Derivative

- Function $\mathbb{R}^n \to \mathbb{R}$: $f(\mathbf{x}) = a_1 x_1 + a_2 x_2$

$$\nabla_{\mathbf{x}} f(\mathbf{x}) \triangleq \begin{bmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \dfrac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

  - $\dfrac{\partial f(\mathbf{x})}{\partial x_i}$: partial derivative of element i of vector **x**
  - If there is no variable but **x**, $\nabla_{\mathbf{x}} f(\mathbf{x})$ can be written $\nabla f(\mathbf{x})$
  - If $\mathbf{x} \in \mathbb{R}^n$, $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^n$

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 x_n} \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

  - Second-order gradient is also called Hessian and symmetric matrix size n

- Function $f(X) : \mathbb{R}^{n \times m} \to \mathbb{R}$

$$\nabla f(X) = \begin{bmatrix} \dfrac{\partial f(X)}{\partial x_{11}} & \dfrac{\partial f(X)}{\partial x_{12}} & \cdots & \dfrac{\partial f(X)}{\partial x_{1m}} \\ \dfrac{\partial f(X)}{\partial x_{21}} & \dfrac{\partial f(X)}{\partial x_{22}} & \cdots & \dfrac{\partial f(X)}{\partial x_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f(X)}{\partial x_{n1}} & \dfrac{\partial f(X)}{\partial x_{n2}} & \cdots & \dfrac{\partial f(X)}{\partial x_{nm}} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

- Function $v(x) : \mathbb{R} \to \mathbb{R}^n$

$$v(x) = \begin{bmatrix} v_1(x) \\ v_2(x) \\ \vdots \\ v_n(x) \end{bmatrix}$$

- Derivative of the function by x is the **row vector**

$$\nabla_x v(x) \triangleq \begin{bmatrix} \frac{\partial v_1(x)}{\partial x} & \frac{\partial v_2(x)}{\partial x} & \cdots & \frac{\partial v_n(x)}{\partial x} \end{bmatrix} \in \mathbb{R}^{1\times n}$$

$$\nabla_x^2 v(x) \triangleq \begin{bmatrix} \frac{\partial v_1(x)}{\partial x^2} & \frac{\partial v_2(x)}{\partial x^2} & \cdots & \frac{\partial v_n(x)}{\partial x^2} \end{bmatrix}$$

- Example: Given $\mathbf{a} \in \mathbb{R}^n$ and vector-valued $v(x) = x\mathbf{a}$, $\begin{cases} \nabla v(x) = \mathbf{a}^T \\ \nabla^2 v(x) = 0 \end{cases}$

- Function $h(\mathbf{x}) : \mathbb{R}^k \to \mathbb{R}^n$

$$h(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) : \mathbb{R}^k \to \mathbb{R} \\ h_2(\mathbf{x}) : \mathbb{R}^k \to \mathbb{R} \\ \vdots \\ h_n(\mathbf{x}) : \mathbb{R}^k \to \mathbb{R} \end{bmatrix}$$

$$\nabla h(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial h_1(\mathbf{x})}{\partial x_1} & \frac{\partial h_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_1} \\ \frac{\partial h_1(\mathbf{x})}{\partial x_2} & \frac{\partial h_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_1(\mathbf{x})}{\partial x_k} & \frac{\partial h_2(\mathbf{x})}{\partial x_k} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_k} \end{bmatrix} \in \mathbb{R}^{k\times n}$$

Derivative Properties

- Product Rule

  - Assume matrix X is the input and all functions are in proper size to multiply

$$\nabla\left(f(X)^T g(X)\right) = \left(\nabla f(X)\right) g(X) + \left(\nabla g(X)\right) f(X)$$

  - Above rule is just another way to represent: $\left(f(x) g(x)\right)' = f'(x) g(x) + g'(x) f(x)$

- Chain Rule

$$\nabla_X g\left(f(X)\right) = \left(\nabla_X f\right)^T \left(\nabla_f g\right)$$

Derivative of common function

- $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x}$

  - Assume $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$, we rewrite $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$
  - $\dfrac{\partial f(\mathbf{x})}{\partial x_i} = a_i, \forall i = 1,2,3,\ldots,n$

$$\to \nabla f(\mathbf{x}) = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^T = \mathbf{a}$$

- $f(\mathbf{x}) = A\mathbf{x}$
  - This is the function $f:\mathbb{R}^n \to \mathbb{R}^m$ with $\mathbf{x} \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}$, so $A\mathbf{x} = \begin{bmatrix} \mathbf{a}_1\mathbf{x} \\ \mathbf{a}_2\mathbf{x} \\ \vdots \\ \mathbf{a}_m\mathbf{x} \end{bmatrix}$
  - $\nabla_{\mathbf{x}}(A\mathbf{x}) = \begin{bmatrix} \mathbf{a}_1^T & \mathbf{a}_2^T & \cdots & \mathbf{a}_m^T \end{bmatrix} = A^T$
- $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$
  - Assume $\mathbf{x} \in \mathbb{R}^n, \mathrm{A} \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned}
\nabla \mathbf{x}^T A\mathbf{x} = \nabla f(\mathbf{x}) &= \nabla\left(\left(\mathbf{x}^T\right)(A\mathbf{x})\right) \\
&= \left(\nabla(\mathbf{x}) A\mathbf{x}\right) + \left(\nabla(A\mathbf{x})\right)\mathbf{x} \\
&= IA\mathbf{x} + A^T\mathbf{x} \\
&= \left(A + A^T\right)\mathbf{x}
\end{aligned}$$

$$\nabla_{\mathbf{x}}^2 \mathbf{x}^T A\mathbf{x} = A + A^T$$

- $f(\mathbf{x}) = \|A\mathbf{x} - b\|_2^2$

$$\left.\begin{aligned} \nabla(A\mathbf{x} - b) &= A^T \\ \nabla\|\mathbf{x}\|_2^2 &= 2\mathbf{x} \end{aligned}\right\} \to \nabla\|A\mathbf{x} - b\|_2^2 = 2A^T(A\mathbf{x} - b)$$

- $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}$
  - Rewrite as $f(\mathbf{x}) = \left(\mathbf{x}^T\mathbf{b}\right)\left(\mathbf{a}^T\mathbf{x}\right)$ and apply product rule:

$$\nabla\left(\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}\right) = \mathbf{b}\mathbf{a}^T\mathbf{x} + \mathbf{x}^T\mathbf{b}\mathbf{a} = \mathbf{b}\mathbf{a}^T\mathbf{x} + \mathbf{b}^T\mathbf{x}\mathbf{a} = \left(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T\right)\mathbf{x}$$

- $f(X) = trace(AX)$
  - Assume $A \in \mathbb{R}^{n\times m}, X \in \mathbb{R}^{m\times n}, B = AX \in \mathbb{R}^{n\times n}$,based on the definition of trace:

$$f(X) = trace(AX) = trace(B) = \sum_{j=1}^{n} b_{jj} = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ji}x_{ij}$$

$$\to \frac{\partial f(X)}{\partial x_{ij}} = a_{ji} \to \nabla_{\mathbf{x}} trace(AX) = A^T$$

- $f(X) = \mathbf{a}^T X\mathbf{b}$
  - Assume $\mathbf{a} \in \mathbb{R}^m, X \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^n$,after multiplying, the result: $f(X) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}a_i b_j$

$$\rightarrow \nabla_X \left(\mathbf{a}^T X\mathbf{b}\right) = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_mb_1 & a_mb_2 & \cdots & a_mb_n \end{bmatrix} = \mathbf{ab}^T$$

- $f\left(X\right) = \|X\|_F^2$
  - Assume $X \in \mathbb{R}^{n\times n}$,rewrite $\|X\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^2$, so $\frac{\partial f}{\partial x_{ij}} = 2x_{ij}$

$$\rightarrow \nabla \|X\|_F^2 = 2X$$

$f\left(\mathbf{x}\right)$	$\nabla f\left(\mathbf{x}\right)$	$f\left(X\right)$	$\nabla_X f\left(X\right)$
$\mathbf{x}$	$\mathbf{I}$	$trace\left(X\right)$	$\mathbf{I}$
$\mathbf{a}^T\mathbf{x}$	$\mathbf{a}$	$trace\left(A^TX\right)$	A
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\left(A+A^T\right)\mathbf{x}$	$trace\left(X^TAX\right)$	$\left(A+A^T\right)X$
$\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2$	$2\mathbf{x}$	$trace\left(X^TX\right) = \|X\|_F^2$	$2X$
$\|A\mathbf{x}-\mathbf{b}\|_2^2$	$2A^T\left(A\mathbf{x}-\mathbf{b}\right)$	$\|AX-B\|_F^2$	$2A^T\left(AX-B\right)$
$\mathbf{a}^T\mathbf{x}^T\mathbf{x}\mathbf{b}$	$2\mathbf{a}^T\mathbf{b}\mathbf{x}$	$\mathbf{a}^TX\mathbf{b}$	$\mathbf{ab}^T$
$\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}$	$\left(\mathbf{ab}^T+\mathbf{ba}^T\right)\mathbf{x}$	$trace\left(A^TXB\right)$	AB^T

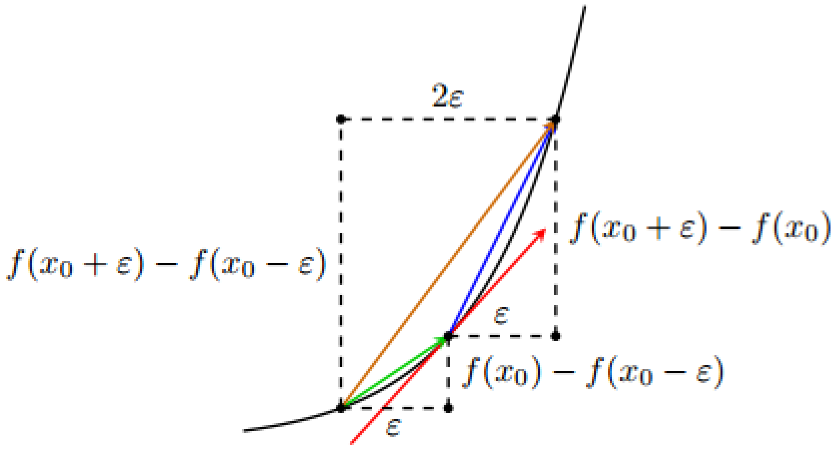
Numerical Method of Derivative

- Based on definition, $f'\left(x\right) = \lim_{\varepsilon\to 0}\frac{f\left(x+\varepsilon\right)-f\left(x\right)}{\varepsilon}$, ε usually gets very small or = 10^{-6}, $f'\left(x\right) \approx \frac{f\left(x+\varepsilon\right)-f\left(x-\varepsilon\right)}{2\varepsilon}$
- Above f'(x) can be solve by either Analysis (Taylor Series) or Geometry
  - Taylor Series $\begin{cases} f\left(x+\varepsilon\right) \approx f\left(x\right) + f'\left(x\right)\varepsilon + \frac{f''\left(x\right)}{2}\varepsilon^2 + \frac{f^{(3)}}{6}\varepsilon^3 + \ldots \\ f\left(x-\varepsilon\right) \approx f\left(x\right) - f'\left(x\right)\varepsilon + \frac{f''\left(x\right)}{2}\varepsilon^2 - \frac{f^{(3)}}{6}\varepsilon^3 + \ldots \end{cases}$

$$\rightarrow \begin{cases} \frac{f\left(x+\varepsilon\right)-f\left(x\right)}{\varepsilon} \approx f'\left(x\right) + \frac{f''\left(x\right)}{2}\varepsilon + \ldots = f'\left(x\right) + O\left(\varepsilon\right) & (2.21) \\ \frac{f\left(x+\varepsilon\right)-f\left(x-\varepsilon\right)}{2\varepsilon} \approx f'\left(x\right) + \frac{f^{(3)}\left(x\right)}{6}\varepsilon^2 + \ldots = f'\left(x\right) + O\left(\varepsilon^2\right) & (2.22) \end{cases}$$

(2.21) has the error of O(ϵ), but 2.22 just has the error of O(ϵ^2), and when ε→0, O(ϵ^2) ≪ O(ϵ), so (2.22) is much better

  - Geometry



Hình 2.1: Giải thích cách xấp xỉ đạo hàm bằng hình học.

Easily see that $f(x_0+\varepsilon)-f(x_0-\varepsilon)$ is more similar with the real derivative than $f(x_0)-f(x_0-\varepsilon)$ and $f(x_0+\varepsilon)-f(x_0)$