$$

$$= x_0^\mathsf{T} Ax_0 + x_0^\mathsf{T} Ah + h^\mathsf{T} Ax_0 + h^\mathsf{T} Ah \tag{1.3}$$

$$= f(x_0) + x_0^\mathsf{T} Ah + x_0^\mathsf{T} A^\mathsf{T} h + h^\mathsf{T} Ah \tag{1.4}$$

$$= f(x_0) + x_0^\mathsf{T}(A + A^\mathsf{T})h + h^\mathsf{T} Ah \tag{1.5}$$

$$= f(x_0) + M(h) + h^\mathsf{T} Ah \tag{1.6}$$

(1.7)

where we used $(h^\mathsf{T} Ax_0)^\mathsf{T} = x_0^\mathsf{T} A^\mathsf{T} h$. Now, we have to show that $\lim_{||h||\to 0} \frac{|h^\mathsf{T} Ah|}{||h||} = 0$ where we use just absolute value instead of norm because $h^\mathsf{T} Ah$ is a number.

Now, notice that $|h^\mathsf{T} Ah| \leq ||h||||A||_2||h||$ where $||A||_2$ is the spectral norm of $A$. Hence, $\lim_{||h||\to 0} \frac{|h^\mathsf{T} Ah|}{||h||} \leq \lim_{||h||\to 0} \frac{||h||||A||_2||h||}{||h||} \leq \lim_{||h||\to 0} ||A||_2||h|| = 0$ □

## 2. Properties of Matrix Derivative

**Theorem 5.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a multivariate vector-valued function. Suppose $f$ is Frechet-differentiable at $x_0$ and let $M$ be its derivative. Then the matrix representation of $M$ is*

$$M = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \cdots & & \frac{\partial f_2}{\partial x_n} \\ \cdots & & & \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

*where $f_1(x), f_2(x), ..., f_m(x)$ are defined such that $f(x) = (f_1(x), ..., f_m(x))$.*

In other word, the theorem states that the Frechet Derivative coincides with the Jacobian Derivative. Hence, we will refer to both as matrix derivative.

**Note:** To simplify notation, when we say that the derivative derivative of $f : \mathbb{R}^n \to \mathbb{R}^m$ at $x_0$ is a matrix $M$, we mean that derivative is a function $M : \mathbb{R}^n \to \mathbb{R}^m$ such that $M(\Delta) = M\Delta$

Next, we list the important properties of matrix derivative. These are analogous to the properties of scalar derivative.

**Theorem 6.** *(Properties)*

1. ***Addition*** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions. Let $A, B$ be the derivative at $x_0$ of $f, g$ respectively, then the derivative of $f + g$ at $x_0$ is $A + B$.*
2. ***Composition*** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^d$ be two differentiable functions. Let $A, B$ be the derivative of $f, g$ at $x_0 \in R^n$, $y_0 \in \mathbb{R}^m$ respectively and let $f(x_0) = y_0$. Then the derivative of $g \circ f$ at $x_0$ is $BA$.*

(3) *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ with derivatives $A, B$ at $x_0$.*
***Inner Product*** *Define $h : \mathbb{R}^n \to \mathbb{R}$ such that $h(x) = f(x)^\mathsf{T} g(x)$. Then the derivative of $h$ is $x_0$ is $f(x_0)^\mathsf{T} B + g(x_0)^\mathsf{T} A$*
***Outer Product*** *Define $h : \mathbb{R}^n \to \mathbb{R}^{m \times m}$ such that $h(x) = f(x)g(x)^\mathsf{T}$. Then the derivative of $h$ at $x_0$ is a function $\Delta \mapsto A\Delta g(x_0)^\mathsf{T} + B\Delta f(x_0)^\mathsf{T}$*

*Proof.* TODO:FILL □

## 3. Simple Examples

3.1. **Matrix Multiplican.** Let $f : \mathbb{R}^{q \times p} \to \mathbb{R}^{a \times b}$ be defined as $f(M) = AMB$ where matrix $A \in \mathbb{R}^{a \times q}$ and matrix $B \in \mathbb{R}^{a \times q}$

$$\begin{aligned} f(M + \Delta) &= A(M + \Delta)B \\ &= AMB + A\Delta B \end{aligned}$$

Hence, the derivative simply is $\Delta \mapsto A\Delta B$

3.2. **Frobenius Norm.** Let $f : \mathbb{R}^{q \times p} \to \mathbb{R}$ be defined as $f(B) = ||B||_F^2$.

$$\begin{aligned} f(B + \Delta) &= \langle B + \Delta, B + \Delta \rangle \\ &= \langle B, B \rangle + 2\langle B, \Delta \rangle + \langle \Delta, \Delta \rangle \\ &= f(B) + 2\langle B, \Delta \rangle + ||\Delta||_F^2 \end{aligned}$$

Thus, we see that derivative of $f$ at $B$ is a function $\Delta \mapsto 2\langle B, \Delta \rangle$.

3.3. **Matrix Mahalanobis Norm.** Let $f : \mathbb{R}^{q \times p} \to \mathbb{R}$ be defined as $f(B) = \langle B, \Omega B \rangle$ where $\Omega \in \mathbb{R}^{q \times p}$.

$$\begin{aligned} f(B + \Delta) &= \langle B + \Delta, \Omega(B + \Delta) \rangle \\ &= \langle B, \Omega B \rangle + 2\langle \Omega B, \Delta \rangle + \langle \Omega\Delta, \Delta \rangle \\ &= f(B) + 2\langle \Omega B, \Delta \rangle + \langle \Omega\Delta, \Delta \rangle \end{aligned}$$

Thus, the derivative of $f$ at $B$ is $\Delta \mapsto 2\langle \Omega B, \Delta \rangle$

3.4. **Duplication Operation.** We will now take derivative of $x^3$ with respect to $x$ in a way that is excessively complicated but illustrates the subtleties in the chain rule.

We break down $f(x) = x^3$ as $f = g \circ h$ where $h(x) = (x, x, x)$ and $g(x, y, z) = xyz$.

The derivative of $h$ is a function $\mathbb{R}^1 \to \mathbb{R}^3$ and is $\Delta \mapsto (\Delta, \Delta, \Delta)$, represented as a row vector $(1, 1, 1)$ in Jacobian form.

The derivative of $g$ is a function $\mathbb{R}^3 \to \mathbb{R}^1$ and is $(\Delta_1, \Delta_2, \Delta_3) \mapsto \Delta_1 yz + \Delta_2 xz + \Delta_3 xy$, represented as a column vector $(yz, xz, xy)$ in Jacobian form.

The derivative of composition is then just $\Delta \mapsto \Delta(yz + xz + xy)$. Since in our composition, $x = y = z$, we get that the derivative is $\Delta \mapsto 3x^2\Delta$.

If we use Chain rule and work with Jacobian form, we get $3x^2$ as our answer, consistent with the other approach.

## 4. Examples

4.1. **Matrix Regression.** Let $Y \in \mathbb{R}^{q\times n}$ and $X \in \mathbb{R}^{p\times n}$. Define function $f : \mathbb{R}^{q\times p} \to \mathbb{R}$

$$f(B) = ||Y - BX||_F^2$$

We know that the derivative of $B \mapsto Y - BX$ with respective to $B$ is $\Delta \mapsto -\Delta X$.

And that the derivative of $Y - BX \mapsto ||Y - BX||_F^2$ with respect to $Y - BX$ is $\Delta \mapsto 2\langle Y - BX, \Delta\rangle$.

Compose the two derivatives and we get the overall derivative is

$$\begin{aligned}
\Delta &\mapsto 2\langle Y - BX, -\Delta X\rangle \\
&= -2\text{tr}((\Delta X)^\mathsf{T}(Y - BX)) \\
&= -2\text{tr}(X^\mathsf{T}\Delta^\mathsf{T}(Y - BX)) \\
&= -2\text{tr}(\Delta^\mathsf{T}(Y - BX)X^\mathsf{T}) \\
&= -2\text{tr}(\Delta^\mathsf{T}YX^\mathsf{T} - \Delta^\mathsf{T}BXX^\mathsf{T}) \\
&= -2\text{tr}(\Delta^\mathsf{T}(YX^\mathsf{T} - BXX^\mathsf{T})) \\
&= 2\langle \Delta, -YX^\mathsf{T} + BXX^\mathsf{T}\rangle
\end{aligned}$$

4.2. **Matrix Regression 2.** Let $Y \in \mathbb{R}^{n\times q}$ and $X \in \mathbb{R}^{n\times p}$. Define function $f : \mathbb{R}^{p\times q} \to \mathbb{R}$

$$f(B) = ||Y - XB||_F^2$$

where $B \in \mathbb{R}^{p\times q}$. Note that in the case $q = 1$, this is exactly linear regression.

Since $(Y - XB)^\mathsf{T} = Y^\mathsf{T} - B^\mathsf{T}X^\mathsf{T}$, we can directly apply the previous example and get the derivative is $\Delta \mapsto 2\langle \Delta, -Y^\mathsf{T}X + B^\mathsf{T}X^\mathsf{T}X\rangle$