

# General rules for backprop:

input — module — output

\* How does output change w.r.t module  $\times$  signal from output  $\leftarrow$  how module changes w.r.t signal

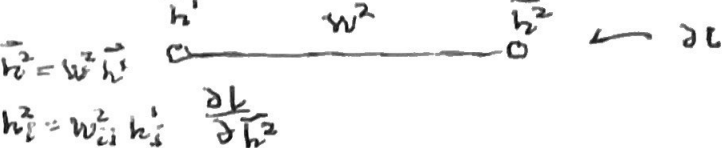
\* How does output change w.r.t input  $\times$  signal from output  $\leftarrow$  how signal propagates backwards

\* Note that the "multiplication" rule is just a rule. For general interaction, it's a personal responsibility to determine the order of multiplication for shape compatibility

## Rules for shape

Let  $s$  be a scalar,  $\vec{v} \in \mathbb{R}^{k \times 1}$ ,  $\vec{w} \in \mathbb{R}^{m \times 1}$ ,  $\vec{x} \in \mathbb{R}^{n \times 1}$ , and  $W \in \mathbb{R}^{m \times n}$   
then  $\frac{\partial s}{\partial v} \in \mathbb{R}^{1 \times k}$ ,  $\frac{\partial s}{\partial \vec{w}} \in \mathbb{R}^{m \times 1}$ , and  $\frac{\partial s}{\partial W} \in \mathbb{R}^{m \times n}$

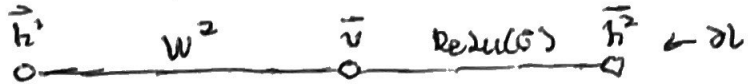
## Dealing with pure linear modules



$$\frac{\partial L}{\partial W^2} = \frac{\partial \vec{h}_2}{\partial W^2} \frac{\partial L}{\partial \vec{h}_2} \rightarrow \frac{\partial L}{\partial W_{21}^2} = \frac{\partial L}{\partial h_2^2} \times \frac{\partial h_2^2}{\partial W_{21}^2} = \frac{\partial L}{\partial h_2^2} \cdot h_1^1 = \left( \frac{\partial L}{\partial \vec{h}_2} \right)^T (\vec{h}_1)^T$$

$$\frac{\partial L}{\partial \vec{h}_1} = \frac{\partial \vec{h}_2}{\partial \vec{h}_1} \frac{\partial L}{\partial \vec{h}_2} = \frac{\partial L}{\partial \vec{h}_2} W^2 \quad \text{"switch for compatibility"}$$

## Dealing with functional and linear modules



\* Use  $\vec{v}$  as an intermediary variable

$$\begin{aligned} \vec{v} &= W^2 \vec{h}_1 \\ \vec{h}_2 &= \text{Relu}(\vec{v}) \\ v_i &= W_{21}^2 h_1^1 \\ h_2^2 &= \text{Relu}(W_{21}^2 h_1^1) \\ h_1^1 &= \text{Relu}(v_i) \\ \frac{\partial h_2^2}{\partial v_i} &= \delta_{21} \mathbb{1}(v_i > 0) \end{aligned}$$

ensure row vector shape

$$\begin{aligned} \frac{\partial L}{\partial \vec{v}} &= \frac{\partial \vec{h}_2}{\partial \vec{v}} \frac{\partial L}{\partial \vec{h}_2} = \text{Relu}'(\vec{v}) \frac{\partial L}{\partial \vec{h}_2} \\ \rightarrow \frac{\partial L}{\partial v_k} &= \frac{\partial h_2^2}{\partial v_k} \frac{\partial L}{\partial h_2^2} = \frac{\partial L}{\partial h_2^2} \frac{\partial h_2^2}{\partial v_k} = \frac{\partial L}{\partial h_2^2} \text{Relu}'(\vec{v}) \end{aligned}$$

ensure matrix shape

$$\begin{aligned} \frac{\partial L}{\partial W^2} &= \frac{\partial L}{\partial \vec{v}} \frac{\partial \vec{v}}{\partial W^2} \rightarrow \frac{\partial L}{\partial v_i} \frac{\partial v_i}{\partial W_{21}^2} \rightarrow \frac{\partial L}{\partial v_i} \cdot h_1^1 \\ &= \left( \frac{\partial L}{\partial \vec{v}} \right)^T (\vec{h}_1)^T = \left( \frac{\partial L}{\partial h_2^2} \cdot \text{diag}(\mathbb{1}(\vec{v} > 0)) \right)^T (\vec{h}_1)^T = \text{diag}(\mathbb{1}(\vec{v} > 0)) \left( \frac{\partial L}{\partial h_2^2} \right)^T (\vec{h}_1)^T \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial \vec{h}_1} &= \frac{\partial L}{\partial \vec{v}} \frac{\partial \vec{v}}{\partial \vec{h}_1} = \frac{\partial L}{\partial v_i} \frac{\partial v_i}{\partial h_1^1} = \frac{\partial L}{\partial h_2^2} \frac{\partial h_2^2}{\partial v_i} \frac{\partial v_i}{\partial h_1^1} \sim (1 \times k) \times (k \times 1) \times (1 \times 1) \rightarrow 1 \times 1 \\ &= \frac{\partial L}{\partial h_2^2} \cdot \text{Relu}'(\vec{v}) W^2 = \frac{\partial L}{\partial h_2^2} \cdot \text{diag}(\mathbb{1}(\vec{v} > 0)) \cdot W^2 \end{aligned}$$

For linear modules (and functional modules)

signal from output  $\rightarrow$  how output changes w.r.t input  $=$  how signal propagates back

how output changes w.r.t module  $\rightarrow$  signal from output  $=$  how module changes w.r.t module

Bonus: updating bias

$$\vec{y} = W\vec{x} + \vec{b} \quad \vec{x} \quad \vec{y} \quad \leftarrow \partial L$$

$$y_i = w_{i3}x_3 + b_i \quad \frac{\partial L}{\partial y}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial b} \rightarrow \frac{\partial L}{\partial b_k} = \frac{\partial}{\partial y_i} \frac{\partial y_i}{\partial b_k} = \frac{\partial L}{\partial y_i} \delta_{ik}$$

- \* Note that bias follows a different multiplication ordering
- \* Best to use general rule and use shape compatibility to understand ordering