Differentiation rule

- Constant

$$f(x) = c \quad \rightarrow \quad f'(x) = 0$$

- Factors

$$f(x) = c \cdot g(x) \rightarrow f'(x) = c \cdot g'(x)$$

- Power rule

$$f(x) = x^n \rightarrow f'(x) = n \cdot x^{n-1}$$

- Sum rule

$$f(x) = g(x) + h(x)$$

$$\rightarrow f'(x) = g'(x) + h'(x)$$

- Product rule

$$f(x) = g(x) \cdot h(x)$$

$$\rightarrow f'(x) = g'(x) h(x) + g(x) h'(x)$$

- Quotient rule

$$f(x) = \frac{g(x)}{h(x)}$$

$$\rightarrow f'(x) = \frac{h(x) g'(x) - g(x) h'(x)}{(h(x))^2}$$

- Chain rule

$f(x) = g(h(x))$

$\rightarrow f'(x) = g'(h(x)) \cdot h'(x)$

e.g. $y = (3x+5)^5$

$$\frac{dy}{dx} = 5(3x+5)^4 \cdot 3$$

- Important derivatives

  - $(\log x)' = \frac{1}{x}$

  - $(e^x)' = e^x$

- Partial Derivatives

Derivative of a scalar-valued function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ of d variables with respect to one of those variables

$z = f(x,y) = x^2 + 3xy + 4y^3$

$$\rightarrow \frac{\partial z}{\partial x} = 2x + 3y$$

- Gradients

  Derivative of a scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$ of d variables w.r.t. all of those variables

$$z = f(x,y) = (x^2 + 3xy + 4y^3)$$

$$\to \quad \nabla z = \begin{pmatrix} \frac{\partial z}{\partial x} \\ \frac{\partial z}{\partial y} \end{pmatrix} = \begin{pmatrix} 2x + 3y \\ 3x + 12y^2 \end{pmatrix}$$

- Jacobian Matrices

  Derivative of a vector-valued function $F : \mathbb{R}^d \to \mathbb{R}^n$, ie. gradient of each element of the output vector $F(x) \in \mathbb{R}^n$ w.r.t. the input vector $x \in \mathbb{R}^d$

$$\frac{\partial F}{\partial x} = J_F = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_1} & \cdots & \frac{\partial F_n}{\partial x_d} \end{pmatrix} \in \mathbb{R}^{n \times d}$$

e.g.

$$Z = \begin{pmatrix} 2x_1 + 3x_3 \\ 3x_2 \\ 4x_3 \\ x_2 - x_3 \end{pmatrix} ; \underbrace{Z \in \mathbb{R}^4}_{\text{output vector}} , \underbrace{x \in \mathbb{R}^3}_{\text{input vector}}$$

$$\rightarrow J_Z = \begin{pmatrix} 2 & 0 & 3 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \\ 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{4 \times 3}$$

· Some other examples

∘ $x \in \mathbb{R}^n$, $\underbrace{F(x)}_{\text{scalar-valued function}} = x^T x \in \mathbb{R}$

$$x^T x = x_1^2 + x_2^2 + \dots + x_n^2$$

$$\rightarrow \frac{\partial x^T x}{\partial x} = \begin{pmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{pmatrix} = 2x$$

- $w^T x = w_1 x_1 + \dots + w_n x_n$

$$\to \frac{\partial w^T x}{\partial x} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = w$$

$$\frac{\partial w^T x}{\partial w} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x$$

- $x^T w = x_1 w_1 + \dots + x_n w_n$

$$\to \frac{\partial x^T w}{\partial x} = w$$

$$\to \frac{\partial x^T w}{\partial w} = x$$

Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Schubert
Philipp Pfefferkorn, Zongyue Li

Deep Learning and Artificial Intelligence

WS 2025/26

Exercise 2: Math Primer

Exercise 2-1 Partial Derivative of Cross-Entropy Loss

Given a prediction $\hat{\mathbf{y}} = \text{softmax}(\mathbf{z})$ with $\mathbf{z}, \hat{\mathbf{y}} \in \mathbb{R}^K$, where K is the number of classes of a classification problem. The i-th element of $\hat{\mathbf{y}}$ is defined as $\hat{y}_i = \text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{k=1}^K e^{z_k}}$ for $i = 1, \ldots K$.

Calculate the partial derivative $\frac{\partial \hat{y}_i}{\partial z_j}$!

when $i = j$, $g = e^{z_j}$, $h = \sum_{k=1}^K e^{z_k}$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{g'h - gh'}{h^2} = \frac{e^{z_j} \cdot \sum_{k=1}^K e^{z_k} - e^{z_j} \cdot e^{z_j}}{\sum_{k=1}^K e^{z_k}} = \text{softmax}(z)_j \cdot (1 - \text{softmax}(z)_j)$$

Quotient rule

when $i \neq j$, $g = e^{z_i}$, $h = \sum_{k=1}^K e^{z_k}$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{g'h - gh'}{h^2} = \frac{0 \cdot \sum_{k=1}^K e^{z_k} - e^{z_i} e^{z_j}}{(\sum_{k=1}^K e^{z_k})^2} = -\text{softmax}(z)_i \cdot \text{softmax}(z)_j$$

Exercise 2-2 Gradients

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a scalar function that maps an input vector $\mathbf{x} \in \mathbb{R}^d$ to a scalar output $f(\mathbf{x}) \in \mathbb{R}$. The gradient $\frac{\partial f}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} f$ (or ∇f for short) is defined as:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} \in \mathbb{R}^d.$$

$$\nabla y = \begin{pmatrix} 2 \\ 6x_2 \\ \frac{-3}{x_3^2} \end{pmatrix}$$

(a) Given $y = 2x_1 + 3(x_2)^2 + \frac{3}{x_3}$. with $x \in \mathbb{R}^3$. Calculate the gradient $\nabla_{\mathbf{x}} y$!

log + chain rule

(b) Given $z = 3 + \log(2a_1) + e^{2a_3}$. with $a \in \mathbb{R}^3$. Calculate the gradient $\nabla_{\mathbf{a}} z$!

$$\nabla_a z = \begin{pmatrix} 0 + \frac{1}{2a_1} \cdot 2 + 0 \\ 0 + 0 + 0 \\ 0 + 0 + e^{2a_3} \cdot 2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2a_1} \\ 0 \\ 2e^{2a_3} \end{pmatrix}$$

e^x + chain rule

Exercise 2-3 Jacobian Matrices

Suppose $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^n$ is a vector-valued function that maps an input vector $\mathbf{x} \in \mathbb{R}^d$ to an output vector $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^n$. The Jacobian matrix $\frac{\partial \mathbf{F}}{\partial \mathbf{x}} = J_{\mathbf{F}}$ is defined as:

$$J_{\mathbf{F}} := \begin{pmatrix} \frac{\partial \mathbf{F}}{\partial x_1} & \cdots & \frac{\partial \mathbf{F}}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_1} & \cdots & \frac{\partial F_n}{\partial x_d} \end{pmatrix} \in \mathbb{R}^{n \times d},$$

i.e. it contains the derivatives of each output with regard to each input: $(J_{\mathbf{F}})_{ij} = \frac{\partial F_i}{\partial x_j}$.

In the following, let $\mathbf{x} \in \mathbb{R}^d$ be a vector with d elements and $\mathbf{W} \in \mathbb{R}^{n \times d}$ be a matrix with n rows and d columns.

(a) The i-th element of z : $z_i = \sum_{k=1}^{d} w_{ik} \cdot x_k$

$$(J_z)_{ij} = \frac{\partial z_i}{\partial x_j} = w_{ij} \Rightarrow J_z = \frac{\partial z}{\partial x} = w$$

(a) Given $\mathbf{z} = \mathbf{W}\mathbf{x}$. Calculate the Jacobian matrix $J_\mathbf{z} = \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$! *Hint:* Interpret $\mathbf{z}$ as a vector-function that maps $\mathbf{x}$ to an n-dimensional vector.

(b) Given $\mathbf{z} = f(\mathbf{x})$, where f is applied elementwise to the vector $\mathbf{x}$, i.e. $z_i = f(x_i)$.

Calculate the Jacobian matrix $J_\mathbf{z} = \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ (not the gradient)!

$$(J_z)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} f(x_i) \Rightarrow \begin{cases} f'(x_i) & \text{when } i = j \\ 0 & \text{when } i \neq j \end{cases} \Rightarrow \frac{\partial z}{\partial x} = \begin{pmatrix} f'(x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f'(x_d) \end{pmatrix} = \operatorname{diag}(f'(x))$$

Exercise 2-4 Mean Squared Error

Consider the input dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ with n samples of size d, a target vector $\mathbf{y} \in \mathbb{R}^n$, a weight vector $\mathbf{w} \in \mathbb{R}^d$ and a prediction $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$. The mean squared error (MSE) is defined as the sum over the squared differences between the prediction $\hat{y}_i$ and the true values y_i for each instance:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2,$$

where $\hat{y}_i = \mathbf{w}^T \mathbf{x}_i$ and $\mathbf{x}_i \in \mathbb{R}^d$ is one sample of the dataset (corresponding to one row in $\mathbf{X}$).

Find the vector $\hat{\mathbf{w}}$ that minimizes the MSE loss function!

☆ *Hint:* Write the sum above as a vector product $\frac{1}{n}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})$. Moreover, you can use the following: $(\mathbf{A}\mathbf{x})^T = \mathbf{x}^T \mathbf{A}^T$, $\frac{\partial \mathbf{x}^T \mathbf{A} x}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$ and $\frac{\partial \mathbf{x}^T \mathbf{b}}{\partial \mathbf{x}} = \frac{\partial \mathbf{b}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{b}$ for vectors $\mathbf{x}$ and $\mathbf{b}$ and matrices $\mathbf{A}$.

must remember !!!

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$a^T b = b^T a$

$$= \frac{1}{n} (Xw - y)^T (Xw - y)$$

$$= \frac{1}{n} \left((Xw)^T (Xw) - (Xw)^T y - y^T (Xw) + y^T y \right)$$

$$= \frac{1}{n} \left(w^T X^T X w - 2 (Xw)^T y + y^T y \right)$$

$$= \frac{1}{n} \left(w^T \boxed{X^T X} w - 2 w^T X^T y + y^T y \right)$$

$$\frac{\partial L}{\partial w} = \nabla_w L = \frac{1}{n} \left(2 X^T X w - 2 X^T y \right)$$

$$\text{Let } \frac{\partial L}{\partial w} = \nabla_w L = 0$$

$$\Rightarrow \frac{1}{n} \left(2 X^T X w - 2 X^T y \right) = 0$$

$$X^T X w - X^T y = 0$$

$$X^T X w = X^T y$$

$$w = (X^T X)^{-1} X^T y$$