



AI and Deep Learning

2-1-S1 Derivatives and the Chain Rule

Zhonglei Wang

WISE, SOE, CHOW

XMU

(Version v1)

Contents

1. Derivatives

2. Chain rule

Derivatives

1. Recall the following rules for $x \in \mathbb{R}$:

- Let $a \in \mathbb{R}$ be constant. If $f(x) = ax$, then $f'(x) = a$
- For differentiable univariate functions $u = u(x)$ and $v = v(x)$, we have

$$(u \cdot v)' = u' \cdot v + v' \cdot u$$

2. For a vector input $\mathbf{x} = (x_1, \dots, x_m)^T \in \mathbb{R}^{m \times 1}$, use the following derivative notation and layout conventions for differentiable functions

- For a scalar-valued function $f(\mathbf{x})$, $\partial f / \partial \mathbf{x} = (\partial f / \partial x_1, \dots, \partial f / \partial x_m)^T$ has the same shape as $\mathbf{x}$
- For a vector-valued function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_k(\mathbf{x}))^T$, we have

$$\partial \mathbf{f}^T / \partial \mathbf{x} = (\partial f_1 / \partial \mathbf{x}, \dots, \partial f_k / \partial \mathbf{x}) \in \mathbb{R}^{m \times k}$$

Examples

1. **Example 1:** For a constant vector $\mathbf{a} \in \mathbb{R}^{m \times 1}$, consider $f(\mathbf{x}) = \mathbf{a}^T \cdot \mathbf{x}$ with

input $\mathbf{x} \in \mathbb{R}^{m \times 1}$

2. Then:

$$f(\mathbf{x}) = \sum_{j=1}^m a_j x_j$$

- $\mathbf{a} = (a_1, \dots, a_m)^T, \mathbf{x} = (x_1, \dots, x_m)^T$

3. Thus, $\partial f / \partial \mathbf{x} = (a_1, \dots, a_m)^T = \mathbf{a}$

Examples

1. **Example 1.1:** For $\boldsymbol{x} \in \mathbb{R}^{m \times 1}$ and $\boldsymbol{y} \in \mathbb{R}^{m \times 1}$, consider $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \cdot \boldsymbol{y}$

2. Then:

$$\begin{aligned} \frac{\partial f}{\partial \boldsymbol{x}} &= \boldsymbol{y}, & \frac{\partial f}{\partial \boldsymbol{y}} &= \boldsymbol{x}, \\ \frac{\partial f}{\partial \boldsymbol{x}^T} &= \boldsymbol{y}^T, & \frac{\partial f}{\partial \boldsymbol{y}^T} &= \boldsymbol{x}^T \end{aligned}$$

3. See also Section 2.1.

Examples

1. **Example 2:** For two vector-valued functions $\mathbf{u}(\mathbf{x}) \in \mathbb{R}^{k \times 1}$ and $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^{k \times 1}$,

consider $f(\mathbf{x}) = \mathbf{u}(\mathbf{x})^T \cdot \mathbf{v}(\mathbf{x})$

2. Then:

$$f(\mathbf{x}) = \sum_{j=1}^k u_j(\mathbf{x})v_j(\mathbf{x})$$

- $\mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), \dots, u_k(\mathbf{x}))^T, \mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), \dots, v_k(\mathbf{x}))^T$

3. Verify that

$$\frac{\partial f}{\partial \mathbf{x}} = \sum_{j=1}^k \left\{ \frac{\partial u_j(\mathbf{x})}{\partial \mathbf{x}} \cdot v_j(\mathbf{x}) + \frac{\partial v_j(\mathbf{x})}{\partial \mathbf{x}} \cdot u_j(\mathbf{x}) \right\} = \frac{\partial \mathbf{u}^T}{\partial \mathbf{x}} \cdot \mathbf{v}(\mathbf{x}) + \frac{\partial \mathbf{v}^T}{\partial \mathbf{x}} \cdot \mathbf{u}(\mathbf{x})$$

Examples

1. **Example 3:** For $\boldsymbol{x} \in \mathbb{R}^{m \times 1}$, consider the identity function $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$

2. By definition:

$$\partial \boldsymbol{f}^T / \partial \boldsymbol{x} = (\partial f_1 / \partial \boldsymbol{x}, \dots, \partial f_m / \partial \boldsymbol{x}) = (\partial x_1 / \partial \boldsymbol{x}, \dots, \partial x_m / \partial \boldsymbol{x}) = \boldsymbol{I}$$

- $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x}))^T = (x_1, \dots, x_m)^T$
- $\boldsymbol{I}$: the $m \times m$ identity matrix
- Verify the calculation above.

Examples

1. **Example 4:** For a constant matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{x} \in \mathbb{R}^{m \times 1}$, consider $\mathbf{f}(\mathbf{x}) = \mathbf{A} \cdot \mathbf{x}$

2. Note that:

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{pmatrix} \cdot \mathbf{x} = \begin{pmatrix} \mathbf{a}_1^T \cdot \mathbf{x} \\ \vdots \\ \mathbf{a}_n^T \cdot \mathbf{x} \end{pmatrix}$$

- $\mathbf{a}_i^T$: the i th row of $\mathbf{A}$

3. Thus:

$$\frac{\partial \mathbf{f}^T}{\partial \mathbf{x}} = \left(\frac{\partial(\mathbf{a}_1^T \cdot \mathbf{x})}{\partial \mathbf{x}}, \dots, \frac{\partial(\mathbf{a}_n^T \cdot \mathbf{x})}{\partial \mathbf{x}} \right) = (\mathbf{a}_1, \dots, \mathbf{a}_n) = \mathbf{A}^T$$

- The second equality follows from Example 1.
- Verify the calculation above.

Examples

1. **Example 5:** For a constant matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ and $\mathbf{x} \in \mathbb{R}^{m \times 1}$, consider

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

2. Let $\mathbf{u}(\mathbf{x}) = \mathbf{x}$ and $\mathbf{v}(\mathbf{x}) = \mathbf{A} \cdot \mathbf{x}$, so $f(\mathbf{x}) = \mathbf{u}(\mathbf{x})^T \cdot \mathbf{v}(\mathbf{x})$

3. By Examples 2–4:

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}^T}{\partial \mathbf{x}} \cdot \mathbf{v} + \frac{\partial \mathbf{v}^T}{\partial \mathbf{x}} \cdot \mathbf{u} = \mathbf{A} \cdot \mathbf{x} + \mathbf{A}^T \cdot \mathbf{x} = (\mathbf{A} + \mathbf{A}^T) \cdot \mathbf{x}$$

4. Verify the calculation above.

Examples

1. **Example 5.1:** For $\boldsymbol{x} \in \mathbb{R}^{m \times 1}$, consider $f(\boldsymbol{x}) = \boldsymbol{x}^T \cdot \boldsymbol{x}$
2. With $\boldsymbol{A} = \boldsymbol{I}$, Example 5 gives $\partial f / \partial \boldsymbol{x} = 2\boldsymbol{x}$
3. **Example 5.2:** For a constant symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ satisfying $\boldsymbol{A}^T = \boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^{m \times 1}$, consider $f(\boldsymbol{x}) = \boldsymbol{x}^T \cdot \boldsymbol{A} \cdot \boldsymbol{x}$
4. By Example 5, $\partial f / \partial \boldsymbol{x} = 2\boldsymbol{A} \cdot \boldsymbol{x}$

Chain rule

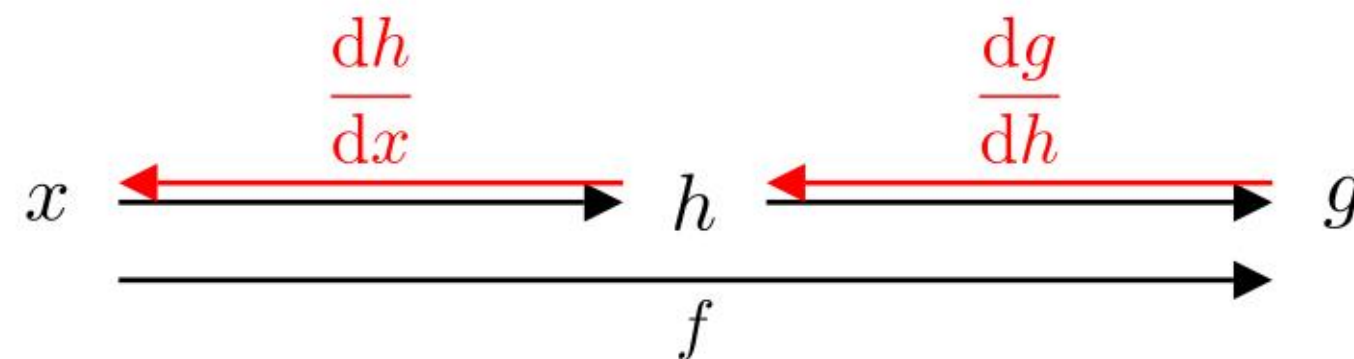
1. First, consider $f(x) = g \circ h(x)$

- $x \in \mathbb{R}$
- $h(x) \in \mathbb{R}$
- $g(h) \in \mathbb{R}$
- Assume the inner and outer functions are differentiable at the relevant points.

2. The computational graph is shown on the right.

3. By the chain rule:

$$\frac{df}{dx} = \frac{dh}{dx} \cdot \frac{dg}{dh}$$



Chain rule

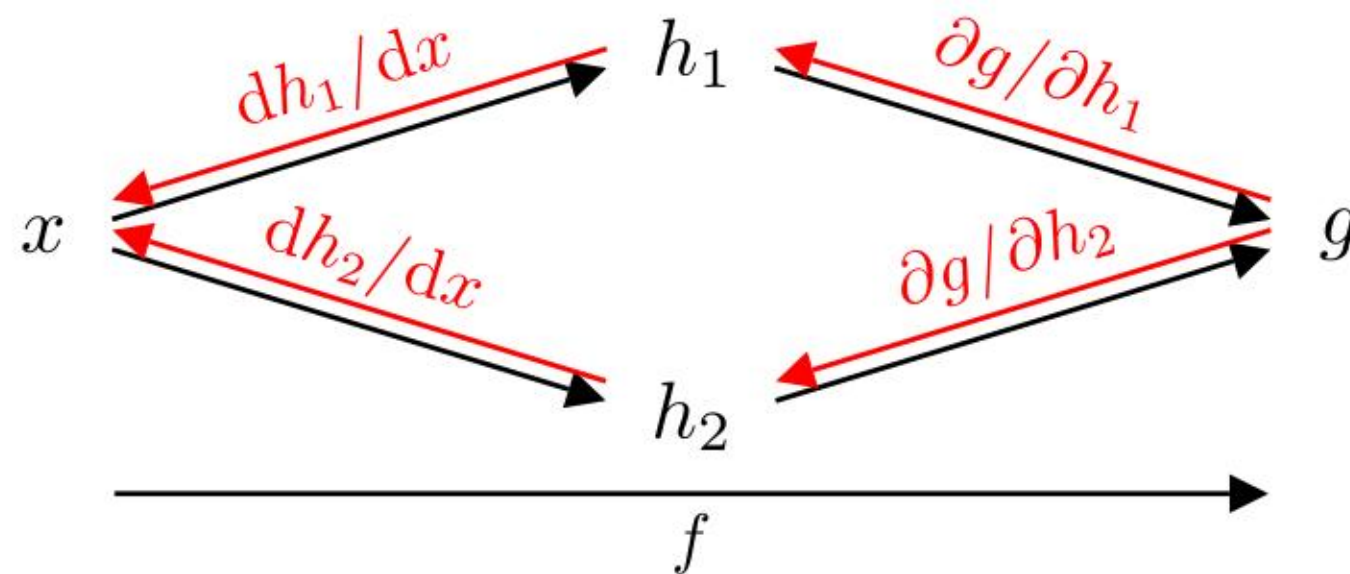
1. Next, consider $f(x) = g \circ \mathbf{h}(x)$

- $x \in \mathbb{R}$
- $\mathbf{h}(x) = (h_1(x), h_2(x))^T \in \mathbb{R}^2$
- $g(\mathbf{h}) \in \mathbb{R}$
- Assume the inner and outer functions are differentiable at the relevant points.

2. The computational graph is shown on the right.

3. By the chain rule:

$$\frac{df}{dx} = \frac{dh_1}{dx} \cdot \frac{\partial g}{\partial h_1} + \frac{dh_2}{dx} \cdot \frac{\partial g}{\partial h_2}$$



Chain rule

1. More generally, consider $f(x) = g \circ \mathbf{h}(x)$

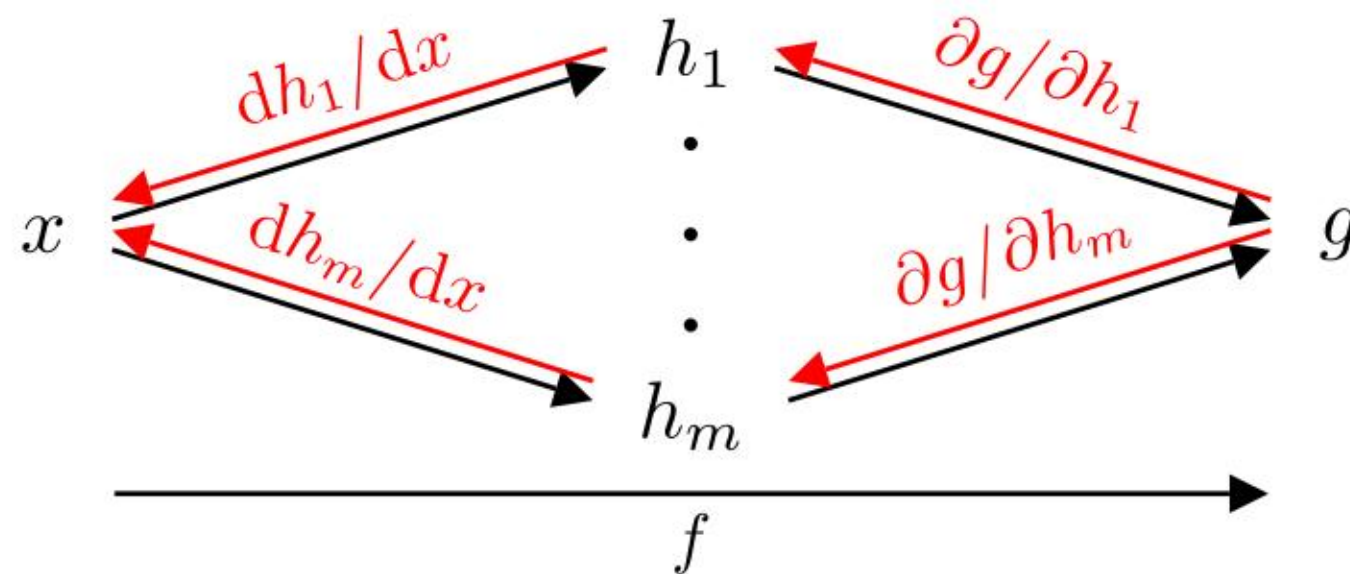
- $x \in \mathbb{R}$
- $\mathbf{h}(x) = (h_1(x), \dots, h_m(x))^T \in \mathbb{R}^m$
- $g(\mathbf{h}) \in \mathbb{R}$
- Assume the inner and outer functions are differentiable at the relevant points.

2. The computational graph is shown on the right.

3. By the chain rule:

$$\frac{df}{dx} = \sum_{j=1}^m \frac{dh_j}{dx} \cdot \frac{\partial g}{\partial h_j}$$

4. See Section 2.1 for the general case.



Summary

1. Vector differentiation

- Identify input and output shapes; keep the derivative layout consistent
- Decompose into inner products, linear transformations, and other basic operations; combine known derivative results

2. The chain rule for composite functions

- Identify dependency paths from input to output
- Multiply local derivatives along each path, then sum the path contributions