Deep Learning

Feed Forward Neural Networks and Backpropagation

Puneet Kumar Jain

CSE Department

National Institute of Technology Rourkela

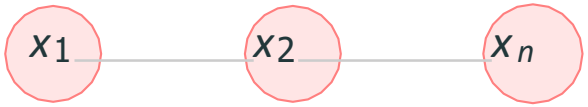
References:

The Slides are prepared from the following major source:

- **"CS7105-Deep Learning" by Mitesh M. Khapra, IIT Madras. http://www.cse.iitm.ac.in/~miteshk/CS7015_2018.html.**
- See the excellent videos by Hugo Larochelle on Backpropagation

Feedforward Neural Networks(multilayered network of neurons)

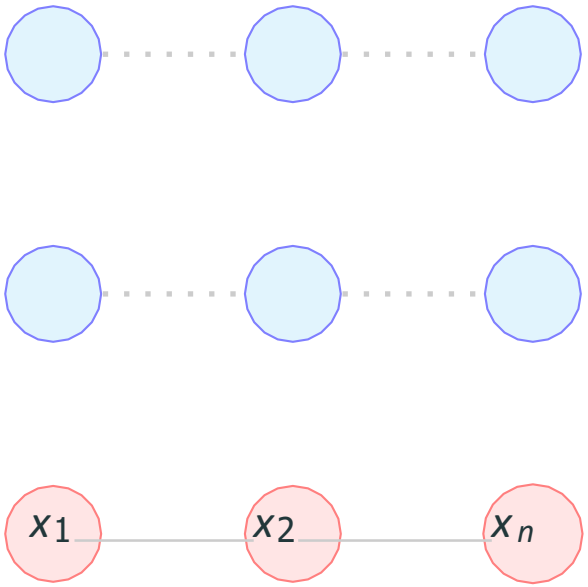
The input to the network is an **n** -dimensional vector



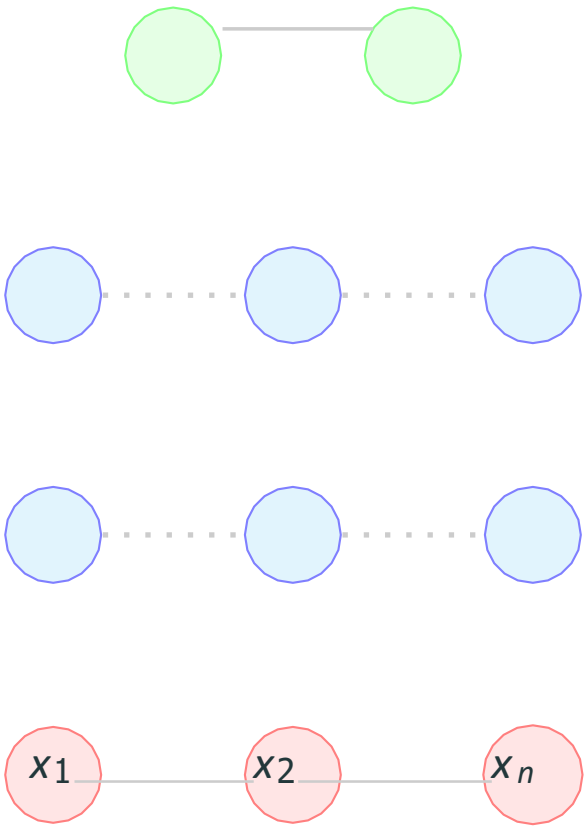
Feedforward Neural Networks (multilayered network of neurons)

The input to the network is an n-dimensional vector

The network contains $L-1$ hidden layers (2, in this case) having n neurons each



Feedforward Neural Networks (multilayered network of neurons)

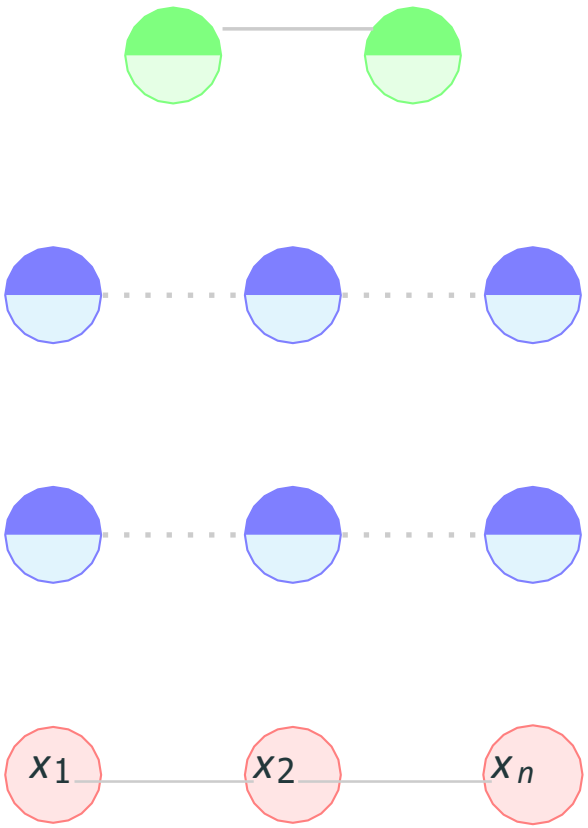


The input to the network is an **n**-dimensional vector

The network contains **L −1** hidden layers (2, in this case) having **n** neurons each

Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Feedforward Neural Networks(multilayered network of neurons)



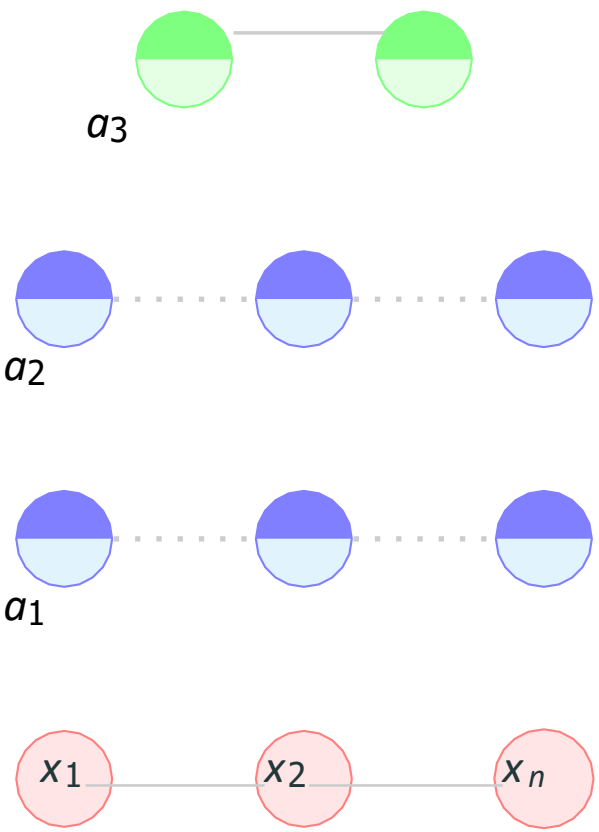
The input to the network is an **n**-dimensional vector

The network contains **L −1** hidden layers (2, in this case) having **n** neurons each

Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Each neuron in the hidden layer and output layer can be split into two parts :

Feedforward Neural Networks(multilayered network of neurons)



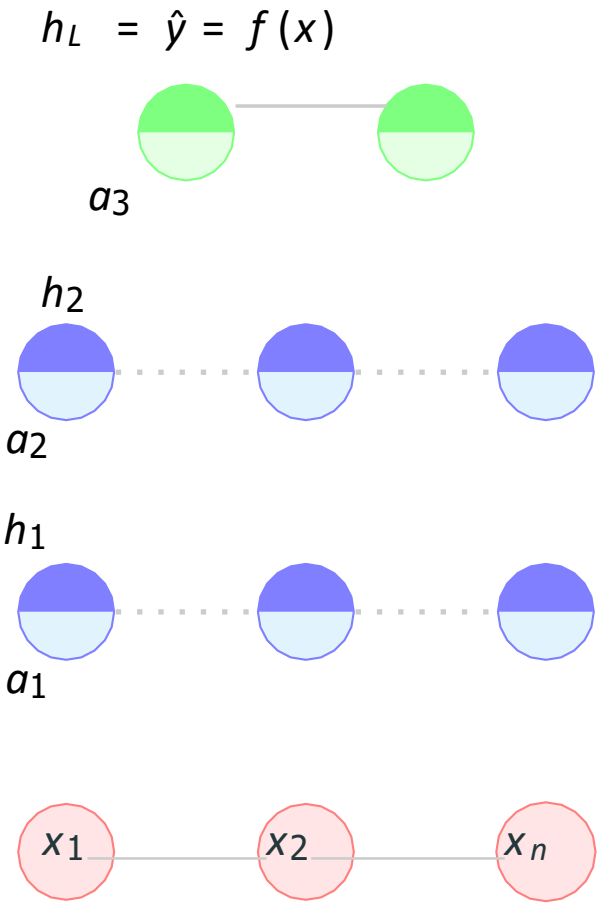
The input to the network is an **n**-dimensional vector

The network contains **L −1** hidden layers (2, in this case) having **n** neurons each

Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Each neuron in the hidden layer and output layer can be split into two parts : pre-activation

Feedforward Neural Networks(multilayered network of neurons)



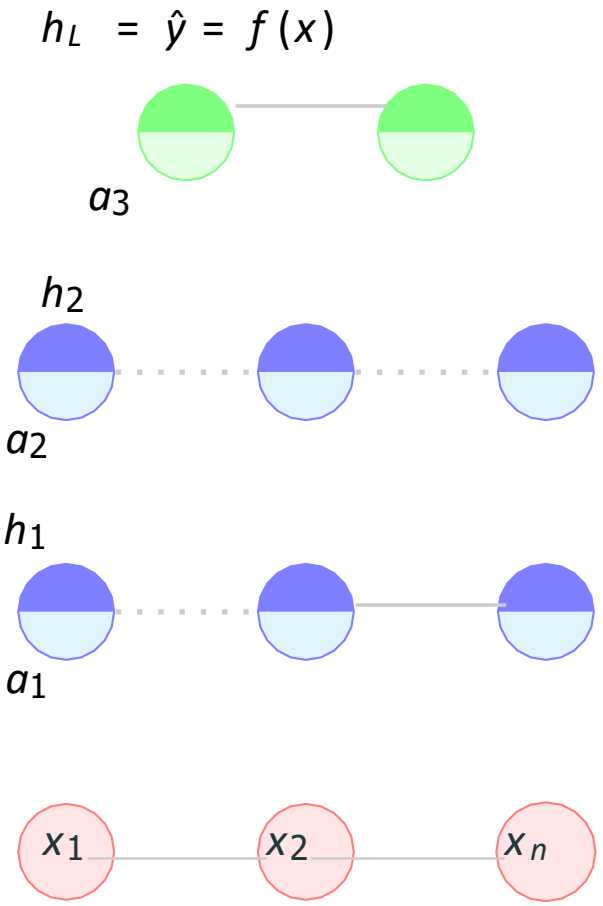
The input to the network is an **n**-dimensional vector

The network contains **L −1** hidden layers (2, in this case) having **n** neurons each

Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Each neuron in the hidden layer and output layer can be split into two parts : pre-activation and activation (a_i and h_i are vectors)

Feedforward Neural Networks(multilayered network of neurons)



The input to the network is an **n**-dimensional vector

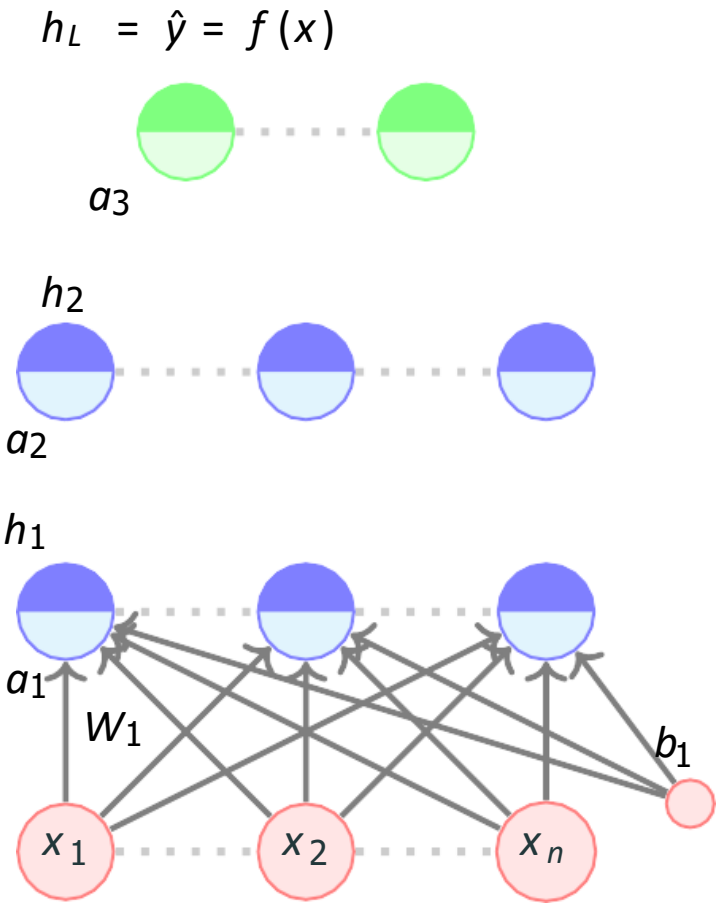
The network contains **L −1** hidden layers (2, in this case) having **n** neurons each

Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Each neuron in the hidden layer and output layer can be split into two parts : pre-activation and activation (a_i and h_i are vectors)

The input layer can be called the 0-th layer and the output layer can be called the (L)-th layer

Feedforward Neural Networks(multilayered network of neurons)



The input to the network is an **n**-dimensional vector

The network contains **L** -1 hidden layers (2, in this case) having **n** neurons each

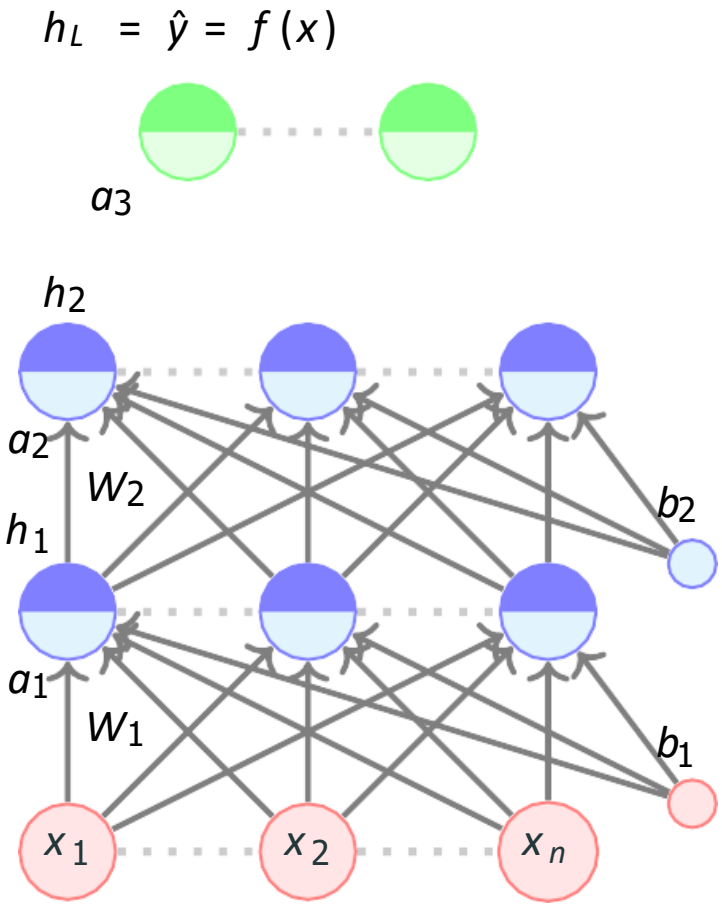
Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Each neuron in the hidden layer and output layer can be split into two parts : pre-activation and activation (a_i and h_i are vectors)

The input layer can be called the 0-th layer and the output layer can be called the (L)-th layer

$W_i \in \mathbb{R}^{n \times n}$ and $b_i \in \mathbb{R}^n$ are the weight and bias between layers $i-1$ and i ($0 < i < L$)

Feedforward Neural Networks(multilayered network of neurons)



The input to the network is an **n**-dimensional vector

The network contains **L** -1 hidden layers (2, in this case) having **n** neurons each

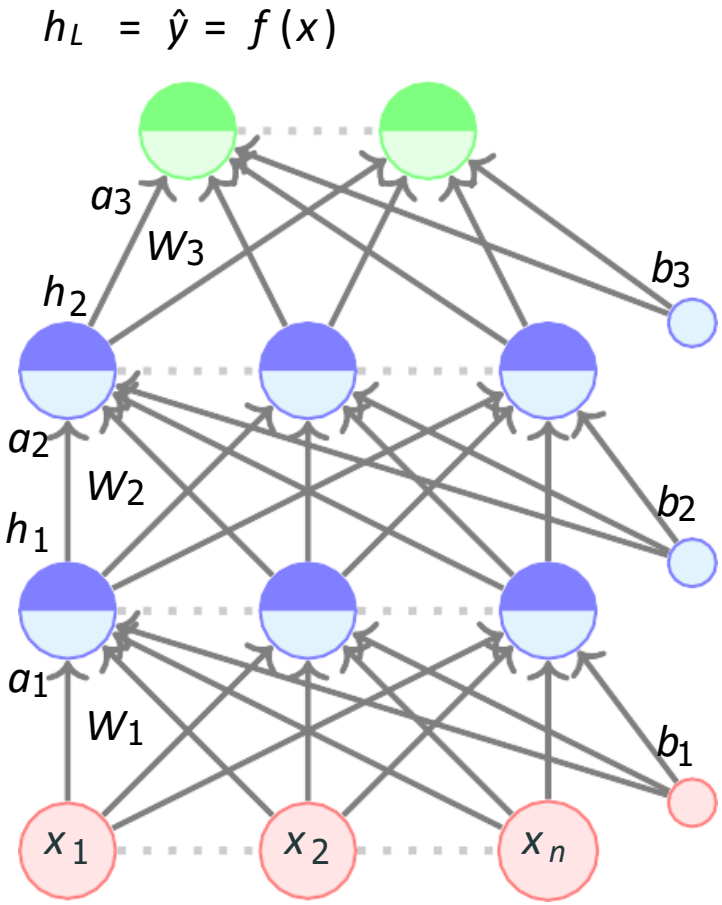
Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Each neuron in the hidden layer and output layer can be split into two parts : pre-activation and activation (a_i and h_i are vectors)

The input layer can be called the 0-th layer and the output layer can be called the (L)-th layer

$W_i \in \mathbb{R}^{n \times n}$ and $b_i \in \mathbb{R}^n$ are the weight and bias between layers $i-1$ and i ($0 < i < L$)

Feedforward Neural Networks(multilayered network of neurons)



The input to the network is an **n**-dimensional vector

The network contains **L** −**1** hidden layers (2, in this case) having **n** neurons each

Finally, there is one output layer containing **k** neurons (say, corresponding to **k** classes)

Each neuron in the hidden layer and output layer can be split into two parts : pre-activation and activation (a_i and h_i are vectors)

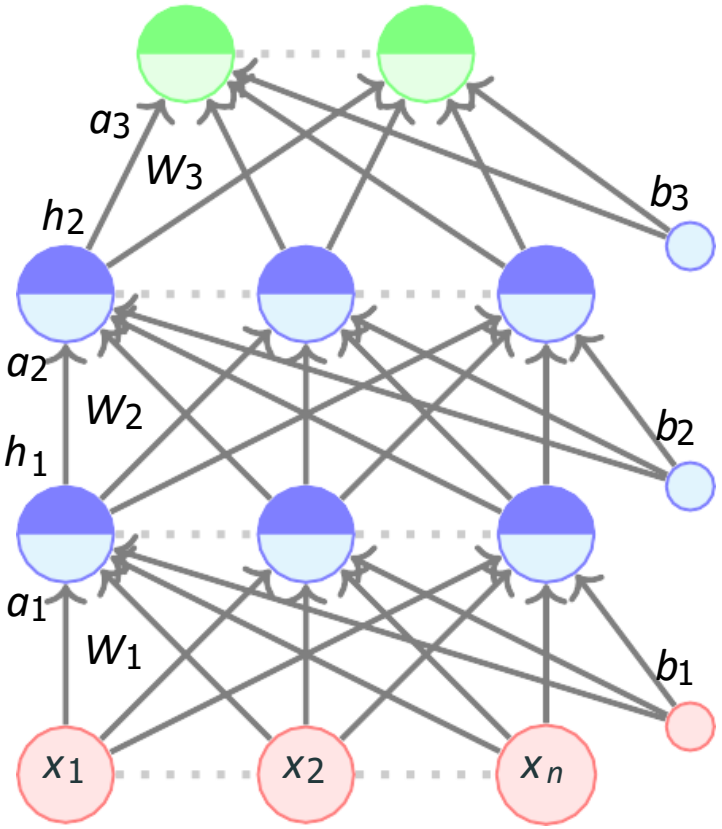
The input layer can be called the 0-th layer and the output layer can be called the (L)-th layer

$W_i \in \mathbb{R}^{n \times n}$ and $b_i \in \mathbb{R}^n$ are the weight and bias between layers $i-1$ and i ($0 < i < L$)

$W_L \in \mathbb{R}^{n \times k}$ and $b_L \in \mathbb{R}^k$ are the weight and bias between the last hidden layer and the output layer ($L = 3$ in this case)

Feedforward Neural Networks(multilayered network of neurons)

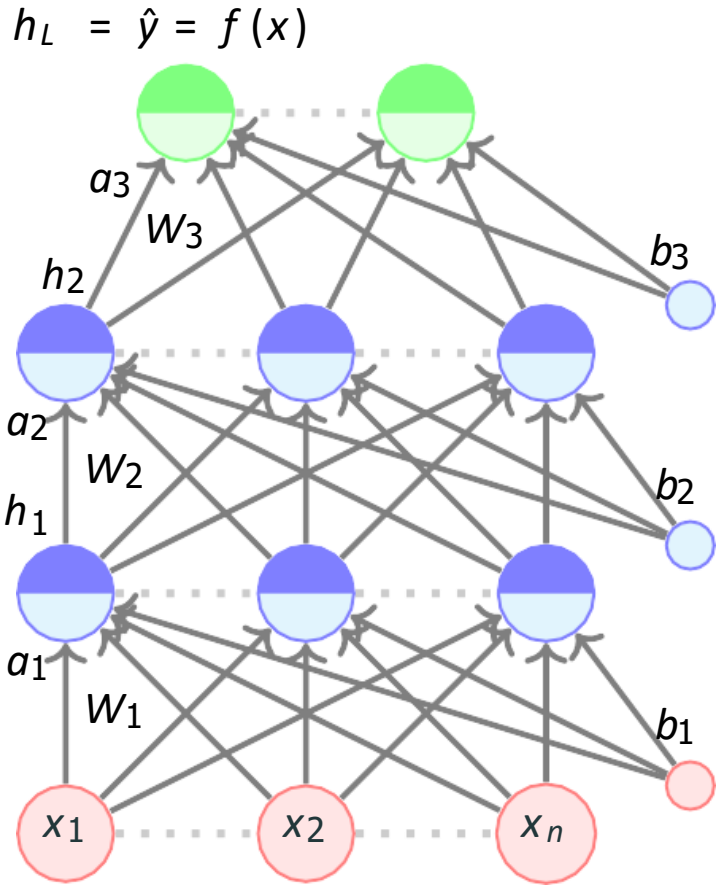
$h_L = \hat{y} = f(x)$



The pre-activation at layer i is given by

$$a_i(x) = b_i + W_i h_{i-1}(x)$$

Feedforward Neural Networks(multilayered network of neurons)



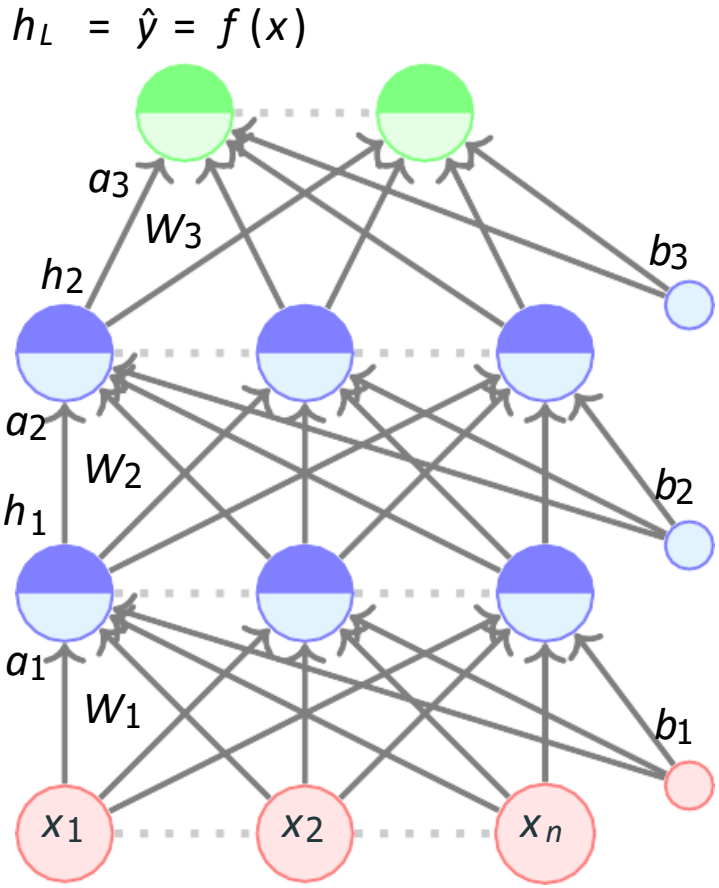
The pre-activation at layer i is given by

$$a_i(x) = b_i + W_i h_{i-1}(x)$$

The activation at layer i is given by

$$h_i(x) = g(a_i(x))$$

Feedforward Neural Networks (multilayered network of neurons)



The pre-activation at layer i is given by

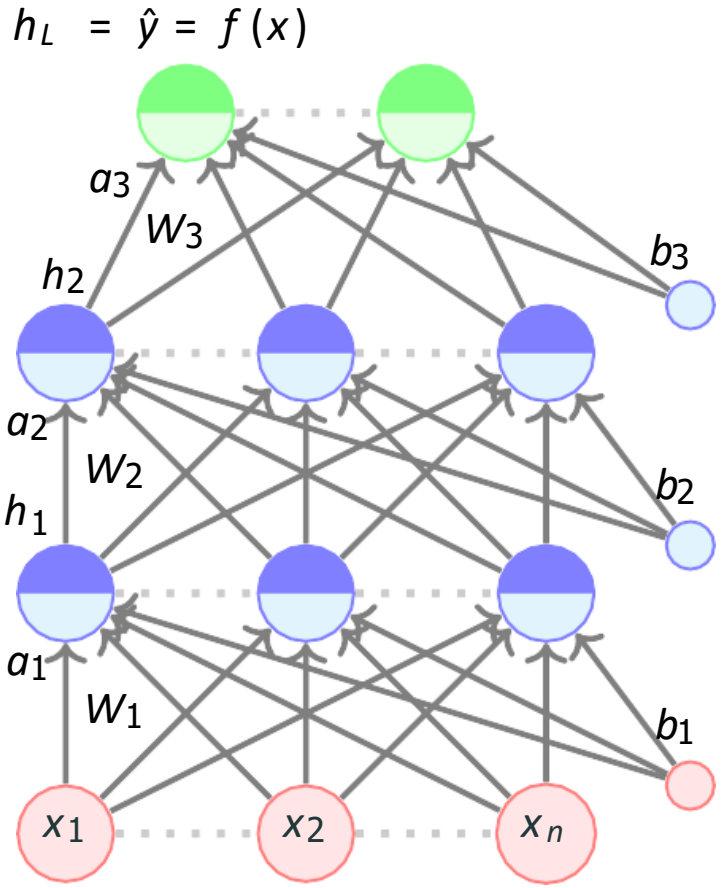
$$a_i(x) = b_i + W_i h_{i-1}(x)$$

The activation at layer i is given by

$$h_i(x) = g(a_i(x))$$

where g is called the activation function (for example, logistic, tanh, linear, *etc.*)

Feedforward Neural Networks (multilayered network of neurons)



The pre-activation at layer i is given by

$$a_i(x) = b_i + W_i h_{i-1}(x)$$

The activation at layer i is given by

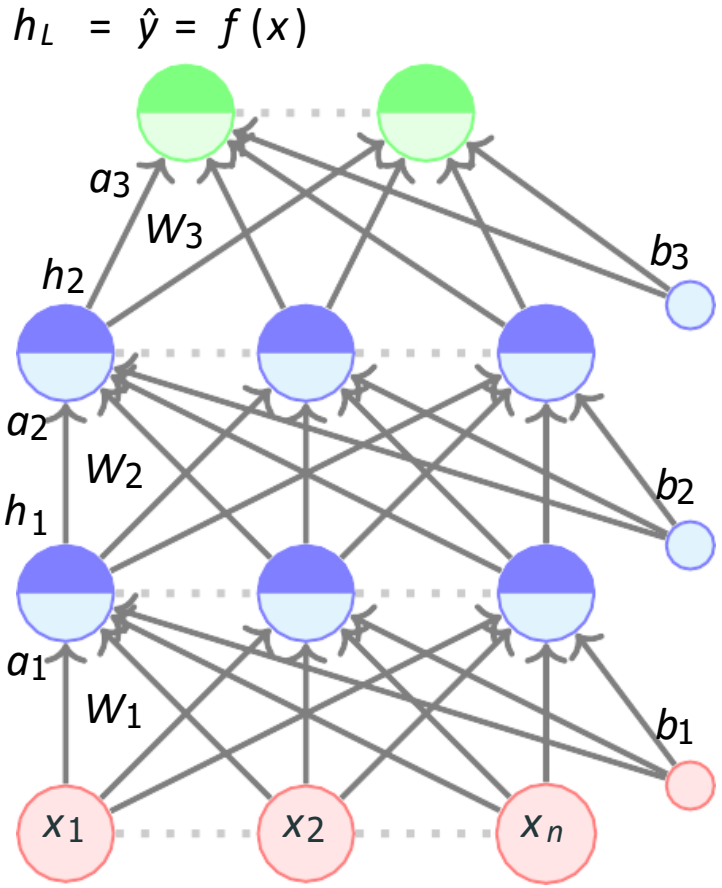
$$h_i(x) = g(a_i(x))$$

where g is called the activation function (for example, logistic, tanh, linear, *etc.*)

The activation at the output layer is given by

$$f(x) = h_L(x) = O(a_L(x))$$

Feedforward Neural Networks(multilayered network of neurons)



The pre-activation at layer i is given by

$$a_i(x) = b_i + W_i h_{i-1}(x)$$

The activation at layer i is given by

$$h_i(x) = g(a_i(x))$$

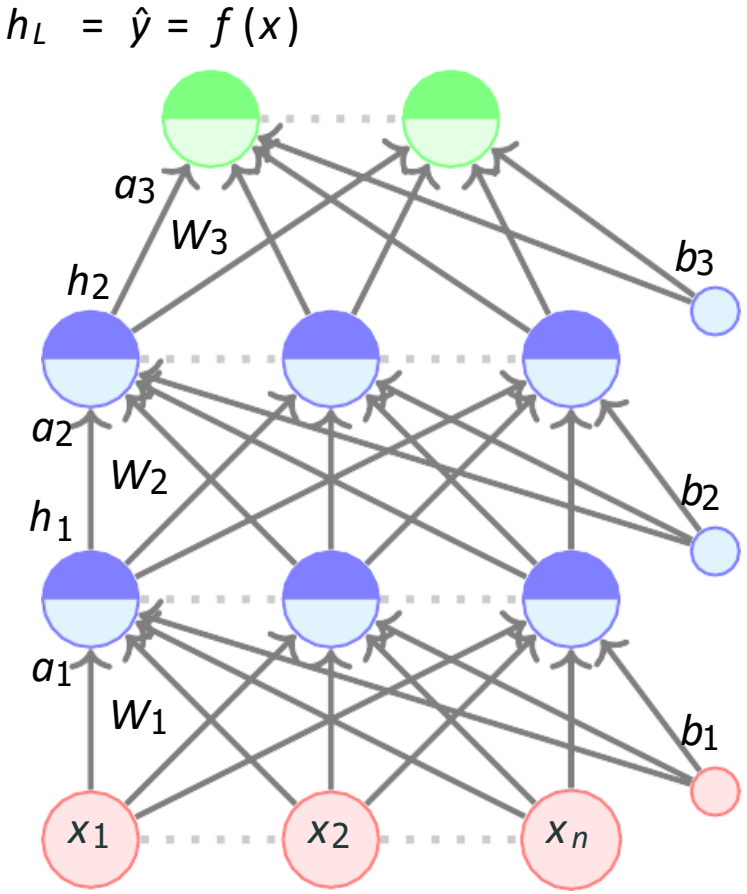
where g is called the activation function (for example, logistic, tanh, linear, *etc.*)

The activation at the output layer is given by

$$f(x) = h_L(x) = O(a_L(x))$$

where O is the output activation function (for example, softmax, linear, *etc.*)

Feedforward Neural Networks (multilayered network of neurons)



The pre-activation at layer i is given by

$$a_i(x) = b_i + W_i h_{i-1}(x)$$

The activation at layer i is given by

$$h_i(x) = g(a_i(x))$$

where g is called the activation function (for example, logistic, tanh, linear, *etc.*)

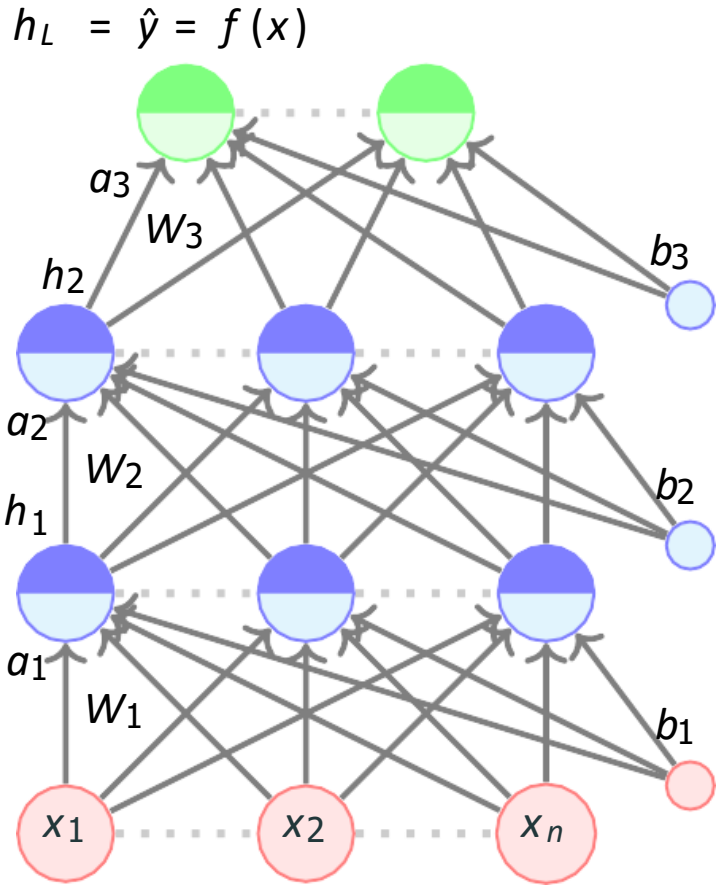
The activation at the output layer is given by

$$f(x) = h_L(x) = O(a_L(x))$$

where O is the output activation function (for example, softmax, linear, *etc.*)

To simplify notation we will refer to $a_i(x)$ as a_i and

Feedforward Neural Networks (multilayered network of neurons)



The pre-activation at layer i is given by

$$a_i = b_i + W_i h_{i-1}$$

The activation at layer i is given by

$$h_i = g(a_i)$$

where g is called the activation function (for example, logistic, tanh, linear, *etc.*)

The activation at the output layer is given by

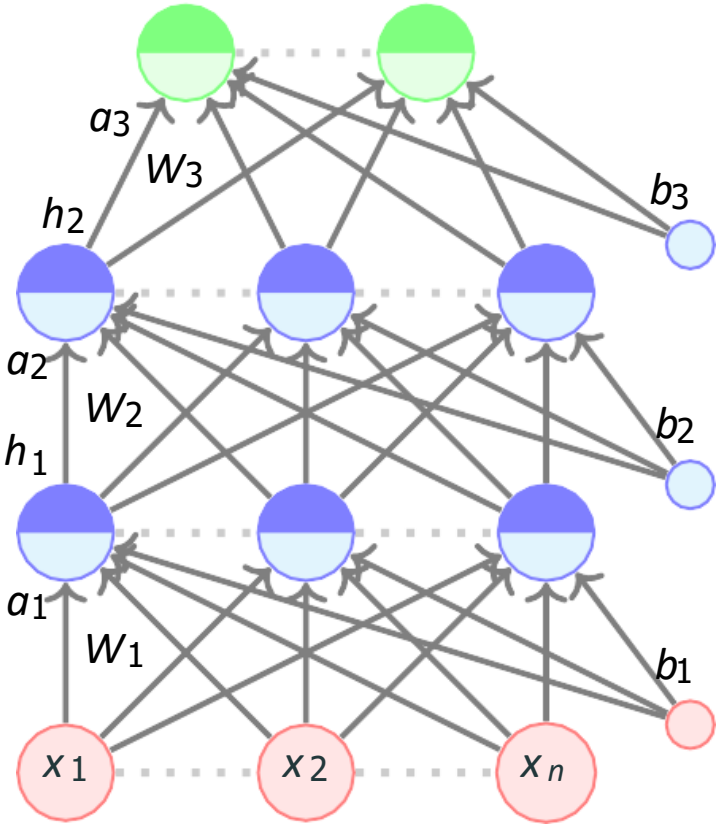
$$f(x) = h_L = O(a_L)$$

where O is the output activation function (for example, softmax, linear, *etc.*)

Feedforward Neural Networks(multilayered network of neurons)

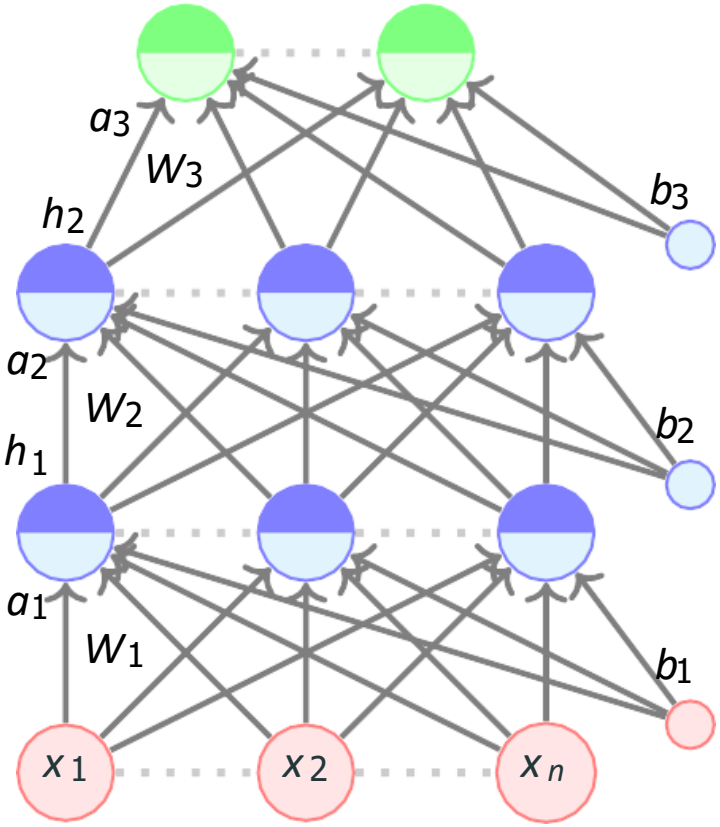
$h_L = \hat{y} = f(x)$

Data: $\{x_i, y_i\}_{i=1}^{N}$



Feedforward Neural Networks(multilayered network of neurons)

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



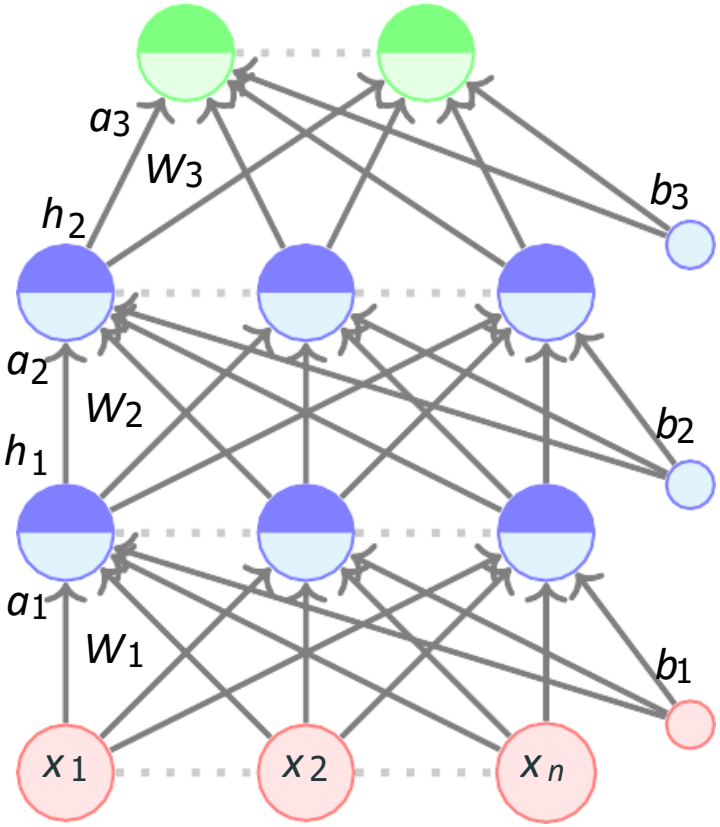
Data: $\{x_i, y_i\}_{i=1}^{N}$

Model:

$$\hat{y_i} = f(x_i) = O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)$$

Feedforward Neural Networks (multilayered network of neurons)

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



Data: $\{x_i, y_i\}_{i=1}^{N}$

Model:

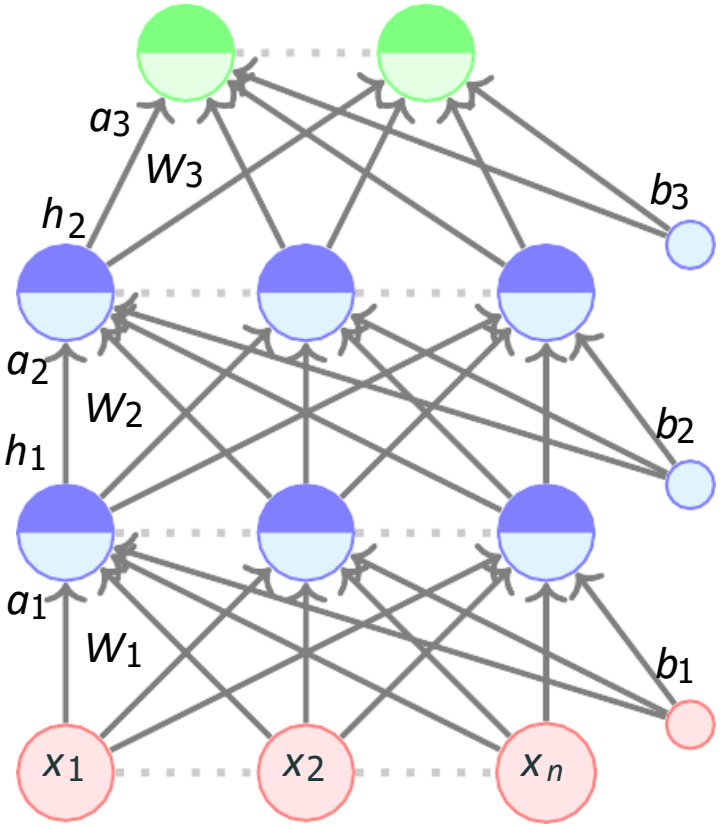
$$\hat{y_i} = f(x_i) = O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)$$

Parameters:

$$\vartheta = W_1, .., W_L, b_1, b_2, ..., b_L \ (L = 3)$$

Feedforward Neural Networks (multilayered network of neurons)

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



Data: $\{x_i, y_i\}_{i=1}^{N}$

Model:

$$\hat{y_i} = f(x_i) = O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)$$

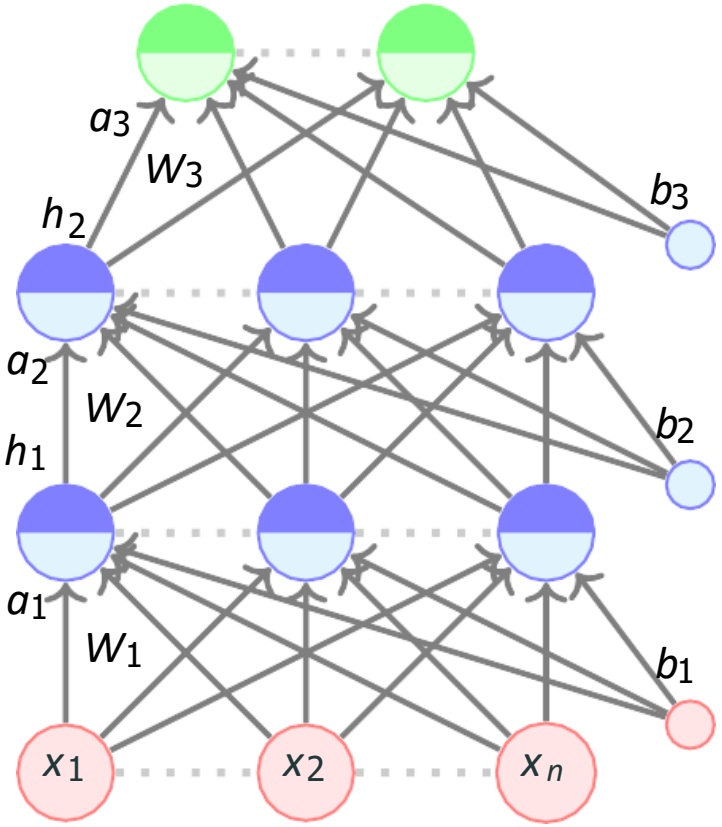
Parameters:

$\vartheta = W_1, .., W_L, b_1, b_2, ..., b_L \ (L = 3)$

Algorithm: Gradient Descent with Back-propagation (we will see soon)

Feedforward Neural Networks (multilayered network of neurons)

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



Data: $\{x_i, y_i\}_{i=1}^{N}$

Model:

$$\hat{y_i} = f(x_i) = O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)$$

Parameters:

$$\vartheta = W_1, .., W_L, b_1, b_2, ..., b_L \ (L = 3)$$

Algorithm: Gradient Descent with Back-propagation (we will see soon)

Objective/Loss/Error function: Say,

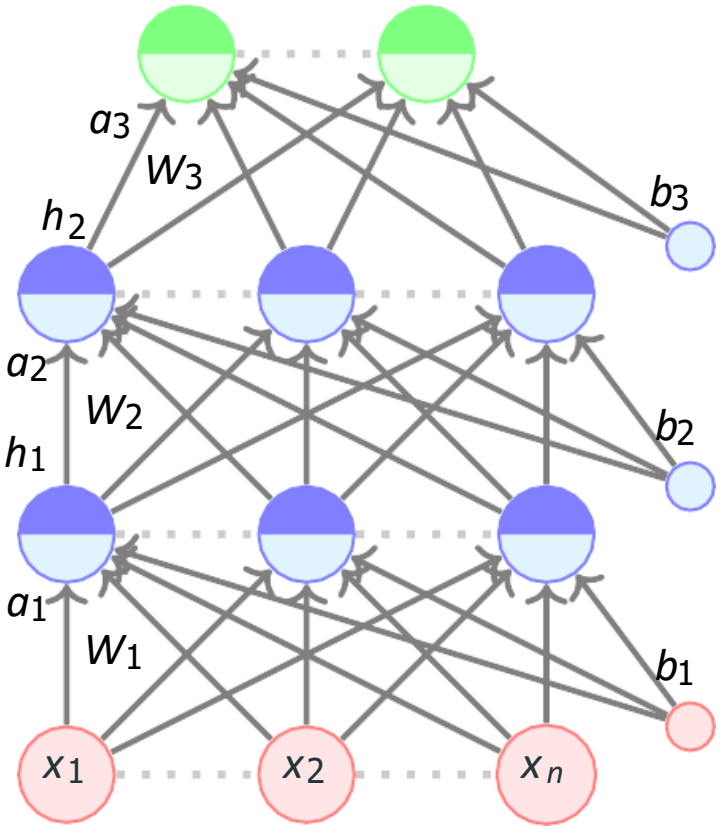
$$min \ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} (\hat{y}_{ij} - y_{ij})^2$$

$$In\ general, min\ L(\vartheta)$$

where $L(\vartheta)$ is some function of the parameters

- **Learning Parameters of Feedforward Neural Networks (Intuition)**

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



Recall our gradient descent algorithm

Algorithm: gradient_descent()

$t \leftarrow 0$;

$max\_iterations \leftarrow 1000$;

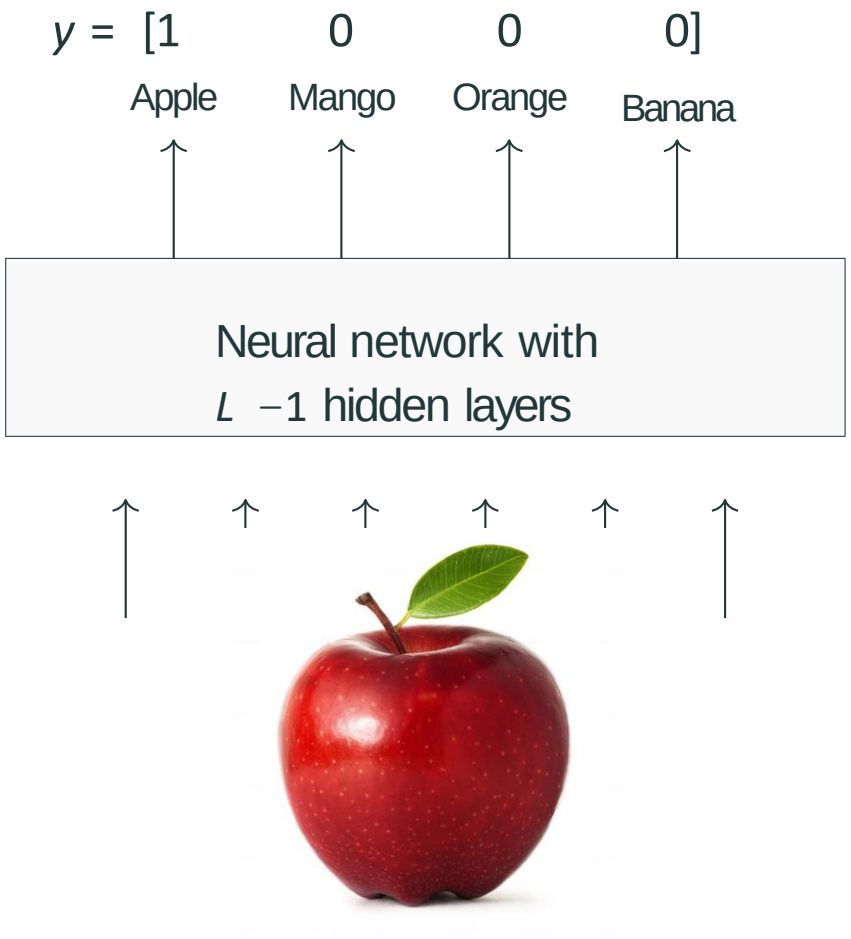
$Initialize \quad w_0, b_0$;

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

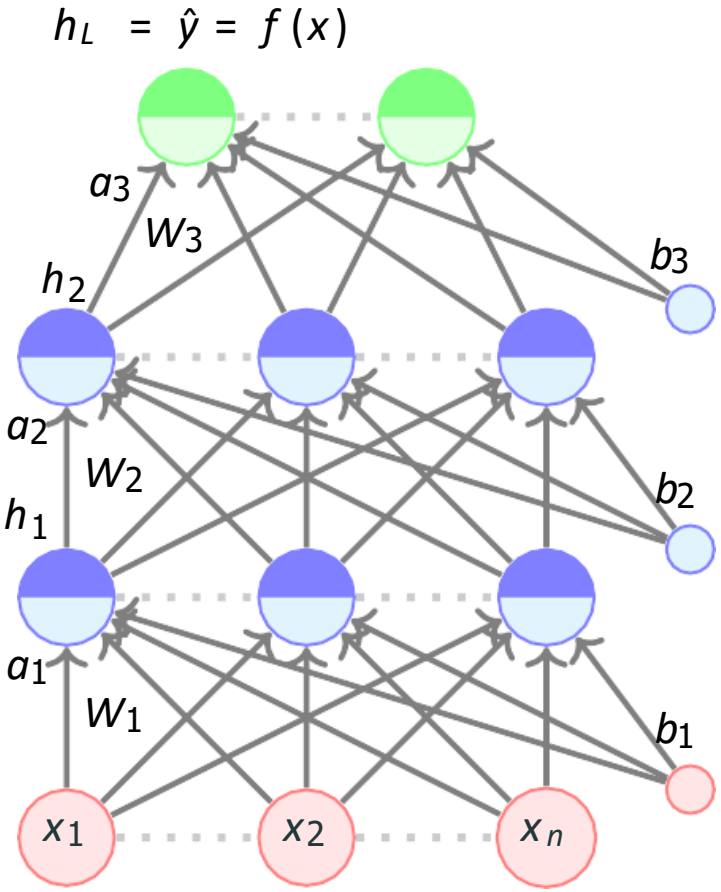


Now let us consider another problem for which a different loss function would be appropriate

Suppose we want to classify an image into 1 of k classes

Here again we could use the squared error loss to capture the deviation

But can you think of a better function?



Notice that y is a probability distribution

Therefore we should also ensure that $\hat{y}$ is a probability distribution

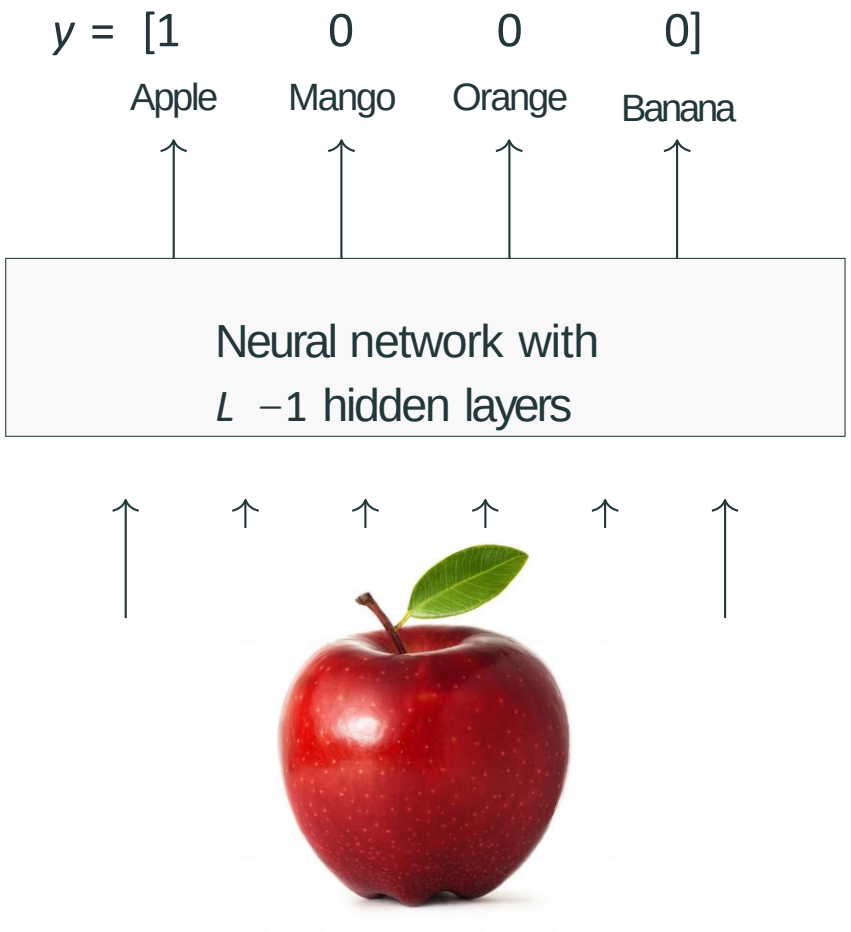
What choice of the output activation 'O' will ensure this ?

$$a_L = W_L h_{L-1} + b_L$$

$$\hat{y_j} = O(a_L)_j = \frac{e^{a_{L,j}}}{\sum_{i=1}^{k} e^{a_{L,i}}}$$

$O(a_L)_j$ is the j^{th} element of $\hat{y}$ and $a_{L,j}$ is the j^{th} element of the vector a_L .

This function is called the *softmax* function



Now that we have ensured that both y & $\hat{y}$ are probability distributions can you think of a function which captures the difference between them?

Cross-entropy

$$L(\vartheta) = -\sum_{c=1}^{k} y_c \log \hat{y}_c$$

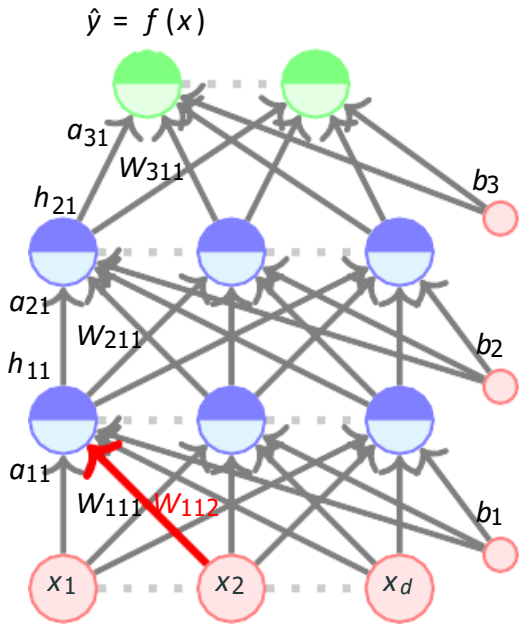
Notice that

$y_c = 1$ if $c = l$ (the true class label)

$= 0$ otherwise

$\therefore L(\vartheta) = -\log \hat{y}_l$

Let us focus on this one weight (W_{112}).



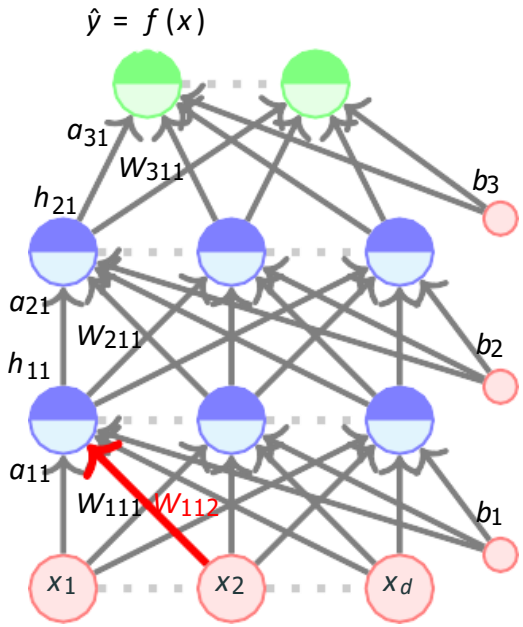
Algorithm: Gradient descent()

$t \leftarrow 0$;
max_iterations $\leftarrow$ 1000;
Initialize ϑ_0;
while t++ < *max_iterations* **do**
  $\vartheta_{t+1} \leftarrow \vartheta_t - \eta \nabla \vartheta_t$;
end

Let us focus on this one weight (W_{112}).

To learn this weight using SGD we need a formula for $\frac{\partial L(\vartheta)}{\partial W_{112}}$.

We will see how to calculate this.



Algorithm: Gradient descent()

$t \leftarrow 0$;

$max\_iterations \leftarrow$ 1000;

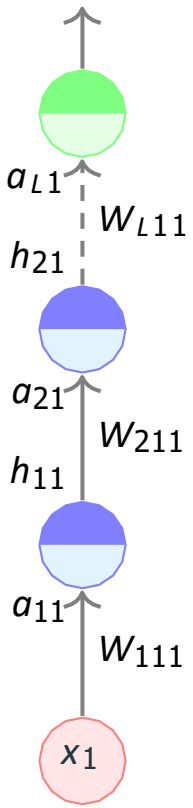
Initialize ϑ_0;

while t++ < $max\_iterations$ **do**

  $\vartheta_{t+1} \leftarrow \vartheta_t - \eta \nabla \vartheta_t$;

end

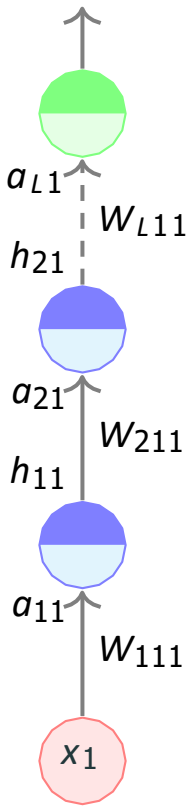
First let us take the simple case when we have a deep but thin network.



First let us take the simple case when we have a deep but thin network.

In this case it is easy to find the derivative by chain rule.

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$



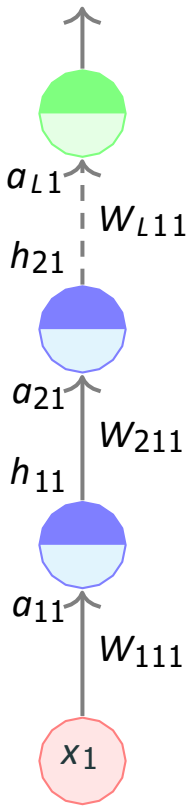
First let us take the simple case when we have a deep but thin network.

In this case it is easy to find the derivative by chain rule.

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial h_{11}} \frac{\partial h_{11}}{\partial W_{111}}$$

(just compressing the chain rule)



First let us take the simple case when we have a deep but thin network.

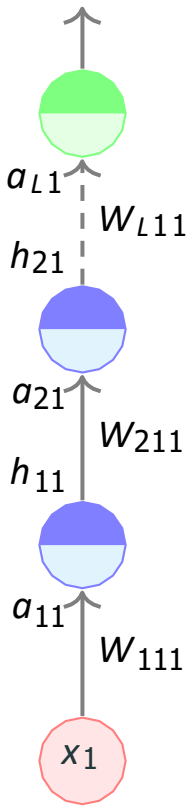
In this case it is easy to find the derivative by chain rule.

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial h_{11}} \frac{\partial h_{11}}{\partial W_{111}}$$

$$\frac{\partial L(\vartheta)}{\partial W_{211}} = \frac{\partial L(\vartheta)}{\partial h_{21}} \frac{\partial h_{21}}{\partial W_{211}}$$

(just compressing the chain rule)



First let us take the simple case when we have a deep but thin network.

In this case it is easy to find the derivative by chain rule.

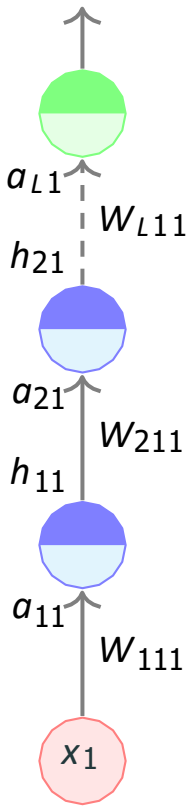
$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial h_{11}} \frac{\partial h_{11}}{\partial W_{111}}$$

$$\frac{\partial L(\vartheta)}{\partial W_{211}} = \frac{\partial L(\vartheta)}{\partial h_{21}} \frac{\partial h_{21}}{\partial W_{211}}$$

$$\frac{\partial L(\vartheta)}{\partial W_{L11}} = \frac{\partial L(\vartheta)}{\partial a_{L1}} \frac{\partial a_{L1}}{\partial W_{L11}}$$

(just compressing the chain rule)



Quantities of interest

Gradient w.r.t. output units

Gradient w.r.t. hidden units

Gradient w.r.t. weights and biases

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3} \frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2} \frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1} \frac{\partial a_1}{\partial W_{111}}$$

- $\frac{\partial L(\vartheta)}{\partial W_{111}}$: Talk to the weight directly
- $\frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}$: Talk to the output layer
- $\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}$: Talk to the previous hidden layer
- $\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}$: Talk to the previous hidden layer
- $\frac{\partial a_1}{\partial W_{111}}$: and now talk to the weights

Our focus is on *Cross entropy loss* and *Softmax* output.

Backpropagation

Computing Gradients w.r.t. the Output Units

Quantities of interest

Gradient w.r.t. output units

Gradient w.r.t. hidden units

Gradient w.r.t. weights

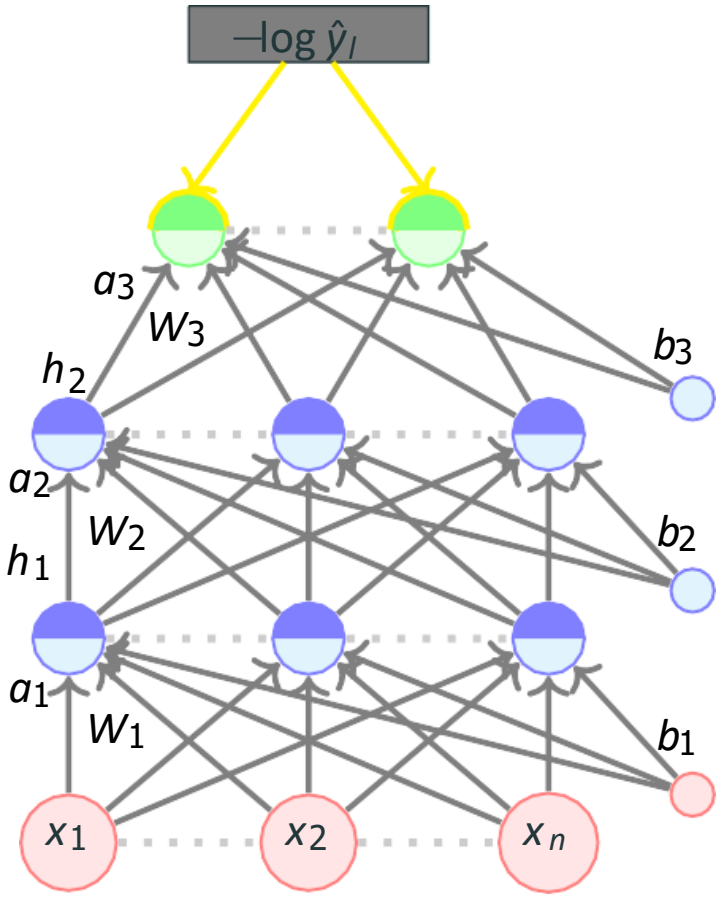
$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \underbrace{\frac{\partial L(\vartheta)}{\partial \hat{y}}}_{\text{Talk to the output layer}} \underbrace{\frac{\partial \hat{y}}{\partial a_3}}_{\text{Talk to the output layer}} \underbrace{\frac{\partial a_3}{\partial h_2}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial h_2}{\partial a_2}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_2}{\partial h_1}}_{\text{and now talk to the weights}} \underbrace{\frac{\partial h_1}{\partial a_1}}_{\text{and now talk to the weights}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\text{and now talk to the weights}}$$

Talk to the weight directly

Our focus is on *Cross entropy loss* and *Softmax* output.

Let us first consider the partial derivative w.r.t. i-th output

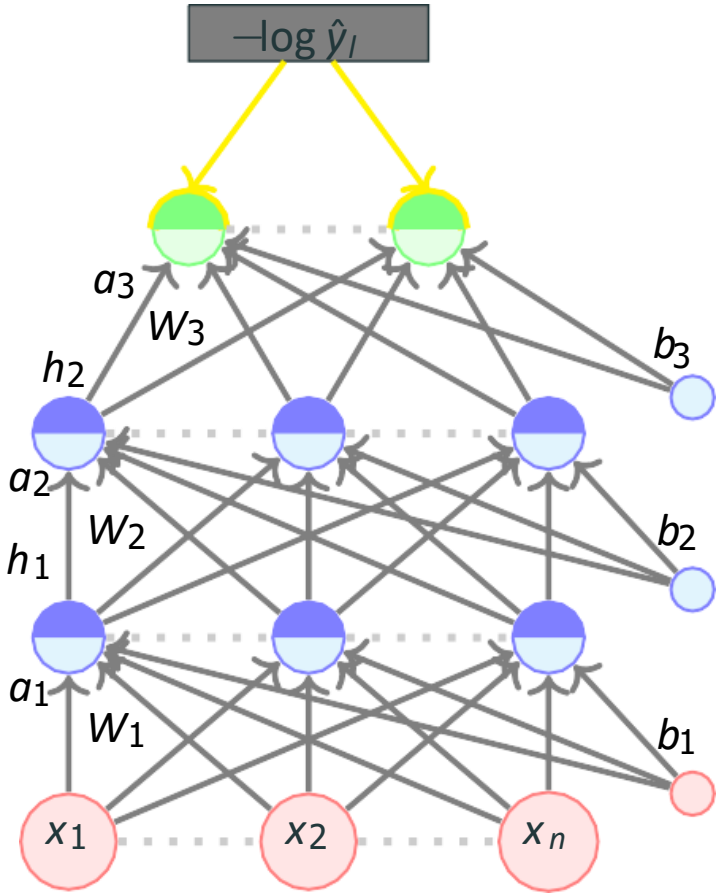
$$L(\vartheta) = -\log \hat{y}_l \quad (l = \text{true class label})$$



Let us first consider the partial derivative w.r.t. i-th output

$$L(\vartheta) = -\log \hat{y}_l \quad (l = \text{true class label})$$

$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = \frac{\partial}{\partial \hat{y}_i}(-\log \hat{y}_l)$$



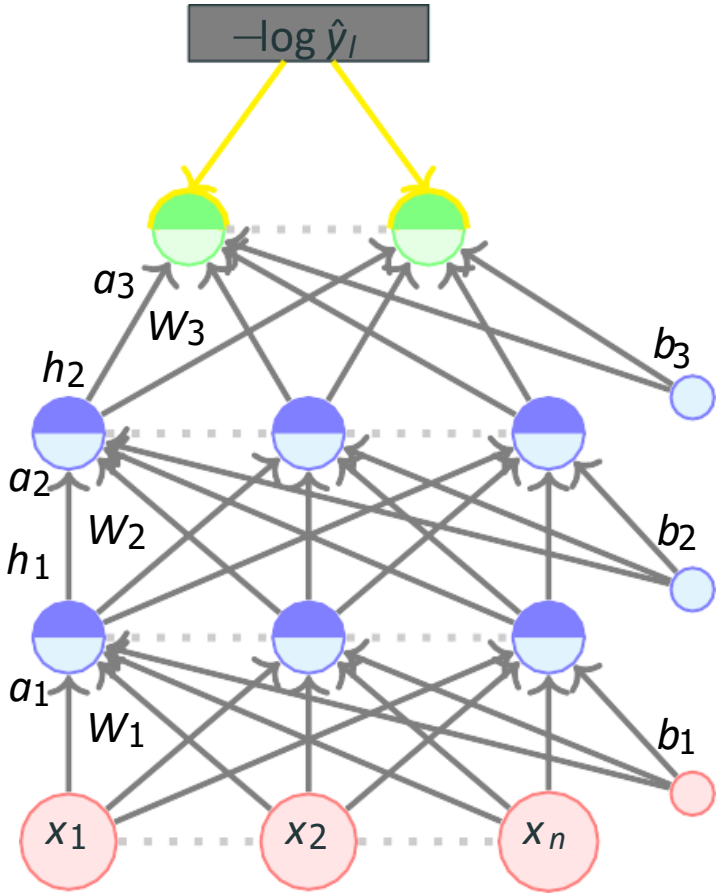
Let us first consider the partial derivative w.r.t. i-th output

$$L(\vartheta) = -\log \hat{y}_l \quad (l = \text{true class label})$$

$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = \frac{\partial}{\partial \hat{y}_i}(-\log \hat{y}_l)$$

$$= -\frac{1}{\hat{y}_l} \quad \text{if } i = l$$

$$= 0 \quad \text{otherwise}$$



Let us first consider the partial derivative w.r.t. i-th output

$$L(\vartheta) = -\log \hat{y}_l \quad (l = \text{true class label})$$

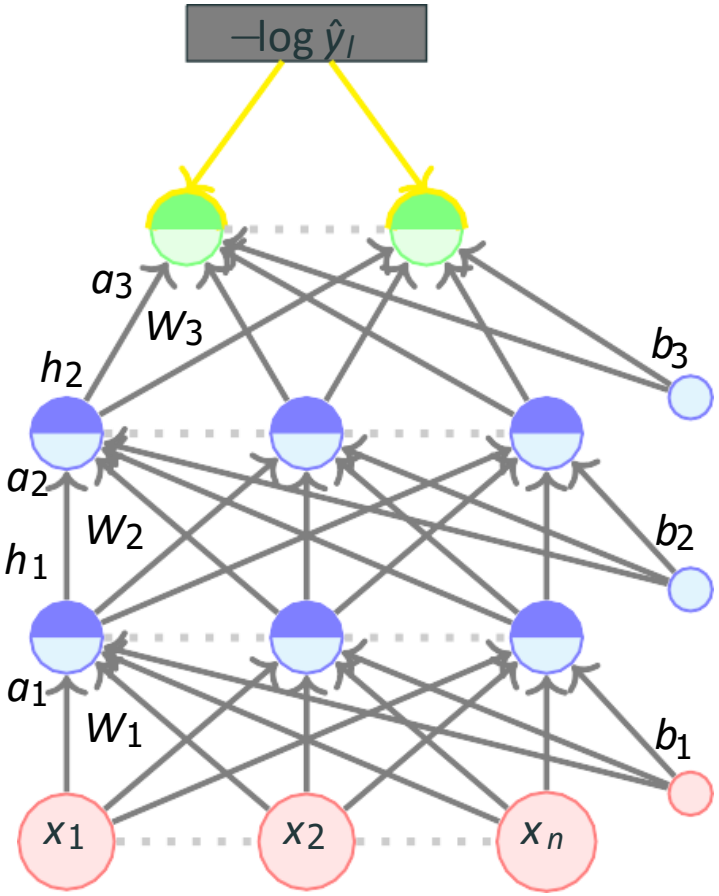
$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = \frac{\partial}{\partial \hat{y}_i}(-\log \hat{y}_l)$$

$$= -\frac{1}{\hat{y}_l} \quad \text{if } i = l$$

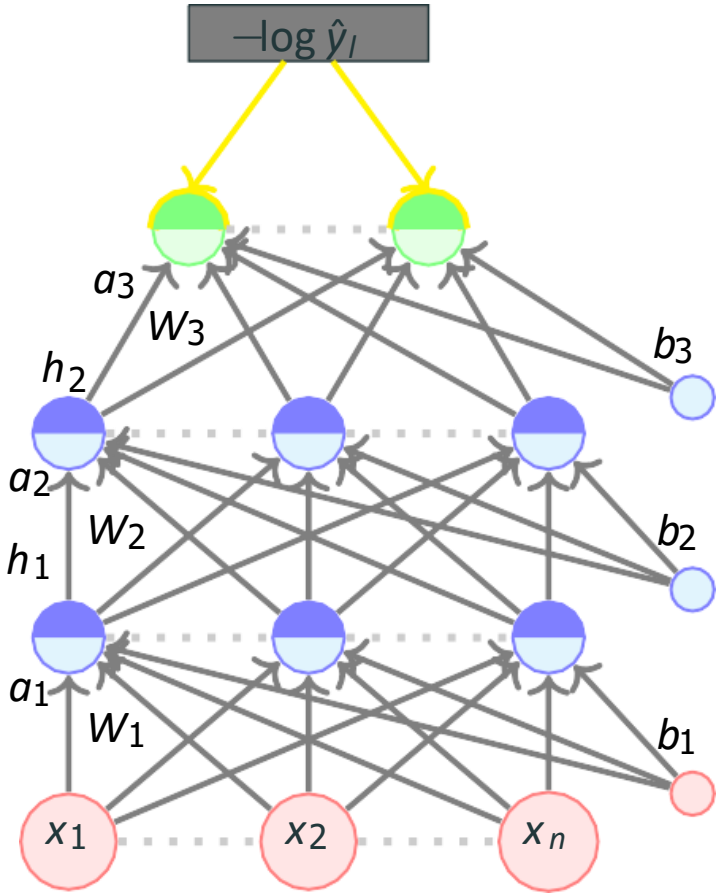
$$= 0 \quad \text{otherwise}$$

More compactly,

$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = -\frac{1(i=l)}{\hat{y}_l}$$



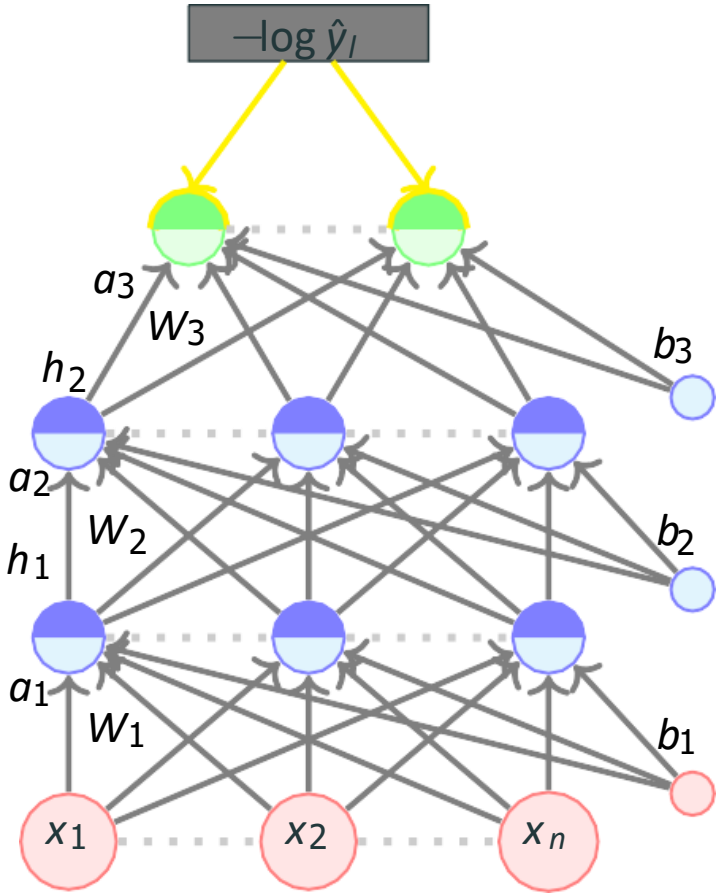
$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = -\frac{1(l=i)}{\hat{y}_l}$$



$$\frac{\partial}{\partial \hat{y}_i}(L\ (\vartheta)) = -\frac{1(l=i)}{\hat{y}_l}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

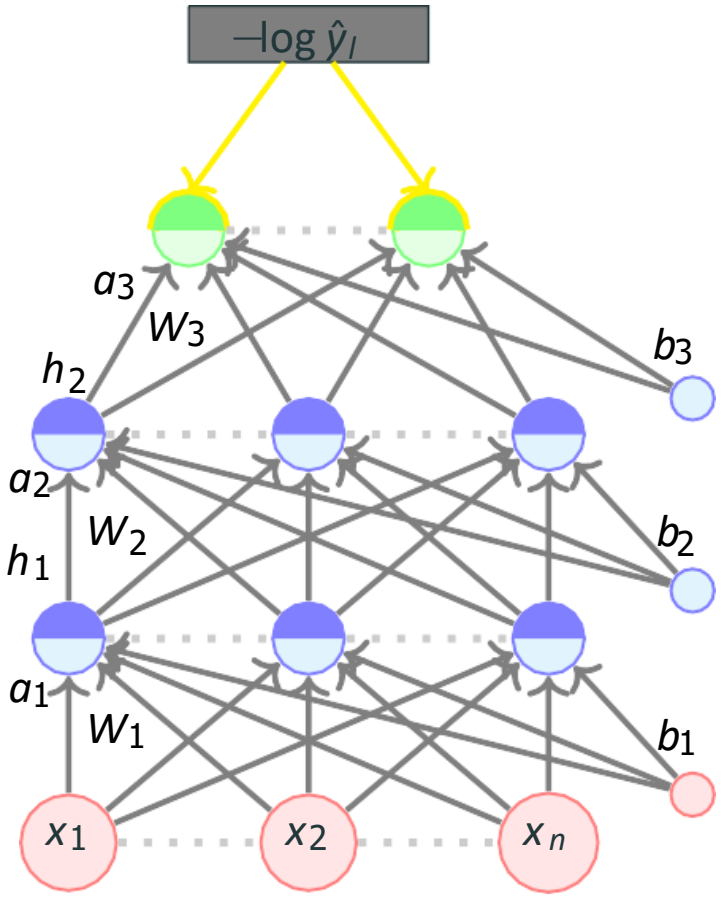
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \\ \\ \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = -\frac{\mathbb{1}(l=i)}{\hat{y}_l}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

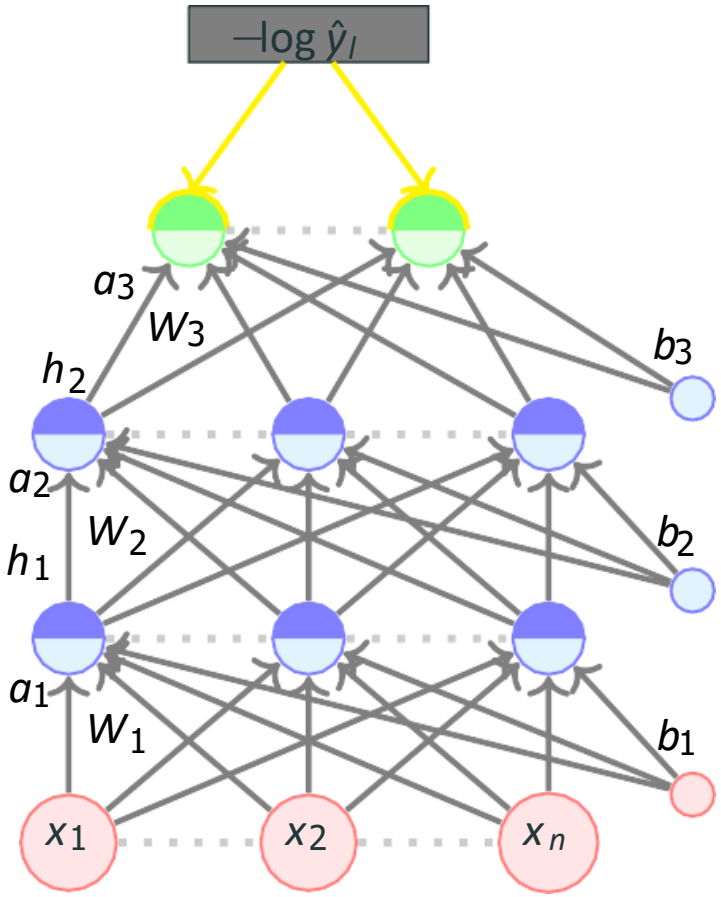
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \\ \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}(L\ (\vartheta)) = -\frac{1(l=i)}{\hat{y}_l}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

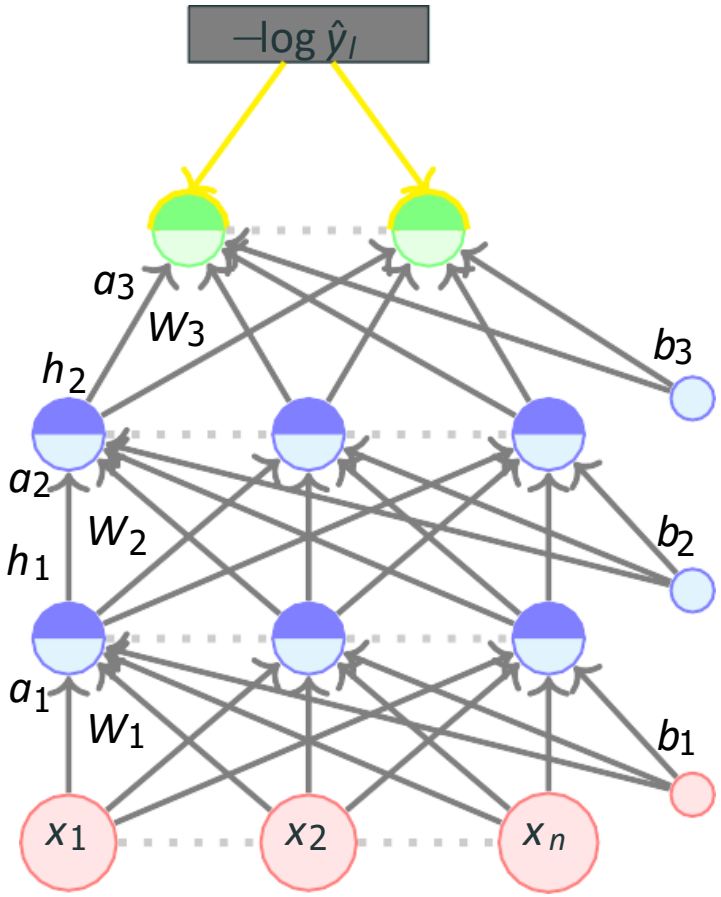
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = -\frac{\mathbb{1}(l=i)}{\hat{y}_l}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

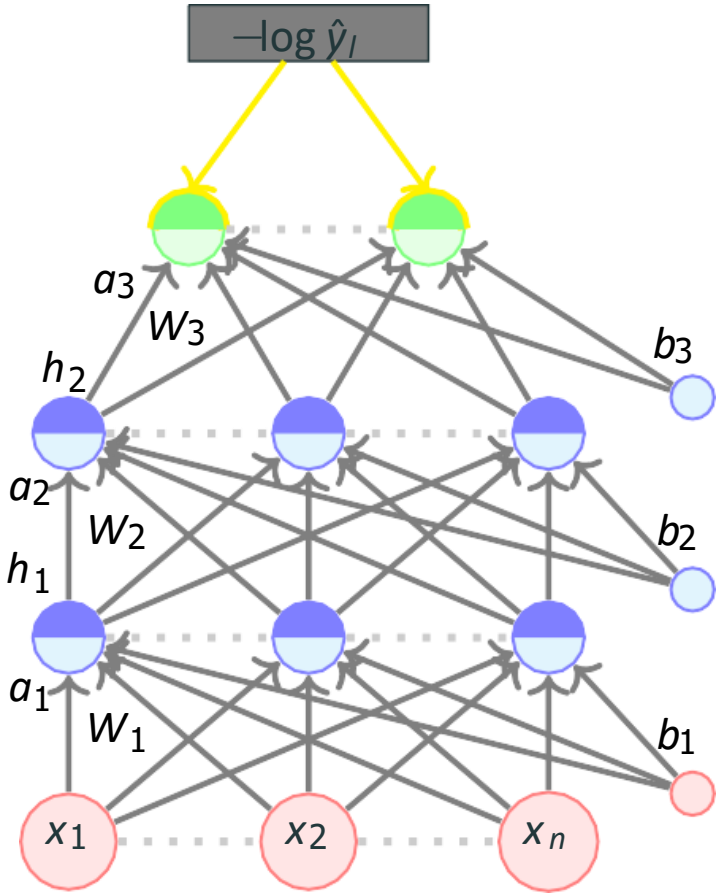
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}(L(\vartheta)) = -\frac{\mathbb{1}(l=i)}{\hat{y}_l}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \\ \mathbb{1}_{\ell=k} \end{bmatrix}$$

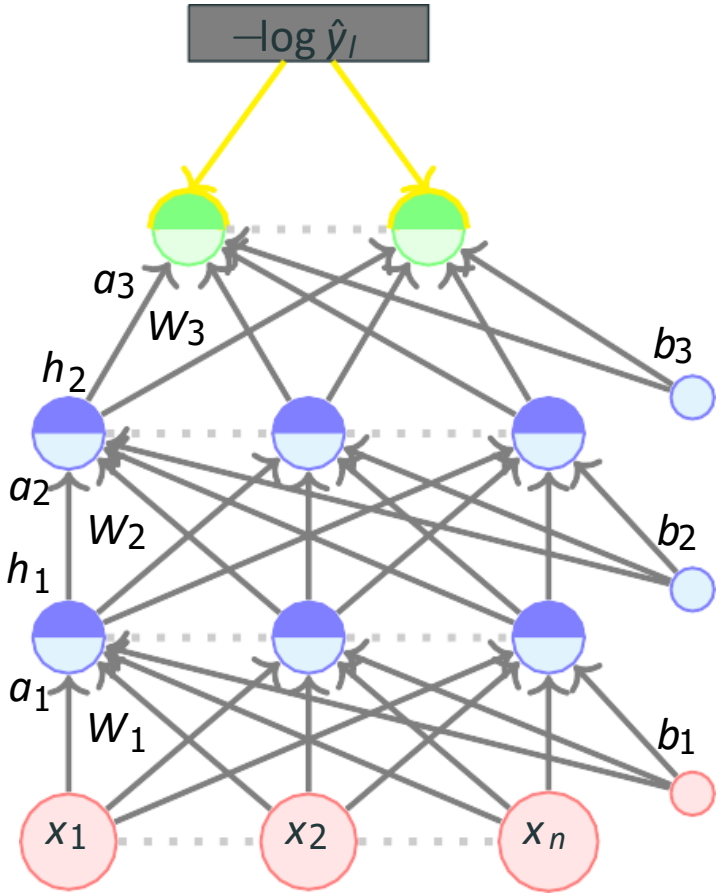


$$\frac{\partial}{\partial \hat{y}_i}(L\ (\vartheta)) = -\frac{\mathbb{1}(l=i)}{\hat{y}_l}$$

We can now talk about the gradient w.r.t. the vector $\hat{y}$

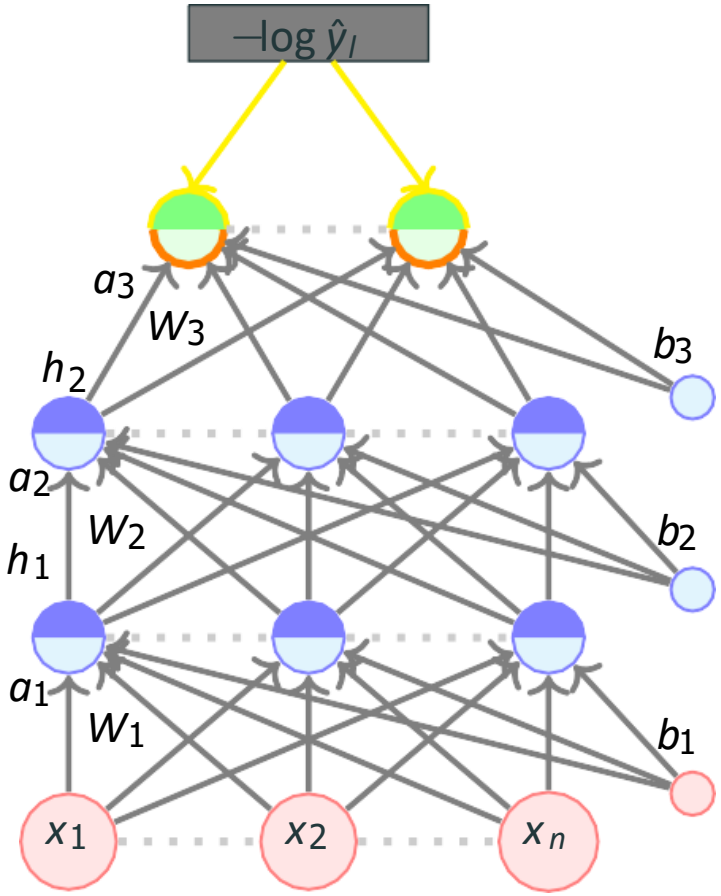
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}\begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \\ \mathbb{1}_{\ell=k} \end{bmatrix}$$
$$= -\frac{1}{\hat{y}_\ell} e_\ell$$

where $e(l)$ is a k-dimensional vector whose l-th element is 1 and all other elements are 0.



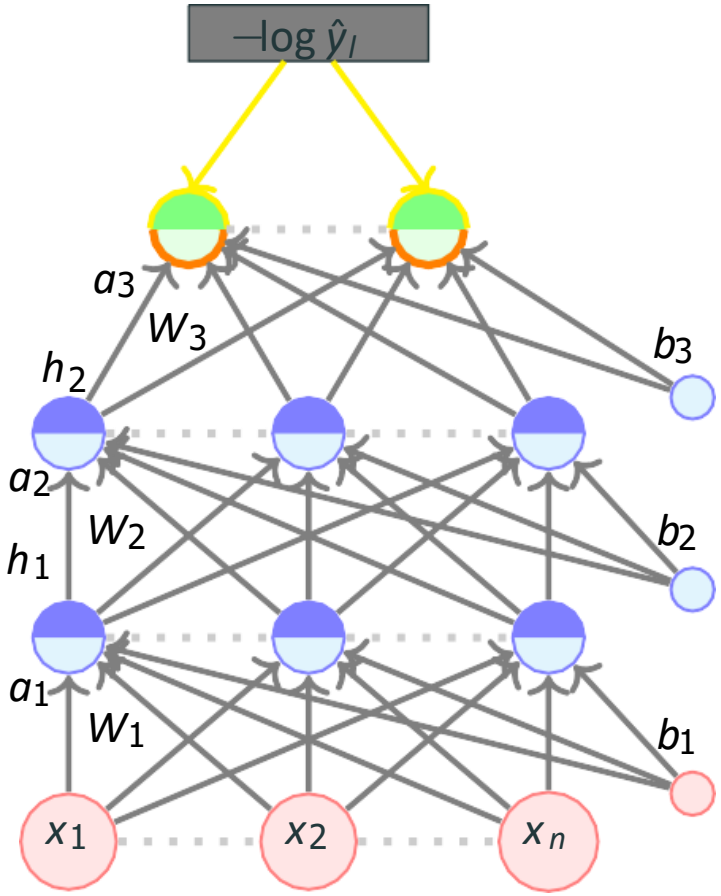
What we are actually interested in is

$$\frac{\partial L(\vartheta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial a_{Li}}$$



What we are actually interested in is

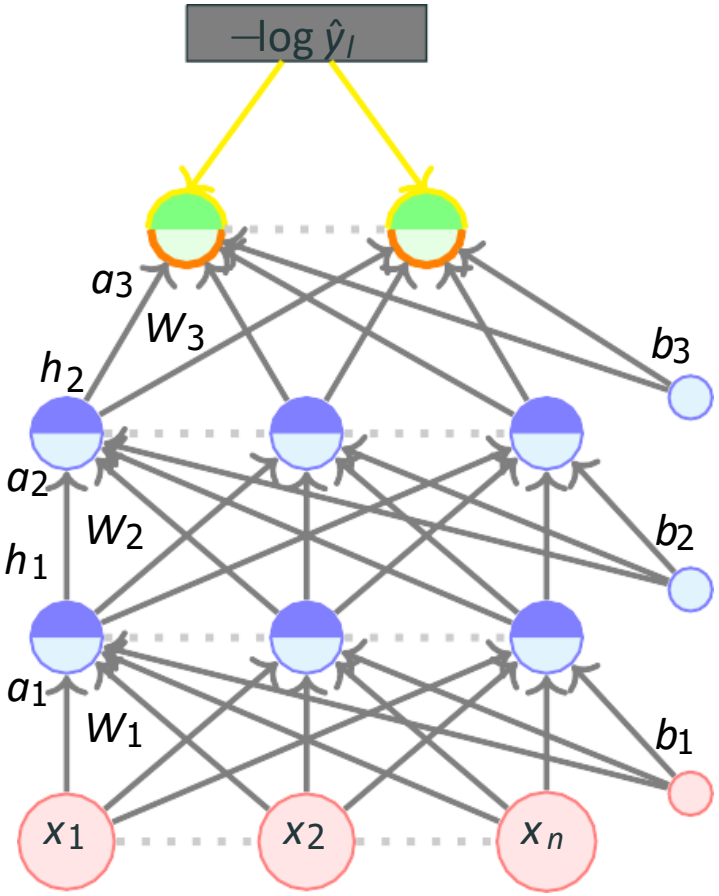
$$\frac{\partial L(\vartheta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_{Li}}$$



What we are actually interested in is

$$\frac{\partial L(\vartheta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_{Li}}$$

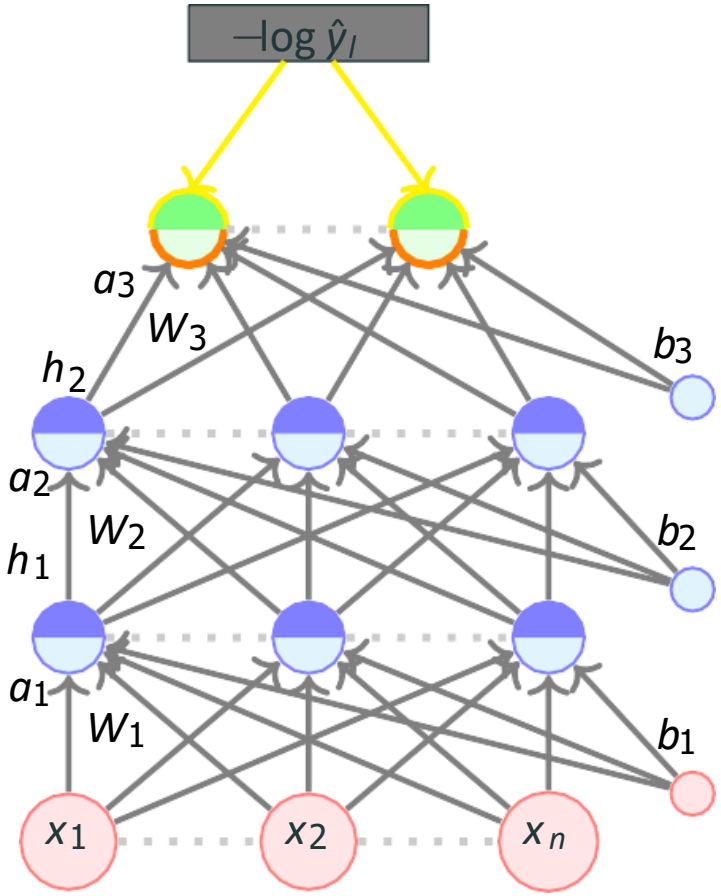
Does $\hat{y}_l$ depend on a_{Li} ?



What we are actually interested in is

$$\frac{\partial L(\vartheta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_{Li}}$$

Does $\hat{y}_l$ depend on a_{Li} ? Indeed, it does.

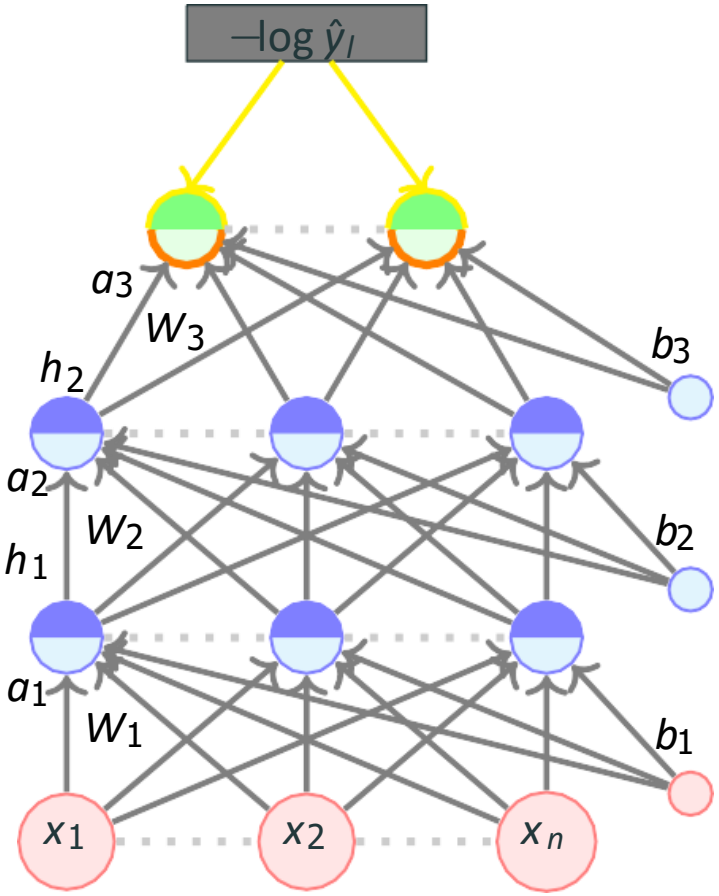


What we are actually interested in is

$$\frac{\partial L(\vartheta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_{Li}}$$

Does $\hat{y}_l$ depend on a_{Li} ? Indeed, it does.

$$\hat{y}_l = \frac{exp(a_{Ll})}{\sum_i exp(a_{Li})}$$



What we are actually interested in is

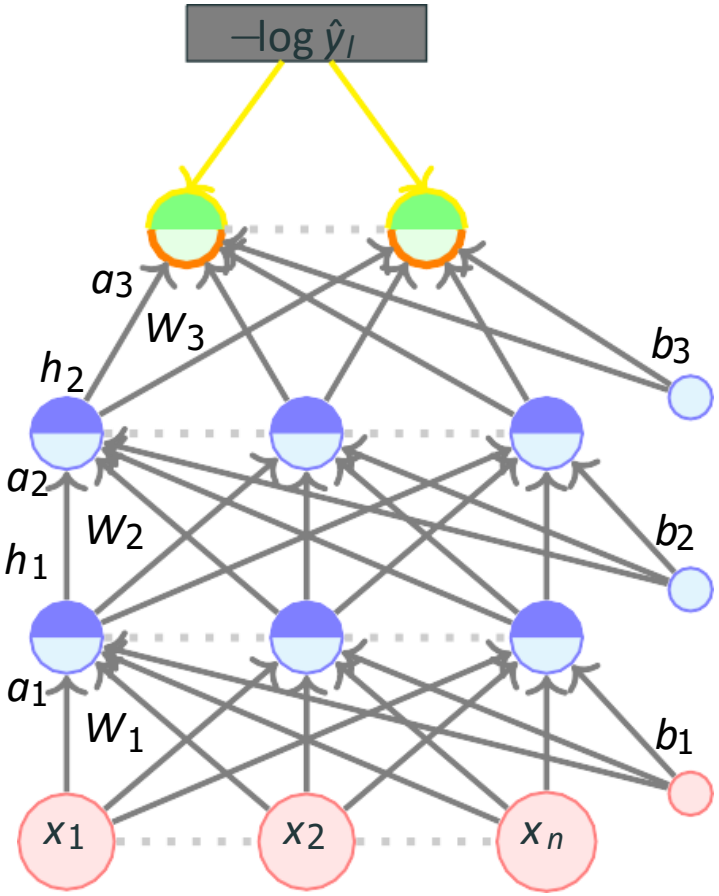
$$\frac{\partial L(\vartheta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial a_{Li}}$$

$$= \frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_{Li}}$$

Does $\hat{y}_l$ depend on a_{Li} ? Indeed, it does.

$$\hat{y}_l = \frac{exp(a_{Ll})}{\sum_i exp(a_{Li})}$$

Having established this, we will now derive the full expression on the next slide



$$\frac{\partial}{\partial a_{L\,i}} -\log \hat{y}_A =$$

$L\ i$

$$\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell = \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell$$

$$= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} softmax(\mathbf{a}_L)_\ell$$

$$= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_\ell}$$

$$= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} (\exp(\mathbf{a}_L)_{i'})^2} \right)$$

$$= \frac{-1}{\hat{y}_\ell} \left(\frac{\mathbb{1}_{(\ell=i)} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \frac{\exp(\mathbf{a}_L)_i}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \right)$$

$$= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} softmax(\mathbf{a}_L)_\ell - softmax(\mathbf{a}_L)_\ell softmax(\mathbf{a}_L)_i \right)$$

$$\frac{\partial \frac{g(x)}{h(x)}}{\partial x} = \frac{\partial g(x)}{\partial x} \frac{1}{h(x)} - \frac{g(x)}{h(x)^2} \frac{\partial h(x)}{\partial x}$$

$$
\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} softmax(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_\ell} \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} (\exp(\mathbf{a}_L)_{i'})^2} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\mathbb{1}_{(\ell=i)} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \frac{\exp(\mathbf{a}_L)_i}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} softmax(\mathbf{a}_L)_\ell - softmax(\mathbf{a}_L)_\ell softmax(\mathbf{a}_L)_i \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} \hat{y}_\ell - \hat{y}_\ell \hat{y}_i \right)
\end{aligned}
$$

$$
\frac{\partial \frac{g(x)}{h(x)}}{\partial x} = \frac{\partial g(x)}{\partial x} \frac{1}{h(x)} - \frac{g(x)}{h(x)^2} \frac{\partial h(x)}{\partial x}
$$

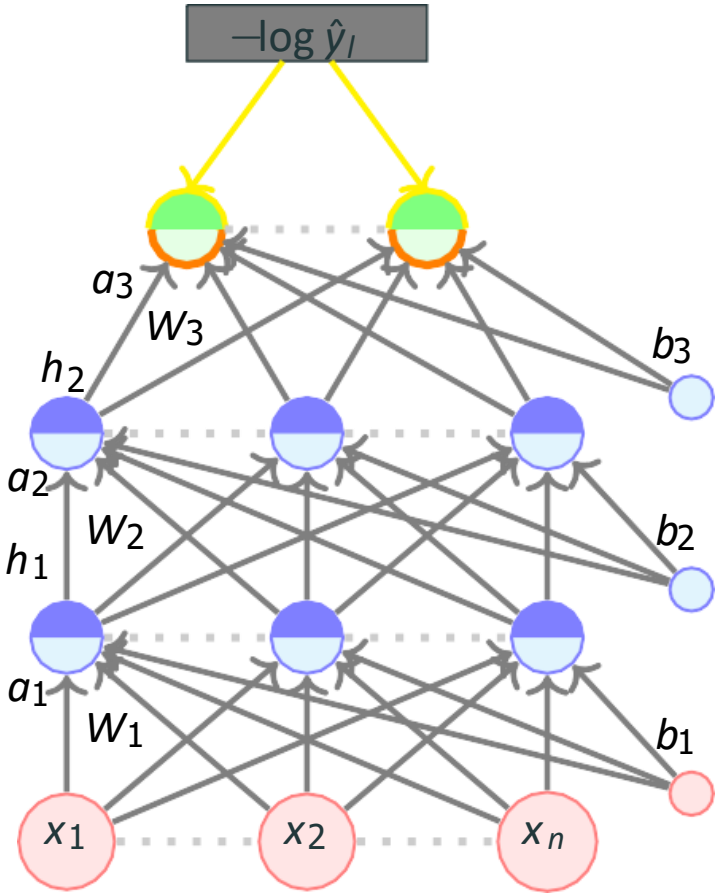
$$
\begin{aligned}
\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell &= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} softmax(\mathbf{a}_L)_\ell \\
&= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_\ell} \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(\mathbf{a}_L)_{i'} \right)}{(\sum_{i'} (\exp(\mathbf{a}_L)_{i'})^2} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\frac{\mathbb{1}_{(\ell=i)} \exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} - \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \frac{\exp(\mathbf{a}_L)_i}{\sum_{i'} \exp(\mathbf{a}_L)_{i'}} \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} softmax(\mathbf{a}_L)_\ell - softmax(\mathbf{a}_L)_\ell softmax(\mathbf{a}_L)_i \right) \\
&= \frac{-1}{\hat{y}_\ell} \left(\mathbb{1}_{(\ell=i)} \hat{y}_\ell - \hat{y}_\ell \hat{y}_i \right) \\
&= -\left(\mathbb{1}_{(\ell=i)} - \hat{y}_i \right)
\end{aligned}
$$

$$
\frac{\partial \frac{g(x)}{h(x)}}{\partial x} = \frac{\partial g(x)}{\partial x} \frac{1}{h(x)} - \frac{g(x)}{h(x)^2} \frac{\partial h(x)}{\partial x}
$$

So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(1_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

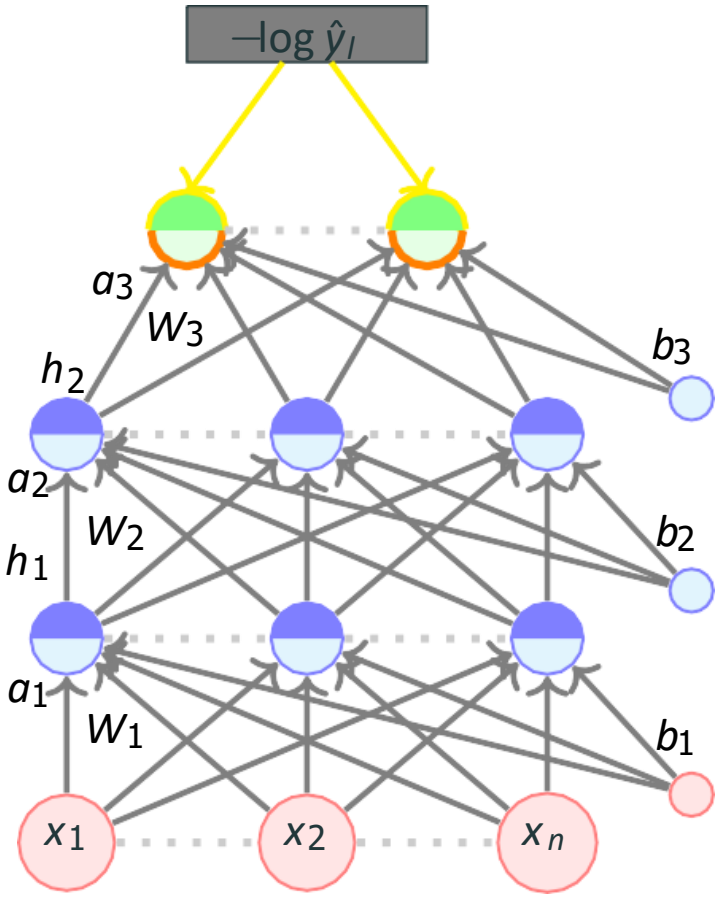


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(1_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} L(\vartheta)$$

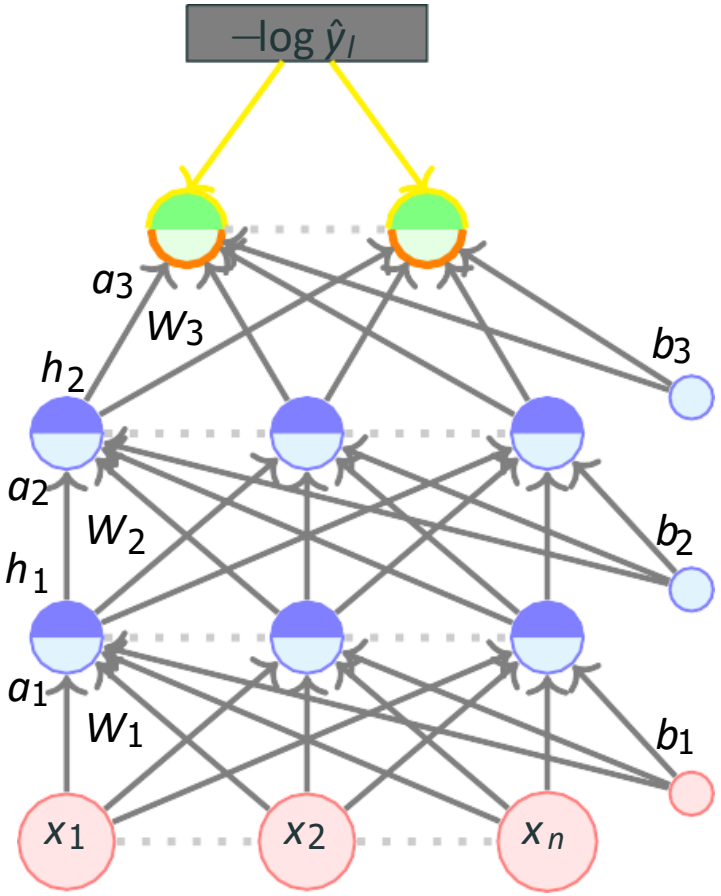


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(\mathbb{1}_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \\ \end{bmatrix}$$

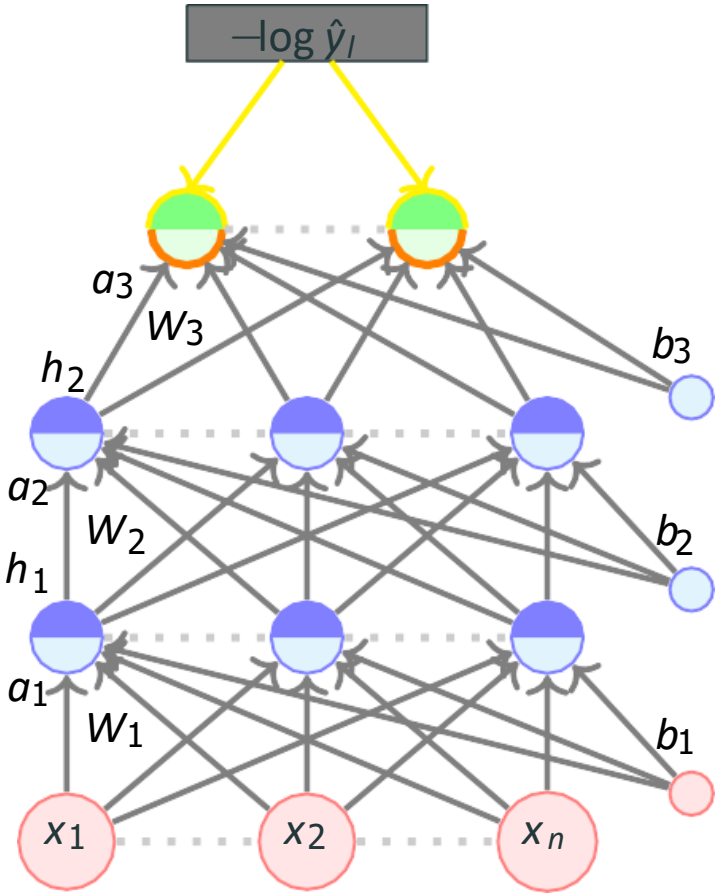


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(\mathbb{1}_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \end{bmatrix}$$

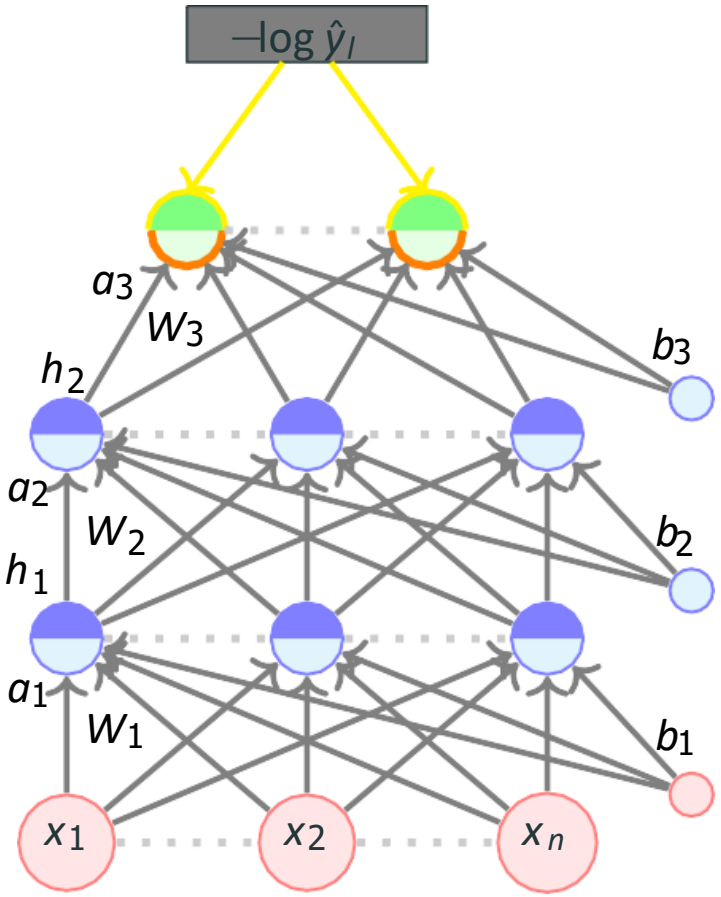


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(1_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix}$$

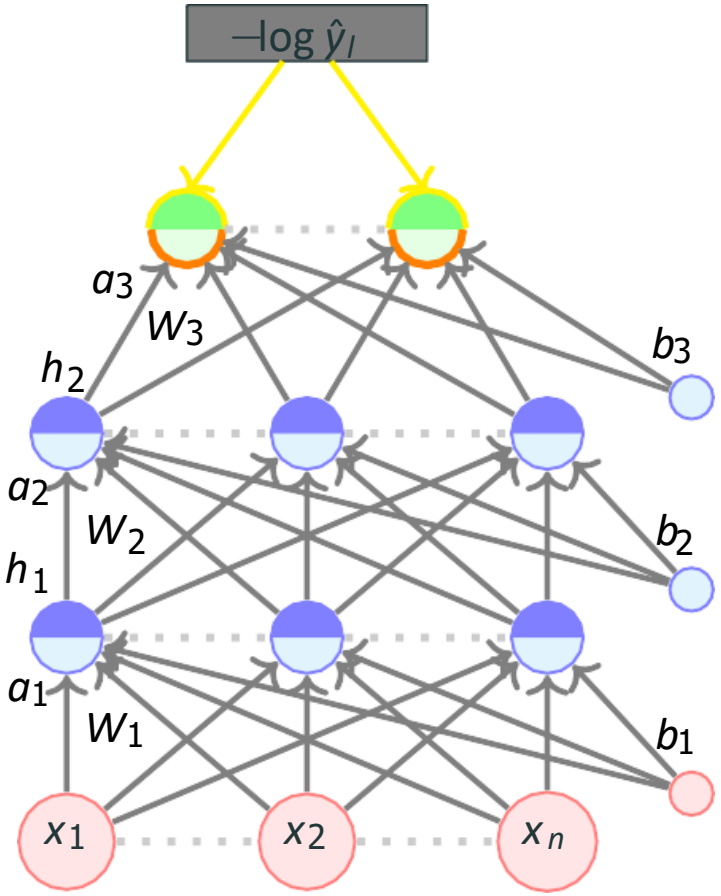


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(1_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

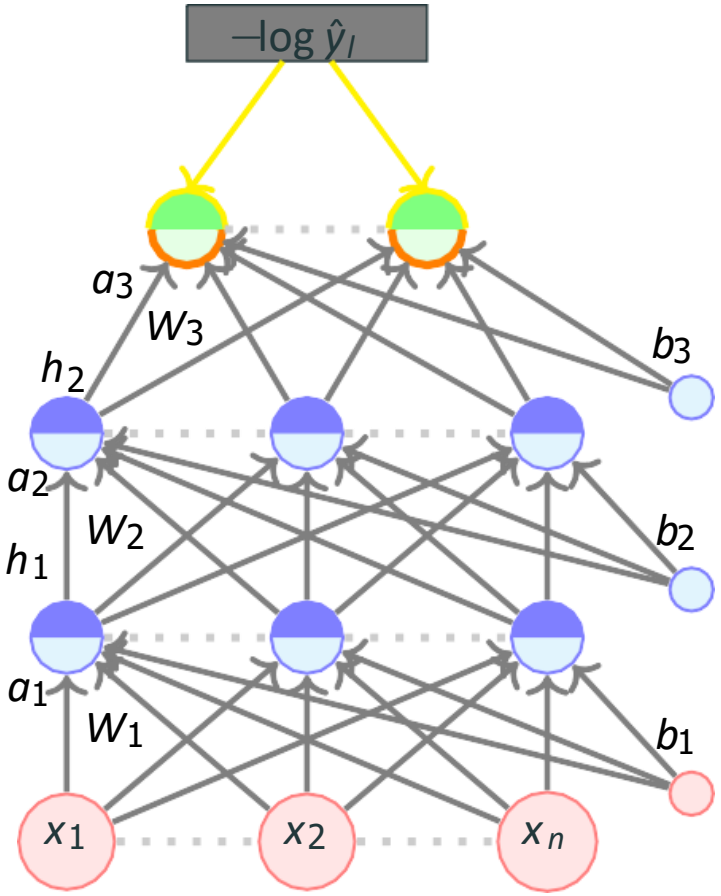


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(\mathbb{1}_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ \\ \\ \end{bmatrix}$$

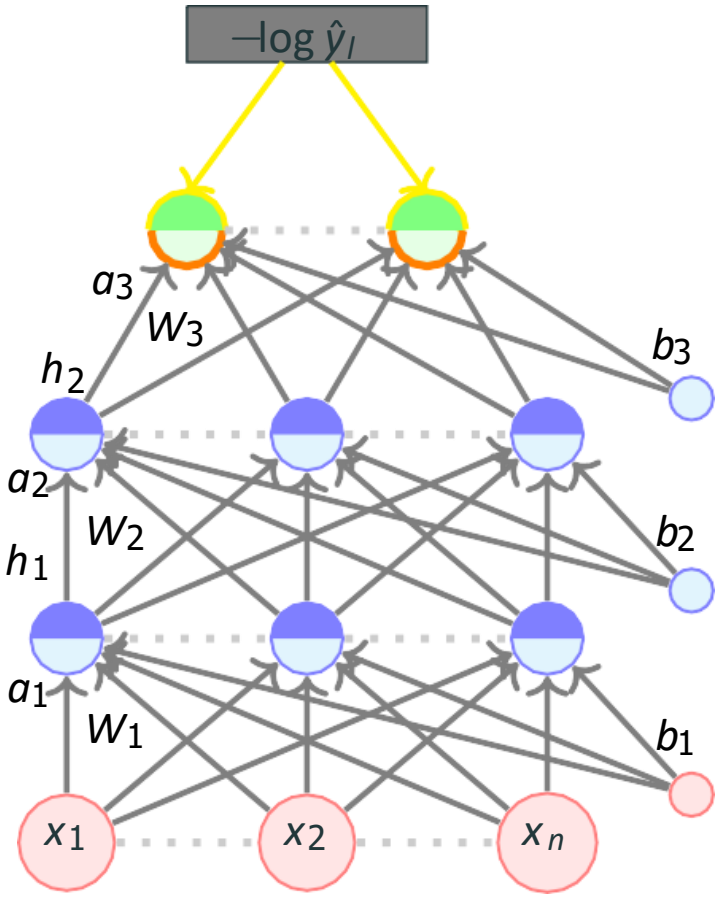


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(\mathbb{1}_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ -(\mathbb{1}_{\ell=2} - \hat{y}_2) \\ \\ \end{bmatrix}$$

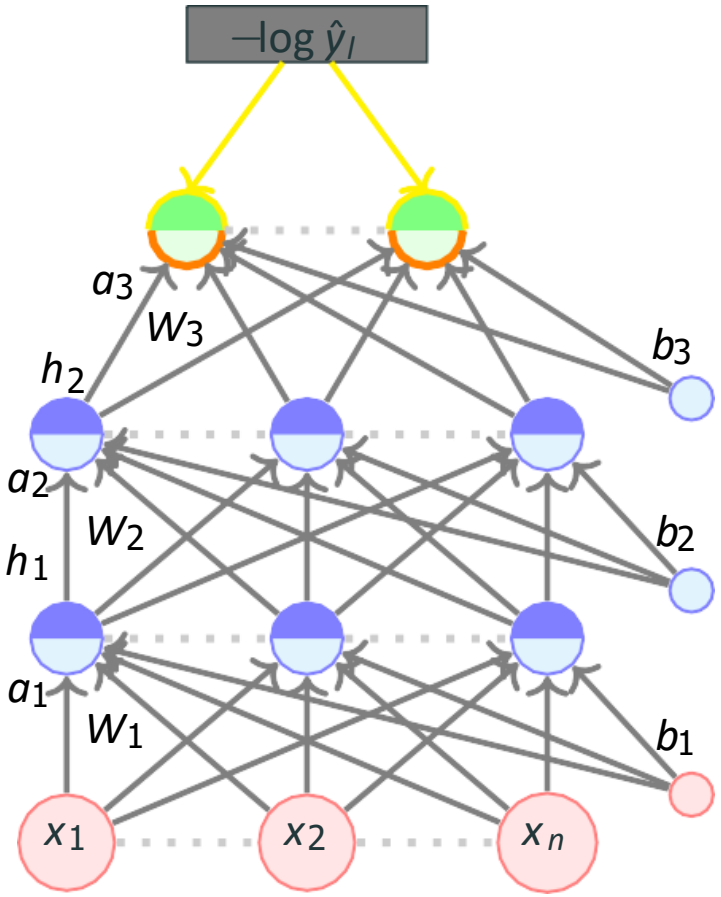


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(\mathbb{1}_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ -(\mathbb{1}_{\ell=2} - \hat{y}_2) \\ \vdots \\ \end{bmatrix}$$

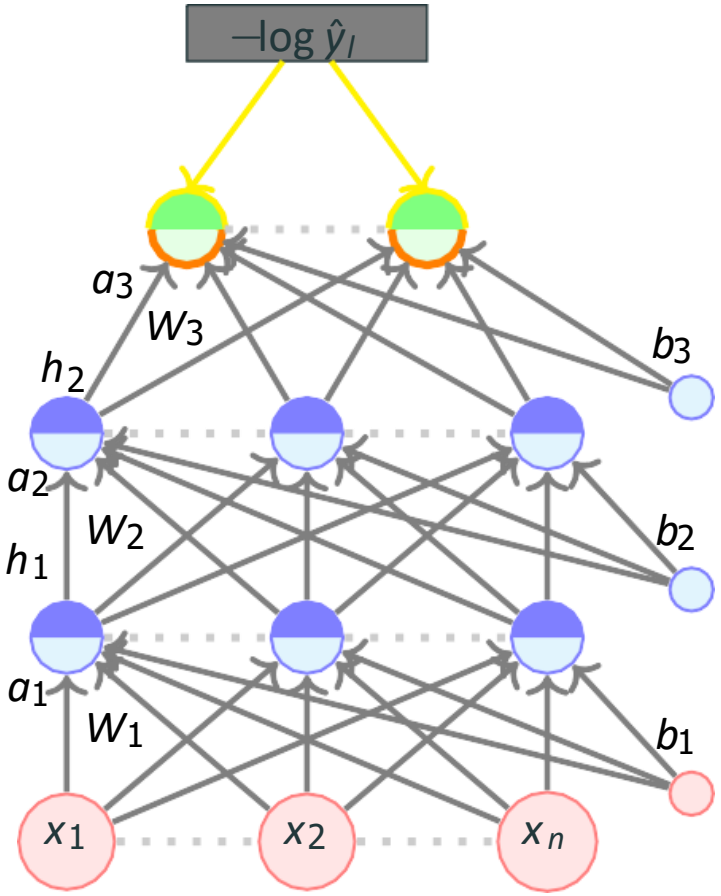


So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(\mathbb{1}_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ -(\mathbb{1}_{\ell=2} - \hat{y}_2) \\ \vdots \\ -(\mathbb{1}_{\ell=k} - \hat{y}_k) \end{bmatrix}$$



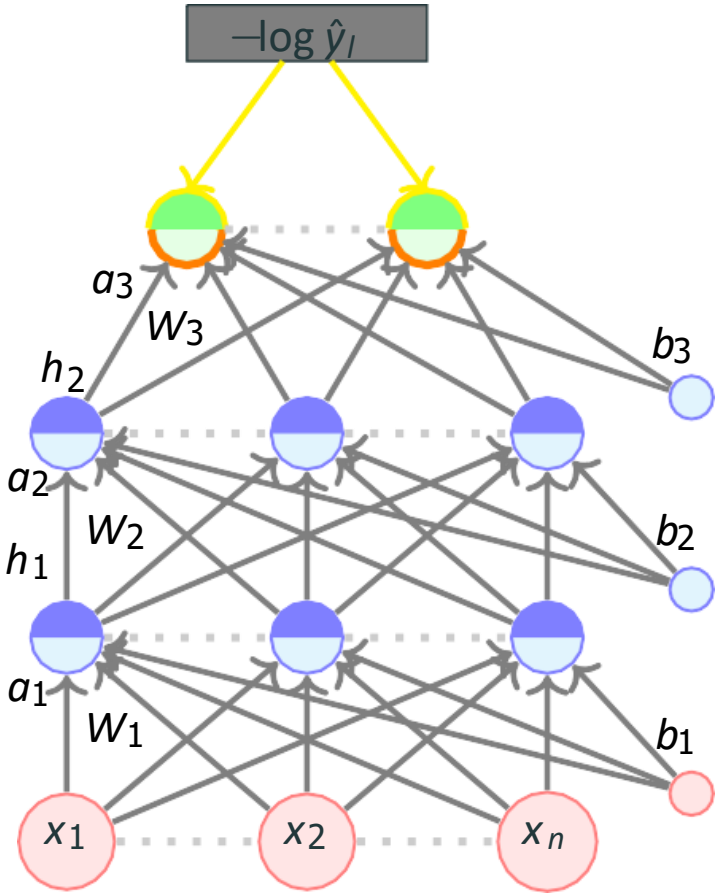
So far we have derived the partial derivative w.r.t. the i-th element of $\mathbf{a}_L$

$$\frac{\partial L(\vartheta)}{\partial a_{L,i}} = -(\mathbb{1}_{l=i} - \hat{y}_i)$$

We can now write the gradient w.r.t. the vector $\mathbf{a}_L$

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{Lk}} \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ -(\mathbb{1}_{\ell=2} - \hat{y}_2) \\ \vdots \\ -(\mathbb{1}_{\ell=k} - \hat{y}_k) \end{bmatrix}$$

$$= -(\mathbf{e}(\ell) - \hat{y})$$



• Backpropagation

Computing Gradients w.r.t. Hidden Units

Quantities of interest (roadmap for the remaining part):

Gradient w.r.t. output units

Gradient w.r.t. hidden units

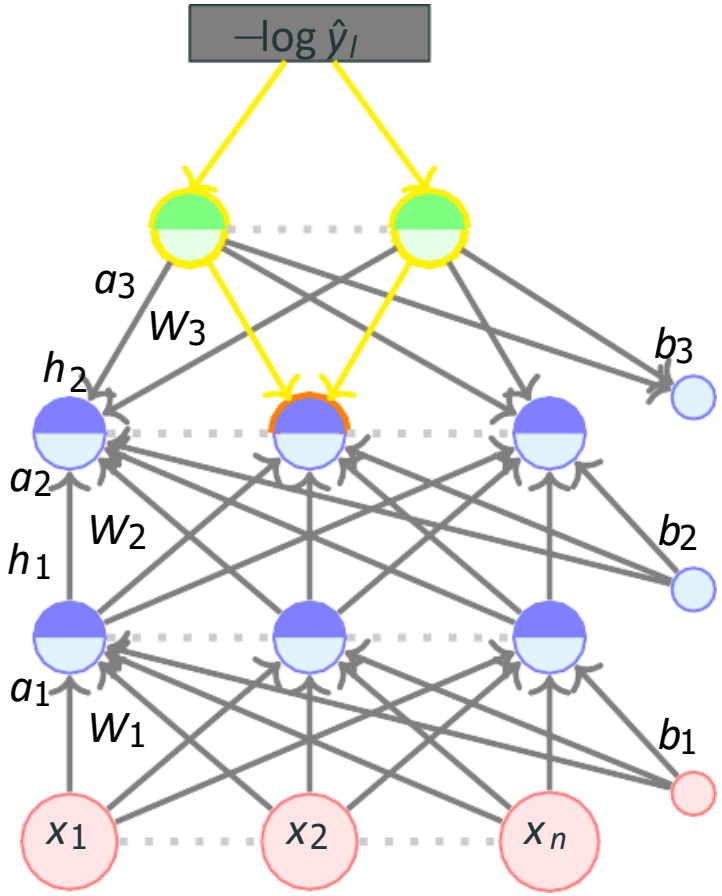
Gradient w.r.t. weights and biases

$$\frac{\partial L(\vartheta)}{\partial W_{111}} = \frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3} \frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2} \frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1} \frac{\partial a_1}{\partial W_{111}}$$

- $\frac{\partial L(\vartheta)}{\partial W_{111}}$: Talk to the weight directly
- $\frac{\partial L(\vartheta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}$: Talk to the output layer
- $\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}$: Talk to the previous hidden layer
- $\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}$: Talk to the previous hidden layer
- $\frac{\partial a_1}{\partial W_{111}}$: and now talk to the weights

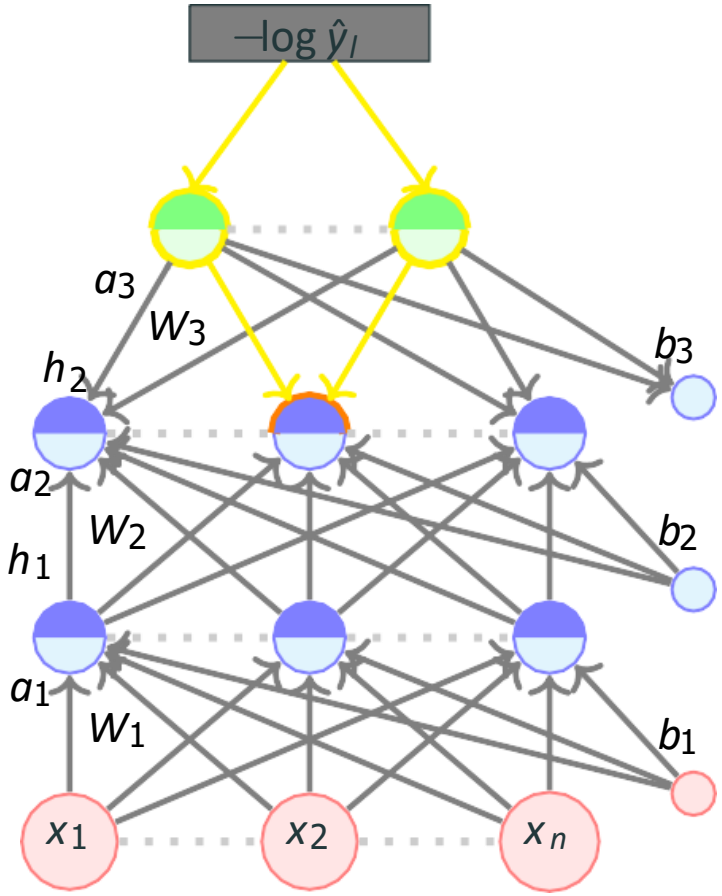
Our focus is on *Cross entropy loss* and *Softmax* output.

Chain rule along multiple paths: If a function $p(z)$ can be written as a function of intermediate results $q_i(z)$ then we have:



Chain rule along multiple paths: If a function $p(z)$ can be written as a function of intermediate results $q_i(z)$ then we have:

$$\frac{\partial p(z)}{\partial z} = \sum_m \frac{\partial p(z)}{\partial q_m(z)} \frac{\partial q_m(z)}{\partial z}$$

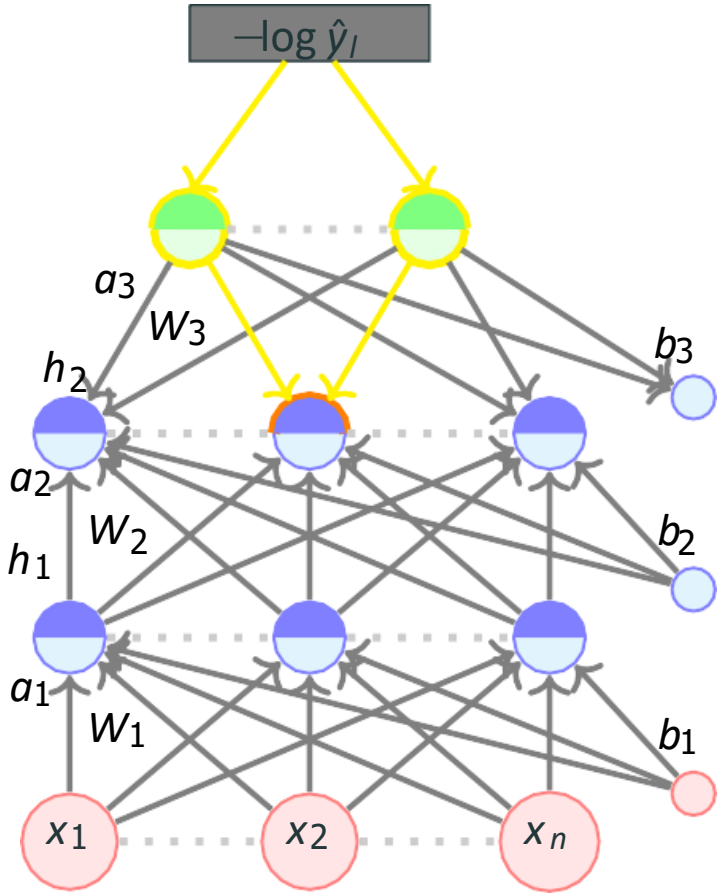


Chain rule along multiple paths: If a function $p(z)$ can be written as a function of intermediate results $q_i(z)$ then we have:

$$\frac{\partial p(z)}{\partial z} = \sum_m \frac{\partial p(z)}{\partial q_m(z)} \frac{\partial q_m(z)}{\partial z}$$

In our case:

$p(z)$ is the loss function $L(\vartheta)$



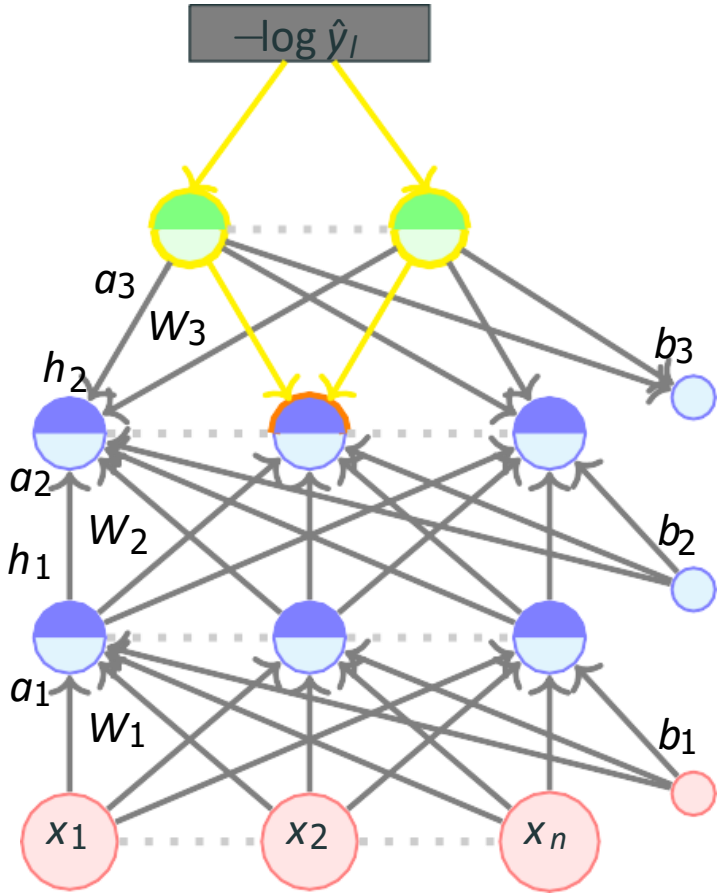
Chain rule along multiple paths: If a function $p(z)$ can be written as a function of intermediate results $q_i(z)$ then we have:

$$\frac{\partial p(z)}{\partial z} = \sum_m \frac{\partial p(z)}{\partial q_m(z)} \frac{\partial q_m(z)}{\partial z}$$

In our case:

$p(z)$ is the loss function $L(\vartheta)$

$z = h_{ij}$



Chain rule along multiple paths: If a function $p(z)$ can be written as a function of intermediate results $q_i(z)$ then we have:

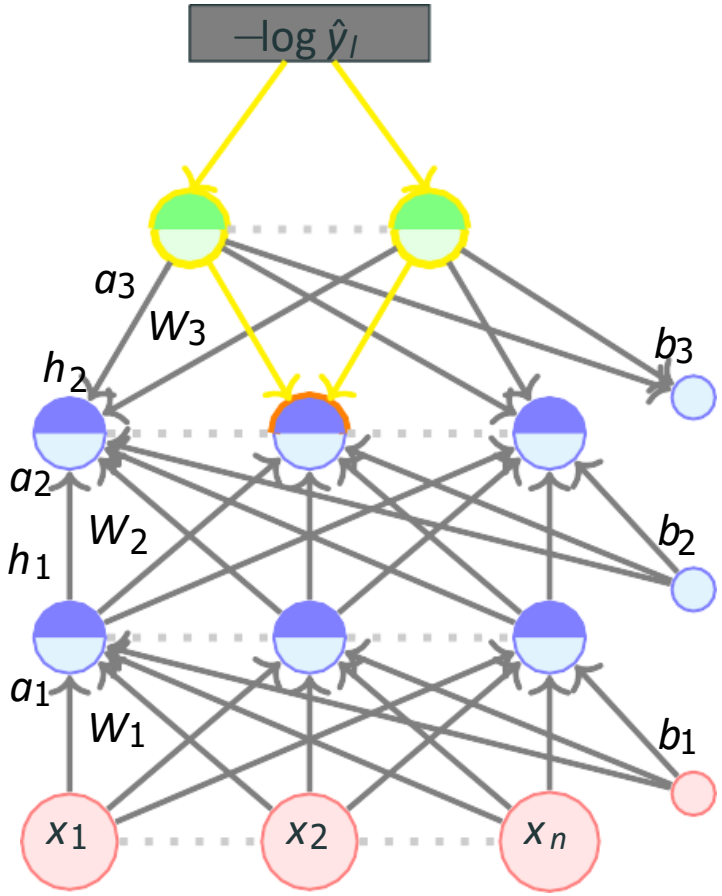
$$\frac{\partial p(z)}{\partial z} = \sum_m \frac{\partial p(z)}{\partial q_m(z)} \frac{\partial q_m(z)}{\partial z}$$

In our case:

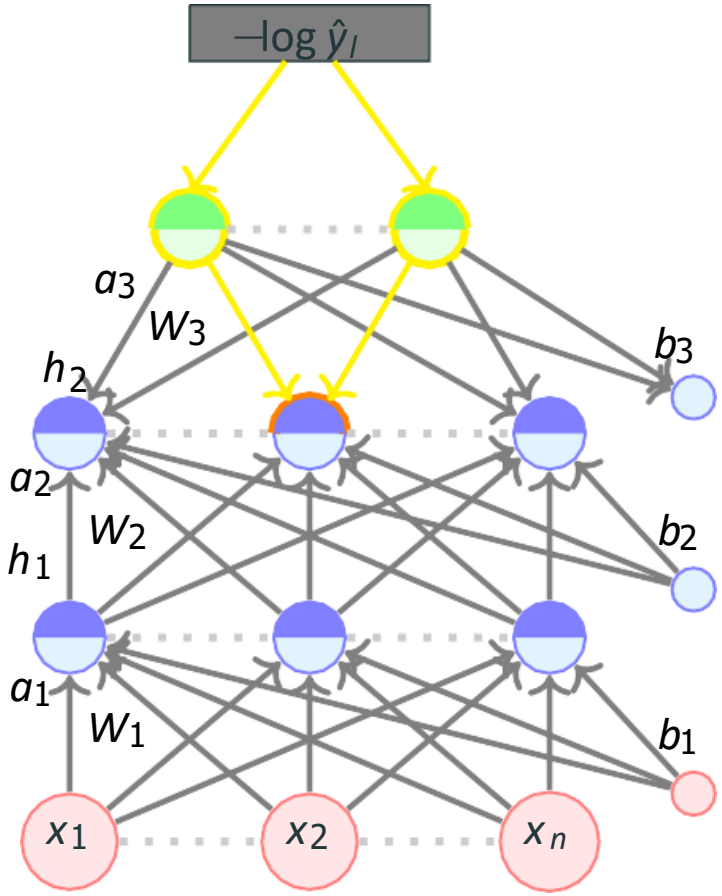
$p(z)$ is the loss function $L(\vartheta)$

$z = h_{ij}$

$q_m(z) = a_{Lm}$

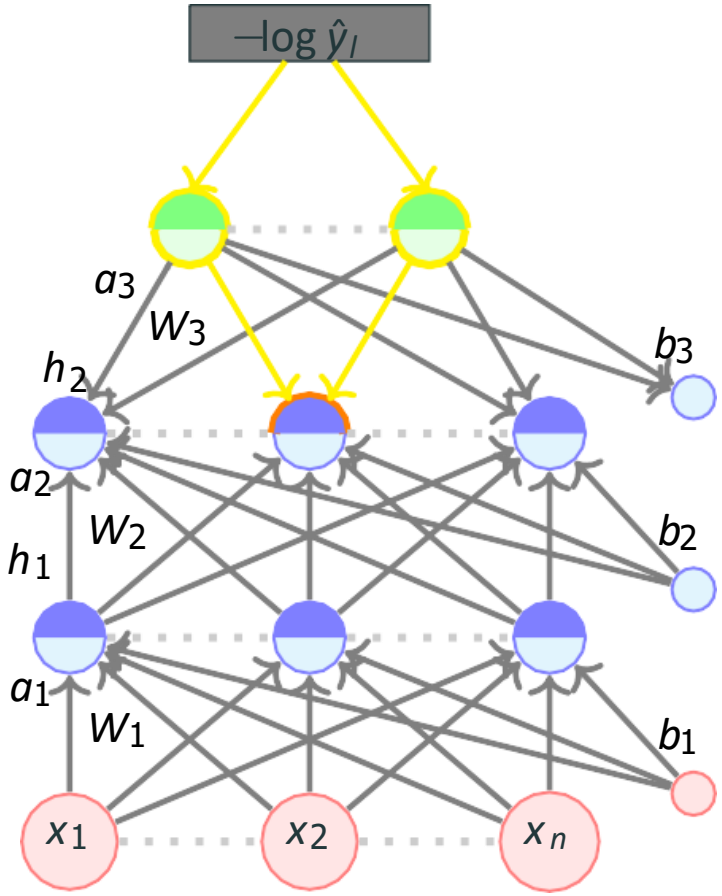


$$\frac{\partial L(\vartheta)}{\partial h_{ij}}$$



$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

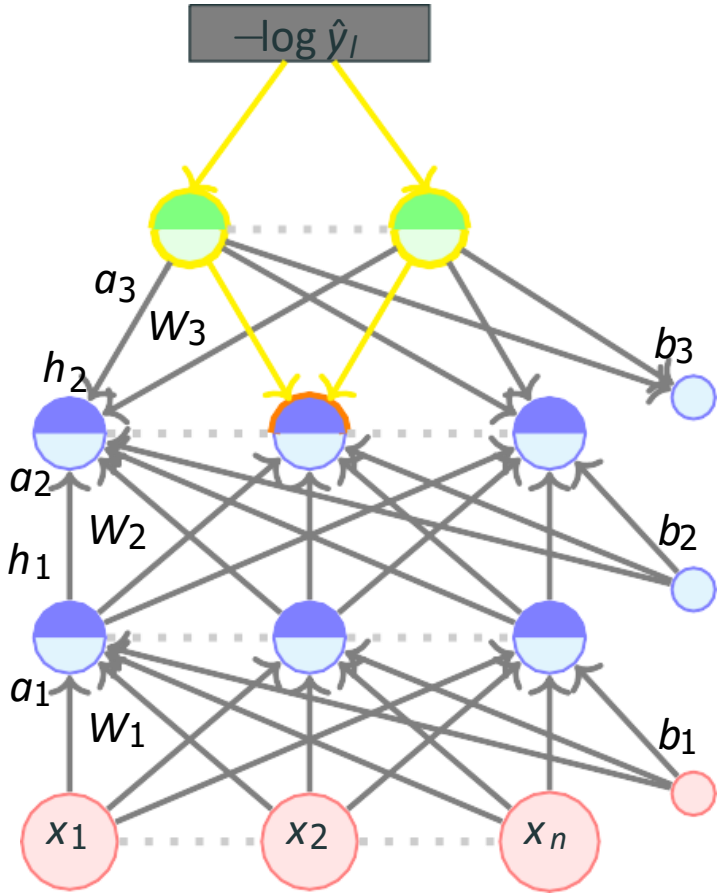
$$\frac{\partial L(\vartheta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial L(\vartheta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$



$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial L(\vartheta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial L(\vartheta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial L(\vartheta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

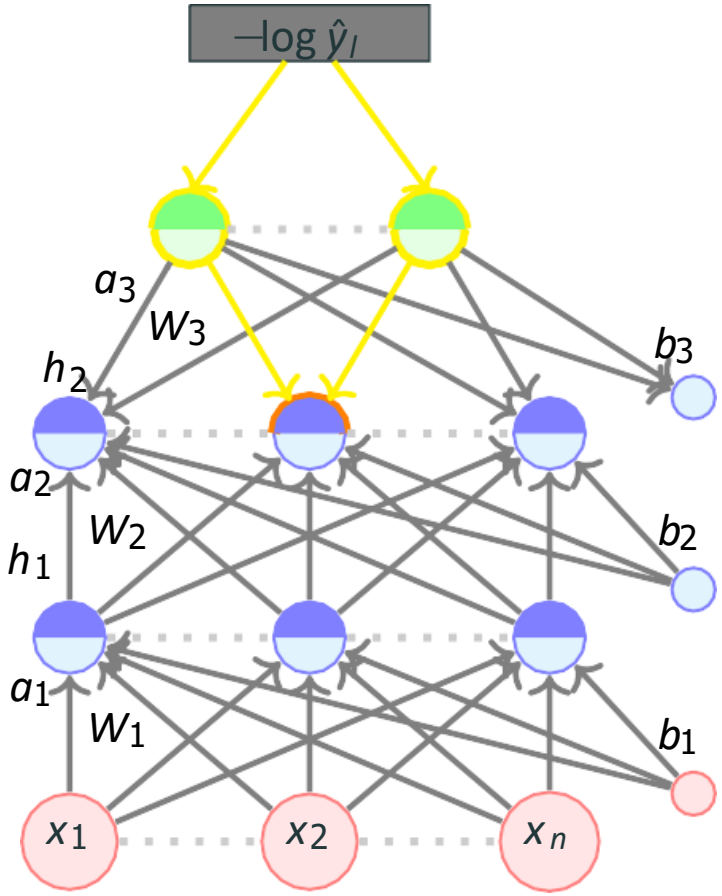


$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial L(\vartheta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial L(\vartheta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial L(\vartheta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,



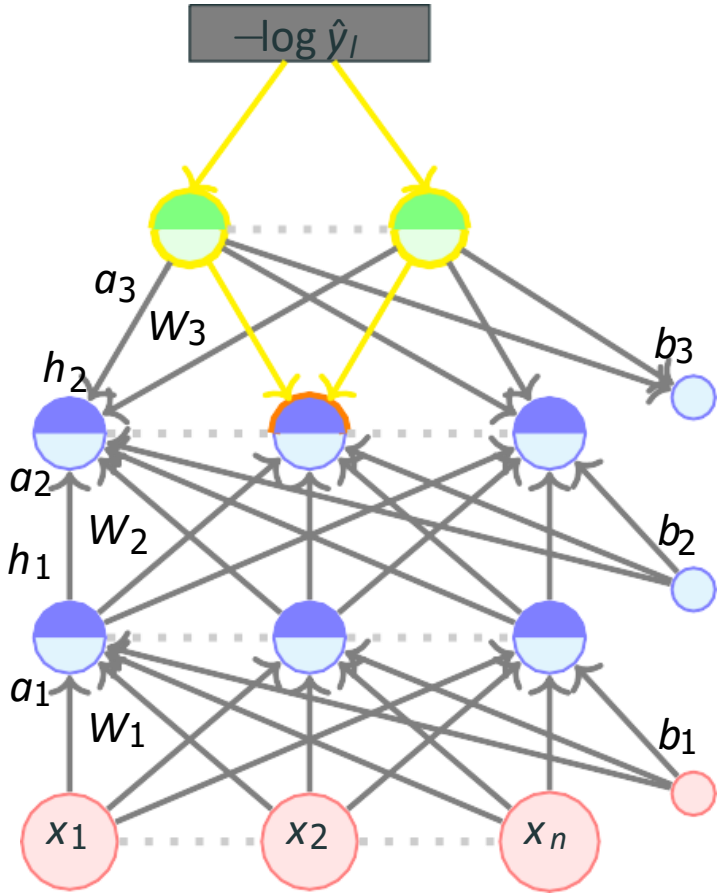
$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial L(\vartheta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial L(\vartheta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial L(\vartheta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \\ \\ \end{bmatrix}; W_{i+1,\cdot,j} = \begin{bmatrix} \\ \\ \end{bmatrix}$$



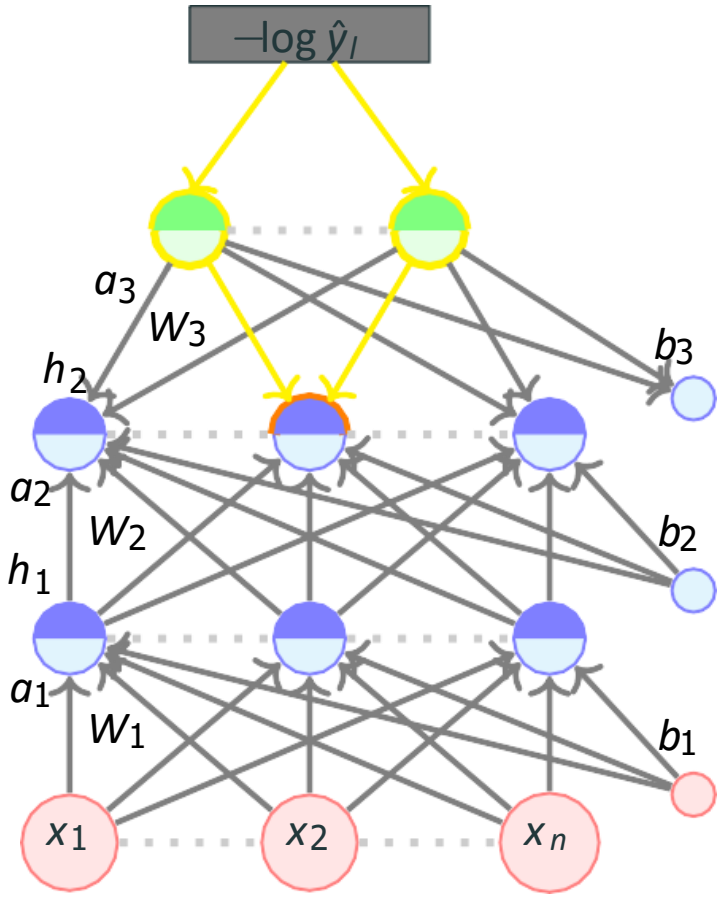
$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \\ \\ \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$



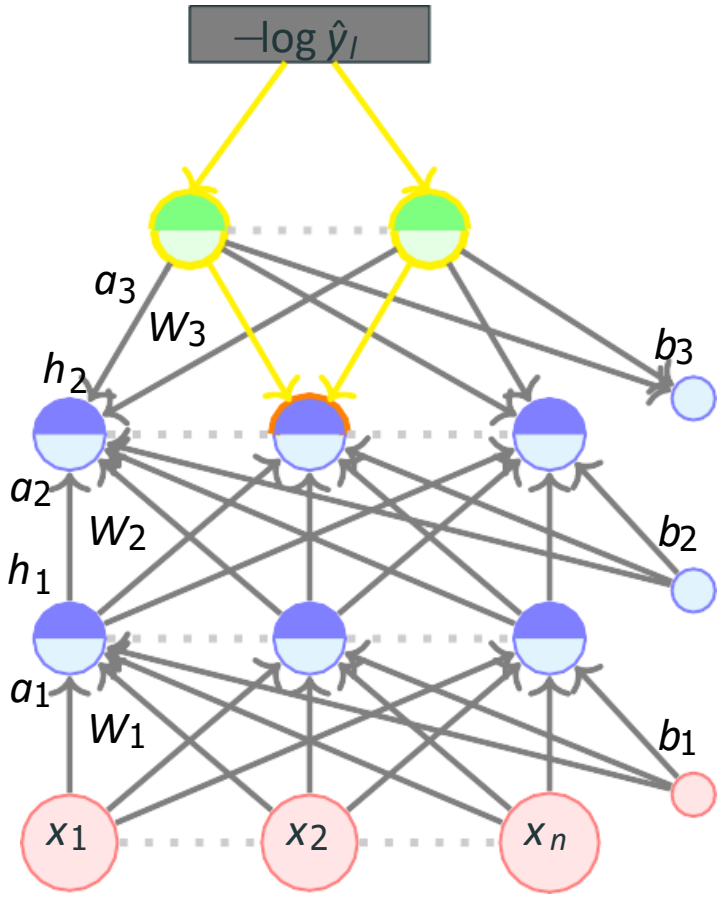
$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \\ \\ \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \\ \\ \end{bmatrix}$$



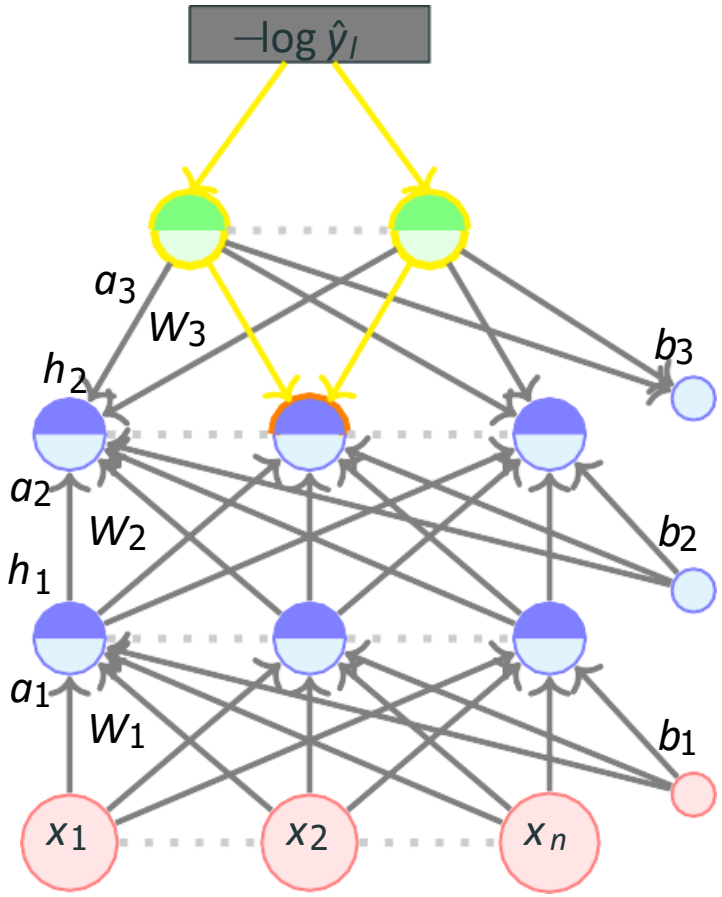
$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \end{bmatrix}; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \end{bmatrix}$$



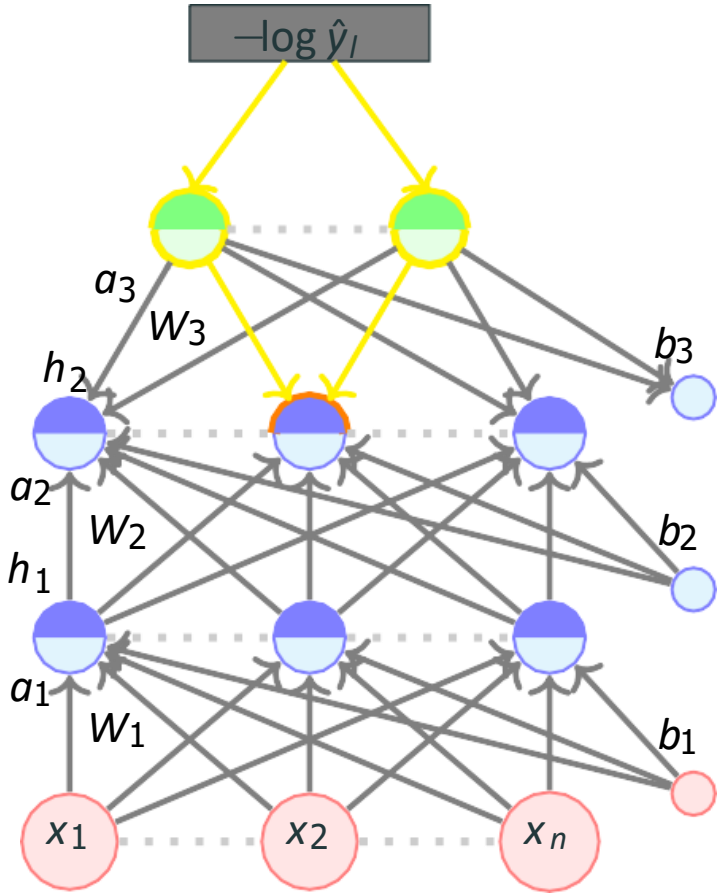
$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ \end{bmatrix}$$



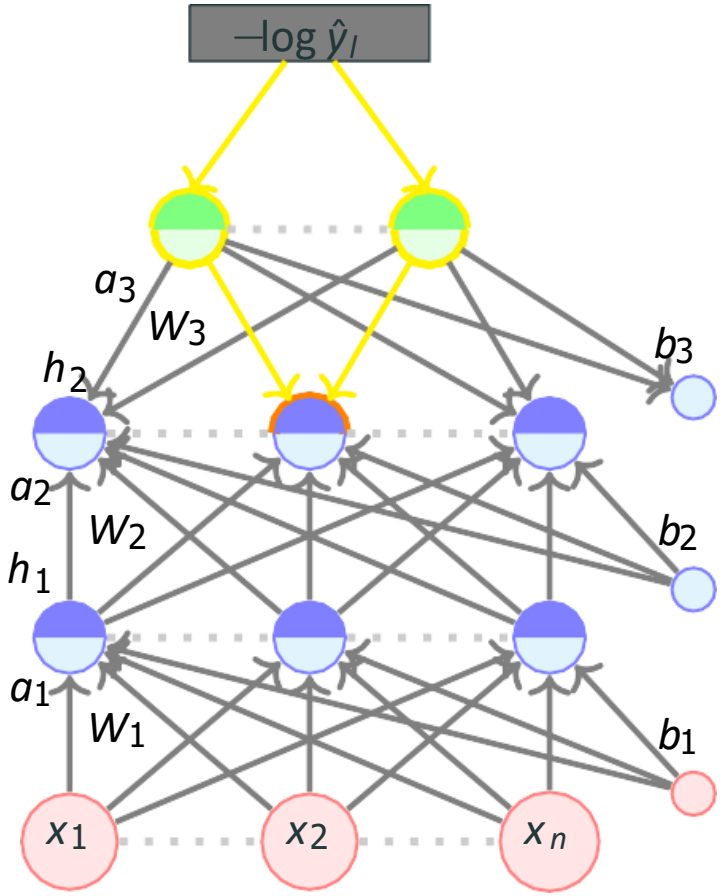
$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$



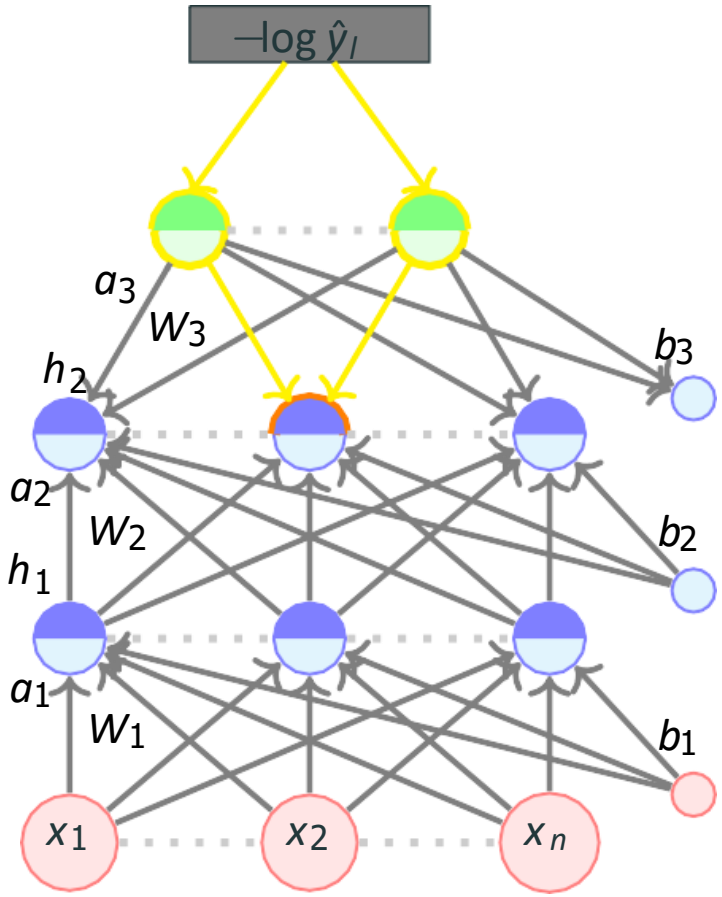
$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$



$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$

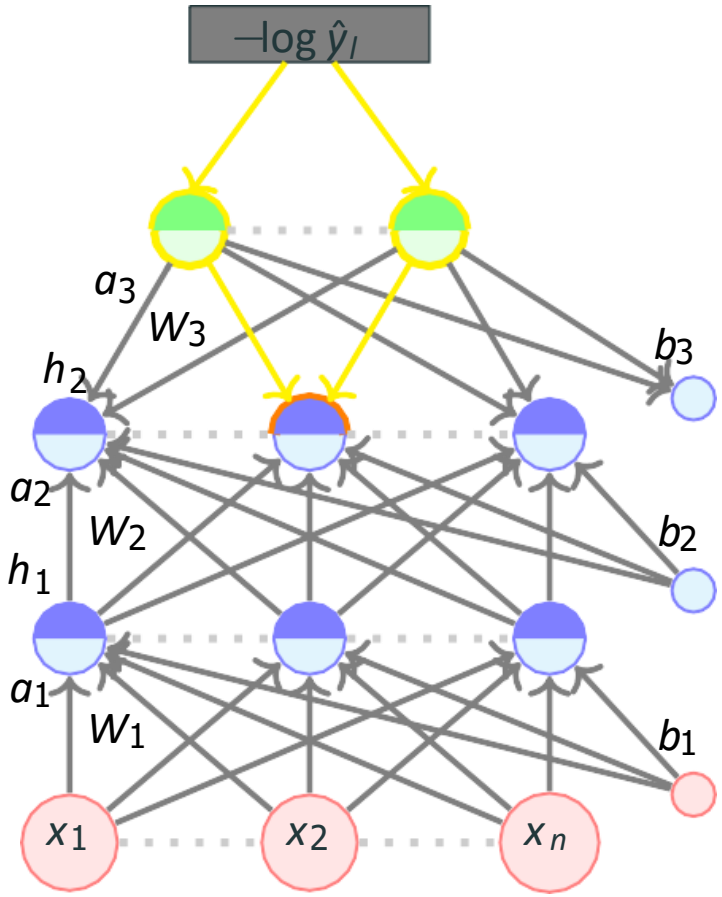
$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix}; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$

$W_{i+1,\cdot,j}$ is the j-th column of W_{i+1};



$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

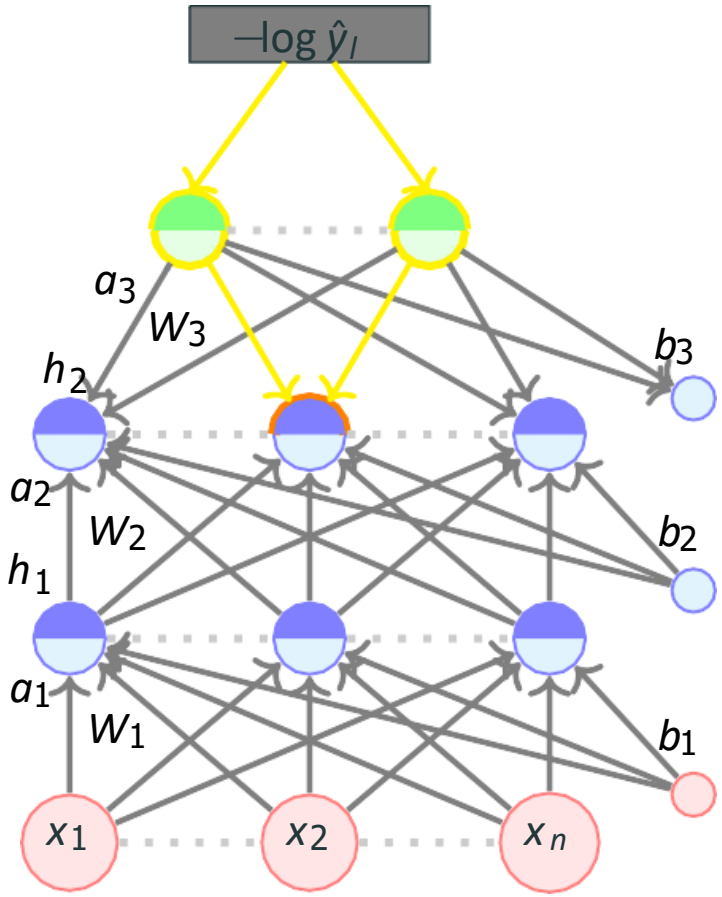
$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$

$W_{i+1,\cdot,j}$ is the j-th column of W_{i+1}; see that,



$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

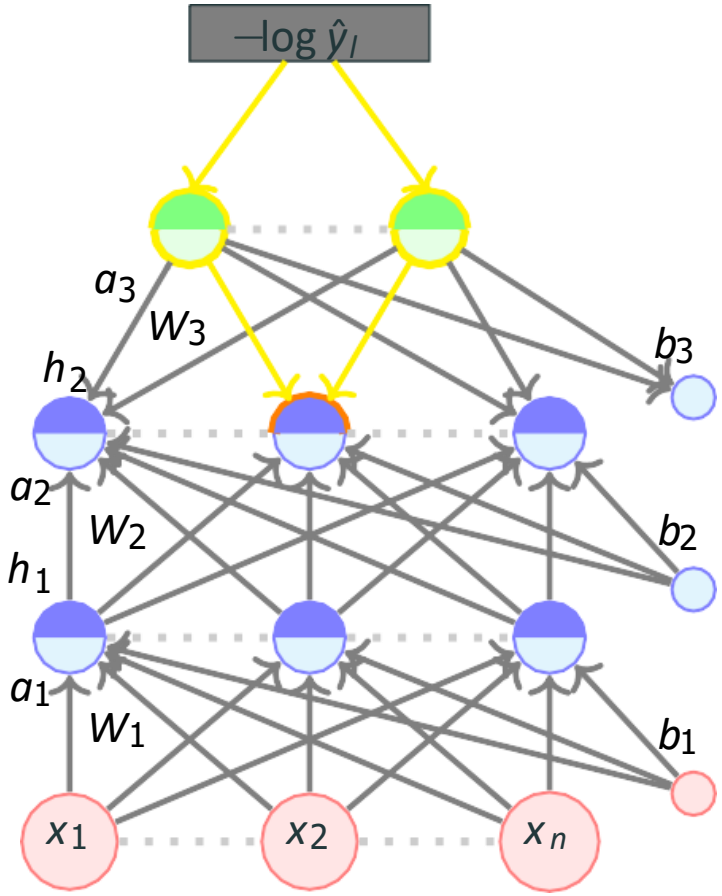
$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$

$W_{i+1,\cdot,j}$ is the j-th column of W_{i+1}; see that,

$$(W_{i+1,\cdot,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) =$$



$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$

$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

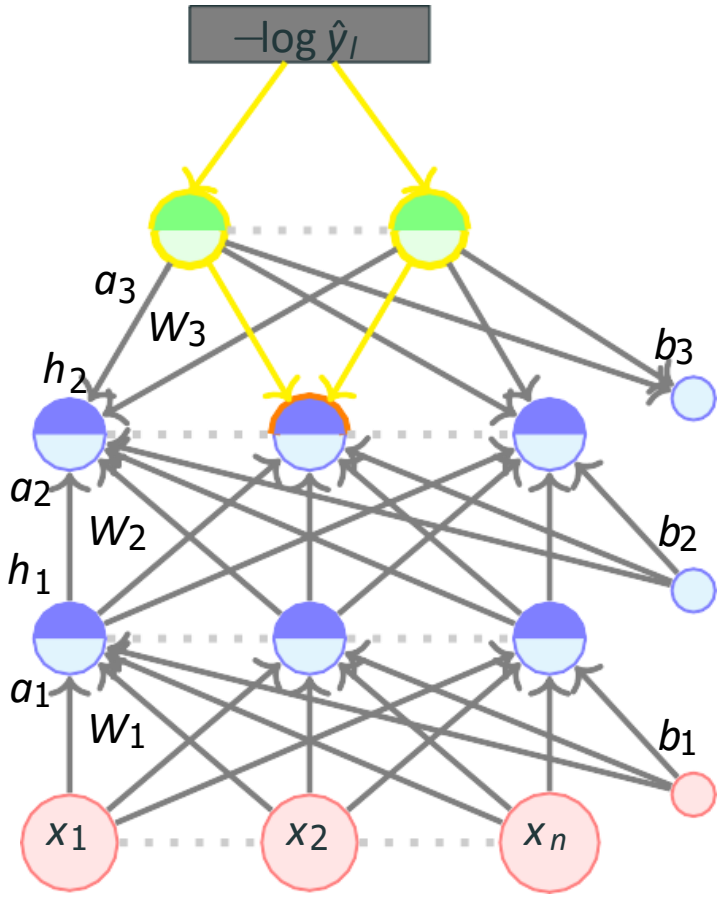
$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

Now consider these two vectors,

$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix} ; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$

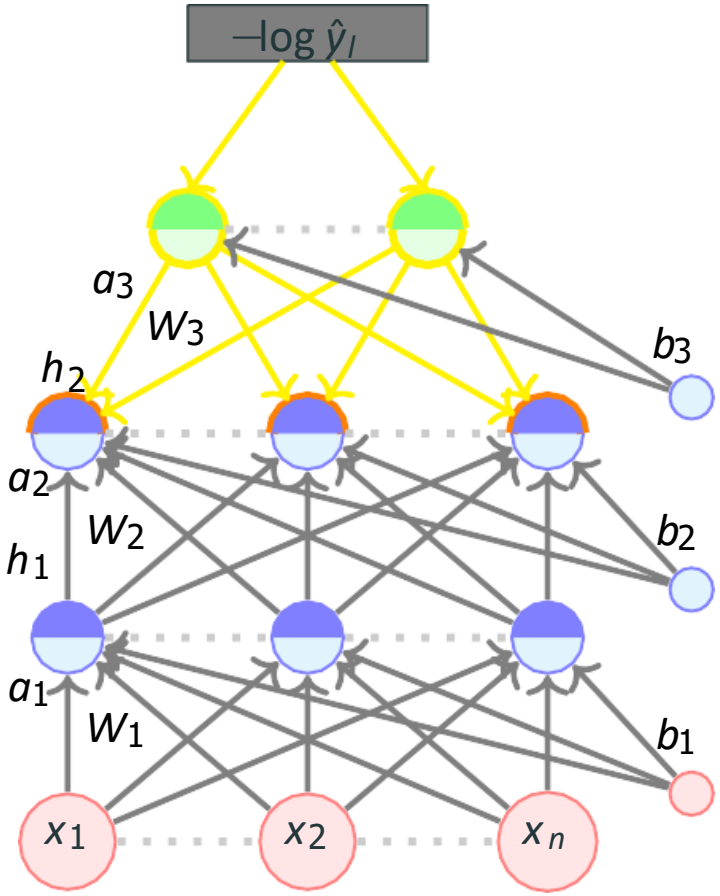
$W_{i+1,\cdot,j}$ is the j-th column of W_{i+1}; see that,

$$(W_{i+1,\cdot,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$



$$a_{i+1} = W_{i+1} h_{ij} + b_{i+1}$$

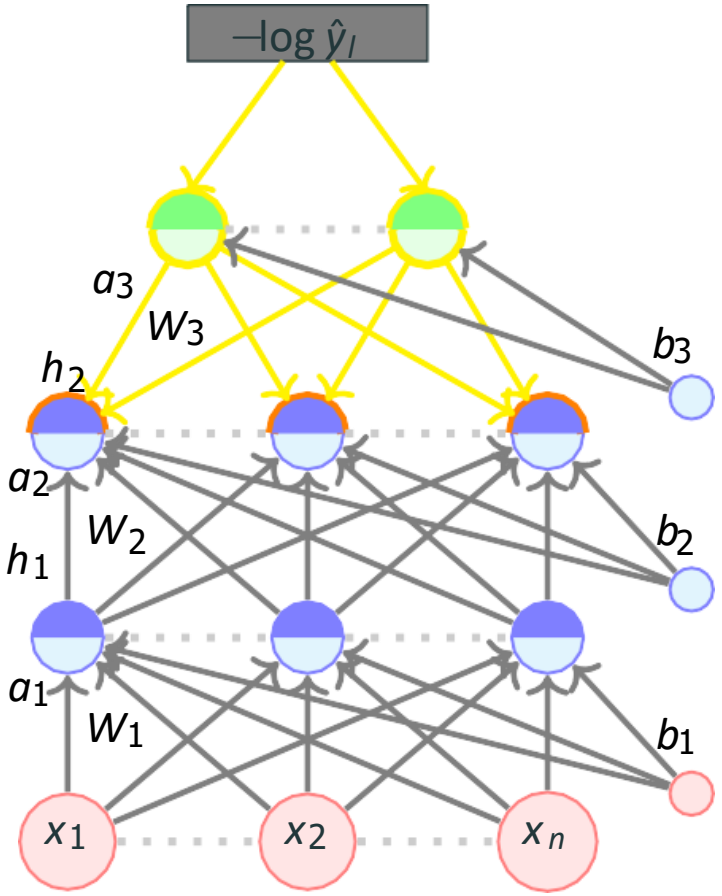
$$We\ have, \frac{\partial L\ (\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L\ (\vartheta)$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

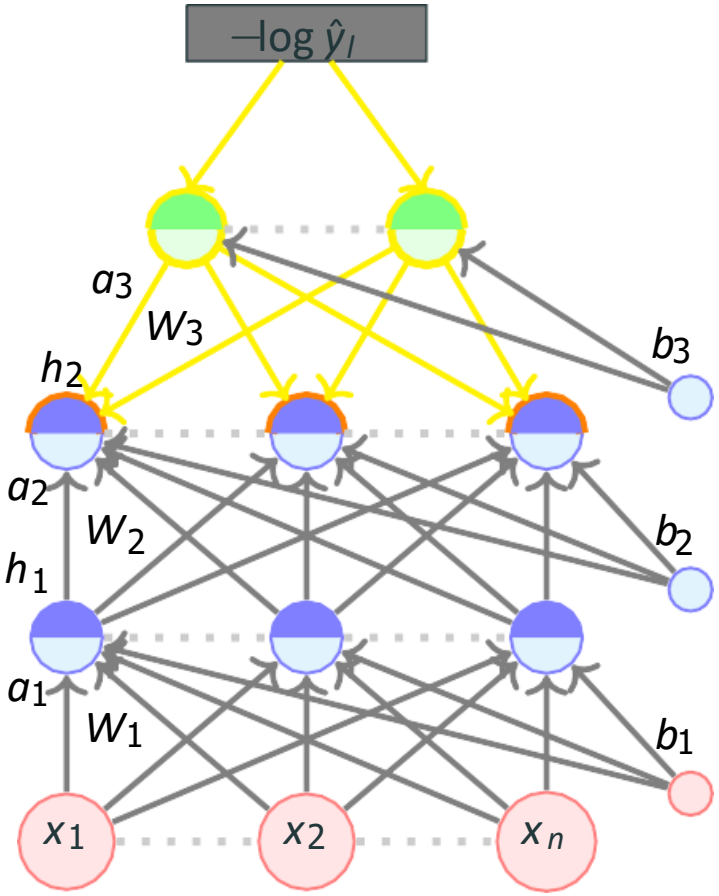
$$\nabla_{h_i} L(\vartheta)$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

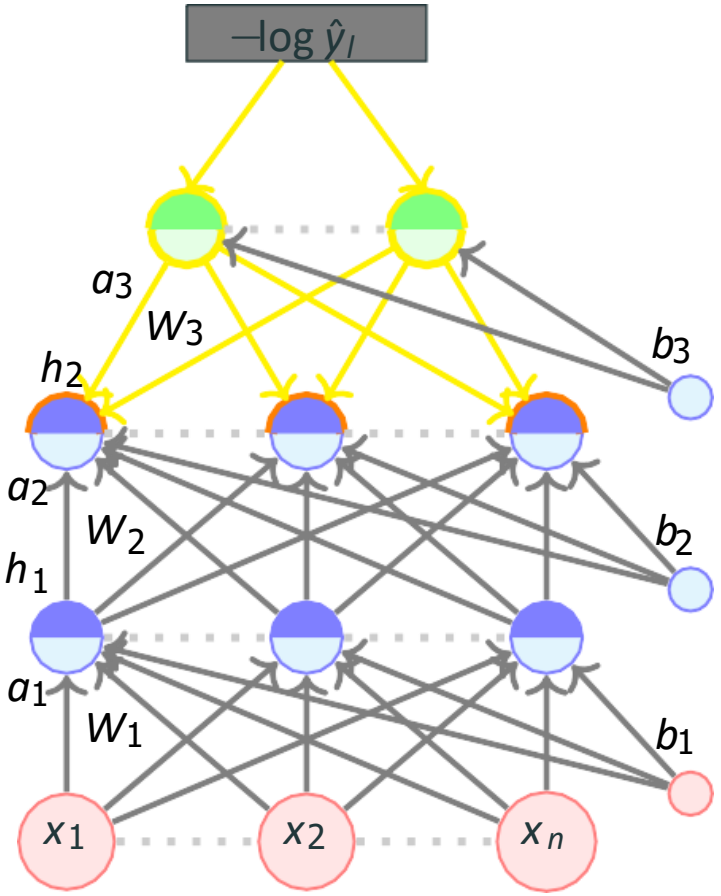
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \\ \end{bmatrix} = \begin{bmatrix} \\ \end{bmatrix}$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

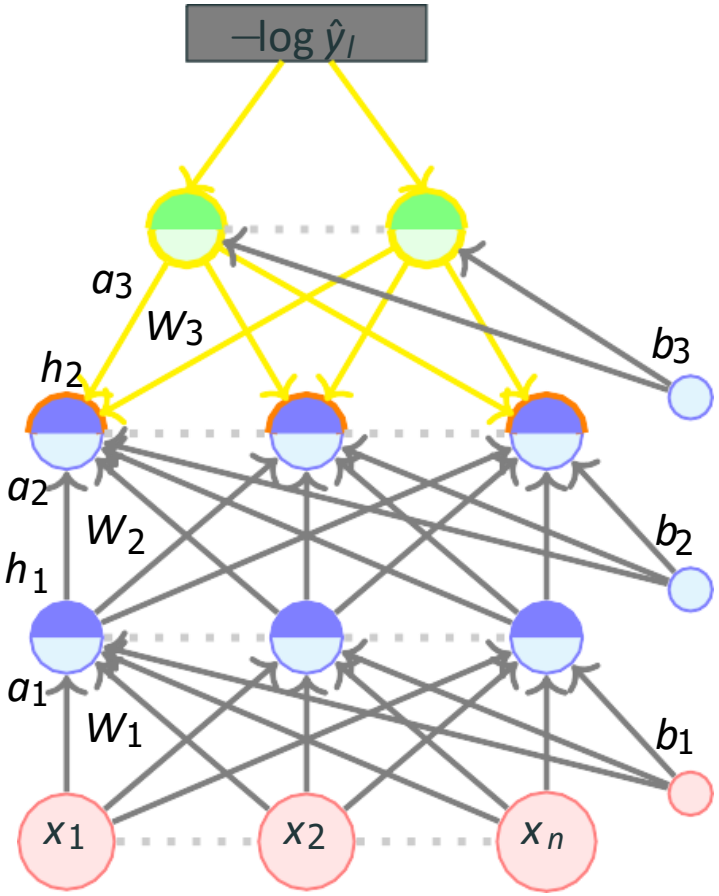
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \end{bmatrix}$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

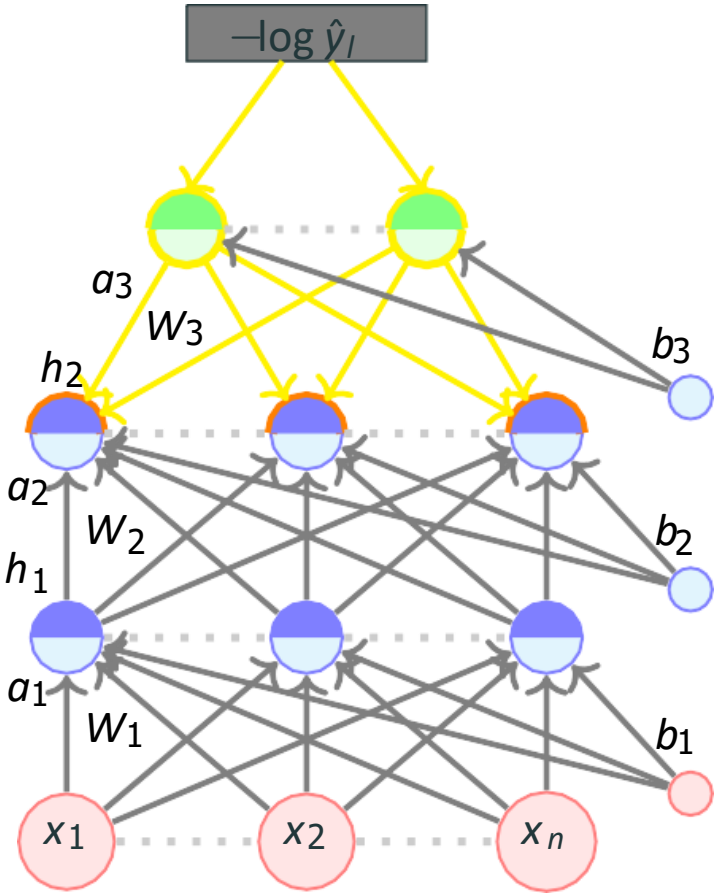
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \\ \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \\ \end{bmatrix}$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

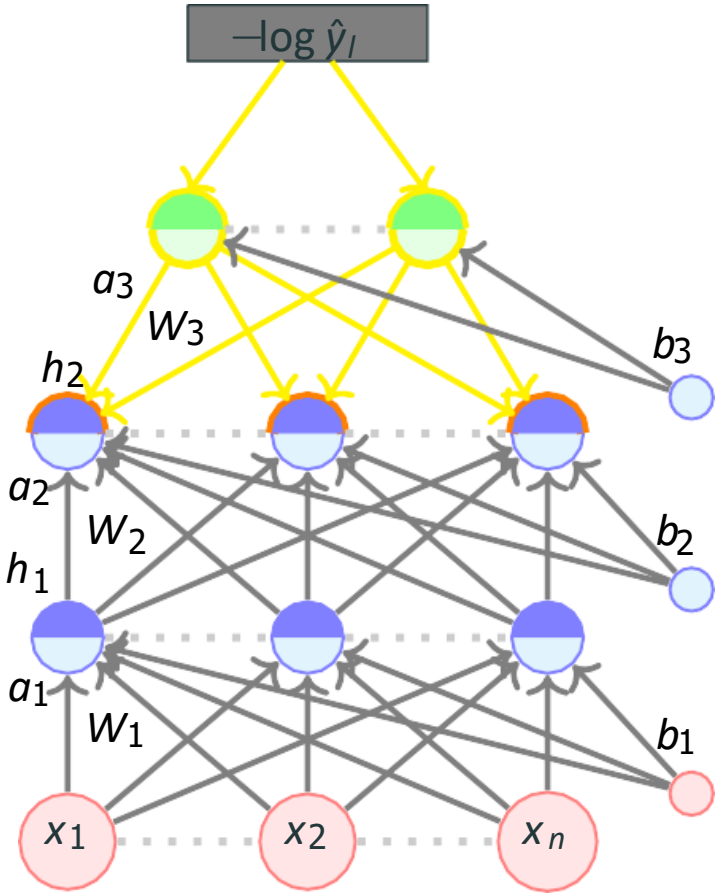
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \\ \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \\ \\ \end{bmatrix}$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

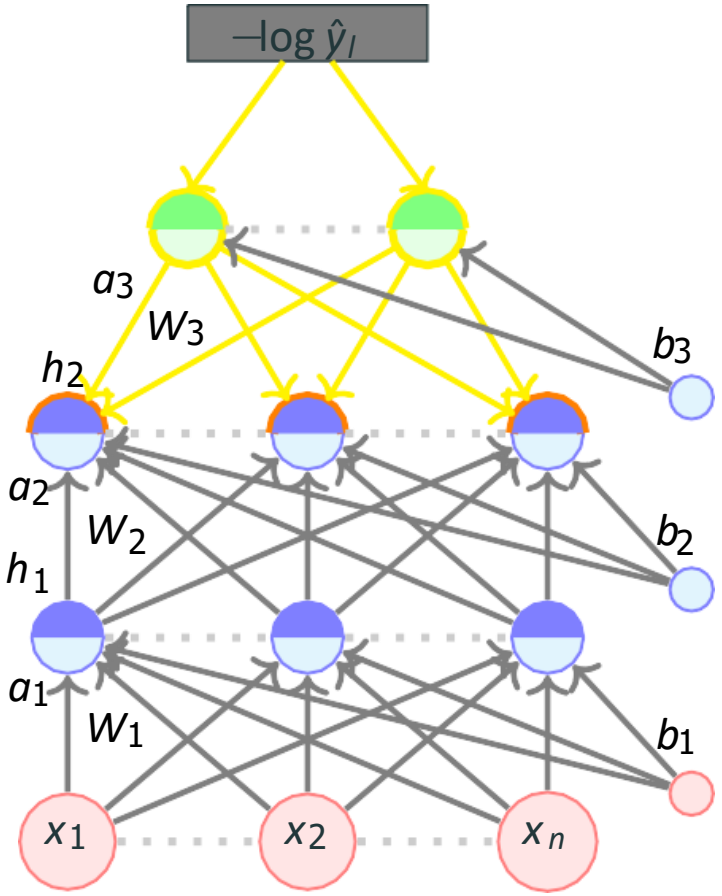
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \\ \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \\ \end{bmatrix}$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

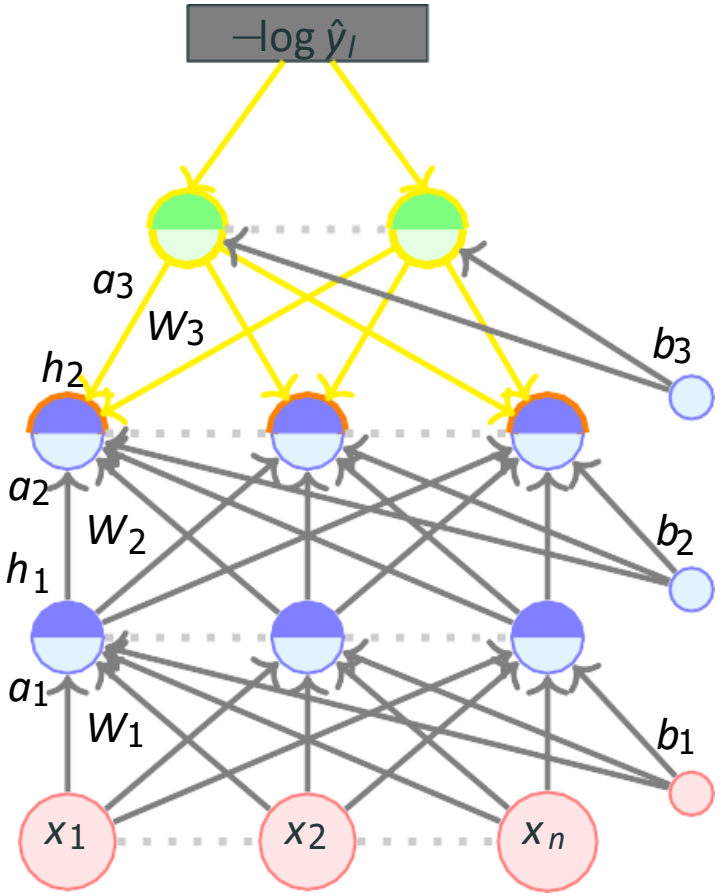
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \vdots \end{bmatrix}$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

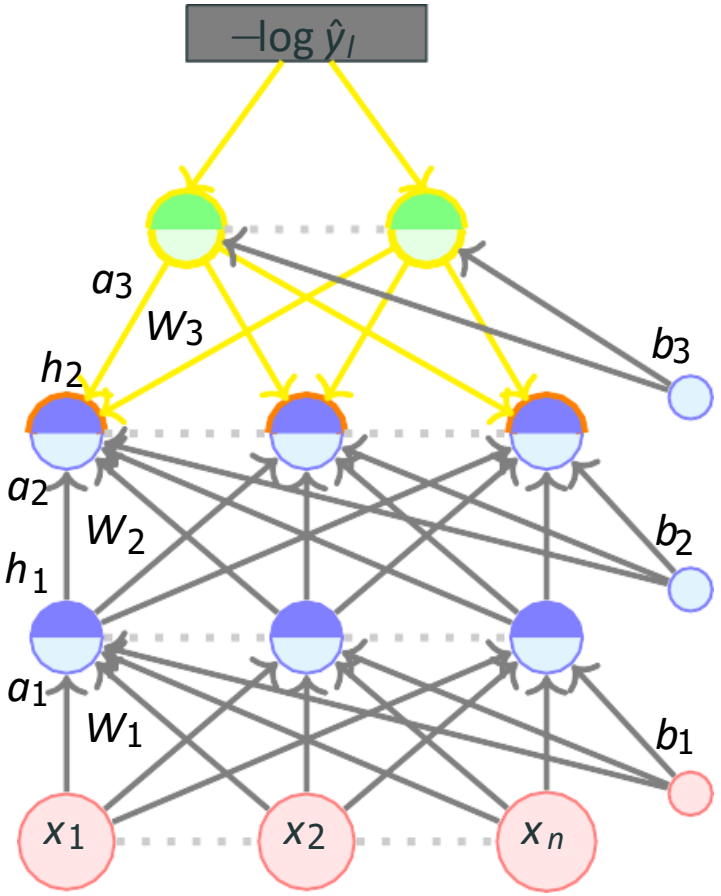
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \vdots \end{bmatrix}$$



$$We\ have, \frac{\partial L(\vartheta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} L(\vartheta)$$

We can now write the gradient w.r.t. h_i

$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \vdots \\ (W_{i+1,\cdot,n})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \end{bmatrix}$$

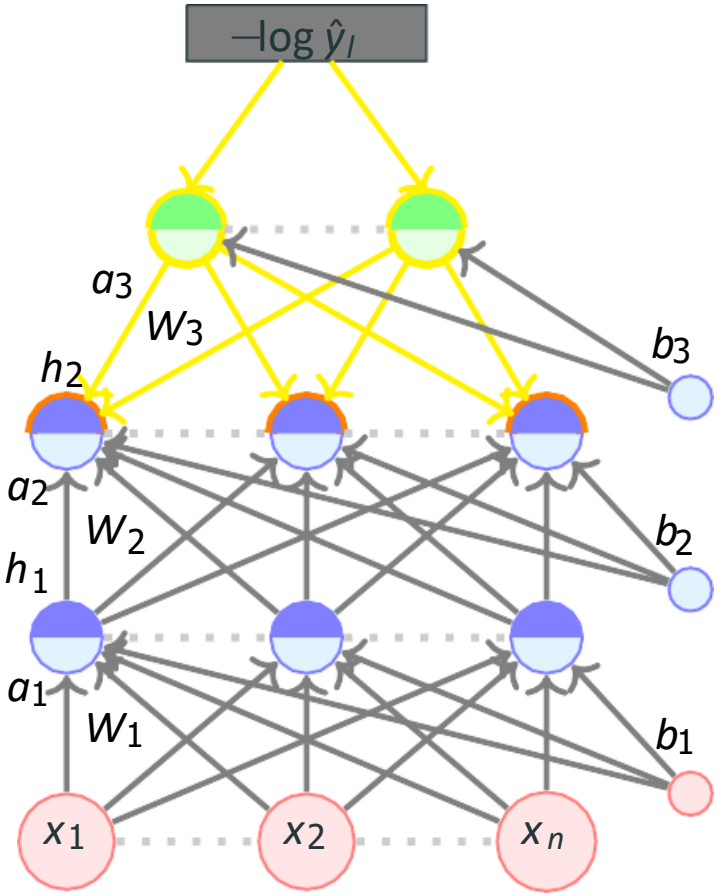


We have, $\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta)$

We can now write the gradient w.r.t. h_i

$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \vdots \\ (W_{i+1,\cdot,n})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \end{bmatrix}$$

$$= (W_{i+1})^T (\nabla_{a_{i+1}} \mathscr{L}(\theta))$$



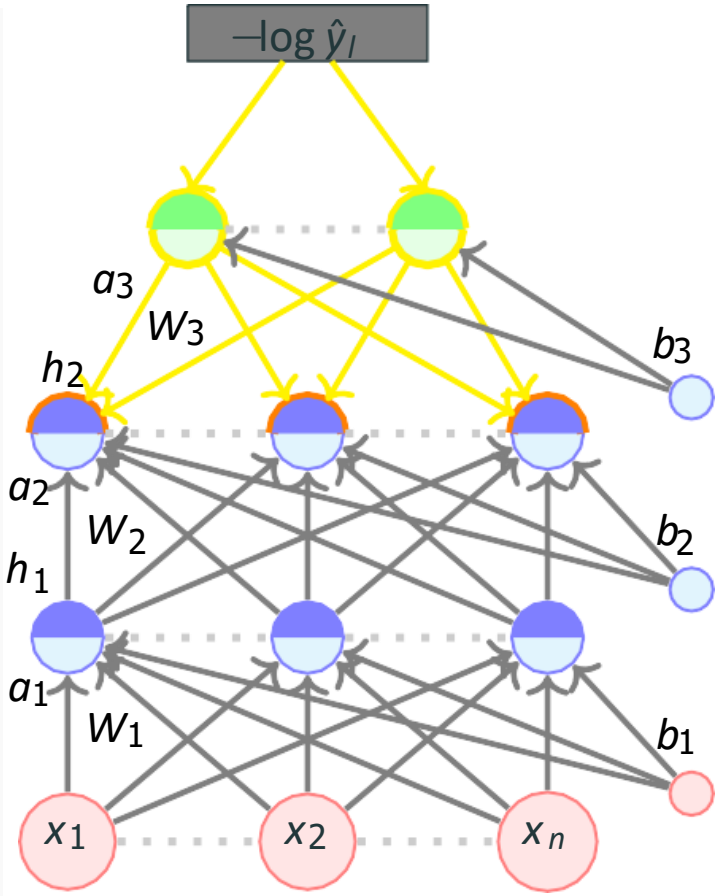
$$\text{We have, } \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta)$$

We can now write the gradient w.r.t. h_i

$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \vdots \\ (W_{i+1,\cdot,n})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \end{bmatrix}$$

$$= (W_{i+1})^T (\nabla_{a_{i+1}} \mathscr{L}(\theta))$$

- We are almost done except that we do not know how to calculate $\nabla_{a_{i+1}} \mathscr{L}(\theta)$ for $i < L - 1$



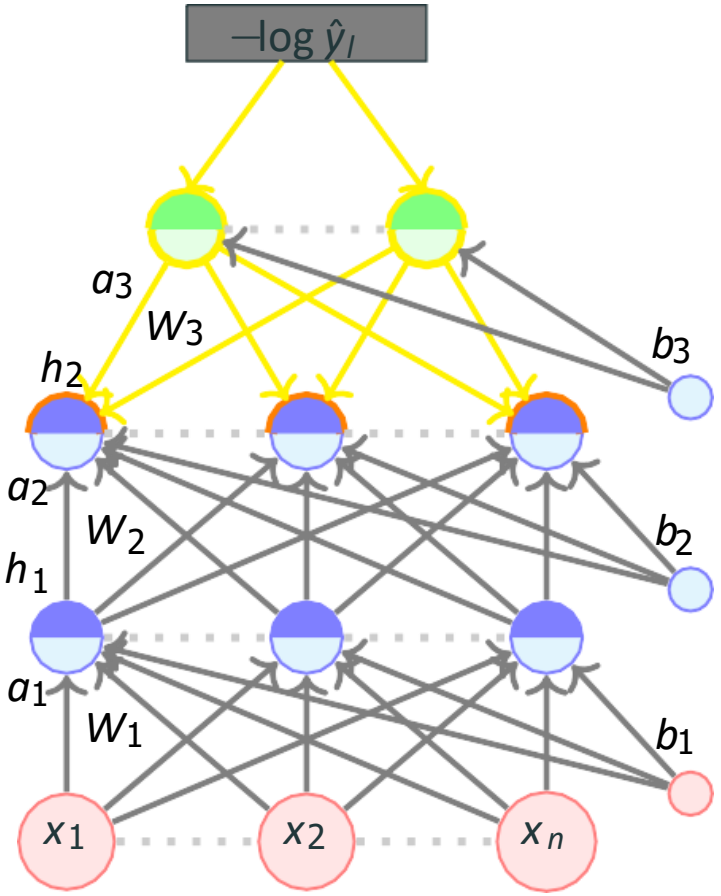
We have, $\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta)$

We can now write the gradient w.r.t. h_i

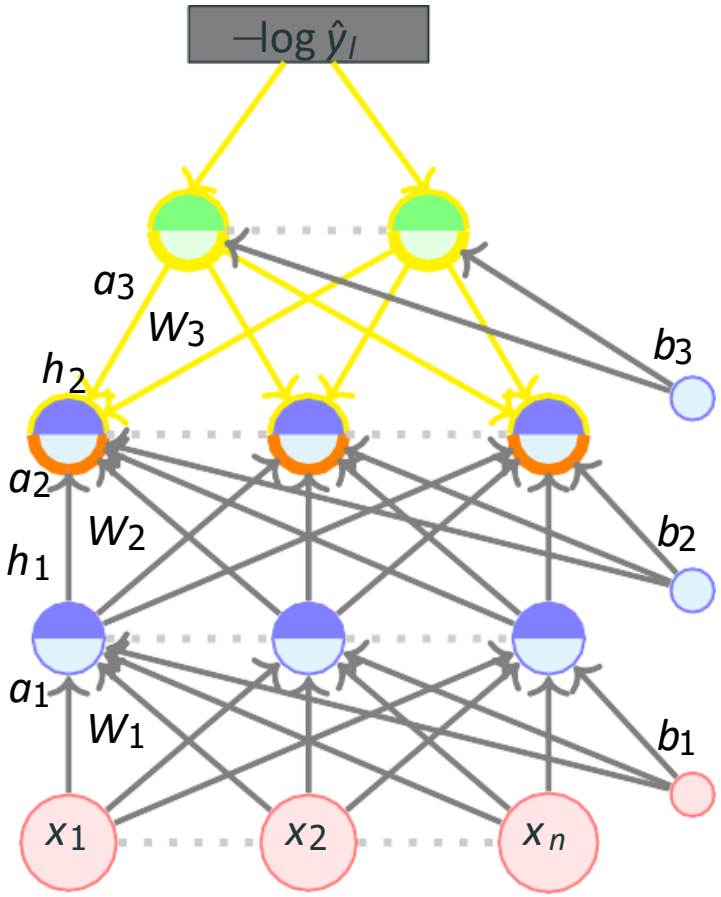
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \vdots \\ (W_{i+1,\cdot,n})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \end{bmatrix}$$

$$= (W_{i+1})^T (\nabla_{a_{i+1}} \mathscr{L}(\theta))$$

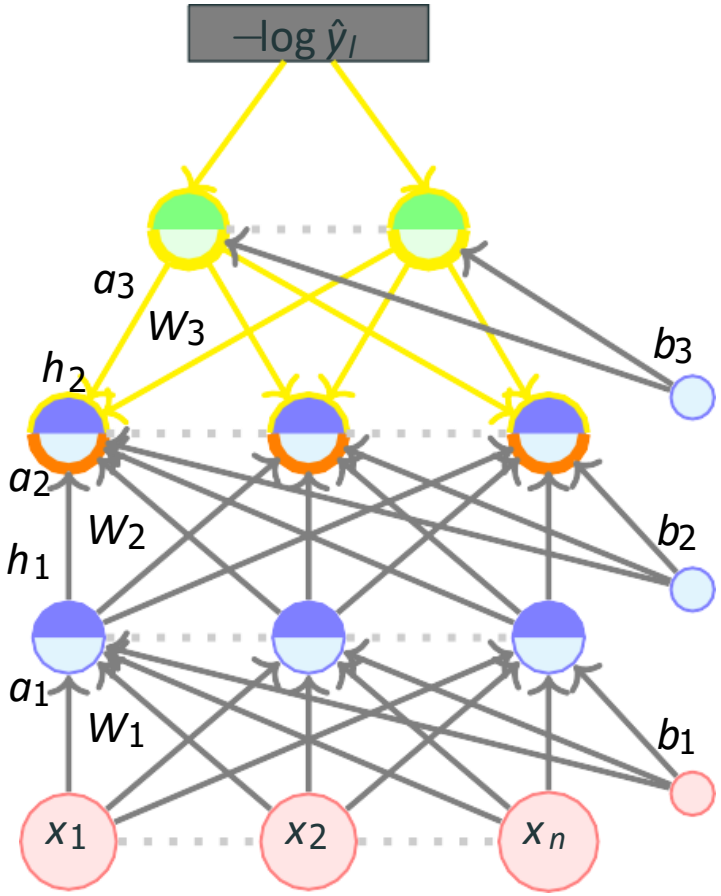
- We are almost done except that we do not know how to calculate $\nabla_{a_{i+1}} \mathscr{L}(\theta)$ for $i < L - 1$
- We will see how to compute that



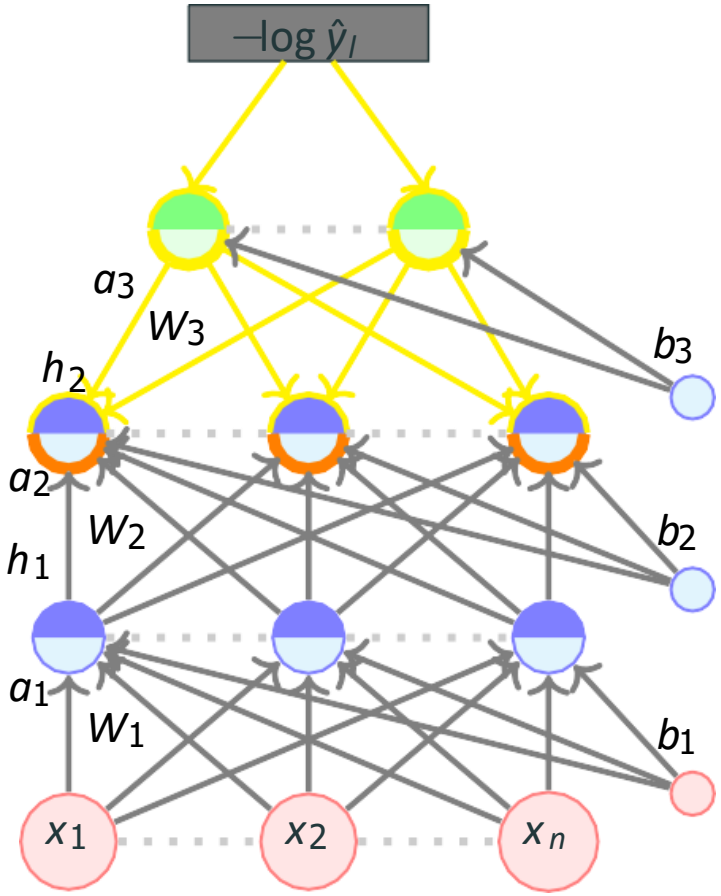
$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta)$



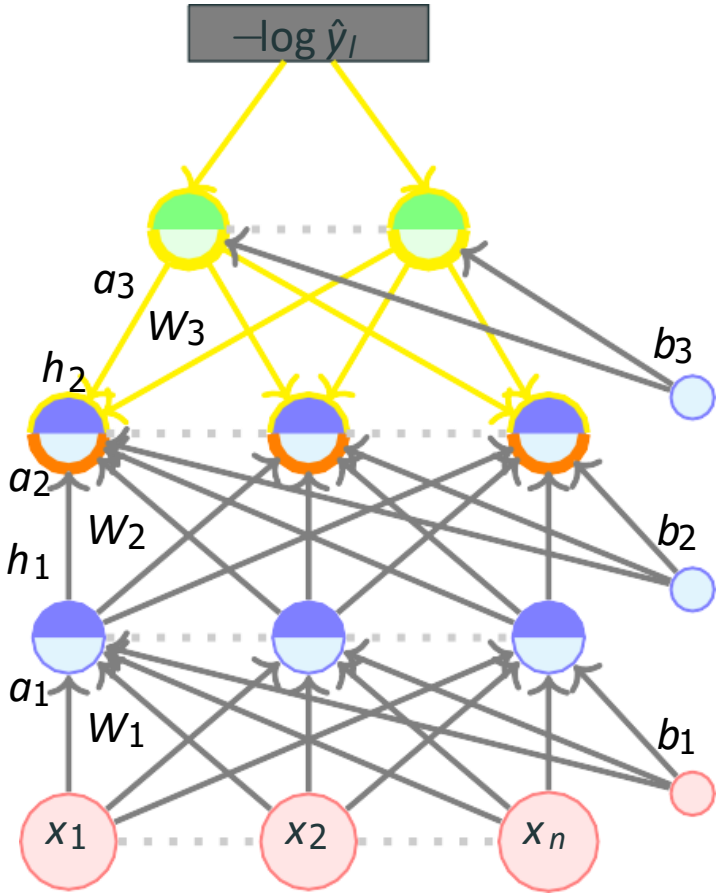
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \\ \\ \end{bmatrix}$$



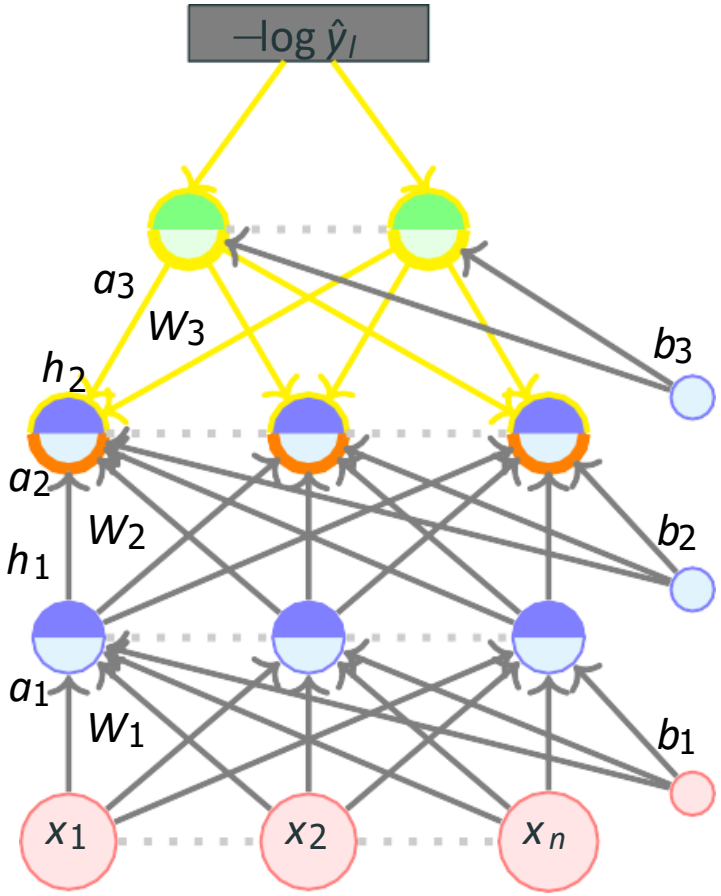
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \\ \\ \end{bmatrix}$$



$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \end{bmatrix}$$

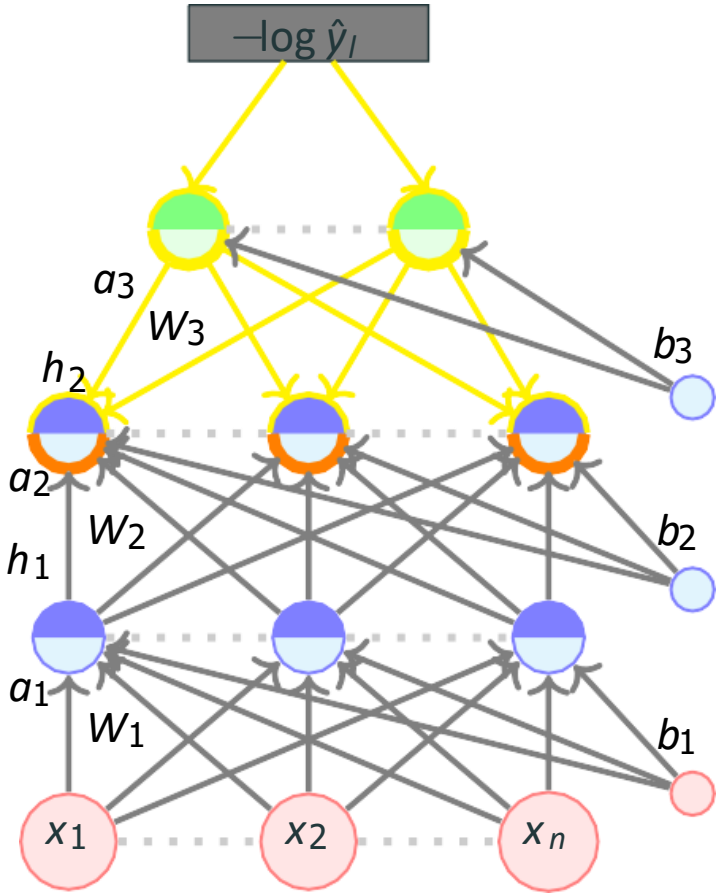


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$



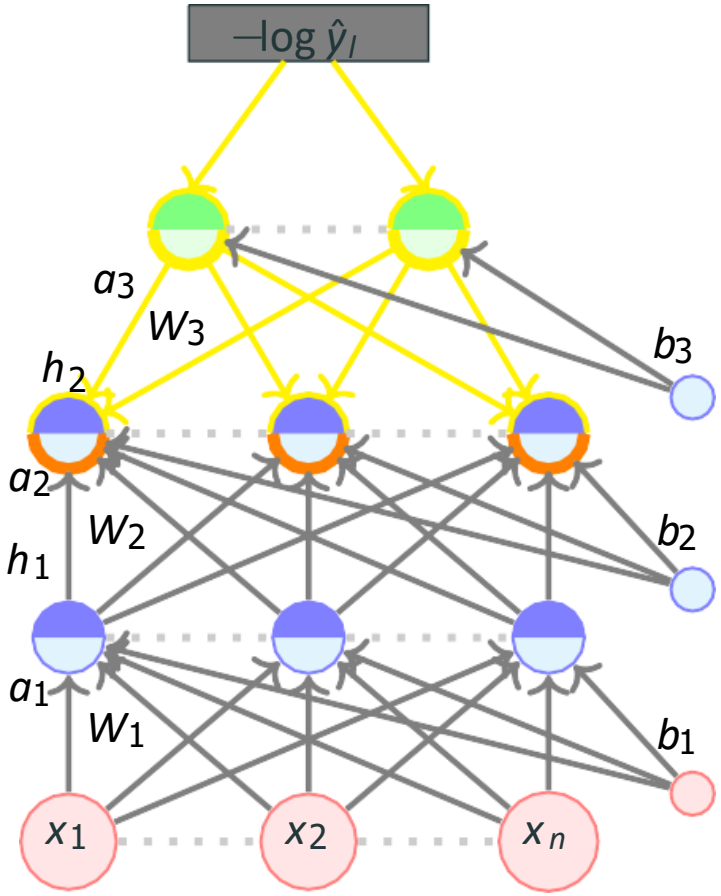
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}}$$



$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

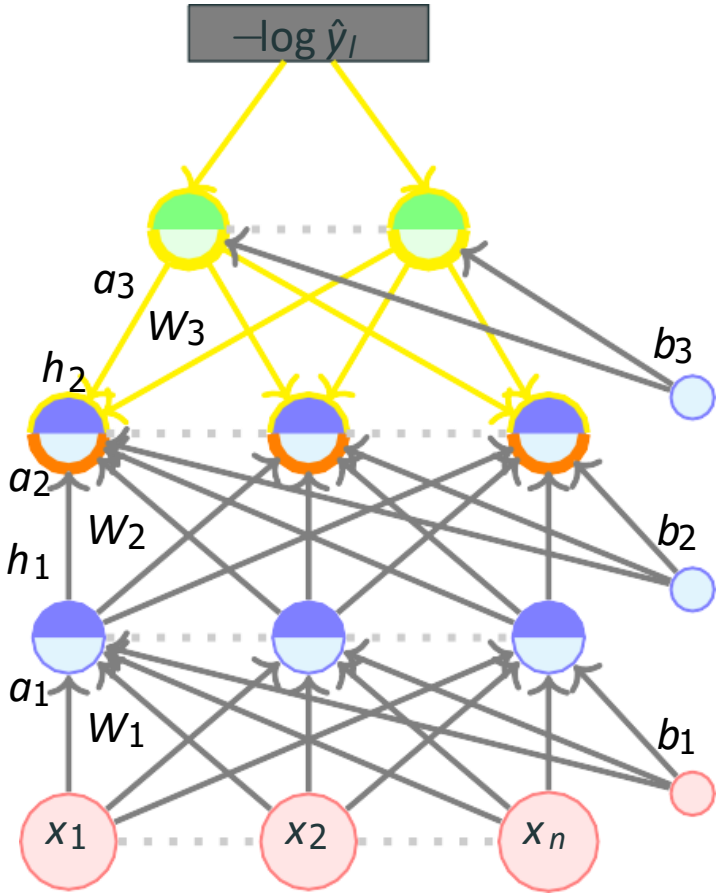
$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$



$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

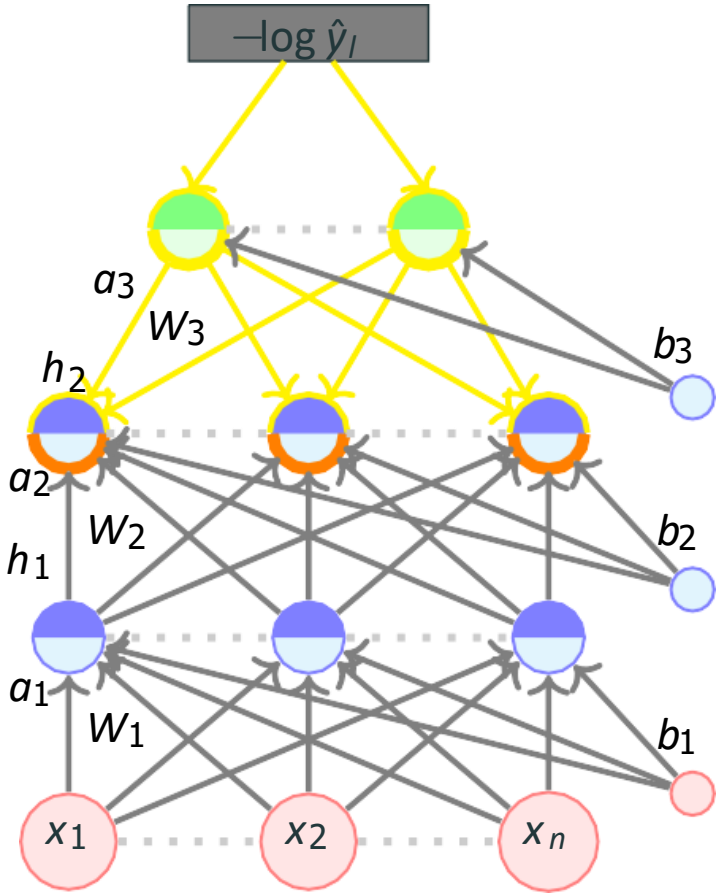


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta)$$

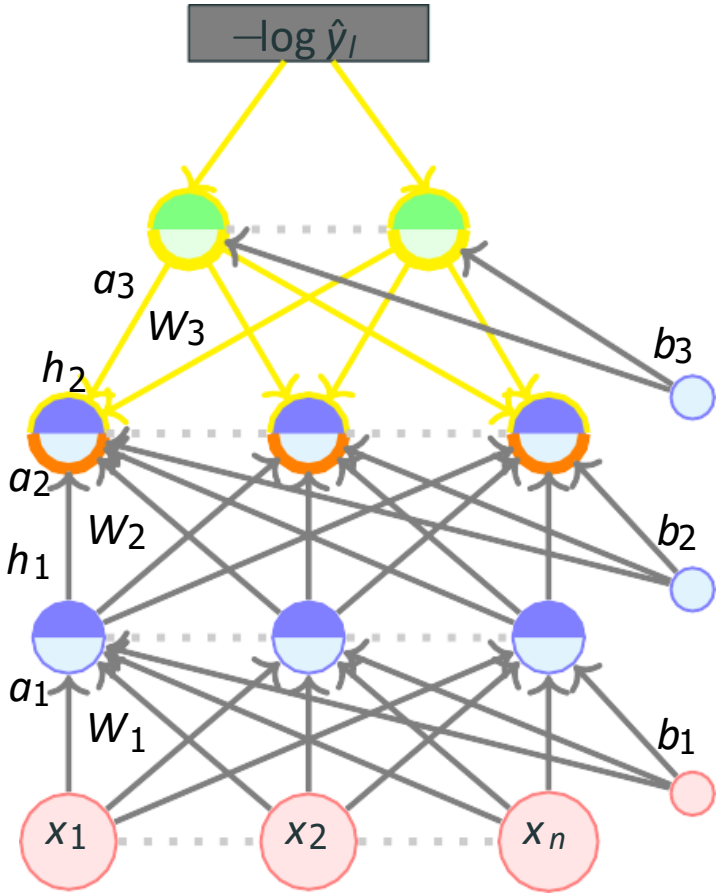


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

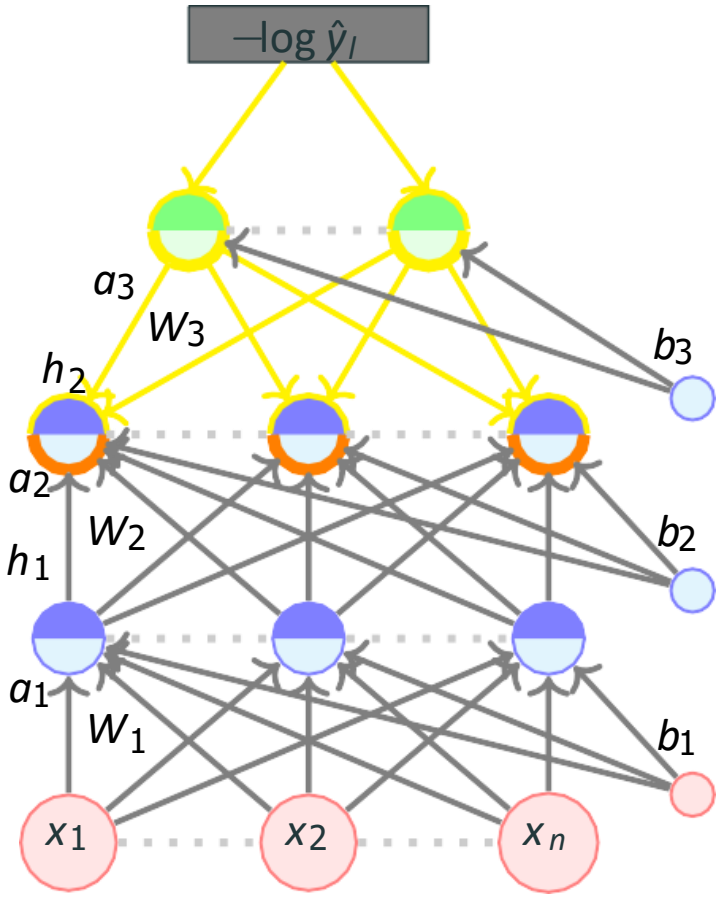


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} g'(a_{i1}) \\ \\ \end{bmatrix}$$

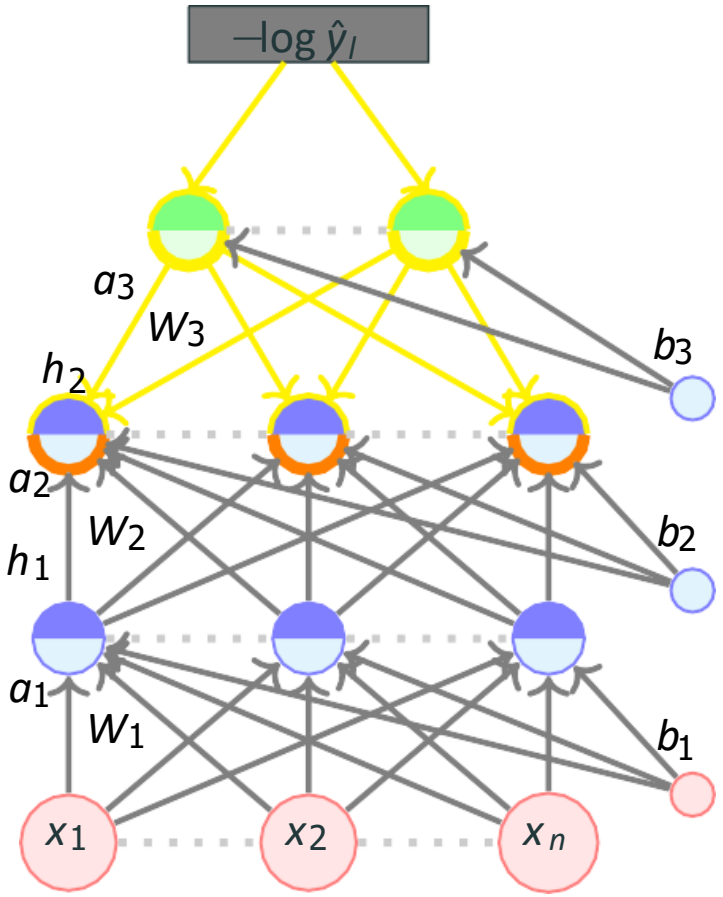


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} g'(a_{i1}) \\ \vdots \\ \end{bmatrix}$$

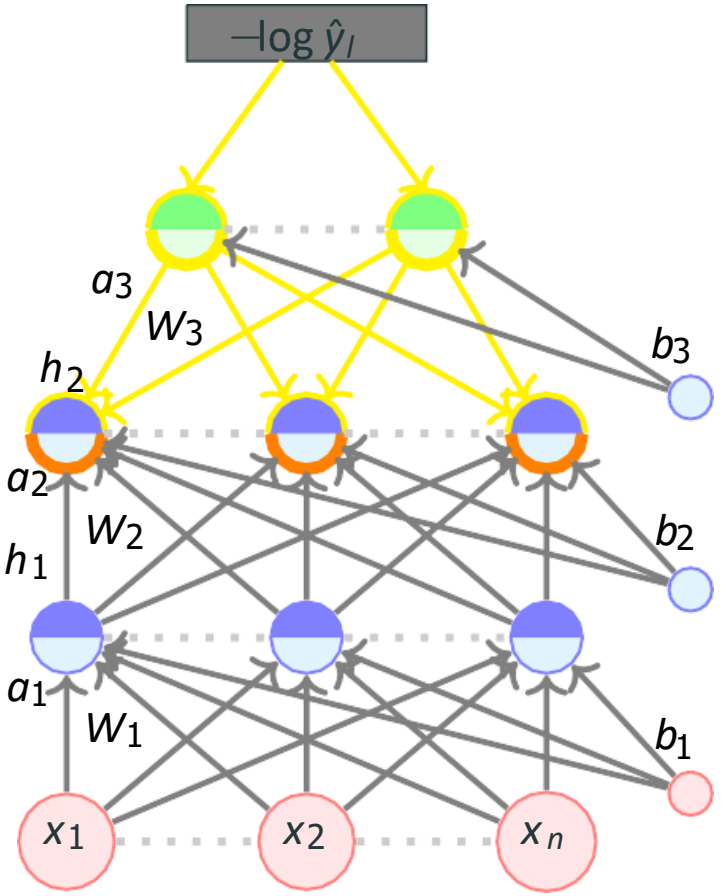


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} g'(a_{i1}) \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} g'(a_{in}) \end{bmatrix}$$



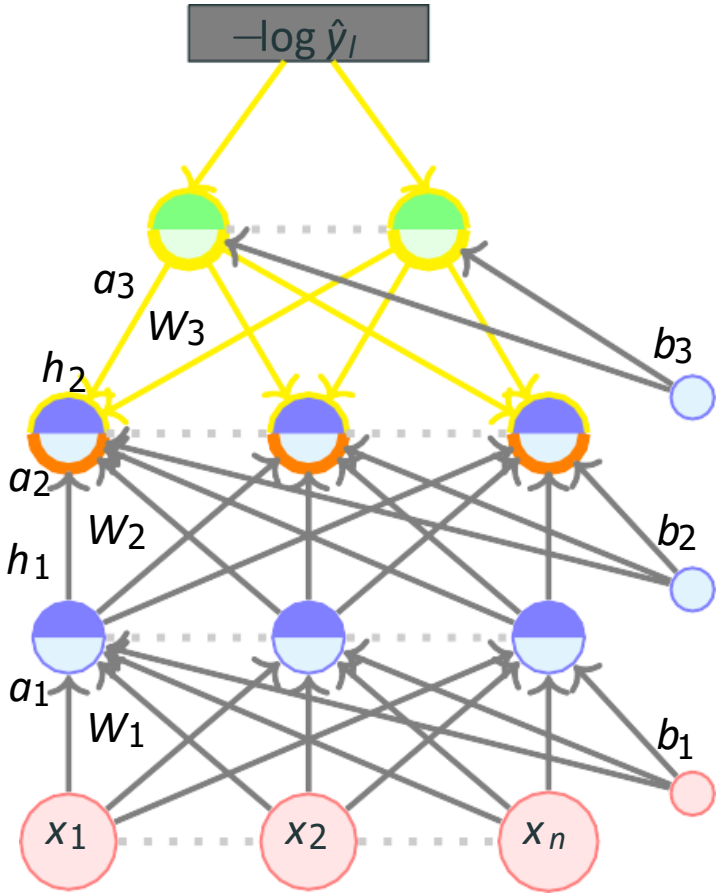
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} g'(a_{i1}) \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} g'(a_{in}) \end{bmatrix}$$

$$= \nabla_{h_i} \mathscr{L}(\theta) \odot [\ldots, g'(a_{ik}), \ldots]$$



• Backpropagation

Computing Gradients w.r.t. Parameters

Quantities of interest (roadmap for the remaining part):

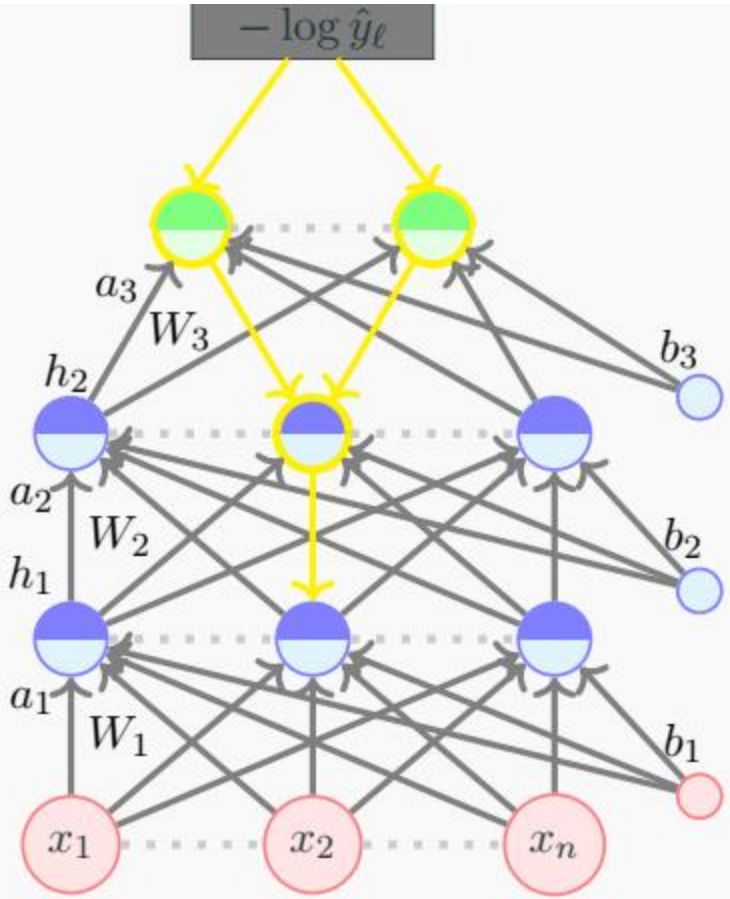
- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights and biases

$$\underbrace{\frac{\partial \mathcal{L}(\theta)}{\partial W_{111}}}_{\text{Talk to the weight directly}} = \underbrace{\frac{\partial \mathcal{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\text{Talk to the output layer}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\text{and now talk to the weights}}$$

- Our focus is on *Cross entropy loss* and *Softmax* output.

Recall that,

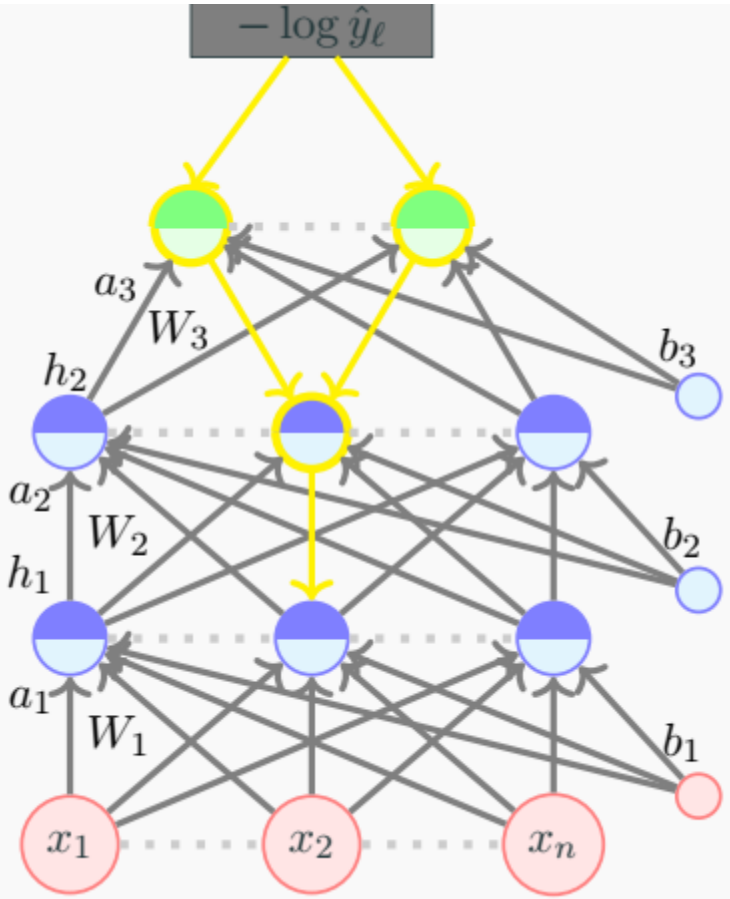
$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$



Recall that,

$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

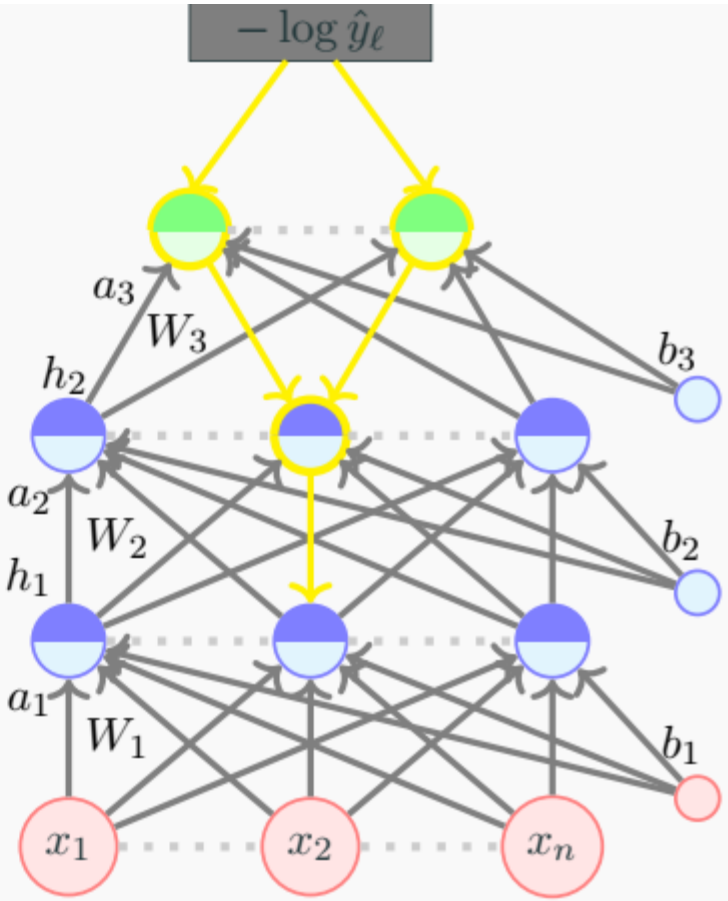


Recall that,

$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}}$$

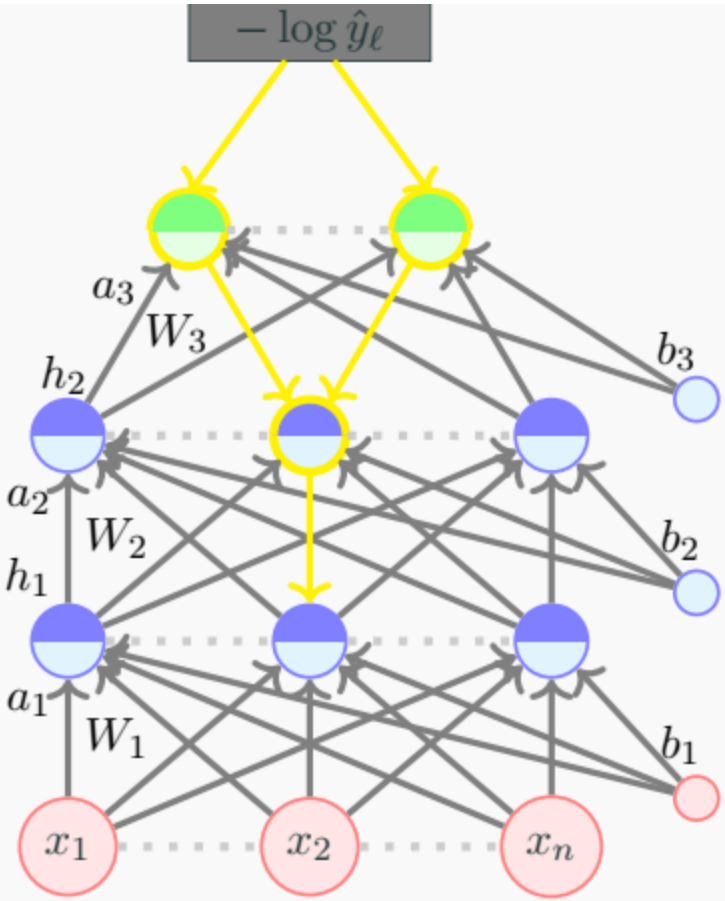


Recall that,

$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$



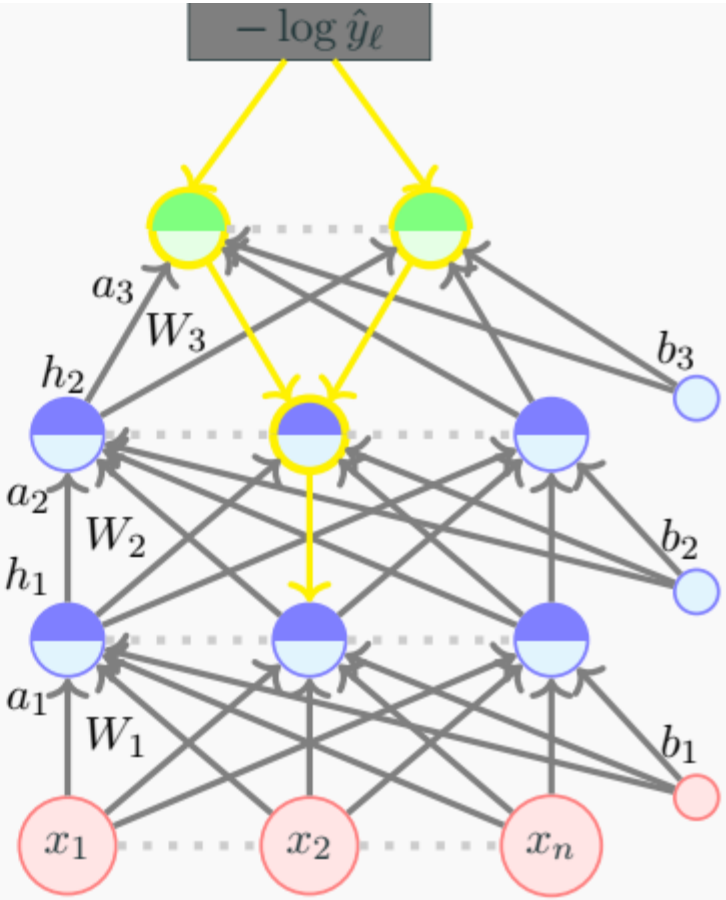
Recall that,

$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}$$



Recall that,

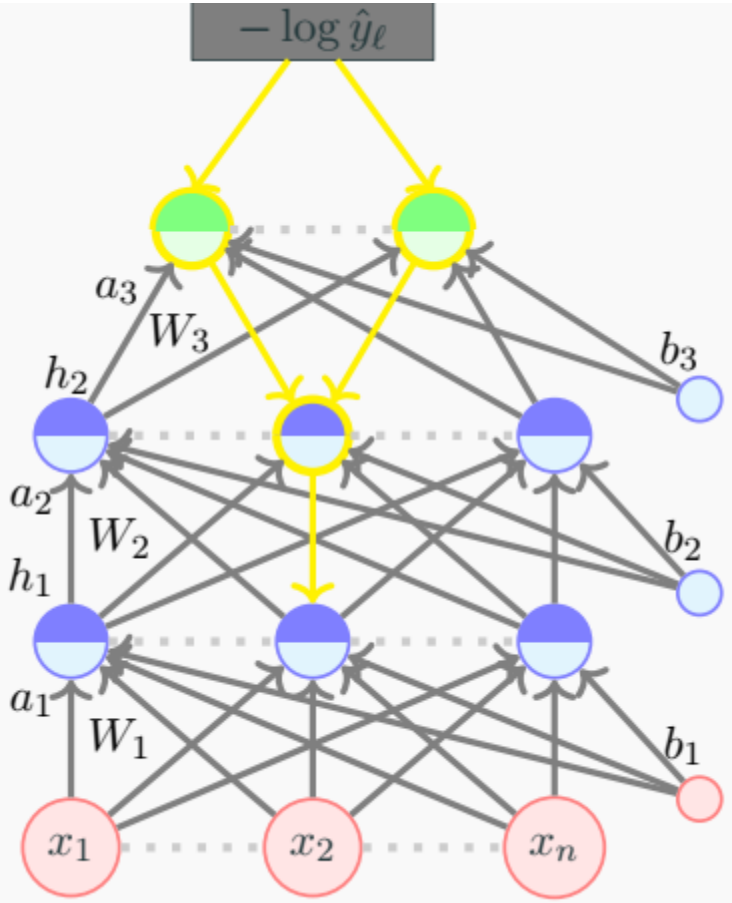
$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}$$

$$\nabla_{W_k} \mathscr{L}(\theta) =$$



Recall that,

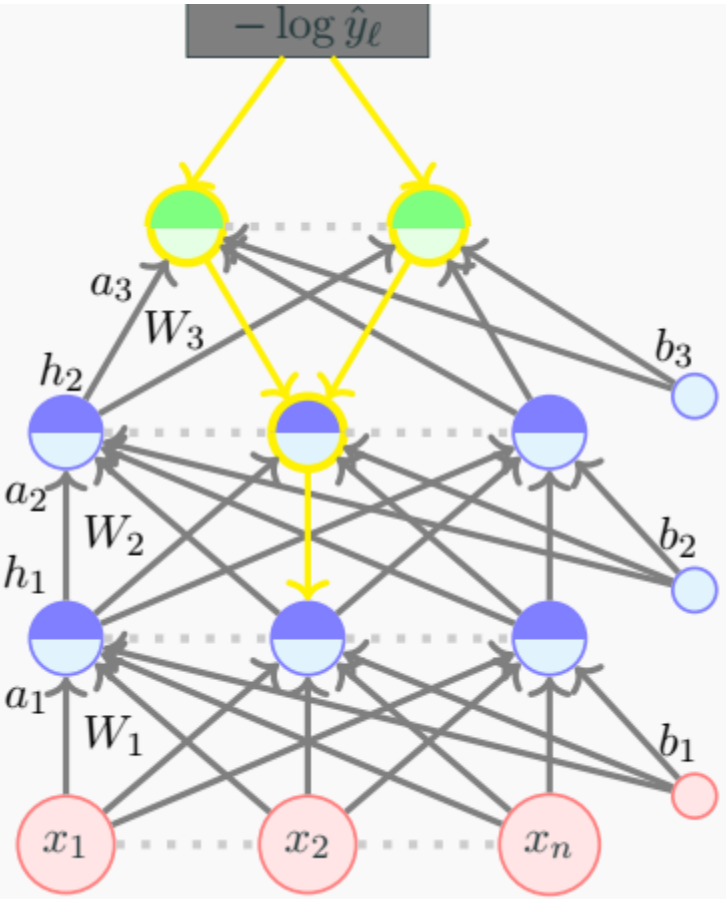
$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}$$

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \cdots & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k1n}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{knn}} \end{bmatrix}$$



Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

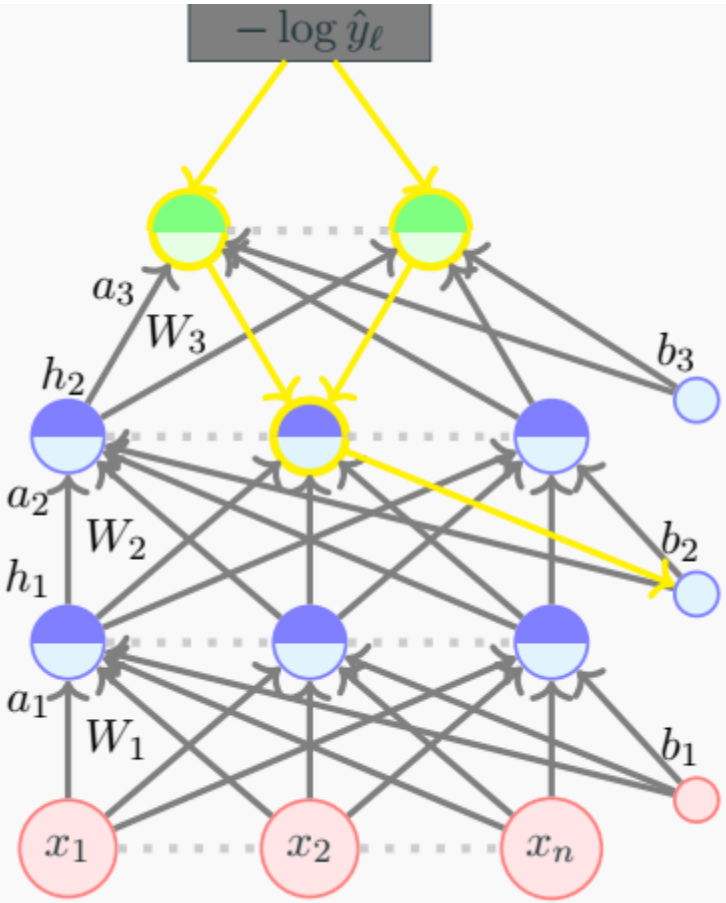
$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} =$$

Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k13}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k23}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k31}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k32}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k33}} \end{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{kij}}$$

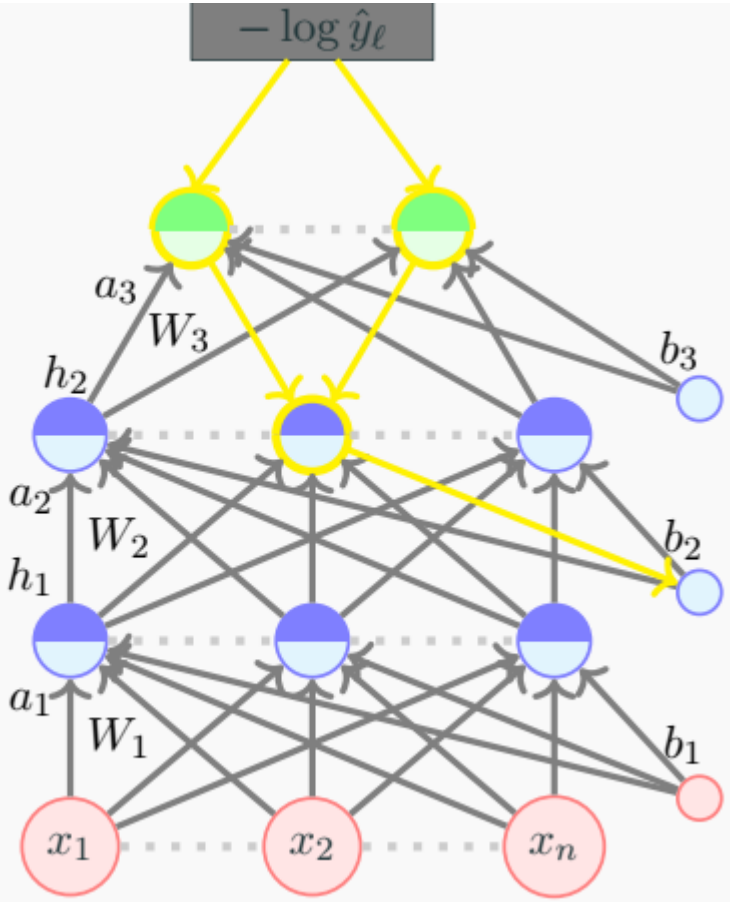
$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k3}} h_{k-1,3} \end{bmatrix} = \nabla_{a_k}\mathscr{L}(\theta) \cdot \mathbf{h_{k-1}}^T$$

Finally, coming to the biases



Finally, coming to the biases

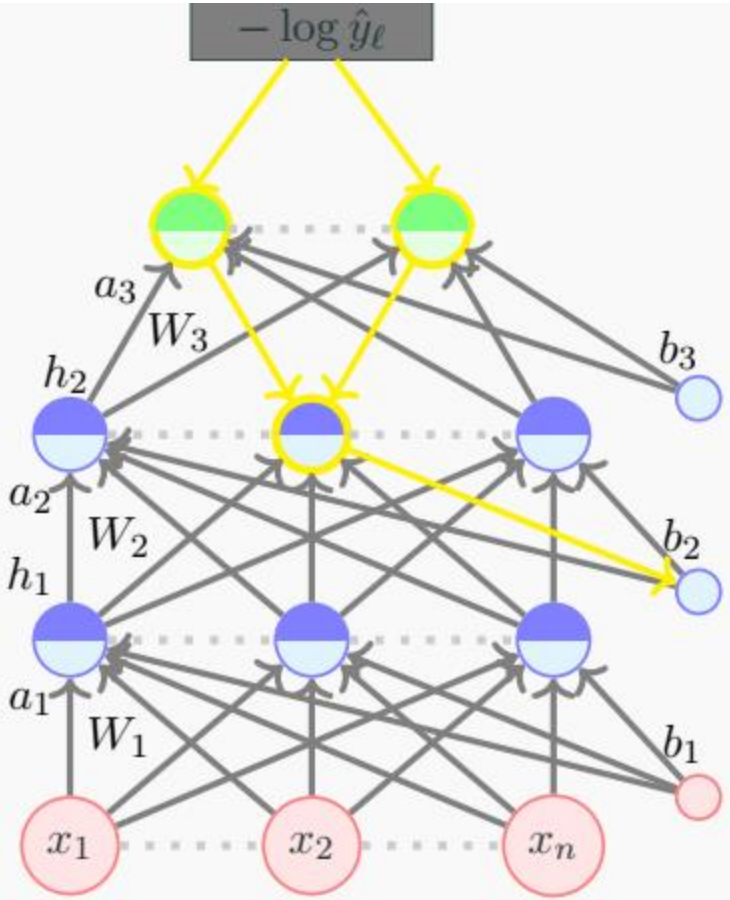
$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$



Finally, coming to the biases

$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$



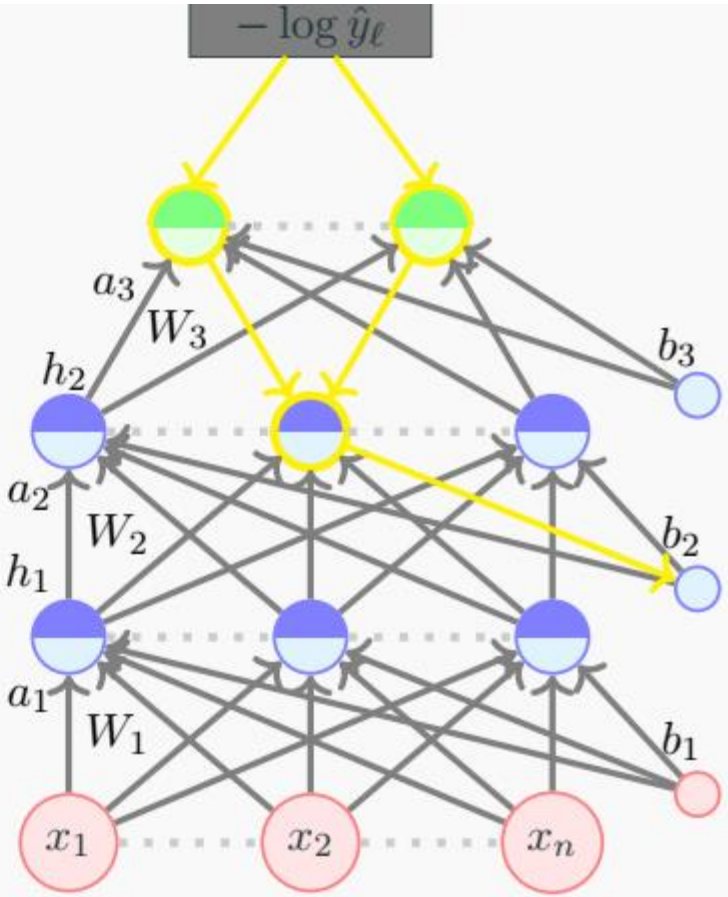
Finally, coming to the biases

$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}}$$

We can now write the gradient w.r.t. the vector b_k



Finally, coming to the biases

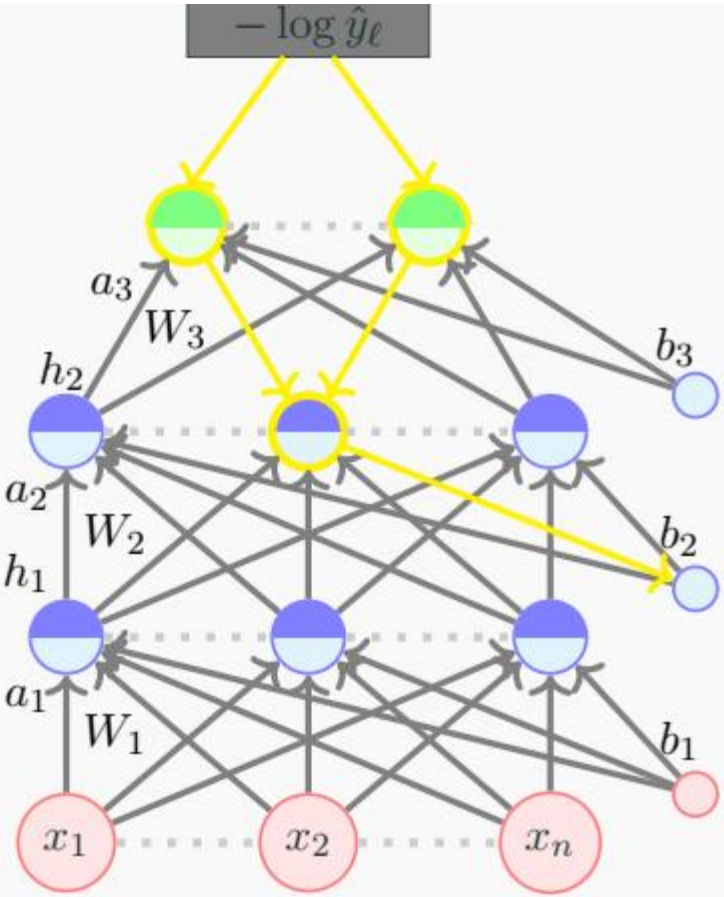
$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}}$$

We can now write the gradient w.r.t. the vector b_k

$$\nabla_{\mathbf{b_k}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{a_{k1}} \\ \frac{\partial \mathscr{L}(\theta)}{a_{k2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{a_{kn}} \end{bmatrix}$$



Finally, coming to the biases

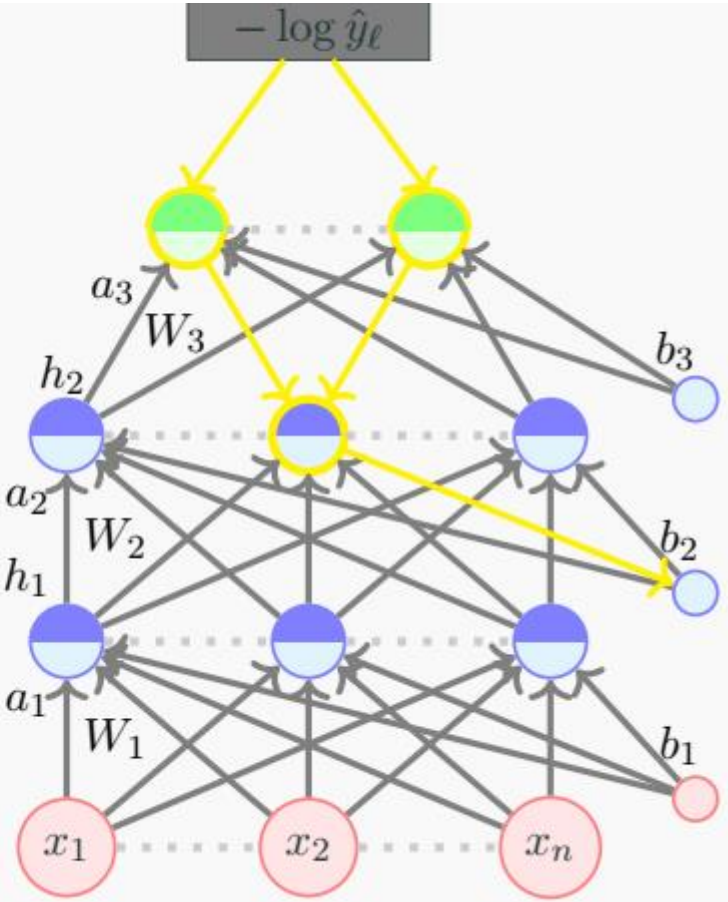
$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}}$$

We can now write the gradient w.r.t. the vector b_k

$$\nabla_{\mathbf{b_k}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{a_{k1}} \\ \frac{\partial \mathscr{L}(\theta)}{a_{k2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{a_{kn}} \end{bmatrix} = \nabla_{\mathbf{a_k}} \mathscr{L}(\theta)$$



- **Backpropagation: Pseudo code**

Algorithm: gradient_descent()

$t \leftarrow 0$;

$max\_iterations \leftarrow 1000$;

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

while $t{+}{+} < max\_iterations$ **do**

$h_1, h_2, ..., h_{L-1}, a_1, a_2, ..., a_L, \hat{y} = forward\_propagation(\theta_t)$;

$\nabla\theta_t = backward\_propagation(h_1, h_2, ..., h_{L-1}, a_1, a_2, ..., a_L, y, \hat{y})$;

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

end

Algorithm: forward_propagation(θ)

for $k = L\text{-}1$ to 1 **do**

$a_k = b_k + W_k h_{k-1}$;

$h_k = g(a_k)$;

end

$a_L = b_L + W_L h_{L-1}$;

$\hat{y} = O(a_L)$;

Just do a forward propagation and compute all h_i's, a_i's, and $\hat{y}$

Algorithm: back_propagation($h_1, h_2, ..., h_{L-1}, a_1, a_2, ..., a_L, y, \hat{y}$)

//Compute output gradient ;

$\nabla_{a_L} \mathscr{L}(\theta) = -(e(y) - \hat{y})$;

for $k = L$ *to* 1 **do**

  // Compute gradients w.r.t. parameters ;

  $\nabla_{W_k} \mathscr{L}(\theta) = \nabla_{a_k} \mathscr{L}(\theta) h_{k-1}^T$;

  $\nabla_{b_k} \mathscr{L}(\theta) = \nabla_{a_k} \mathscr{L}(\theta)$;

  // Compute gradients w.r.t. layer below ;

  $\nabla_{h_{k-1}} \mathscr{L}(\theta) = W_k^T (\nabla_{a_k} \mathscr{L}(\theta))$;

  // Compute gradients w.r.t. layer below (pre-activation);

  $\nabla_{a_{k-1}} \mathscr{L}(\theta) = \nabla_{h_{k-1}} \mathscr{L}(\theta) \odot [\ldots, g'(a_{k-1,j}), \ldots]$;

end

- **Thank You**