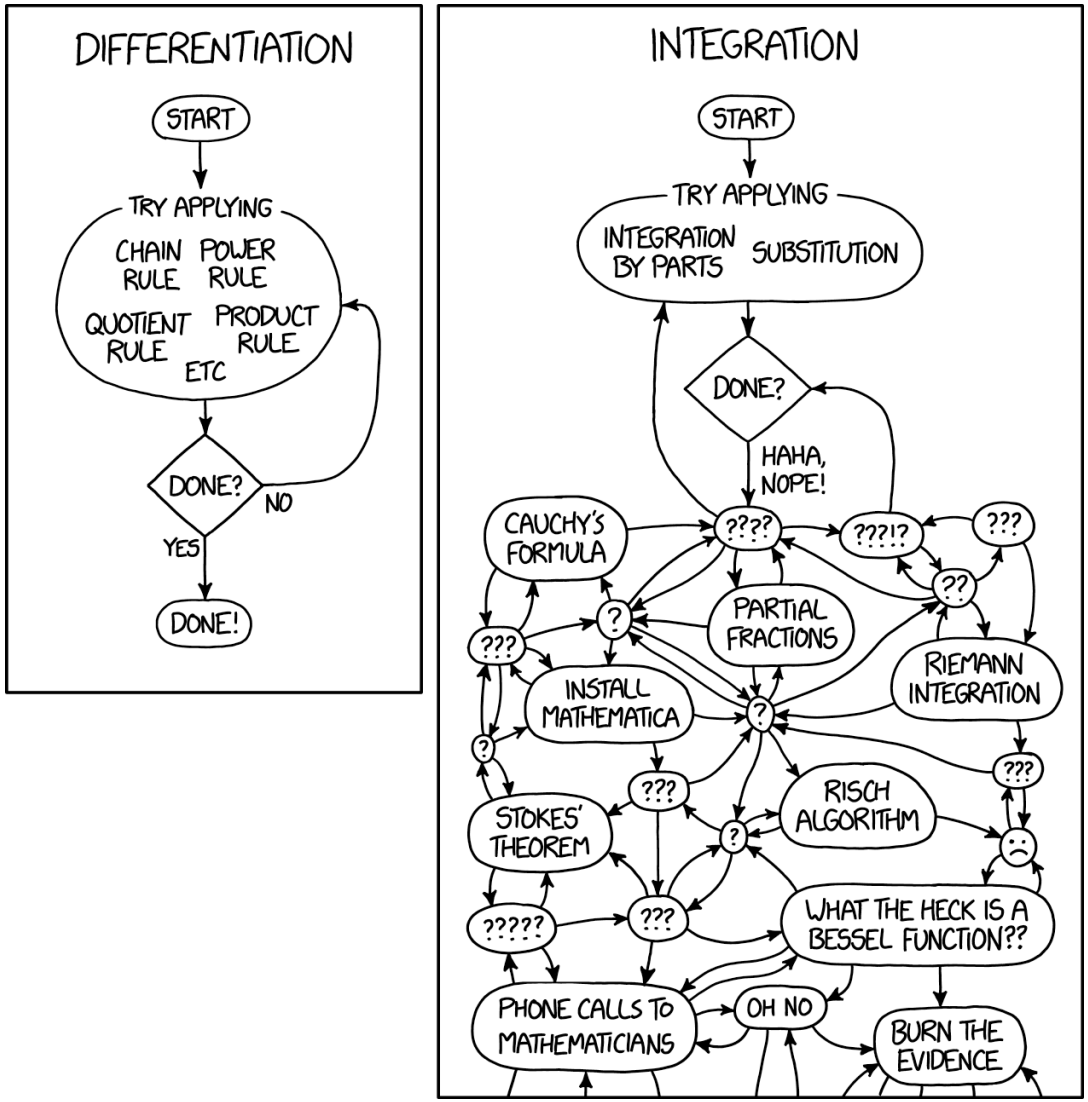
Matrix Differentiation

Wen Perng

NTUEE

2024 September 18

Derivatives are easy! (Source: XKCD)



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:

$$\left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x^*} = 0,$$

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

$$\left.\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\right|_{x^*} > 0.$$

We denote the point of minima as

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

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:

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

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

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

$$\frac{\mathrm{d}f(A)}{\mathrm{d}A} \text{ and } \frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}},$$

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

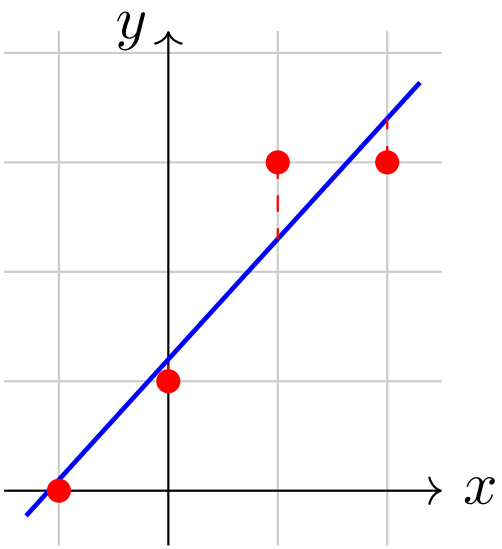
What do the derivatives mean and how do they behave? Note that we will work in cases where matrices are treated as *matrices* instead of *arrays*.

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

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

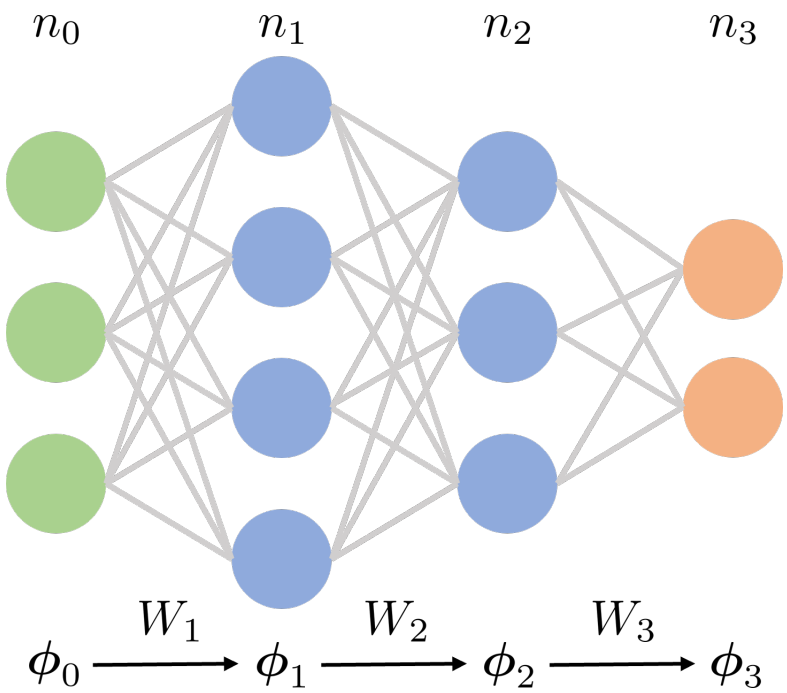
where $X \in \mathbb{R}^{m \times n}$,$\boldsymbol{w} \in \mathbb{R}^n$,$\boldsymbol{y} \in \mathbb{R}^m$. We would like to find a $\boldsymbol{w}$ that *best estimates* $\boldsymbol{y}$ by the approximation:

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



$$X = [1, \boldsymbol{x}]$$
$$\boldsymbol{w} = [a_0, a_1]^\mathsf{T}$$
$$\hat{y} = a_0 + a_1 x$$

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.



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

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

Lastly, others might want to differentiate functions such as

$$f = \boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x},\ \det(A),\ \mathrm{tr}(A),\ \lambda_i(A),\ \cdots .$$

Or even crazier, the QR-decomposition:

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

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

Def. (Transpose)

Flip the matrix along its diagonal:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}. \tag{1}$$

What number when transposed doubles?

$$午^{\mathsf{T}} = 什$$

Def. (Inner Product)

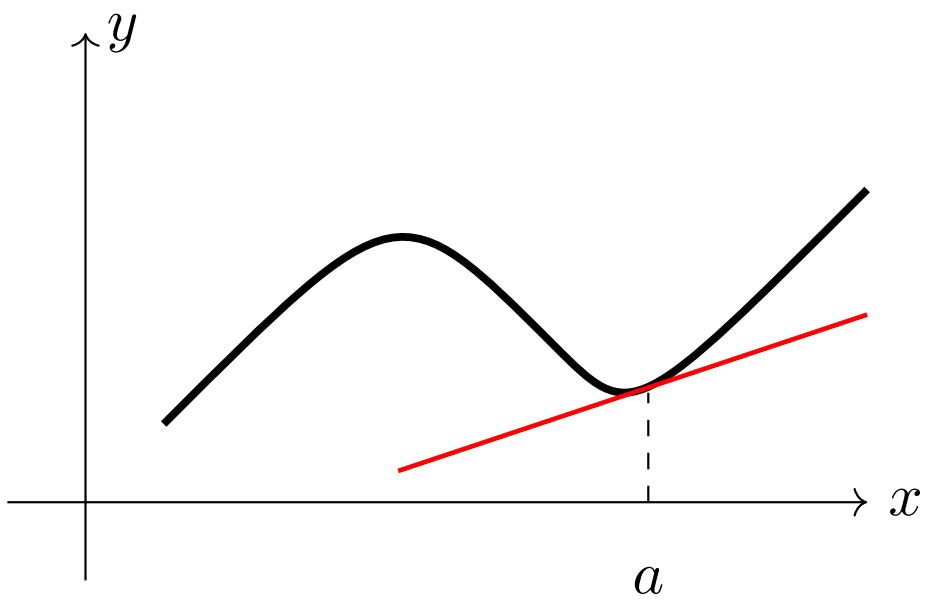
We denote the *Euclidean* inner product of two column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ as:

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{y} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle.$$

Table of Contents

Table of Contents

Derivatives are linearizations. Consider the easiest case below.



We know that the first-order approximation will be

$$y(x) = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_a (x-a) + y(a).$$

We can rewrite the previous equation into the form below:

$$y(a + \mathrm{d}x) = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_a \mathrm{d}x + y(a)$$

$$\mathrm{d}y = y(a + \mathrm{d}x) - y(a) = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_a \mathrm{d}x$$

Thus obtaining:

$$\mathrm{d}y = \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) \mathrm{d}x \tag{2}$$

We don't need to "divide" the $\mathrm{d}x$ over to the other side. The equation above will generalize to future cases.

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$\mathrm{d}y = \frac{\mathrm{d}y}{\mathrm{d}x}\mathrm{d}x.$$

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

e.g.

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

By plugging in the definiton above:

$$\mathrm{d}\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}) = A(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - A\boldsymbol{x} = A\,\mathrm{d}\boldsymbol{x}$$

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} = A$$

Prop. (Derivatives as Linear Approximation)

For $y = f(x)$,

$$\mathrm{d}y = \frac{\mathrm{d}y}{\mathrm{d}x}\mathrm{d}x.$$

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

e.g.

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

By plugging in the definiton above:

$$\mathrm{d}\boldsymbol{f} = \boldsymbol{f}(A + \mathrm{d}A) - \boldsymbol{f}(A) = (A + \mathrm{d}A)\boldsymbol{x} - A\boldsymbol{x} = \mathrm{d}A\ \boldsymbol{x}$$

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}A} \stackrel{?}{=} \boldsymbol{x}$$

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

$$\mathrm{d}y = \mathcal{L}\{\mathrm{d}x\}. \tag{3}$$

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

$$\begin{aligned} \mathcal{L}\{\mathrm{d}x\} &= f'\mathrm{d}x \\ \mathcal{L}\{\mathrm{d}x\} &= \mathrm{d}x f' \\ \mathcal{L}\{\mathrm{d}x\} &= \sqrt{f'}\mathrm{d}x\sqrt{f'}. \end{aligned}$$

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

$$\mathcal{L}\{a\mathrm{d}x_1 + \mathrm{d}x_2\} = a\mathcal{L}\{\mathrm{d}x_1\} + \mathcal{L}\{\mathrm{d}x_2\}.$$

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]

[1] Since 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.

Table of Contents

Two methods are possible:

1. Term-wise differentiation,
2. **Directional derivatives** (if it exists): the directional derivative of $f(x)$ by x at x_0 in the direction of h will be

$$\mathrm{d}f = \frac{\partial f(x)}{\partial x}(x_0)[h] := \frac{\mathrm{d}}{\mathrm{d}t} f(x_0 + th)\bigg|_{t=0}, \tag{4}$$

   or equivalently, find a linear $\mathcal{L}$ such that

$$f(x) = f(x_0) + \mathcal{L}\{x - x_0\} + o(\|x - x_0\|). \tag{5}$$

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 $\boldsymbol{x} \in \mathbb{R}^n$:
Let us suppose

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

[Sol.]

$$\begin{aligned}\frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x})[\boldsymbol{h}] &= \frac{\mathrm{d}}{\mathrm{d}t}\left((\boldsymbol{x}+t\boldsymbol{h})^{\mathsf{T}} A(\boldsymbol{x}+t\boldsymbol{h})\right)\Big|_{t=0} \\ &= \boldsymbol{h}^{\mathsf{T}} A\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}} A\boldsymbol{h} \\ &= 2\boldsymbol{x}^{\mathsf{T}} A\boldsymbol{h}.\end{aligned}$$

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

$$\frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x}) \equiv \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \equiv \frac{\partial f}{\partial \boldsymbol{x}} = 2\boldsymbol{x}A. \tag{6}$$

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

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{n \times 1},$$

we know from calculus that its infinitesimal change is:

$$\mathrm{d}f = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \cdots + \frac{\partial f}{\partial x_n}\mathrm{d}x_n. \tag{7}$$

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

Notice that

$$\mathrm{d}f = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \cdots + \frac{\partial f}{\partial x_n}\mathrm{d}x_n$$

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

$$\mathrm{d}f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \begin{bmatrix} \mathrm{d}x_1 \\ \vdots \\ \mathrm{d}x_n \end{bmatrix} = (\nabla f)^{\mathsf{T}}\mathrm{d}\boldsymbol{x}.$$

Def. (Gradient)

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

$$\nabla f(\boldsymbol{x}) \in \mathbb{R}^n \text{ s.t. } \mathrm{d}f = \langle \nabla f(\boldsymbol{x}), \boldsymbol{h} \rangle. \tag{8}$$

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

$$(\nabla f)^{\mathsf{T}} \mathrm{d}\boldsymbol{x} = \mathrm{d}f = \frac{\partial f}{\partial \boldsymbol{x}} \mathrm{d}\boldsymbol{x}$$

as operation by *left-multiplication*, then

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

And if we define

$$\mathrm{d}f = \mathrm{d}\boldsymbol{x}^{\mathsf{T}} \frac{\partial f}{\partial \boldsymbol{x}^{\mathsf{T}}}$$

by *right-multiplication*, then

$$\frac{\partial f}{\partial \boldsymbol{x}^{\mathsf{T}}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \nabla f.$$

Back to our example above:

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x}$$

[Sol.]

$$\mathrm{d}f = \frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x})[\boldsymbol{h}] = 2\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{h} = \langle 2A\boldsymbol{x}, \boldsymbol{h} \rangle$$
$$\nabla f(\boldsymbol{x}) = 2A\boldsymbol{x}.$$

Notice that:

$$1 \times n \rightarrow \frac{\partial f}{\partial \boldsymbol{x}} \begin{matrix} \leftarrow 1 \times 1 \\ \leftarrow n \times 1 \end{matrix}$$

$$n \times 1 \rightarrow \frac{\partial f}{\partial \boldsymbol{x}^{\mathsf{T}}} \begin{matrix} \leftarrow 1 \times 1 \\ \leftarrow 1 \times n \end{matrix}$$

We can even have:

$$m \times n \rightarrow \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \begin{matrix} \leftarrow m \times 1 \\ \leftarrow n \times 1 \end{matrix}$$

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

(2) Differentiate vector function $\boldsymbol{f}$ by another vector $\boldsymbol{x}$:
If $\boldsymbol{x}$ is a column vector and $\boldsymbol{f}$ is of the form:

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

then we have:

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

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

1. Gradient (tranpose):

$$\nabla f := \frac{\partial f}{\partial \boldsymbol{x}^\mathsf{T}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \nabla^\mathsf{T} f := \frac{\partial f}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (9)$$

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

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

2. Hessian:

$$\mathsf{H}_f := \nabla\nabla^\mathsf{T} f = \frac{\partial^2 f}{\partial \boldsymbol{x}^\mathsf{T} \partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \tag{10}$$

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

For a vector function $\boldsymbol{y}(\boldsymbol{x}) = [y_1(\boldsymbol{x}), \cdots, y_n(\boldsymbol{x})]^{\mathsf{T}}$ and position vector $\boldsymbol{x} = [x_1, \cdots, x_n]^{\mathsf{T}}$,

3. Jacobian:

$$J := \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_n}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}, \tag{11}$$

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

$$\int \cdots \int_{\boldsymbol{y}(\mathcal{X})} f(\boldsymbol{y}) \mathrm{d}y_1 \cdots \mathrm{d}y_n = \int \cdots \int_{\mathcal{X}} f(\boldsymbol{y}(\boldsymbol{x})) |J| \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

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

$$\mathrm{d}f = \frac{\partial f}{\partial X}(X_0)[H] = \left.\frac{\mathrm{d}}{\mathrm{d}t} f(X_0 + tH)\right|_{t=0}.$$

Some examples are as follows:

1. $f(A) = A\boldsymbol{x}$

$$\mathrm{d}f = (\mathrm{d}A)\boldsymbol{x}$$

2. $f(A) = A^{\mathsf{T}}A$

$$\mathrm{d}f = (A + \mathrm{d}A)^{\mathsf{T}}(A + \mathrm{d}A) - A^{\mathsf{T}}A = (\mathrm{d}A)^{\mathsf{T}}A + A^{\mathsf{T}}\mathrm{d}A$$

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

Def. (Trace)

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

$$\mathrm{Tr}\left\{A\right\} = \sum_{i=1}^{n} a_{ii}$$

[Sol.]

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

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

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

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 $V \times V \to \mathbb{R}$ 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:

$$\langle X, Y \rangle = \mathrm{Tr}\left\{ X^{\mathsf{T}} Y \right\}. \tag{13}$$

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

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

$$
\frac{\partial f}{\partial X}[H] = 2\mathrm{Tr}\left\{H^{\mathsf{T}} A X B\right\}, \ \nabla f(X) = 2AXB.
$$

Remember that the following properties of derivatives still applies:

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

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

2. Product Rule:

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

Table of Contents

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

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

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

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{h}) &= f(\boldsymbol{x}) + \frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x})[\boldsymbol{h}] + \frac{1}{2}\frac{\partial^2 f}{\partial \boldsymbol{x}^2}(\boldsymbol{x})[\boldsymbol{h}] + \cdots \\ &= f(\boldsymbol{x}) + (\nabla f(\boldsymbol{x}))^{\mathsf{T}} \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^{\mathsf{T}} \mathsf{H}_f(\boldsymbol{x}) \boldsymbol{h} + \cdots , \end{aligned} \tag{17}$$

where the matrix $\mathsf{H}_f(\boldsymbol{x})$ is called the **Hessian matrix** of f at $\boldsymbol{x}$. We have seen that the Hessian matrix $\mathsf{H}_f = \frac{\partial^2 f}{\partial \boldsymbol{x}^{\mathsf{T}} \partial \boldsymbol{x}}$ determines the behavior of the extrema.

Def. (Definiteness)

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

$$\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x} > 0 \Longleftrightarrow A \succ 0.$$

Moreover, we have:

(positive semi-definite)	$\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x} \geq 0 \Longleftrightarrow A \succeq 0$
(negative definite)	$\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x} < 0 \Longleftrightarrow A \prec 0$
(negative semi-definite)	$\boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x} \leq 0 \Longleftrightarrow A \preceq 0$

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 $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, then

$$\begin{aligned} f(X+H) &= f(X) + \frac{\partial f}{\partial X}(X)[H] + \frac{1}{2}\frac{\partial^2 f}{\partial X^2}(X)[H] + \cdots \\ &= f(X) + \langle \nabla f(X), H\rangle + \frac{1}{2}\langle \mathsf{H}_f(X)[H], H\rangle + \cdots . \end{aligned} \tag{18}$$

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

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

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

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

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

So how do we caculate the second derivative?
The second derivative can be derived via directional derivatives:

$$\frac{\partial^2 f}{\partial x^2}[h,k] = \left.\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}s} f(x+th+sk)\right|_{t=0,s=0}. \tag{21}$$

Or gaining idea from Taylor series, one can first calculate

$$\frac{\partial^2 f}{\partial x^2}[h] = \left.\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x+th)\right|_{t=0}, \tag{22}$$

then utilize the **polarization identity**:

$$\frac{\partial^2 f}{\partial x^2}[h,k] = \frac{1}{4}\left(\frac{\partial^2 f}{\partial x^2}[h+k] - \frac{\partial^2 f}{\partial x^2}[h-k]\right). \tag{23}$$

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

[Sol.]

$$\frac{\partial f}{\partial X}[H] = 2\mathrm{Tr}\left\{H^{\mathsf{T}} A X B\right\}.$$

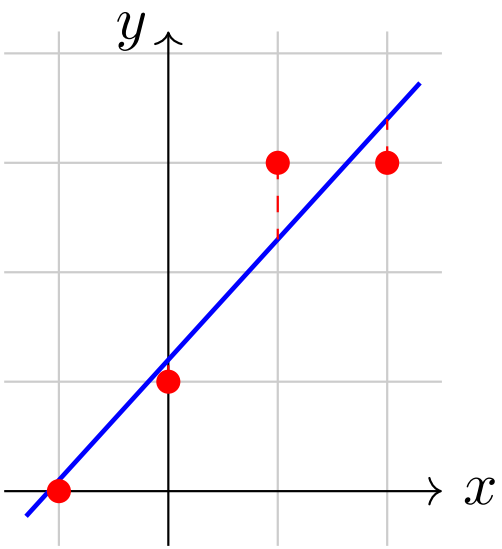
$$\begin{aligned}
\frac{\partial^2 f}{\partial X^2}[H,K] &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}s} f(X + tH + sK)\Big|_{t=0,s=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\partial f}{\partial X}(X + sK)[H]\right)\Big|_{s=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}s} 2\mathrm{Tr}\left\{H^{\mathsf{T}}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

Let us also try using the polarization identity.
[Sol.]

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

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

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}}_{\boldsymbol{w}} \approx \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}}_{\boldsymbol{y}}$$

Def. (LLMSE problem)

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

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

such that the mean square error (variance)

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

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:

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

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

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

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

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

or simply, the derivative is zero:

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

[Sol.]

$$\frac{\partial}{\partial \boldsymbol{w}}||X\boldsymbol{w}-\boldsymbol{y}||^2 = \frac{\partial}{\partial \boldsymbol{w}}(X\boldsymbol{w}-\boldsymbol{y})^{\mathsf{T}}(X\boldsymbol{w}-\boldsymbol{y})$$
$$= 2(X\boldsymbol{w}-\boldsymbol{y})^{\mathsf{T}}X$$

$$(X\boldsymbol{w}^*-\boldsymbol{y})^{\mathsf{T}}X = 0$$
$$X^{\mathsf{T}}X\boldsymbol{w}^* = X^{\mathsf{T}}\boldsymbol{y}$$
$$\rightarrow \boldsymbol{w}^* = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{y}. \quad \blacksquare$$

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

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

We've only checked that $\boldsymbol{w}^* = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}$ is an extrema, but we have yet to check whether its a maxima or a minima. A second derivative test is needed:

$$\frac{\partial}{\partial \boldsymbol{w}}||X\boldsymbol{w} - \boldsymbol{y}||^2 = 2(X\boldsymbol{w} - \boldsymbol{y})^\mathsf{T}X$$

$$\frac{\partial^2}{\partial \boldsymbol{w}^\mathsf{T}\partial \boldsymbol{w}}||X\boldsymbol{w} - \boldsymbol{y}||^2 = \frac{\partial}{\partial \boldsymbol{w}^\mathsf{T}}2(X\boldsymbol{w} - \boldsymbol{y})^\mathsf{T}X = 2X^\mathsf{T}X$$

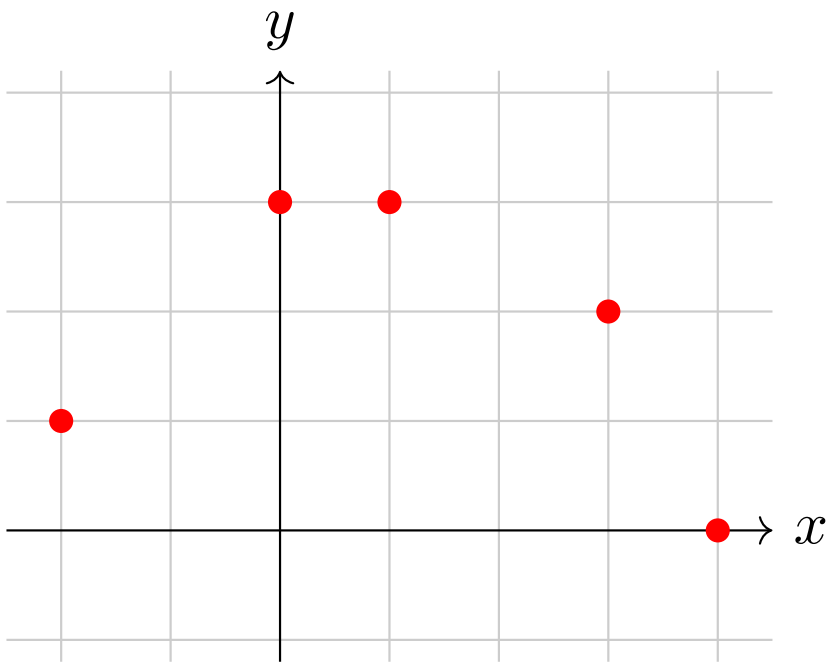
Hence we know that the second derivative is positive definite, i.e. for all $\mathrm{d}\boldsymbol{w} \neq 0$,

$$||X(\boldsymbol{w}^* + \mathrm{d}\boldsymbol{w}) - \boldsymbol{y}||^2 - ||X\boldsymbol{w}^* - \boldsymbol{y}||^2 = (\mathrm{d}\boldsymbol{w})^\mathsf{T}(2X^\mathsf{T}X)\mathrm{d}\boldsymbol{w} > 0,$$

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:

$$\hat{y} = a_0 + a_1 x + a_2 x^2$$

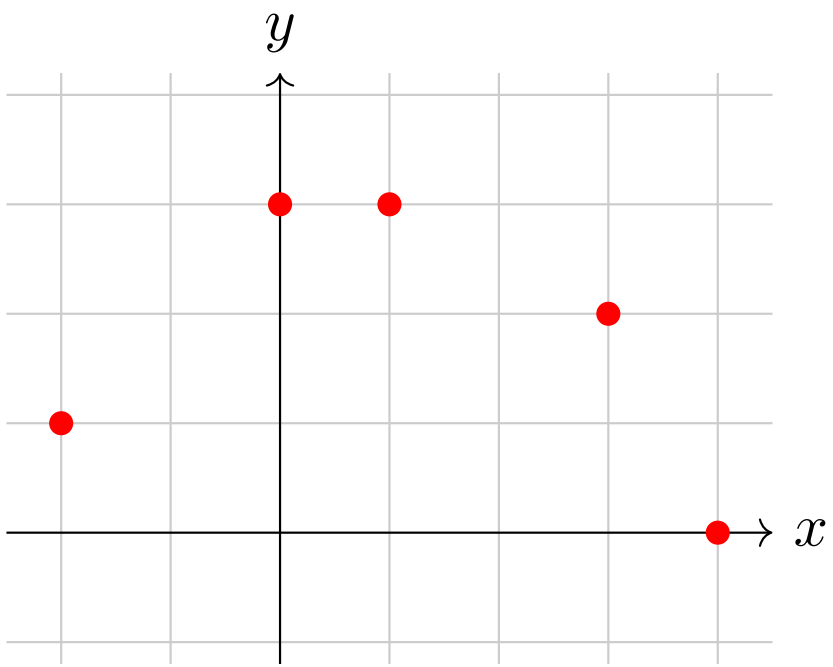
such that the mean square error is minimized.

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:

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

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



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

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

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

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

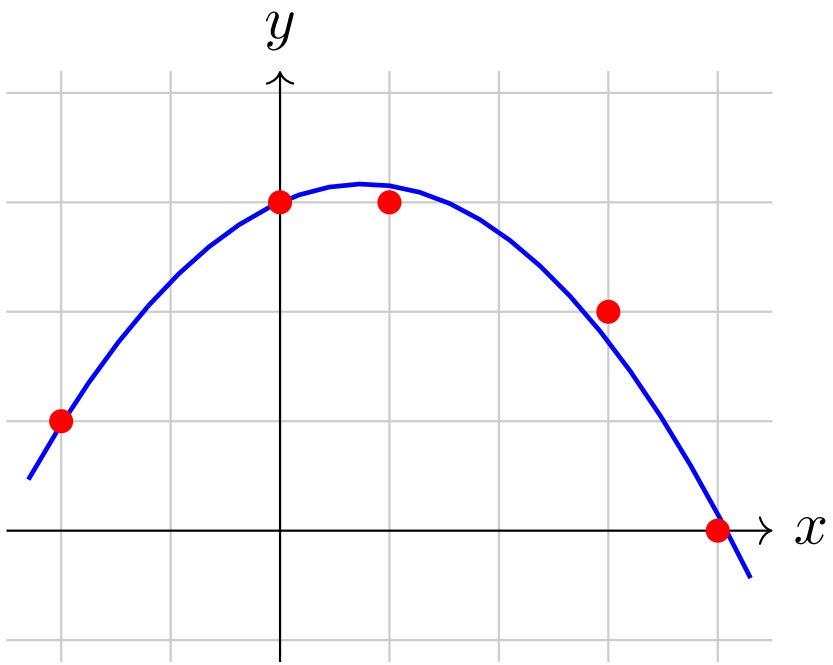


Table of Contents

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

$$\frac{\mathrm{d}A^{-1}}{\mathrm{d}t}$$

expressed using $\frac{\mathrm{d}A}{\mathrm{d}t}$?
[Sol.] Since we have

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

by product rule:

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

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

$$\Rightarrow \frac{\mathrm{d}A^{-1}}{\mathrm{d}t} = \frac{\partial A^{-1}}{\partial A}\left[\frac{\mathrm{d}A}{\mathrm{d}t}\right] = -A^{-1}\frac{\mathrm{d}A}{\mathrm{d}t}A^{-1} \quad \blacksquare$$

Lemma

The following identity holds for all square matrices A:

$$\det(\mathrm{e}^{A}) = \mathrm{e}^{\mathrm{Tr}\{A\}}. \tag{24}$$

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

e.g. $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$

Hence if the matrix A is expressible as an exponential:

$$A = \mathrm{e}^{B},$$

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

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

Def. (Eigenvalues and Eigenvectors)

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

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

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

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

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

For a time varying $A(t)$, what is $\frac{\mathrm{d}\lambda_i(t)}{\mathrm{d}t}$?
[Sol.]
Let us consider $\lambda_i(t)$ with its associated right and left eigenvectors: $\boldsymbol{v}_i(t)$ and $\boldsymbol{u}_i(t)$ that has length satisfying:

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

Then,

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

$$\frac{\mathrm{d}\lambda_i}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{u}_i^{\mathsf{T}}}{\mathrm{d}t}A\boldsymbol{v}_i + \boldsymbol{u}_i^{\mathsf{T}}\frac{\mathrm{d}A}{\mathrm{d}t}\boldsymbol{v}_i + \boldsymbol{u}_i^{\mathsf{T}}A\frac{\mathrm{d}\boldsymbol{v}_i}{\mathrm{d}t}$$

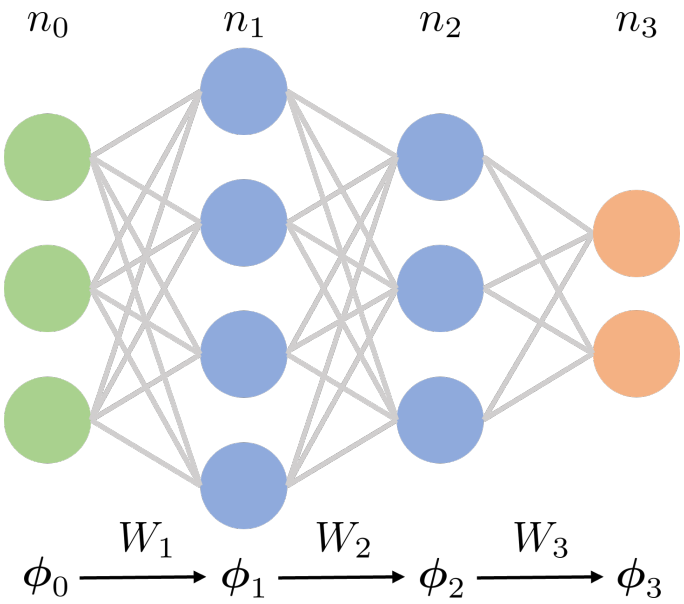
For a time varying $A(t)$, what is $\frac{\mathrm{d}\lambda_i(t)}{\mathrm{d}t}$?
[Sol.]

$$\frac{\mathrm{d}\lambda_i}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{u}_i^\mathsf{T}}{\mathrm{d}t} A\boldsymbol{v}_i + \boldsymbol{u}_i^\mathsf{T} \frac{\mathrm{d}A}{\mathrm{d}t}\boldsymbol{v}_i + \boldsymbol{u}_i^\mathsf{T} A \frac{\mathrm{d}\boldsymbol{v}_i}{\mathrm{d}t}$$

But we also have

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

$$\rightarrow \frac{\mathrm{d}\lambda_i}{\mathrm{d}t} = \boldsymbol{u}_i^\mathsf{T} \frac{\mathrm{d}A}{\mathrm{d}t}\boldsymbol{v}_i$$



$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_n \end{bmatrix} \rightarrow \boxed{\boldsymbol{w}_i \mid \sigma} \rightarrow \phi_i = \sigma(\boldsymbol{w}_i^{\mathsf{T}} \boldsymbol{\phi})$$

For layer-ℓ with n_ℓ perceptrons,

$$\mathbb{R}^{n_\ell} \ni \boldsymbol{\phi}_\ell = \sigma(W_\ell^{\mathsf{T}} \boldsymbol{\phi}_{\ell-1}),$$

the weight matrix

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

Hence if we denote $\sigma(W^{\mathsf{T}}\boldsymbol{\phi}) = f(W, \boldsymbol{\phi})$, the network above can be reduced to

$$\boldsymbol{\phi}_3 = f(W_3, \cdot) \circ f(W_2, \cdot) \circ f(W_1, \boldsymbol{\phi}_0).$$

Our goal is to do backpropagation: given a cost function $\mathcal{E}(\boldsymbol{\phi}_3)$, 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 $f(W, \boldsymbol{\phi})$.

- Differentiate w.r.t. W:

$$\frac{\partial f}{\partial W}(W_\ell, \boldsymbol{\phi}_{\ell-1})[H] = \begin{bmatrix} \frac{\partial}{\partial W}\sigma(W_{\ell,1}^\mathsf{T}\boldsymbol{\phi}_{\ell-1})[H] \\ \vdots \end{bmatrix} = \begin{bmatrix} \sigma'(W_\ell^\mathsf{T}\boldsymbol{\phi}_{\ell-1})H_1^\mathsf{T}\boldsymbol{\phi} \\ \vdots \end{bmatrix}$$
$$= \begin{bmatrix} \sigma'(W_{\ell,1}^\mathsf{T}\boldsymbol{\phi}_{\ell-1}) & 0 \\ 0 & \ddots \end{bmatrix} \begin{bmatrix} H_1^\mathsf{T}\boldsymbol{\phi} \\ \vdots \end{bmatrix} = \mathsf{diag}\left(\sigma'(W_\ell^\mathsf{T}\boldsymbol{\phi}_{\ell-1})\right) H^\mathsf{T}\boldsymbol{\phi}_{\ell-1}.$$

- Differentiate w.r.t. $\boldsymbol{\phi}$:

$$\frac{\partial f}{\partial \boldsymbol{\phi}}(W_\ell, \boldsymbol{\phi}_{\ell-1})[\boldsymbol{h}] = \mathsf{diag}\left(\sigma'(W_\ell^\mathsf{T}\boldsymbol{\phi}_{\ell-1})\right) W_\ell^\mathsf{T}\boldsymbol{h}.$$

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

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

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.

Notes: