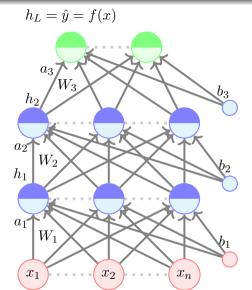
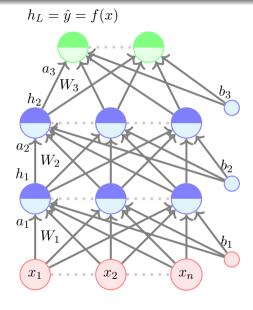
Module 4.2: Learning Parameters of Feedforward Neural Networks (Intuition)

The story so far...

- We have introduced feedforward neural networks
- We are now interested in finding an algorithm for learning the parameters of this model



• Recall our gradient descent algorithm



• Recall our gradient descent algorithm

Algorithm: gradient_descent()

$$t \leftarrow 0;$$

 $max_iterations \leftarrow 1000;$

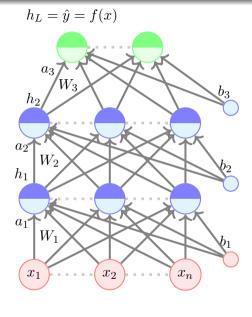
Initialize $w_0, b_0;$

while $t++ < max_iterations$ do

$$w_{t+1} \leftarrow w_t - \eta \nabla w_t;$$

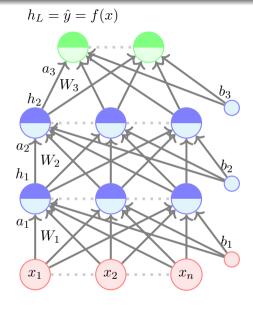
$$b_{t+1} \leftarrow b_t - \eta \nabla b_t;$$

end



- Recall our gradient descent algorithm
- We can write it more concisely as

```
t \leftarrow 0;
max\_iterations \leftarrow 1000;
Initialize \quad w_0, b_0;
\mathbf{while} \ t++ < max\_iterations \ \mathbf{do}
\mid w_{t+1} \leftarrow w_t - \eta \nabla w_t;
\mid b_{t+1} \leftarrow b_t - \eta \nabla b_t;
\mathbf{end}
```



- Recall our gradient descent algorithm
- We can write it more concisely as

$$t \leftarrow 0;$$

 $max_iterations \leftarrow 1000;$
 $Initialize \quad \theta_0 = [w_0, b_0];$
while $t++ < max_iterations$ do
 $\mid \theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t;$
end

$$h_L = \hat{y} = f(x)$$

$$a_3$$

$$h_2$$

$$h_1$$

$$W_2$$

$$h_1$$

$$W_2$$

$$h_1$$

$$W_2$$

$$h_1$$

$$W_2$$

$$h_2$$

$$h_2$$

$$h_3$$

$$W_2$$

$$h_4$$

$$W_2$$

$$h_4$$

$$W_2$$

$$h_4$$

$$W_2$$

$$h_4$$

$$W_3$$

$$h_2$$

$$W_3$$

$$h_2$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_5$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_5$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_5$$

$$h_4$$

$$W_5$$

$$h_4$$

$$W_7$$

$$h_4$$

$$W_8$$

$$h_4$$

$$W_8$$

$$h_4$$

$$W_1$$

$$W_2$$

$$h_4$$

$$W_3$$

$$h_4$$

$$W_4$$

$$h_4$$

$$W_4$$

$$W_4$$

$$H_4$$

$$W_4$$

$$W_5$$

$$H_7$$

$$W_8$$

$$W_8$$

$$H_7$$

$$W_8$$

$$W_$$

- Recall our gradient descent algorithm
- We can write it more concisely as

$$\begin{split} t &\leftarrow 0; \\ max_iterations &\leftarrow 1000; \\ Initialize &\quad \theta_0 = [w_0, b_0]; \\ \mathbf{while} \ t++ &< max_iterations \ \mathbf{do} \\ &\quad \mid \ \theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t; \\ \mathbf{end} \end{split}$$

• where
$$\nabla \theta_t = \left[\frac{\partial \mathcal{L}(\theta)}{\partial w_t}, \frac{\partial \mathcal{L}(\theta)}{\partial b_t}\right]^T$$

$$h_L = \hat{y} = f(x)$$

$$a_3$$

$$h_2$$

$$h_1$$

$$W_2$$

$$h_1$$

$$W_2$$

$$h_1$$

$$W_2$$

$$h_2$$

$$W_3$$

$$h_2$$

$$W_3$$

$$h_2$$

$$h_3$$

$$h_2$$

$$h_3$$

$$h_2$$

$$h_4$$

$$W_2$$

$$h_1$$

$$W_3$$

$$h_2$$

$$h_3$$

$$h_2$$

$$h_3$$

$$h_4$$

$$h_2$$

$$h_3$$

$$h_4$$

$$h_3$$

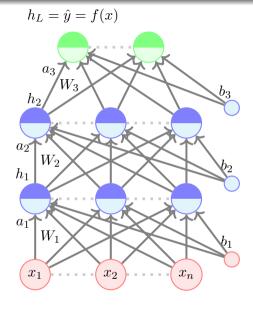
$$h_4$$

$$h_$$

- Recall our gradient descent algorithm
- We can write it more concisely as

$$\begin{array}{l} t \leftarrow 0; \\ max_iterations \leftarrow 1000; \\ Initialize \quad \theta_0 = [w_0, b_0]; \\ \mathbf{while} \ t++ < max_iterations \ \mathbf{do} \\ \mid \ \theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t; \\ \mathbf{end} \end{array}$$

- where $\nabla \theta_t = \left[\frac{\partial \mathcal{L}(\theta)}{\partial w_t}, \frac{\partial \mathcal{L}(\theta)}{\partial b_t}\right]^T$
- Now, in this feedforward neural network, instead of $\theta = [w, b]$ we have $\theta = W_1, W_2, ..., W_L, b_1, b_2, ..., b_L$

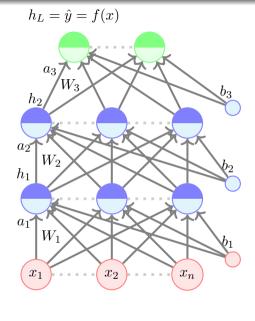


- Recall our gradient descent algorithm
- We can write it more concisely as

$$\begin{array}{l} t \leftarrow 0; \\ max_iterations \leftarrow 1000; \\ Initialize \quad \theta_0 = [w_0, b_0]; \\ \textbf{while } t++ < max_iterations \ \textbf{do} \\ \mid \ \theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t; \\ \textbf{end} \end{array}$$

- where $\nabla \theta_t = \left[\frac{\partial \mathcal{L}(\theta)}{\partial w_t}, \frac{\partial \mathcal{L}(\theta)}{\partial b_t} \right]^T$
- Now, in this feedforward neural network, instead of $\theta = [w, b]$ we have $\theta = W_1, W_2, ..., W_L, b_1, b_2, ..., b_L$
- We can still use the same algorithm for learning the parameters of our model





- Recall our gradient descent algorithm
- We can write it more concisely as

$$\begin{array}{l} t \leftarrow 0; \\ max_iterations \leftarrow 1000; \\ Initialize \quad \theta_0 = [W_1^0,...,W_L^0,b_1^0,...,b_L^0]; \\ \textbf{while } t++ < max_iterations \ \textbf{do} \\ \mid \ \theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t; \\ \textbf{end} \end{array}$$

- where $\nabla \theta_t = \left[\frac{\partial \mathcal{L}(\theta)}{\partial w_t}, \frac{\partial \mathcal{L}(\theta)}{\partial b_t} \right]^T$
- Now, in this feedforward neural network, instead of $\theta = [w, b]$ we have $\theta = W_1, W_2, ..., W_L, b_1, b_2, ..., b_L$
- We can still use the same algorithm for learning the parameters of our model



•	Except	that	now	our	$\nabla \theta$	looks	much	more	nasty
	$ \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} $								

• Except that now our $\nabla \theta$ looks much more nasty $\lceil \frac{\partial \mathcal{L}(\theta)}{} \rceil$

 $\frac{\partial \mathcal{L}(\theta)}{\partial W_{111}}$...

• Except that now our $\nabla \theta$ looks much more nasty $\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}} \end{bmatrix}$

$$\frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} \cdots \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}}$$

• Except that now our $\nabla \theta$ looks much more nasty $\int \partial \mathcal{L}(\theta) \partial \mathcal{L}(\theta)$

$$\frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} \cdots \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial W_{121}} \cdots \frac{\partial \mathcal{L}(\theta)}{\partial W_{12n}}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial W_{1n1}} \cdots \frac{\partial \mathcal{L}(\theta)}{\partial W_{1nn}}$$

```
\partial \mathscr{L}(\theta)
                                                                                               \partial \mathcal{L}(\theta)
                                                            \partial \mathscr{L}(\theta)
                                                                                                                                                           \partial \mathscr{L}(\theta)
                                                                                               \partial W_{211}
                                                                                                                                                                                               . . .
  \overline{\partial W_{111}}
                                     . . .
                                                            \overline{\partial W_{11n}}
                                                                                                                                   . . .
                                                                                                                                                           \overline{\partial W_{21n}}
                                                            \partial \mathscr{L}(\theta)
                                                                                               \partial \mathscr{L}(\theta)
                                                                                                                                                           \partial \mathscr{L}(\theta)
 \partial \mathscr{L}(\theta)
\overline{\partial W_{121}}
                                                                                                                                                          \partial W_{22n}
                                     . . .
                                                            \overline{\partial W_{12n}}
                                                                                                \overline{\partial W_{221}}
                                                                                                                                                                                               . . .
\overline{\partial W_{1n1}}
                                                            \overline{\partial W_{1nn}}
                                                                                                \overline{\partial W_{2n1}}
                                                                                                                                                           \overline{\partial W_{2nn}}
                                                                                                                                                                                               . . .
```

_	ACCPU	01160 11	ow our	V U TOOK	b IIIu	cii iiioi c	TIGOU.	y		
	$ \begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} \end{bmatrix} $		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}}$	$\frac{\partial \mathcal{L}(\theta)}{\partial W_{211}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{21n}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{L,11}}$		$\frac{\partial \mathscr{L}(\theta)}{\partial W_{L,1k}}$
	$\frac{\partial \mathcal{L}(\theta)}{\partial W_{121}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{12n}}$	$\frac{\partial \mathcal{L}(\theta)}{\partial W_{221}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{22n}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{L,21}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{L,2k}}$
	:	:	:	:	÷	÷	÷	÷	÷	:
	$\frac{\partial \mathcal{L}(\theta)}{\partial W_{1n1}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{1nn}}$	$\frac{\partial \mathcal{L}(\theta)}{\partial W_{2n1}}$		$\frac{\partial \mathcal{L}(\theta)}{\partial W_{2nn}}$		$\frac{\partial \mathscr{L}(\theta)}{\partial W_{L,n1}}$		$\frac{\partial \mathscr{L}(\theta)}{\partial W_{L,nk}}$

$$\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{211}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{21n}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,11}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,1k}} \\ \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{121}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{12n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{221}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{22n}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,21}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,2k}} \\ \\ \vdots & \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{1n1}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2n1}} & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2nn}} & & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,n1}} & & & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,nk}} \end{bmatrix}$$

• ... and similar entries for partial derivatives w.r.t. the elements of $b_1, b_2, ..., b_L$

$$\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{211}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{21n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,11}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,1k}} \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{121}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{12n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{221}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{22n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,21}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,2k}} \\ \vdots & \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{1n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{1nn}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2nn}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,nk}} \end{bmatrix}$$

- ... and similar entries for partial derivatives w.r.t. the elements of $b_1, b_2, ..., b_L$
- $\nabla \theta$ is thus composed of $\nabla W_1, \nabla W_2, ... \nabla W_L \in \mathbb{R}^{n \times n}, \nabla W_L \in \mathbb{R}^{n \times k}, \nabla b_1, \nabla b_2, ..., \nabla b_n \in \mathbb{R}^n \text{ and } \nabla b_L \in \mathbb{R}^k$

We need to answer two questions

We need to answer two questions

• How to choose the loss function $\mathcal{L}(\theta)$?

We need to answer two questions

- How to choose the loss function $\mathcal{L}(\theta)$?
- How to compute $\nabla \theta$ which is composed of $\nabla W_1, \nabla W_2, ..., \nabla W_{L-1} \in \mathbb{R}^{n \times n}, \nabla W_L \in \mathbb{R}^{n \times k}, \nabla b_1, \nabla b_2, ..., \nabla b_{L-1} \in \mathbb{R}^n \text{ and } \nabla b_L \in \mathbb{R}^k$?