EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-
Naturwissenschaftliche Fakultät

DNN supplementary slides

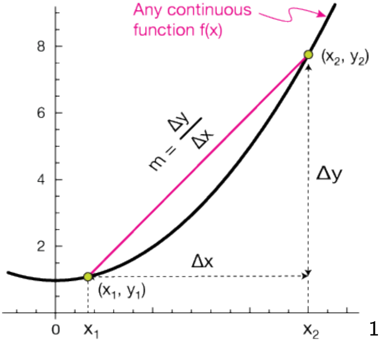
Kevin A. Laube, Maximus Mutschler

$\frac{d}{dx}$ - Derivatives

Derivative: slope in a point

$$\frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

$$\frac{df}{dx} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



[1]

[1] http://xaktly.com/Images/Mathematics/TheDerivative/DerivativeDefinition1.png

$\frac{d}{dx}$ - Known rules

- Product rule: $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$
- Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$
- Sum rule: $(f(x) + g(x))' = f'(x) + g'(x)$
- Chain rule: $(g(f(x)))' = (g \circ f)' = g'(f(x))f'(x)$

$\frac{d}{dx}$ - Example: Chain rule

Chain rule: $(g(f(x)))' = (g \circ f)' = g'(f(x))f'(x)$

- $h(x) = (x^2 + 2)^3 = g(f(x))$

- $g(x) = x^3 \qquad g'(x) = 3x^2$
- $f(x) = x^2 + 2 \qquad f'(x) = 2x$

- $h'(x) = 3(f(x))^2 f'(x) = 3(x^2 + 2)^2 \cdot 2x$

$\frac{\partial}{\partial x}$ - Partial derivatives (multivariate)

Derivative of functions with multiple variables: gradient.
Collect all partial derivatives in a row vector:

$$\mathbf{x} \in \mathbb{R}^n, \qquad f : \mathbb{R}^n \to \mathbb{R}$$

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

$\frac{\partial}{\partial x}$ - Example: Chain rule

Chain rule: $(g(f(x)))' = (g \circ f)' = g'(f(x))f'(x)$

- $h(x, y) = (2x + y^2)^2$
- $\frac{\partial h(x,y)}{\partial x} = 2(2x + y^2) \cdot \frac{\partial}{\partial x}(2x + y^2) = 2(2x + y^2) \cdot 2$
- $\frac{\partial h(x,y)}{\partial y} = 2(2x + y^2) \cdot \frac{\partial}{\partial y}(2x + y^2) = 2(2x + y^2) \cdot 2y$

$$\nabla_{\mathbf{x}} h = \frac{dh}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial h(x,y)}{\partial x} & \frac{\partial h(x,y)}{\partial y} \end{bmatrix} \in \mathbb{R}^{1 \times 2}$$

$\frac{\partial}{\partial x}$ - Rules

The rules from univariate functions apply, but now the order is important. Vector/Matrix multiplication is not commutative.

- Product rule: $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x})g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} g(\mathbf{x}) + f(\mathbf{x}) \frac{\partial g}{\partial \mathbf{x}}$
- Sum rule: $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}}$
- Chain rule: $\frac{\partial}{\partial \mathbf{x}}(g \circ f)(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(g(f(\mathbf{x}))) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial \mathbf{x}}$

$\frac{\partial}{\partial x}$ - Example 2: Chain rule

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Chain rule: $\frac{\partial}{\partial \mathbf{x}}(g \circ f)(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(g(f(\mathbf{x}))) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \mathbf{x}}$

$h(f_1, f_2), \quad f_1(x_1, x_2), \quad f_2(x_1, x_2)$ The inner functions also have multiple inputs now

$$\frac{\partial h}{\partial x_1} = \begin{bmatrix} \frac{\partial h}{\partial f_1} & \frac{\partial h}{\partial f_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \\ \frac{\partial f_2}{\partial x_1} \end{bmatrix} = \sum_i \frac{\partial h}{\partial f_i}\frac{\partial f_i}{\partial x_1} = \frac{\partial h}{\partial f_1}\frac{\partial f_1}{\partial x_1} + \frac{\partial h}{\partial f_2}\frac{\partial f_2}{\partial x_1}$$

$$\frac{\partial h}{\partial x_2} = \begin{bmatrix} \frac{\partial h}{\partial f_1} & \frac{\partial h}{\partial f_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \sum_i \frac{\partial h}{\partial f_i}\frac{\partial f_i}{\partial x_2} = \frac{\partial h}{\partial f_1}\frac{\partial f_1}{\partial x_2} + \frac{\partial h}{\partial f_2}\frac{\partial f_2}{\partial x_2}$$

$$\frac{\partial h}{\partial (x_1, x_2)} = \begin{bmatrix} \frac{\partial h}{\partial f_1} & \frac{\partial h}{\partial f_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} \end{bmatrix}$$

$\frac{\partial a}{\partial \mathbf{M}}$ - Scalar by Matrix derivatives

- For this lecture we will only need scalar by matrix derivatives.

  $\frac{\partial f(\mathbf{M})}{\partial \mathbf{M}} \qquad \mathbf{M} \in \mathbb{R}^{n,m}, \qquad f : \mathbb{R}^{n,m} \to \mathbb{R}$

- To be consistent with automatic differentiation frameworks like TensorFlow or PyTorch:

  $\dim(\frac{\partial f(\mathbf{M})}{\partial \mathbf{M}}) = \dim(M)$

$\frac{\partial a}{\partial \mathbf{M}}$ - Chain rule, indexwise differentiation

- Chain rule:
  $\frac{\partial a}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \mathbf{x}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\partial a}{\partial M_{i,j}} \frac{\partial M_{i,j}}{\partial \mathbf{x}}$
- Usually, we first look at derivatives by index and than reconstruct a matrix expression from the results:

$$(\nabla_{\mathbf{x}} f(\mathbf{M}))_{i,j} = \nabla_{x_{i,j}} f(\mathbf{M}) = \frac{\partial f(\mathbf{M})}{\partial M_{i,j}}$$

To simplify expressions we use the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Which has useful properties:

- $\sum_j \delta_{ij} x_j = x_i$
- $\frac{\partial}{\partial A_{ij}} A_{kl} = \delta_{ik}\delta_{jl}$
- $\delta_{ij} = \delta_{0,j-i}$

$\frac{\partial a}{\partial \mathbf{M}}$ - Example 1

We are often interested in derivatives when matrices are involved:
$\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x} \in \mathbb{R} \quad \boldsymbol{x} \in \mathbb{R}^n \quad \boldsymbol{M} \in \mathbb{R}^{n \times n}$

$(\nabla_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x})_i$

$= \frac{\partial}{\partial x_i} \sum_k \sum_l x_k x_l M_{k,l}$

$= \sum_k \sum_l \frac{\partial}{\partial x_i} x_k x_l M_{k,l}$ sum rule

$= \sum_k \sum_l \frac{\partial}{\partial x_i} \delta_{i,k} x_l M_{k,l} + \sum_k \sum_l \frac{\partial}{\partial x_i} x_k \delta_{i,l} M_{k,l}$ product rule

$= \sum_l x_l M_{i,l} + \sum_k x_k M_{k,i}$ kronecker delta property

$= 2 \sum_k x_k M_{k,i}$ requires M to be symmetric

$= 2\boldsymbol{x}^T \boldsymbol{M}_{:,i}$

$\nabla_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}^T = 2\boldsymbol{x}\boldsymbol{M}$

$\frac{\partial a}{\partial \mathbf{M}}$ - Example 2

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

$\nabla_{\boldsymbol{A}} tr(\boldsymbol{AB})$ $\quad$ $tr(\boldsymbol{A}) = \sum_i A_{ii}$

$$(\nabla_{\boldsymbol{A}} tr(\boldsymbol{AB}))_{ij}$$

$$= \frac{\partial}{\partial A_{ij}} \left(\sum_k \sum_l A_{kl} B_{lk}\right)$$

$$= \sum_k \sum_l \frac{\partial}{\partial A_{ij}} A_{kl} B_{lk}$$ sum rule

$$= \sum_k \sum_l \delta_{ik} \delta_{jl} B_{lk}$$ kronecker delta property

$$= B_{ji}$$

$$= (\boldsymbol{B}^T)_{ij}$$