Calculus

AI ToolKit

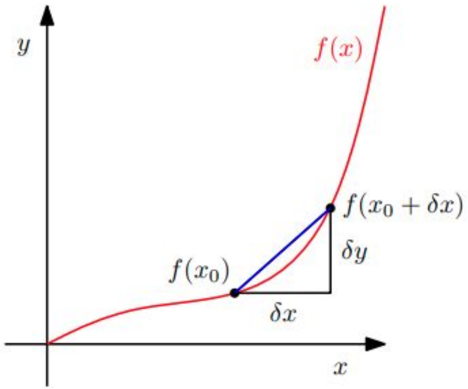
Fall Semester, 2023

Function and Machine Learning

- Goal of machine learning is to approximate the unknown real function $\mathbf{y} = f^*(\mathbf{x})$ with $f_\theta(\mathbf{x})$,
  - where $\theta \in \mathbb{R}^d$ is a model parameter.
  - We want to find optimal θ, which can optimize the objective function.
    - What kinds of objective function?

      $\Rightarrow$

- If objective functions are differentiable, $\dfrac{df}{dx} := \lim_{\delta x \to 0} \dfrac{f(x+\delta x) - f(x)}{\delta x}$ can we find optimal θ easily?



Differentiation Rules

- Product rule

  $\Rightarrow$ $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$

- Quotient rule

  $\Rightarrow$ $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$

- Sum rule

  $\Rightarrow$ $(f(x) + g(x))' = f'(x) + g'(x)$

- Chain rule

  $\Rightarrow$ $(g(f(x)))' = g'(f(x))f'(x)$

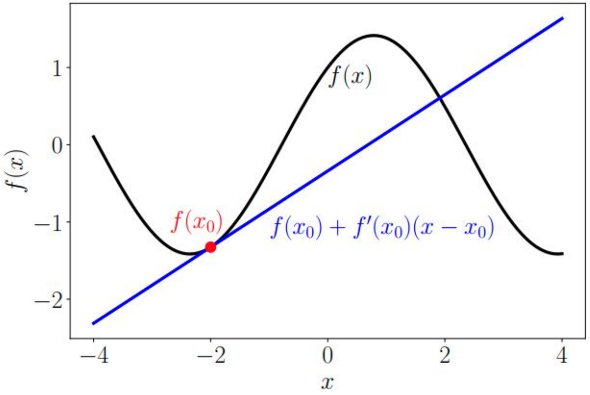
- $\frac{d}{dx}x^n = nx^{n-1}$

  $\frac{d}{dx}e^x = e^x$

  $\frac{d}{dx}\ln x = \frac{1}{x}$

Derivative and Approximation

- $\frac{df}{dx}(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \Rightarrow \frac{df}{dx}(x) \approx \frac{f(x+\epsilon) - f(x)}{\epsilon} \Rightarrow \epsilon \frac{df}{dx}(x) \approx f(x+\epsilon) - f(x) \Rightarrow f(x+\epsilon) \approx \epsilon \frac{df}{dx}(x) + f(x)$



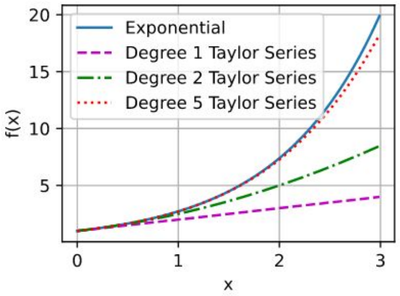
⇒ Approximate the value of f by a line which passes $(x, f(x))$ with slope of $\frac{df}{dx}(x)$.

⇒ Can we expand this approximation into n-th order polynomial?

Taylor Series

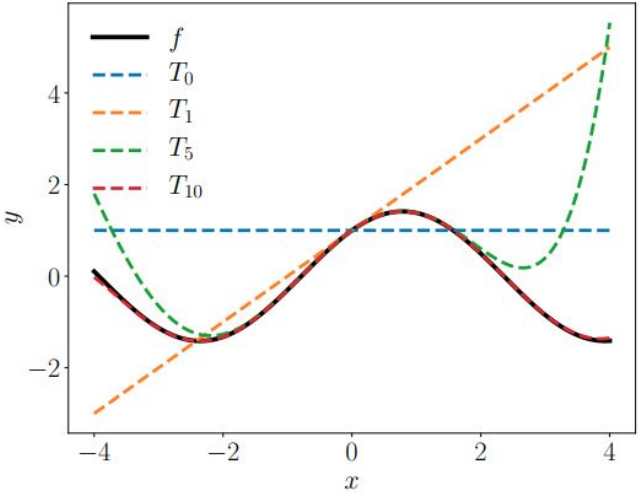
- The Taylor polynomial of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as:

$$\Rightarrow T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$



- If $n \to \infty$ the Taylor series of f at x_0 is defined as:

$$\Rightarrow T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$



Ahri Lab

Partial Differentiation and Gradient

- For a function $f : \mathbb{R}^n \to \mathbb{R}$, partial derivative is defined as:

$$\Rightarrow \frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x))}{h}$$

- If partial derivatives are collected into the row vector, gradient of a function is obtained.

$$\Rightarrow \nabla_{\mathbf{x}} f = \frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x}))}{\partial x_1} & \frac{\partial f(\mathbf{x}))}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x}))}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}$$

Gradients of Vector-Valued Functions

- For $\mathbf{x} \in \mathbb{R}^n$, a vector-valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^m$.
- Its partial derivative is given as a vector,

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h\to 0} \frac{f_1(x_1,\ldots,x_{i-1},x_i+h,x_{i+1},\ldots x_n)-f_1(\boldsymbol{x})}{h} \\ \vdots \\ \lim_{h\to 0} \frac{f_m(x_1,\ldots,x_{i-1},x_i+h,x_{i+1},\ldots x_n)-f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathbb{R}^m .$$

- And its gradient is given as a matrix, by collecting these partial derivatives:

$$\frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1}} & \cdots & \boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Rules of Partial Differentiation

- Product rule, sum rule, chain rule also applies for partial differentiation.
  - $$\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x})g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} g(\mathbf{x}) + f(\mathbf{x}) \frac{\partial g}{\partial \mathbf{x}}$$
    $$\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}}$$
    $$\frac{\partial}{\partial \mathbf{x}}(g(f(\mathbf{x})) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial \mathbf{x}}$$

- If $f(x_1, x_2)$ and $x_1(s,t)$ & $x_2(s,t)$,

$$\Rightarrow \quad \frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}$$
$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

$$\Rightarrow \quad \frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial (s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{= \frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{= \frac{\partial \boldsymbol{x}}{\partial (s,t)}}$$

⇒ The chain rule can be written as a matrix multiplication

Jacobian and Hessian

- Jacobian: the collection of all 1st-order partial derivatives of a function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

$$\Rightarrow \boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix},$$

- Hessian: the collection of all second-order partial derivatives
  - If $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$
  - If $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, the Hessian is an (m x n x n) tensor.

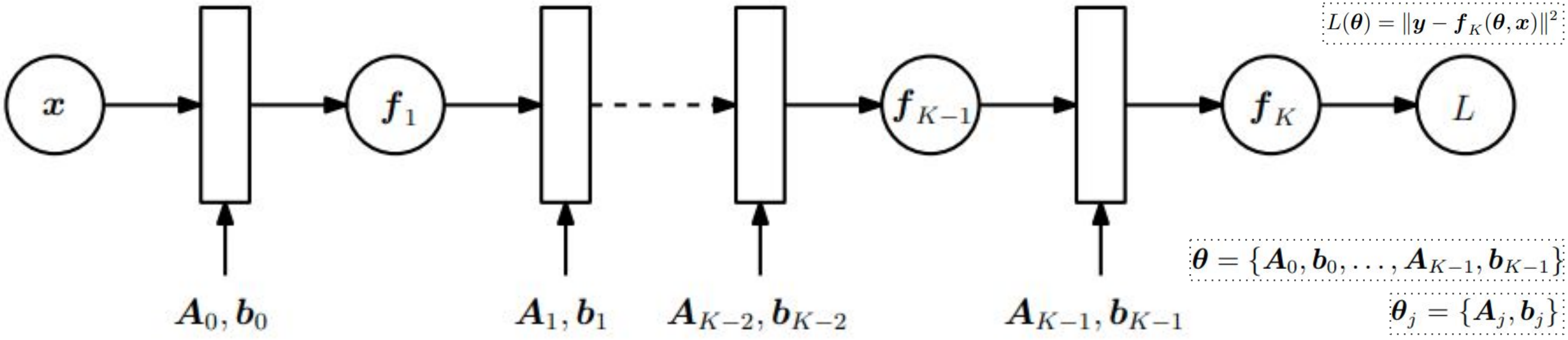
Higher-Dimension Differentiation and Approximation

- Recall $f(x+\epsilon) \approx \epsilon \frac{df}{dx}(x) + f(x)$.
- Similarly, $f(x_1+\epsilon_1, x_2, \ldots, x_n) \approx f(x_1, x_2, \ldots, x_n) + \epsilon_1 \frac{\partial}{\partial x_1} f(x_1, x_2, \ldots, x_n)$
  - If we repeat this, $f(x_1+\epsilon_1, x_2+\epsilon_2, \ldots, x_n+\epsilon_n) \approx f(x_1, x_2, \ldots, x_n) + \sum_i \epsilon_i \frac{\partial}{\partial x_i} f(x_1, x_2, \ldots, x_n)$.
  - Let $\boldsymbol{\epsilon} = [\epsilon_1, \ldots \epsilon_n]^\top$ and $\nabla_{\mathbf{x}} f = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right]^\top$, then $\boldsymbol{f(\mathbf{x}+\epsilon) \approx f(\mathbf{x}) + \epsilon^T \nabla_{\mathbf{x}} f(\mathbf{x})}$.
  - What happens if we consider $f(\cdot)$ as our objective function $L(\cdot)$, and $\|\boldsymbol{\epsilon}\| = 1$?

    $\Rightarrow$ $\boldsymbol{L(\theta+\epsilon) \approx L(\theta) + \epsilon^T \nabla_\theta L(\theta) = L(\theta) + \|\nabla_\theta L(\theta)\| cos(\alpha)}$

    $\Rightarrow$ Can we relate this with optimization?

Gradients, Backpropagation, Chain Rule, and Neural Network



$$\frac{\partial L}{\partial \boldsymbol{\theta}_{K-1}} = \frac{\partial L}{\partial \boldsymbol{f}_K} \frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{\theta}_{K-1}}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{K-2}} = \frac{\partial L}{\partial \boldsymbol{f}_K} \boxed{\frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}} \frac{\partial \boldsymbol{f}_{K-1}}{\partial \boldsymbol{\theta}_{K-2}}}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{K-3}} = \frac{\partial L}{\partial \boldsymbol{f}_K} \frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}} \boxed{\frac{\partial \boldsymbol{f}_{K-1}}{\partial \boldsymbol{f}_{K-2}} \frac{\partial \boldsymbol{f}_{K-2}}{\partial \boldsymbol{\theta}_{K-3}}}$$

$$\Rightarrow \quad \frac{\partial L}{\partial \boldsymbol{\theta}_i} = \frac{\partial L}{\partial \boldsymbol{f}_K} \frac{\partial \boldsymbol{f}_K}{\partial \boldsymbol{f}_{K-1}} \cdots \boxed{\frac{\partial \boldsymbol{f}_{i+2}}{\partial \boldsymbol{f}_{i+1}} \frac{\partial \boldsymbol{f}_{i+1}}{\partial \boldsymbol{\theta}_i}}$$

Any Questions?