



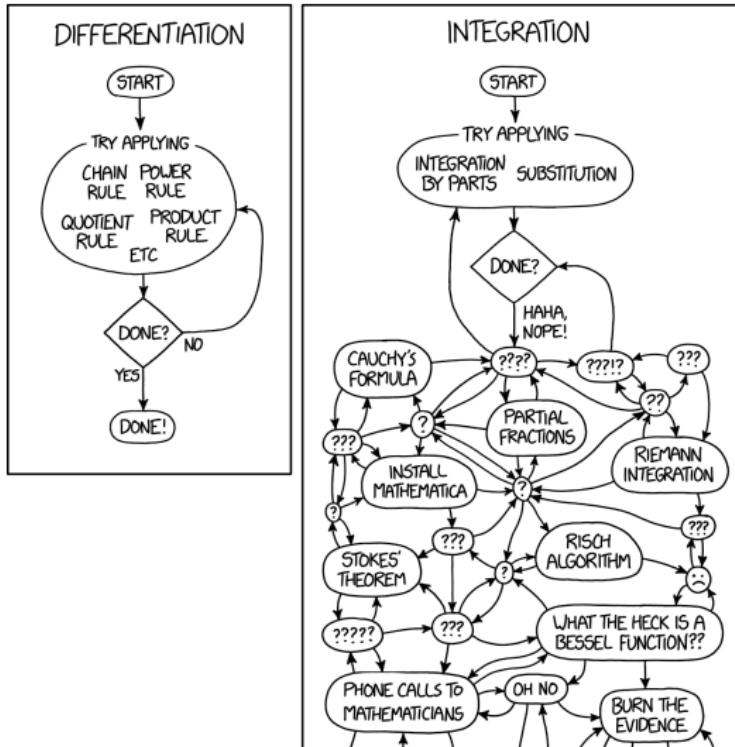
Matrix Differentiation

Wen Perng

NTUEE

2024 September 18

Derivatives are easy! (Source: XKCD)



What do we want?

In machine learning and many other studies, we often want to optimize functions:

$$\min_x f(x).$$

We know that **extremas** (max,min,saddle points) occur at places with zero derivatives:

$$\frac{df(x)}{dx} \Big|_{x^*} = 0,$$

whether it is a minima or not can be determined by a second order derivative test:

$$\frac{d^2f(x)}{dx^2} \Big|_{x^*} > 0.$$

We denote the point of minima as

$$x^* = \arg \min_x f(x).$$

What do we want?



What if we want to minimize a function of two variables?

$$\min_{x_1, x_2} f(x_1, x_2)$$

What do we want?

What if we want to minimize a function of two variables?

$$\min_{x_1, x_2} f(x_1, x_2)$$

We do the same test for finding extremas: by **partial derivatives** of first and second order:

$$\frac{\partial f}{\partial x_1} = 0, \quad \frac{\partial f}{\partial x_2} = 0, \quad \det \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right) > 0.$$

This requires the notions of **gradient** and **Hessian**, which will be discussed later.

What do we want?

But instead of writing the derivatives out term-by-term, is there a

1. coordinate free,
2. compact way

of writing out the derivatives? I.e., we would want to justify the notations of

$$\frac{df(A)}{dA} \text{ and } \frac{df(\mathbf{x})}{d\mathbf{x}},$$

where A is a matrix and \mathbf{x} is a vector.

What do the derivatives mean and how do they behave? Note that we will work in cases where matrices are treated as *matrices* instead of *arrays*.

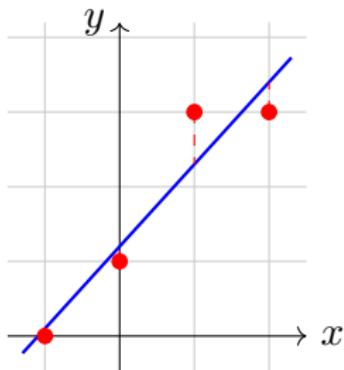
A glimpse of what is to come

So... what kind of functions will we encounter? A prime example is **regression** by minimizing the **squared error**:

$$\mathcal{E} = \|X\mathbf{w} - \mathbf{y}\|^2,$$

where $X \in \mathbb{R}^{m \times n}$, $\mathbf{w} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$. We would like to find a \mathbf{w} that *best estimates* \mathbf{y} by the approximation:

$$\hat{\mathbf{y}} = X\mathbf{w}.$$



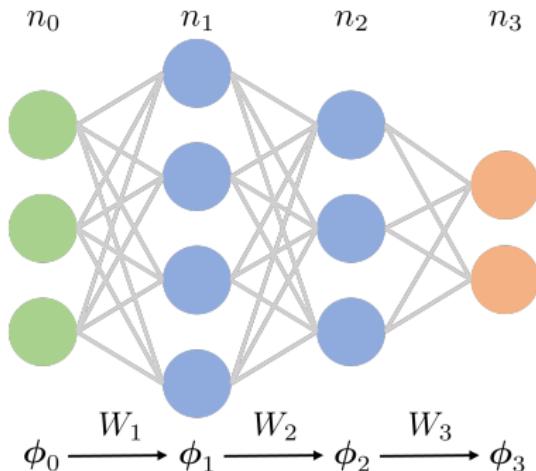
$$X = [1, \mathbf{x}]$$

$$\mathbf{w} = [a_0, a_1]^\top$$

$$\hat{y} = a_0 + a_1 x$$

A glimpse of what is to come

In machine learning, when we want to calculate the backpropagation of a multilayer perceptron (MLP), we also need to learn how to differentiate with respect to matrix variables.



$$f(W, \phi) := \sigma(W^T \phi) \quad (\text{perceptron})$$

$$\phi_3 = f(W_3, \cdot) \circ f(W_2, \cdot) \circ f(W_1, \phi_0)$$

Lastly, others might want to differentiate functions such as

$$f = \mathbf{x}^\top A \mathbf{x}, \det(A), \text{Tr}\{A\}, \lambda_i(A), \dots.$$

A glimpse of what is to come

Lastly, others might want to differentiate functions such as

$$f = \mathbf{x}^\top A \mathbf{x}, \det(A), \text{Tr}\{A\}, \lambda_i(A), \dots.$$

Or even crazier, the QR-decomposition:

$$\begin{aligned} X(t) &= Q(t)R(t) \\ \Rightarrow \frac{dQ(t)}{dt} &\stackrel{?}{=} \frac{\partial Q}{\partial X} \cdot \frac{dX}{dt}. \end{aligned}$$

This last part will be introduced in the next course. So stay tuned.

Before we go...

Def. (Transpose)

Flip the matrix along its diagonal:

$$\begin{bmatrix} \textcolor{red}{a_{11}} & a_{12} & a_{13} \\ a_{21} & \textcolor{red}{a_{22}} & a_{23} \end{bmatrix}^T = \begin{bmatrix} \textcolor{red}{a_{11}} & a_{21} \\ a_{12} & \textcolor{red}{a_{22}} \\ a_{13} & a_{23} \end{bmatrix}. \quad (1)$$

Before we go...

Def. (Transpose)

Flip the matrix along its diagonal:

$$\begin{bmatrix} \textcolor{red}{a_{11}} & a_{12} & a_{13} \\ a_{21} & \textcolor{red}{a_{22}} & a_{23} \end{bmatrix}^T = \begin{bmatrix} \textcolor{red}{a_{11}} & a_{21} \\ a_{12} & \textcolor{red}{a_{22}} \\ a_{13} & a_{23} \end{bmatrix}. \quad (1)$$

What number when transposed doubles?

Before we go...

Def. (Transpose)

Flip the matrix along its diagonal:

$$\begin{bmatrix} \textcolor{red}{a_{11}} & a_{12} & a_{13} \\ a_{21} & \textcolor{red}{a_{22}} & a_{23} \end{bmatrix}^T = \begin{bmatrix} \textcolor{red}{a_{11}} & a_{21} \\ a_{12} & \textcolor{red}{a_{22}} \\ a_{13} & a_{23} \end{bmatrix}. \quad (1)$$

What number when transposed doubles?

$$\text{午}^T = \text{什}$$

Before we go...

Def. (Transpose)

Flip the matrix along its diagonal:

$$\begin{bmatrix} \textcolor{red}{a_{11}} & a_{12} & a_{13} \\ a_{21} & \textcolor{red}{a_{22}} & a_{23} \end{bmatrix}^T = \begin{bmatrix} \textcolor{red}{a_{11}} & a_{21} \\ a_{12} & \textcolor{red}{a_{22}} \\ a_{13} & a_{23} \end{bmatrix}. \quad (1)$$

What number when transposed doubles?

$$\text{午}^T = \text{什}$$

Def. (Inner Product)

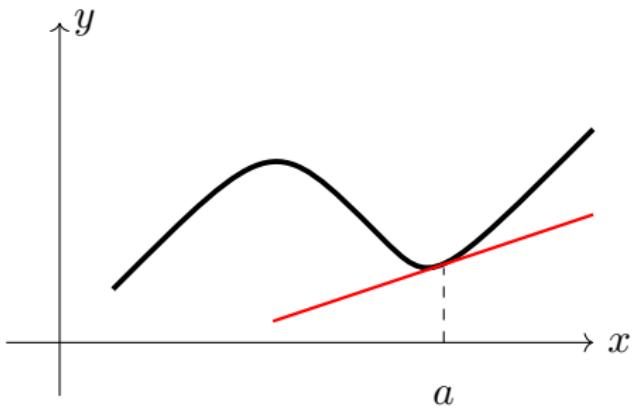
We denote the *Euclidean* inner product of two column vectors x and y as:

$$x \cdot y = x^T y = \langle x, y \rangle.$$

- 1 Differentiation Revisited**
- 2 Matrix Differentiation**
- 3 Second Derivative**
- 4 Examples of Matrix Derivatives**

- 1 Differentiation Revisited
 - What is Derivative ■ Derivative as Linear Operator
- 2 Matrix Differentiation
- 3 Second Derivative
- 4 Examples of Matrix Derivatives

Derivatives are linearizations. Consider the easiest case below.



We know that the first-order approximation will be

$$y(x) = \frac{dy}{dx} \Big|_a (x - a) + y(a).$$

What is a derivative?

We can rewrite the previous equation into the form below:

$$y(a + dx) = \frac{dy}{dx} \Big|_a dx + y(a)$$

$$\cancel{\frac{dy}{dx}} \quad dy = y(a + dx) - y(a) = \frac{dy}{dx} \Big|_a dx$$

Thus obtaining:

$$dy = \left(\frac{dy}{dx} \right) dx \quad (2)$$

We don't need to "divide" the dx over to the other side. The equation above will generalize to future cases.

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$dy = \frac{dy}{dx} dx.$$

The above definition holds for x and y being scalars, vectors or matrices so long as the dimensions are compatible.

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$dy = \frac{dy}{dx} dx.$$

The above definition holds for x and y being scalars, vectors or matrices so long as the dimensions are compatible.

e.g.

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x}, \quad A \in \mathbb{R}^{m \times n}, \quad \mathbf{x} \in \mathbb{R}^{n \times 1}$$

$$\begin{aligned} d\mathbf{f} &= f(\mathbf{x} + d\mathbf{x}) - f(\mathbf{x}) = A(\mathbf{x} + d\mathbf{x}) - A\mathbf{x} \\ &= A d\mathbf{x} \end{aligned}$$

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$dy = \frac{dy}{dx} dx.$$

The above definition holds for x and y being scalars, vectors or matrices so long as the dimensions are compatible.

e.g.

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x}, \quad A \in \mathbb{R}^{m \times n}, \quad \mathbf{x} \in \mathbb{R}^{n \times 1}$$

By plugging in the definition above:

$$d\mathbf{f} = \mathbf{f}(x + dx) - \mathbf{f}(x) =$$

Derivatives as Linear Approximation?

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$dy = \frac{dy}{dx} dx.$$

The above definition holds for x and y being scalars, vectors or matrices so long as the dimensions are compatible.

e.g.

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x}, \quad A \in \mathbb{R}^{m \times n}, \quad \mathbf{x} \in \mathbb{R}^{n \times 1}$$

By plugging in the definition above:

$$d\mathbf{f} = \mathbf{f}(x + dx) - \mathbf{f}(x) = A(x + dx) - Ax = A dx$$

$$\frac{d\mathbf{f}}{dx} = A$$

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$dy = \frac{dy}{dx} dx.$$

The above definition holds for x and y being scalars, vectors or matrices so long as the dimensions are compatible.

e.g.

$$\mathbf{f}(\mathbf{A}) = \mathbf{A}\mathbf{x}, \quad \mathbf{A} \in \mathbb{R}^{m \times n}, \quad \mathbf{x} \in \mathbb{R}^{n \times 1}$$

Derivatives as Linear Approximation?

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$dy = \frac{dy}{dx} dx.$$

The above definition holds for x and y being scalars, vectors or matrices so long as the dimensions are compatible.

e.g.

$$\mathbf{f}(A) = A\mathbf{x}, \quad A \in \mathbb{R}^{m \times n}, \quad \mathbf{x} \in \mathbb{R}^{n \times 1}$$

By plugging in the definition above:

$$d\mathbf{f} = \mathbf{f}(A + dA) - \mathbf{f}(A) = (A + dA)\mathbf{x} - A\mathbf{x} = dA \mathbf{x}$$

$$\frac{d\mathbf{f}}{dA} \stackrel{?}{=} \mathbf{x}$$

Perhaps we shouldn't stick to the *left-multiplications*, but simply view derivatives as a **linear operator**:

$$dy = \mathcal{L}\{dx\}. \quad (3)$$

Perhaps we shouldn't stick to the *left-multiplications*, but simply view derivatives as a **linear operator**:

$$dy = \mathcal{L}\{dx\}. \quad (3)$$

For $y = f(x)$, the differentiation linear operator can have many forms:

$$\mathcal{L}\{dx\} = f'dx$$

$$\mathcal{L}\{dx\} = dx f'$$

$$\mathcal{L}\{dx\} = \sqrt{f'}dx\sqrt{f'}.$$

All works as long as it is **linear**:

$$\mathcal{L}\{adx_1 + dx_2\} = a\mathcal{L}\{dx_1\} + \mathcal{L}\{dx_2\}.$$

Prop. (Derivatives as Linear Operators (Fréchet Derivative))

For $y = f(x)$,

$$\mathrm{d}y = \mathcal{L}\{\mathrm{d}x\}.$$

Where $\mathcal{L}\{\cdot\}$ is a linear operator, denoting the notion of a derivative.

The derivative is a **linear operator (function)** that maps the changes in the variable to the first-order changes in the function.¹

¹Since all linear operators can be represented by matrices, we can, still, represent derivatives as matrices and vectors alike. But it is often not needed and might further complicate the results.

- 1 Differentiation Revisited**
- 2 Matrix Differentiation**
 - How to Calculate ■ First Derivative
- 3 Second Derivative**
- 4 Examples of Matrix Derivatives**

Two methods are possible:

1. Term-wise differentiation,

How to Calculate Derivative?

Two methods are possible:

1. Term-wise differentiation,
2. **Directional derivatives** (if it exists): the directional derivative of $f(\mathbf{x})$ by \mathbf{x} at \mathbf{x}_0 in the direction of \mathbf{h} will be

$$\begin{aligned} df &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}(\mathbf{x}_0)[\mathbf{h}] := \left. \frac{d}{dt} f(\mathbf{x}_0 + t\mathbf{h}) \right|_{t=0} \\ &= \lim_{t \rightarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{h}) - f(\mathbf{x}_0)}{t}. \end{aligned} \tag{4}$$

The first method is quite boring, we will dive into the second method and some examples in the next page.

(1) Differentiate scalar function f by vector $\mathbf{x} \in \mathbb{R}^n$:

Let us suppose

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}, \text{ where } A \text{ is symmetric.}$$

[Sol.]

(1) Differentiate scalar function f by vector $x \in \mathbb{R}^n$:

Let us suppose

$$f(x) = x^T A x, \text{ where } A \text{ is symmetric.}$$

[Sol.]

$$\begin{aligned}
 \frac{\partial f}{\partial x}(x)[h] &= \frac{d}{dt} (\vec{x} + t\vec{h})^T A (\vec{x} + t\vec{h}) \Big|_{t=0} \\
 &= \vec{h}^T A (\vec{x} + t\vec{h}) + (\vec{x} + t\vec{h})^T A \vec{h} \Big|_{t=0} \\
 &= \boxed{\vec{h}^T A \vec{x}} + \vec{x}^T A \vec{h} \\
 &\quad \text{(} n \times n \text{) } \\
 &= (\vec{h}^T A \vec{x})^T + \vec{x}^T A \vec{h} \\
 &= \vec{x}^T A^T \vec{h} + \vec{x}^T A \vec{h} = \vec{x}^T (A + A^T) \vec{h} = 2\vec{x}^T A \vec{h}
 \end{aligned}$$

Differentiate by Vectors

(1) Differentiate scalar function f by vector $\mathbf{x} \in \mathbb{R}^n$:

Let us suppose

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}, \text{ where } A \text{ is symmetric.}$$

[Sol.]

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x})[\mathbf{h}] &= \left. \frac{d}{dt} ((\mathbf{x} + t\mathbf{h})^\top A (\mathbf{x} + t\mathbf{h})) \right|_{t=0} \\ &= \mathbf{h}^\top A \mathbf{x} + \mathbf{x}^\top A \mathbf{h} \\ &= 2\mathbf{x}^\top A \mathbf{h}.\end{aligned}$$

Differentiate by Vectors

(1) Differentiate scalar function f by vector $\mathbf{x} \in \mathbb{R}^n$:

Let us suppose

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}, \text{ where } A \text{ is symmetric.}$$

[Sol.]

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x})[\mathbf{h}] &= \left. \frac{d}{dt} ((\mathbf{x} + t\mathbf{h})^\top A (\mathbf{x} + t\mathbf{h})) \right|_{t=0} \\ &= \mathbf{h}^\top A \mathbf{x} + \mathbf{x}^\top A \mathbf{h} \\ &= 2\mathbf{x}^\top A \mathbf{h}.\end{aligned}$$

Since the final result is of the **matrix-vector multiplication form**, we can have the following statement:

$$\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}) \equiv \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \equiv \frac{\partial f}{\partial \mathbf{x}} = 2\mathbf{x}^\top A. \quad (5)$$

Further, given a function $f(\mathbf{x}) = f(x_1, \dots, x_n)$ with

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = [x_1 \quad \cdots \quad x_n]^T \in \mathbb{R}^{n \times 1},$$

we know from calculus that its **infinitesimal** change is:

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n. \quad (6)$$

This result is obtained by term-wise differentiation. Intimidating?
But this is just the dot product of two vectors!

Differentiate by Vectors

Notice that

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n$$

is just the dot product of ∇f and $d\mathbf{x}$:

$$df = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \begin{bmatrix} dx_1 \\ \vdots \\ dx_n \end{bmatrix} = (\nabla f)^T d\mathbf{x}.$$

nabla

Notice that

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n$$

is just the dot product of ∇f and $d\mathbf{x}$:

$$df = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \begin{bmatrix} dx_1 \\ \vdots \\ dx_n \end{bmatrix} = (\nabla f)^\top d\mathbf{x}.$$

Def. (Gradient)

Given an inner product $\langle \cdot, \cdot \rangle$, the **gradient** of a function $f(\mathbf{x})$ ($\mathbf{x} \in \mathbb{R}^n$) is defined as

$$\nabla f(\mathbf{x}) \in \mathbb{R}^n \text{ s.t. } df = \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x})[\mathbf{h}] = \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle. \quad (7)$$

Differentiate by Vectors

For $f = f(\mathbf{x})$, if we have

$$(\nabla f)^\top d\mathbf{x} = df = \frac{\partial f}{\partial \mathbf{x}} d\mathbf{x}$$

as operation by *left-multiplication*, then

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} = (\nabla f)^\top \equiv \nabla^\top f.$$

Differentiate by Vectors

For $f = f(\mathbf{x})$, if we have

$$(\nabla f)^\top d\mathbf{x} = df = \frac{\partial f}{\partial \mathbf{x}} d\mathbf{x}$$

as operation by *left-multiplication*, then

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} = (\nabla f)^\top \equiv \nabla^\top f.$$

And if we define

$$df = d\mathbf{x}^\top \frac{\partial f}{\partial \mathbf{x}^\top}$$

by *right-multiplication*, then

$$\frac{\partial f}{\partial \mathbf{x}^\top} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \nabla f.$$

Back to our example above:

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$$

[Sol.]

$$\begin{aligned} df &= \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x})[\mathbf{h}] = 2\mathbf{x}^\top A \mathbf{h} = \langle 2A\mathbf{x}, \mathbf{h} \rangle \\ &\quad \text{(Note: } \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}) = 2A \text{)} \\ \nabla f(\mathbf{x}) &= 2A\mathbf{x}. \end{aligned}$$

Notice that:

$$1 \times n \rightarrow \frac{\partial f}{\partial \vec{x}} \leftarrow 1 \times 1 \quad df = \frac{\partial f}{\partial \vec{x}} d\vec{x}$$

$$n \times 1 \rightarrow \frac{\partial f}{\partial \vec{x}^T} \leftarrow 1 \times 1 \quad df = d\vec{x}^T \frac{\partial f}{\partial \vec{x}^T}$$

Differentiate by Vectors

Notice that:

$$\begin{aligned} 1 \times n &\rightarrow \frac{\partial f}{\partial \mathbf{x}} \leftarrow \underset{n \times 1}{n \times 1} \\ n \times 1 &\rightarrow \frac{\partial f}{\partial \mathbf{x}^T} \leftarrow \underset{1 \times n}{1 \times n} \end{aligned}$$

We can even have:

$$\vec{df} = \frac{\partial \vec{f}}{\partial \vec{x}} d\vec{x} \quad m \times n \rightarrow \frac{\partial \vec{f}}{\partial \vec{x}} \leftarrow \underset{m \times 1}{m \times 1}$$

Notice that:

$$\begin{aligned} 1 \times n &\rightarrow \frac{\partial f}{\partial \mathbf{x}} \leftarrow \underset{n \times 1}{\textcolor{red}{n \times 1}} \\ n \times 1 &\rightarrow \frac{\partial f}{\partial \mathbf{x}^T} \leftarrow \underset{1 \times n}{\textcolor{red}{1 \times n}} \end{aligned}$$

We can even have:

$$m \times n \rightarrow \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \leftarrow \underset{n \times 1}{\textcolor{red}{m \times 1}}$$

But remember, this is just a convention. And the main takeaway is that **derivatives are linear operators**.

(2) Differentiate vector function \mathbf{f} by another vector \mathbf{x} :

If \mathbf{x} is a column vector and \mathbf{f} is of the form:

$$\mathbf{f} = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix},$$

(2) Differentiate vector function \mathbf{f} by another vector \mathbf{x} :

If \mathbf{x} is a column vector and \mathbf{f} is of the form:

$$\mathbf{f} = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix},$$

then we have:

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial f_m}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n},$$

$$d\vec{f} = \frac{\partial \vec{f}}{\partial \vec{x}} d\vec{x}$$

Some terms and notations to introduce: for a scalar function f and position vector $\mathbf{x} = [x_1, \dots, x_n]^T$,

Some terms and notations to introduce: for a scalar function f and position vector $\mathbf{x} = [x_1, \dots, x_n]^\top$,

1. Gradient (transpose):

$$\nabla f := \frac{\partial f}{\partial \mathbf{x}^\top} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \quad \nabla^\top f := \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (8)$$

Some terms and notations to introduce: for a scalar function f and position vector $\mathbf{x} = [x_1, \dots, x_n]^\top$,

1. Gradient (transpose):

$$\nabla f := \frac{\partial f}{\partial \mathbf{x}^\top} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \quad \nabla^\top f := \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (8)$$

e.g. $f(\mathbf{x}) = f(x_1, x_2) = x_1^2 x_2$

$$\nabla f(\vec{x}) = \begin{bmatrix} 2x_1 x_2 \\ x_1^2 \cdot 1 \end{bmatrix}$$

Some terms to introduce: for a scalar function f and position vector $\mathbf{x} = [x_1, \dots, x_n]^\top$,

2. Hessian:

$$\mathsf{H}_f := \nabla \nabla^\top f = \frac{\partial^2 f}{\partial \mathbf{x}^\top \partial \mathbf{x}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \boxed{\frac{\partial^2 f}{\partial x_1 \partial x_2}} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \quad (9)$$

Some terms to introduce: for a scalar function f and position vector $\mathbf{x} = [x_1, \dots, x_n]^\top$,

2. Hessian:

$$\mathsf{H}_f := \nabla \nabla^\top f = \frac{\partial^2 f}{\partial \mathbf{x}^\top \partial \mathbf{x}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \quad (9)$$

e.g. $f(\mathbf{x}) = f(x_1, x_2) = x_1^2 x_2$

$$\nabla f(\vec{x}) = \begin{bmatrix} 2x_1 x_2 \\ x_1^2 \end{bmatrix}, \quad \nabla^\top \nabla f = \begin{bmatrix} 2x_2 & 2x_1 \\ 2x_1 & 0 \end{bmatrix}$$

For a vector function $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), \dots, y_n(\mathbf{x})]^\top$ and position vector $\mathbf{x} = [x_1, \dots, x_n]^\top$,

3. Jacobian:

$$J := \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial y_n}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}, \quad (10)$$

Differentiate by Vectors

For a vector function $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), \dots, y_n(\mathbf{x})]^\top$ and position vector $\mathbf{x} = [x_1, \dots, x_n]^\top$,

3. Jacobian:

$$J := \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial y_n}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}, \quad (10)$$

The Jacobian is useful in coordinate transformations, like the change of variables in integration.

$$\int \cdots \int_{\mathbf{y}(\mathcal{X})} f(\mathbf{y}) dy_1 \cdots dy_n = \int \cdots \int_{\mathcal{X}} f(\mathbf{y}(\mathbf{x})) |J| dx_1 \cdots dx_n.$$

$$\int_{\mathbf{y}} f(\vec{y}) d\vec{y} = \int_{\mathcal{X}} f(\vec{y}(\vec{x})) \left| \frac{\partial \vec{y}}{\partial \vec{x}} \right| d\vec{x}$$

(3) Differentiate function $f(X)$ by a matrix $X \in \mathbb{R}^{m \times n}$:

$$df = \frac{\partial f}{\partial X}(X_0)[H] = \left. \frac{d}{dt} f(X_0 + tH) \right|_{t=0} \quad (H \in \mathbb{R}^{m \times n}).$$

(3) Differentiate function $f(X)$ by a matrix $X \in \mathbb{R}^{m \times n}$:

$$df = \frac{\partial f}{\partial X}(X_0)[H] = \left. \frac{d}{dt} f(X_0 + tH) \right|_{t=0} \quad (H \in \mathbb{R}^{m \times n}).$$

Some examples are as follows:

1. $f(A) = Ax$

(3) Differentiate function $f(X)$ by a matrix $X \in \mathbb{R}^{m \times n}$:

$$df = \frac{\partial f}{\partial X}(X_0)[H] = \left. \frac{d}{dt} f(X_0 + tH) \right|_{t=0} \quad (H \in \mathbb{R}^{m \times n}).$$

Some examples are as follows:

1. $f(A) = Ax$

$$df = (dA)x$$

(3) Differentiate function $f(X)$ by a matrix $X \in \mathbb{R}^{m \times n}$:

$$df = \frac{\partial f}{\partial X}(X_0)[H] = \left. \frac{d}{dt} f(X_0 + tH) \right|_{t=0} \quad (H \in \mathbb{R}^{m \times n}).$$

Some examples are as follows:

1. $f(A) = Ax$

$$df = (dA)x$$

2. $f(A) = A^T A$

$$df = (A+dA)^T (A+dA) - A^T A = \underline{A^T dA} + \underline{dA^T A} + \cancel{dA^T dA}$$

(3) Differentiate function $f(X)$ by a matrix $X \in \mathbb{R}^{m \times n}$:

$$df = \frac{\partial f}{\partial X}(X_0)[H] = \left. \frac{d}{dt} f(X_0 + tH) \right|_{t=0} \quad (H \in \mathbb{R}^{m \times n}).$$

Some examples are as follows:

1. $f(A) = Ax$

$$df = (dA)x$$

2. $f(A) = A^T A$

$$df = (A + dA)^T (A + dA) - A^T A = (dA)^T A + A^T dA$$

$$\stackrel{?}{=} \left. \frac{d}{dt} (A + t dA)^T (A + t dA) \right|_{t=0}$$

(3) Differentiate function $f(X)$ by a matrix $X \in \mathbb{R}^{m \times n}$:

$$df = \frac{\partial f}{\partial X}(X_0)[H] = \left. \frac{d}{dt} f(X_0 + tH) \right|_{t=0} \quad (H \in \mathbb{R}^{m \times n}).$$

Some examples are as follows:

1. $f(A) = Ax$

$$df = (dA)x$$

2. $f(A) = A^T A$

$$\begin{aligned} df &= (A + dA)^T (A + dA) - A^T A = (dA)^T A + A^T dA \\ &= \left. \frac{d}{dt} (A + t dA)^T (A + t dA) \right|_{t=0} \end{aligned}$$

3. $f(X) = \text{Tr}\{X^T A X B\}$, with symmetric A and B .

Def. (Trace)

The trace of a square matrix is the sum of its diagonals:

$$\text{Tr}\left\{\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\right\} = 5 \quad \text{Tr}\{A\} = \sum_{i=1}^n a_{ii}, \quad (11)$$

satisfying:

$$\begin{aligned} 3.1 \quad \text{Tr}\{A^T\} &= \text{Tr}\{A\}, \\ 3.2 \quad \text{Tr}\{ABC\} &= \text{Tr}\{CAB\} \quad (\text{cyclic permutation}). \end{aligned}$$

$$\frac{d}{dt} \text{Tr}\{A(t)\} = \lim_{dt \rightarrow 0} \frac{\text{Tr}\{A(t+dt)\} - \text{Tr}\{A(t)\}}{dt}$$

$$= \text{Tr}\left\{\lim_{dt \rightarrow 0} \frac{A(t+dt) - A(t)}{dt}\right\}$$

[Sol.]

$$\begin{aligned} \frac{\partial f}{\partial X}[H] &= \text{Tr}\left\{\frac{d}{dt}(x+tH)^T A(x+tH)B\Big|_0\right\} = \text{Tr}\{H^T A X B + \underline{x^T A H B}\} \\ &= \text{Tr}\{H^T A X B + H^T A X B\} = \text{Tr}\{H^T(2AXB)\} \end{aligned}$$

3. $f(X) = \text{Tr} \{X^T A X B\}$, with symmetric A and B .

Def. (Trace)

The trace of a square matrix is the sum of its diagonals:

$$\text{Tr} \{A\} = \sum_{i=1}^n a_{ii}, \quad (11)$$

satisfying:

3.1 $\text{Tr} \{A^T\} = \text{Tr} \{A\}$,

3.2 $\text{Tr} \{ABC\} = \text{Tr} \{CAB\}$ (cyclic permutation).

[Sol.]

$$\begin{aligned} \frac{\partial f}{\partial X}[H] &= \frac{d}{dt} \text{Tr} \{(X + tH)^T A (X + tH) B\} \Big|_{t=0} \\ &= 2 \text{Tr} \{H^T A X B\} \end{aligned}$$

Differentiate by Matrices

For $X \in \mathbb{R}^{m \times n}$, $f(X) = f(x_{11}, \dots, x_{mn})$, we have its **infinitesimal** change as:

$$df = \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f}{\partial x_{ij}} dx_{ij}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

vectorization

For $X \in \mathbb{R}^{m \times n}$, $f(X) = f(x_{11}, \dots, x_{mn})$, we have its infinitesimal change as:

$$\begin{aligned} df &= \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f}{\partial x_{ij}} dx_{ij} \\ &= \text{Tr} \left\{ \left[\begin{array}{cccc} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{m1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \frac{\partial f}{\partial x_{2n}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{array} \right] \left[\begin{array}{c|c|c} dx_{11} & \cdots & dx_{1n} \\ \hline dx_{21} & \cdots & dx_{2n} \\ \vdots & \ddots & \vdots \\ dx_{m1} & \cdots & dx_{mn} \end{array} \right] \right\} \end{aligned}$$

$$\begin{aligned} &= \frac{\partial f}{\partial x_{11}} dx_{11} + \frac{\partial f}{\partial x_{21}} dx_{21} + \cdots + \frac{\partial f}{\partial x_{m1}} dx_{m1} + \cdots \\ &\quad + \frac{\partial f}{\partial x_{1n}} dx_{1n} + \frac{\partial f}{\partial x_{2n}} dx_{2n} + \cdots + \frac{\partial f}{\partial x_{mn}} dx_{mn} \end{aligned}$$

Differentiate by Matrices

For $X \in \mathbb{R}^{m \times n}$, $f(X) = f(x_{11}, \dots, x_{mn})$, we have its **infinitesimal change** as:

$$\begin{aligned} df &= \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f}{\partial x_{ij}} dx_{ij} \\ &= \text{Tr} \left\{ \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{21}} & \dots & \frac{\partial f}{\partial x_{m1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \frac{\partial f}{\partial x_{2n}} & \dots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \begin{bmatrix} dx_{11} & \dots & dx_{1n} \\ dx_{21} & \dots & dx_{2n} \\ \vdots & \ddots & \vdots \\ dx_{m1} & \dots & dx_{mn} \end{bmatrix} \right\} \\ &= \frac{\partial f}{\partial X} [dX]. \end{aligned} \tag{12}$$

Hence, it is often denoted that for $X \in \mathbb{R}^{m \times n}$, $\frac{\partial f}{\partial X} \in \mathbb{R}^{n \times m}$, acting on the differential change by the **Frobenius inner product**.

Def. (Inner Product)

An inner product $\langle \cdot, \cdot \rangle$ on a vector space V over the field \mathbb{R} is a map $V \times V \rightarrow \mathbb{R}$ that satisfies: for $x, y, z \in V$ and $\lambda, \mu \in \mathbb{R}$

1. Symmetry: $\langle x, y \rangle = \langle y, x \rangle$,
2. Linearity: $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$,
3. Positive-definiteness: $\langle x, x \rangle > 0$ for all $x \neq 0$

Def. (Frobenius Inner Product)

For real matrices in $\mathbb{R}^{m \times n}$, we have the following inner product on them:

$$\langle X, Y \rangle = \text{Tr} \{ X^T Y \}. \quad (13)$$

Differentiate by Matrices

$$df = \text{Tr} \left\{ \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{m1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \frac{\partial f}{\partial x_{2n}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \begin{bmatrix} dx_{11} & \cdots & dx_{1n} \\ dx_{21} & \cdots & dx_{2n} \\ \vdots & \ddots & \vdots \\ dx_{m1} & \cdots & dx_{mn} \end{bmatrix} \right\}$$

Differentiate by Matrices

$$\begin{aligned}
 df &= \text{Tr} \left\{ \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{m1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \frac{\partial f}{\partial x_{2n}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \begin{bmatrix} dx_{11} & \cdots & dx_{1n} \\ dx_{21} & \cdots & dx_{2n} \\ \vdots & \ddots & \vdots \\ dx_{m1} & \cdots & dx_{mn} \end{bmatrix} \right\} \\
 &= \text{Tr} \{ (\nabla f)^T dX \} = \langle \nabla f, dX \rangle \\
 \rightarrow \nabla f(X) &= \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \tag{14}
 \end{aligned}$$

Differentiate by Matrices

$$\begin{aligned}
 df &= \text{Tr} \left\{ \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{m1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \frac{\partial f}{\partial x_{2n}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \begin{bmatrix} dx_{11} & \cdots & dx_{1n} \\ dx_{21} & \cdots & dx_{2n} \\ \vdots & \ddots & \vdots \\ dx_{m1} & \cdots & dx_{mn} \end{bmatrix} \right\} \\
 &= \text{Tr} \{ (\nabla f)^T dX \} = \langle \nabla f, dX \rangle \\
 \rightarrow \nabla f(X) &= \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}
 \end{aligned} \tag{14}$$

Let us look back at the example: for $f(X) = \text{Tr} \{ X^T A X B \}$,

$$\begin{aligned}
 \frac{\partial f}{\partial X}[H] &= \langle H, 2AXB \rangle \\
 \frac{\partial f}{\partial X}[H] &= 2\text{Tr} \{ H^T A X B \}, \quad \nabla f(X) = 2AXB.
 \end{aligned}$$

Remember that the following properties of derivatives still applies:

1. Chain Rule: for $f = f(y)$ and $y = g(x)$

$$\begin{aligned}\frac{\partial}{\partial x} (\underbrace{f \circ g}_{f(g(x))})[h] &= \frac{\partial f}{\partial y}(g(x)) \left[\frac{\partial g}{\partial x}[h] \right] \equiv \frac{\partial f}{\partial y} \Big|_{y=g(x)} \circ \frac{\partial g}{\partial x}[h] \\ &\equiv \frac{\partial f}{\partial y} \circ \frac{\partial g}{\partial x}[h]\end{aligned}\tag{15}$$

2. Product Rule:

$$\frac{\partial}{\partial x} (f(x) \cdot g(x))[h] = \frac{\partial f}{\partial x}[h] \cdot g(x) + f(x) \cdot \frac{\partial g}{\partial x}[h]\tag{16}$$

- 1 Differentiation Revisited
- 2 Matrix Differentiation
- 3 Second Derivative
 - Hessian ■ LLMSE Problem
- 4 Examples of Matrix Derivatives

For a given function $f : \mathbb{R} \rightarrow \mathbb{R}$, we can find its second-order approximation as

$$f(x + h) = f(x) + \frac{\partial f}{\partial x}(x) \cdot h + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} h^2 + \dots$$

For a given function $f : \mathbb{R} \rightarrow \mathbb{R}$, we can find its second-order approximation as

$$f(x + h) = f(x) + \frac{\partial f}{\partial x}(x) \cdot h + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} h^2 + \dots$$

But if we have $f : \mathbb{R}^n \rightarrow \mathbb{R}$, then

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + (\nabla f(\mathbf{x}))^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \mathbf{H}_f(\mathbf{x}) \mathbf{h} + \dots$$

For a given function $f : \mathbb{R} \rightarrow \mathbb{R}$, we can find its second-order approximation as

$$f(x + h) = f(x) + \frac{\partial f}{\partial x}(x) \cdot h + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} h^2 + \dots$$

But if we have $f : \mathbb{R}^n \rightarrow \mathbb{R}$, then

$$\begin{aligned} f(\mathbf{x} + \mathbf{h}) &= f(\mathbf{x}) + (\nabla f(\mathbf{x}))^\top \mathbf{h} + \boxed{\frac{1}{2} \mathbf{h}^\top \mathbf{H}_f(\mathbf{x}) \mathbf{h}} + \dots \\ &= f(\mathbf{x}) + \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x})[\mathbf{h}] + \frac{1}{2} \frac{\partial^2 f}{\partial \mathbf{x}^2}(\mathbf{x})[\mathbf{h}] + \dots, \end{aligned} \quad (17)$$

quadratic

where the matrix $\mathbf{H}_f(\mathbf{x})$ is called the **Hessian matrix** of f at \mathbf{x} .

We have seen that the Hessian matrix $\mathbf{H}_f = \boxed{\frac{\partial^2 f}{\partial \mathbf{x}^\top \partial \mathbf{x}}}$ determines the behavior of the extrema.

$$\vec{h}^\top \left(\frac{\partial^2 f}{\partial \vec{x}^\top \partial \vec{x}} \right) \vec{h}$$

Def. (Definiteness)

A symmetric matrix A is called **positive definite** if for any non-zero vector x it satisfies:

正定

$$x^T A x > 0 \iff A \succ 0.$$

Def. (Definiteness)

A symmetric matrix A is called **positive definite** if for any non-zero vector \mathbf{x} it satisfies:

$$\mathbf{x}^T A \mathbf{x} > 0 \iff A \succ 0.$$

Moreover, we have:

(positive semi-definite) $\mathbf{x}^T A \mathbf{x} \geq 0 \iff A \succeq 0$

(negative definite) $\mathbf{x}^T A \mathbf{x} < 0 \iff A \prec 0$

(negative semi-definite) $\mathbf{x}^T A \mathbf{x} \leq 0 \iff A \preceq 0$

If none of the above are satisfied, then the matrix is **indefinite**.

Def. (Definiteness)

A symmetric matrix A is called **positive definite** if for any non-zero vector \mathbf{x} it satisfies:

$$\mathbf{x}^T A \mathbf{x} > 0 \iff A \succ 0.$$

Moreover, we have:

- | | |
|--------------------------|---|
| (positive semi-definite) | $\mathbf{x}^T A \mathbf{x} \geq 0 \iff A \succeq 0$ |
| (negative definite) | $\mathbf{x}^T A \mathbf{x} < 0 \iff A \prec 0$ |
| (negative semi-definite) | $\mathbf{x}^T A \mathbf{x} \leq 0 \iff A \preceq 0$ |

If none of the above are satisfied, then the matrix is **indefinite**.

Positive definite \Rightarrow minima; negative definite \Rightarrow maxima.

And we can make the series expansion even more general to have
 $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$, then

$$f(X + H) = f(X) + \frac{\partial f}{\partial X}(X)[H] + \frac{1}{2} \frac{\partial^2 f}{\partial X^2}(X)[H] + \dots$$

Taylor Series Expansion

And we can make the series expansion even more general to have
 $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$, then

$$\begin{aligned} f(X + H) &= f(X) + \frac{\partial f}{\partial X}(X)[H] + \frac{1}{2} \frac{\partial^2 f}{\partial X^2}(X)[H] + \dots \\ &= f(X) + \langle \nabla f(X), H \rangle + \frac{1}{2} \langle \underset{\text{Frobenius}}{H_f(X)[H]}, H \rangle + \dots . \end{aligned} \tag{18}$$

The linear function $H_f(X)$ that maps $H \in \mathbb{R}^{m \times n}$ to $\mathbb{R}^{m \times n}$ is called the **Hessian operator**.

$$h^T H_f h = \langle H_f h, h \rangle$$

Like the gradient, the definition of the Hessian also relies on the inner product.

Def. (Gradient)

Given an inner product $\langle \cdot, \cdot \rangle$, the **gradient** of a function $f(x)$ ($x \in \mathbb{R}$, \mathbb{R}^n or $\mathbb{R}^{m \times n}$) is defined as

$$\nabla f(x) \text{ s.t. } \frac{\partial f}{\partial x}(x)[h] = \langle \nabla f(x), h \rangle. \quad (19)$$

Def. (Hessian)

Given an inner product $\langle \cdot, \cdot \rangle$, the **Hessian** of a function $f(x)$ ($x \in \mathbb{R}$, \mathbb{R}^n , or $\mathbb{R}^{m \times n}$) is defined as

$$\mathbf{H}_f(x) \text{ s.t. } \frac{\partial^2 f}{\partial x^2}(x)[h] = \langle \mathbf{H}_f(x)[h], h \rangle. \quad (20)$$

So how to calculate second derivative? Let x be a scalar in \mathbb{R} or a vector in \mathbb{R}^n or a matrix in $\mathbb{R}^{m \times n}$. Given a function $f(x)$, if we set $x = \gamma(t) = x_0 + th$, we can Taylor expand it about $t = 0$ into:

So how to calculate second derivative? Let x be a scalar in \mathbb{R} or a vector in \mathbb{R}^n or a matrix in $\mathbb{R}^{m \times n}$. Given a function $f(x)$, if we set $x = \gamma(t) = x_0 + th$, we can Taylor expand it about $t = 0$ into:

$$f \circ \gamma(t) = f \circ \gamma(0) + \frac{df \circ \gamma}{dt}(0)t + \frac{1}{2} \frac{d^2 f \circ \gamma}{dx^2} t^2 + \dots$$

So how to calculate second derivative? Let x be a scalar in \mathbb{R} or a vector in \mathbb{R}^n or a matrix in $\mathbb{R}^{m \times n}$. Given a function $f(x)$, if we set $x = \gamma(t) = x_0 + th$, we can Taylor expand it about $t = 0$ into:

$$f \circ \gamma(t) = f \circ \gamma(0) + \boxed{\frac{df \circ \gamma}{dt}(0)t + \frac{1}{2} \frac{d^2 f \circ \gamma}{dx^2} t^2 + \dots} \quad (21)$$

$$= f(x_0) + \frac{\partial f}{\partial x}(x_0)[h]t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2}(x_0)[h]t^2 + \dots$$



$$\frac{d}{dt} f(x_0 + th) \Big|_{t=0}$$

So how to calculate second derivative? Let x be a scalar in \mathbb{R} or a vector in \mathbb{R}^n or a matrix in $\mathbb{R}^{m \times n}$. Given a function $f(x)$, if we set $x = \gamma(t) = x_0 + th$, we can Taylor expand it about $t = 0$ into:

$$\begin{aligned} f \circ \gamma(t) &= f \circ \gamma(0) + \frac{df \circ \gamma}{dt}(0)t + \frac{1}{2} \frac{d^2 f \circ \gamma}{dx^2} t^2 + \dots \\ &= f(x_0) + \frac{\partial f}{\partial x}(x_0)[h]t + \cancel{\frac{1}{2}} \frac{\partial^2 f}{\partial x^2}(x_0)[h]t^2 + \dots \end{aligned} \quad (21)$$

Thus, we have the first and second derivatives as:

$$\frac{\partial f}{\partial x}(x)[h] = \left. \frac{d}{dt} f(x + th) \right|_{t=0}, \quad (22)$$

$$\frac{\partial^2 f}{\partial x^2}(x)[h] = \left. \frac{d^2}{dt^2} f(x + th) \right|_{t=0}. \quad (23)$$

Gaining idea from Taylor series, one can first calculate

$$\begin{array}{c} \vec{h}^\top H_f \vec{h} \\ \parallel \\ \langle H_f(\vec{h}), \vec{h} \rangle \end{array} \quad \frac{\partial^2 f}{\partial x^2}[h] = \left. \frac{d^2}{dt^2} f(x + th) \right|_{t=0}. \quad (24)$$

$$\begin{array}{c} \vec{h}^\top H_f \vec{k} \\ \parallel \\ \langle H_f(\vec{k}), \vec{k} \rangle \\ \parallel \\ \langle H_f(\vec{h}), \vec{k} \rangle \end{array}$$

Gaining idea from Taylor series, one can first calculate

$$\frac{\partial^2 f}{\partial x^2}[h] = \left. \frac{d^2}{dt^2} f(x + th) \right|_{t=0}. \quad (24)$$

By utilizing the **polarization identity**, we can define the *bilinear form*:

$$\langle H_f[h], k \rangle = \frac{\partial^2 f}{\partial x^2}[h, k] := \frac{1}{4} \left(\frac{\partial^2 f}{\partial x^2}[h+k] - \frac{\partial^2 f}{\partial x^2}[h-k] \right), \quad (25)$$

where $\frac{\partial^2 f}{\partial x^2}[h, h] \equiv \frac{\partial^2 f}{\partial x^2}[h]$.

ex. $Q(\vec{x}) = \vec{x}^T A \vec{x}$

$$Q(\vec{x} + \vec{y}) = \vec{x}^T A \vec{x} + \vec{y}^T A \vec{x} + \vec{x}^T A \vec{y} + \vec{y}^T A \vec{y}$$

$$\rightarrow Q(\vec{x} - \vec{y}) = \vec{x}^T A \vec{x} - \vec{y}^T A \vec{x} - \vec{x}^T A \vec{y} + \vec{y}^T A \vec{y}$$

$$2(\vec{y}^T A \vec{x} + \vec{x}^T A \vec{y}) = 4(\vec{x}^T A \vec{y})$$

Gaining idea from Taylor series, one can first calculate

$$\frac{\partial^2 f}{\partial x^2}[h] = \left. \frac{d^2}{dt^2} f(x + th) \right|_{t=0}. \quad (24)$$

By utilizing the **polarization identity**, we can define the *bilinear form*:

$$\frac{\partial^2 f}{\partial x^2}[h, k] := \frac{1}{4} \left(\frac{\partial^2 f}{\partial x^2}[h+k] - \frac{\partial^2 f}{\partial x^2}[h-k] \right), \quad (25)$$

where $\frac{\partial^2 f}{\partial x^2}[h, h] \equiv \frac{\partial^2 f}{\partial x^2}[h]$.

The second derivative can also be derived via directional derivatives:

$$\frac{\partial^2 f}{\partial x^2}[h, k] = \left. \frac{d}{dt} \frac{d}{ds} f(x + th + sk) \right|_{t=0, s=0}. \quad (26)$$

An example is as follows: $f(X) = \text{Tr} \{ X^T A X B \}$, with A, B symmetric.

[Sol.]

$$\frac{\partial f}{\partial X}[H] = 2\text{Tr} \{ H^T A X B \}.$$

$$\frac{\partial^2 f}{\partial X^2}[H, K] = \left. \frac{d}{dt} \frac{d}{ds} f(X + tH + sK) \right|_{t=0, s=0}$$

$$\begin{aligned}
 \frac{\partial^2 f}{\partial X^2}[K, H] &= \frac{d}{dt} \left. \frac{\partial f(X+tH)}{\partial X} \right|_{t=0} [K] \\
 &= \frac{d}{dt} \left. 2\text{Tr} \{ K^T A (X+tH) B \} \right|_{t=0} \\
 &= 2\text{Tr} \{ K^T A H B \} \\
 &= 2\text{Tr} \{ H^T A K B \}
 \end{aligned}$$

An example is as follows: $f(X) = \text{Tr} \{ X^T A X B \}$, with A, B symmetric.

[Sol.]

$$\frac{\partial f}{\partial X}[H] = 2\text{Tr} \{ H^T A X B \}.$$

$$\begin{aligned}\frac{\partial^2 f}{\partial X^2}[H, K] &= \left. \frac{d}{dt} \frac{d}{ds} f(X + tH + sK) \right|_{t=0, s=0} \\ &= \left. \frac{d}{ds} \left(\frac{\partial f}{\partial X}(X + sK)[H] \right) \right|_{s=0} \\ &= \left. \frac{d}{ds} 2\text{Tr} \{ H^T A (X + sK) B \} \right|_{s=0} \\ &= \text{Tr} \{ H^T A K B \} = \text{Tr} \{ K^T \underset{“}{A} H B \} \\ \Rightarrow \mathsf{H}_f(X)[H] &= AHB.\end{aligned}$$

$\langle K, AHB \rangle$

Let us also try using the polarization identity.

[Sol.]

$$\frac{\partial^2 f}{\partial X^2}[H] =$$

Let us also try using the polarization identity.

[Sol.]

$$\begin{aligned}\frac{\partial^2 f}{\partial X^2}[H] &= \left. \frac{d^2}{dt^2} \text{Tr} \{(X + tH)^\top A (X + tH) B\} \right|_{t=0} \\ &= 2 \text{Tr} \{ H^\top A H B \}\end{aligned}$$

Let us also try using the polarization identity.

[Sol.]

$$\begin{aligned}\frac{\partial^2 f}{\partial X^2}[H] &= \left. \frac{d^2}{dt^2} \text{Tr} \{(X + tH)^T A (X + tH) B\} \right|_{t=0} \\ &= 2 \text{Tr} \{ H^T A H B \}\end{aligned}$$

$$\frac{\partial^2 f}{\partial X^2}[H, K] = \frac{1}{4} \left(2 \text{Tr} \{ (H+K)^T A (H+K) B \} - 2 \text{Tr} \{ (H-K)^T A (H-K) B \} \right)$$

Let us also try using the polarization identity.

[Sol.]

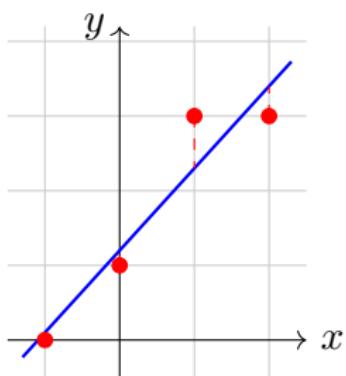
$$\begin{aligned}
 \frac{\partial^2 f}{\partial X^2}[H] &= \left. \frac{d^2}{dt^2} \text{Tr} \{(X + tH)^T A(X + tH)B\} \right|_{t=0} \\
 &= 2\text{Tr} \{H^T AHB\} \\
 \frac{\partial^2 f}{\partial X^2}[H, K] &= \frac{1}{4} \left(2\text{Tr} \{(H + K)^T A(H + K)B\} \right. \\
 &\quad \left. - 2\text{Tr} \{(H - K)^T A(H - K)B\} \right) \\
 &= 2\text{Tr} \{H^T AKB\}
 \end{aligned}$$

= $\text{Tr}\{H^T AKB\}$ + $\text{Tr}\{K^T AHB\}$

Thus, we can see that the results are identical.

Now we turn to the least mean square problem as another example.

Suppose the regression line for the data points is



$$\hat{y} = a_0 + a_1 x,$$

we can record the relationship as follows

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}}_X \underbrace{\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}}_w \approx \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}}_y$$

Def. (LLMSE problem)

The problem of finding the *linear least mean square estimate* is stated as below: given measurements \mathbf{y} over sample points X , find the optimal coefficients (weights) \mathbf{w} that gives the estimate

$$\hat{\mathbf{y}} = X\mathbf{w},$$

such that the mean square error (variance)

$$\mathcal{E} = \|\hat{\mathbf{y}} - \mathbf{y}\|^2$$

is minimized.

It is often used as a cost function in filtering and machine learning. For our talk, we will be focusing on the optimization problem of:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2.$$

By our knowledge of extrema occurs at stationary points, we know that

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

occurs when the directional derivative is always zero at that point, i.e.

$$\left. \left(\frac{\partial}{\partial \mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right) \right|_{\mathbf{w}^*} [\mathbf{h}] = 0 \quad (\forall \mathbf{h}),$$

or simply, the derivative is zero:

$$\left. \left(\frac{\partial}{\partial \mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right) \right|_{\mathbf{w}^*} = 0.$$

[Sol.] Recall that $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top A \mathbf{x} = 2\mathbf{x}^\top A$ (pg.20),

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 &= \frac{\partial}{\partial \mathbf{w}} (\mathbf{X}\vec{\mathbf{w}} - \vec{\mathbf{y}})^\top (\mathbf{X}\vec{\mathbf{w}} - \vec{\mathbf{y}}) \\ &= \left(\frac{\partial}{\partial (\mathbf{X}\vec{\mathbf{w}} - \vec{\mathbf{y}})} (\mathbf{X}\vec{\mathbf{w}} - \vec{\mathbf{y}})^\top (\mathbf{X}\vec{\mathbf{w}} - \vec{\mathbf{y}}) \right) \left[\frac{\partial (\mathbf{X}\vec{\mathbf{w}} - \vec{\mathbf{y}})}{\partial \mathbf{w}} \right] \\ &= 2(\mathbf{X}\vec{\mathbf{w}} - \vec{\mathbf{y}})^\top \mathbf{X}\end{aligned}$$

[Sol.] Recall that $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top A \mathbf{x} = 2\mathbf{x}^\top A$ (pg.20),

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 &= \frac{\partial}{\partial \mathbf{w}} (\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= 2(\mathbf{X}\mathbf{w} - \mathbf{y})^\top \mathbf{X}\end{aligned}$$

[Sol.] Recall that $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top A \mathbf{x} = 2A$ (pg.20),

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 &= \frac{\partial}{\partial \mathbf{w}} (X\mathbf{w} - \mathbf{y})^\top (X\mathbf{w} - \mathbf{y}) \\ &= 2(X\mathbf{w} - \mathbf{y})^\top X\end{aligned}$$

$$\begin{aligned}(X\mathbf{w}^* - \mathbf{y})^\top X &= 0 \\ X^\top X\mathbf{w}^* &= X^\top \mathbf{y} \\ \rightarrow \mathbf{w}^* &= \underline{(X^\top X)^{-1} X^\top \mathbf{y}}. \quad \blacksquare\end{aligned}$$

$$X\mathbf{w}^* \approx \mathbf{y}$$

[Sol.] Recall that $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top A \mathbf{x} = 2A$ (pg.20),

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 &= \frac{\partial}{\partial \mathbf{w}} (X\mathbf{w} - \mathbf{y})^\top (X\mathbf{w} - \mathbf{y}) \\ &= 2(X\mathbf{w} - \mathbf{y})^\top X\end{aligned}$$

$$\begin{aligned}(X\mathbf{w}^* - \mathbf{y})^\top X &= 0 \\ X^\top X\mathbf{w}^* &= X^\top \mathbf{y} \\ \rightarrow \mathbf{w}^* &= (X^\top X)^{-1} X^\top \mathbf{y}. \quad \blacksquare\end{aligned}$$

The term $(X^\top X)^{-1} X^\top$ is coined the **matrix pseudo-inverse**.

[Sol.] Recall that $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top A \mathbf{x} = 2A$ (pg.20),

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 &= \frac{\partial}{\partial \mathbf{w}} (\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= 2(\mathbf{X}\mathbf{w} - \mathbf{y})^\top \mathbf{X}\end{aligned}$$

$$\begin{aligned}(\mathbf{X}\mathbf{w}^* - \mathbf{y})^\top \mathbf{X} &= 0 \\ \mathbf{X}^\top \mathbf{X}\mathbf{w}^* &= \mathbf{X}^\top \mathbf{y} \\ \rightarrow \mathbf{w}^* &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad \blacksquare\end{aligned}$$

The term $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is coined the **matrix pseudo-inverse**.

Remark: What is the criterion for the existence of the pseudo-inverse?

\mathbf{X} : full rank \Leftrightarrow sample \mathbf{x} 's doesn't repeat

We've only checked that $\mathbf{w}^* = (X^\top X)^{-1} X^\top \mathbf{y}$ is an extrema, but we have yet to check whether its a maxima or a minima. A second derivative test is needed:

$$\frac{\partial}{\partial \mathbf{w}} ||X\mathbf{w} - \mathbf{y}||^2 = 2(X\mathbf{w} - \mathbf{y})^\top X$$

We've only checked that $\mathbf{w}^* = (X^\top X)^{-1} X^\top \mathbf{y}$ is an extrema, but we have yet to check whether its a maxima or a minima. A second derivative test is needed:

$$\frac{\partial}{\partial \mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 = 2(X\mathbf{w} - \mathbf{y})^\top X$$

$$\frac{\partial^2}{\partial \mathbf{w}^\top \partial \mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 = \frac{\partial}{\partial \mathbf{w}^\top} 2(X\mathbf{w} - \mathbf{y})^\top X = 2X^\top X$$

We've only checked that $\mathbf{w}^* = (X^\top X)^{-1} X^\top \mathbf{y}$ is an extrema, but we have yet to check whether its a maxima or a minima. A second derivative test is needed:

$$\frac{\partial}{\partial \mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 = 2(X\mathbf{w} - \mathbf{y})^\top X$$

$$\frac{\partial^2}{\partial \mathbf{w}^\top \partial \mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2 = \frac{\partial}{\partial \mathbf{w}^\top} 2(X\mathbf{w} - \mathbf{y})^\top X = 2X^\top X$$

Hence we know that the second derivative is **positive definite**, i.e. for all \mathbf{h} ,

$$\mathbf{h}^\top (2X^\top X) \mathbf{h} = 2(X\mathbf{h})^\top (X\mathbf{h}) > 0,$$

it is therefore a minima.

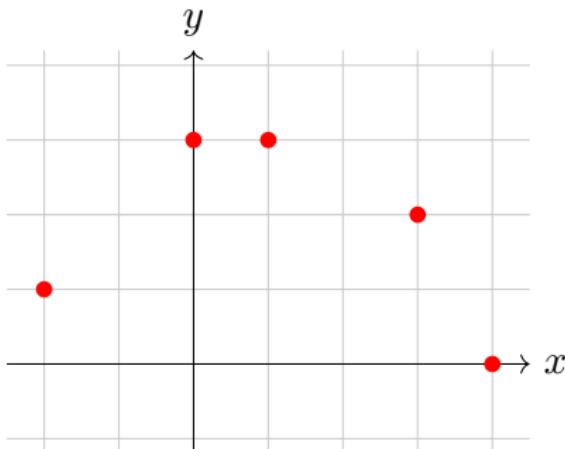
Given a set of data points (x, y) , find a function that interpolates them with the least mean square error.

x	y
-2	1
0	3
1	3
3	2
4	0

Find a quadratic:

$$\hat{y} = a_0 + a_1x + a_2x^2$$

such that the mean square error is minimized.



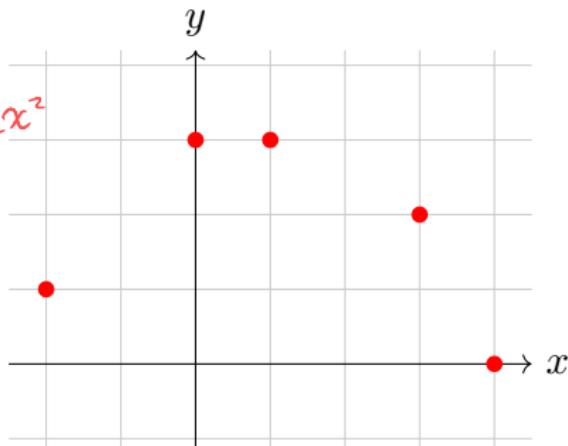
Given a set of data points (x, y) , find a function that interpolates them with the least mean square error.

We can rewrite the estimation equation as:

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} \begin{bmatrix} i \end{bmatrix} = a_0 + a_1 x + a_2 x^2$$

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 & -2 & 4 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \mathbf{X}\mathbf{w}$$

$$\mathbf{y} = [1, 3, 3, 2, 0]^T$$

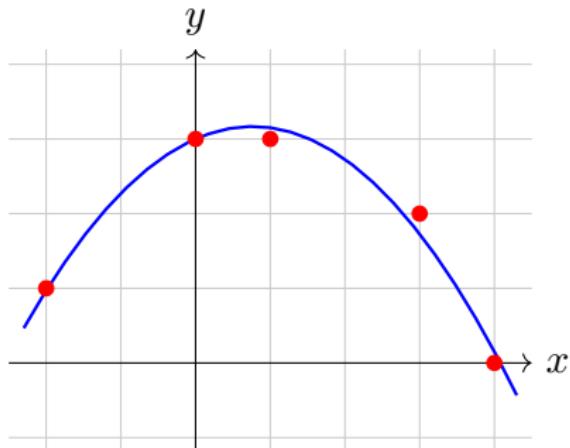


Given a set of data points (x, y) , find a function that interpolates them with the least mean square error.

$$(X^T X)^{-1} X^T = \begin{bmatrix} 0.14 & 0.43 & 0.43 & 0.14 & -0.14 \\ -0.28 & 0.047 & 0.12 & 0.10 & 0.00 \\ 0.06 & -0.05 & -0.06 & -0.01 & 0.06 \end{bmatrix}$$

$$\mathbf{w}^* = \begin{bmatrix} 3.0000 \\ 0.4394 \\ -0.2879 \end{bmatrix}$$

$$\hat{y} = 3 + 0.44x - 0.29x^2$$



- 1 Differentiation Revisited
- 2 Matrix Differentiation
- 3 Second Derivative
- 4 Examples of Matrix Derivatives
 - Derivatives of Matrix Functions ■ MLP ■ Summary

Given $A = A(t)$, what is

$$\frac{dA^{-1}}{dt}$$

expressed using $\frac{dA}{dt}$?

[Sol.]

$$\frac{dA^{-1}}{dt} = \frac{\partial A^{-1}}{\partial A} \left[\frac{dA}{dt} \right]$$

$$A^{-1} A = \mathbb{1}$$

$$\left[\frac{\partial A^{-1}}{\partial A} [H] \right] A + A^{-1} \frac{\partial A}{\partial A} [H] = 0$$

H

$$\frac{\partial A^{-1}}{\partial A} [H] = -A^{-1} H A^{-1}$$

Given $A = A(t)$, what is

$$\frac{dA^{-1}}{dt}$$

expressed using $\frac{dA}{dt}$?

[Sol.] Since we have

$$AA^{-1} = \mathbb{1},$$

Given $A = A(t)$, what is

$$\frac{dA^{-1}}{dt}$$

expressed using $\frac{dA}{dt}$?

[Sol.] Since we have

$$AA^{-1} = \mathbb{1},$$

by product rule:

$$\frac{\partial}{\partial A} (AA^{-1}) [H] =$$

Given $A = A(t)$, what is

$$\frac{dA^{-1}}{dt}$$

expressed using $\frac{dA}{dt}$?

[Sol.] Since we have

$$AA^{-1} = \mathbb{1},$$

by product rule:

$$\frac{\partial}{\partial A} (AA^{-1}) [H] = \cancel{\frac{\partial A}{\partial A}[H]}^H \cdot A^{-1} + A \cdot \frac{\partial A^{-1}}{\partial A}[H] = 0$$

$$\frac{\partial A^{-1}}{\partial A}[H] = -A^{-1} H A^{-1}$$

Given $A = A(t)$, what is

$$\frac{dA^{-1}}{dt}$$

expressed using $\frac{dA}{dt}$?

[Sol.] Since we have

$$AA^{-1} = \mathbb{1},$$

by product rule:

$$\begin{aligned} \frac{\partial}{\partial A} (AA^{-1}) [H] &= \cancel{\frac{\partial A}{\partial A}}^H [H] \cdot A^{-1} + A \cdot \frac{\partial A^{-1}}{\partial A} [H] = 0 \\ \frac{\partial A^{-1}}{\partial A} [H] &= -A^{-1} H A^{-1} \\ \Rightarrow \frac{dA^{-1}}{dt} &= \frac{\partial A^{-1}}{\partial A} \left[\frac{dA}{dt} \right] = -A^{-1} \frac{dA}{dt} A^{-1} \quad \blacksquare \end{aligned}$$

Lemma

The following identity holds for all square matrices A :

$$\det(e^A) = e^{\text{Tr}\{A\}}. \quad (27)$$

It can be immediately proven by Jordan canonical form of matrices.

Lemma

The following identity holds for all square matrices A :

$$\det(\underbrace{e^A}) = e^{\text{Tr}\{A\}}. \quad (27)$$

It can be immediately proven by Jordan canonical form of matrices.

e.g. $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ $A^2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = -\mathbb{1}$, $A^3 = -A$, $A^4 = \mathbb{1}$

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} = \mathbb{1} \underbrace{\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}}_{\text{identity matrix}} + A \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} = \mathbb{1} \cdot \cos(1) + A \cdot \sin(1) = \begin{bmatrix} \cos 1 & \sin 1 \\ -\sin 1 & \cos 1 \end{bmatrix}$$

$$\det(e^A) = \cos^2 1 + \sin^2 1 = 1 , \text{Tr}\{A\} = 0 , e^{\text{Tr}\{A\}} = 1 \quad \blacksquare$$

Hence if the matrix A is expressible as an exponential:

$$A = e^B,$$

i.e., it is invertible, then a solution is as below:

[Sol.]

$$\frac{d}{dt} \det(A(t)) = \frac{d}{dt} e^{\text{Tr}\{B(t)\}} = e^{\text{Tr}\{B(t)\}} \text{Tr}\left\{\frac{d}{dt} B\right\}$$

$$\begin{aligned}
 \text{Tr}\left\{ A^{-1} \frac{dA}{dt} \right\} &= \text{Tr}\left\{ e^B \frac{d}{dt} \sum_{n=0}^{\infty} \frac{B^n}{n!} \right\} = \text{Tr}\left\{ e^B \sum_{n=0}^{\infty} \underbrace{\frac{1}{n!} \left(\frac{d}{dt} B^n \right)}_{\frac{d}{dt} B^n} \right\} \\
 &= \text{Tr}\left\{ e^B \sum_{n=0}^{\infty} \underbrace{\frac{1}{n!} (nB^{n-1})}_{e^B} \frac{dB}{dt} \right\} \quad \frac{dB}{dt} B^{n-1} + B \frac{dB}{dt} B^{n-2} + \dots + B^{n-1} \frac{dB}{dt} \\
 &= \text{Tr}\left\{ \frac{dB}{dt} \right\}
 \end{aligned}$$

Hence if the matrix A is expressible as an exponential:

$$A = e^B,$$

i.e., it is invertible, then a solution is as below:

[Sol.]

$$\begin{aligned} \frac{d}{dt} \det(A(t)) &= \frac{d}{dt} e^{\text{Tr}\{B(t)\}} \\ &= e^{\text{Tr}\{B(t)\}} \text{Tr} \left\{ \frac{d}{dt} B(t) \right\} \\ &= \det(A(t)) \text{Tr} \left\{ A^{-1} \frac{dA(t)}{dt} \right\}. \quad \blacksquare \end{aligned}$$

$A =$

\det

Def. (Eigenvalues and Eigenvectors)

For a given square matrix A , it has **eigenvalues** $\{\lambda_i\}$ such that they satisfy:

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i,$$

where \mathbf{v}_i is the associated (right) **eigenvector**.

Def. (Eigenvalues and Eigenvectors)

For a given square matrix A , it has **eigenvalues** $\{\lambda_i\}$ such that they satisfy:

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i,$$

where \mathbf{v}_i is the associated (right) **eigenvector**. And if \mathbf{u}_i satisfies:

$$\mathbf{u}_i^T A = \lambda_i \mathbf{u}_i^T,$$

then \mathbf{u}_i is the associated left eigenvector.

For a time varying $A(t)$, what is $\frac{d\lambda_i(t)}{dt}$?
[Sol.]

For a time varying $A(t)$, what is $\frac{d\lambda_i(t)}{dt}$?

[Sol.]

Let us consider $\lambda_i(t)$ with its associated right and left eigenvectors: $v_i(t)$ and $u_i(t)$ that has length satisfying:

$$u_i^T v_i = 1.$$

For a time varying $A(t)$, what is $\frac{d\lambda_i(t)}{dt}$?

[Sol.]

Let us consider $\lambda_i(t)$ with its associated right and left eigenvectors: $v_i(t)$ and $u_i(t)$ that has length satisfying:

$$u_i^T v_i = 1.$$

Then,

$$\underline{u_i^T(t) A(t) v_i(t)} = \lambda_i(t)$$

$\lambda_i \vec{v}_i$

For a time varying $A(t)$, what is $\frac{d\lambda_i(t)}{dt}$?

[Sol.]

Let us consider $\lambda_i(t)$ with its associated right and left eigenvectors: $\mathbf{v}_i(t)$ and $\mathbf{u}_i(t)$ that has length satisfying:

$$\mathbf{u}_i^T \mathbf{v}_i = 1.$$

Then,

$$\mathbf{u}_i^T(t) A(t) \mathbf{v}_i(t) = \lambda_i(t)$$

$$\frac{d\lambda_i}{dt} = \frac{d\mathbf{u}_i^T}{dt} A \mathbf{v}_i + \mathbf{u}_i^T \frac{dA}{dt} \mathbf{v}_i + \mathbf{u}_i^T A \frac{d\mathbf{v}_i}{dt}$$

For a time varying $A(t)$, what is $\frac{d\lambda_i(t)}{dt}$?

[Sol.]

$$\frac{d\lambda_i}{dt} = \frac{d\mathbf{u}_i^T}{dt} A \mathbf{v}_i + \mathbf{u}_i^T \frac{dA}{dt} \mathbf{v}_i + \mathbf{u}_i^T A \frac{d\mathbf{v}_i}{dt}$$

But we also have

$$\begin{aligned} \frac{d\mathbf{u}_i^T}{dt} A \mathbf{v}_i + \underbrace{\mathbf{u}_i^T A \frac{d\mathbf{v}_i}{dt}}_{\lambda_i \mathbf{u}_i} &= \lambda_i \left(\frac{d\mathbf{u}_i^T}{dt} \mathbf{v}_i + \mathbf{u}_i^T \frac{d\mathbf{v}_i}{dt} \right) \\ &= \lambda_i \frac{d}{dt} \underbrace{(\mathbf{u}_i^T \mathbf{v}_i)}_1 = 0 \end{aligned}$$

Eigenvalues

For a time varying $A(t)$, what is $\frac{d\lambda_i(t)}{dt}$?

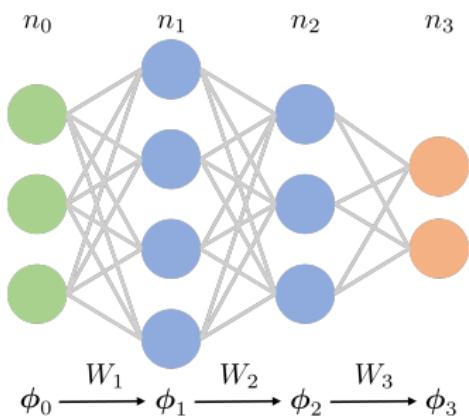
[Sol.]

$$\frac{d\lambda_i}{dt} = \frac{d\mathbf{u}_i^T}{dt} A \mathbf{v}_i + \mathbf{u}_i^T \frac{dA}{dt} \mathbf{v}_i + \mathbf{u}_i^T A \frac{d\mathbf{v}_i}{dt}$$

But we also have

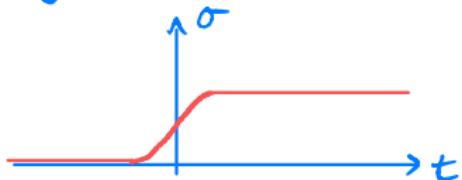
$$\begin{aligned}\frac{d\mathbf{u}_i^T}{dt} A \mathbf{v}_i + \mathbf{u}_i^T A \frac{d\mathbf{v}_i}{dt} &= \lambda_i \left(\frac{d\mathbf{u}_i^T}{dt} \mathbf{v}_i + \mathbf{u}_i^T \frac{d\mathbf{v}_i}{dt} \right) \\ &= \lambda_i \frac{d}{dt} (\mathbf{u}_i^T \mathbf{v}_i) = 0\end{aligned}$$

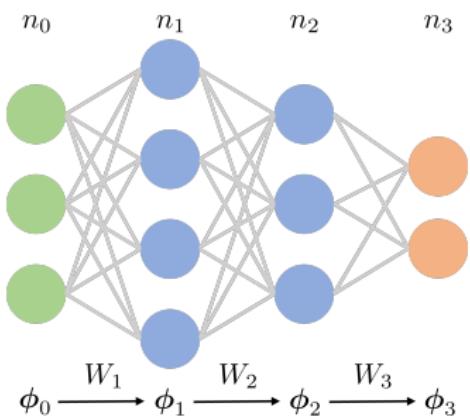
$$\rightarrow \frac{d\lambda_i}{dt} = \mathbf{u}_i^T \frac{dA}{dt} \mathbf{v}_i$$



$$\phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_n \end{bmatrix} \xrightarrow{\mathbf{w}_i} \text{circle split} \xrightarrow{\sigma} \phi_i = \sigma(\mathbf{w}_i^\top \phi)$$

Sigmoid: $\sigma(t) = \frac{1}{1+e^{-t}}$





$$\phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_n \end{bmatrix} \xrightarrow{\omega_i} \sigma \xrightarrow{\phi_i = \sigma(\omega_i^\top \phi)}$$

For layer- ℓ with n_ℓ perceptrons,

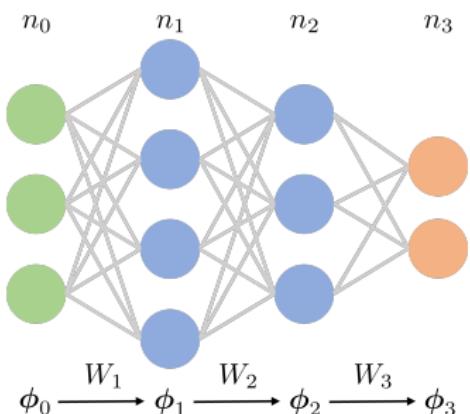
$$\mathbb{R}^{n_\ell} \ni \phi_\ell = \sigma(W_\ell^\top \phi_{\ell-1}),$$

the weight matrix

$$W_\ell = [\omega_{\ell,1} \cdots \omega_{\ell,n_\ell}] \in \mathbb{R}^{n_\ell \times n_{\ell+1}}.$$

abuse of notation

$$\vec{\phi}_\ell = \sigma \left(\begin{bmatrix} \vec{w}_{\ell,1}^\top \vec{\phi}_{\ell-1} \\ \vdots \\ \vec{w}_{\ell,n_\ell}^\top \vec{\phi}_{\ell-1} \end{bmatrix} \right) \stackrel{?}{=} \begin{bmatrix} \sigma(\vec{w}_{\ell,1}^\top \vec{\phi}_{\ell-1}) \\ \vdots \\ \sigma(\vec{w}_{\ell,n_\ell}^\top \vec{\phi}_{\ell-1}) \end{bmatrix}$$



$$\phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_n \end{bmatrix} \xrightarrow{\omega_i} \sigma \xrightarrow{\phi_i = \sigma(\omega_i^\top \phi)}$$

For layer- ℓ with n_ℓ perceptrons,

$$\mathbb{R}^{n_\ell} \ni \phi_\ell = \sigma(W_\ell^\top \phi_{\ell-1}),$$

the weight matrix

$$W_\ell = [\omega_{\ell,1} \cdots \omega_{\ell,n_\ell}] \in \mathbb{R}^{n_\ell \times n_{\ell+1}}.$$

Hence if we denote $\sigma(W^\top \phi) = f(W, \phi)$, the network above can be reduced to

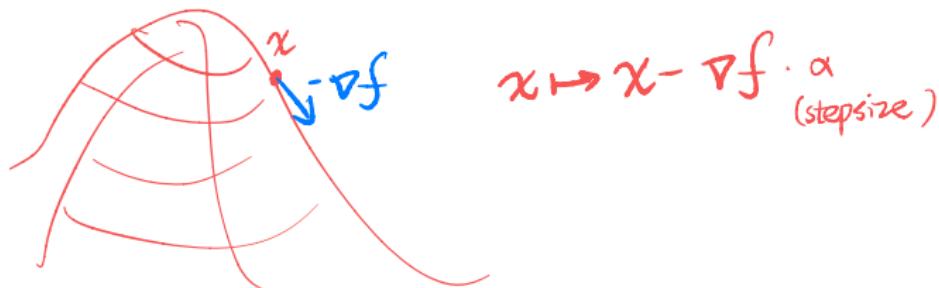
$$\phi_3 = f(W_3, \cdot) \circ f(W_2, \cdot) \circ f(W_1, \phi_0).$$



Our goal is to do backpropagation: given a cost function $\mathcal{E}(\phi_3)$, find the optimal weights W_1, W_2, W_3 such that the cost is minimized.

Hence, we need to first calculate the derivatives of $f(W, \phi)$.

Gradient Descent



Our goal is to do backpropagation: given a cost function $\mathcal{E}(\phi_3)$, find the optimal weights W_1, W_2, W_3 such that the cost is minimized.

Hence, we need to first calculate the derivatives of $f(W, \phi)$.

- ▶ Differentiate w.r.t. W : for $W_\ell = [W_{\ell,1}, W_{\ell,2}, \dots]$ and $H = [H_1, H_2, \dots]$

$$\frac{\partial f}{\partial W}(W_\ell, \phi_{\ell-1})[H] =$$

Our goal is to do backpropagation: given a cost function $\mathcal{E}(\phi_3)$, find the optimal weights W_1, W_2, W_3 such that the cost is minimized.

Hence, we need to first calculate the derivatives of $f(W, \phi)$.

- ▶ Differentiate w.r.t. W : for $W_\ell = [W_{\ell,1}, W_{\ell,2}, \dots]$ and $H = [H_1, H_2, \dots]$

$$\frac{\partial f}{\partial W}(W_\ell, \phi_{\ell-1})[H] = \begin{bmatrix} \frac{\partial}{\partial W} \sigma(W_{\ell,1}^\top \phi_{\ell-1})[H] \\ \frac{\partial}{\partial W} \sigma(W_{\ell,2}^\top \phi_{\ell-1})[H] \\ \vdots \end{bmatrix} = \begin{bmatrix} \sigma'(W_{\ell,1}^\top \phi_{\ell-1}) H_1^\top \phi_{\ell-1} \\ \sigma'(W_{\ell,2}^\top \phi_{\ell-1}) H_2^\top \phi_{\ell-1} \\ \vdots \end{bmatrix}$$

Our goal is to do backpropagation: given a cost function $\mathcal{E}(\phi_3)$, find the optimal weights W_1, W_2, W_3 such that the cost is minimized.

Hence, we need to first calculate the derivatives of $f(W, \phi)$.

- ▶ Differentiate w.r.t. W : for $W_\ell = [W_{\ell,1}, W_{\ell,2}, \dots]$ and $H = [H_1, H_2, \dots]$

$$\begin{aligned}\frac{\partial f}{\partial W}(W_\ell, \phi_{\ell-1})[H] &= \begin{bmatrix} \frac{\partial}{\partial W} \sigma(W_{\ell,1}^\top \phi_{\ell-1})[H] \\ \vdots \end{bmatrix} = \begin{bmatrix} \sigma'(W_{\ell,1}^\top \phi_{\ell-1}) H_1^\top \phi_{\ell-1} \\ \vdots \end{bmatrix} \\ &= \begin{bmatrix} \sigma'(W_{\ell,1}^\top \phi_{\ell-1}) & 0 \\ 0 & \ddots \end{bmatrix} \begin{bmatrix} H_1^\top \phi_{\ell-1} \\ \vdots \end{bmatrix} \\ &= \text{diag}(\sigma'(W_\ell^\top \phi_{\ell-1})) H^\top \phi_{\ell-1}.\end{aligned}$$

$$\text{diag}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right) = \begin{bmatrix} x_1 & & & \\ & x_2 & \ddots & \\ & & \ddots & x_n \end{bmatrix} \quad \text{abuse of notation}$$

Our goal is to do backpropagation: given a cost function $\mathcal{E}(\phi_3)$, find the optimal weights W_1, W_2, W_3 such that the cost is minimized.

Hence, we need to first calculate the derivatives of $f(W, \phi)$.

- ▶ Differentiate w.r.t. ϕ :

$$\frac{\partial f}{\partial \phi}(W_\ell, \phi_{\ell-1})[\mathbf{h}] =$$

Our goal is to do backpropagation: given a cost function $\mathcal{E}(\phi_3)$, find the optimal weights W_1, W_2, W_3 such that the cost is minimized.

Hence, we need to first calculate the derivatives of $f(W, \phi)$.

- ▶ Differentiate w.r.t. ϕ :

$$\frac{\partial f}{\partial \phi}(W_\ell, \phi_{\ell-1})[\mathbf{h}] = \begin{bmatrix} \frac{\partial}{\partial \phi} \sigma(W_{\ell,1}^\top \phi_{\ell-1})[\mathbf{h}] \\ \vdots \end{bmatrix} = \begin{bmatrix} \sigma'(W_{\ell,1}^\top \phi_{\ell-1}) W_{\ell,1}^\top \mathbf{h} \\ \vdots \end{bmatrix}$$

Our goal is to do backpropagation: given a cost function $\mathcal{E}(\phi_3)$, find the optimal weights W_1, W_2, W_3 such that the cost is minimized.

Hence, we need to first calculate the derivatives of $f(W, \phi)$.

- ▶ Differentiate w.r.t. ϕ :

$$\begin{aligned}
 \frac{\partial f}{\partial \phi}(W_\ell, \phi_{\ell-1})[\mathbf{h}] &= \begin{bmatrix} \frac{\partial}{\partial \phi} \sigma(W_{\ell,1}^\top \phi_{\ell-1})[\mathbf{h}] \\ \vdots \end{bmatrix} = \begin{bmatrix} \sigma'(W_{\ell,1}^\top \phi_{\ell-1}) W_{\ell,1}^\top \mathbf{h} \\ \vdots \end{bmatrix} \\
 &= \begin{bmatrix} \sigma'(W_{\ell,1}^\top \phi_{\ell-1}) & 0 \\ 0 & \ddots \end{bmatrix} \begin{bmatrix} W_{\ell,1}^\top \mathbf{h} \\ \vdots \end{bmatrix} \\
 &= \text{diag}(\sigma'(W_\ell^\top \phi_{\ell-1})) W_\ell^\top \mathbf{h}.
 \end{aligned}$$

MLP - Backpropagation

Suppose we want to alter W_1 by the methods of **gradient descent**. The gradient to W_1 can be calculated:

$$\frac{\partial \mathcal{E}}{\partial W_1}(\phi_3)[H] = \frac{\partial \mathcal{E}}{\partial \phi_3} \circ \left. \frac{\partial f}{\partial \phi_2} \right|_{(W_3, \phi_2)} \circ \left. \frac{\partial f}{\partial \phi_1} \right|_{(W_2, \phi_1)} \circ \left. \frac{\partial f}{\partial W_1} \right|_{(W_1, \phi_0)} [H]$$

$$\mathcal{E}(\phi_3) = \mathcal{E} \circ f(w_3, \cdot) \circ f(w_2, \cdot) \circ f(w_1, \phi_0)$$

MLP - Backpropagation

Suppose we want to alter W_1 by the methods of **gradient descent**. The gradient to W_1 can be calculated:

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial W_1}(\phi_3)[H] &= \frac{\partial \mathcal{E}}{\partial \phi_3} \circ \left. \frac{\partial f}{\partial \phi_2} \right|_{(W_3, \phi_2)} \circ \left. \frac{\partial f}{\partial \phi_1} \right|_{(W_2, \phi_1)} \circ \left. \frac{\partial f}{\partial W_1} \right|_{(W_1, \phi_0)} [H] \\ &= \frac{\partial \mathcal{E}}{\partial \phi_3} [\text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \cdot \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \cdot \text{diag}(\sigma'(W_1^T \phi_0)) H^T \phi_0]\end{aligned}$$

MLP - Backpropagation

Suppose we want to alter W_1 by the methods of **gradient descent**. The gradient to W_1 can be calculated:

$$\begin{aligned} \frac{\partial \mathcal{E}}{\partial W_1}(\phi_3)[H] &= \frac{\partial \mathcal{E}}{\partial \phi_3} \circ \left. \frac{\partial f}{\partial \phi_2} \right|_{(W_3, \phi_2)} \circ \left. \frac{\partial f}{\partial \phi_1} \right|_{(W_2, \phi_1)} \circ \left. \frac{\partial f}{\partial W_1} \right|_{(W_1, \phi_0)} [H] \\ &\stackrel{\text{def } \mathcal{E}}{=} \text{Tr} \left\{ \frac{\partial \mathcal{E}}{\partial \phi_3} [\text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \cdot \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \cdot \text{diag}(\sigma'(W_1^T \phi_0)) H^T \phi_0] \right\} \\ &= \text{Tr} \left\{ H^T \phi_0 \frac{\partial \mathcal{E}}{\partial \phi_3} \text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \text{diag}(\sigma'(W_1^T \phi_0)) \right\} \\ &= \text{Tr} \{ H^T \nabla_{W_1} \mathcal{E} \} = \langle H, \nabla_{W_1} \mathcal{E} \rangle \end{aligned}$$

$\therefore \vec{p}_3$: column vector

$\therefore \frac{\partial \mathcal{E}}{\partial \phi_3}$: row vector

$$\therefore \frac{\partial \mathcal{E}}{\partial \vec{p}_3}[\vec{h}] = \frac{\partial \mathcal{E}}{\partial \vec{p}_3} \cdot \vec{h}$$

Suppose we want to alter W_1 by the methods of **gradient descent**. The gradient to W_1 can be calculated:

$$\begin{aligned}
 \frac{\partial \mathcal{E}}{\partial W_1}(\phi_3)[H] &= \frac{\partial \mathcal{E}}{\partial \phi_3} \circ \left. \frac{\partial f}{\partial \phi_2} \right|_{(W_3, \phi_2)} \circ \left. \frac{\partial f}{\partial \phi_1} \right|_{(W_2, \phi_1)} \circ \left. \frac{\partial f}{\partial W_1} \right|_{(W_1, \phi_0)} [H] \\
 &= \frac{\partial \mathcal{E}}{\partial \phi_3} [\text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \cdot \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \cdot \text{diag}(\sigma'(W_1^T \phi_0)) H^T \phi_0] \\
 &= \text{Tr} \left\{ H^T \phi_0 \frac{\partial \mathcal{E}}{\partial \phi_3} \text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \text{diag}(\sigma'(W_1^T \phi_0)) \right\} \\
 &= \text{Tr} \{ H^T \nabla_{W_1} \mathcal{E} \} = \langle H, \nabla_{W_1} \mathcal{E} \rangle \\
 \Rightarrow \nabla_{W_1} \mathcal{E} &= \phi_0 \frac{\partial \mathcal{E}}{\partial \phi_3} \text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \text{diag}(\sigma'(W_1^T \phi_0))
 \end{aligned}$$

MLP - Backpropagation

Suppose we want to alter W_1 by the methods of **gradient descent**. The gradient to W_1 can be calculated:

$$\begin{aligned}
 \frac{\partial \mathcal{E}}{\partial W_1}(\phi_3)[H] &= \frac{\partial \mathcal{E}}{\partial \phi_3} \circ \left. \frac{\partial f}{\partial \phi_2} \right|_{(W_3, \phi_2)} \circ \left. \frac{\partial f}{\partial \phi_1} \right|_{(W_2, \phi_1)} \circ \left. \frac{\partial f}{\partial W_1} \right|_{(W_1, \phi_0)} [H] \\
 &= \frac{\partial \mathcal{E}}{\partial \phi_3} [\text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \cdot \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \cdot \text{diag}(\sigma'(W_1^T \phi_0)) H^T \phi_0] \\
 &= \text{Tr} \left\{ H^T \phi_0 \frac{\partial \mathcal{E}}{\partial \phi_3} \text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \text{diag}(\sigma'(W_1^T \phi_0)) \right\} \\
 &= \text{Tr} \{ H^T \nabla_{W_1} \mathcal{E} \} = \langle H, \nabla_{W_1} \mathcal{E} \rangle \\
 \Rightarrow \nabla_{W_1} \mathcal{E} &= \phi_0 \frac{\partial \mathcal{E}}{\partial \phi_3} \text{diag}(\sigma'(W_3^T \phi_2)) W_3^T \text{diag}(\sigma'(W_2^T \phi_1)) W_2^T \text{diag}(\sigma'(W_1^T \phi_0))
 \end{aligned}$$

After each iteration, W_1 is updated via

$$W_1 \mapsto W_1 - \alpha \nabla_{W_1} \mathcal{E},$$

with α a suitable decreasing step size.

Derivatives are linear operators.

Derivatives

Let x be a scalar in \mathbb{R} or a vector in \mathbb{R}^n or a matrix in $\mathbb{R}^{m \times n}$.

Given a function $f(x)$, we have the first and second derivatives as:

$$\frac{\partial f}{\partial x}(x)[h] = \left. \frac{d}{dt} f(x + th) \right|_{t=0}, \quad (28)$$

$$\frac{\partial^2 f}{\partial x^2}(x)[h] = \left. \frac{d^2}{dt^2} f(x + th) \right|_{t=0}. \quad (29)$$

Def. (Gradient)

Given an inner product $\langle \cdot, \cdot \rangle$, the **gradient** of a function $f(x)$ ($x \in \mathbb{R}$, \mathbb{R}^n or $\mathbb{R}^{m \times n}$) is defined as

$$\nabla f(x) \text{ s.t. } \frac{\partial f}{\partial x}(x)[h] = \langle \nabla f(x), h \rangle. \quad (30)$$

Def. (Hessian)

Given an inner product $\langle \cdot, \cdot \rangle$, the **Hessian** of a function $f(x)$ ($x \in \mathbb{R}$, \mathbb{R}^n , or $\mathbb{R}^{m \times n}$) is defined as

$$\mathsf{H}_f(x) \text{ s.t. } \frac{\partial^2 f}{\partial x^2}(x)[h] = \langle \mathsf{H}_f(x)[h], h \rangle. \quad (31)$$

Notes: