Chain Rule Example - Jan 19, 2026

$$\sigma(x) = \frac{1}{1+e^{-x}} = \underbrace{(1+e^{-x})}_{g(x)}{}^{-1} \qquad g(x) = 1 + (e^{x})^{-1} = 1 + h(x)^{-1}$$

$$\frac{d\sigma}{dx} = \frac{d\sigma}{dg} \cdot \frac{dg}{dh} \cdot \frac{dh}{dx}$$

$$\frac{dh}{dx} = e^{x}, \quad \frac{dg}{dh} = 0 + (-h(x)^{-2}), \quad \frac{d\sigma}{dg} = -g(x)^{-2}$$

subbing back in:

$$\frac{d\sigma}{dx} = -g(x)^{-2} \cdot -h(x) \cdot e^{x}$$

$$= \cancel{-}(1+e^{-x})^{-2} \cdot \cancel{-}(e^{x})^{-2\,!} \cdot \cancel{e}^{x}$$

$$= (1+e^{-x})^{-2} \cdot e^{-x} = \frac{e^{-x}}{(1+e^{-x})^{2}} \;/\!/$$

→ commonly written as $\sigma(x)(1-\sigma(x))$

Linear Regression - Jan 21

$\frac{\partial MSE}{\partial \theta_0} = 0$ to find θ_0 that minimizes MSE

$$\frac{\partial}{\partial \theta_0}\left(\frac{1}{m}\sum_i \underbrace{(\theta_0 + \theta_1 x_i - y_i)}_{g}^2\right) \qquad \frac{\partial g}{\partial \theta_0} = 1 \;,\; \frac{\partial g}{\partial \theta_1} = x_i$$

$$= \frac{1}{m}\sum_i 2g^1 \cdot \frac{\partial g}{\partial \theta_0} = \frac{2}{m}\sum_i (\theta_0 + \theta_1 x_i - y_i)$$

$$0 = \cancel{\frac{2}{m}}\left(\sum_i \theta_0 + \sum_i (\theta_1 x_i - y_i)\right)$$

$$\hookrightarrow \theta_0 = \frac{\sum_i y_i}{m} - \theta_1 \frac{\sum x_i}{m} = \mu_y - \theta_1 \mu_x \;//$$

$$\frac{\partial}{\partial \theta_1} = \frac{1}{m}\sum_i 2g \cdot \frac{\partial g}{\partial \theta_1} = \cancel{\frac{2}{m}}\left(\sum_i (\theta_0 + \theta_1 x_i - y_i) \cdot x_i\right.$$

$$0 = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2 - \sum x_i y_i$$

$$= (\mu_y - \theta_1 \mu_x)\sum_i x_i + \theta_1 \sum_i x_i^2 - \sum_i x_i y_i$$

$$= \mu_y \sum x_i - \sum x_i y_i - \theta_1\left(\mu_x \sum x_i - \sum x_i^2\right)$$

$$\therefore \theta_1 = \frac{\mu_y \sum x_i - \sum x_i y_i}{\mu_x \sum x_i - \sum x_i^2} \;//$$

Matrix Form

∇_θ : just like scalar, $\frac{d}{dx}(\vec{a}\vec{x}) = \vec{a}$, $\frac{d}{dx}(\vec{a}\vec{x}^T\vec{x}) = 2\vec{a}\vec{x}$, etc

First, expand

$$\frac{1}{m}(X\vec{\theta} - \vec{y})^T(X\vec{\theta} - \vec{y})$$

$(X\vec{\theta})^T = \vec{\theta}^T X^T$

$$\frac{1}{m}(\vec{\theta}^T X^T - \vec{y}^T)(X\vec{\theta} - \vec{y}) = \vec{\theta}^T X^T X\vec{\theta} \underbrace{- \vec{\theta}^T X^T \vec{y} - \vec{y}^T X\vec{\theta}}_{\text{can group, both end up as vectors}} + \vec{y}^T\vec{y}$$

$$= \vec{\theta}^T X^T X\vec{\theta} - 2\vec{\theta}^T X^T \vec{y} + \cancel{\vec{y}^T\vec{y}}$$

$\frac{\partial}{\partial\theta} = 0$

$\vec{\theta}^T\vec{\theta} = \sum_i \theta_i^2$, $\frac{\partial}{\partial\theta} = 2X^T X\vec{\theta}$

$$\nabla_\theta = \frac{1}{m}(2X^T X\vec{\theta} - 2X^T\vec{y}) = \frac{2X^T}{m}(X\vec{\theta} - \vec{y})$$

The gradient to descend

↳ can set $= 0$ to get:

$$\vec{\theta} = (X^T X)^{-1}(X^T\vec{y})$$

(closed form solution)