

Mathematics for Machine Learning

Multivariate Calculus

Formula sheet

Dr Samuel J. Cooper
Prof. David Dye
Dr A. Freddie Page

Definition of a derivative

$$f'(x) = \frac{df(x)}{dx} = \lim_{\Delta x \rightarrow 0} \left(\frac{f(x + \Delta x) - f(x)}{\Delta x} \right)$$

Time saving rules

- *Sum Rule:*

$$\frac{d}{dx} (f(x) + g(x)) = \frac{d}{dx} (f(x)) + \frac{d}{dx} (g(x))$$

- *Power Rule:*

$$\begin{aligned} \text{Given } f(x) &= ax^b, \\ \text{then } f'(x) &= abx^{(b-1)} \end{aligned}$$

- *Product Rule:*

$$\begin{aligned} \text{Given } A(x) &= f(x)g(x), \\ \text{then } A'(x) &= f'(x)g(x) + f(x)g'(x) \end{aligned}$$

- *Chain Rule:*

$$\begin{aligned} \text{Given } h &= h(p) \text{ and } p = p(m), \\ \text{then } \frac{dh}{dm} &= \frac{dh}{dp} \times \frac{dp}{dm} \end{aligned}$$

- *Total derivative:*

For the function $f(x, y, z, \dots)$, where each variable is a function of parameter t , the total derivative is

$$\frac{df}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} + \frac{\partial f}{\partial z} \frac{dz}{dt} + \dots$$

Derivatives of named functions

$$\frac{d}{dx} \left(\frac{1}{x} \right) = -\frac{1}{x^2}$$

$$\frac{d}{dx} (\sin(x)) = \cos(x)$$

$$\frac{d}{dx} (\cos(x)) = -\sin(x)$$

$$\frac{d}{dx} (\exp(x)) = \exp(x)$$

Derivative structures

$$\text{Given } f = f(x, y, z)$$

- *Jacobian:*

$$\mathbf{J}_f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right]$$

- *Hessian:*

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix}$$

Neural networks

- *Activation function:*

$$\sigma(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$
$$\frac{d}{dx}(\sigma(x)) = \frac{1}{\cosh^2(x)} = \frac{4}{(e^x + e^{-x})^2}$$

Taylor Series

- *Univariate:*

$$f(x) = f(c) + f'(c)(x - c) + \frac{1}{2}f''(c)(x - c)^2 + \dots$$
$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!} (x - c)^n$$

- *Multivariate:*

$$f(\mathbf{x}) = f(\mathbf{c}) + \mathbf{J}_f(\mathbf{c})(\mathbf{x} - \mathbf{c}) +$$
$$\frac{1}{2}(\mathbf{x} - \mathbf{c})^t \mathbf{H}_f(\mathbf{c})(\mathbf{x} - \mathbf{c}) + \dots$$

Optimization and Vector Calculus

- *Newton-Raphson:*

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

- *Grad:*

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix}$$

- *Directional Gradient:*

$$\nabla f \cdot \hat{r}$$

- *Gradient Descent:*

$$s_{n+1} = s_n - \gamma \nabla f$$

- *Lagrange Multipliers λ :*

$$\nabla f = \lambda \nabla g$$

- *Least Squares - χ^2 minimization:*

$$\chi^2 = \sum_i^n \frac{(y_i - y(x_i; a_k))^2}{\sigma_i}$$

$$\text{criterion: } \nabla \chi^2 = 0$$

$$a_{\text{next}} = a_{\text{cur}} - \gamma \nabla \chi^2$$
$$= a_{\text{cur}} + \gamma \sum_i^n \frac{(y_i - y(x_i; a_k))}{\sigma_i} \frac{\partial y}{\partial a_k}$$