

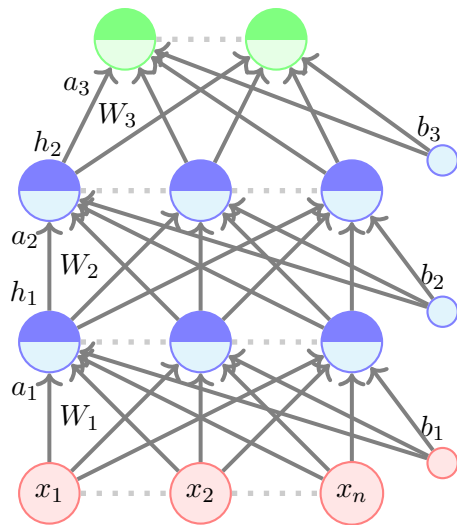
Module 4.2: Learning Parameters of Feedforward Neural Networks (Intuition)

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

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

- Recall our gradient descent algorithm



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

- Recall our gradient descent algorithm

Algorithm: gradient_descent()

$t \leftarrow 0$;

$max_iterations \leftarrow 1000$;

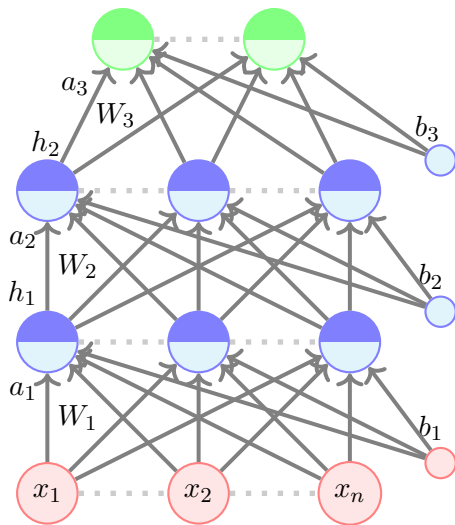
Initialize w_0, b_0 ;

while $t++ < max_iterations$ **do**

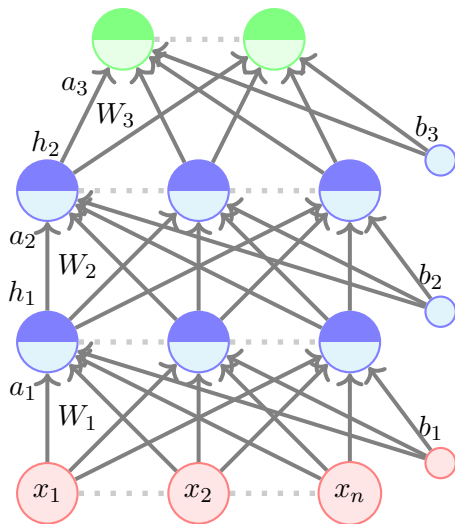
$w_{t+1} \leftarrow w_t - \eta \nabla w_t$;

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

end



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



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

Algorithm: `gradient_descent()`

$t \leftarrow 0$;

$max_iterations \leftarrow 1000$;

Initialize w_0, b_0 ;

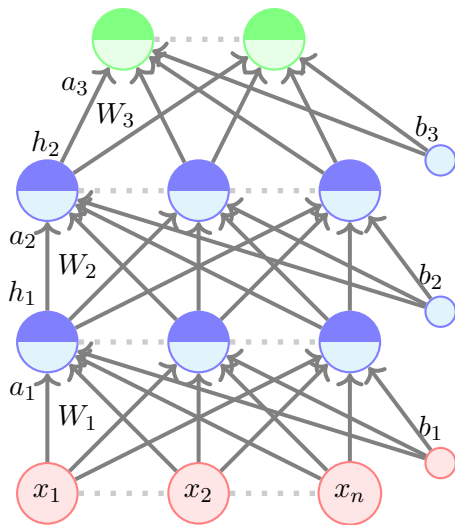
while $t++ < max_iterations$ **do**

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

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

end

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



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

Algorithm: `gradient_descent()`

$t \leftarrow 0$;

$max_iterations \leftarrow 1000$;

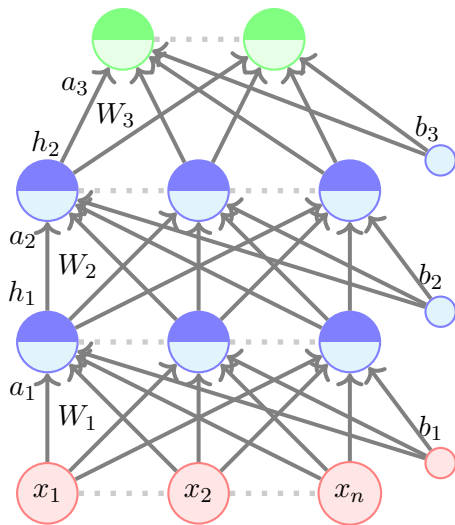
Initialize $\theta_0 = [w_0, b_0]$;

while $t++ < max_iterations$ **do**

$\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;

end

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



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

Algorithm: `gradient_descent()`

$t \leftarrow 0;$

$max_iterations \leftarrow 1000;$

Initialize $\theta_0 = [w_0, b_0];$

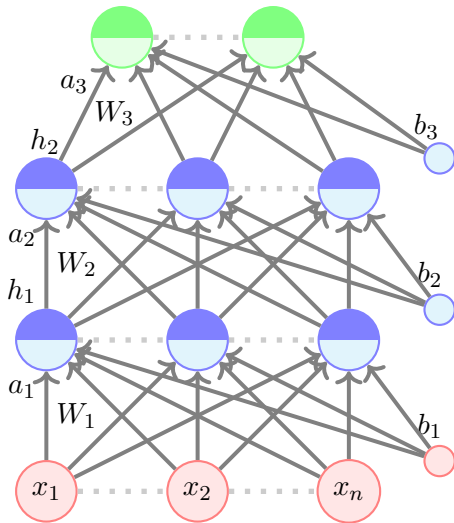
while $t++ < max_iterations$ **do**

$\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t;$

end

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

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



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

Algorithm: `gradient_descent()`

$t \leftarrow 0;$

$max_iterations \leftarrow 1000;$

Initialize $\theta_0 = [w_0, b_0];$

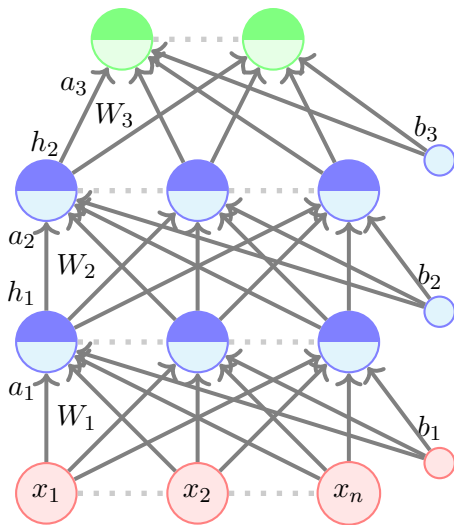
while $t++ < max_iterations$ **do**

$\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t;$

end

- 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 $\theta = [W_1, W_2, \dots, W_L, b_1, b_2, \dots, b_L]$

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



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

Algorithm: `gradient_descent()`

$t \leftarrow 0;$

$max_iterations \leftarrow 1000;$

Initialize $\theta_0 = [w_0, b_0];$

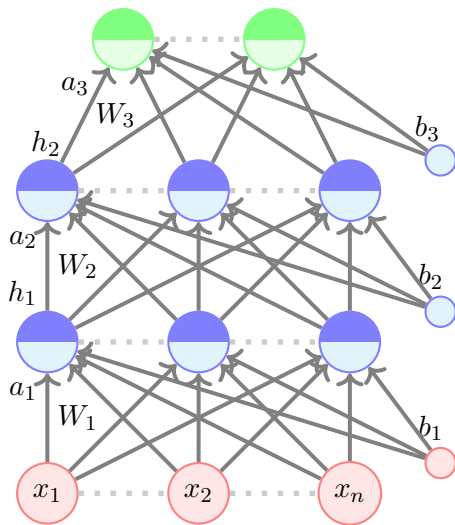
while $t++ < max_iterations$ **do**

$\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t;$

end

- 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 $\theta = [W_1, W_2, \dots, W_L, b_1, b_2, \dots, b_L]$
- We can still use the same algorithm for learning the parameters of our model

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



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

Algorithm: gradient_descent()

$t \leftarrow 0$;

$max_iterations \leftarrow 1000$;

Initialize $\theta_0 = [W_1^0, \dots, W_L^0, b_1^0, \dots, b_L^0]$;

while $t++ < max_iterations$ **do**

$\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;

end

- where $\nabla \theta_t = [\frac{\partial \mathcal{L}(\theta)}{\partial W_{1,t}}, \dots, \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,t}}, \frac{\partial \mathcal{L}(\theta)}{\partial b_{1,t}}, \dots, \frac{\partial \mathcal{L}(\theta)}{\partial b_{L,t}}]^T$
- Now, in this feedforward neural network, instead of $\theta = [w, b]$ we have $\theta = [W_1, W_2, \dots, W_L, b_1, b_2, \dots, b_L]$
- We can still use the same algorithm for learning the parameters of our model

- Except that now our $\nabla\theta$ looks much more nasty

- Except that now our $\nabla\theta$ looks much more nasty

$$\left[\begin{array}{c} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \right]$$

- Except that now our $\nabla\theta$ looks much more nasty

$$\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & \cdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix}$$

- Except that now our $\nabla\theta$ looks much more nasty

$$\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}} \\ \vdots & & \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{n11}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{nnn}} \end{bmatrix}$$

- Except that now our $\nabla\theta$ looks much more nasty

$$\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}} \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{121}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{12n}} \\ \vdots & \vdots & \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{1n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{1nn}} \end{bmatrix}$$

- Except that now our $\nabla\theta$ looks much more nasty

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

- Except that now our $\nabla\theta$ looks much more nasty

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

- Except that now our $\nabla\theta$ looks much more nasty

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

- Except that now our $\nabla\theta$ looks much more nasty

$$\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{211}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{21n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,11}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,1k}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,1k}} & \frac{\partial \mathcal{L}(\theta)}{\partial b_{11}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial b_{L1}} \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{121}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{12n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{221}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{22n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,21}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,2k}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,2k}} & \frac{\partial \mathcal{L}(\theta)}{\partial b_{12}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial b_{L2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{1n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{1nn}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2nn}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,nk}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,nk}} & \frac{\partial \mathcal{L}(\theta)}{\partial b_{1n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial b_{Lk}} \end{bmatrix}$$

- Except that now our $\nabla\theta$ looks much more nasty

$$\begin{bmatrix} \frac{\partial \mathcal{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{11n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{211}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{21n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,11}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,1k}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,1k}} & \frac{\partial \mathcal{L}(\theta)}{\partial b_{11}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial b_{L1}} \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{121}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{12n}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{221}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{22n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,21}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,2k}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,2k}} & \frac{\partial \mathcal{L}(\theta)}{\partial b_{12}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial b_{L2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \mathcal{L}(\theta)}{\partial W_{1n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{1nn}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{2nn}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,n1}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,nk}} & \frac{\partial \mathcal{L}(\theta)}{\partial W_{L,nk}} & \frac{\partial \mathcal{L}(\theta)}{\partial b_{1n}} & \cdots & \frac{\partial \mathcal{L}(\theta)}{\partial b_{Lk}} \end{bmatrix}$$

- $\nabla\theta$ is thus composed of
 $\nabla W_1, \nabla W_2, \dots, \nabla W_{L-1} \in \mathbb{R}^{n \times n}, \nabla W_L \in \mathbb{R}^{n \times k},$
 $\nabla b_1, \nabla b_2, \dots, \nabla b_{L-1} \in \mathbb{R}^n$ 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
 $\nabla W_1, \nabla W_2, \dots, \nabla W_{L-1} \in \mathbb{R}^{n \times n}, \nabla W_L \in \mathbb{R}^{n \times k}$
 $\nabla b_1, \nabla b_2, \dots, \nabla b_{L-1} \in \mathbb{R}^n$ and $\nabla b_L \in \mathbb{R}^k$?