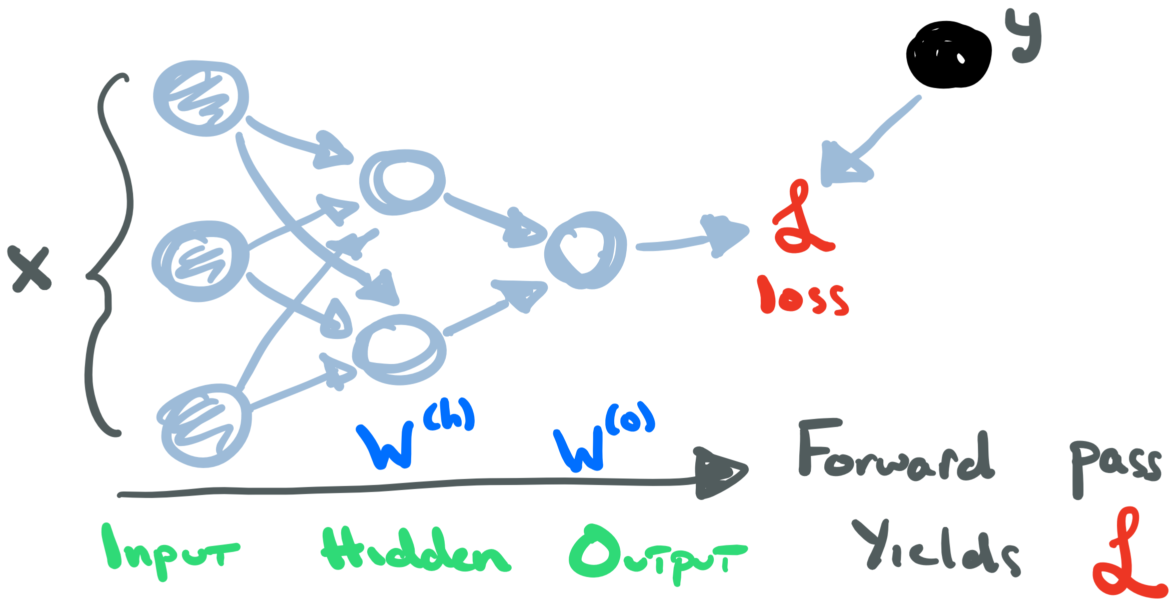
DS 4440 Backpropagation (1)

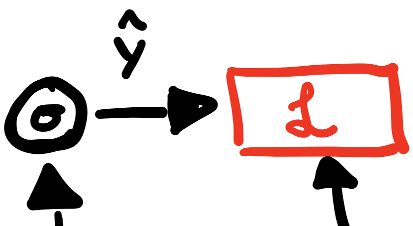
Last time: Computation graphs comprise layers to form Networks



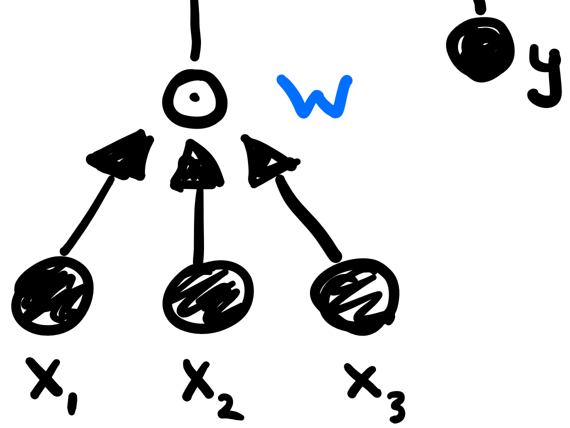
Today: How to adjust parameters in Computation graphs to minimize $\mathcal{L}$

Backprop. This extends SGD (L1; HW2).

Consider the graph for logistic regression.



$\mathcal{L}$oss $-P(y \mid x, \hat{w})$



$$\nabla_w \mathcal{L} = \underbrace{(y-\sigma(\hat{w}\cdot x))x}_{\text{for one instance } (x,y).}$$

The Recipe

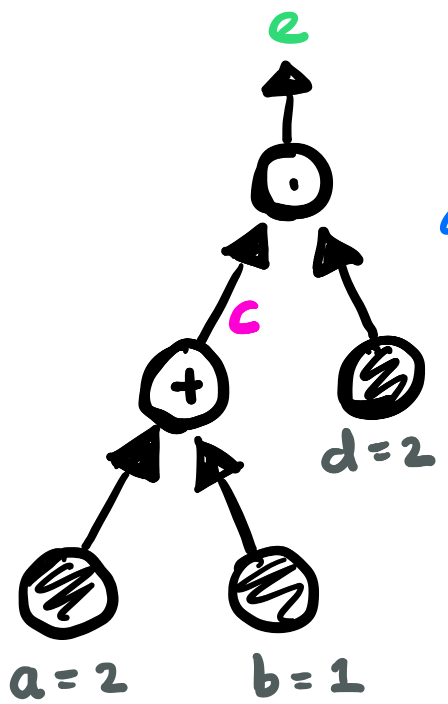
- Build graph (model) w/ parameters w
- Define (differentiable) $\mathcal{L}$
- Use $\nabla_w \mathcal{L}$ to find $\hat{w}$ via SGD

Problem Finding ∇_w is painful to do manually for big models.

Solution Use the computation graph to auto-diff via backprop.

Consider a simple scalar example

$$e = \underbrace{(a+b)}_{c}\cdot d$$



How is e affected by a?

Chain Rule!

$$\frac{d}{dx} f(\overbrace{g(x)}^{u}) = \frac{d}{du} f(u) \cdot \frac{d}{dx} u$$

$$\frac{\partial e}{\partial a} = \frac{\partial e}{\partial c} \cdot \frac{\partial c}{\partial a} = 2.$$

$$\frac{\partial}{\partial a}(a+b) = 1$$

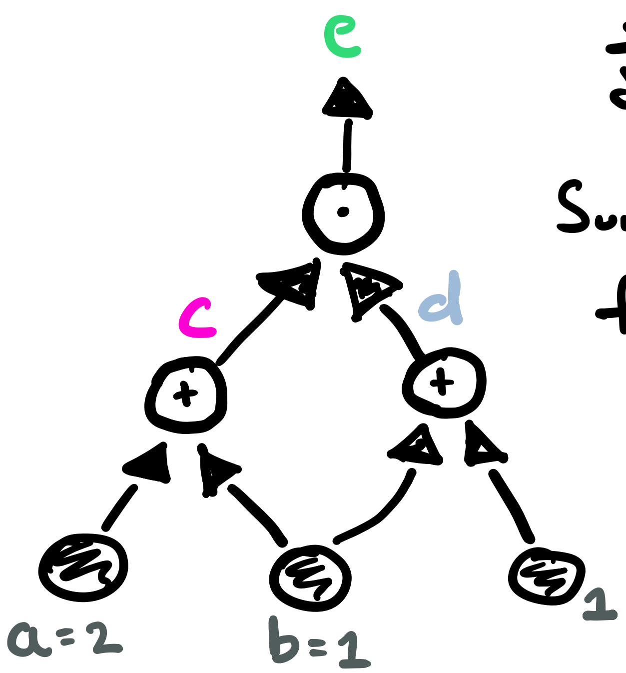
$$\frac{\partial}{\partial c} c \cdot d = d = 2.$$

$$\frac{\partial e}{\partial b} = \frac{\partial e}{\partial c} \frac{\partial c}{\partial b}$$

Re-use!

A slightly trickier example

$$e = (\underbrace{a+b}_{c})(\underbrace{b+1}_{d})$$



$$\frac{\partial}{\partial b} e = ?$$

Sum over multiple paths from $b \to e$

$$= \frac{\partial e}{\partial c} \frac{\partial c}{\partial b} + \frac{\partial e}{\partial d} \frac{\partial d}{\partial b}$$

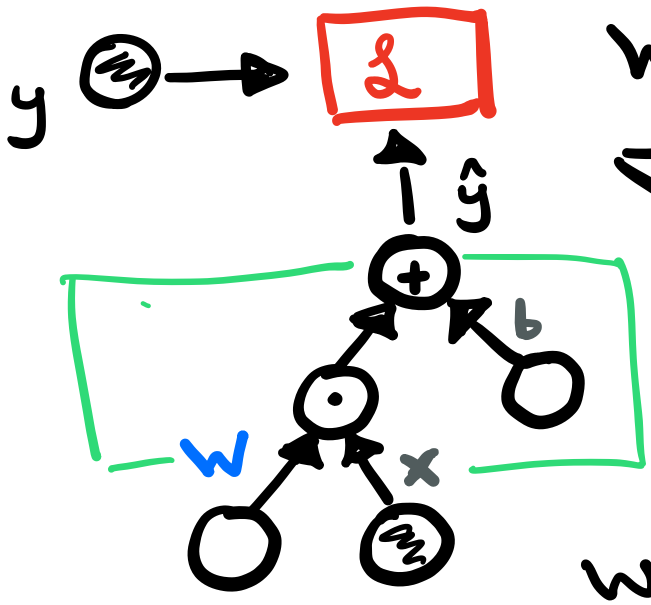
$$= (b+1) + (a+b)$$

$$= 5$$

Let's add a Loss. Assume regression.

$$\hat{y} = \hat{w}x + \hat{b}$$

$$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2 = (y - [\hat{w}x + \hat{b}])^2$$

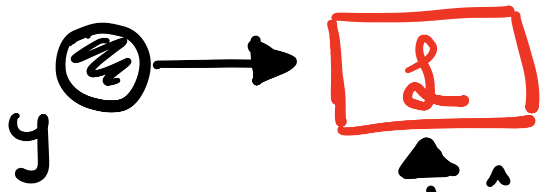


We previously derived ∇_w for this single layer model.

More interesting when we introduce hidden layers.

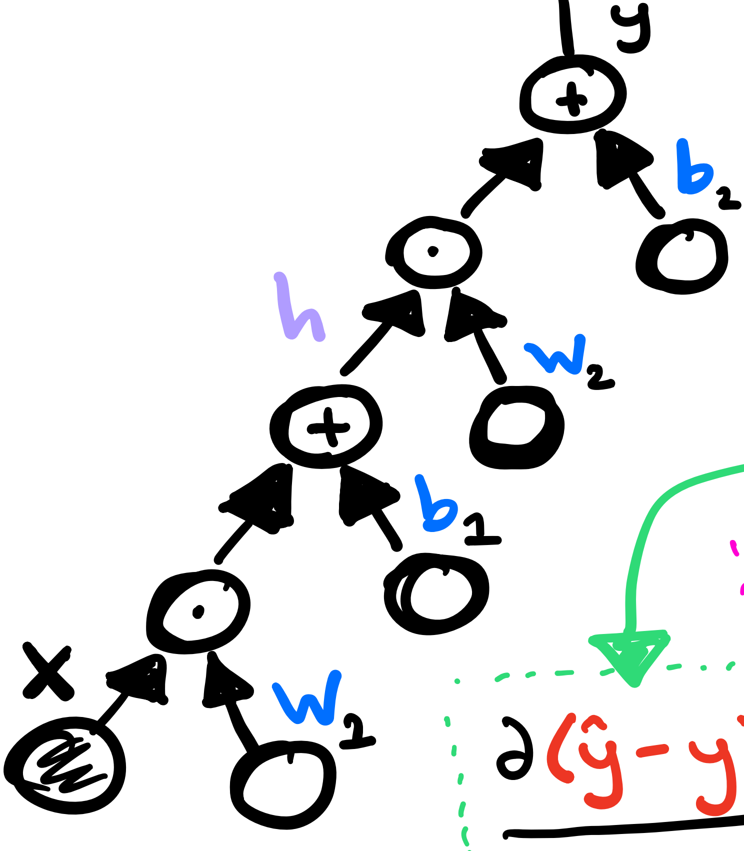
(See in-class exercise!)

$$\hat{y} = W_2 \cdot (\underbrace{W_1 \cdot x + b_1}_{h}) + b_2$$



$$\frac{\partial \mathcal{L}}{\partial} = ?$$

∂h



$$= \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h}$$

$$\frac{\partial}{\partial h} W_2 \cdot h + b_2 = W_2$$

$$\frac{\partial (\hat{y} - y)^2}{\partial \hat{y}} = 2(\hat{y} - y)$$

$$= 2(\hat{y} - y) W_2 .$$

—

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial h}{\partial W_1} \cdot \frac{\partial \mathcal{L}}{\partial h} = x \cdot 2(\hat{y} - y) W_2$$

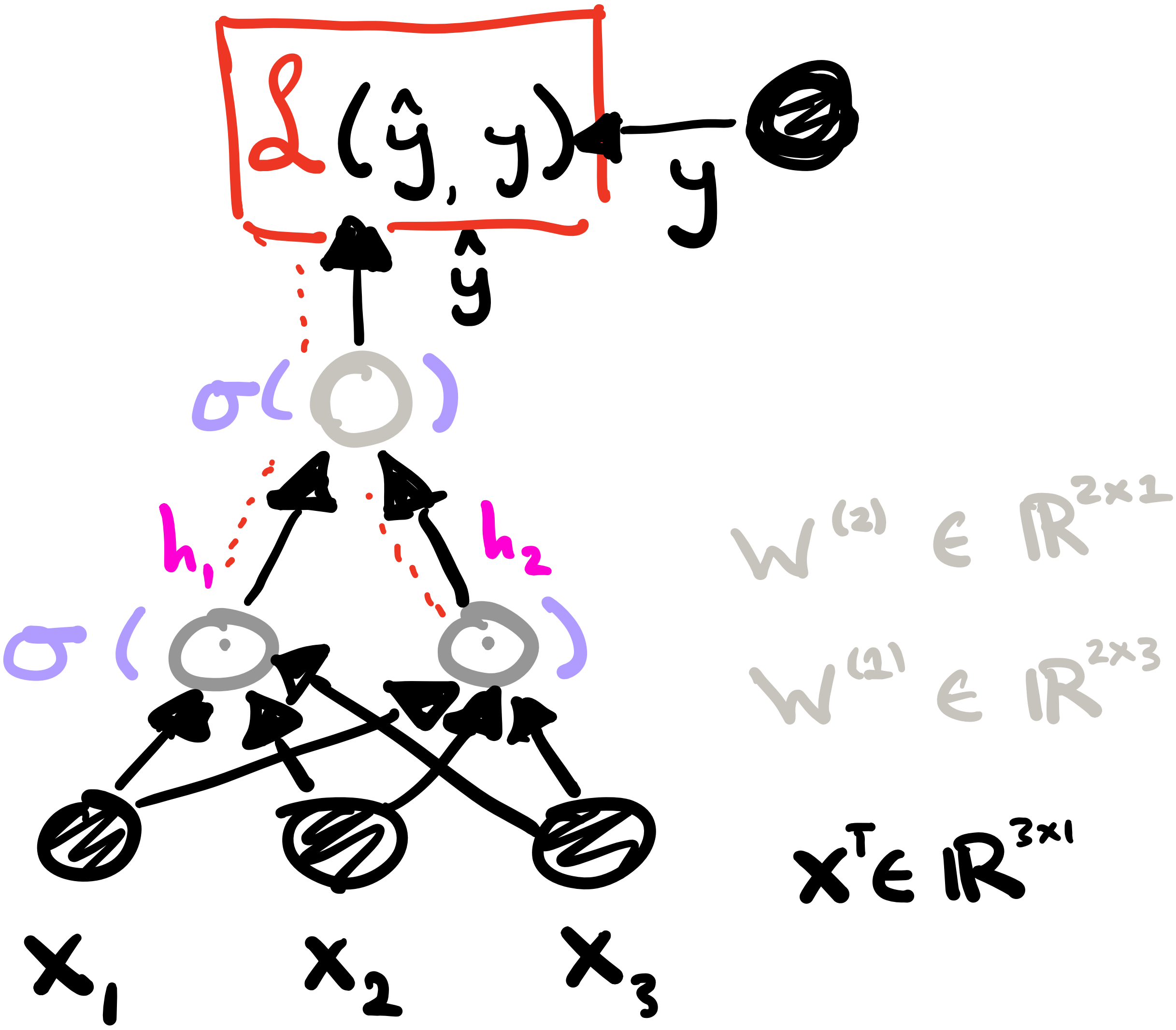
$$\frac{\partial}{\partial W_1} W_1 \cdot x + b_1 = x$$

—

$$\frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial h}{\partial b_1} \cdot \frac{\partial \mathcal{L}}{\partial h} = 2(\hat{y} - y) W_2$$

$= 1$ (See Colab!)

How about This NN ?



$\nabla_{W^{(2)}} \mathcal{L}$ we have already seen

$\nabla_{W^{(1)}} \mathcal{L} = ?$

$$= \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h} \cdot \frac{\partial h}{\partial W^{(1)}}$$ | Chain rule!

$$[1]$$

Scalar

$$\begin{bmatrix} \frac{\partial \hat{y}}{\partial h_1} & \frac{\partial \hat{y}}{\partial h_2} \end{bmatrix}$$

$$\delta_1 \quad \delta_2$$

$$h = \sigma(W^{(1)} \cdot x)$$

$$\nabla_{W_i^{(1)}} h \approx x$$

Ignoring element wise σ

$$\begin{bmatrix} \frac{\partial h_1}{\partial W^{(1)}} \\ \frac{\partial h_2}{\partial W^{(1)}} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_1 & x_2 & x_3 \end{bmatrix}$$

Local error at h_1 and h_2, after $\cdot$ with $\frac{\partial \mathcal{L}}{\partial \hat{y}}$.

$$= \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \begin{bmatrix} \frac{\partial \hat{y}}{\partial h_1} x_1 & \frac{\partial \hat{y}}{\partial h_1} x_2 & \frac{\partial \hat{y}}{\partial h_1} x_3 \\ \frac{\partial \hat{y}}{\partial h_2} x_1 & \frac{\partial \hat{y}}{\partial h_2} x_2 & \frac{\partial \hat{y}}{\partial h_2} x_3 \end{bmatrix}$$

$$= \begin{bmatrix} \delta_1 \vec{x} \\ \delta_2 \vec{x} \end{bmatrix}$$

More Next Time!