CHAIN RULE

Recall this fundamental result.

Given a differentiable function $f: \mathbb{R}^n \to \mathbb{R}^m$, its differential is a linear map $\frac{\partial f}{\partial x}(x_0): \mathbb{R}^n \to \mathbb{R}^m$.

It can be represented by a matrix $\frac{\partial f}{\partial x}(x_0) \in \mathbb{R}^{m \times n}$

$$\frac{\partial f}{\partial x}(x_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

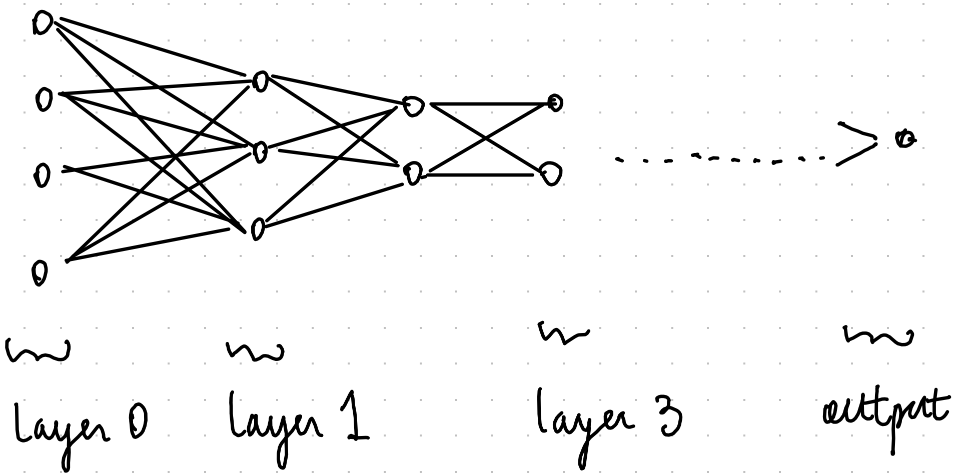
Given two differentiable functions

$$f: \mathbb{R}^n \to \mathbb{R}^m, \quad g: \mathbb{R}^m \to \mathbb{R}^k$$
$$x \mapsto f(x) \qquad\qquad y \mapsto g(y)$$

then $g(f(x))$ is differentiable and

$$\frac{\partial}{\partial x}(g \circ f)(x_0) = \frac{\partial g}{\partial y}(f(x_0)) \frac{\partial f}{\partial x}(x_0) \in \mathbb{R}^{k \times n}$$

BACKPROPAGATION



$x^0 \in \mathbb{R}^{N \times M^0}$

$y^0 = y^0(x^0; W^1) \in \mathbb{R}^{N \times M^1}$

$x^1 = y^0$ new input for next layer

$\vdots$

$y^l = y^l(x^l; W^l) \in \mathbb{R}^{N \times M^l}$

$\vdots$

$y^K = y(x^K; W^K)$

Loss is computed on output $L(y^K)$.

<u>Aim</u>: Compute

$$\nabla_{W^\ell} L$$

i.e., the gradient of the loss with respect to all parameters.
This is needed for optimization algorithms like gradient descent.

We first compute

$$\frac{\partial L}{\partial y^K} = \text{differential of cost with respect to output of last layer}$$

<u>Ex</u>: if output is $y^K \in \mathbb{R}^{N \times M_{out}}$, then

$$\frac{\partial L}{\partial y^K} \in \mathbb{R}^{M_{out} \times N}$$

Then we need to apply the chain rule.

Assume that we have computed $\frac{\partial L}{\partial y^\ell}$.
Then

$$\frac{\partial L}{\partial W^\ell} = \text{"}\frac{\partial L}{\partial y^\ell} \cdot \frac{\partial y^\ell}{\partial W^\ell}\text{"}, \quad \frac{\partial L}{\partial x^\ell} = \text{"}\frac{\partial L}{\partial y^\ell} \cdot \frac{\partial y^\ell}{\partial x^\ell}\text{"}$$

The quotation marks mean:
one must pay attention in writing the multiplication since the differential can be tensors and everything must be dimensionally consistent.