Gradients

Deep Networks: Large Nested Functions

$$f(x) = g_2(g_1(x))$$

$$y = g_1(x)$$

$$f(x) = z = g_2(y)$$

Training Deep Networks: Compute Partial Derivatives

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x})$$

Gradient: Partial Derivative of a Scalar Function

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\begin{aligned} \nabla_{\mathbf{x}} f(\mathbf{x}) &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \\ &= \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \end{aligned}$$

Jacobian: Partial Derivative of a Vector-Valued Function

$$J_f = \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \nabla_{\mathbf{x}} f_1(\mathbf{x}) \\ \nabla_{\mathbf{x}} f_2(\mathbf{x}) \\ \cdots \\ \nabla_{\mathbf{x}} f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \frac{\partial f_m(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Size of Gradients

- Gradients of $f : \mathbb{R}^n \to \mathbb{R}$

  size-n row vectors

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

- Partial derivatives of $f : \mathbb{R} \to \mathbb{R}^m$

  size-m column vectors

$$\frac{\partial}{\partial x} f(x) = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

- Jacobians of functions $f : \mathbb{R}^n \to \mathbb{R}^m$

  $m \times n$ matrices

$$J_f = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Chain Rule

$g_1 : \mathbb{R}^n \to \mathbb{R}^m$
$g_2 : \mathbb{R}^m \to \mathbb{R}^k$

The Jacobian of $f(\mathbf{x}) = g_2(g_1(\mathbf{x}))$:

$$\begin{aligned} J_f &= \nabla_{\mathbf{x}} g_2(g_1(\mathbf{x})) \\ &= \underbrace{\nabla_{\mathbf{y}} g_2(\mathbf{y})}_{J_{g_2} \in \mathbb{R}^{k \times m}} \underbrace{\nabla_{\mathbf{x}} g_1(\mathbf{x})}_{J_{g_1} \in \mathbb{R}^{m \times n}} \quad \text{where} \quad \mathbf{y} = g_2(\mathbf{x}) \end{aligned}$$

Gradients - TL;DR

Gradients are row vectors

Chain rule: gradient of a nested function is the (matrix) product of the gradients of its individual functions