SI231 Matrix Analysis and Computations
Matrix Calculus and Derivatives

Zepeng Zhang and Ziping Zhao

Spring Term 2022–2023

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

http://si231.sist.shanghaitech.edu.cn

Outline

- Basics of Matrix Calculus and Derivatives
- Examples
- Complex Derivatives

Matrix Calculus and Derivatives

- Matrix calculus is a specialized notation for doing multivariable calculus.

- For multivariable calculus, we have

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = \left(\frac{\partial f}{\partial \mathbf{x}} \right)^T d\mathbf{x} = \nabla f(\mathbf{x})^T d\mathbf{x},$$

  where $f(\mathbf{x})$ is a scalar function of vector $\mathbf{x} \in \mathbb{R}^n$.

- For matrix calculus, we have

$$df = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial f}{\partial x_{ij}} dx_{ij} = \mathrm{tr}\left(\left(\frac{\partial f}{\partial \mathbf{X}} \right)^T d\mathbf{X} \right) = \mathrm{tr}\left(\nabla f(\mathbf{X})^T d\mathbf{X} \right),$$

  where $f(\mathbf{X})$ is a scalar function of matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$.

Second-Oder Gradient of Functions of Matrices

- Let $f(\mathbf{X}) : \mathbb{R}^{m\times n} \to \mathbb{R}$. The second-order gradient of f can be expressed as

$$\nabla^2 f(X) \triangleq \begin{bmatrix} \nabla\frac{\partial f(X)}{\partial X_{11}} & \nabla\frac{\partial f(X)}{\partial X_{12}} & \cdots & \nabla\frac{\partial f(X)}{\partial X_{1n}} \\ \nabla\frac{\partial f(X)}{\partial X_{21}} & \nabla\frac{\partial f(X)}{\partial X_{22}} & \cdots & \nabla\frac{\partial f(X)}{\partial X_{2n}} \\ \vdots & \vdots & & \vdots \\ \nabla\frac{\partial f(X)}{\partial X_{m1}} & \nabla\frac{\partial f(X)}{\partial X_{m2}} & \cdots & \nabla\frac{\partial f(X)}{\partial X_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n\times m\times n}.$$

Basic Rules for Matrix Calculus

Consider two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$, we introduce some basic rules for matrix calculus in the following.

- Addition and subtraction: $d\left(\mathbf{X} \pm \mathbf{Y}\right) = d\mathbf{X} \pm d\mathbf{Y}$;
- Multiplication: $d\left(\mathbf{X}\mathbf{Y}\right) = \left(d\mathbf{X}\right)\mathbf{Y} + \mathbf{X}d\mathbf{Y}$;
- Transpose: $d\left(\mathbf{X}^T\right) = \left(d\mathbf{X}\right)^T$;
- Trace: $d\mathrm{tr}\left(\mathbf{X}\right) = \mathrm{tr}\left(d\mathbf{X}\right)$;
- Element-wise multiplication: $d\left(\mathbf{X} \odot \mathbf{Y}\right) = d\mathbf{X} \odot \mathbf{Y} + \mathbf{X} \odot d\mathbf{Y}$;
- Element-wise function: $d\sigma\left(\mathbf{X}\right) = \sigma'\left(\mathbf{X}\right) \odot d\mathbf{X}$, where $\sigma\left(\mathbf{X}\right) = \left[\sigma\left(x_{ij}\right)\right]$ and $\sigma'\left(\mathbf{X}\right) = \left[\sigma'\left(x_{ij}\right)\right]$ are element-wise functions.

Basic Rules for Matrix Calculus

- Determinant: $d\left|\mathbf{X}\right| = \mathrm{tr}\left(\mathbf{X}^{\sharp} d\mathbf{X}\right) = \left|\mathbf{X}\right| \mathrm{tr}\left(\mathbf{X}^{-1} d\mathbf{X}\right)$. [1]

- Inverse: $d\mathbf{X}^{-1} = -\mathbf{X}^{-1}\left(d\mathbf{X}\right)\mathbf{X}^{-1}$;

Proof. For $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}$, we have $d\mathbf{X}\mathbf{X}^{-1} = d\mathbf{I} = \mathbf{0}$, hence

$$d\mathbf{X}\mathbf{X}^{-1} = \left(d\mathbf{X}\right)\mathbf{X}^{-1} + \mathbf{X} d\mathbf{X}^{-1} = 0,$$

which leads to $d\mathbf{X}^{-1} = -\mathbf{X}^{-1}\left(d\mathbf{X}\right)\mathbf{X}^{-1}$. $\square$

[1] $\mathbf{X}^{\sharp}$ represents the adjugate matrix.

Some Properties of Trace

- For scalar x, we have $x = \mathrm{tr}(x)$.
- For matrix $\mathbf{X}$, we have $\mathrm{tr}(\mathbf{X}) = \mathrm{tr}(\mathbf{X}^T)$.
- For matrix $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$, we have $\mathrm{tr}(\mathbf{X} \pm \mathbf{Y}) = \mathrm{tr}(\mathbf{X}) \pm \mathrm{tr}(\mathbf{Y})$.
- For matrix $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{m \times n}$, we have $\mathrm{tr}\left(\mathbf{A}^T (\mathbf{B} \odot \mathbf{C})\right) = \mathrm{tr}\left((\mathbf{A} \odot \mathbf{B})^T \mathbf{C}\right)$.

With the above rules and properties, for a scalar function of matrix, we can calculate its differential and rewrite it in the form of

$$df = \mathrm{tr}\left(\frac{\partial f}{\partial \mathbf{X}}^T d\mathbf{X}\right),$$

through which we can get the derivatives by calculating differentials.

Chain Rule

- If $f : \mathbb{R}^d \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}$, then the derivative of $h(\mathbf{x}) = g(f(\mathbf{x}))$ is

$$\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)^T \frac{\partial g(f(\mathbf{x}))}{\partial f(\mathbf{x})}$$

- Let $\mathbf{U} = f(\mathbf{X})$ and the derivative of the function $g(\mathbf{U})$ with respect to $\mathbf{X}$ is

$$\frac{\partial g(\mathbf{U})}{\partial \mathbf{X}} = \frac{\partial g(f(\mathbf{X}))}{\partial \mathbf{X}}.$$

Then chain rule can be applied as follows:

$$\frac{\partial g(\mathbf{U})}{\partial x_{ij}} = \sum_{k=1}^{M} \sum_{l=1}^{N} \frac{\partial g(\mathbf{U})}{\partial u_{kl}} \frac{\partial u_{kl}}{\partial x_{ij}} = \mathrm{tr}\left((\frac{\partial g(\mathbf{U})}{\partial \mathbf{U}})^T \frac{\partial \mathbf{U}}{\partial x_{ij}} \right),$$

where M and N are the dimensions of rows and columns of $\mathbf{U}$.

Example 1

Calculate the derivative of $f(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{b}$, where $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{X} \in \mathbb{R}^{m\times n}$, and $\mathbf{b} \in \mathbb{R}^n$.

- First calculate the differential of f as

$$df = d(\mathbf{a}^T\mathbf{X}\mathbf{b}) = \mathbf{a}^T d(\mathbf{X}\mathbf{b}) = \mathbf{a}^T d(\mathbf{X})\mathbf{b}.$$

- Then rewrite df as follows:

$$df = \mathrm{tr}(df) = \mathrm{tr}\left(\mathbf{a}^T d(\mathbf{X})\mathbf{b}\right) = \mathrm{tr}\left(\mathbf{b}\mathbf{a}^T d(\mathbf{X})\right) = \mathrm{tr}\left((\mathbf{a}\mathbf{b}^T)^T d(\mathbf{X})\right).$$

- Observe that $df = \mathrm{tr}\left(\left(\frac{\partial f}{\partial \mathbf{X}}\right)^T d\mathbf{X}\right) = \mathrm{tr}\left((\mathbf{a}\mathbf{b}^T)^T d(\mathbf{X})\right)$, we can conclude that $\frac{\partial f}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$.

Example 2

Calculate the derivative of $f(\mathbf{X}) = \mathbf{a}^T e^{\mathbf{X}\mathbf{b}}$, where $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^n$, and $e^{\mathbf{X}\mathbf{b}}$ is applied element-wise.

- First calculate the differential of f as

$$df = d(\mathbf{a}^T e^{\mathbf{X}\mathbf{b}}) = \mathbf{a}^T d(e^{\mathbf{X}\mathbf{b}}) = \mathbf{a}^T \big(e^{\mathbf{X}\mathbf{b}} \odot d(\mathbf{X}\mathbf{b})\big) = \mathbf{a}^T \big(e^{\mathbf{X}\mathbf{b}} \odot d(\mathbf{X})\mathbf{b}\big).$$

- Then rewrite df as follows:

$$\begin{aligned} df &= \mathrm{tr}(df) = \mathrm{tr}\left(\mathbf{a}^T \big(e^{\mathbf{X}\mathbf{b}} \odot d(\mathbf{X})\mathbf{b}\big)\right) = \mathrm{tr}\left((\mathbf{a} \odot e^{\mathbf{X}\mathbf{b}})^T d(\mathbf{X})\mathbf{b}\right) \\ &= \mathrm{tr}\left(\left((\mathbf{a} \odot e^{\mathbf{X}\mathbf{b}})\mathbf{b}^T\right)^T d(\mathbf{X})\right). \end{aligned}$$

- Observe that $df = \mathrm{tr}\big((\frac{\partial f}{\partial \mathbf{X}})^T d\mathbf{X}\big) = \mathrm{tr}\left(\left((\mathbf{a} \odot e^{\mathbf{X}\mathbf{b}})\mathbf{b}^T\right)^T d(\mathbf{X})\right)$, we can conclude that $\frac{\partial f}{\partial \mathbf{X}} = (\mathbf{a} \odot e^{\mathbf{X}\mathbf{b}})\mathbf{b}^T$.

Example 3

Calculate the derivative of $f = \mathrm{tr}((\sigma(\mathbf{WX}))^T\mathbf{M}\sigma(\mathbf{WX}))$, where $\mathbf{W} \in \mathbb{R}^{l\times m}$, $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{M} \in \mathbb{S}^{l\times l}$, and $\sigma(\mathbf{WX})$ is a element-wise function.

- Denote $\mathbf{Y} = \sigma(\mathbf{WX})$, first calculate the differential of f as

$$
\begin{aligned}
df &= d\big(\mathrm{tr}(\mathbf{Y}^T\mathbf{MY})\big) = \mathrm{tr}\big(d(\mathbf{Y}^T\mathbf{MY})\big) = \mathrm{tr}\big((d\mathbf{Y}^T)\mathbf{MY}\big) + \mathrm{tr}(\mathbf{Y}^T\mathbf{M}d\mathbf{Y}) \\
&= \mathrm{tr}(\mathbf{Y}^T\mathbf{M}^Td\mathbf{Y}) + \mathrm{tr}(\mathbf{Y}^T\mathbf{M}d\mathbf{Y}) = \mathrm{tr}\big(\mathbf{Y}^T(\mathbf{M}^T+\mathbf{M})d\mathbf{Y}\big),
\end{aligned}
$$

  hence $\frac{\partial f}{\partial \mathbf{Y}} = 2\mathbf{MY}$.

- Observe that $df = \mathrm{tr}((\frac{\partial f}{\partial \mathbf{Y}})^T d\mathbf{Y})$, we have

$$
df = \mathrm{tr}\Big((\frac{\partial f}{\partial \mathbf{Y}})^T(\sigma'(\mathbf{WX})\odot(\mathbf{W}d\mathbf{X}))\Big) = \mathrm{tr}\Big((\frac{\partial f}{\partial \mathbf{Y}}\odot\sigma'(\mathbf{WX}))^T\mathbf{W}d\mathbf{X}\Big),
$$

  which means $\frac{\partial f}{\partial \mathbf{X}} = \mathbf{W}^T\left(\frac{\partial f}{\partial \mathbf{Y}}\odot\sigma'(\mathbf{WX})\right) = \mathbf{W}^T((2\mathbf{M}\sigma(\mathbf{WX}))\odot\sigma'(\mathbf{WX}))$.

Example 4

Calculate the derivative of $f(\mathbf{x}) = \left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{y} \in \mathbb{R}^m$.

- The variable is a vector, but we can take it as a special case of matrix.

- First calculate the differential of f as

$$\begin{aligned} df &= d\left((\mathbf{A}\mathbf{x} - \mathbf{y})^T (\mathbf{A}\mathbf{x} - \mathbf{y})\right) = (\mathbf{A}d\mathbf{x})^T (\mathbf{A}\mathbf{x} - \mathbf{y}) + (\mathbf{A}\mathbf{x} - \mathbf{y})^T \mathbf{A}d\mathbf{x} \\ &= \mathrm{tr}\left(2 (\mathbf{A}\mathbf{x} - \mathbf{y})^T \mathbf{A}d\mathbf{x}\right). \end{aligned}$$

- Then we can conclude that $\frac{\partial f}{\partial \mathbf{x}} = 2\mathbf{A}^T (\mathbf{A}\mathbf{x} - \mathbf{y})$.

Example 5: Two-Layer Neural Network

Consider a classification problem, where we have samples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)$ with $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in \mathbb{R}^m$. Note that $\mathbf{y}_i$ is a zero vector with one entry equals one. The loss function of a two layer neural networks can be defined as

$$\ell(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2) = -\sum_{i=1}^{N} \mathbf{y}_i^T \log\left(\mathsf{softmax}\left(\mathbf{W}_2 \sigma\left(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1\right) + \mathbf{b}_2\right)\right),$$

where $\mathbf{W}_2 \in \mathbb{R}^{m \times p}$, $\mathbf{W}_1 \in \mathbb{R}^{p \times n}$, $\mathbf{b}_1 \in \mathbb{R}^p$, $\mathbf{b}_2 \in \mathbb{R}^m$, softmax$(\mathbf{x}) = \frac{e^{\mathbf{x}}}{\mathbf{1}^T e^{\mathbf{x}}}$, and $\sigma(x) = \frac{1}{1+e^{-x}}$. In the following, we will derive the derivative of ℓ.

- Let $\mathbf{a}_{1,i} = \mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1$, $\mathbf{h}_{1,i} = \sigma\left(\mathbf{a}_{1,i}\right)$, and $\mathbf{a}_{2,i} = \mathbf{W}_2 \mathbf{h}_{1,i} + \mathbf{b}_2$, then

$$\ell = -\sum_{i=1}^{N} \mathbf{y}_i^T \log\left(\mathsf{softmax}\left(\mathbf{a}_{2,i}\right)\right).$$

Example 5: Two-Layer Neural Network

- Loss function ℓ can be rewritten as

$$\ell = -\sum_{i=1}^{N} \mathbf{y}_i^T \log\left(\frac{e^{\mathbf{a}_{2,i}}}{\mathbf{1}^T e^{\mathbf{a}_{2,i}}}\right) = -\sum_{i=1}^{N} \mathbf{y}_i^T \left(\log(e^{\mathbf{a}_{2,i}}) - \mathbf{1}\log(\mathbf{1}^T e^{\mathbf{a}_{2,i}})\right)$$

$$= -\sum_{i=1}^{N} \mathbf{y}_i^T \mathbf{a}_{2,i} + \sum_{i=1}^{N} \log(\mathbf{1}^T e^{\mathbf{a}_{2,i}}).$$

- We will first calculate $\frac{\partial \ell}{\partial \mathbf{a}_{2,i}}$. The differential of ℓ is

$$d\ell = d\Big(-\sum_{i=1}^{N} \mathbf{y}_i^T \mathbf{a}_{2,i} + \sum_{i=1}^{N} \log(\mathbf{1}^T e^{\mathbf{a}_{2,i}})\Big) = -\sum_{i=1}^{N} \mathbf{y}_i^T d\mathbf{a}_{2,i} + \sum_{i=1}^{N} \frac{d(\mathbf{1}^T e^{\mathbf{a}_{2,i}})}{\mathbf{1}^T e^{\mathbf{a}_{2,i}}}$$

$$= -\sum_{i=1}^{N} \mathbf{y}_i^T d\mathbf{a}_{2,i} + \sum_{i=1}^{N} \frac{\mathbf{1}^T d\left(e^{\mathbf{a}_{2,i}}\right)}{\mathbf{1}^T e^{\mathbf{a}_{2,i}}} = -\sum_{i=1}^{N} \mathbf{y}_i^T d\mathbf{a}_{2,i} + \sum_{i=1}^{N} \frac{\mathbf{1}^T \left(e^{\mathbf{a}_{2,i}} \odot d\mathbf{a}_{2,i}\right)}{\mathbf{1}^T e^{\mathbf{a}_{2,i}}}$$

Example 5: Two-Layer Neural Network

- The differential of ℓ can be rewritten as

$$
\begin{aligned}
d\ell &= \mathrm{tr}\Big(-\sum_{i=1}^{N}\mathbf{y}_i^T d\mathbf{a}_{2,i} + \sum_{i=1}^{N}\frac{\mathbf{1}^T(e^{\mathbf{a}_{2,i}} \odot d\mathbf{a}_{2,i})}{\mathbf{1}^T e^{\mathbf{a}_{2,i}}}\Big) \\
&= -\sum_{i=1}^{N}\mathbf{y}_i^T d\mathbf{a}_{2,i} + \mathrm{tr}\big(\sum_{i=1}^{N}\frac{(e^{\mathbf{a}_{2,i}})^T d\mathbf{a}_{2,i}}{\mathbf{1}^T e^{\mathbf{a}_{2,i}}}\big) \\
&= -\sum_{i=1}^{N}\mathbf{y}_i^T d\mathbf{a}_{2,i} + \sum_{i=1}^{N}\mathsf{softmax}(\mathbf{a}_{2,i})^T d\mathbf{a}_{2,i} \\
&= \sum_{i=1}^{N}(\mathsf{softmax}(\mathbf{a}_{2,i}) - \mathbf{y}_i)^T d\mathbf{a}_{2,i} \\
&= \mathrm{tr}\Big(\sum_{i=1}^{N}(\mathsf{softmax}(\mathbf{a}_{2,i}) - \mathbf{y}_i)^T d\mathbf{a}_{2,i}\Big),
\end{aligned}
$$

which means $\frac{\partial \ell}{\partial \mathbf{a}_{2,i}} = \mathsf{softmax}(\mathbf{a}_{2,i}) - \mathbf{y}_i$.

Example 5: Two-Layer Neural Network

- Observe that

$$
\begin{aligned}
d\ell = \operatorname{tr}\Big(\sum_{i=1}^{N}(\frac{\partial \ell}{\partial \mathbf{a}_{2,i}})^T d\mathbf{a}_{2,i}\Big) &= \operatorname{tr}\Big(\sum_{i=1}^{N}(\frac{\partial \ell}{\partial \mathbf{a}_{2,i}})^T d(\mathbf{W}_2\mathbf{h}_{1,i} + \mathbf{b}_2)\Big) \\
&= \operatorname{tr}\Big(\sum_{i=1}^{N}(\frac{\partial \ell}{\partial \mathbf{a}_{2,i}})^T d(\mathbf{W}_2)\mathbf{h}_{1,i}\Big) + \operatorname{tr}\Big(\sum_{i=1}^{N}(\frac{\partial \ell}{\partial \mathbf{a}_{2,i}})^T \mathbf{W}_2 d(\mathbf{h}_{1,i})\Big) \\
&\quad + \operatorname{tr}\Big(\sum_{i=1}^{N}(\frac{\partial \ell}{\partial \mathbf{a}_{2,i}})^T d\mathbf{b}_2\Big),
\end{aligned}
$$

from which we can get $\frac{\partial \ell}{\partial \mathbf{W}_2} = \sum_{i=1}^{N} \frac{\partial \ell}{\partial \mathbf{a}_{2,i}} \mathbf{h}_{1,i}^T$, $\frac{\partial \ell}{\partial \mathbf{h}_{1,i}} = \mathbf{W}_2^T \frac{\partial \ell}{\partial \mathbf{a}_{2,i}}$, and $\frac{\partial \ell}{\partial \mathbf{b}_2} = \sum_{i=1}^{N} \frac{\partial \ell}{\partial \mathbf{a}_{2,i}}$.

Example 5: Two-Layer Neural Network

- Since $\mathbf{h}_{1,i} = \sigma(\mathbf{a}_{1,i})$, we have

$$\frac{\partial \ell}{\partial \mathbf{a}_{1,i}} = \frac{\partial \ell}{\partial \mathbf{h}_{1,i}} \odot \sigma'(\mathbf{a}_{1,i}).$$

- Considering that

$$\begin{aligned} d\ell &= \mathrm{tr}\Big(\sum_{i=1}^{N} (\frac{\partial \ell}{\partial \mathbf{a}_{1,i}})^T d\mathbf{a}_{1,i}\Big) = \mathrm{tr}\Big(\sum_{i=1}^{N} (\frac{\partial \ell}{\partial \mathbf{a}_{1,i}})^T d(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1)\Big) \\ &= \mathrm{tr}\Big(\sum_{i=1}^{N} (\frac{\partial \ell}{\partial \mathbf{a}_{1,i}})^T d(\mathbf{W}_1)\mathbf{x}_i\Big) + \mathrm{tr}\Big(\sum_{i=1}^{N} (\frac{\partial \ell}{\partial \mathbf{a}_{1,i}})^T d\mathbf{b}_1\Big), \end{aligned}$$

we can get $\frac{\partial \ell}{\partial \mathbf{W}_1} = \sum_{i=1}^{N} \frac{\partial \ell}{\partial \mathbf{a}_{1,i}} \mathbf{x}_i^T$ and $\frac{\partial \ell}{\partial \mathbf{b}_1} = \sum_{i=1}^{N} \frac{\partial \ell}{\partial \mathbf{a}_{1,i}}$.

Example 5: Two-Layer Neural Network

- Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, $\mathbf{A}_1 = [\mathbf{a}_{1,1}, \ldots, \mathbf{a}_{1,N}]$, $\mathbf{H}_1 = [\mathbf{h}_{1,1}, \ldots, \mathbf{h}_{1,N}]$, and $\mathbf{A}_2 = [\mathbf{a}_{2,1}, \ldots, \mathbf{a}_{2,N}]$.

- Then we have

$$\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{W}_2} &= \frac{\partial \ell}{\partial \mathbf{A}_2}\mathbf{H}_1^T \\
\frac{\partial \ell}{\partial \mathbf{H}_1} &= \mathbf{W}_2^T\frac{\partial \ell}{\partial \mathbf{A}_2} \\
\frac{\partial \ell}{\partial \mathbf{b}_2} &= \frac{\partial \ell}{\partial \mathbf{A}_2}\mathbf{1} \\
\frac{\partial \ell}{\partial \mathbf{A}_1} &= \frac{\partial \ell}{\partial \mathbf{H}_1} \odot \sigma'(\mathbf{A}_1) \\
\frac{\partial \ell}{\partial \mathbf{W}_1} &= \sum_{i=1}^{N} \frac{\partial \ell}{\partial \mathbf{A}_1}\mathbf{X}^T \\
\frac{\partial \ell}{\partial \mathbf{b}_1} &= \frac{\partial \ell}{\partial \mathbf{A}_1}.
\end{aligned}$$

Complex-Differentiable

- Similar to real functions, for a complex function that is continuous at point z, we can define its complex derivative as

$$f'(z) = \frac{df}{dz} = \lim_{\delta z \to 0} \frac{f(z + \delta z) - f(z)}{\delta z}.$$

- In principle, we might get different results from the above formula when we plug in different infinitesimals δz (e.g., $f(z) = z^*$).

- A complex function is complex-differentiable at z is the above definition gives the same answer regardless of the argument of δz.

- Besides, if a complex function is complex-differentiable at all points in some domain, then it is said to be analytic in that domain.

Cauchy-Riemann Equations

- Let $f(z = x + \mathrm{j}y) = u(x, y) + \mathrm{j}v(x, y)$ be a complex function where $u(x, y)$ and $v(x, y)$ are real functions. If f is complex-differentiable at a given $z = x + \mathrm{j}y$, then we have

$$\begin{cases} \mathrm{Re}\{f'(z)\} = \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \\ \mathrm{Im}\{f'(z)\} = \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}. \end{cases} \qquad \text{(Cauchy-Riemann Equations)}$$

- Conversely, if the Cauchy-Riemann Equations holds at point z, then the function f is complex-differentiable at z.

Differentials of Complex Matrix

- $d\mathbf{Z} = d\mathsf{Re}\{\mathbf{Z}\} + \mathrm{j}d\mathsf{Im}\{\mathbf{Z}\}$
- $d\mathbf{Z}^* = d\mathsf{Re}\{\mathbf{Z}\} - \mathrm{j}d\mathsf{Im}\{\mathbf{Z}\}$
- $d\mathsf{Re}\{\mathbf{Z}\} = \frac{1}{2}(d\mathbf{Z} + d\mathbf{Z}^*)$
- $d\mathsf{Im}\{\mathbf{Z}\} = \frac{1}{2\mathrm{j}}(d\mathbf{Z} - d\mathbf{Z}^*)$

Basic Rules for Matrix Calculus

Consider two matrices $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{C}^{m \times n}$, we introduce some basic rules for matrix calculus in the following.

- Addition and subtraction: $d\left(\mathbf{Z}_1 \pm \mathbf{Z}_2\right) = d\mathbf{Z}_1 \pm d\mathbf{Z}_2$;
- Multiplication: $d\left(\mathbf{Z}_1 \mathbf{Z}_2\right) = \left(d\mathbf{Z}_1\right)\mathbf{Z}_2 + \mathbf{Z}_1 d\mathbf{Z}_2$;
- Transpose: $d\left(\mathbf{Z}_1^H\right) = \left(d\mathbf{Z}_1\right)^H$;
- Trace: $d\mathrm{tr}\left(\mathbf{Z}_1\right) = \mathrm{tr}\left(d\mathbf{Z}_1\right);$
- Element-wise multiplication: $d\left(\mathbf{Z}_1 \odot \mathbf{Z}_2\right) = d\mathbf{Z}_1 \odot \mathbf{Z}_2 + \mathbf{Z}_1 \odot d\mathbf{Z}_2$;
- Inverse: $d\mathbf{Z}_1^{-1} = -\mathbf{Z}_1^{-1}\left(d\mathbf{Z}_1\right)\mathbf{Z}_1^{-1}$;
- Determinant: $d\left|\mathbf{Z}_1\right| = \mathrm{tr}\left(\mathbf{Z}_1^{\sharp} d\mathbf{Z}_1\right) = \left|\mathbf{Z}_1\right|\mathrm{tr}\left(\mathbf{Z}_1^{-1} d\mathbf{Z}_1\right)$.

Thanks!