

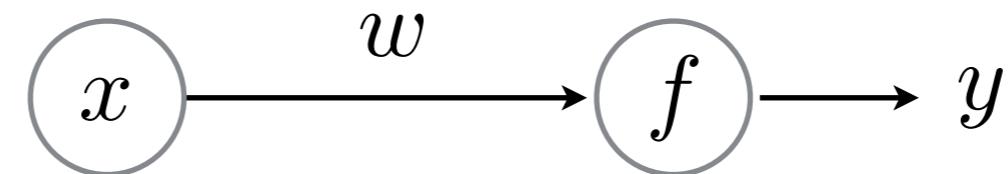


# Back-Propagation

16-385 Computer Vision (Kris Kitani)  
**Carnegie Mellon University**

back to the...

# World's Smallest Perceptron!



$$y = wx$$

(a.k.a. line equation, linear regression)

function of **ONE** parameter!

# Training the world's smallest perceptron

This is just gradient descent, that means...



this should be the gradient of the loss function

Now where does this come from?

$$\frac{d\mathcal{L}}{dw}$$

...is the rate at which **this** will change...

$$\mathcal{L} = \frac{1}{2}(y - \hat{y})^2$$

the loss function

... per unit change of **this**

$$y = w x$$

the weight parameter

Let's compute the derivative...

Compute the derivative

$$\begin{aligned}\frac{d\mathcal{L}}{dw} &= \frac{d}{dw} \left\{ \frac{1}{2} (y - \hat{y})^2 \right\} \\ &= -(y - \hat{y}) \frac{dwx}{dw} \\ &= -(y - \hat{y})x = \nabla w \quad \text{just shorthand}\end{aligned}$$

That means the weight update for **gradient descent** is:

$$\begin{aligned}w &= w - \nabla w \quad \text{move in direction of negative gradient} \\ &= w + (y - \hat{y})x\end{aligned}$$

# Gradient Descent (world's smallest perceptron)

For each sample

$$\{x_i, y_i\}$$

1. Predict

a. Forward pass

$$\hat{y} = wx_i$$

b. Compute Loss

$$\mathcal{L}_i = \frac{1}{2}(y_i - \hat{y})^2$$

2. Update

a. Back Propagation

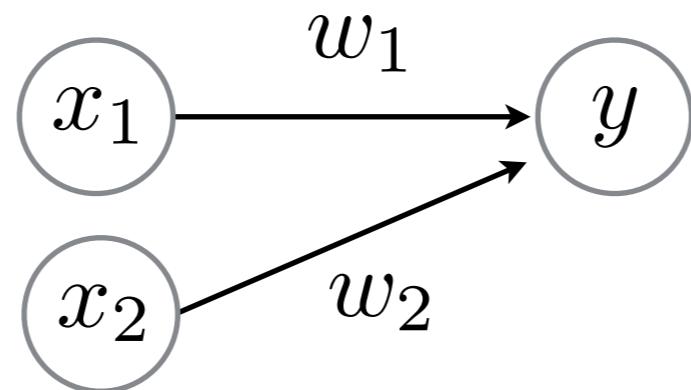
$$\frac{d\mathcal{L}_i}{dw} = -(y_i - \hat{y})x_i = \nabla w$$

b. Gradient update

$$w = w - \nabla w$$

Training the world's smallest perceptron

# world's (second) smallest perceptron!



function of **two** parameters!

# Gradient Descent

For each sample

$$\{x_i, y_i\}$$

1. Predict

a. Forward pass

b. Compute Loss

2. Update

a. Back Propagation

b. Gradient update

we just need to compute partial derivatives for this network

# Back-Propagation

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_1} &= \frac{\partial}{\partial w_1} \left\{ \frac{1}{2} (y - \hat{y})^2 \right\} & \frac{\partial \mathcal{L}}{\partial w_2} &= \frac{\partial}{\partial w_2} \left\{ \frac{1}{2} (y - \hat{y})^2 \right\} \\ &= -(y - \hat{y}) \frac{\partial \hat{y}}{\partial w_1} & &= -(y - \hat{y}) \frac{\partial \hat{y}}{\partial w_2} \\ &= -(y - \hat{y}) \frac{\partial \sum_i w_i x_i}{\partial w_1} & &= -(y - \hat{y}) \frac{\partial \sum_i w_i x_i}{\partial w_1} \\ &= -(y - \hat{y}) \frac{\partial w_1 x_1}{\partial w_1} & &= -(y - \hat{y}) \frac{\partial w_2 x_2}{\partial w_2} \\ &= -(y - \hat{y}) x_1 = \nabla w_1 & &= -(y - \hat{y}) x_2 = \nabla w_2\end{aligned}$$

*Why do we have partial derivatives now?*

# Back-Propagation

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial}{\partial w_1} \left\{ \frac{1}{2} (y - \hat{y})^2 \right\}$$

$$= -(y - \hat{y}) \frac{\partial \hat{y}}{\partial w_1}$$

$$= -(y - \hat{y}) \frac{\partial \sum_i w_i x_i}{\partial w_1}$$

$$= -(y - \hat{y}) \frac{\partial w_1 x_1}{\partial w_1}$$

$$= -(y - \hat{y}) x_1 = \nabla w_1$$

$$\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial}{\partial w_2} \left\{ \frac{1}{2} (y - \hat{y})^2 \right\}$$

$$= -(y - \hat{y}) \frac{\partial \hat{y}}{\partial w_2}$$

$$= -(y - \hat{y}) \frac{\partial \sum_i w_i x_i}{\partial w_1}$$

$$= -(y - \hat{y}) \frac{\partial w_2 x_2}{\partial w_2}$$

$$= -(y - \hat{y}) x_2 = \nabla w_2$$

## Gradient Update

$$w_1 = w_1 - \eta \nabla w_1$$

$$= w_1 + \eta (y - \hat{y}) x_1$$

$$w_2 = w_2 - \eta \nabla w_2$$

$$= w_2 + \eta (y - \hat{y}) x_2$$

# Gradient Descent

For each sample

$$\{x_i, y_i\}$$

(since gradients approximated from stochastic sample)

1. Predict

a. Forward pass  $\hat{y} = f_{\text{MLP}}(x_i; \theta)$

b. Compute Loss  $\mathcal{L}_i = \frac{1}{2}(y_i - \hat{y})$

2. Update

two BP lines now

$$\begin{aligned}\nabla w_{1i} &= -(y_i - \hat{y})x_{1i} \\ \nabla w_{2i} &= -(y_i - \hat{y})x_{2i}\end{aligned}$$

a. Back Propagation

$$\begin{aligned}w_{1i} &= w_{1i} + \eta(y - \hat{y})x_{1i} \\ w_{2i} &= w_{2i} + \eta(y - \hat{y})x_{2i}\end{aligned}$$

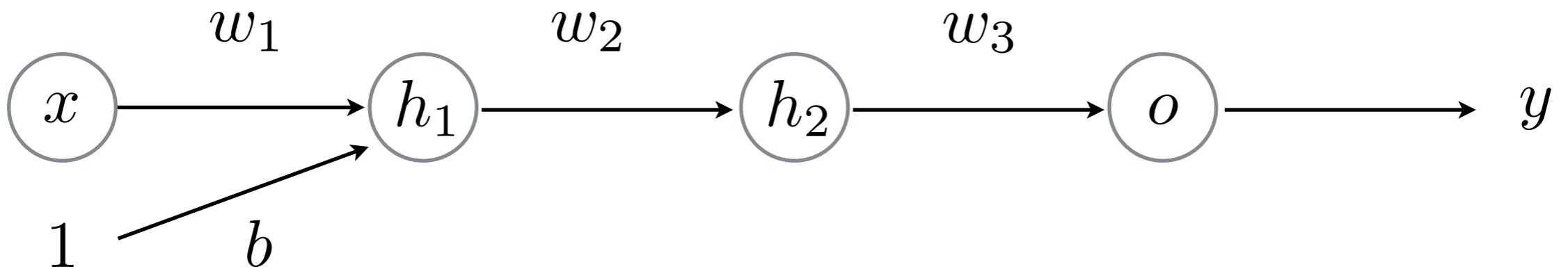
b. Gradient update

(adjustable step size)

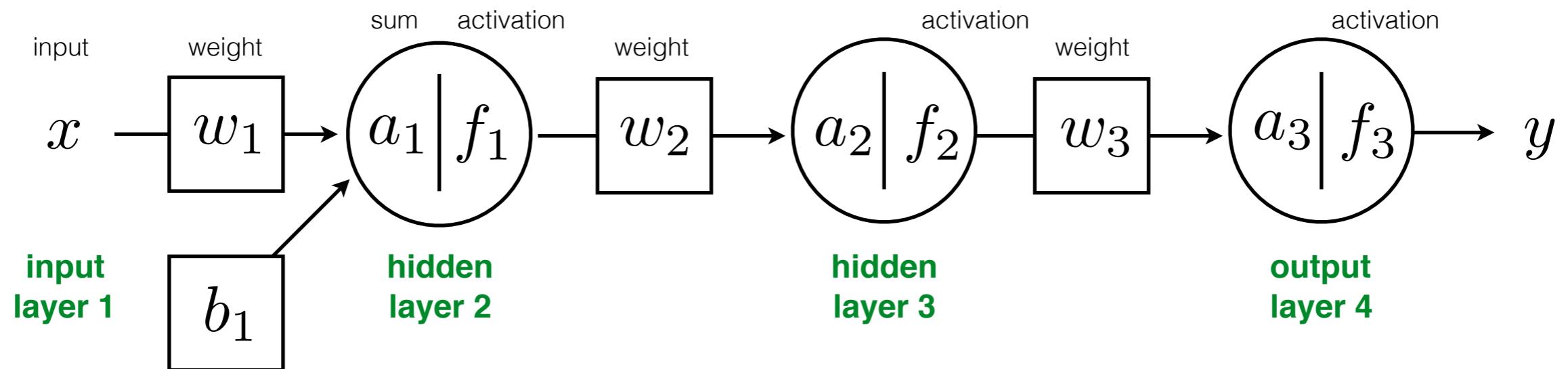


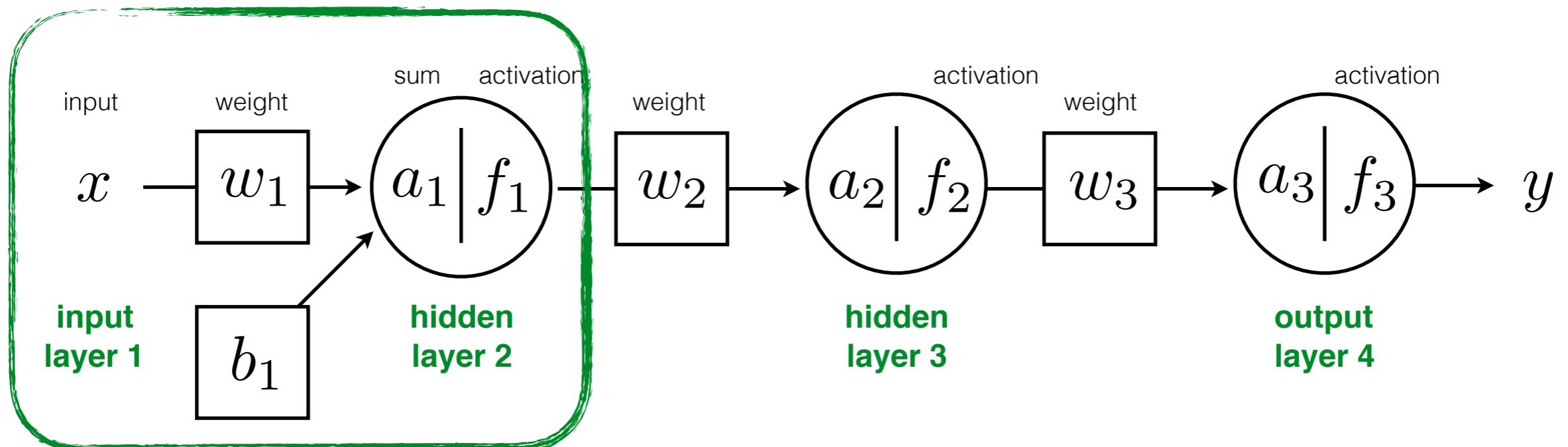
We haven't seen a lot of 'propagation' yet because our perceptrons only had one layer...

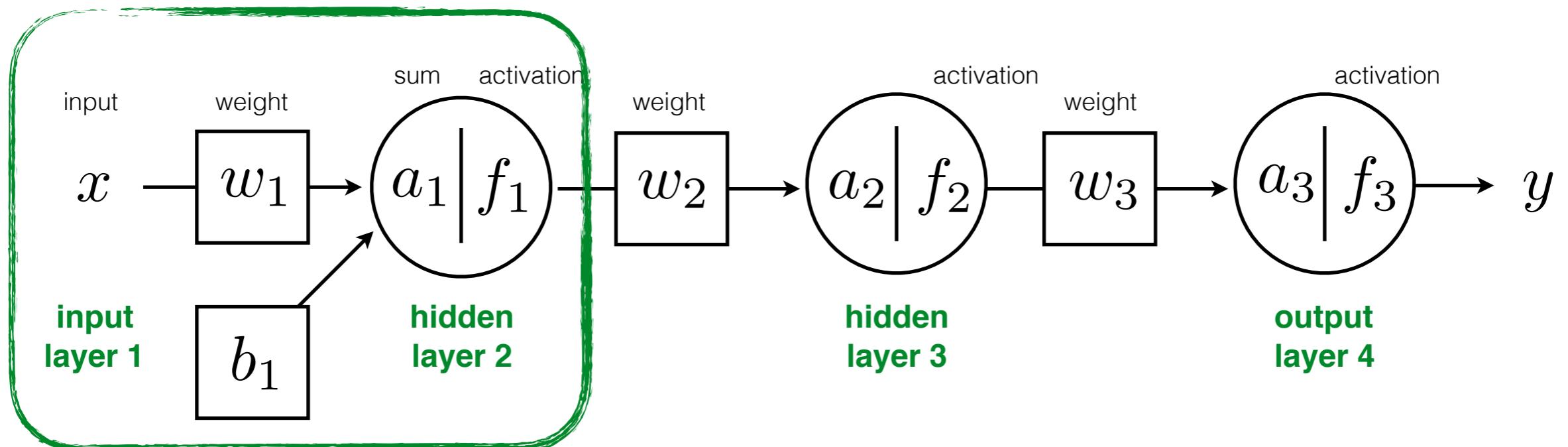
# multi-layer perceptron



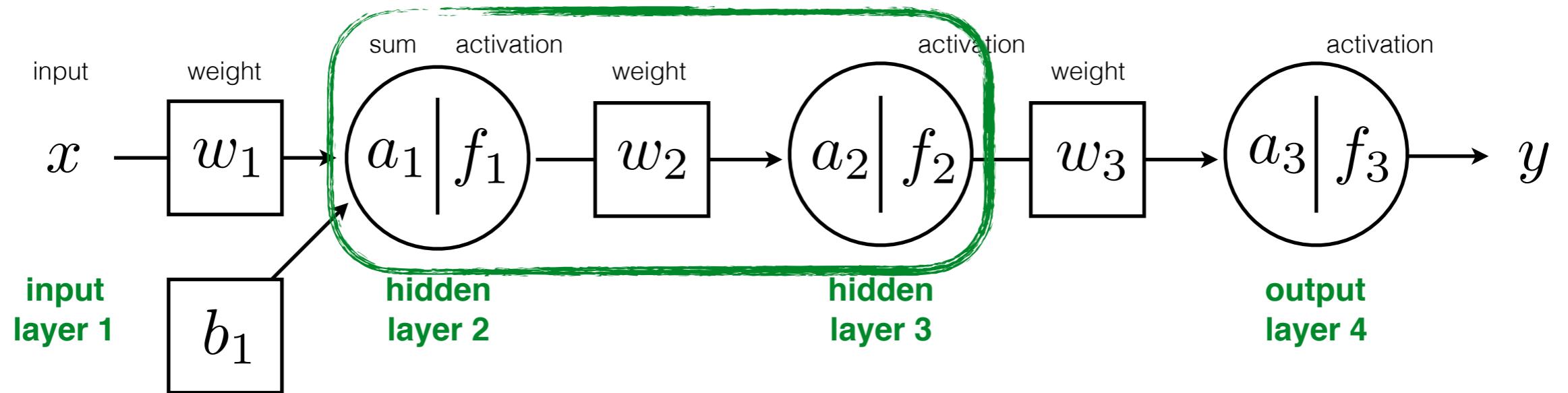
function of **FOUR** parameters and **FOUR** layers!



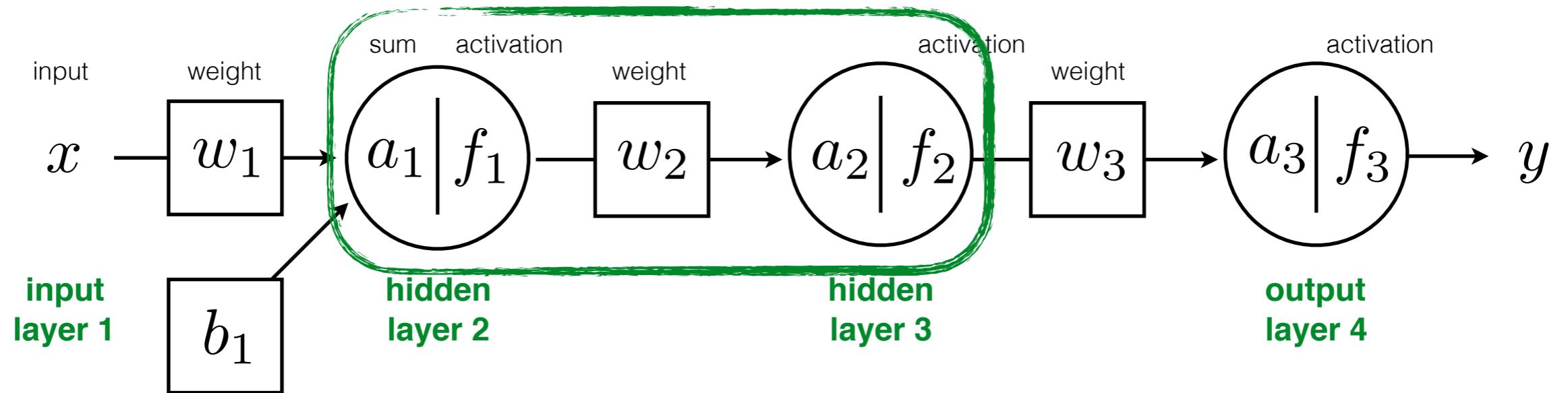




$$a_1 = w_1 \cdot x + b_1$$

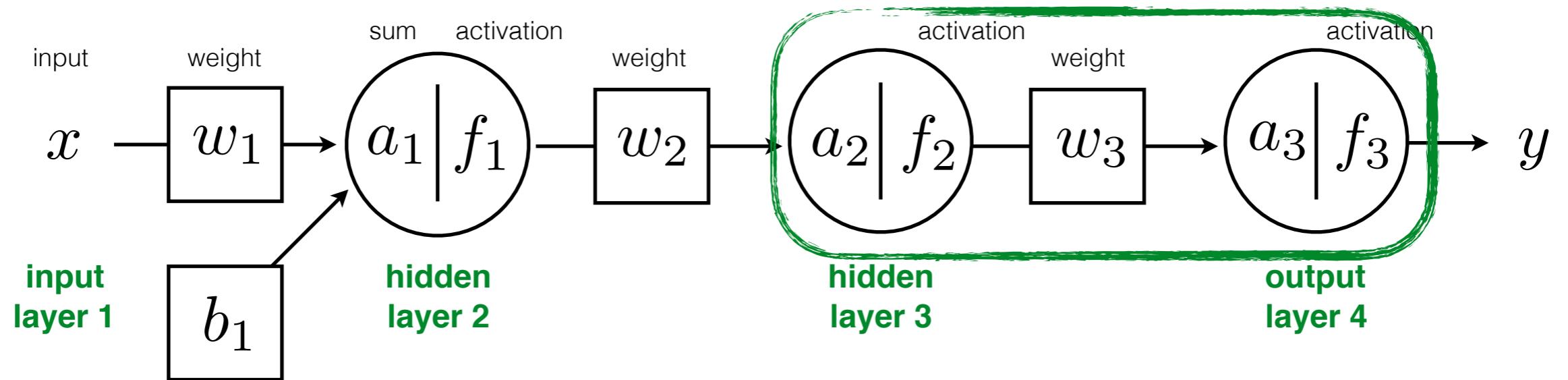


$$a_1 = w_1 \cdot x + b_1$$



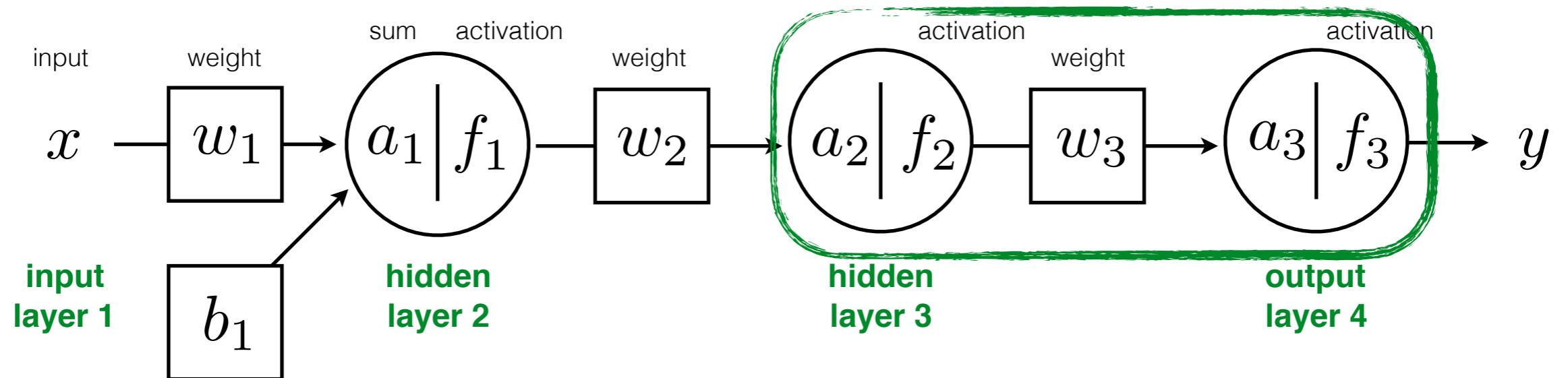
$$a_1 = w_1 \cdot x + b_1$$

$$a_2 = w_2 \cdot f_1(w_1 \cdot x + b_1)$$



$$a_1 = w_1 \cdot x + b_1$$

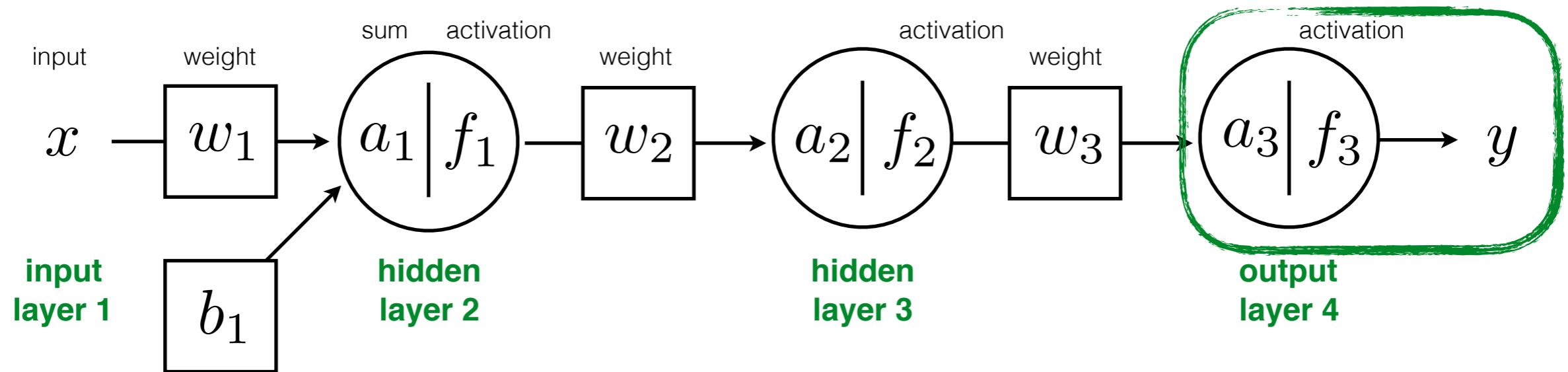
$$a_2 = w_2 \cdot f_1(w_1 \cdot x + b_1)$$



$$a_1 = w_1 \cdot x + b_1$$

$$a_2 = w_2 \cdot f_1(w_1 \cdot x + b_1)$$

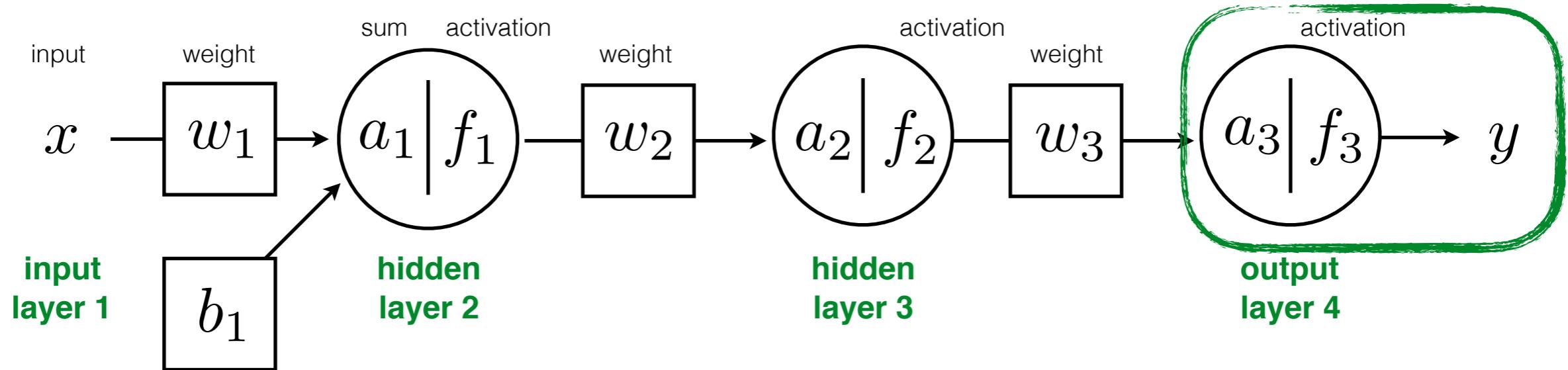
$$a_3 = w_3 \cdot f_2(w_2 \cdot f_1(w_1 \cdot x + b_1))$$



$$a_1 = w_1 \cdot x + b_1$$

$$a_2 = w_2 \cdot f_1(w_1 \cdot x + b_1)$$

$$a_3 = w_3 \cdot f_2(w_2 \cdot f_1(w_1 \cdot x + b_1))$$



$$a_1 = w_1 \cdot x + b_1$$

$$a_2 = w_2 \cdot f_1(w_1 \cdot x + b_1)$$

$$a_3 = w_3 \cdot f_2(w_2 \cdot f_1(w_1 \cdot x + b_1))$$

$$y = f_3(w_3 \cdot f_2(w_2 \cdot f_1(w_1 \cdot x + b_1)))$$

Entire network can be written out as one long equation

$$y = f_3(w_3 \cdot f_2(w_2 \cdot f_1(w_1 \cdot x + b_1)))$$

We need to train the network:

*What is known? What is unknown?*

Entire network can be written out as a long equation

$$y = f_3(w_3 \cdot f_2(w_2 \cdot f_1(w_1 \cdot x + b_1)))$$



We need to train the network:

*What is known? What is unknown?*

Entire network can be written out as a long equation

$$y = f_3(w_3 \cdot f_2(w_2 \cdot f_1(w_1 \cdot x + b_1)))$$

activation function  
sometimes has  
parameters

unknown

We need to train the network:

*What is known? What is unknown?*

# Learning an MLP

Given a set of samples and a MLP

$$\{x_i, y_i\}$$

$$y = f_{\text{MLP}}(x; \theta)$$

Estimate the parameters of the MLP

$$\theta = \{f, w, b\}$$

# Stochastic Gradient Descent

For each **random** sample  $\{x_i, y_i\}$

1. Predict

a. Forward pass

$$\hat{y} = f_{\text{MLP}}(x_i; \theta)$$

b. Compute Loss

2. Update

a. Back Propagation

$$\frac{\partial \mathcal{L}}{\partial \theta}$$

vector of parameter partial derivatives

b. Gradient update

$$\theta \leftarrow \theta - \eta \nabla \theta$$

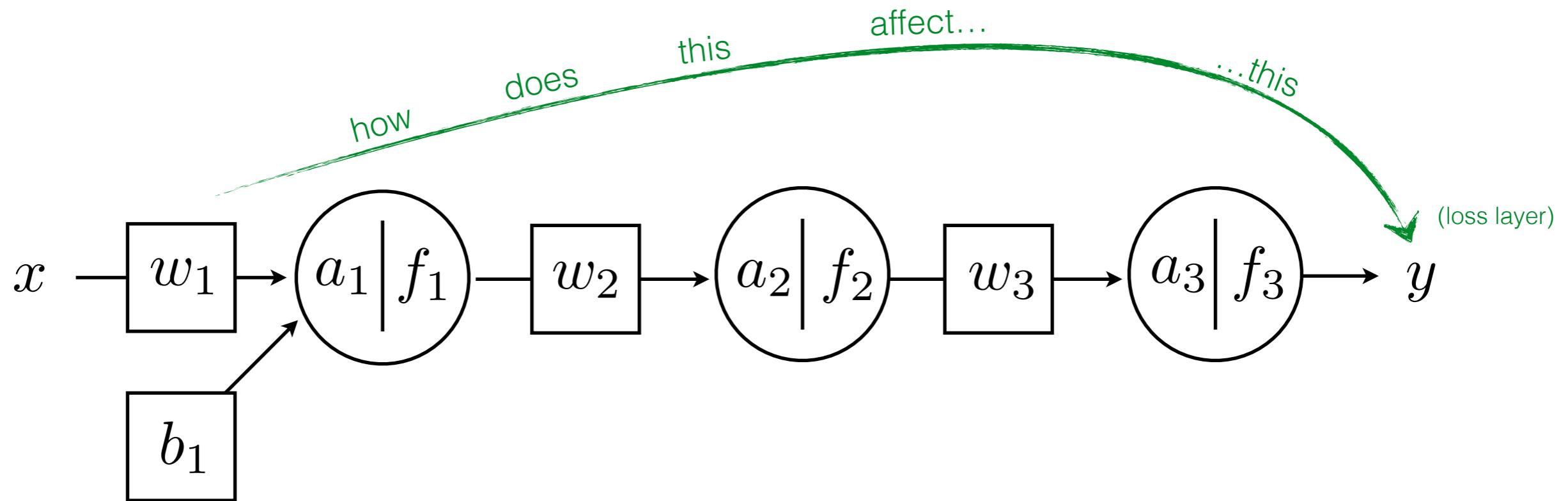
vector of parameter update equations

So we need to compute the partial derivatives

$$\frac{\partial \mathcal{L}}{\partial \theta} = \left[ \frac{\partial \mathcal{L}}{\partial w_3} \frac{\partial \mathcal{L}}{\partial w_2} \frac{\partial \mathcal{L}}{\partial w_1} \frac{\partial \mathcal{L}}{\partial b} \right]$$

Remember,

Partial derivative  $\frac{\partial L}{\partial w_1}$  describes...



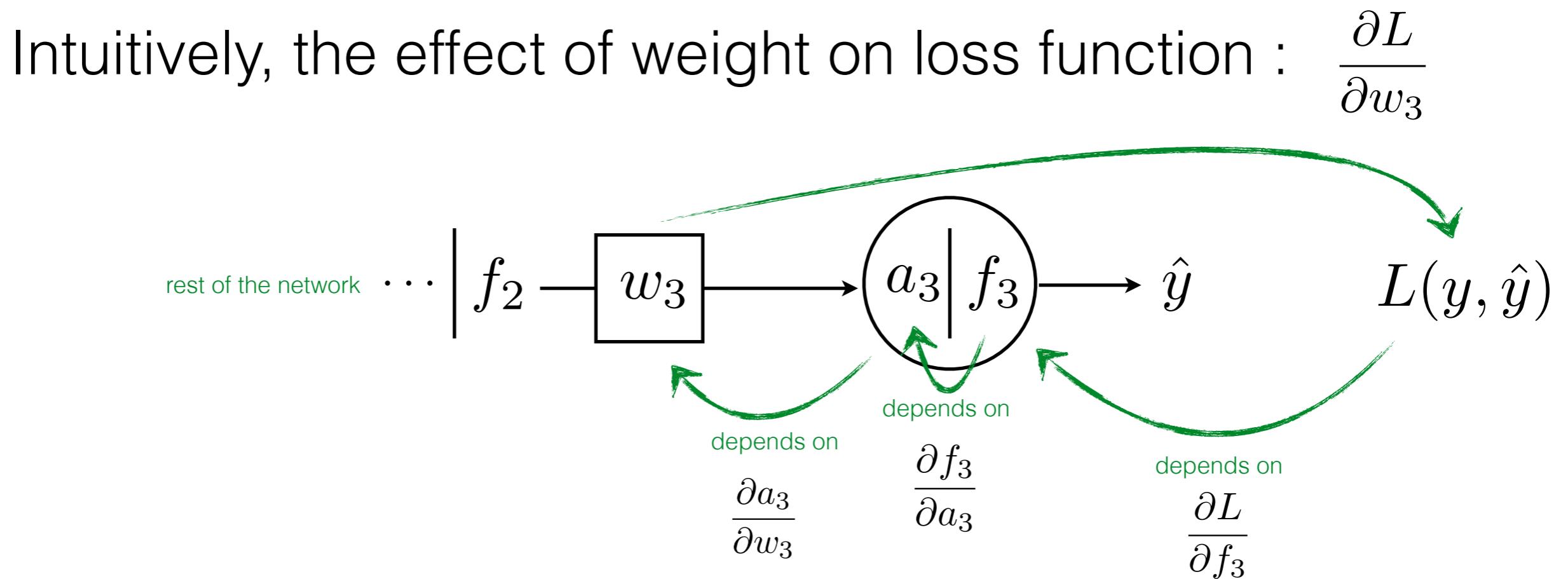
So, how do you compute it?

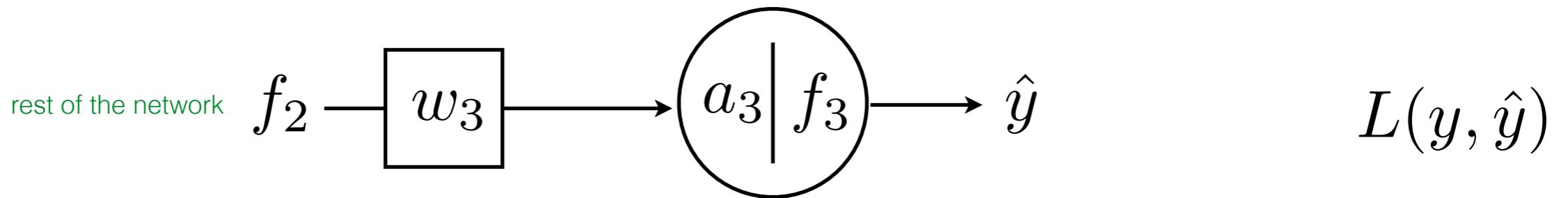
# The Chain Rule



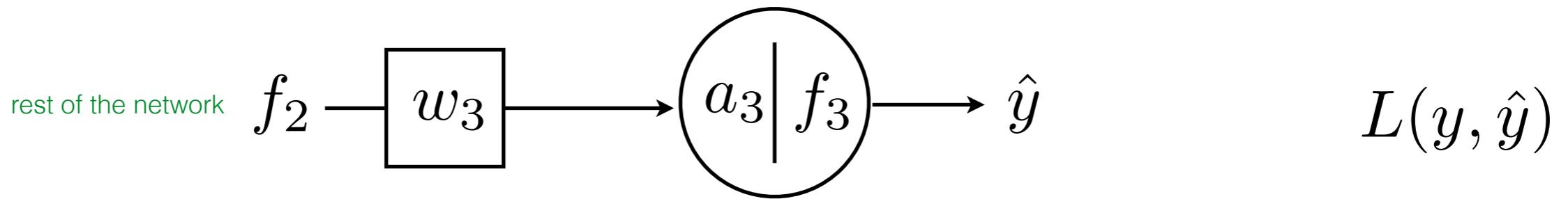
According to the chain rule...

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3}$$





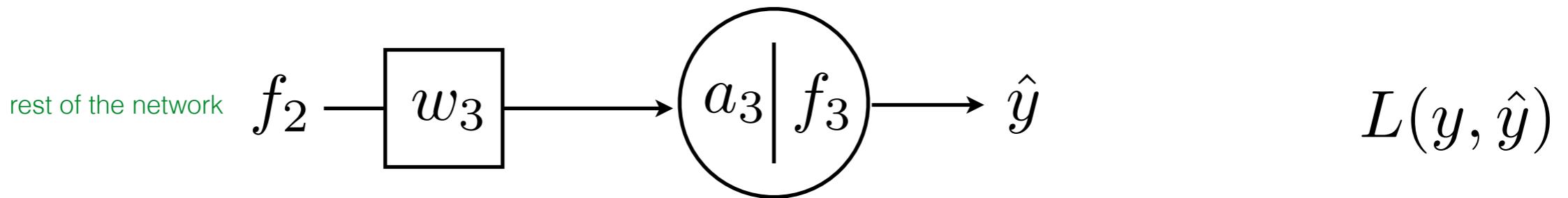
$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \quad \text{Chain Rule!}$$



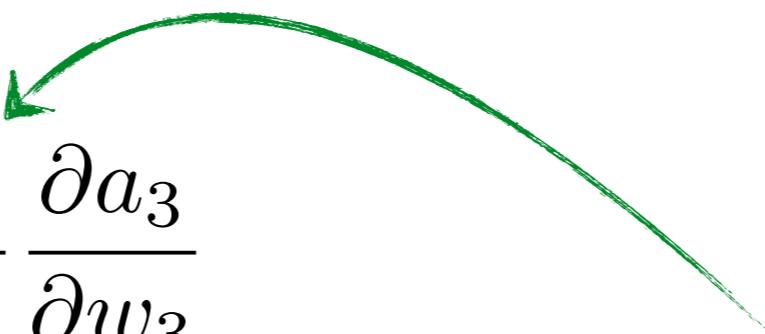
$$\begin{aligned}\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \\ &= -\eta(y - \hat{y}) \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3}\end{aligned}$$

Just the partial derivative of L2 loss



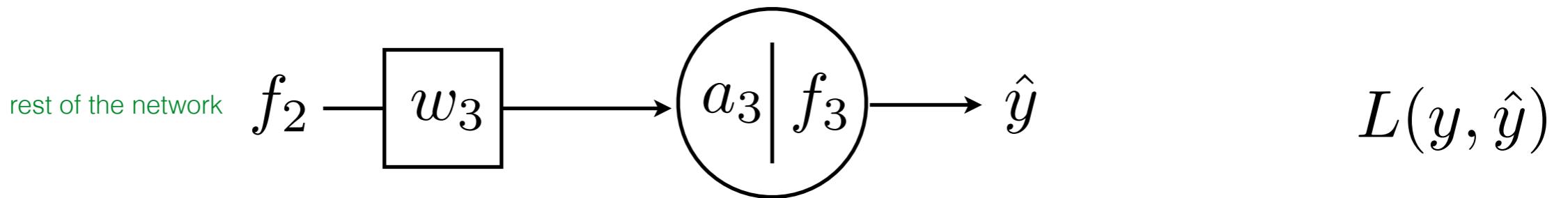


$$\begin{aligned}\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \\ &= -\eta(y - \hat{y}) \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3}\end{aligned}$$



Let's use a Sigmoid function

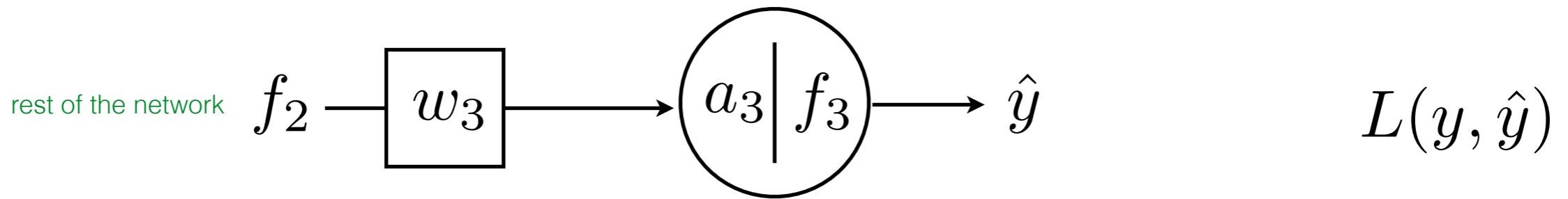
$$\frac{ds(x)}{dx} = s(x)(1 - s(x))$$



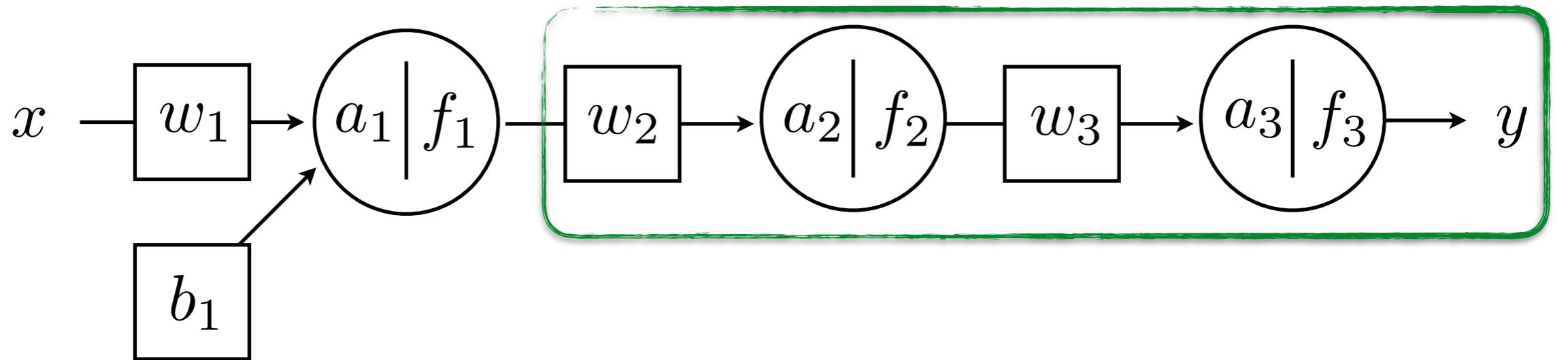
$$\begin{aligned}
 \frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \\
 &= -\eta(y - \hat{y}) \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \\
 &= -\eta(y - \hat{y}) f_3(1 - f_3) \frac{\partial a_3}{\partial w_3}
 \end{aligned}$$

Let's use a Sigmoid function

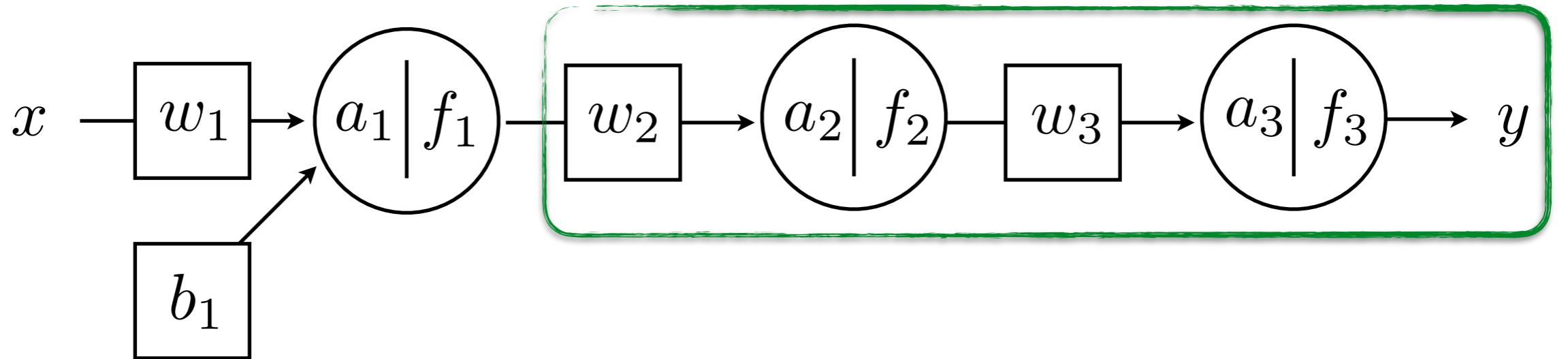
$$\frac{ds(x)}{dx} = s(x)(1 - s(x))$$



$$\begin{aligned}
\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \\
&= -\eta(y - \hat{y}) \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \\
&= -\eta(y - \hat{y}) f_3(1 - f_3) \frac{\partial a_3}{\partial w_3} \\
&= -\eta(y - \hat{y}) f_3(1 - f_3) f_2
\end{aligned}$$



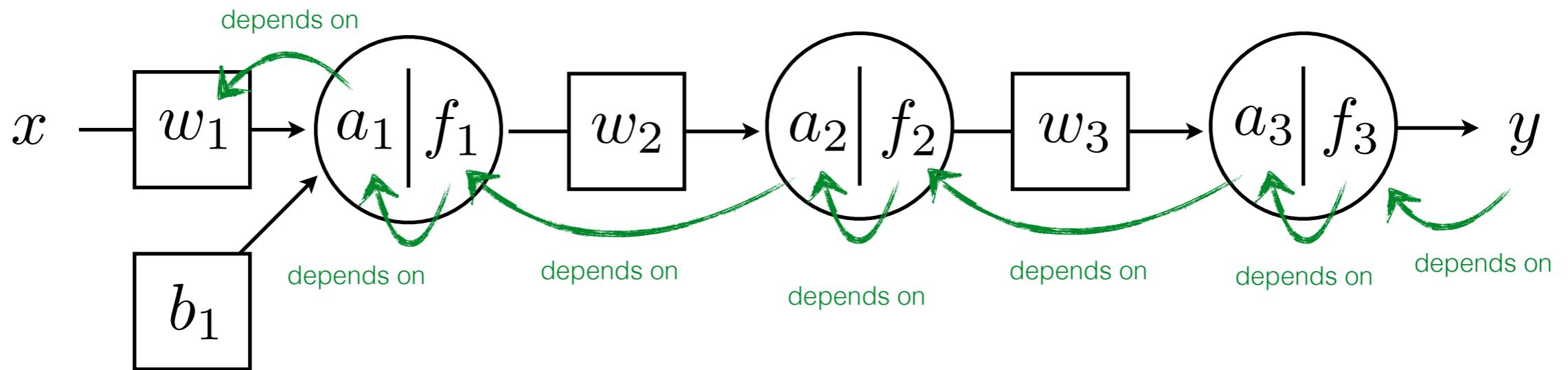
$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial w_2}$$



$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial w_2}$$

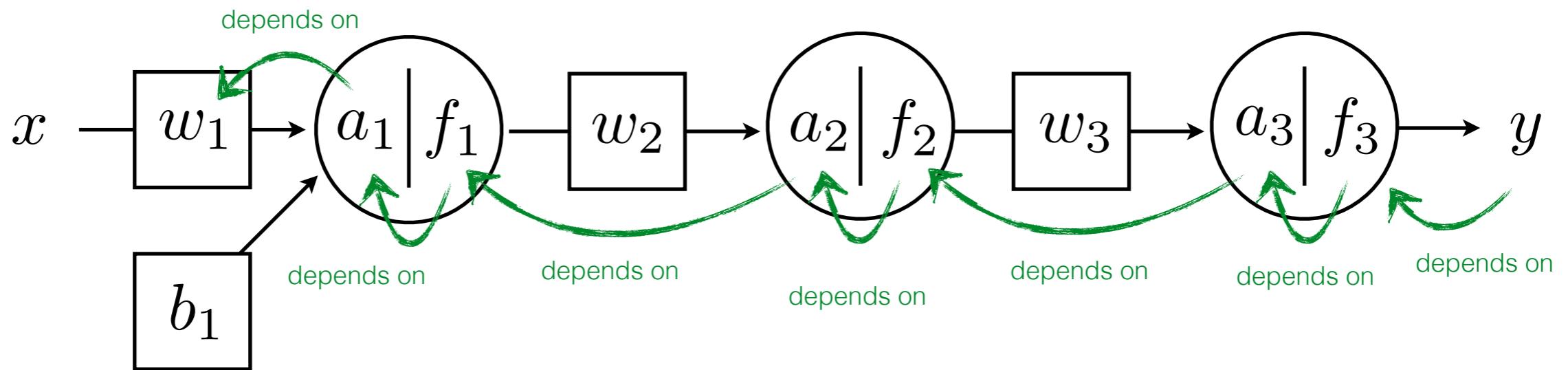
already computed.  
re-use (propagate)!

The Chain rule says...



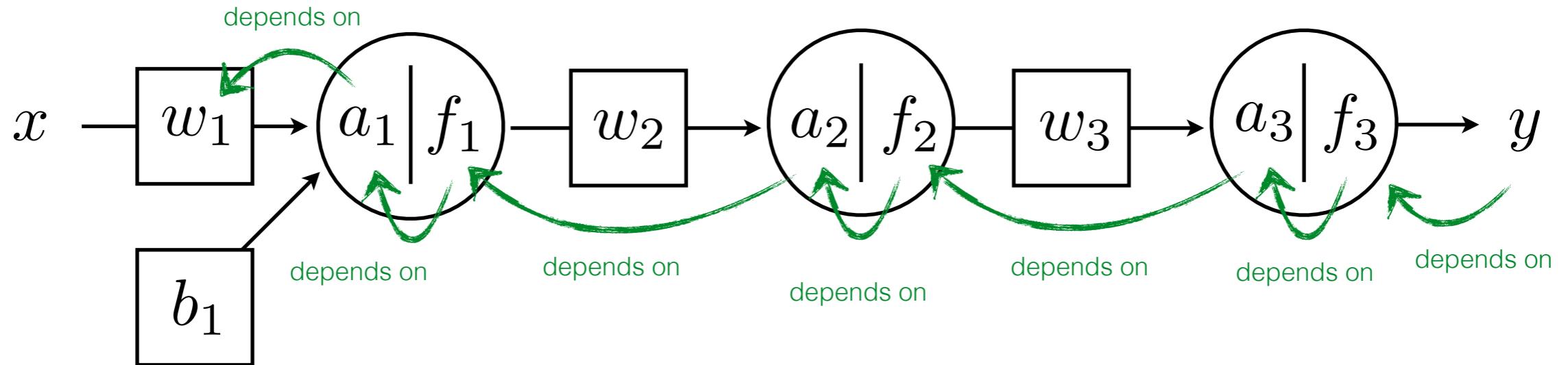
$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial w_1}$$

The Chain rule says...



$$\frac{\partial L}{\partial w_1} = \boxed{\frac{\partial L}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2}} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial w_1}$$

already computed.  
re-use (propagate)!

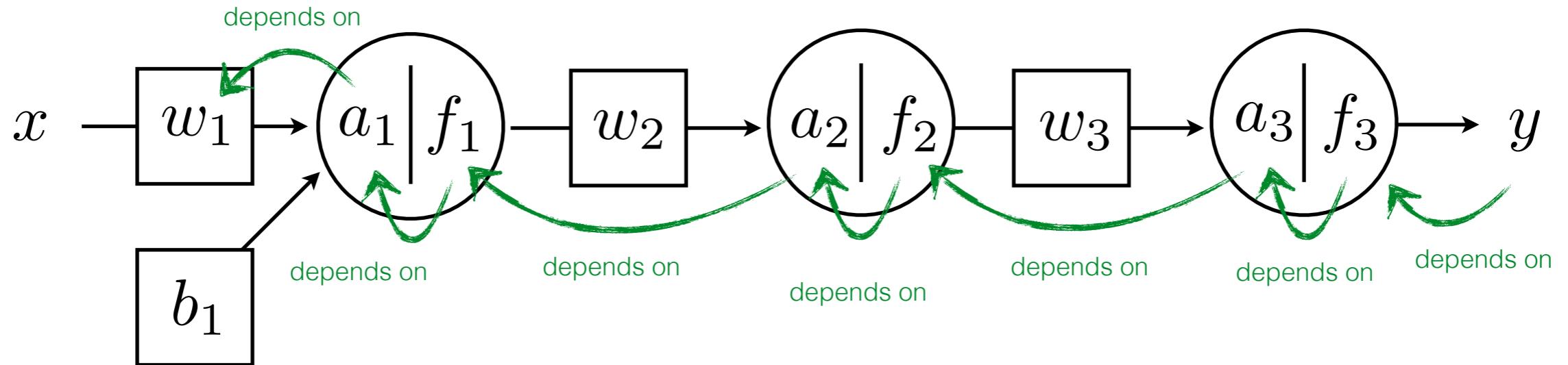


$$\frac{\partial \mathcal{L}}{\partial w_3} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3}$$

$$\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial w_2}$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial w_1}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial b}$$

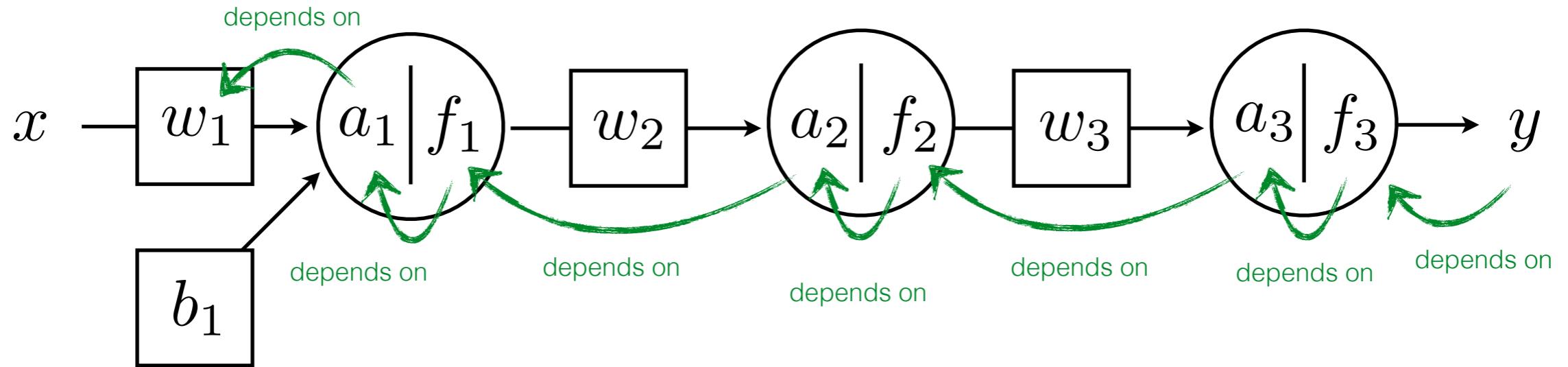


$$\frac{\partial \mathcal{L}}{\partial w_3} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3}$$

$$\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial w_2}$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial w_1}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial b}$$



$$\frac{\partial \mathcal{L}}{\partial w_3} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3}$$

$$\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial w_2}$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial w_1}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial b}$$

# Stochastic Gradient Descent

For each example sample  $\{x_i, y_i\}$

1. Predict

a. Forward pass

$$\hat{y} = f_{\text{MLP}}(x_i; \theta)$$

b. Compute Loss

$$\mathcal{L}_i$$

2. Update

a. Back Propagation

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_3} &= \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial w_3} \\ \frac{\partial \mathcal{L}}{\partial w_2} &= \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial w_2} \\ \frac{\partial \mathcal{L}}{\partial w_1} &= \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial w_1} \\ \frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial \mathcal{L}}{\partial f_3} \frac{\partial f_3}{\partial a_3} \frac{\partial a_3}{\partial f_2} \frac{\partial f_2}{\partial a_2} \frac{\partial a_2}{\partial f_1} \frac{\partial f_1}{\partial a_1} \frac{\partial a_1}{\partial b}\end{aligned}$$

b. Gradient update

$$w_3 = w_3 - \eta \nabla w_3$$

$$w_2 = w_2 - \eta \nabla w_2$$

$$w_1 = w_1 - \eta \nabla w_1$$

$$b = b - \eta \nabla b$$

# Stochastic Gradient Descent

For each example sample  $\{x_i, y_i\}$

1. Predict

a. Forward pass

$$\hat{y} = f_{\text{MLP}}(x_i; \theta)$$

b. Compute Loss

$$\mathcal{L}_i$$

2. Update

a. Back Propagation

$$\frac{\partial \mathcal{L}}{\partial \theta}$$

vector of parameter partial derivatives

b. Gradient update

$$\theta \leftarrow \theta + \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

vector of parameter update equations