Matrix Differentiation

Wen Perng

Electrical Engineering, NTU

2023/11/02

Derivatives are easy!

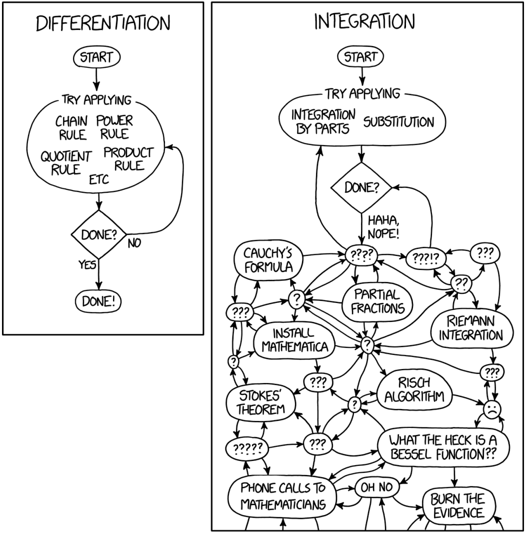


Figure: XKCD: differentiation and integration

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:

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

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 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**

$$\frac{\partial f}{\partial x_1} = 0, \frac{\partial f}{\partial x_2} = 0.$$

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**

$$\frac{\partial f}{\partial x_1} = 0, \frac{\partial f}{\partial x_2} = 0.$$

But if the function to minimize is a function of a **matrix** or a **vector**:

$$\min_{A \in \mathbb{R}^{m \times n}} f(A) \text{ or } \min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}),$$

What do we want?

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

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

We do the same test for finding extremas: by **partial derivatives**

$$\frac{\partial f}{\partial x_1} = 0, \frac{\partial f}{\partial x_2} = 0.$$

But if the function to minimize is a function of a **matrix** or a **vector**:

$$\min_{A \in \mathbb{R}^{m \times n}} f(A) \text{ or } \min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}),$$

how do we find the **derivative of a matrix function**?

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**

$$\frac{\partial f}{\partial x_1} = 0, \frac{\partial f}{\partial x_2} = 0.$$

But if the function to minimize is a function of a **matrix** or a **vector**:

$$\min_{A \in \mathbb{R}^{m \times n}} f(A) \text{ or } \min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}),$$

how do we find the **derivative of a matrix function**? I.e., how do we justify the notation of:

$$\frac{\mathrm{d} f(A)}{\mathrm{d} A} = 0 \text{ or } \frac{\mathrm{d} f(\boldsymbol{x})}{\mathrm{d} \boldsymbol{x}}?$$

A glimpse of what is to come

So... what kind of functions will we encounter? A prime example is 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:

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

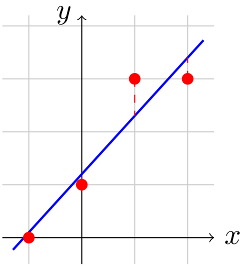
A glimpse of what is to come

So... what kind of functions will we encounter? A prime example is 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:

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



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

A glimpse of what is to come

Others might want to differentiate functions such as

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

A glimpse of what is to come

Others might want to differentiate functions such as

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

Def. (Transpose)

Flip the matrix along its diagonal:

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

A glimpse of what is to come

Others might want to differentiate functions such as

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

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?

A glimpse of what is to come

Others might want to differentiate functions such as

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

Def. (Transpose)

Flip the matrix along its diagonal:

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

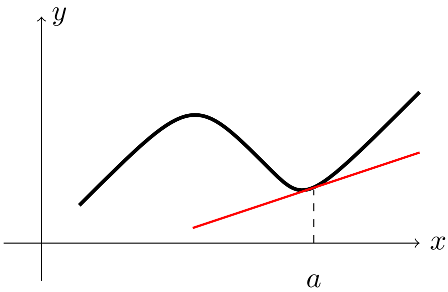
What number when transposed doubles?

$$午^{\mathsf{T}} = 什$$

Table of Contents

What is a derivative?

Derivatives are linearizations. Consider the easiest case below.



We know that the first order approximation will be

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

What is a derivative?

We can rewrite the previous equation into the form below:

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

What is a derivative?

We can rewrite the previous equation into the form below:

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

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

What is a derivative?

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:

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

What is a derivative?

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 works for matrices too.

Table of Contents

Examples

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.

Examples

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.

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

Examples

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:

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

Examples

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:

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

Examples

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:

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

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

Examples

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.

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

Examples

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:

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

Examples

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:

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

Examples

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:

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

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

Derivatives as Linear Operators

Perhaps we shouldn't stick to the matrix algebra,

Derivatives as Linear Operators

Perhaps we shouldn't stick to the matrix algebra, and simply view derivatives as a linear operator:

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

Derivatives as Linear Operators

Perhaps we shouldn't stick to the matrix algebra, and 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:

$$\mathcal{L}\{\mathrm{d}x\} = f'\mathrm{d}x$$

Derivatives as Linear Operators

Perhaps we shouldn't stick to the matrix algebra, and 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:

$$\mathcal{L}\{\mathrm{d}x\} = f'\mathrm{d}x$$
$$\mathcal{L}\{\mathrm{d}x\} = \mathrm{d}x f'$$

Derivatives as Linear Operators

Perhaps we shouldn't stick to the matrix algebra, and 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:

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

Derivatives as Linear Operators

Perhaps we shouldn't stick to the matrix algebra, and 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:

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

Derivatives as Linear Operators

But since all linear operators can be represented by matrices, we can, still, represent derivatives as matrices and vectors alike.

Derivatives as Linear Operators

But since all linear operators can be represented by matrices, we can, still, represent derivatives as matrices and vectors alike.

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

Derivatives as Linear Operators

But since all linear operators can be represented by matrices, we can, still, represent derivatives as matrices and vectors alike.

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

Let us see some examples of functions where the representation of

$$\mathcal{L}\{\cdot\} = A(\cdot)$$

holds, where A is matrix.

Differentiate by Vectors

(1) Differentiate scalar function f by vector $\boldsymbol{x}$:

Differentiate by Vectors

(1) Differentiate scalar function f by vector $\boldsymbol{x}$:
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:

$$\mathrm{d}f = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \cdots + \frac{\partial f}{\partial x_n}\mathrm{d}x_n. \tag{4}$$

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

Differentiate by Vectors

Notice that

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

Differentiate by 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 $\mathrm{d}\boldsymbol{x}$:

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

Where the gradient ∇f is defined as a column vector.

Differentiate by 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 $\mathrm{d}\boldsymbol{x}$:

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

Where the gradient ∇f is defined as a column vector. Compare with

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

Differentiate by Vectors

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

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

then

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

Differentiate by Vectors

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

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

then

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

But if we define

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

Differentiate by Vectors

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

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

then

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

But if we define

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

then

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

Differentiate by Vectors

Notice that:

$$1 \times n \rightarrow \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} \begin{array}{l} \leftarrow 1 \times 1 \\ \leftarrow n \times 1 \end{array}$$

$$n \times 1 \rightarrow \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}^{\mathsf{T}}} \begin{array}{l} \leftarrow 1 \times 1 \\ \leftarrow 1 \times n \end{array}$$

Differentiate by Vectors

Notice that:

$$1 \times n \rightarrow \frac{\mathrm{d}f \leftarrow 1 \times 1}{\mathrm{d}\boldsymbol{x} \leftarrow n \times 1}$$
$$n \times 1 \rightarrow \frac{\mathrm{d}f \leftarrow 1 \times 1}{\mathrm{d}\boldsymbol{x}^\mathsf{T} \leftarrow 1 \times n}$$

Perhaps we could define the derivative as:

$$\underbrace{\mathrm{d}f}_{m \times n} = \underbrace{\frac{\mathrm{d}f}{\mathrm{d}A}}_{m \times s} \underbrace{\mathrm{d}A}_{s \times n} \text{ or } \underbrace{\mathrm{d}B}_{m \times s} \underbrace{\frac{\mathrm{d}f}{\mathrm{d}B}}_{s \times n} .$$

Differentiate by Vectors

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

Differentiate by Vectors

(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:

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

Differentiate by Vectors

(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:

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial f_m}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix},$$

Differentiate by Vectors

(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:

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial f_m}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix},$$

Remark: the notations $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}}$ and $\frac{\partial f}{\partial \boldsymbol{x}}$ are used interchangeably.

Differentiate by Vectors

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

Differentiate by Vectors

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

1. Gradient (tranpose):

$$\nabla f := \frac{\partial f}{\partial \boldsymbol{x}^\mathsf{T}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \nabla^\mathsf{T} f := \frac{\partial f}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (5)$$

Differentiate by Vectors

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

1. Gradient (tranpose):

$$\nabla f := \frac{\partial f}{\partial \boldsymbol{x}^\mathsf{T}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \nabla^\mathsf{T} f := \frac{\partial f}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (5)$$

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

Differentiate by Vectors

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

3. Hessian:

$$\mathsf{H}(f) := \nabla\nabla^\mathsf{T} f = \frac{\partial^2 f}{\partial \boldsymbol{x}^\mathsf{T} \partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \quad (6)$$

Differentiate by Vectors

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

3. Hessian:

$$\mathsf{H}(f) := \nabla \nabla^\mathsf{T} f = \frac{\partial^2 f}{\partial \boldsymbol{x}^\mathsf{T} \partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \tag{6}$$

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

Differentiate by Vectors

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

Differentiate by Vectors

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:

$$J := \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_n}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}, \tag{7}$$

Differentiate by Vectors

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:

$$J := \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_n}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}, \tag{7}$$

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

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

Differentiate by Matrices

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

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

Some examples are as follows:

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

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

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

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

Differentiate by Matrices

For:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m\times n},$$

we have that if f is a scalar function:

$$\frac{\partial f}{\partial A} = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{21}} & \frac{\partial f}{\partial a_{31}} & \cdots & \frac{\partial f}{\partial a_{m1}} \\ \frac{\partial f}{\partial a_{12}} & \frac{\partial f}{\partial a_{22}} & \frac{\partial f}{\partial a_{32}} & \cdots & \frac{\partial f}{\partial a_{m2}} \\ \vdots & \vdots & \ddots & \vdots & \\ \frac{\partial f}{\partial a_{1n}} & \frac{\partial f}{\partial a_{2n}} & \frac{\partial f}{\partial a_{3n}} & \cdots & \frac{\partial f}{\partial a_{mn}} \end{bmatrix}_{n\times m} \tag{8}$$

Time Derivatives

Time derivatives seem like the easiest, just differentiate term by term:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{d}v_1}{\mathrm{d}t} \\ \vdots \\ \frac{\mathrm{d}v_n}{\mathrm{d}t} \end{bmatrix}.$$

Time Derivatives

Time derivatives seem like the easiest, just differentiate term by term:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{d}v_1}{\mathrm{d}t} \\ \vdots \\ \frac{\mathrm{d}v_n}{\mathrm{d}t} \end{bmatrix}.$$

By the chain rule, we also have that

$$\frac{\mathrm{d}}{\mathrm{d}t} f(A(t)) = \frac{\partial f(A)}{\partial A}\frac{\mathrm{d}A}{\mathrm{d}t}.$$

And all the other chain rule, multiplication rule and etc. are satisfied.

Note that, again, the order of multiplication matters. And if the matrix derivative is of other form, the time derivative follow suits.

Time Derivatives

Time derivatives seem like the easiest, just differentiate term by term:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{d}v_1}{\mathrm{d}t} \\ \vdots \\ \frac{\mathrm{d}v_n}{\mathrm{d}t} \end{bmatrix}.$$

By the chain rule, we also have that

$$\frac{\mathrm{d}}{\mathrm{d}t} f(A(t)) = \frac{\partial f(A)}{\partial A}\frac{\mathrm{d}A}{\mathrm{d}t}.$$

E.g.
$f(\boldsymbol{x}) = x_1^2 x_2$
$\boldsymbol{x}(t) = [x_1, x_2]^{\mathsf{T}} = [t, t^2]^{\mathsf{T}}$

(Recap)

(Recap)

1. We should view derivatives of $f(x)$ as a **linear approximation** of $f(x + \mathrm{d}x)$.

(Recap)

1. We should view derivatives of $f(x)$ as a **linear approximation** of $f(x + \mathrm{d}x)$.
2. Hence, we can represent derivatives as a **linear operator** $\mathcal{L}\{\cdot\}$ satisfying:
$$\mathrm{d}f(x) = \mathcal{L}\{\mathrm{d}x\}.$$

(Recap)

1. We should view derivatives of $f(x)$ as a **linear approximation** of $f(x + \mathrm{d}x)$.
2. Hence, we can represent derivatives as a **linear operator** $\mathcal{L}\{\cdot\}$ satisfying:
$$\mathrm{d}f(x) = \mathcal{L}\{\mathrm{d}x\}.$$
3. In working with scalars, vectors or matrices, we can represent derivatives as
$$\mathrm{d}f(x) = \frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x,$$
as long as their dimension matches.

Before we take a break, a short remark is needed.

Before we take a break, a short remark is needed.

Notation

Two conventions exist in when differentiating with matrix/vector: for $\boldsymbol{y}$ of size $m \times 1$ and $\boldsymbol{x}$ of size $n \times 1$,

1. Numerator layout:

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \text{ follows the size of } \boldsymbol{y} \times \boldsymbol{x}^{\mathsf{T}}, \text{ i.e. } m \times n.$$

Before we take a break, a short remark is needed.

Notation

Two conventions exist in when differentiating with matrix/vector: for $\boldsymbol{y}$ of size $m \times 1$ and $\boldsymbol{x}$ of size $n \times 1$,

1. Numerator layout:
$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \text{ follows the size of } \boldsymbol{y} \times \boldsymbol{x}^{\mathsf{T}}, \text{ i.e. } m \times n.$$

2. Denominator layout:
$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \text{ follows the size of } \boldsymbol{x} \times \boldsymbol{y}^{\mathsf{T}}, \text{ i.e. } n \times m.$$

Table of Contents

Vector Functions

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

Vector Functions

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

$$\mathrm{d}f = (\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x} = 1\mathrm{d}x$$

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} = 1$$

Vector Functions

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

$$\mathrm{d}f = (\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x} = 1\mathrm{d}x$$

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} = 1$$

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

Vector Functions

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

$$\mathrm{d}f = (\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x} = 1\mathrm{d}x$$

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} = 1$$

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

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

Vector Functions

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

Vector Functions

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

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

Vector Functions

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

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

Vector Functions

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

$$\begin{aligned} \mathrm{d}f(\boldsymbol{x}) = \mathrm{d}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}) &= (\boldsymbol{x} + \mathrm{d}\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \\ &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\mathrm{d}\boldsymbol{x} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \end{aligned}$$

Vector Functions

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

$$\begin{aligned} \mathrm{d}f(\boldsymbol{x}) = \mathrm{d}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}) &= (\boldsymbol{x} + \mathrm{d}\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \\ &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\mathrm{d}\boldsymbol{x} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \\ &= (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\mathrm{d}\boldsymbol{x} \end{aligned}$$

Vector Functions

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

$$\begin{aligned} \mathrm{d}f(\boldsymbol{x}) = \mathrm{d}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}) &= (\boldsymbol{x} + \mathrm{d}\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \\ &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\mathrm{d}\boldsymbol{x} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \\ &= (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\mathrm{d}\boldsymbol{x} \end{aligned}$$

Ignoring the terms of $O(\mathrm{d}\boldsymbol{x}^2)$ and notice that the transpose of a scalar is still itself ($(\mathrm{d}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{x}^{\mathsf{T}}\mathrm{d}\boldsymbol{x}$), we hence have:

$$\mathrm{d}f = 2\boldsymbol{x}^{\mathsf{T}}\mathrm{d}\boldsymbol{x} = \mathcal{L}\{\mathrm{d}\boldsymbol{x}\}. \quad \blacksquare$$

Quadratic Form

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

Quadratic Form

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

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

Quadratic Form

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

$$\begin{aligned}
\mathrm{d}f &= (\boldsymbol{x} + \mathrm{d}\boldsymbol{x})^\mathsf{T} A(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x}^\mathsf{T} A\boldsymbol{x} \\
&= \boldsymbol{x}^\mathsf{T} A\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A\boldsymbol{x} + \boldsymbol{x}^\mathsf{T} A\mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A\mathrm{d}\boldsymbol{x} - \boldsymbol{x}^\mathsf{T} A\boldsymbol{x}
\end{aligned}$$

Quadratic Form

Find the derivative with respect to x of the function

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

[Sol.]

$$\begin{aligned}
\mathrm{d}f &= (\boldsymbol{x} + \mathrm{d}\boldsymbol{x})^\mathsf{T} A (\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \\
&= \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \boldsymbol{x} + \boldsymbol{x}^\mathsf{T} A \mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \mathrm{d}\boldsymbol{x} - \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \\
&= (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \boldsymbol{x} + \boldsymbol{x}^\mathsf{T} A \mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \mathrm{d}\boldsymbol{x}
\end{aligned}$$

Quadratic Form

Find the derivative with respect to $\boldsymbol{x}$ of the function

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

[Sol.]

$$\begin{aligned}
\mathrm{d}f &= (\boldsymbol{x} + \mathrm{d}\boldsymbol{x})^\mathsf{T} A (\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) - \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \\
&= \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \boldsymbol{x} + \boldsymbol{x}^\mathsf{T} A \mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \mathrm{d}\boldsymbol{x} - \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \\
&= (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \boldsymbol{x} + \boldsymbol{x}^\mathsf{T} A \mathrm{d}\boldsymbol{x} + (\mathrm{d}\boldsymbol{x})^\mathsf{T} A \mathrm{d}\boldsymbol{x}
\end{aligned}$$

Ignoring the terms of $O(\mathrm{d}\boldsymbol{x}^2)$, we hence have:

$$\mathrm{d}f = \boldsymbol{x}^\mathsf{T}(A + A^\mathsf{T})\mathrm{d}\boldsymbol{x} = \mathcal{L}\{\mathrm{d}\boldsymbol{x}\}. \qquad \blacksquare$$

Inverse

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

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

[Sol.]

Inverse

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

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

[Sol.] Since we have

$$A(t)A^{-1}(t) = 1,$$

Inverse

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

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

[Sol.] Since we have

$$A(t)A^{-1}(t) = 1,$$

by chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(AA^{-1}\right) = \frac{\mathrm{d}A}{\mathrm{d}t}A^{-1} + A\frac{\mathrm{d}A^{-1}}{\mathrm{d}t} = 0$$

Inverse

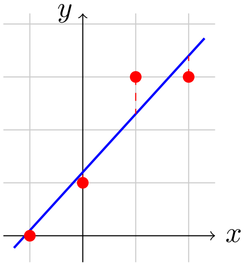
Given $A = A(t)$, what is
$$\frac{\mathrm{d}A^{-1}}{\mathrm{d}t}?$$

[Sol.] Since we have
$$A(t)A^{-1}(t) = 1,$$

by chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(AA^{-1}\right) = \frac{\mathrm{d}A}{\mathrm{d}t}A^{-1} + A\frac{\mathrm{d}A^{-1}}{\mathrm{d}t} = 0$$
$$\frac{\mathrm{d}A^{-1}}{\mathrm{d}t} = -A^{-1}\frac{\mathrm{d}A}{\mathrm{d}t}A^{-1}. \quad \blacksquare$$

Least Mean Square



Suppose the regression line for the data points is

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

we can record the relationship as follows

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

Least Mean Square

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:

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

Least Mean Square

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 derivative is zero at that point, i.e.

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

Least Mean Square

[Sol.]

Least Mean Square

[Sol.]

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

Least Mean Square

[Sol.]

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

Least Mean Square

[Sol.]

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

$$(X\boldsymbol{w}^*-\boldsymbol{y})^\mathsf{T}X = 0$$

Least Mean Square

[Sol.]

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

$$\begin{aligned}(X\boldsymbol{w}^*-\boldsymbol{y})^\top X &= 0 \\ X^\top X\boldsymbol{w}^* &= X^\top \boldsymbol{y}\end{aligned}$$

Least Mean Square

[Sol.]

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

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

Least Mean Square

[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.

Least Mean Square

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

Least Mean Square

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:

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

Least Mean Square

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:

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

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

Least Mean Square

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.

Definiteness

Def. (Definiteness)

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

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

Moreover, we have:

$$\begin{aligned}
&(\text{positive semi-definite}) \qquad && \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \geq 0 \iff A \succeq 0 \\
&(\text{negative definite}) \qquad && \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} < 0 \iff A \prec 0 \\
&(\text{negative semi-definite}) \qquad && \boldsymbol{x}^\mathsf{T} A \boldsymbol{x} \leq 0 \iff A \preceq 0
\end{aligned}$$

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

Taylor Expansion

From the derivation of LLMSE solution above, we can find a second order approximation of a scalar-valued function $f(\boldsymbol{x})$ of vector $\boldsymbol{x}$ by:

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \frac{\partial f(\boldsymbol{a})}{\partial \boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2!}(\boldsymbol{x} - \boldsymbol{a})^{\mathsf{T}} \frac{\partial^2 f(\boldsymbol{a})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathsf{T}}}(\boldsymbol{x} - \boldsymbol{a}) + \cdots , \tag{9}$$

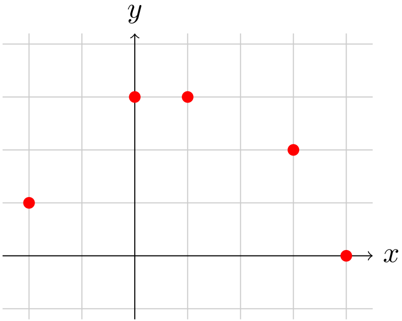
or as

$$f(\boldsymbol{a} + \mathrm{d}\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla^{\mathsf{T}} f(\boldsymbol{a}) \mathrm{d}\boldsymbol{x} + \frac{1}{2}(\mathrm{d}\boldsymbol{x})^{\mathsf{T}} \mathsf{H}(f(\boldsymbol{a})) \mathrm{d}\boldsymbol{x}. \tag{10}$$

Regression

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.

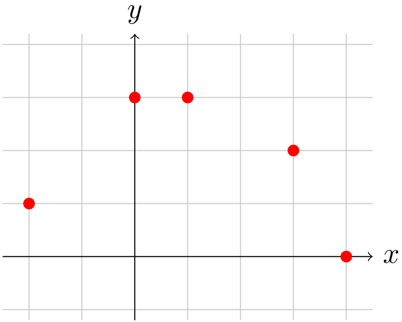
Regression

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:

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

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



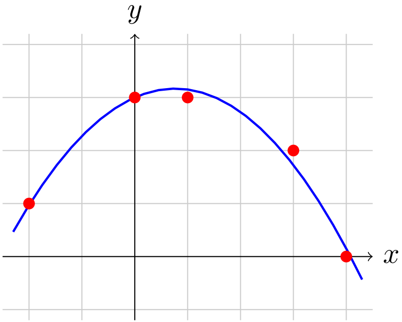
Regression

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

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

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

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



HTML

HTML, HW3

Q6. Let the cross-entropy error function for $E_{\text{in}}(\boldsymbol{w})$ be:

$$E_{\text{in}}(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} \ln\left(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\right).$$

Find the Hessian of the function. Express it in diagonalized form of $\nabla^{\mathsf{T}} \nabla E_{\text{in}} = XDX^{\mathsf{T}}$.

HTML

$$E_{\mathsf{in}}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\right)$$

$$\nabla E_{\mathsf{in}}(\boldsymbol{w}) = \frac{\partial E_{\mathsf{in}}}{\partial \boldsymbol{w}^{\mathsf{T}}} = \frac{\partial}{\partial \boldsymbol{w}^{\mathsf{T}}}\left(\frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\right)\right)$$

HTML

$$E_{\mathsf{in}}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\right)$$

$$\begin{aligned}
\nabla E_{\mathsf{in}}(\boldsymbol{w}) = \frac{\partial E_{\mathsf{in}}}{\partial \boldsymbol{w}^{\mathsf{T}}} &= \frac{\partial}{\partial \boldsymbol{w}^{\mathsf{T}}}\left(\frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\right)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)} \frac{\partial}{\partial \boldsymbol{w}^{\mathsf{T}}}\left(\exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\right)
\end{aligned}$$

HTML

$$E_{\mathsf{in}}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln\Big(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\Big)$$

$$\begin{aligned}
\nabla E_{\mathsf{in}}(\boldsymbol{w}) = \frac{\partial E_{\mathsf{in}}}{\partial \boldsymbol{w}^{\mathsf{T}}} &= \frac{\partial}{\partial \boldsymbol{w}^{\mathsf{T}}} \left(\frac{1}{N}\sum_{n=1}^{N} \ln\Big(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\Big) \right) \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)} \frac{\partial}{\partial \boldsymbol{w}^{\mathsf{T}}} \Big(\exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)\Big) \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{-y_n \boldsymbol{x}_n \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)}{1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)}
\end{aligned}$$

HTML

$$\frac{\partial^2 E_{\mathsf{in}}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^{\mathsf{T}}} = \frac{\partial}{\partial \boldsymbol{w}} \left(\frac{1}{N} \sum_{n=1}^{N} \frac{-y_n \boldsymbol{x}_n \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)}{1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)} \right)$$

HTML

$$\frac{\partial^2 E_{\text{in}}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^\mathsf{T}} = \frac{\partial}{\partial \boldsymbol{w}} \left(\frac{1}{N} \sum_{n=1}^{N} \frac{-y_n \boldsymbol{x}_n \exp(-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n)}{1 + \exp(-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n)} \right)$$

$$= \sum_{n=1}^{N} \frac{-y_n \boldsymbol{x}_n}{N} \frac{-y_n {\boldsymbol{x}_n}^\mathsf{T} \mathrm{e}^{-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n} (1 + \mathrm{e}^{-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n}) + y_n {\boldsymbol{x}_n}^\mathsf{T} (\mathrm{e}^{-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n})^2}{(1 + \mathrm{e}^{-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n})^2}$$

HTML

$$
\begin{aligned}
\frac{\partial^2 E_{\text{in}}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^{\mathsf{T}}} &= \frac{\partial}{\partial \boldsymbol{w}} \left(\frac{1}{N} \sum_{n=1}^{N} \frac{-y_n \boldsymbol{x}_n \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)}{1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)} \right) \\
&= \sum_{n=1}^{N} \frac{-y_n \color{red}{\boldsymbol{x}_n}}{N} \frac{-y_n \color{red}{\boldsymbol{x}_n}^{\mathsf{T}} \color{black} \mathrm{e}^{-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n} (1 + \mathrm{e}^{-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n}) + y_n \color{red}{\boldsymbol{x}_n}^{\mathsf{T}} \color{black} (\mathrm{e}^{-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n})^2}{(1 + \mathrm{e}^{-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n})^2} \\
&= \sum_{i=1}^{N} \left(\frac{y_n^2 \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)}{N(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n))^2} \right) \color{red}{\boldsymbol{x}_n \boldsymbol{x}_n^{\mathsf{T}}} \color{black}.
\end{aligned}
$$

HTML

$$\frac{\partial^2 E_{\mathsf{in}}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^{\mathsf{T}}} = \sum_{i=1}^{N} \left(\frac{y_n^2 \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)}{N(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n))^2} \right) \boldsymbol{x}_n \boldsymbol{x}_n^{\mathsf{T}}$$

HTML

$$\frac{\partial^2 E_{\text{in}}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^{\mathsf{T}}} = \sum_{i=1}^{N} \left(\frac{y_n^2 \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n)}{N(1 + \exp(-y_n \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_n))^2} \right) \boldsymbol{x}_n \boldsymbol{x}_n{}^{\mathsf{T}}$$

Diagonalized (spectral decoposition):

$$E_{\text{in}} = XDX^{\mathsf{T}} = \sum_{n=1}^{N} \lambda_n \boldsymbol{x}_n \boldsymbol{x}_n{}^{\mathsf{T}}. \tag{11}$$

HTML

$$\frac{\partial^2 E_{\text{in}}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^\mathsf{T}} = \sum_{i=1}^{N} \left(\frac{y_n^2 \exp(-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n)}{N(1+\exp(-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n))^2} \right) \boldsymbol{x}_n \boldsymbol{x}_n{}^\mathsf{T}$$

Diagonalized (spectral decoposition):

$$E_{\text{in}} = XDX^\mathsf{T} = \sum_{n=1}^{N} \lambda_n \boldsymbol{x}_n \boldsymbol{x}_n{}^\mathsf{T}. \tag{11}$$

This is the Hessian in its diagonalized form, with $\boldsymbol{x}_n$ being the eigenvectors associated with the eigenvalues of

$$\frac{y_n^2 \exp(-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n)}{N(1+\exp(-y_n \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n))^2}.$$

Table of Contents

Matrix Derivatives

We'll go over the derivative of objects including the trace, the determinant, eigenvalues, and singular values.

Trace

Def. (Trace)

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

Trace

Def. (Trace)

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

$$\operatorname{tr}\left(\begin{bmatrix} a_{11} & * & \cdots & * \\ * & a_{22} & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & a_{nn} \end{bmatrix}\right) = \sum_{i=1}^{n} a_{ii}. \tag{12}$$

Trace

Def. (Trace)

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

$$\operatorname{tr}\left(\begin{bmatrix} a_{11} & * & \cdots & * \\ * & a_{22} & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & a_{nn} \end{bmatrix}\right) = \sum_{i=1}^{n} a_{ii}. \tag{12}$$

It can easily be checked that the derivative operator and trace operator are commutative:

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}(A(t)) = \operatorname{tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}A(t)\right).$$

Determinant

Lemma

The following identity holds for all square matrices A:

$$\det(\mathrm{e}^{A}) = \mathrm{e}^{\mathsf{tr}(A)}. \tag{13}$$

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

Determinant

Lemma

The following identity holds for all square matrices A:

$$\det(\mathrm{e}^{A}) = \mathrm{e}^{\mathsf{tr}(A)}. \tag{13}$$

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

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

Determinant

Hence if the matrix $A(t)$ is expressible as an exponential:

$$A(t) = \mathrm{e}^{B(t)},$$

then a (pseudo-)proof is as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t} \det(A(t)) =$$

Determinant

Hence if the matrix $A(t)$ is expressible as an exponential:

$$A(t) = \mathrm{e}^{B(t)},$$

then a (pseudo-)proof is as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\det(A(t)) = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{\mathsf{tr}B(t)}$$

Determinant

Hence if the matrix $A(t)$ is expressible as an exponential:

$$A(t) = \mathrm{e}^{B(t)},$$

then a (pseudo-)proof is as follows:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\det(A(t)) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{\mathsf{tr}B(t)} \\ &= \mathrm{e}^{\mathsf{tr}B(t)}\mathsf{tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}B(t)\right)\end{aligned}$$

Determinant

Hence if the matrix $A(t)$ is expressible as an exponential:

$$A(t) = \mathrm{e}^{B(t)},$$

then a (pseudo-)proof is as follows:

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

Eigenvalues

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**.

Eigenvalues

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.

Eigenvalues

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.

Written in the language of matrices:

$$AV = A \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} = V\Lambda$$

Eigenvalues

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:

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

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

Written in the language of matrices:

$$U^\mathsf{T} A = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_n \end{bmatrix}^\mathsf{T} A = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_n \end{bmatrix}^\mathsf{T} = \Lambda U^\mathsf{T}$$

Eigenvalues

Some theorems of linear algebra:

Eigenvalues

Some theorems of linear algebra: since we have $AV = V\Lambda$,

Thm. (Diagonalization)

If V is full rank, i.e., V^{-1} exists, then the matrix A can be diagonalized via:

$$\Lambda = V^{-1}AV.$$

Eigenvalues

Some theorems of linear algebra: since we have $AV = V\Lambda$,

Thm. (Diagonalization)

If V is full rank, i.e., V^{-1} exists, then the matrix A can be diagonalized via:

$$\Lambda = V^{-1}AV.$$

By expanding out $A = V\Lambda V^{-1}$ and setting $V = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$, $(V^{-1})^{\mathsf{T}} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n]$,

Thm. (Spectral Decomposition)

The $n \times n$ matrix A can be decomposed by

$$A = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \boldsymbol{u}_i^{\mathsf{T}}.$$

Eigenvalues

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

Eigenvalues

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:

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

Eigenvalues

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,

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

Eigenvalues

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,

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

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

Eigenvalues

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

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

But we also have

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

Eigenvalues

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

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

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

Singular Values

Def. (Singular Value Decomposition, SVD)

For a real (complex) matrix A of size $m \times n$ (WLOG let $m > n$), then it can be decomposed into a sandwich product of diagonal matrix Σ by two orthogonal (unitary) matrices U and V:

$$A = \underbrace{U}_{m\times m}\underbrace{\Sigma}_{m\times n}\underbrace{V}_{n\times n}^{\mathsf{T}} = \begin{bmatrix} | & & | \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_m \\ | & & | \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \begin{bmatrix} \boldsymbol{v}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{v}_n^{\mathsf{T}} \end{bmatrix} \tag{14}$$

where $U^{\mathsf{T}}U = 1$, $V^{\mathsf{T}}V = 1$, and $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$.
Note that for the case of complex A, the transposition are replaced by conjugate-transpose.

Singular Values

Thm. (Spectral Decomposition)

The result of singular value decomposition can also be written as:

$$A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}.$$

Singular Values

Thm. (Spectral Decomposition)

The result of singular value decomposition can also be written as:

$$A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}.$$

The same derivation can be applied for finding the time derivative of singular values.

Singular Values

Thm. (Spectral Decomposition)

The result of singular value decomposition can also be written as:

$$A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}.$$

The same derivation can be applied for finding the time derivative of singular values. By utilizing:

$$AV = U\Sigma V^\mathsf{T} V = U\Sigma, \qquad U^\mathsf{T} A = U^\mathsf{T} U \Sigma V^\mathsf{T} = \Sigma V^\mathsf{T}$$

Singular Values

Thm. (Spectral Decomposition)

The result of singular value decomposition can also be written as:

$$A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}.$$

The same derivation can be applied for finding the time derivative of singular values. By utilizing:

$$AV = U\Sigma V^\mathsf{T} V = U\Sigma, \qquad U^\mathsf{T} A = U^\mathsf{T} U \Sigma V^\mathsf{T} = \Sigma V^\mathsf{T}$$

the solution is given by

$$\frac{\mathrm{d}\sigma_i}{\mathrm{d}t} = \boldsymbol{u}_i^\mathsf{T} \frac{\mathrm{d}A}{\mathrm{d}t} \boldsymbol{v}_i, \tag{15}$$

where $\boldsymbol{u}_i(t)$ and $\boldsymbol{v}_i(t)$ are the left and right singular vector associated with the singular value $\sigma_i(t)$.

That’s all folks.