第四讲：前馈神经网络和反向传播

张盛平

s.zhang@hit.edu.cn

计算学部
哈尔滨工业大学

2021 年秋季学期

致谢

- 讲稿中很多资料或素材来源于网络，包括国外一些大学的相关课程、一些博客、维基百科等。后面不一一列举来源，在此一并表示感谢
- 引用网络资源时，由于本人的理解能力，可能存在一些偏差
- 本讲稿会经常更新

References/Acknowledgments

See the excellent videos by Hugo Larochelle on Backpropagation

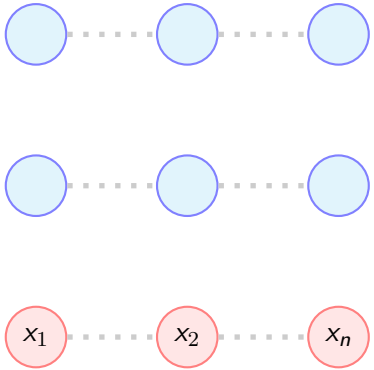
前馈神经网络 (又称多层神经网络)

- 网络的输入是一个 $\mathbf{n}$-D 向量

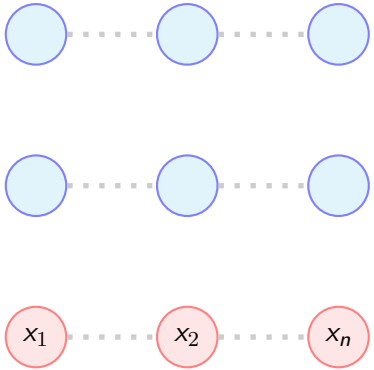
- 网络的输入是一个 n-D 向量

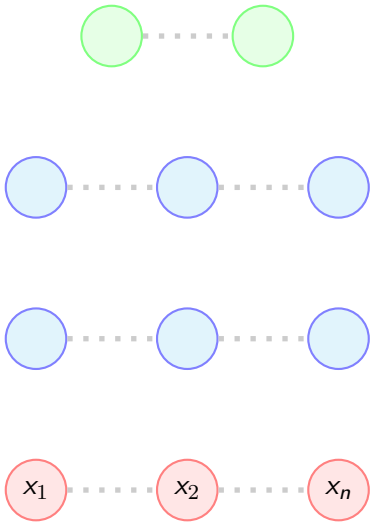
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元

- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元

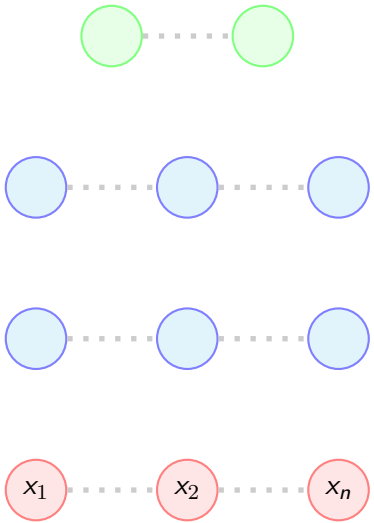


- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)

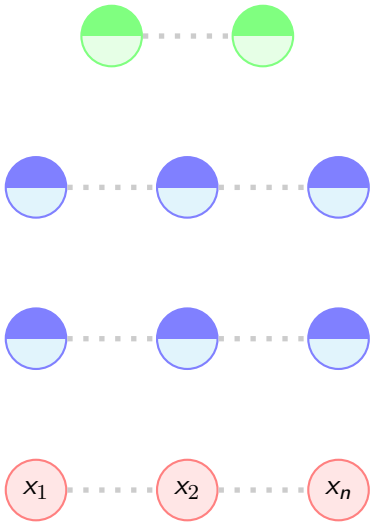




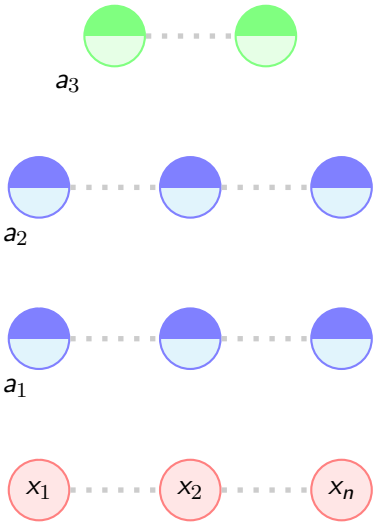
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)



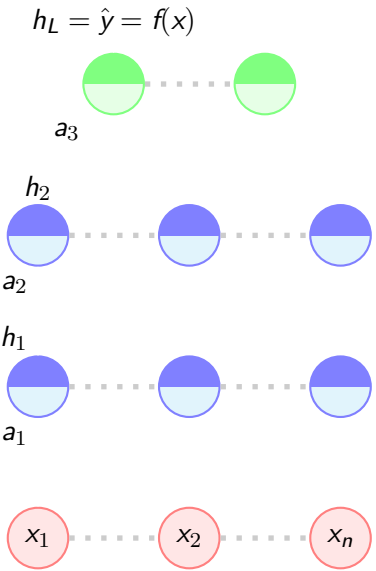
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：



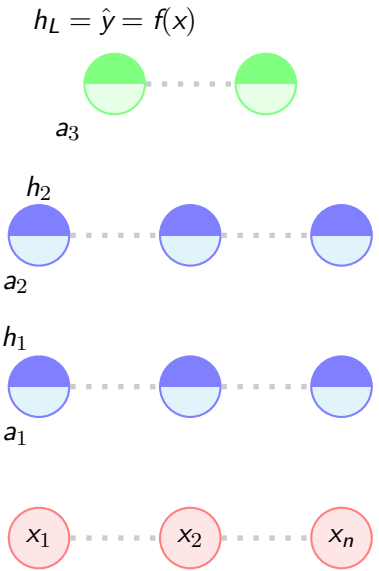
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：



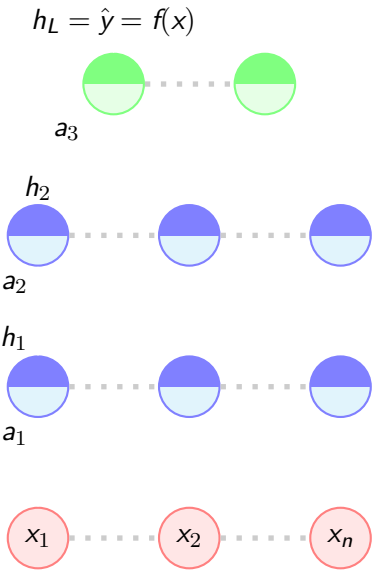
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：预激活（pre-activation）



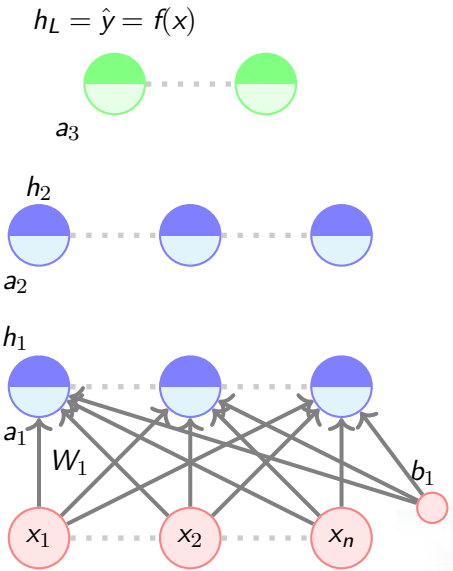
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：预激活 (pre-activation) 和激活 (activation)



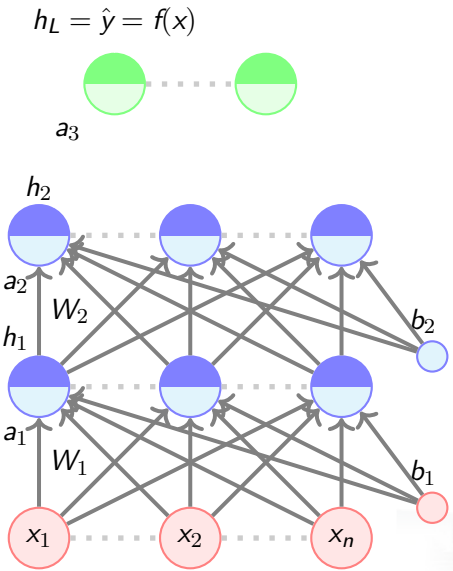
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：预激活 (pre-activation) 和激活 (activation) (a_i 和 h_i 是向量)



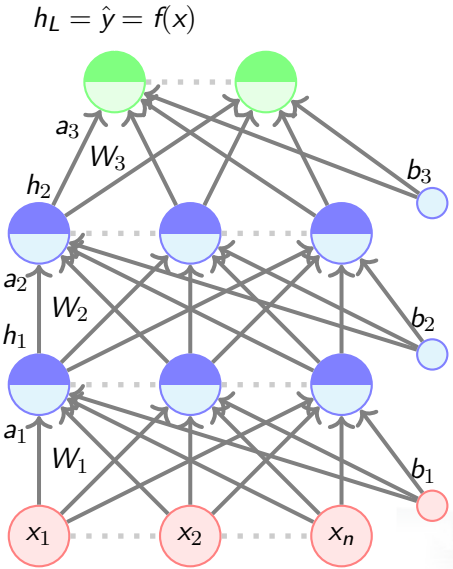
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：预激活 (pre-activation) 和激活 (activation) (a_i 和 h_i 是向量)
- 网络的输入层也称为第 0 层，输出层被称为第 (L) 层



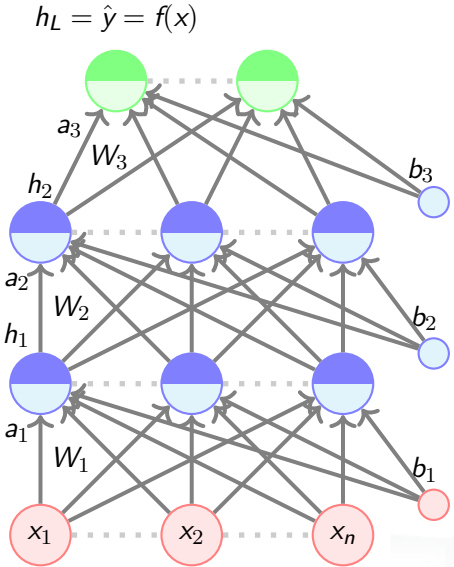
- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：预激活 (pre-activation) 和激活 (activation) (a_i 和 h_i 是向量)
- 网络的输入层也称为第 0 层，输出层被称为第 (L) 层
- $W_i \in \mathbb{R}^{n \times n}$ 和 $b_i \in \mathbb{R}^n$ 分别是层 $i-1$ 到 i $(0 < i < L)$ 的权重和偏置向量



- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：预激活 (pre-activation) 和激活 (activation) (a_i 和 h_i 是向量)
- 网络的输入层也称为第 0 层，输出层被称为第 (L) 层
- $W_i \in \mathbb{R}^{n \times n}$ 和 $b_i \in \mathbb{R}^n$ 分别是层 $i-1$ 到 i $(0 < i < L)$ 的权重和偏置向量

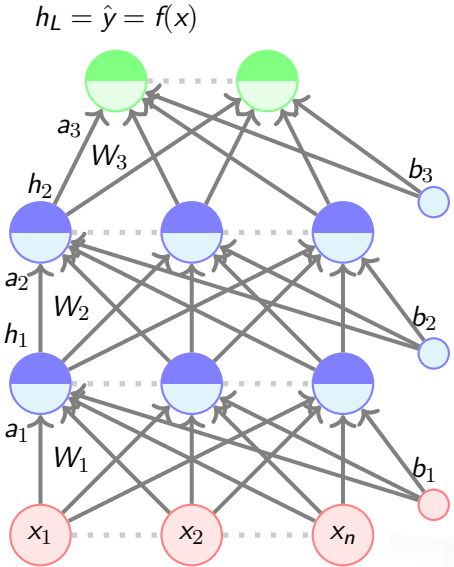


- 网络的输入是一个 $\mathbf{n}$-D 向量
- 网络包含 $\mathbf{L-1}$ 隐含层 (2, in this case)，每个隐含层有 $\mathbf{n}$ 神经元
- 网络有一个输出层，包含 $\mathbf{k}$ 个神经元 (对应分类问题中的 $\mathbf{k}$ 个类别)
- 隐含层和输出层中的每一个神经元可以划分为两部分：预激活 (pre-activation) 和激活 (activation) (a_i 和 h_i 是向量)
- 网络的输入层也称为第 0 层，输出层被称为第 (L) 层
- $W_i \in \mathbb{R}^{n \times n}$ 和 $b_i \in \mathbb{R}^n$ 分别是层 $i-1$ 到 i $(0 < i < L)$ 的权重和偏置向量
- $W_L \in \mathbb{R}^{n \times k}$ 和 $b_L \in \mathbb{R}^k$ 分别是最后一个隐含层到输出层的权重和偏置向量 ($L = 3$ in this case)



- 层 i 的预激活是

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

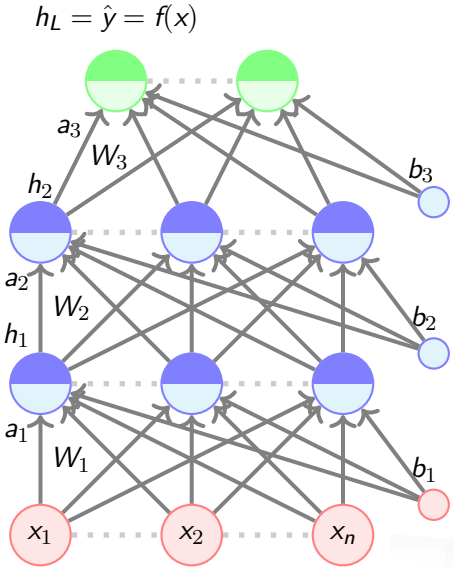


- 层 i 的预激活是

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

- 层 i 的激活是

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



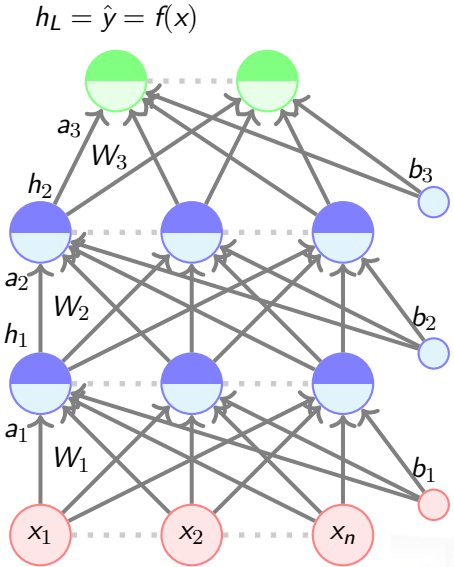
- 层 i 的预激活是

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

- 层 i 的激活是

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

其中 g 称为激活函数（例如，logistic，tanh，linear，*etc.*）



- 层 i 的预激活是

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

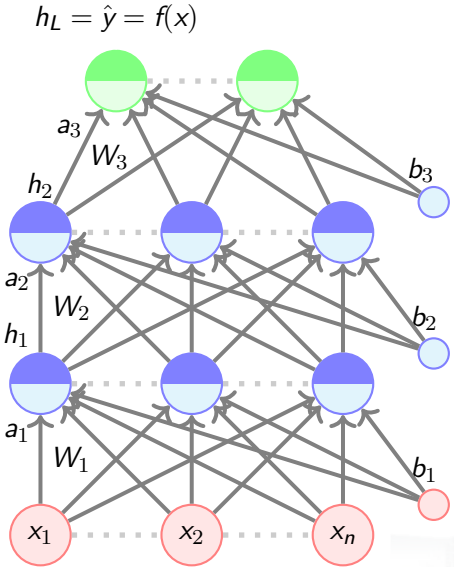
- 层 i 的激活是

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

其中 g 称为激活函数（例如，logistic，tanh，linear，*etc.*）

- 输出层的激活是

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



- 层 i 的预激活是

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

- 层 i 的激活是

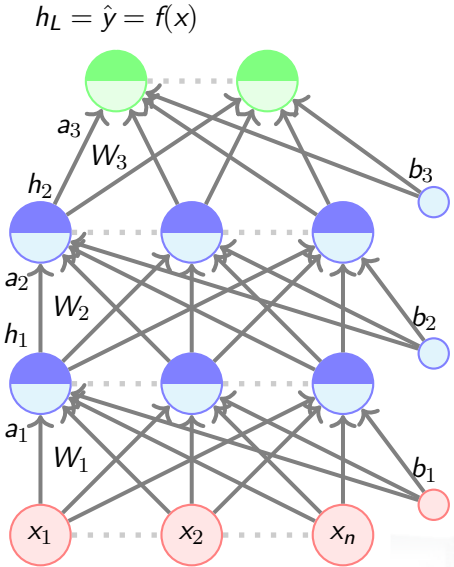
$$h_i(x) = g(a_i(x))$$

其中 g 称为激活函数 (例如, logistic, tanh, linear, *etc.*)

- 输出层的激活是

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

其中 O 是输出层激活函数 (如, softmax, linear, *etc.*)



- 层 i 的预激活是

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

- 层 i 的激活是

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

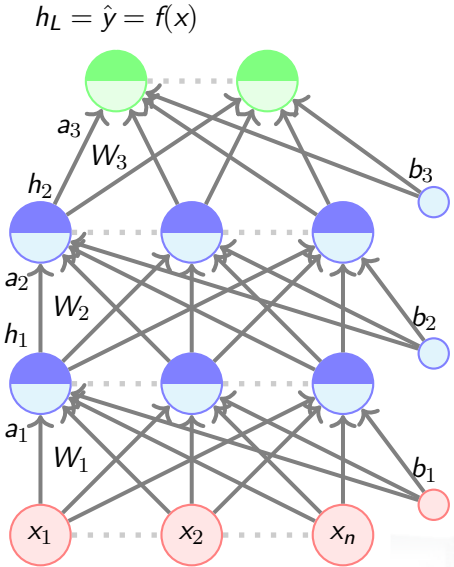
其中 g 称为激活函数 (例如, logistic, tanh, linear, *etc.*)

- 输出层的激活是

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

其中 O 是输出层激活函数 (如, softmax, linear, *etc.*)

- 为了描述简单, 将 $a_i(x)$ 简化为 a_i, 将 $h_i(x)$ 简



- 层 i 的预激活是

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

- 层 i 的激活是

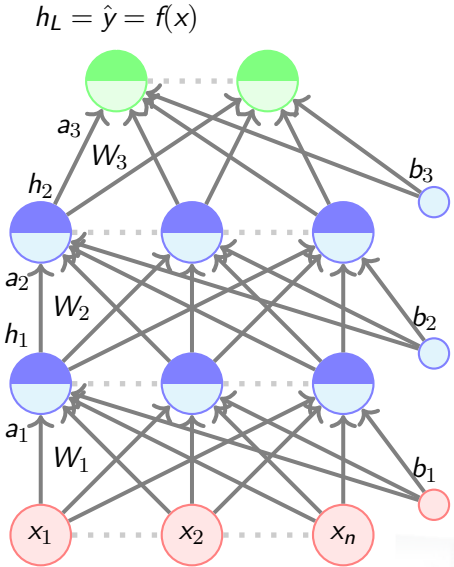
$$h_i = g(a_i)$$

其中 g 称为激活函数（例如，logistic，tanh，linear，*etc.*）

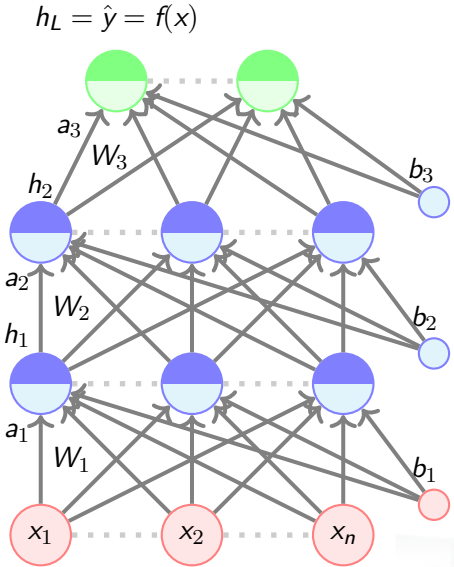
- 输出层的激活函数是

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

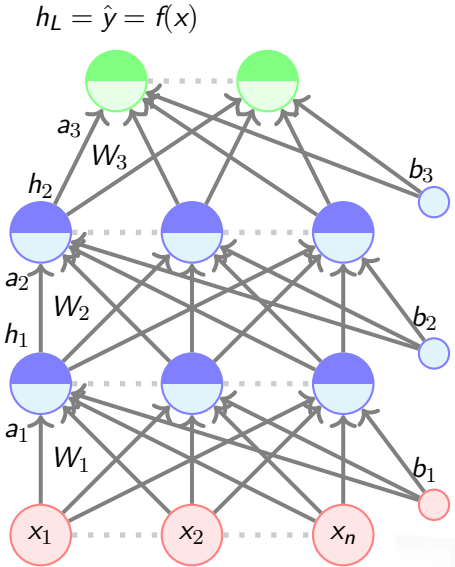
其中 O 是输出层的激活函数（例如，softmax，linear，*etc.*）



- **数据:** $\{x_i, y_i\}_{i=1}^{N}$

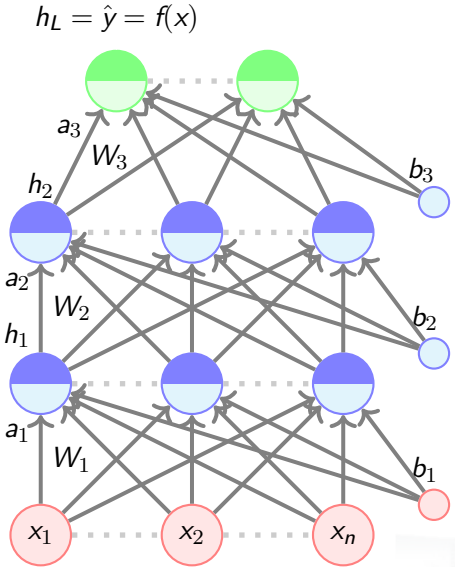


- **数据:** $\{x_i, y_i\}_{i=1}^{N}$
- **模型:**



- **数据:** $\{x_i, y_i\}_{i=1}^{N}$
- **模型:**

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

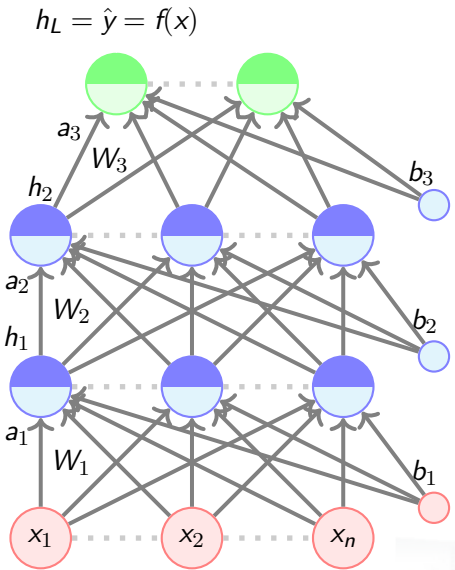


- **数据:** $\{x_i, y_i\}_{i=1}^{N}$
- **模型:**

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

- **参数:**

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



- **数据:** $\{x_i, y_i\}_{i=1}^N$
- **模型:**

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

- **参数:**

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

- **目标/损失/误差函数:**

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

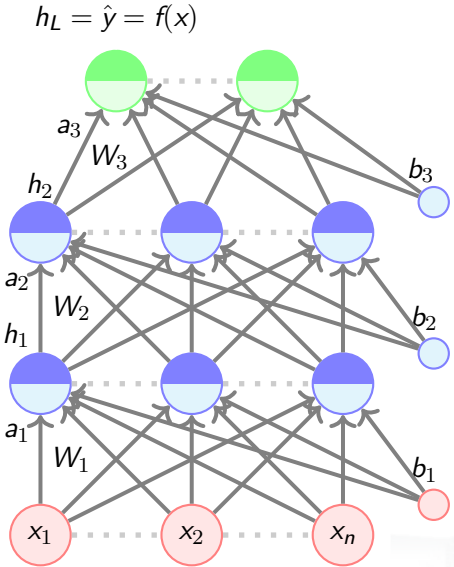
一般地, $min \ \mathscr{L}(\theta)$

其中 $\mathscr{L}(\theta)$ 是损失函数

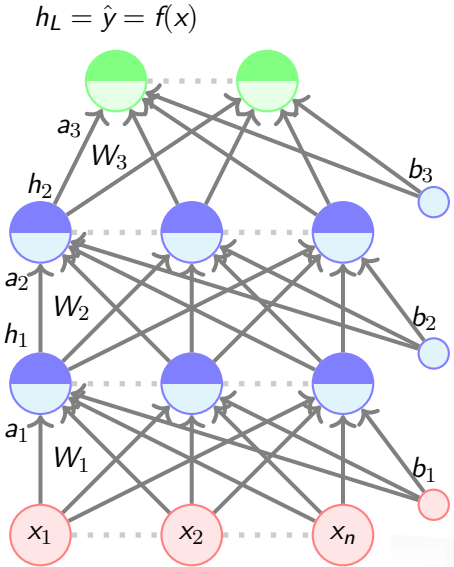
前馈神经网络参数学习 (Intuition)

The story so far...

- 已经介绍了前馈神经网络
- 希望找到一个算法来学习这个网络的参数



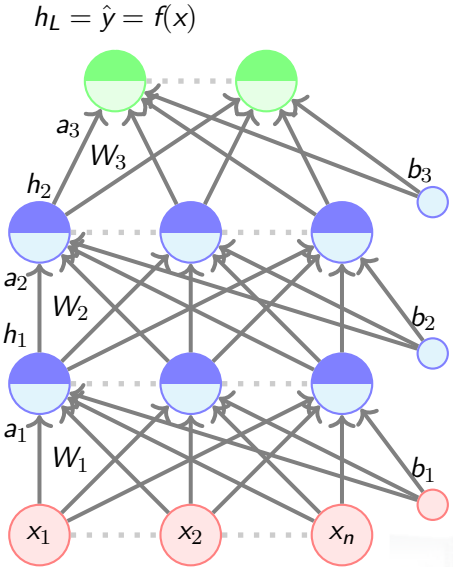
- 回忆一下梯度下降算法



- 回忆一下梯度下降算法

Algorithm: gradient_descent()

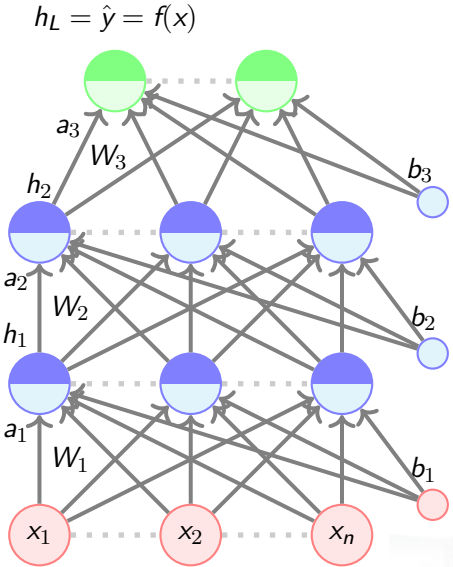
$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize w_0, b_0;
while $t{+}{+} < max\_iterations$ **do**
  $w_{t+1} \leftarrow w_t - \eta \nabla w_t$;
  $b_{t+1} \leftarrow b_t - \eta \nabla b_t$;
end



- 回忆一下梯度下降算法
- 更紧凑的可以写为

Algorithm: gradient_descent()

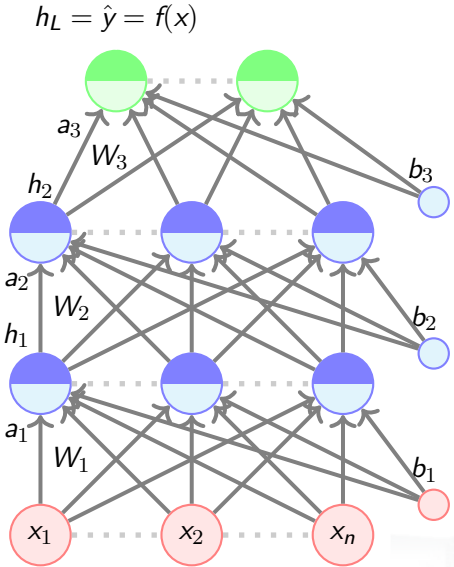
$t \leftarrow 0$;
max_iterations $\leftarrow 1000$;
Initialize w_0, b_0;
while *t++* $<$ *max_iterations* **do**
  $w_{t+1} \leftarrow w_t - \eta \nabla w_t$;
  $b_{t+1} \leftarrow b_t - \eta \nabla b_t$;
end



- 回忆一下梯度下降算法
- 更紧凑的可以写为

Algorithm: gradient_descent()

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

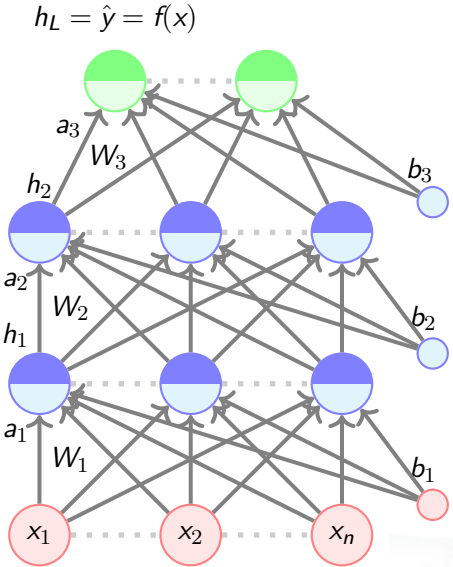


- 回忆一下梯度下降算法
- 更紧凑的可以写为

Algorithm: gradient_descent()

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

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

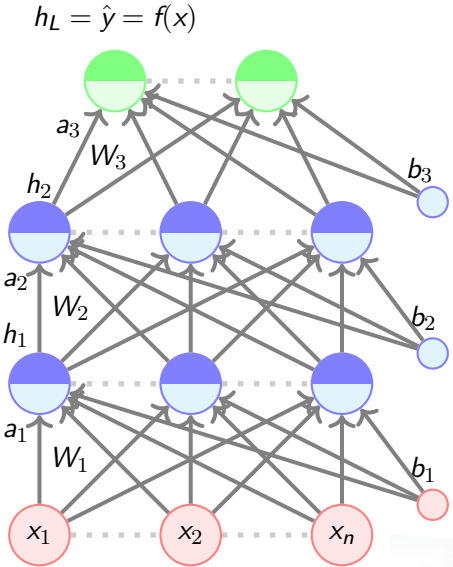


- 回忆一下梯度下降算法
- 更紧凑的可以写为

Algorithm: gradient_descent()

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

- 其中 $\nabla \theta_t = \left[\frac{\partial \mathscr{L}(\theta)}{\partial w_t}, \frac{\partial \mathscr{L}(\theta)}{\partial b_t}\right]^T$
- 现在，在前馈神经网络中不再是 $\theta = [w, b]$，而是 $\theta = [W_1, W_2, .., W_L, b_1, b_2, .., b_L]$

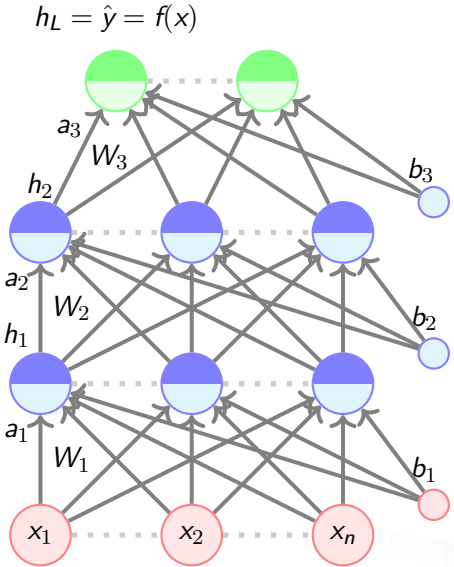


- 回忆一下梯度下降算法
- 更紧凑的可以写为

Algorithm: gradient_descent()

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

- 其中 $\nabla \theta_t = \left[\frac{\partial \mathscr{L}(\theta)}{\partial w_t}, \frac{\partial \mathscr{L}(\theta)}{\partial b_t}\right]^T$
- 现在，在前馈神经网络中不再是 $\theta = [w, b]$，而是 $\theta = [W_1, W_2, .., W_L, b_1, b_2, .., b_L]$
- 仍然可以使用上述梯度下降算法来求解前馈神经网络的参数



- 回忆一下梯度下降算法
- 更紧凑的可以写为

Algorithm: gradient_descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize $\theta_0 = [W_1^0, ..., W_L^0, b_1^0, ..., b_L^0]$;
while $t{+}{+} < max\_iterations$ **do**
  $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;
end

- 其中 $\nabla \theta_t = \left[\frac{\partial \mathscr{L}(\theta)}{\partial W_{1,t}}, .., \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,t}}, \frac{\partial \mathscr{L}(\theta)}{\partial b_{1,t}}, .., \frac{\partial \mathscr{L}(\theta)}{\partial b_{L,t}}\right]^T$
- 现在，在前馈神经网络中不再是 $\theta = [w, b]$，而是 $\theta = [W_1, W_2, .., W_L, b_1, b_2, .., b_L]$
- 仍然可以使用上述梯度下降算法来求解前馈神经网络的参数

- 现在 $\nabla\theta$ 看起来更复杂

- 现在 $\nabla\theta$ 看起来更复杂

$$\begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} \\ \\ \\ \\ \\ \\ \end{bmatrix}$$

- 现在 $\nabla\theta$ 看起来更复杂

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

- 现在 $\nabla\theta$ 看起来更复杂

$$\begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{11n}} \\ & & \\ & & \\ & & \\ & & \end{bmatrix}$$

- 现在 $\nabla\theta$ 看起来更复杂

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

- 现在 $\nabla\theta$ 看起来更复杂

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

- 现在 $\nabla\theta$ 看起来更复杂

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

- 现在 $\nabla\theta$ 看起来更复杂

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

- 现在 $\nabla\theta$ 看起来更复杂

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

- 现在 $\nabla\theta$ 看起来更复杂

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

- $\nabla\theta$ 由
  $\nabla W_1, \nabla W_2, ...\nabla W_{L-1} \in \mathbb{R}^{n\times n}, \nabla W_L \in \mathbb{R}^{n\times k}$,
  $\nabla b_1, \nabla b_2, ..., \nabla b_{L-1} \in \mathbb{R}^n$ and $\nabla b_L \in \mathbb{R}^k$ 组成

需要回答如下两个问题

需要回答如下两个问题

- 如何选择损失函数 $\mathscr{L}(\theta)$?

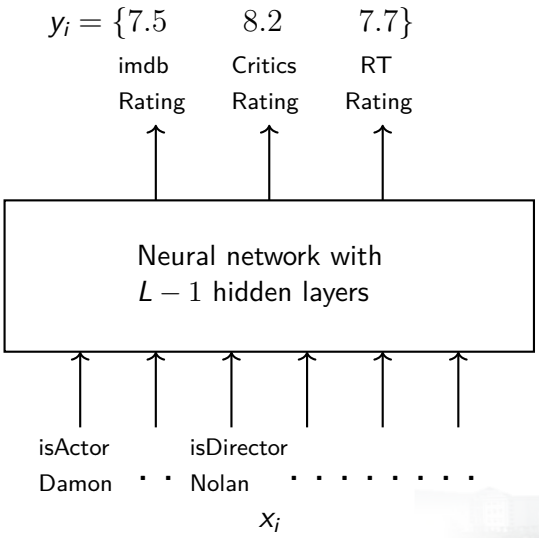
需要回答如下两个问题

- 如何选择损失函数 $\mathscr{L}(\theta)$?
- 如何计算 $\nabla\theta$?

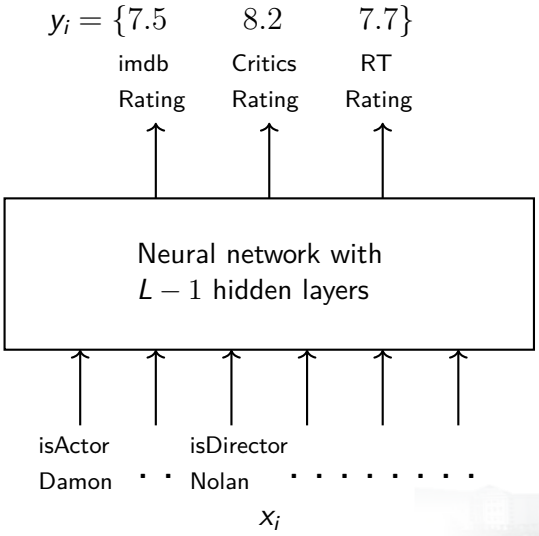
输出函数和损失函数

- 损失函数的选择依赖问题本身

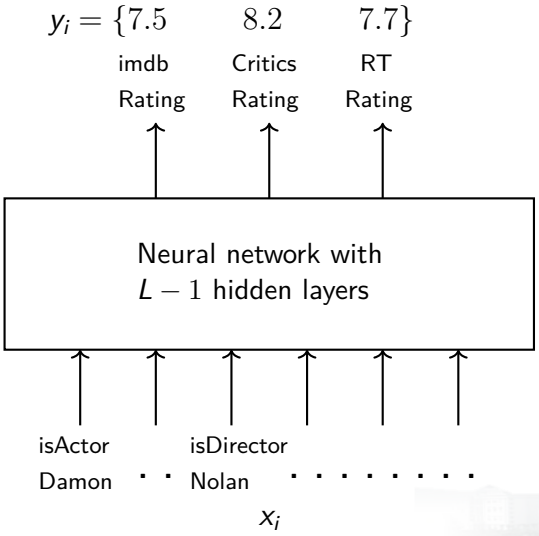
- 损失函数的选择依赖问题本身
- 通过两个例子来阐述



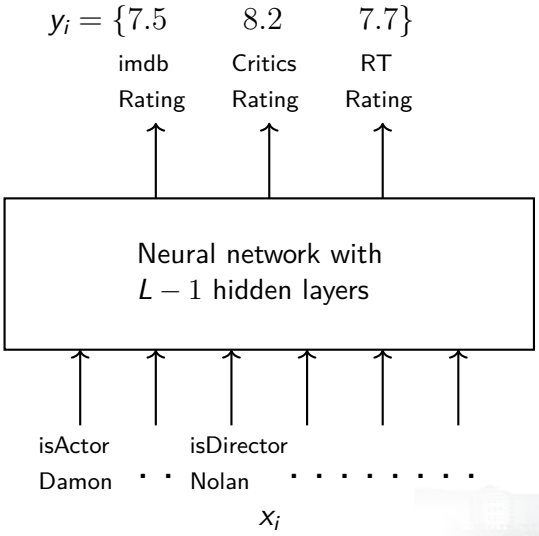
- 损失函数的选择依赖问题本身
- 通过两个例子来阐述
- 考虑之前举过的电影例子，有一点变化，现在我们感兴趣的是预测电影的打分（ratings）



- 损失函数的选择依赖问题本身
- 通过两个例子来阐述
- 考虑之前举过的电影例子，有一点变化，现在我们感兴趣的是预测电影的打分（ratings）
- 因此 $y_i \in \mathbb{R}^3$

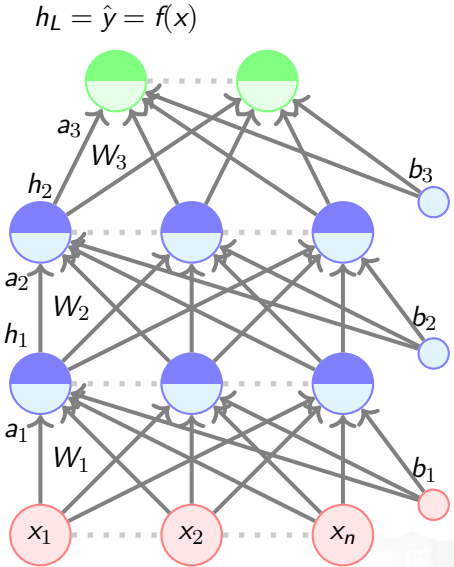


- 损失函数的选择依赖问题本身
- 通过两个例子来阐述
- 考虑之前举过的电影例子，有一点变化，现在我们感兴趣的是预测电影的打分（ratings）
- 因此 $y_i \in \mathbb{R}^3$
- 损失函数应该能刻画 $\hat{y}_i$ 和 y_i 之间的差异

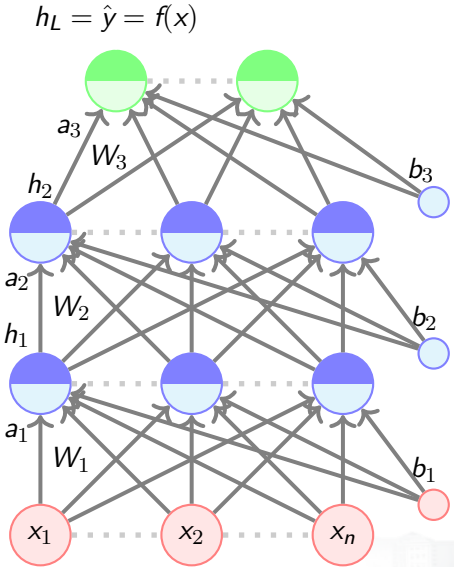


- 损失函数的选择依赖问题本身
- 通过两个例子来阐述
- 考虑之前举过的电影例子，有一点变化，现在我们感兴趣的是预测电影的打分（ratings）
- 因此 $y_i \in \mathbb{R}^3$
- 损失函数应该能刻画 $\hat{y}_i$ 和 y_i 之间的差异
- 如果 $y_i \in \mathbb{R}^n$，平方根损失能够刻画这种差异

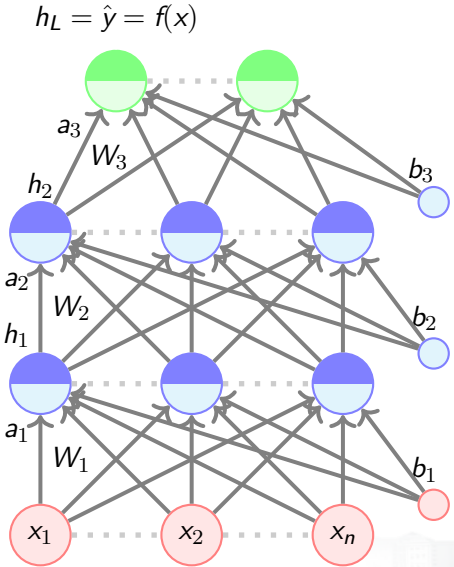
$$\mathscr{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{3} (\hat{y}_{ij} - y_{ij})^2$$



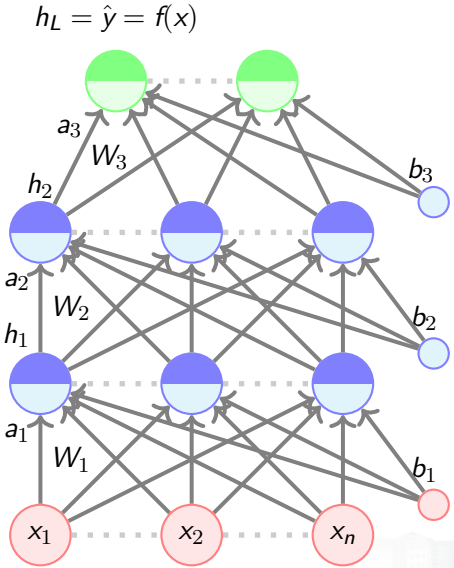
- 一个相关的问题是：如果 $y_i \in \mathbb{R}^3$，输出函数 'O' 应该是什么？



- 一个相关的问题是：如果 $y_i \in \mathbb{R}^3$，输出函数 'O' 应该是什么？
- 可以选择 logistic 函数作为输出函数吗？

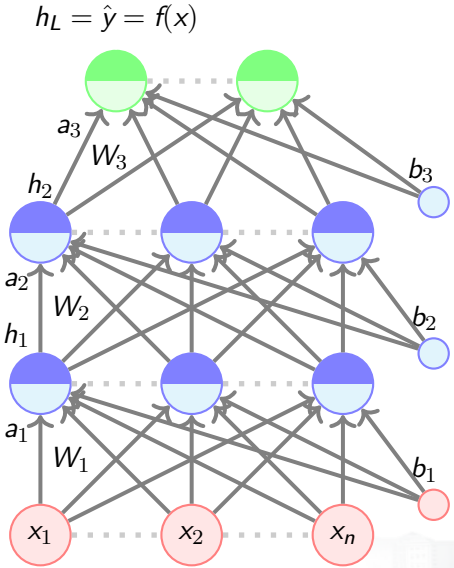


- 一个相关的问题是：如果 $y_i \in \mathbb{R}^3$，输出函数 'O' 应该是什么？
- 可以选择 logistic 函数作为输出函数吗？
- 不行，因为 logistic 函数将 $\hat{y}_i$ 的值限制在 0 和 1 之间，但我们期望 $\hat{y}_i \in \mathbb{R}^3$



- 一个相关的问题是：如果 $y_i \in \mathbb{R}^3$，输出函数 'O' 应该是什么？
- 可以选择 logistic 函数作为输出函数吗？
- 不行，因为 logistic 函数将 $\hat{y}_i$ 的值限制在 0 和 1 之间，但我们期望 $\hat{y}_i \in \mathbb{R}^3$
- 因此，在这种情况下，给 'O' 选择一个线性函数更有道理

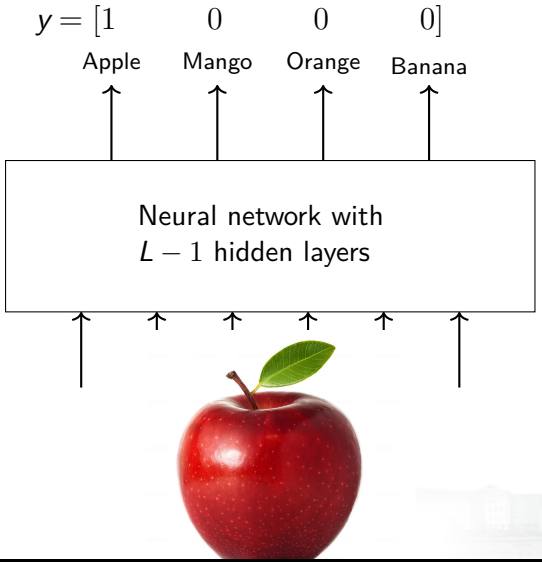
$$\begin{aligned} f(x) &= h_L = O(a_L) \\ &= W_O a_L + b_O \end{aligned}$$



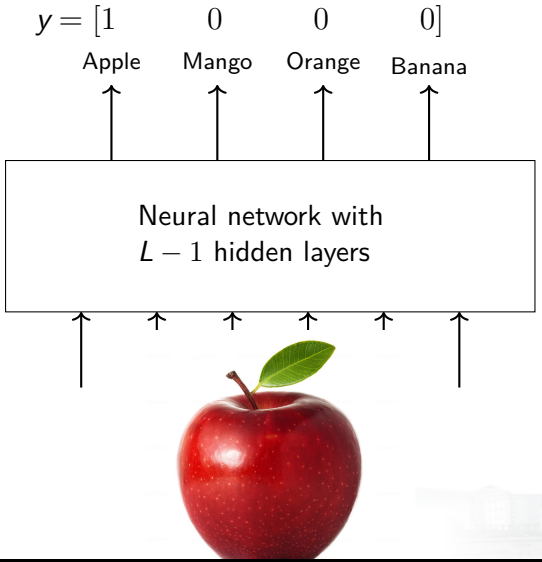
- 一个相关的问题是：如果 $y_i \in \mathbb{R}^3$，输出函数 'O' 应该是什么？
- 可以选择 logistic 函数作为输出函数吗？
- 不行，因为 logistic 函数将 $\hat{y}_i$ 的值限制在 0 和 1 之间，但我们期望 $\hat{y}_i \in \mathbb{R}^3$
- 因此，在这种情况下，给 'O' 选择一个线性函数更有道理

$$\begin{aligned} f(x) = h_L &= O(a_L) \\ &= W_O a_L + b_O \end{aligned}$$

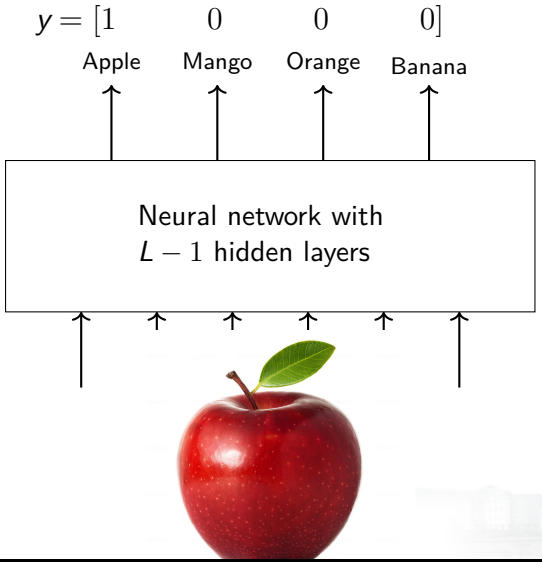
- $\hat{y}_i = f(x_i)$ 不再受限于 0 和 1 之间了


$y = [1 \quad 0 \quad 0 \quad 0]$
Apple
Mango
Orange
Banana
Neural network with
$L-1$ hidden layers

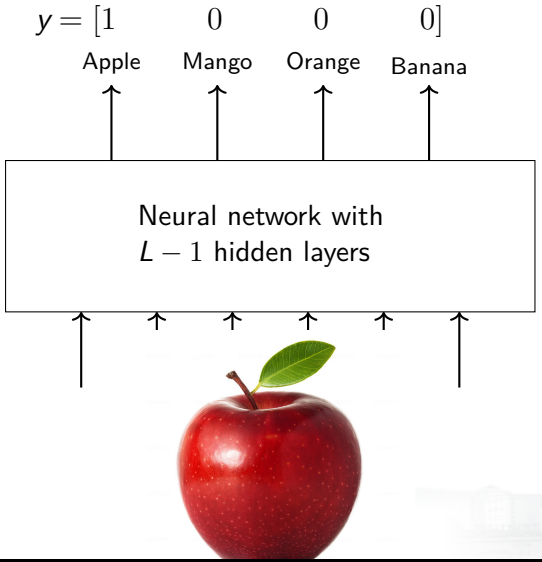
- 考虑另一个问题



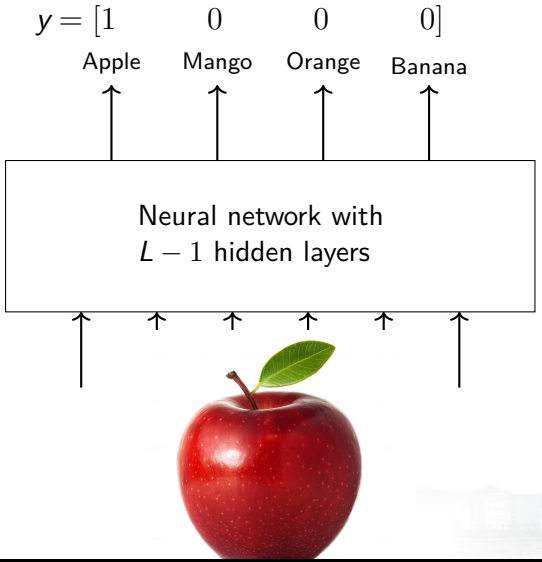
- 考虑另一个问题
- 假定我们希望将一张图像分类成 k 个类中的一个



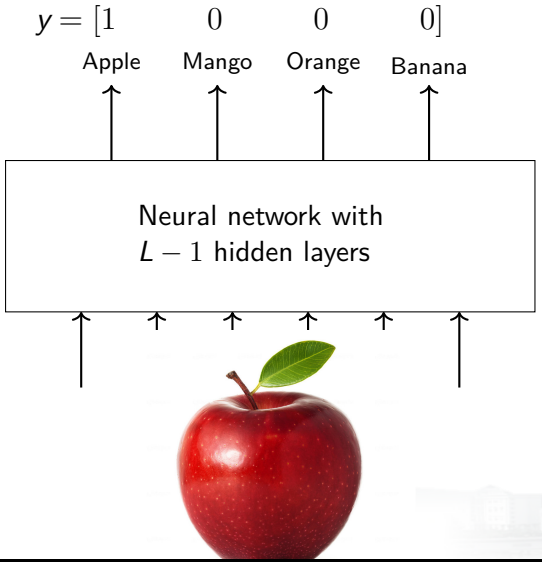
- 考虑另一个问题
- 假定我们希望将一张图像分类成 k 个类中的一个
- 同样，我们也可以使用平方根损失来刻画网络输出与真实标签之间的误差



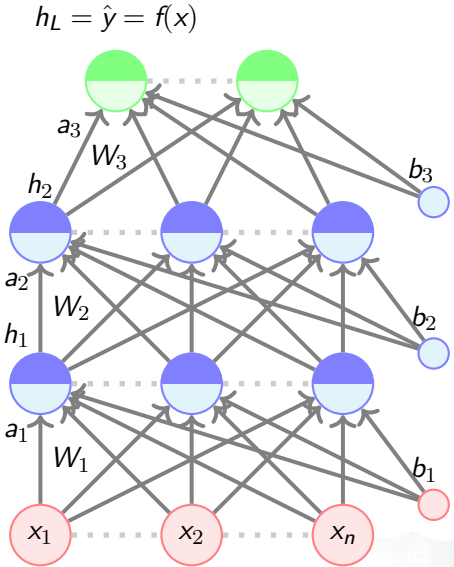
- 考虑另一个问题
- 假定我们希望将一张图像分类成 k 个类中的一个
- 同样，我们也可以使用平方根损失来刻画网络输出与真实标签之间的误差
- 还有更好的损失函数吗?


$y = [1 \quad 0 \quad 0 \quad 0]$
Apple
Mango
Orange
Banana
Neural network with
$L-1$ hidden layers

- 注意：y 是一个概率分布

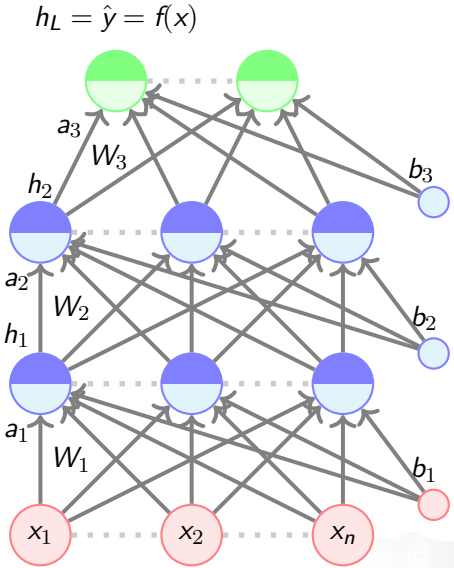


- 注意：y 是一个概率分布
- 因此我们也需要确保 $\hat{y}$ 是一个概率分布



- 注意：y 是一个概率分布
- 因此我们也需要确保 $\hat{y}$ 是一个概率分布
- 选择什么样的输出激活函数 'O' 才能确保输出是一个概率分布?

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

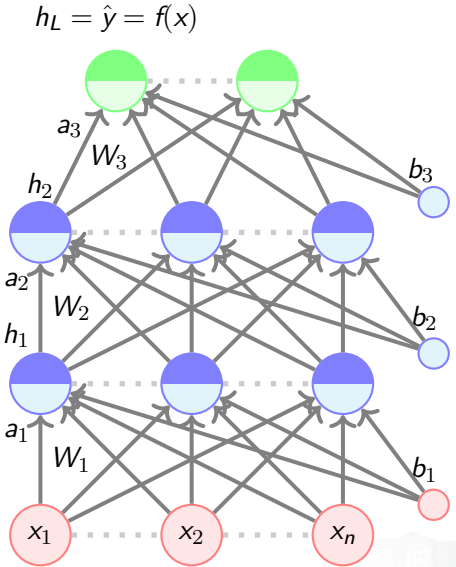


- 注意：y 是一个概率分布
- 因此我们也需要确保 $\hat{y}$ 是一个概率分布
- 选择什么样的输出激活函数 'O' 才能确保输出是一个概率分布?

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

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

$O(a_L)_j$ 是 $\hat{y}$ 的第 j^{th} 个元素，$a_{L,j}$ 是向量 a_L 的第 j^{th} 个元素.



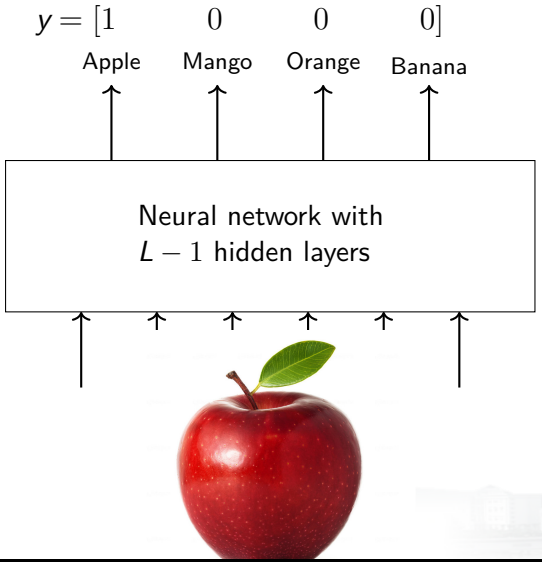
- 注意：y 是一个概率分布
- 因此我们也需要确保 $\hat{y}$ 是一个概率分布
- 选择什么样的输出激活函数 'O' 才能确保输出是一个概率分布?

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

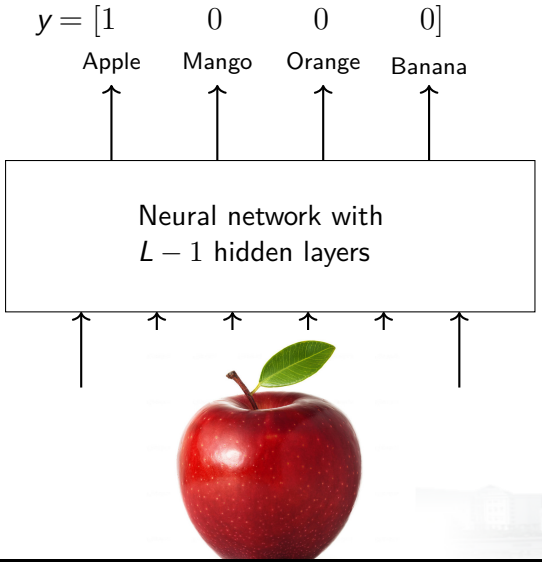
$$\hat{y}_j = O(a_L)_j = \frac{e^{a_{L,j}}}{\sum_{i=1}^{k} e^{a_{L,i}}}$$

$O(a_L)_j$ 是 $\hat{y}$ 的第 j^{th} 个元素，$a_{L,j}$ 是向量 a_L 的第 j^{th} 个元素.

- 这个输出函数称作为 *softmax* 函数

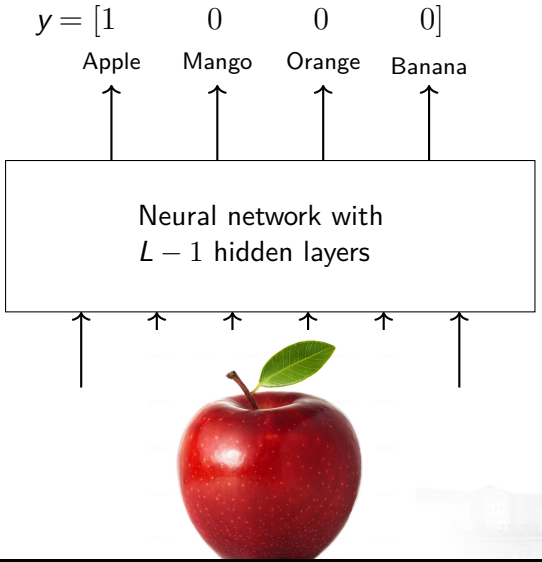


- 现在，我们确保 y 和 $\hat{y}$ 都是概率分布。该如何设计损失函数来刻画两个概率分布之间的差异？



- 现在，我们确保 y 和 $\hat{y}$ 都是概率分布。该如何设计损失函数来刻画两个概率分布之间的差异？
- 交叉熵（Cross-entropy）

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



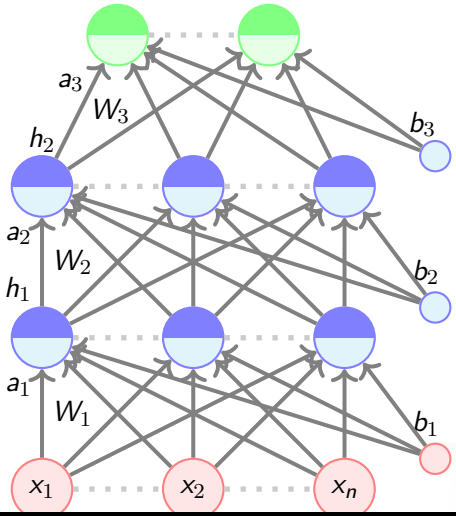
- 现在，我们确保 y 和 $\hat{y}$ 都是概率分布。该如何设计损失函数来刻画两个概率分布之间的差异？
- 交叉熵（Cross-entropy）

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

- 注意：

$$y_c = 1 \quad \text{if } c = \ell \text{ (the true class label)}$$
$$= 0 \quad \text{otherwise}$$
$$\therefore \ \mathscr{L}(\theta) = -\log \hat{y}_\ell$$

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

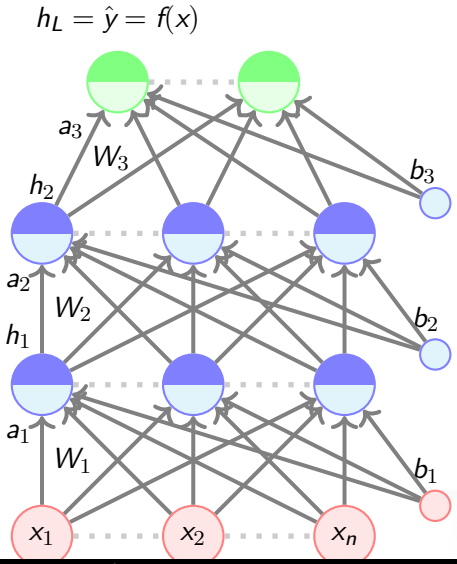


- 因此, 对于分类问题 (从 K 个类中选择一个), 使用如下目标函数

$$\underset{\theta}{\text{minimize}} \quad \mathscr{L}(\theta) = -\log \hat{y}_\ell$$

or

$$\underset{\theta}{\text{maximize}} \quad -\mathscr{L}(\theta) = \log \hat{y}_\ell$$

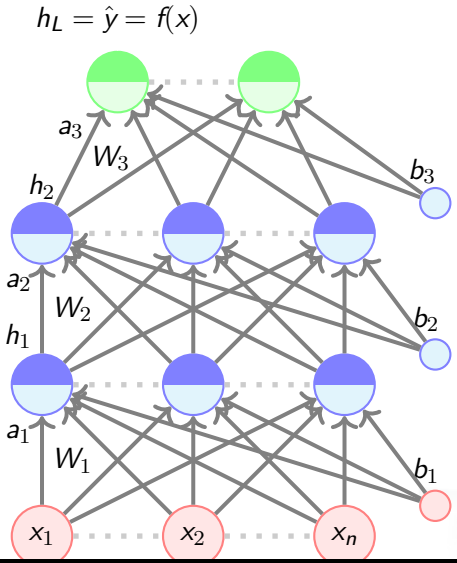


- 因此, 对于分类问题 (从 K 个类中选择一个), 使用如下目标函数

$$\underset{\theta}{\text{minimize}} \quad \mathscr{L}(\theta) = -\log \hat{y}_\ell$$

$$\text{or} \quad \underset{\theta}{\text{maximize}} \quad -\mathscr{L}(\theta) = \log \hat{y}_\ell$$

- 有个问题
  $\hat{y}_\ell$ 是参数 $\theta = [W_1, W_2, ., W_L, b_1, b_2, ., b_L]$ 的函数吗?



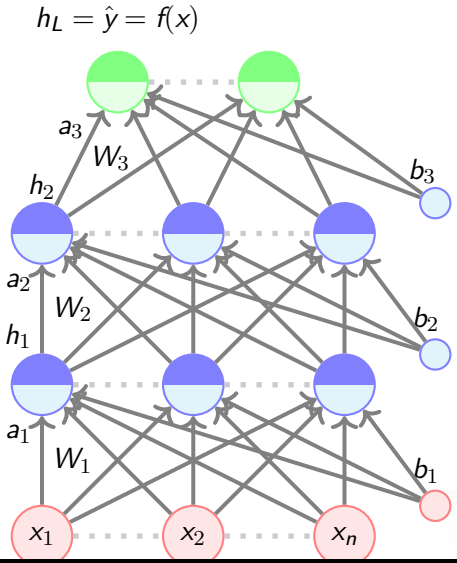
- 因此, 对于分类问题 (从 K 个类中选择一个), 使用如下目标函数

$$\underset{\theta}{\text{minimize}} \quad \mathscr{L}(\theta) = -\log \hat{y}_\ell$$

$$\text{or} \quad \underset{\theta}{\text{maximize}} \quad -\mathscr{L}(\theta) = \log \hat{y}_\ell$$

- 有个问题
  $\hat{y}_\ell$ 是参数 $\theta = [W_1, W_2, ., W_L, b_1, b_2, ., b_L]$ 的函数吗?
- 是的

$$\hat{y}_\ell = [O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)]_\ell$$



- 因此, 对于分类问题 (从 K 个类中选择一个), 使用如下目标函数

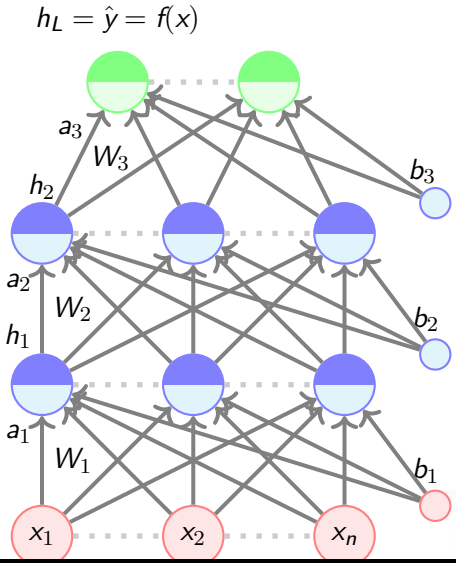
$$\underset{\theta}{\text{minimize}} \quad \mathscr{L}(\theta) = -\log \hat{y}_\ell$$

$$\text{or} \quad \underset{\theta}{\text{maximize}} \quad -\mathscr{L}(\theta) = \log \hat{y}_\ell$$

- 有个问题
  $\hat{y}_\ell$ 是参数 $\theta = [W_1, W_2, ., W_L, b_1, b_2, ., b_L]$ 的函数吗?
- 是的

$$\hat{y}_\ell = [O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)]_\ell$$

- 它是 x 属于第 ℓ^{th} 类的概率.



- 因此, 对于分类问题 (从 K 个类中选择一个), 使用如下目标函数

$$\underset{\theta}{\text{minimize}} \quad \mathscr{L}(\theta) = -\log \hat{y}_\ell$$

$$\text{or} \quad \underset{\theta}{\text{maximize}} \quad -\mathscr{L}(\theta) = \log \hat{y}_\ell$$

- 有个问题
  $\hat{y}_\ell$ 是参数 $\theta = [W_1, W_2, ., W_L, b_1, b_2, ., b_L]$ 的函数吗?
- 是的

$$\hat{y}_\ell = [O(W_3 g(W_2 g(W_1 x + b_1) + b_2) + b_3)]_\ell$$

- 它是 x 属于第 ℓ^{th} 类的概率.
- $\log \hat{y}_\ell$ 被称为数据的 *log-likelihood*

	Outputs	
	Real Values	Probabilities
Output Activation		
Loss Function		

	Outputs	
	Real Values	Probabilities
Output Activation	Linear	
Loss Function		

	Outputs	
	Real Values	Probabilities
Output Activation	Linear	Softmax
Loss Function		

	Outputs	
	Real Values	Probabilities
Output Activation	Linear	Softmax
Loss Function	Squared Error	

	Outputs	
	Real Values	Probabilities
Output Activation	Linear	Softmax
Loss Function	Squared Error	Cross Entropy

	Outputs	
	Real Values	Probabilities
Output Activation	Linear	Softmax
Loss Function	Squared Error	Cross Entropy

- 针对不同的问题，有不同的损失函数，但刚才见到的两个损失函数经常用到

	Outputs	
	Real Values	Probabilities
Output Activation	Linear	Softmax
Loss Function	Squared Error	Cross Entropy

- 针对不同的问题，有不同的损失函数，但刚才见到的两个损失函数经常用到
- 在本讲的后面内容中，将考虑激活函数是 softmax 函数，损失函数是交叉熵的情况

反向传播 (Intuition)

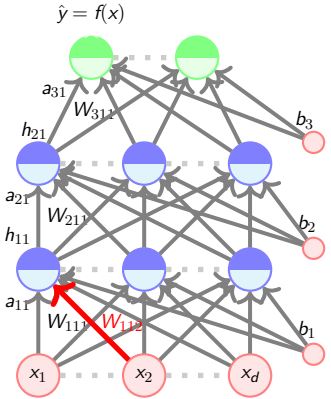
需要回答两个问题

- 如何选择损失函数 $\mathscr{L}(\theta)$?
- 如何计算 $\nabla\theta$?

需要回答两个问题

- 如何选择损失函数 $\mathscr{L}(\theta)$？
- 如何计算 $\nabla\theta$？

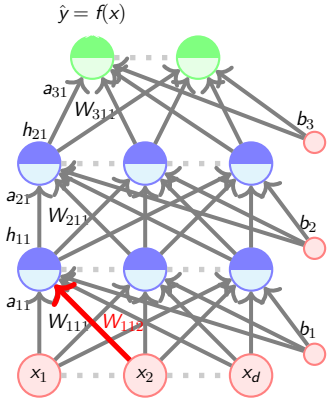
- 考虑这一个权重(W_{112}).



Algorithm 9: gradient descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize θ_0;
while $t{+}{+} < max\_iterations$ **do**
  $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;
end

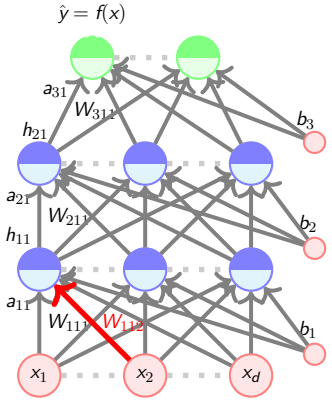
- 考虑这一个权重(W_{112}).
- 使用 SGD 来学习这个权重，需要计算 $\frac{\partial \mathcal{L}(\theta)}{\partial W_{112}}$.



Algorithm 10: gradient descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize θ_0;
while $t{+}{+} < max\_iterations$ **do**
  $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;
end

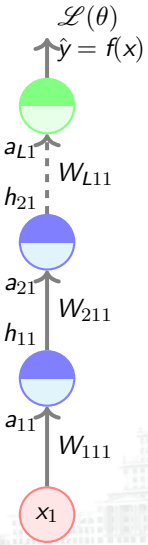
- 考虑这一个权重(W_{112}).
- 使用 SGD 来学习这个权重，需要计算 $\frac{\partial \mathcal{L}(\theta)}{\partial W_{112}}$.
- 如何计算？



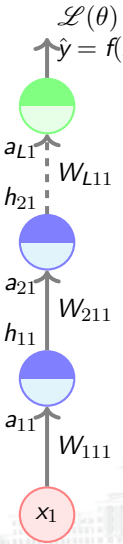
Algorithm 11: gradient descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize θ_0;
while $t{+}{+} < max\_iterations$ **do**
  $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;
end

- 首先，将问题简化，考虑一个『深』但『瘦』的网络

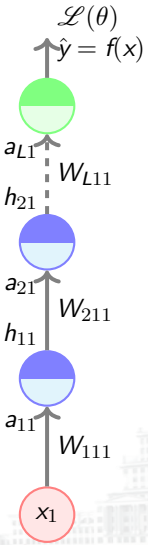


- 首先，将问题简化，考虑一个『深』但『瘦』的网络
- 这种情况，很容易通过链式法则，求得



- 首先，将问题简化，考虑一个『深』但『瘦』的网络
- 这种情况，很容易通过链式法则，求得

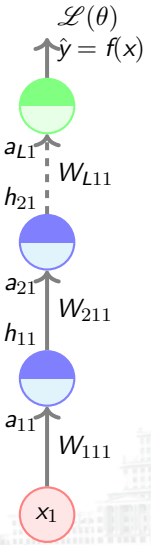
$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial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



- 首先，将问题简化，考虑一个『深』但『瘦』的网络
- 这种情况，很容易通过链式法则，求得

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

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{11}} \frac{\partial h_{11}}{\partial W_{111}}$$ (just compressing the chain rule)



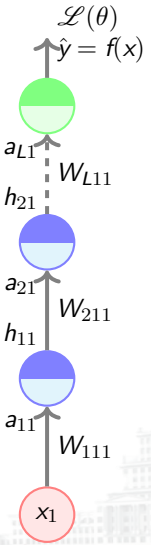
- 首先，将问题简化，考虑一个『深』但『瘦』的网络
- 这种情况，很容易通过链式法则，求得

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

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

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

(just compressing the chain rule)



- 首先，将问题简化，考虑一个『深』但『瘦』的网络
- 这种情况，很容易通过链式法则，求得

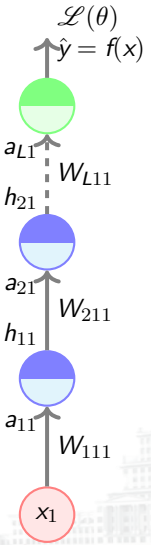
$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial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

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

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

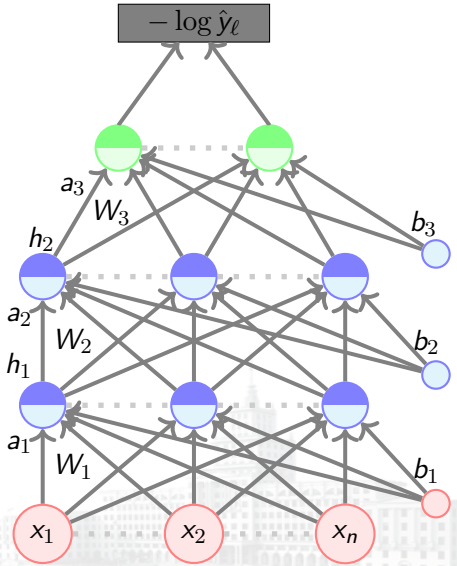
$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{L11}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \frac{\partial a_{L1}}{\partial W_{L11}}$$

(just compressing the chain rule)

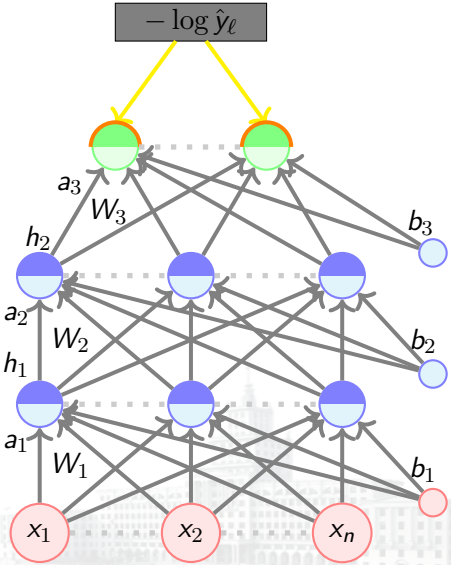


在进入具体的数学推导之前，先看看反向传播的直观解释

- 我们在输出层得到一个损失，我们希望找出谁应该对这个损失负责？

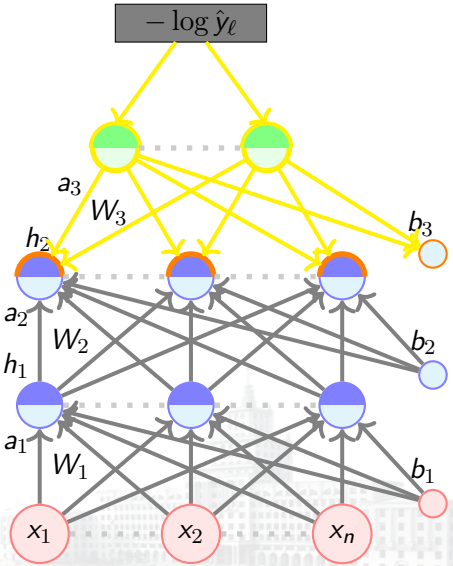


- 我们在输出层得到一个损失，我们希望找出谁应该对这个损失负责？
- 因此，我们对最后一层说“你没有产生我想要的输出，你需要做的更好才行”.

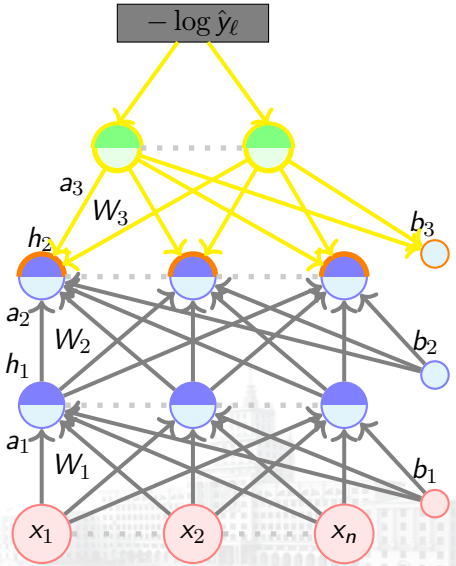


- 我们在输出层得到一个损失，我们希望找出谁应该对这个损失负责？
- 因此，我们对最后一层说“你没有产生我想要的输出，你需要做的更好才行”.
- 最后一层说“我已经尽了我能尽的责任了，请你理解，我只能做得想隐含层和我下面的权重好样好,”. After all ...

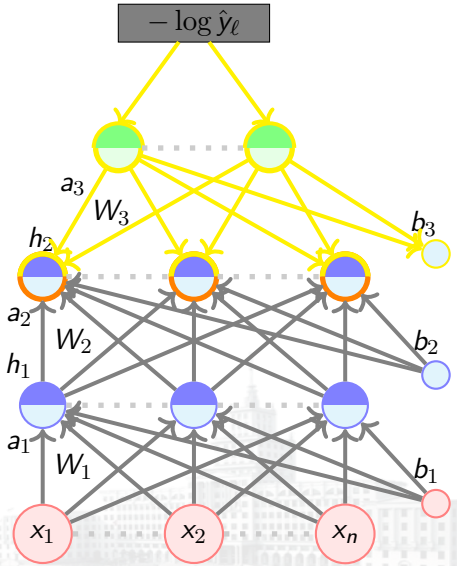
$$f(x) = \hat{y} = O(W_L h_{L-1} + b_L)$$



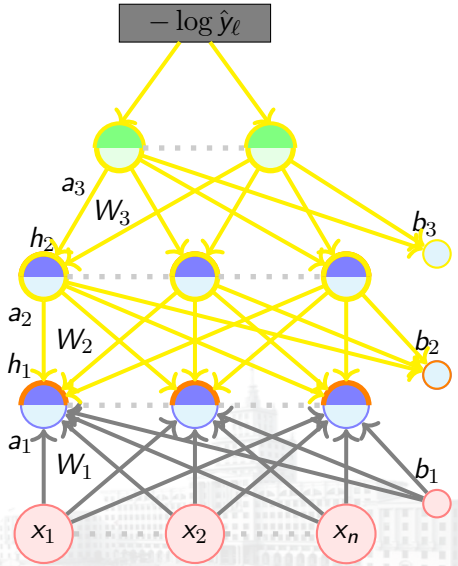
- 因此，我们问 W_L, b_L 和 h_L “你们怎么了?”



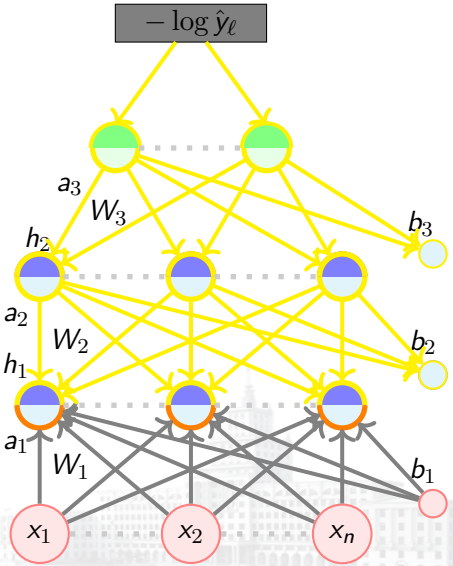
- 因此，我们问 W_L, b_L 和 h_L “你们怎么了?”
- W_L 和 b_L 已经尽了所有的责任，但 h_L 说 “我只能做的像预激活层那样好”



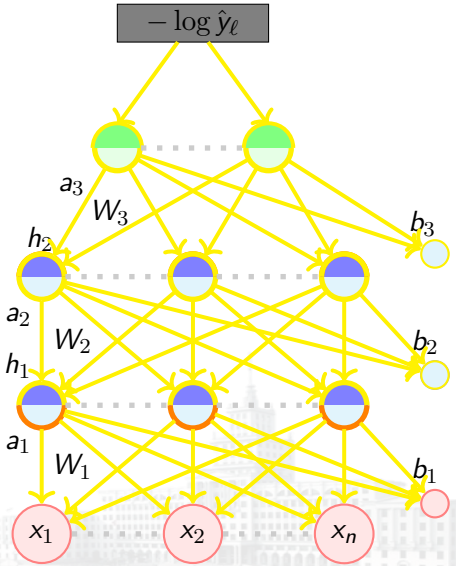
- 因此，我们问 W_L, b_L 和 h_L “你们怎么了?”
- W_L 和 b_L 已经尽了所有的责任，但 h_L 说 “我只能做的像预激活层那样好”
- 预激活层说 “我只能做的像隐含层和我下面的权重那样好”



- 因此，我们问 W_L, b_L 和 h_L “你们怎么了?”
- W_L 和 b_L 已经尽了所有的责任，但 h_L 说 “我只能做的像预激活层那样好”
- 预激活层说 “我只能做的像隐含层和我下面的权重那样好”
- 这个过程继续，我们意识到责任在于所有的权重和偏置（也就是模型的参数）

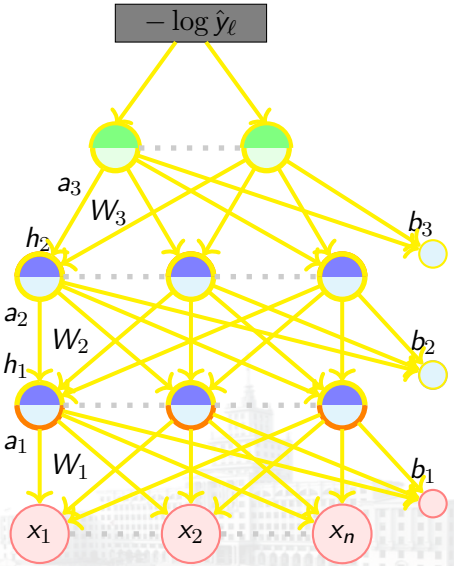


- 因此，我们问 W_L, b_L 和 h_L “你们怎么了?”
- W_L 和 b_L 已经尽了所有的责任，但 h_L 说 “我只能做的像预激活层那样好”
- 预激活层说 “我只能做的像隐含层和我下面的权重那样好”
- 这个过程继续，我们意识到责任在于所有的权重和偏置（也就是模型的参数）
- 不同于直接和它们对话，通过隐含层和输出层来和它们对话会更容易（链式法则也允许我们这样做）



- 因此，我们问 W_L, b_L 和 h_L “你们怎么了?”
- W_L 和 b_L 已经尽了所有的责任，但 h_L 说 “我只能做的像预激活层那样好”
- 预激活层说 “我只能做的像隐含层和我下面的权重那样好”
- 这个过程继续，我们意识到责任在于所有的权重和偏置（也就是模型的参数）
- 不同于直接和它们对话，通过隐含层和输出层来和它们对话会更容易（链式法则也允许我们这样做）

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



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

需要计算的梯度:

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

需要计算的梯度:

- Gradient w.r.t. output units

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

需要计算的梯度:

- Gradient w.r.t. output units
- Gradient w.r.t. hidden units

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

需要计算的梯度:

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

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

需要计算的梯度:

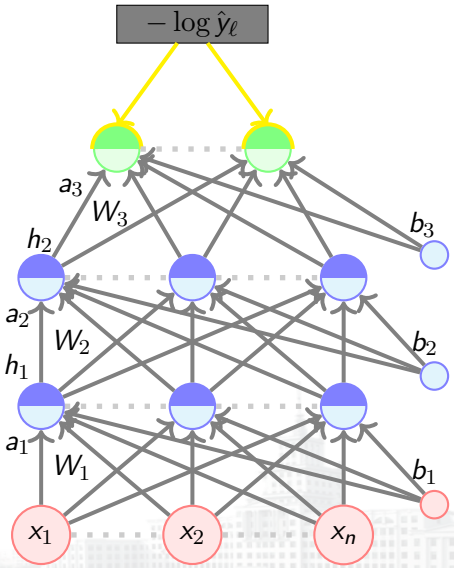
- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights and biases

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

- 我们关注 交叉熵损失函数和 *Softmax* 输出函数

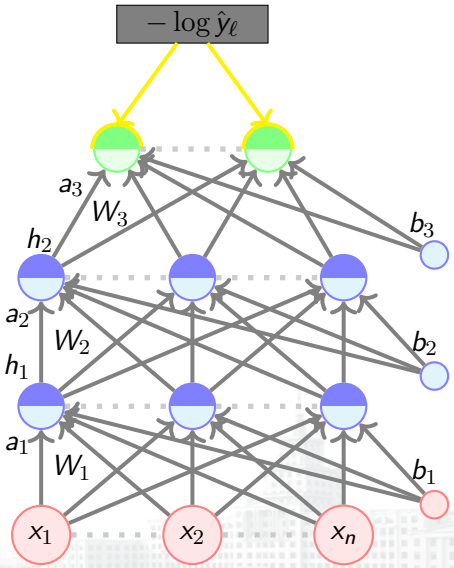
反向传播: 计算损失函数对输出单元的梯度（Gradients w.r.t. the Output Units)

首先，考虑对 i-th 输出的偏导数



首先，考虑对 i-th 输出的偏导数

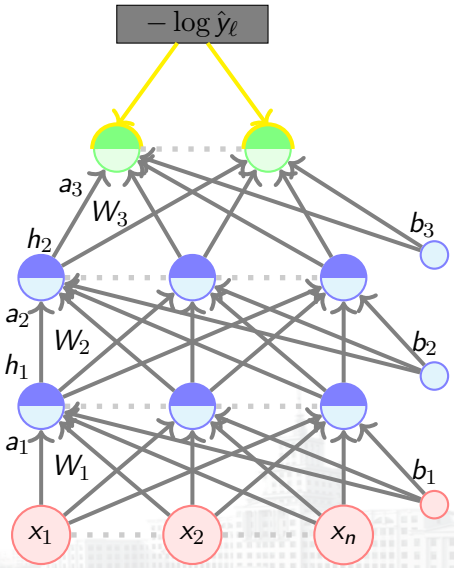
$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$



首先，考虑对 i-th 输出的偏导数

$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$

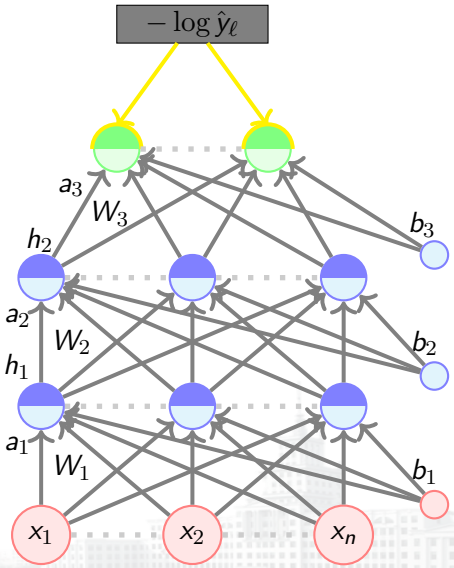
$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) =$$



首先，考虑对 i-th 输出的偏导数

$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$

$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = \frac{\partial}{\partial \hat{y}_i}\left(-\log \hat{y}_\ell\right)$$

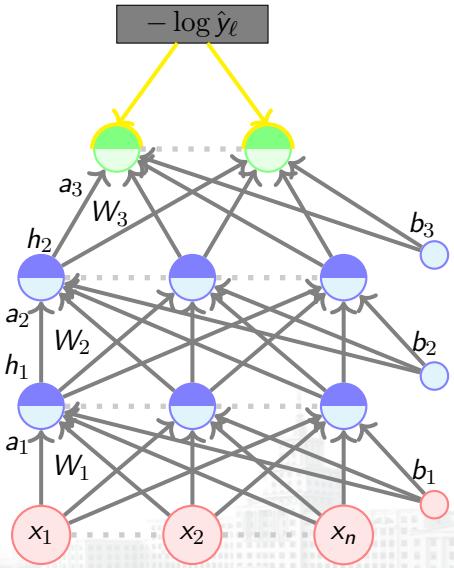


首先，考虑对 i-th 输出的偏导数

$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$

$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = \frac{\partial}{\partial \hat{y}_i}\left(-\log \hat{y}_\ell\right)$$

$$= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell$$



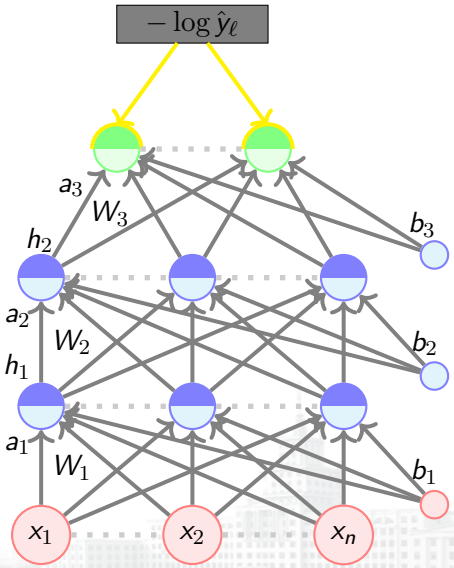
首先，考虑对 i-th 输出的偏导数

$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$

$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = \frac{\partial}{\partial \hat{y}_i}\left(-\log \hat{y}_\ell\right)$$

$$= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell$$

$$= 0 \qquad \textit{otherwise}$$



首先，考虑对 i-th 输出的偏导数

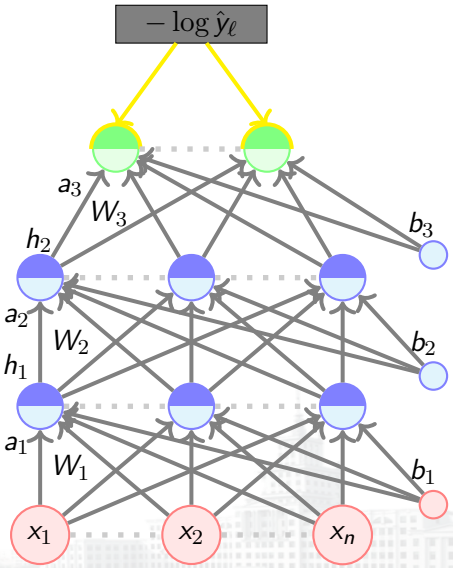
$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$

$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = \frac{\partial}{\partial \hat{y}_i}\left(-\log \hat{y}_\ell\right)$$

$$= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell$$

$$= 0 \qquad \textit{otherwise}$$

更紧凑地，



首先，考虑对 i-th 输出的偏导数

$$\mathscr{L}(\theta) = -\log \hat{y}_\ell \quad (\ell = \text{true class label})$$

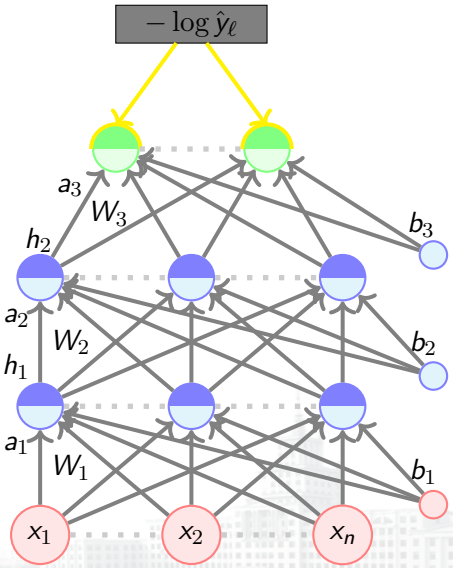
$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = \frac{\partial}{\partial \hat{y}_i}\left(-\log \hat{y}_\ell\right)$$

$$= -\frac{1}{\hat{y}_\ell} \quad \text{if } i = \ell$$

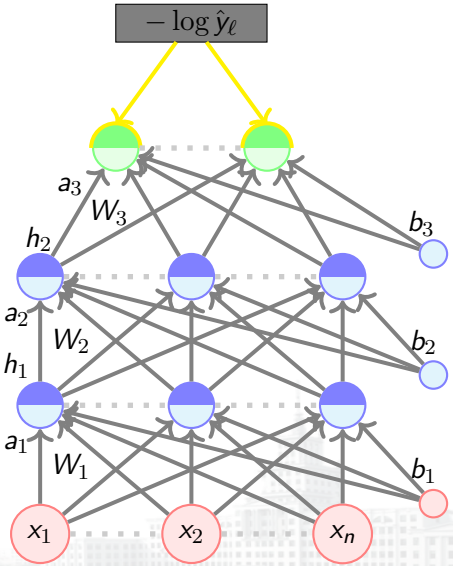
$$= 0 \qquad \textit{otherwise}$$

更紧凑地，

$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(i=\ell)}}{\hat{y}_\ell}$$

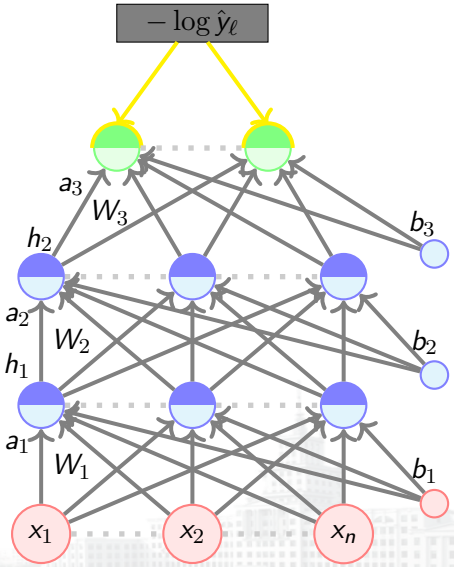


$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

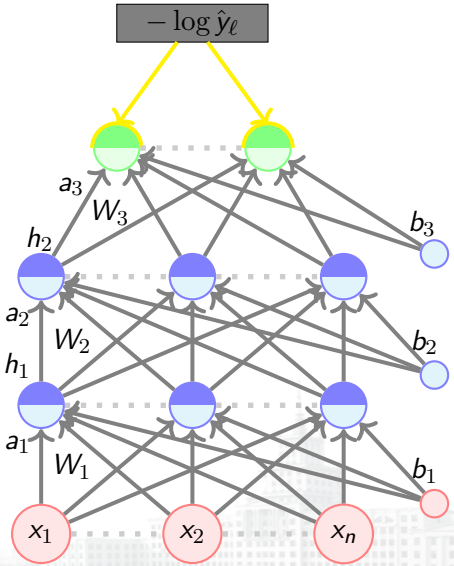
对向量 $\hat{y}$ 的梯度为



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

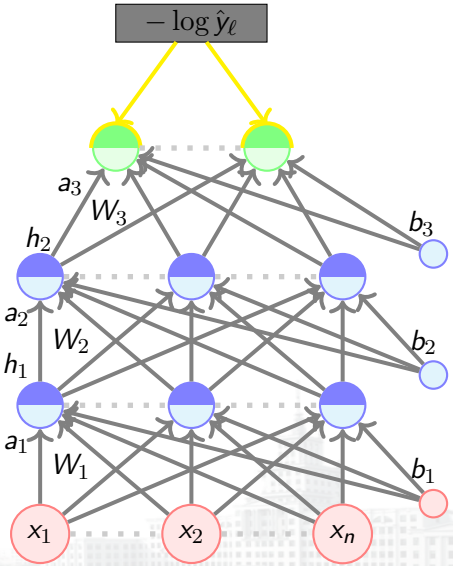
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} & \\ & \\ & \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

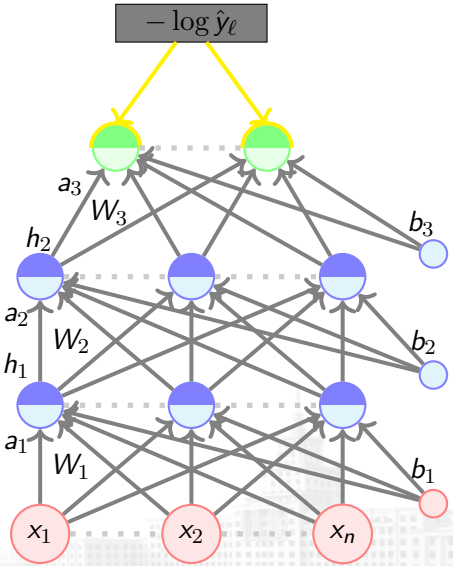
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \\ \\ \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

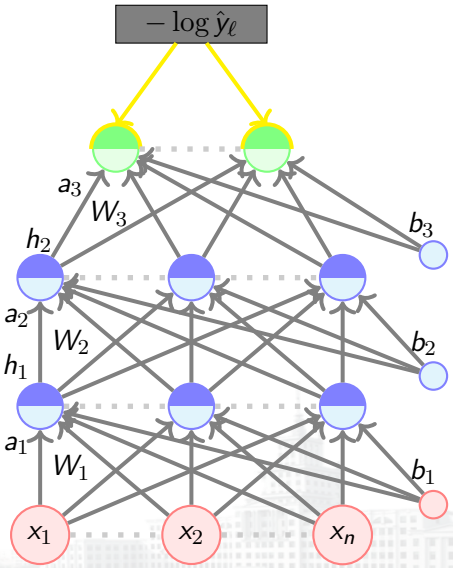
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

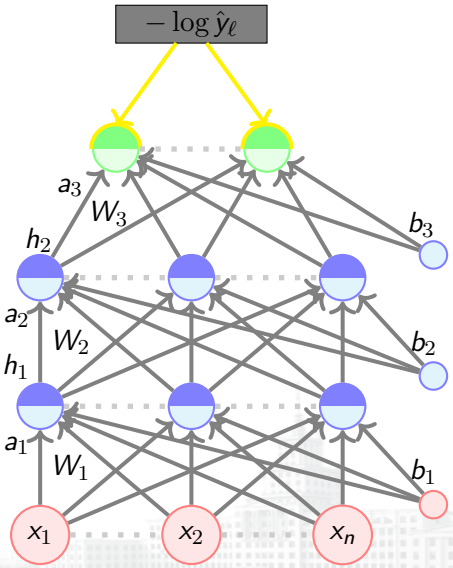
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

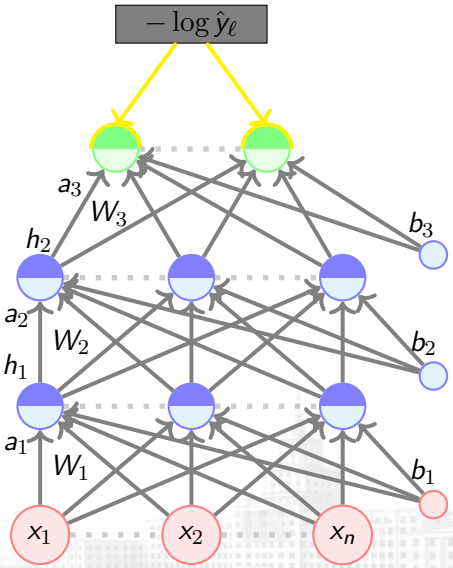
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

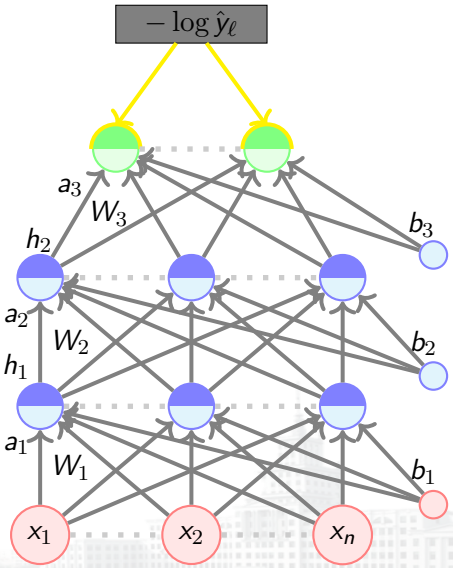
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}\begin{bmatrix} \\ \\ \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

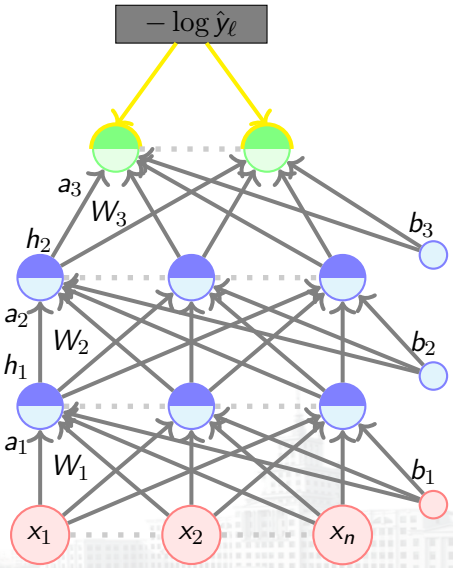
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \\ \\ \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

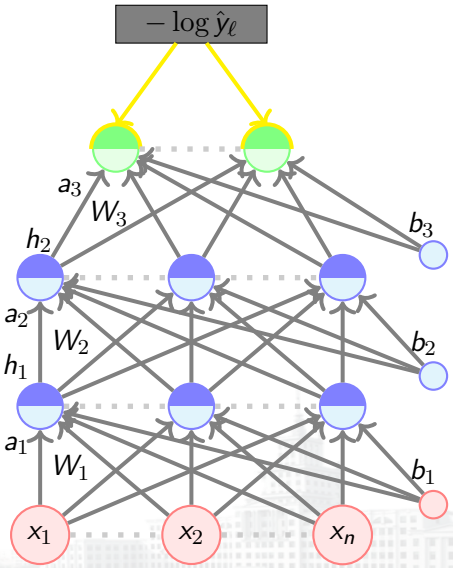
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \\ \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

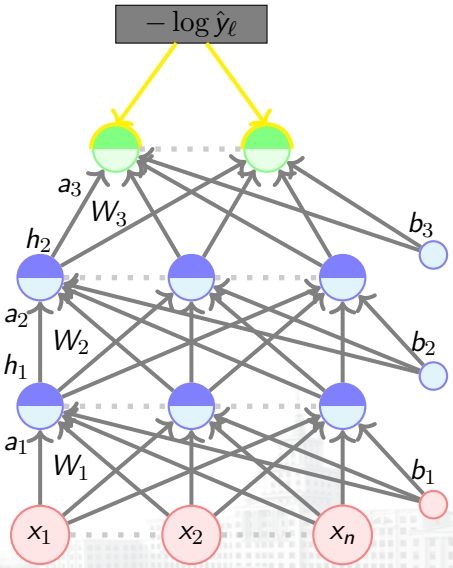
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}\begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \end{bmatrix}$$



$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

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

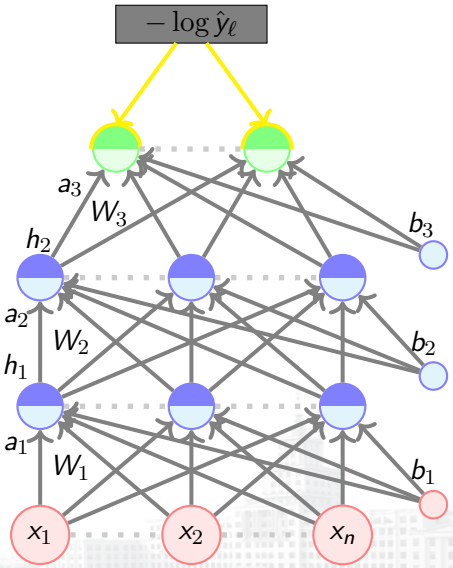


$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

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

$$= -\frac{1}{\hat{y}_\ell}\mathbf{e}_\ell$$

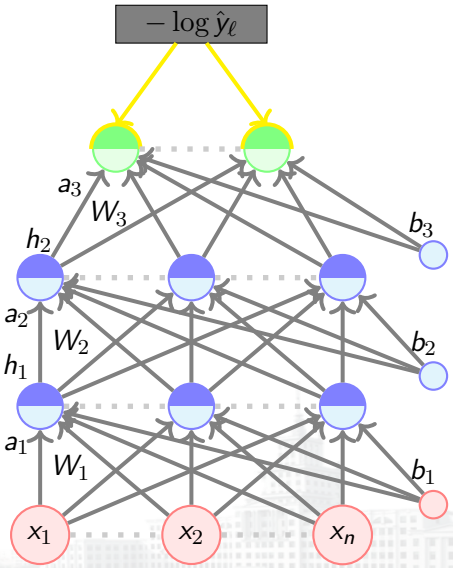


$$\frac{\partial}{\partial \hat{y}_i}\left(\mathscr{L}(\theta)\right) = -\frac{\mathbb{1}_{(\ell=i)}}{\hat{y}_\ell}$$

对向量 $\hat{y}$ 的梯度为

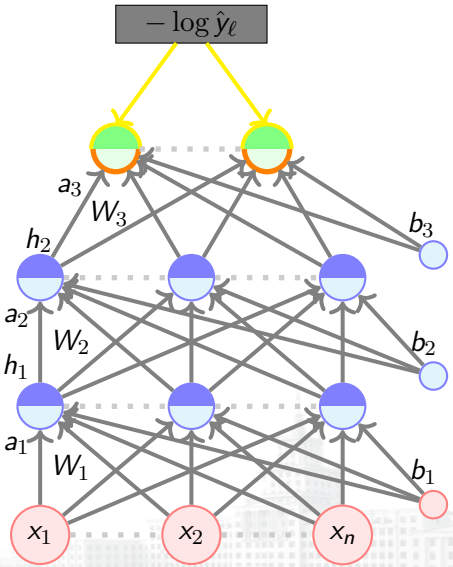
$$\nabla_{\hat{\mathbf{y}}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}_k} \end{bmatrix} = -\frac{1}{\hat{y}_\ell}\begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \\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

其中 $\mathbf{e}_\ell$ 是一个 k 维向量，它的第 ℓ 个元素是 1 其他元素是 0.



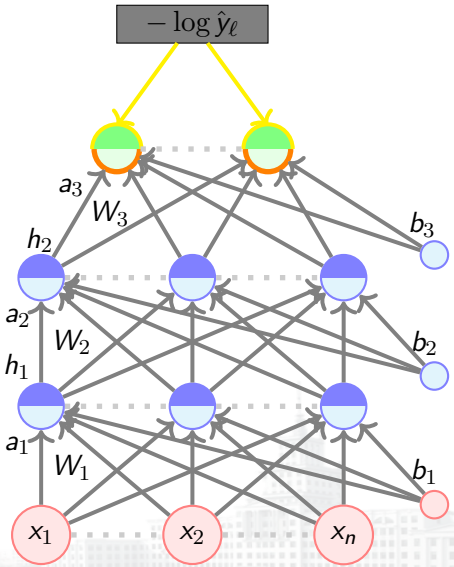
实际上，我们想要计算的是

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}}$$



实际上，我们想要计算的是

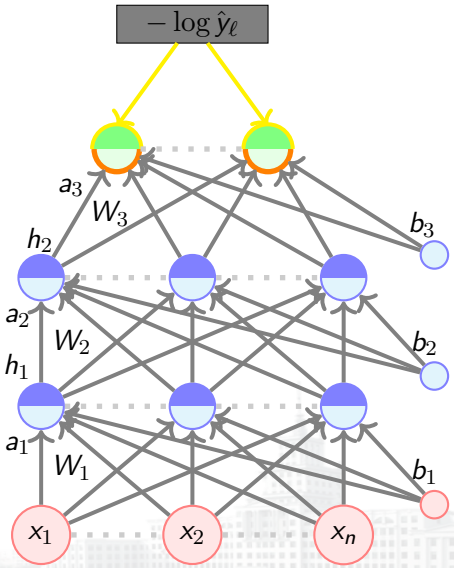
$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}}$$
$$= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}$$



实际上，我们想要计算的是

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}}$$
$$= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}$$

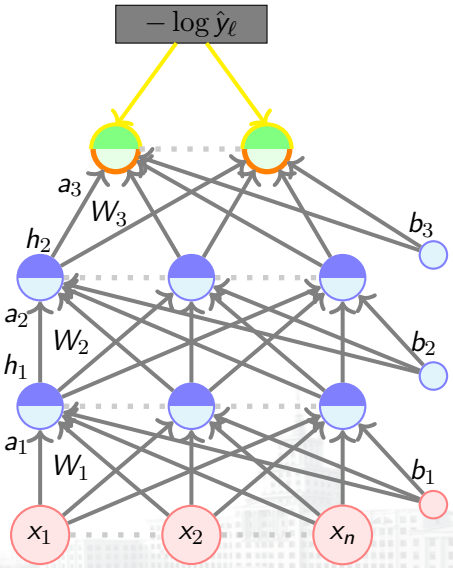
$\hat{y}_\ell$ 依赖于 a_{Li} 吗？



实际上，我们想要计算的是

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}}$$
$$= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}$$

$\hat{y}_\ell$ 依赖于 a_{Li} 吗？ 实事上，是依赖的.

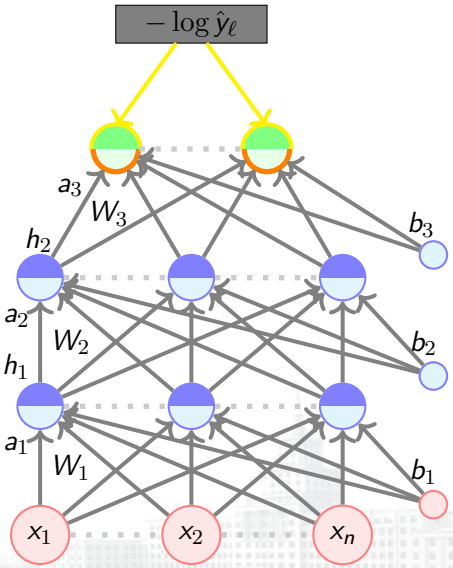


实际上，我们想要计算的是

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}}$$
$$= \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell} \frac{\partial \hat{y}_\ell}{\partial a_{Li}}$$

$\hat{y}_\ell$ 依赖于 a_{Li} 吗？ 实事上，是依赖的.

$$\hat{y}_\ell = \frac{exp(a_{L\ell})}{\sum_i exp(a_{Li})}$$



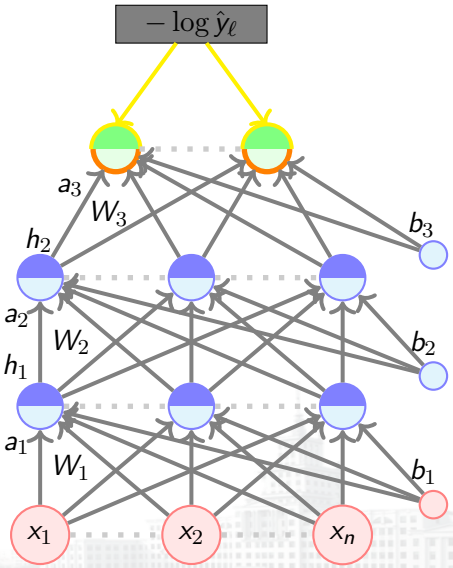
实际上，我们想要计算的是

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_\ell)}{\partial \hat{y}_\ell}\frac{\partial \hat{y}_\ell}{\partial a_{Li}}$$

$\hat{y}_\ell$ 依赖于 a_{Li} 吗？ 实事上，是依赖的.

$$\hat{y}_\ell = \frac{exp(a_{L\ell})}{\sum_i exp(a_{Li})}$$

建立 $\hat{y}_\ell$ 和 a_{Li} 的依赖关系后，下面可以计算出对 a_{Li} 的偏导数



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

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

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

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

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

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

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

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

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

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

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

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

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

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

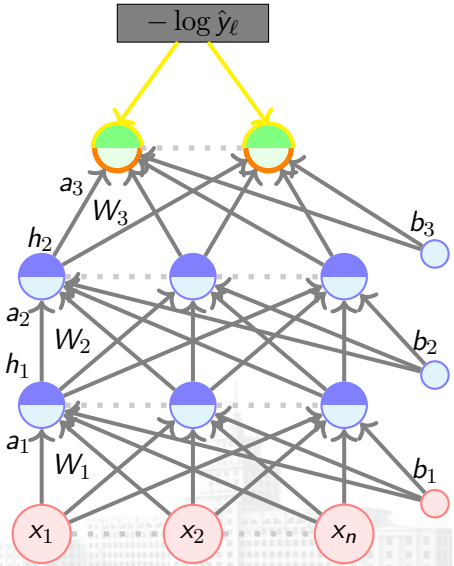
$$\frac{\partial}{\partial a_{Li}} - \log \hat{y}_\ell = \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \hat{y}_\ell$$
$$= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} softmax(\mathbf{a}_L)_\ell$$
$$= \frac{-1}{\hat{y}_\ell} \frac{\partial}{\partial a_{Li}} \frac{\exp(\mathbf{a}_L)_\ell}{\sum_{i'}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

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

现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

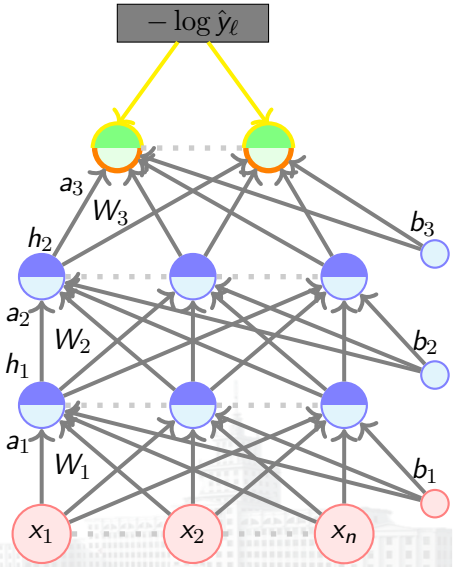


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

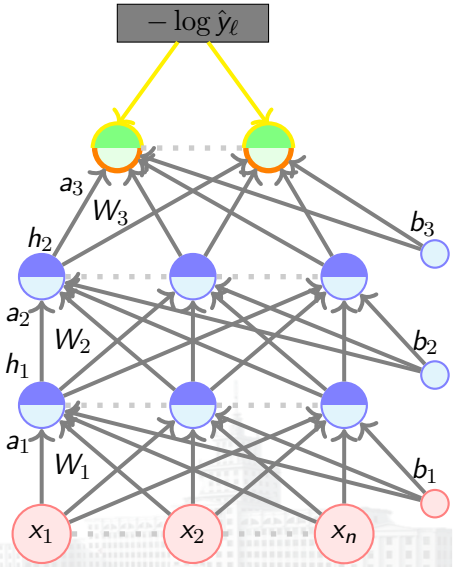


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

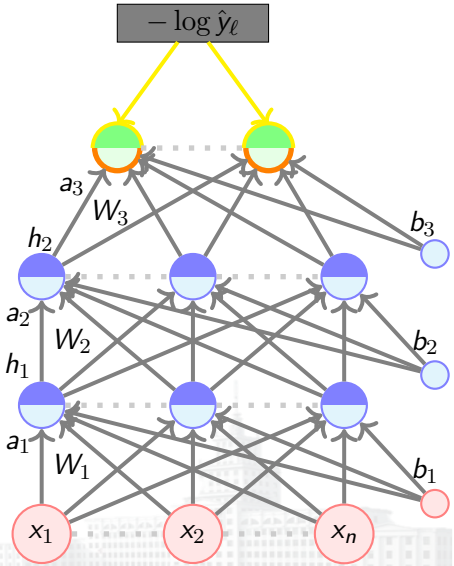


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

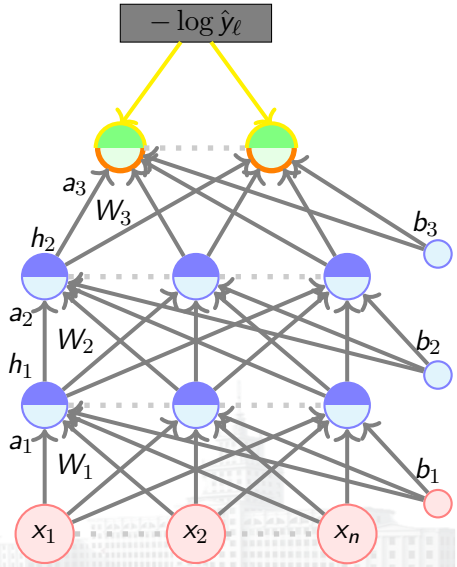


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

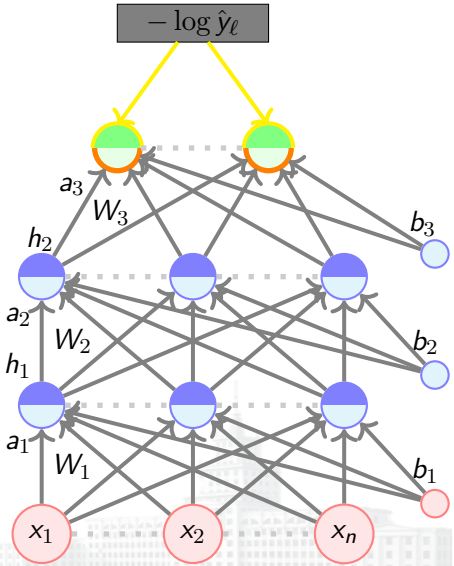


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

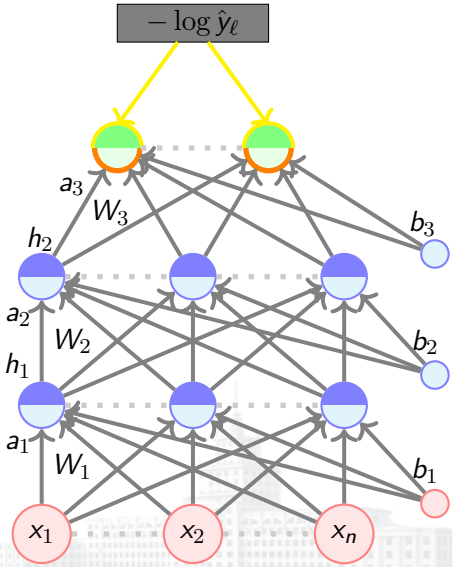


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

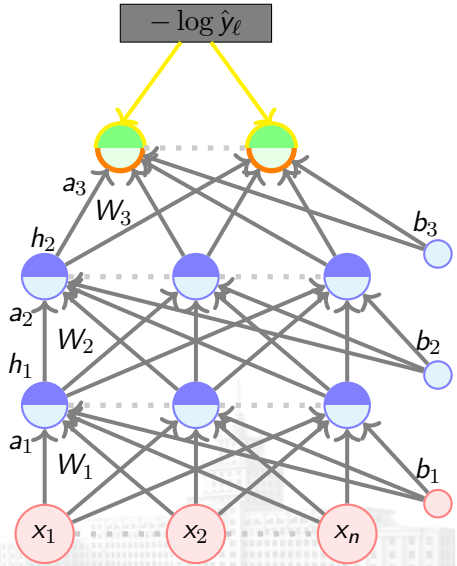


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

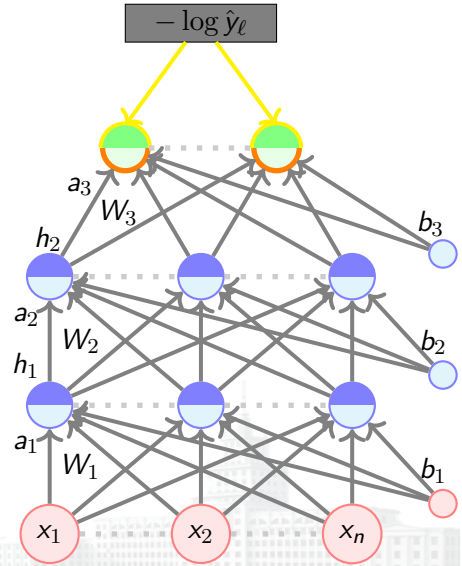


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

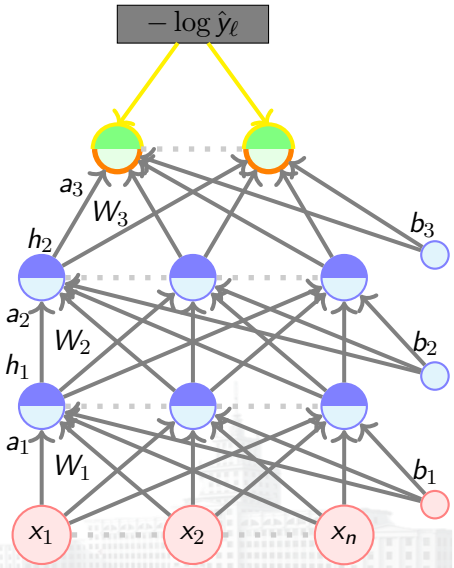


现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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



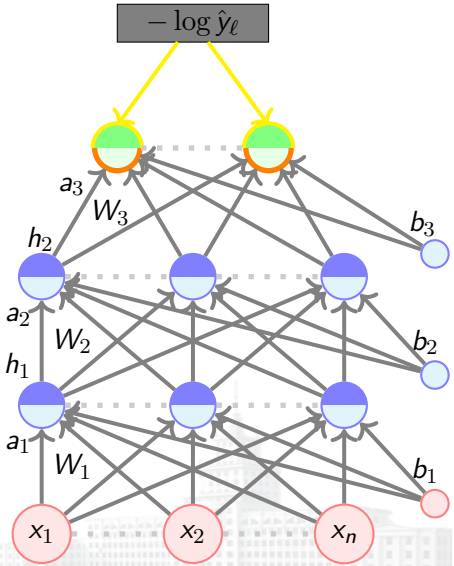
现在，已经得到对 $\mathbf{a}_L$ 的 i 个元素的偏导

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

对向量 $\mathbf{a}_L$ 的导数可以计算为

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

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



反向传播: 计算损失函数对隐含单元的梯度（Gradients w.r.t. Hidden Units)

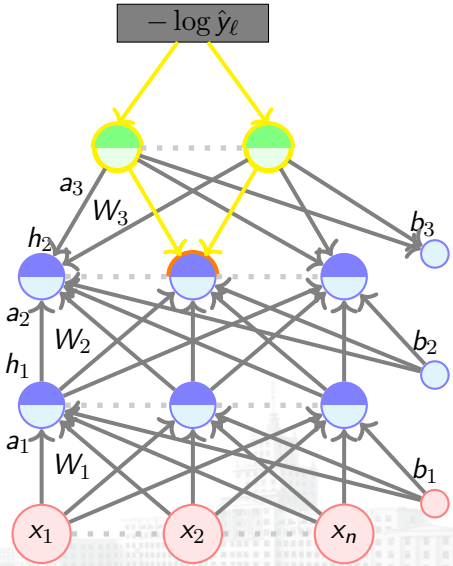
需要计算的梯度:

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

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

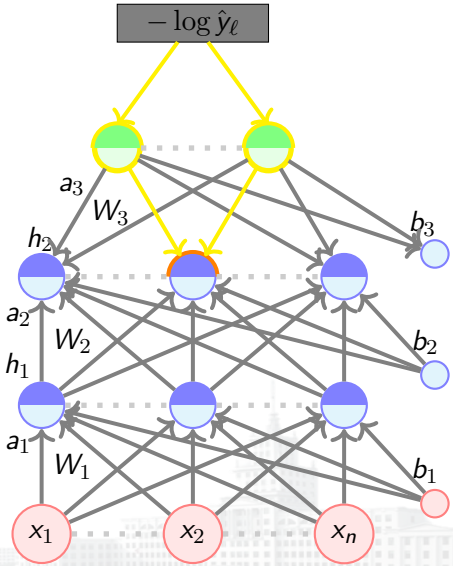
- 我们关注 交叉熵损失函数和 *Softmax* 输出函数

Chain rule along multiple paths: 如果一个函数 $p(z)$ 是中间结果 $q_i(z)$ 的函数:



Chain rule along multiple paths: 如果一个函数 $p(z)$ 是中间结果 $q_i(z)$ 的函数:

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

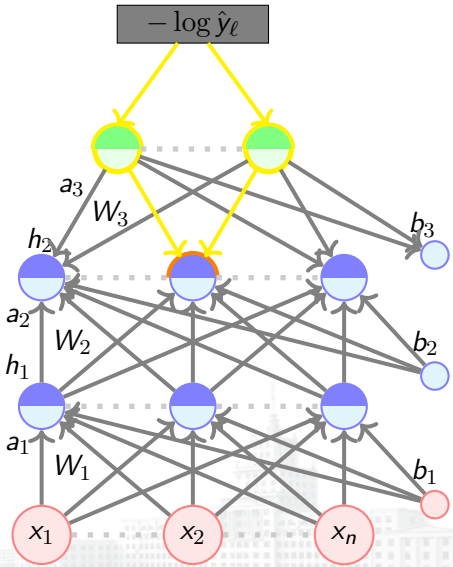


Chain rule along multiple paths: 如果一个函数 $p(z)$ 是中间结果 $q_i(z)$ 的函数:

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

在这里:

- $p(z)$ 就是损失函数 $\mathscr{L}(\theta)$

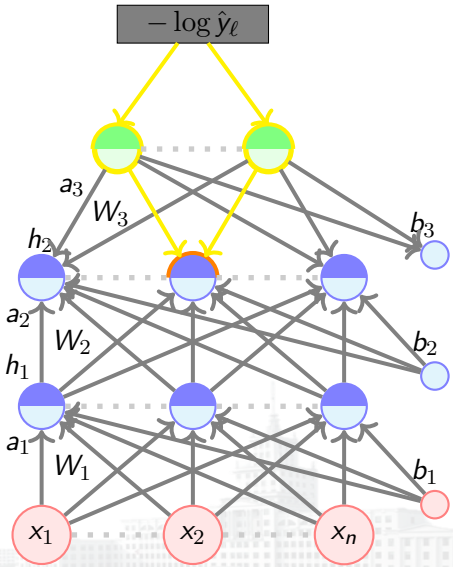


Chain rule along multiple paths: 如果一个函数 $p(z)$ 是中间结果 $q_i(z)$ 的函数:

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

在这里:

- $p(z)$ 就是损失函数 $\mathscr{L}(\theta)$
- $z = h_{ij}$

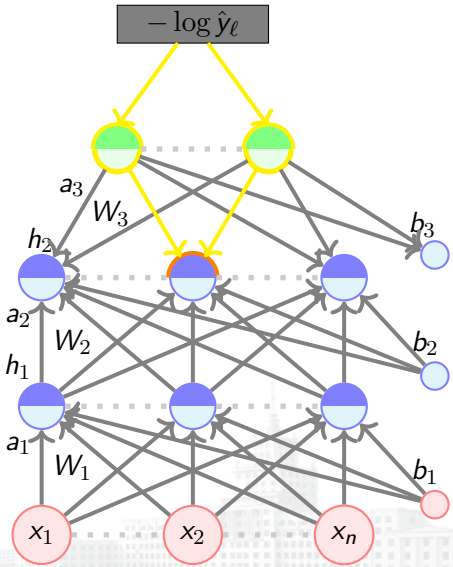


Chain rule along multiple paths: 如果一个函数 $p(z)$ 是中间结果 $q_i(z)$ 的函数:

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

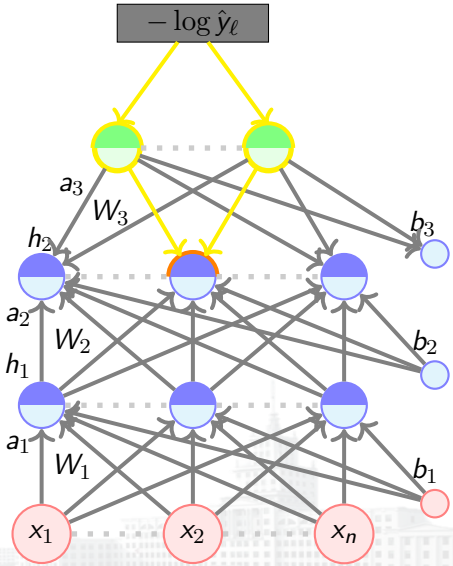
在这里:

- $p(z)$ 就是损失函数 $\mathscr{L}(\theta)$
- $z = h_{ij}$
- $q_m(z) = a_{Lm}$

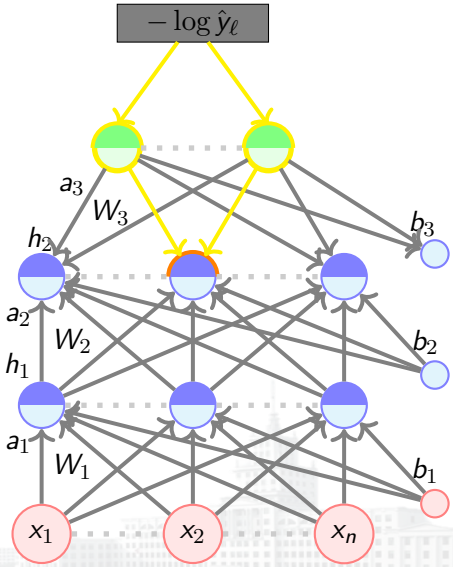


Intentionally left blank

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

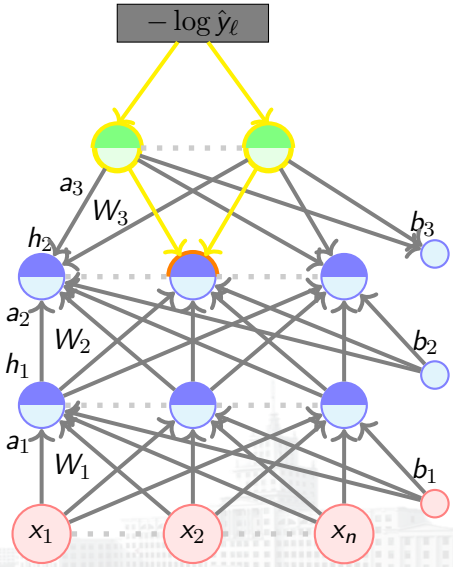


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$



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

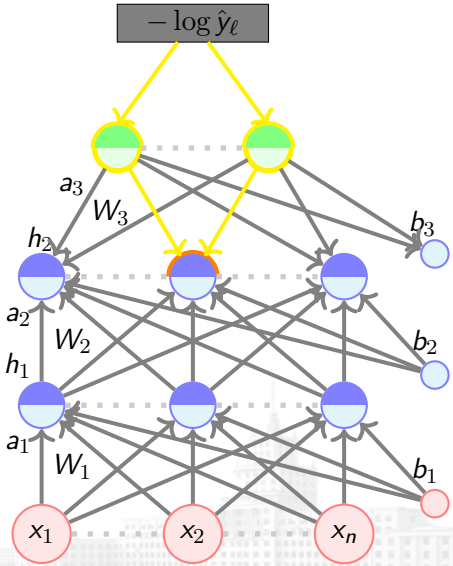
$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$



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

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

考虑如下两个向量：

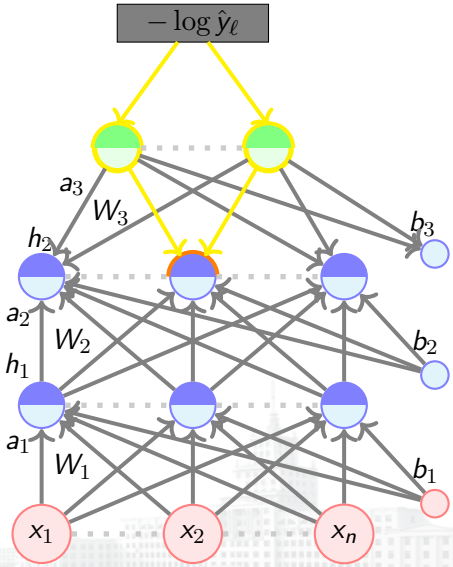


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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

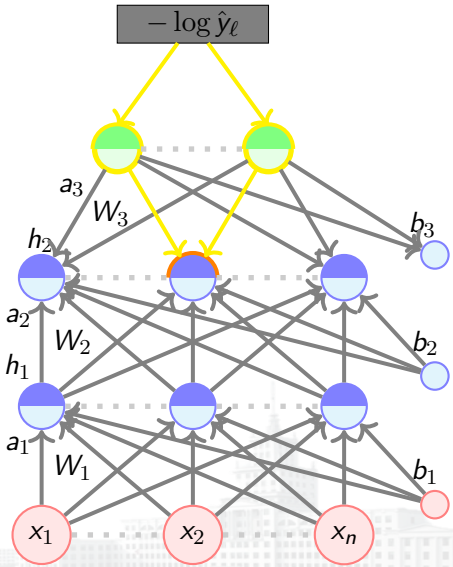


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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

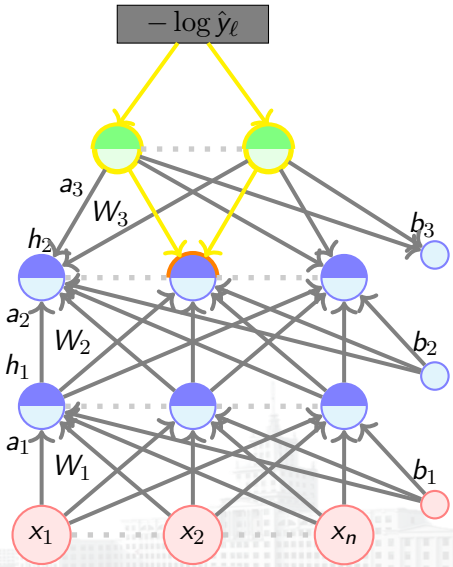


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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

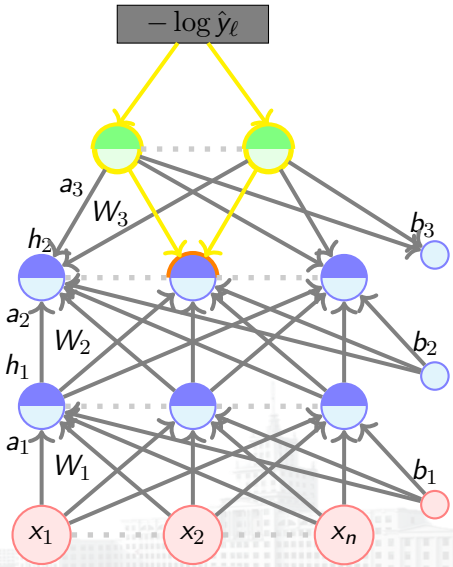


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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

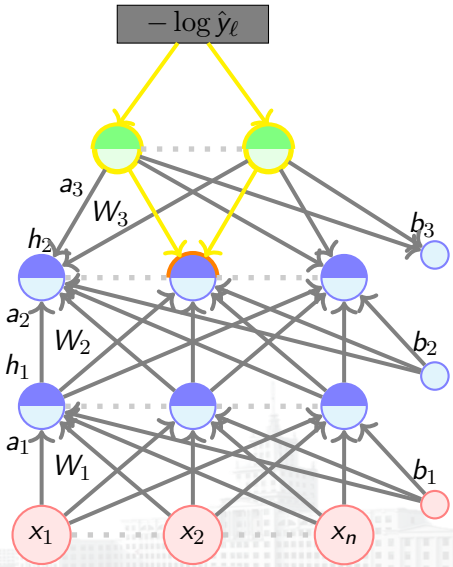


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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

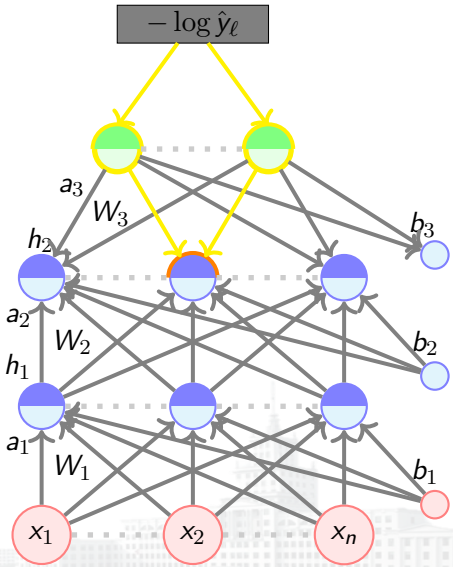


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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

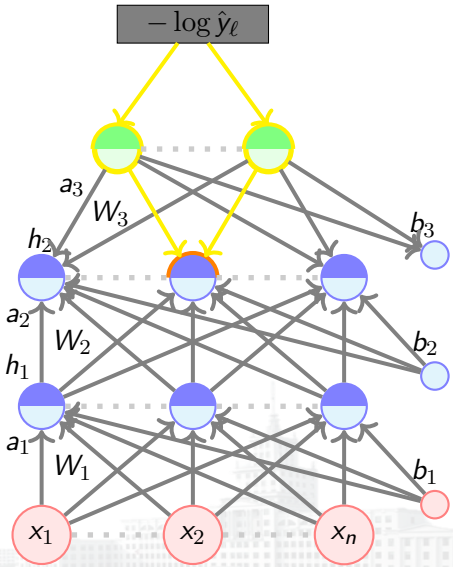


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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



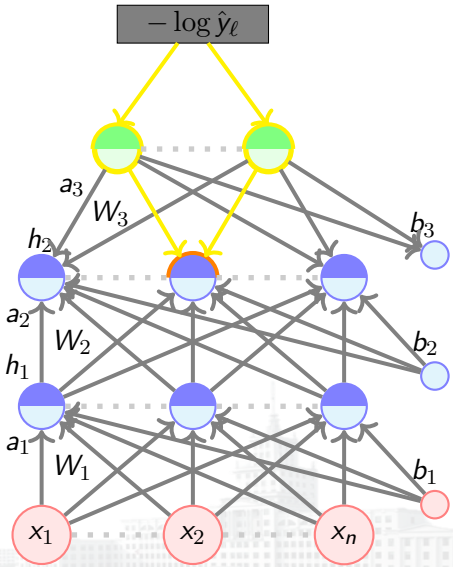
$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

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

考虑如下两个向量：

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

$W_{i+1,\cdot,j}$ 是 W_{i+1} 的第 j 列;

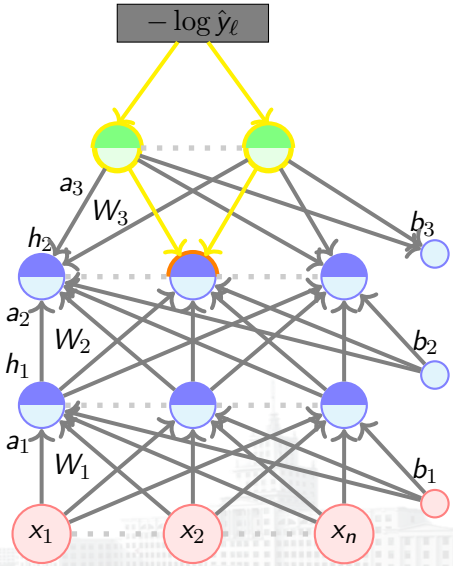


$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$
$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

考虑如下两个向量：

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

$W_{i+1,\cdot,j}$ 是 W_{i+1} 的第 j 列；下面



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

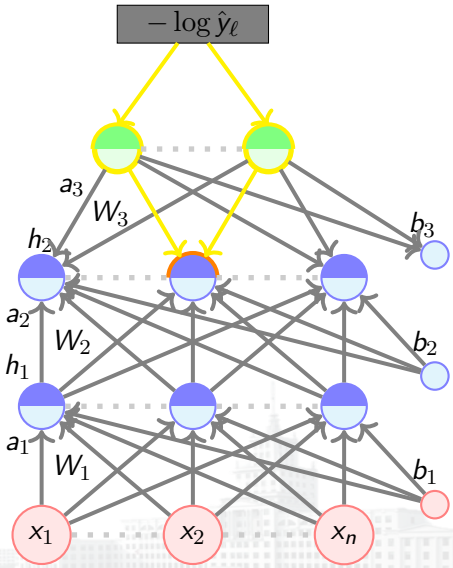
$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

考虑如下两个向量：

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

$W_{i+1,\cdot,j}$ 是 W_{i+1} 的第 j 列；下面

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



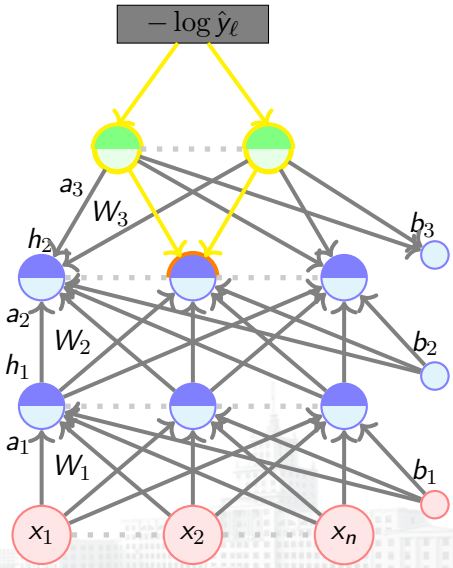
$$\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$
$$= \sum_{m=1}^{k} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

考虑如下两个向量：

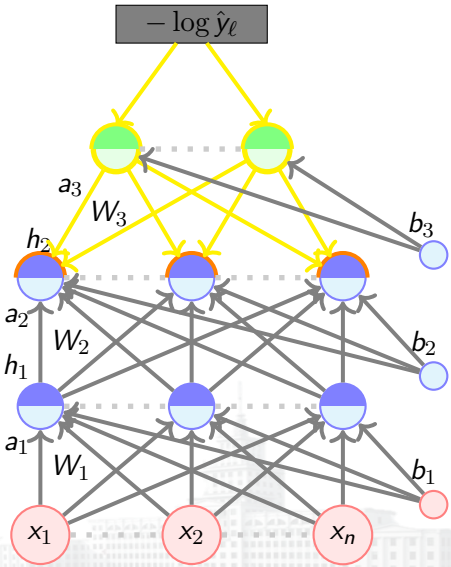
$$\nabla_{a_{i+1}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{i+1,k}} \end{bmatrix}; W_{i+1,\cdot,j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$

$W_{i+1,\cdot,j}$ 是 W_{i+1} 的第 j 列；下面

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



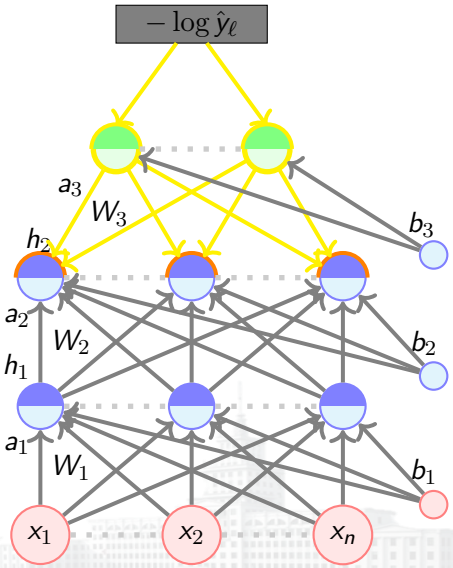
有, $\frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta)$



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

可以重写 gradient w.r.t. h_i

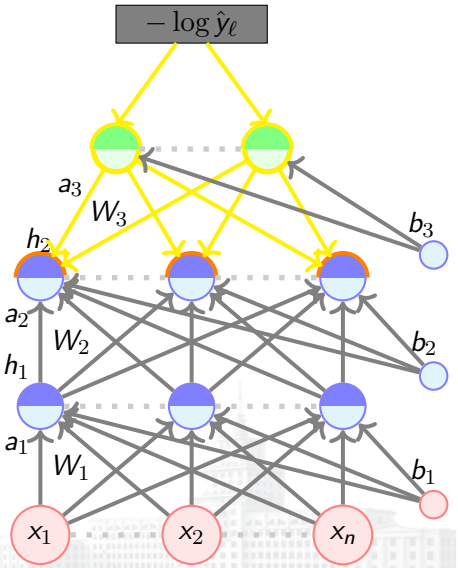
$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta)$



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

可以重写 gradient w.r.t. h_i

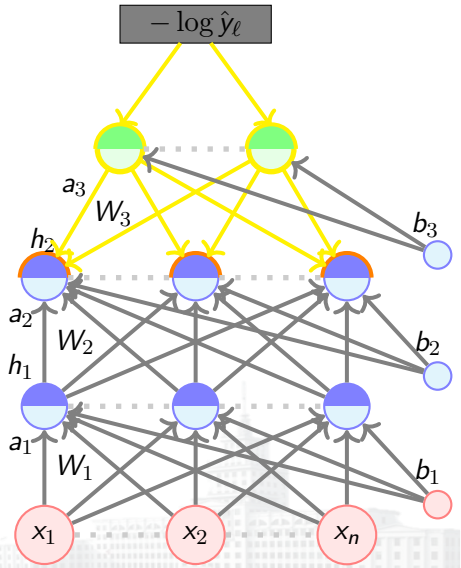
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$



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

可以重写 gradient w.r.t. h_i

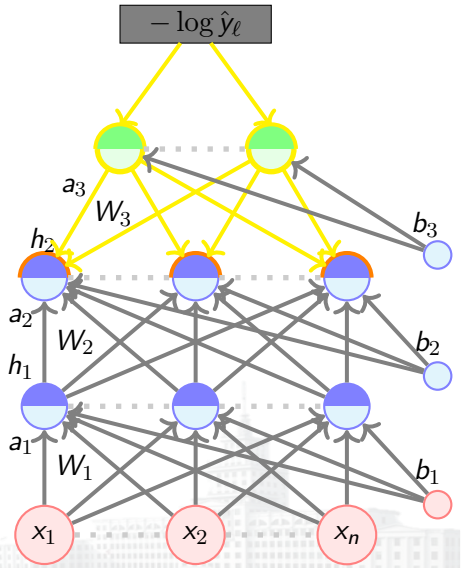
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}$$



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

可以重写 gradient w.r.t. h_i

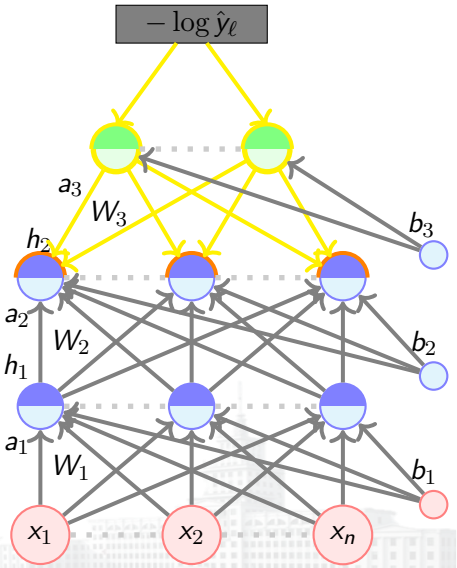
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \\ \\ \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \\ \\ \end{bmatrix}$$



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

可以重写 gradient w.r.t. h_i

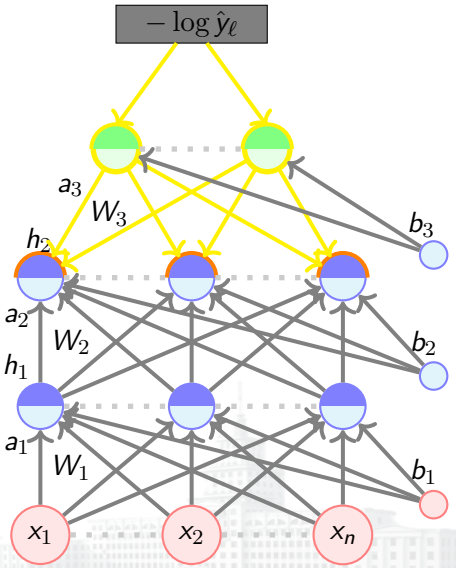
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \\ \\ \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ \\ \\ \\ \end{bmatrix}$$



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

可以重写 gradient w.r.t. h_i

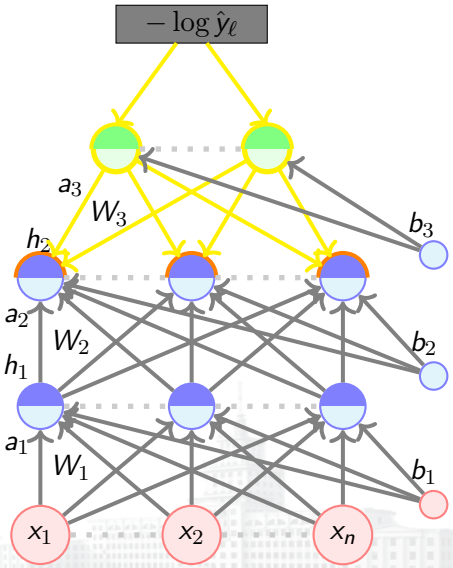
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \\ \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}}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



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

可以重写 gradient w.r.t. h_i

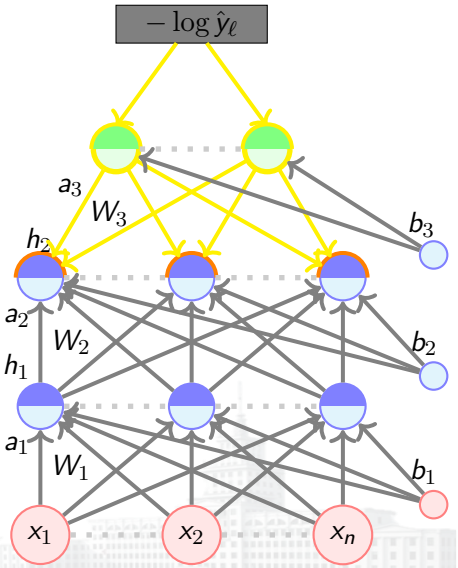
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} \mathscr{L}(\theta) \\ (W_{i+1,\cdot,2})^T \nabla_{a_{i+1}}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



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

可以重写 gradient w.r.t. h_i

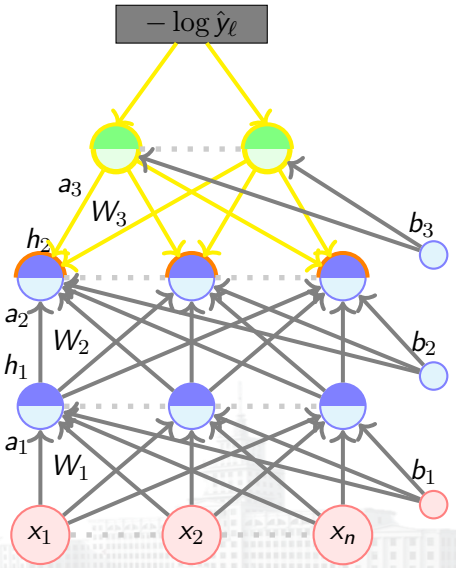
$$\nabla_{\mathbf{h_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{i2}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} \end{bmatrix} = \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}}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



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

可以重写 gradient w.r.t. h_i

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

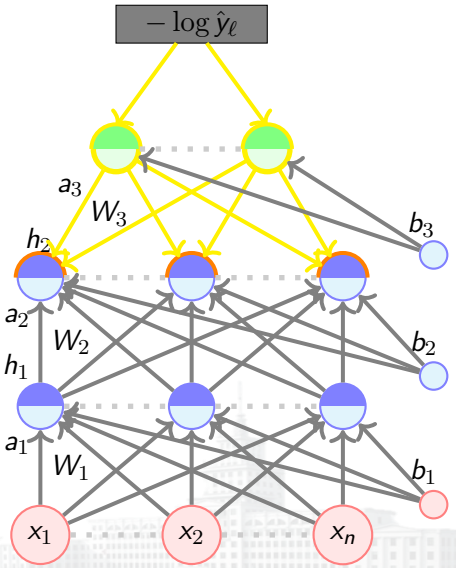


$$\text{有,} \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta)$$

可以重写 gradient w.r.t. h_i

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

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

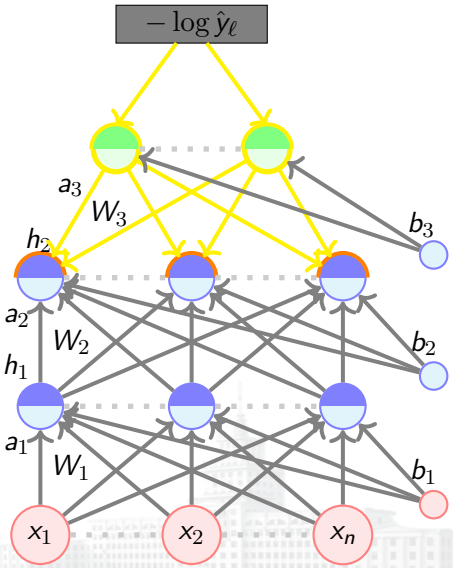


$$\text{有,} \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta)$$

可以重写 gradient w.r.t. h_i

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

- 现在就差计算 $\nabla_{a_{i+1}} \mathscr{L}(\theta)$ for $i < L-1$ 就大功告成了

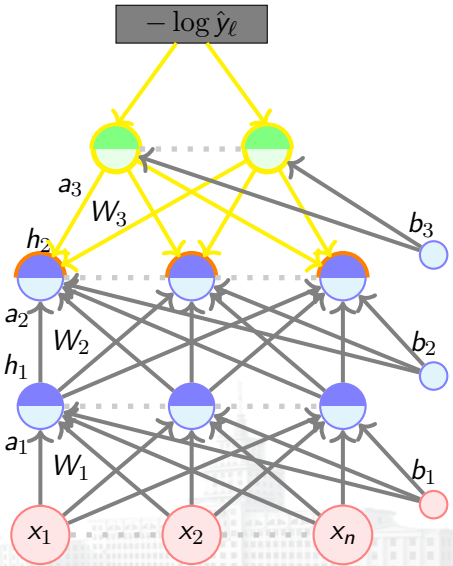


$$\text{有,} \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} = (W_{i+1,.,j})^T \nabla_{a_{i+1}} \mathscr{L}(\theta)$$

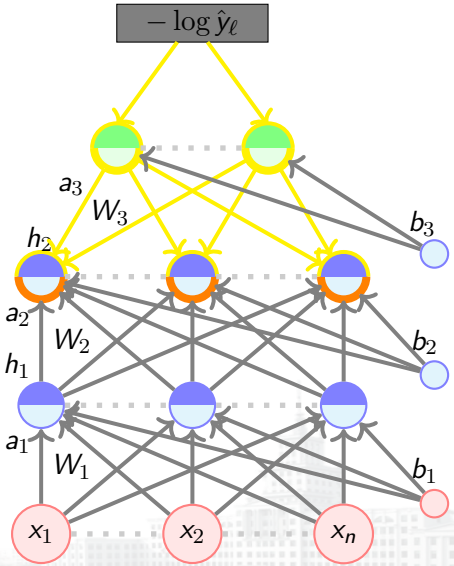
可以重写 gradient w.r.t. h_i

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

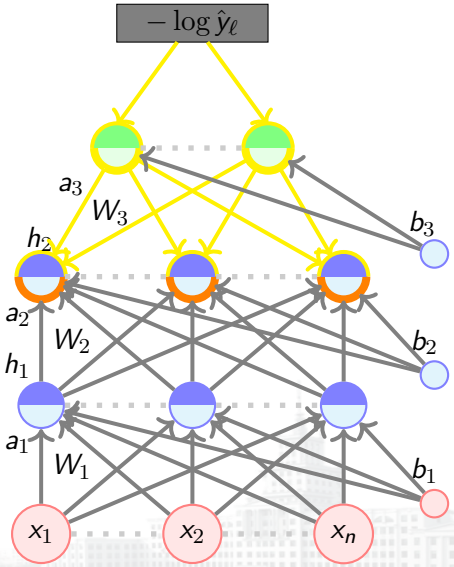
- 现在就差计算 $\nabla_{a_{i+1}} \mathscr{L}(\theta)$ for $i < L-1$ 就大功告成了
- 如何计算？



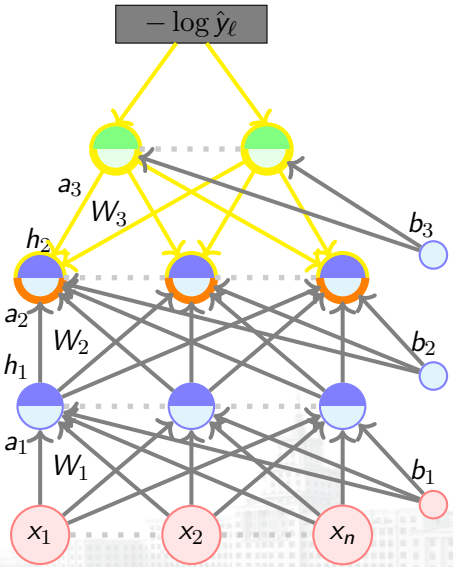
$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta)$



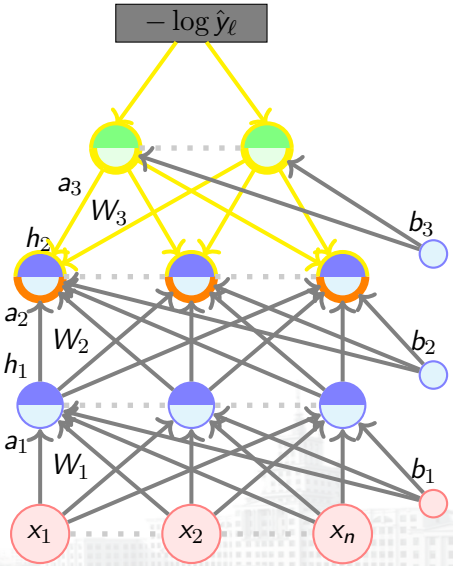
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$



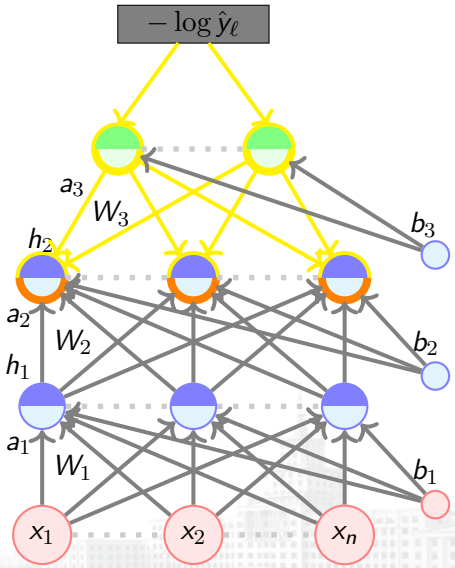
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \\ \\ \end{bmatrix}$$



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

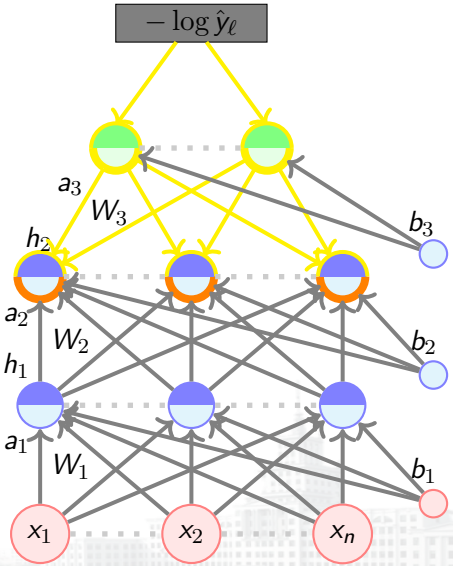


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$



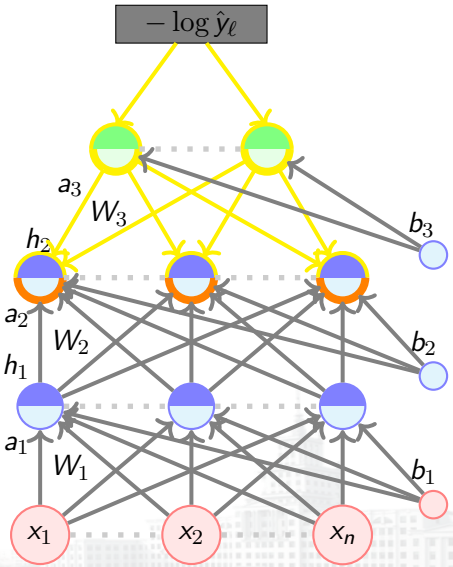
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

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



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

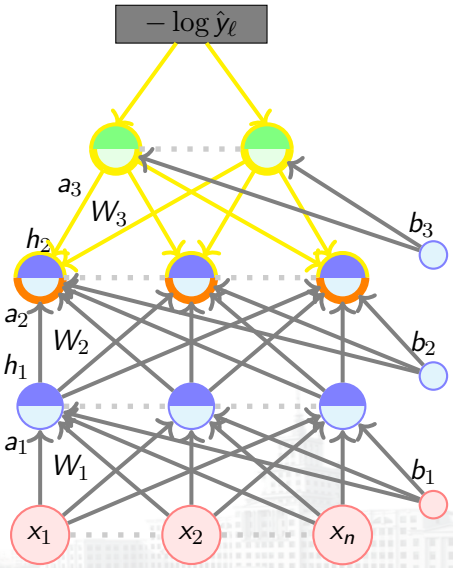
$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$



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

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

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

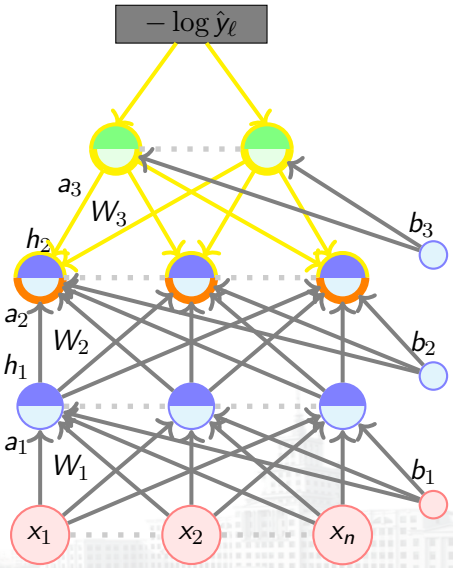


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

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

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

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

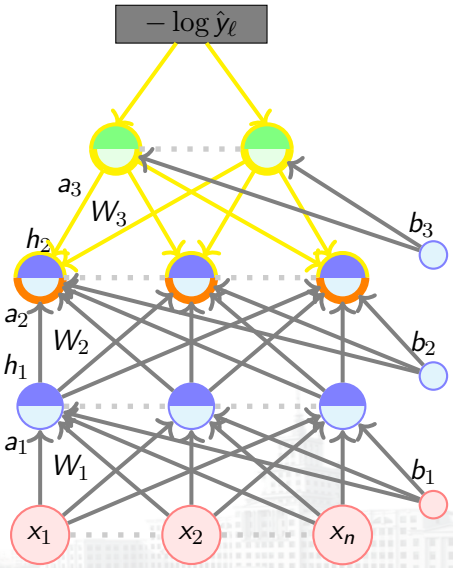


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

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

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

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

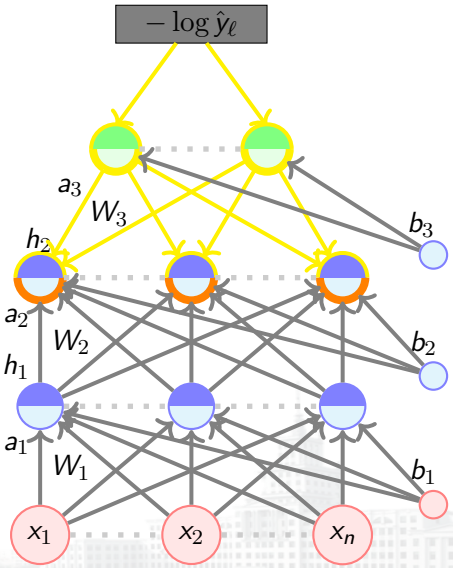


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

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

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

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

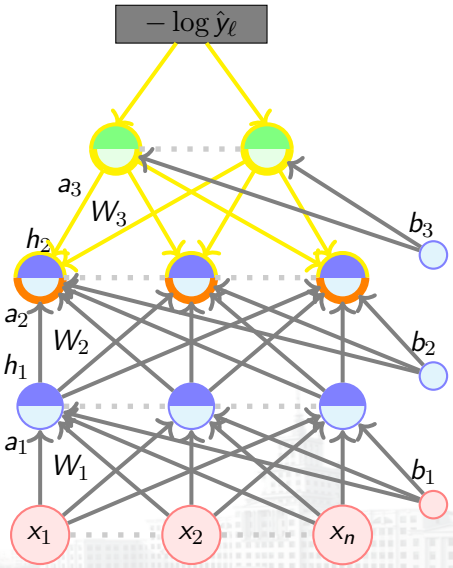


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

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

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

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

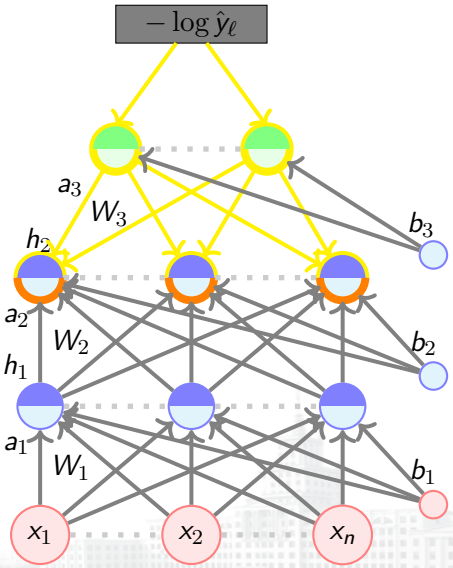


$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

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

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

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



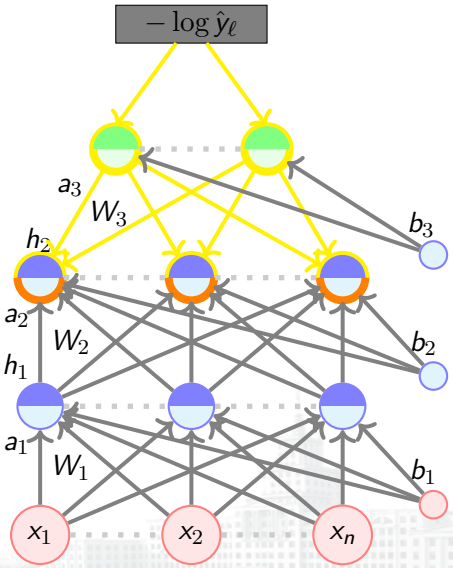
$$\nabla_{\mathbf{a_i}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

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

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

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

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



反向传播: 计算损失函数对参数的梯度（Computing Gradients w.r.t. Parameters）

需要计算的梯度:

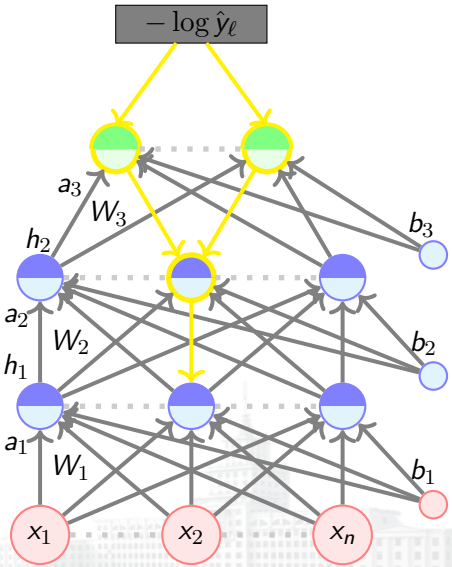
- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights and biases

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

- 我们关注 交叉熵损失函数和 *Softmax* 输出函数

回忆一下

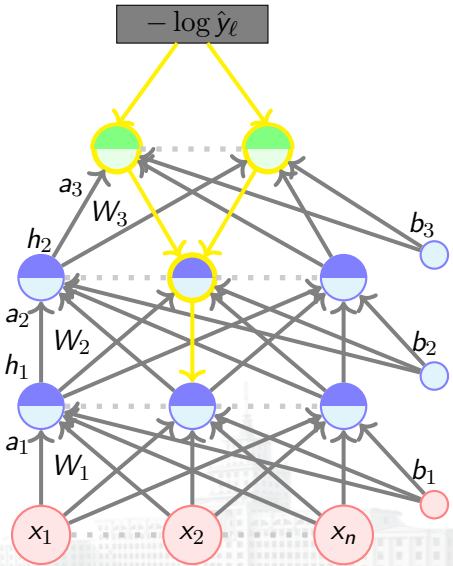
$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$



回忆一下

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

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

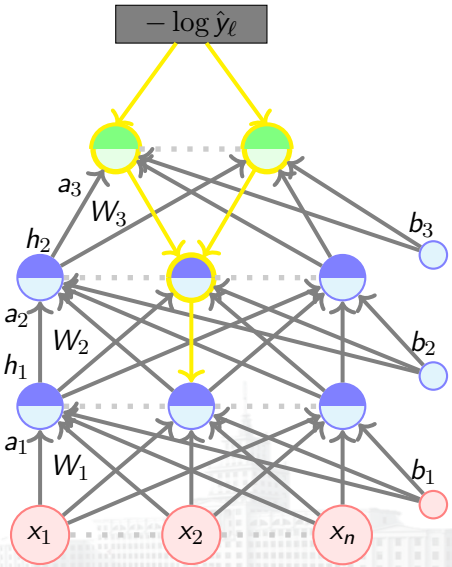


回忆一下

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

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

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

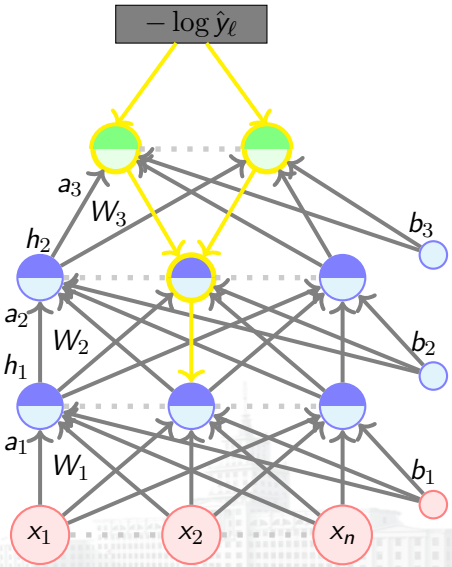


回忆一下

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

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

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



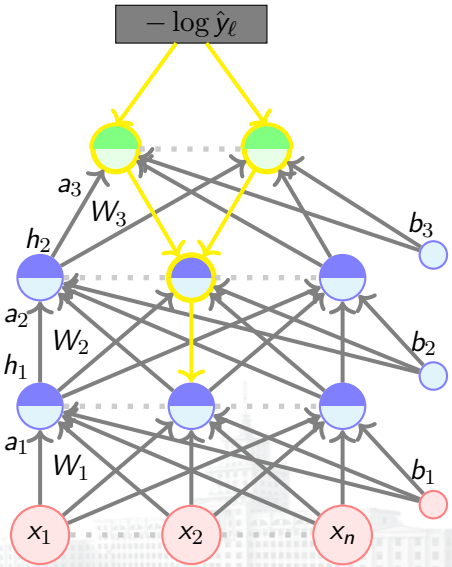
回忆一下

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

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

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

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



回忆一下

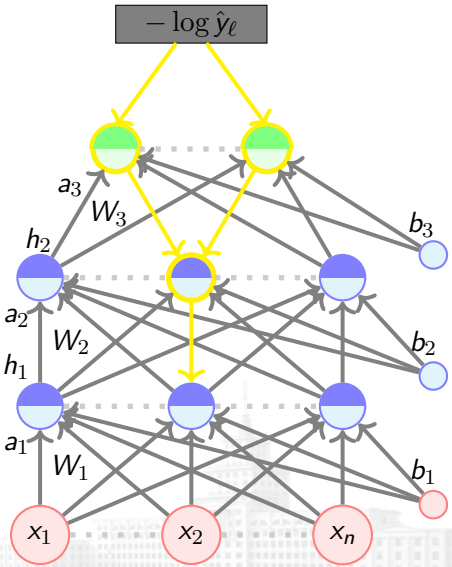
$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

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

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

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

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



回忆一下

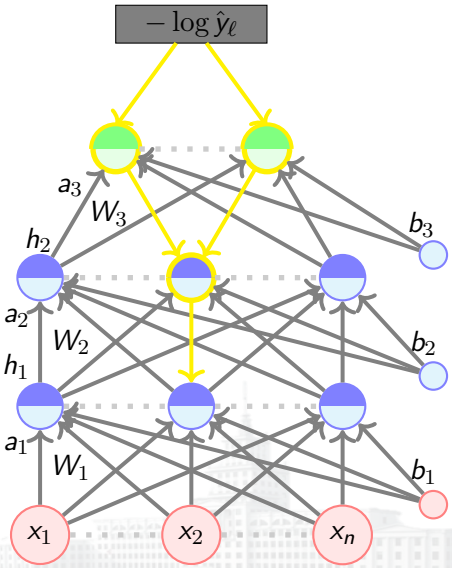
$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

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

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

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

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



Intentionally left blank

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

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

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

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

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

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

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

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

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

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

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

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

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

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

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

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

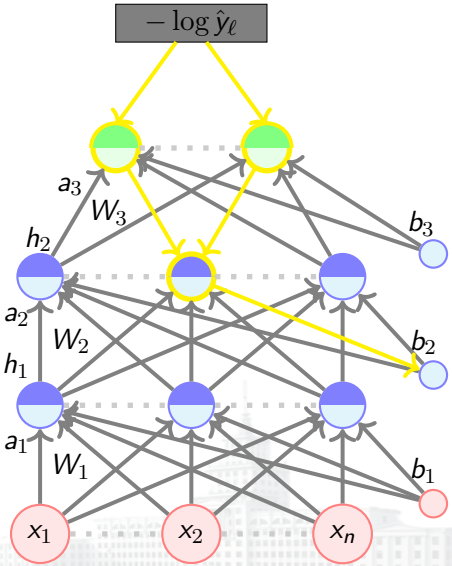
$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}}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

举一个列子 $W_k \in \mathbb{R}^{3\times 3}$ 来看看如何计算

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

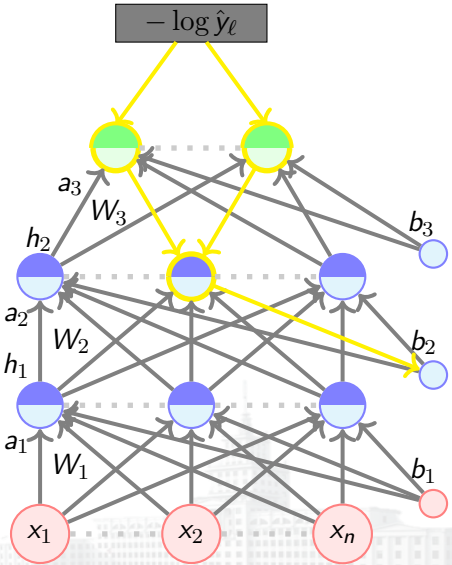
$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}}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

最后，看看对偏置的梯度



最后，看看对偏置的梯度

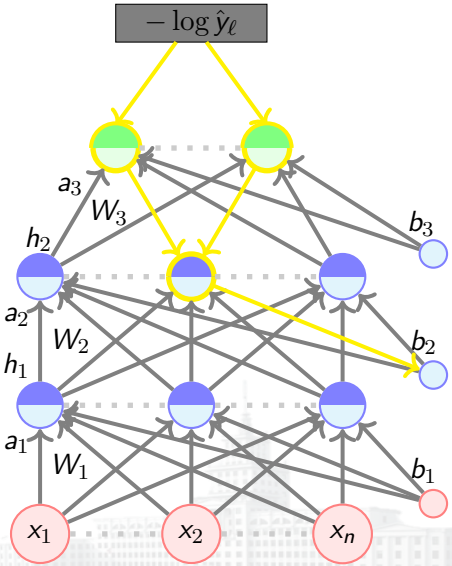
$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$



最后，看看对偏置的梯度

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

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

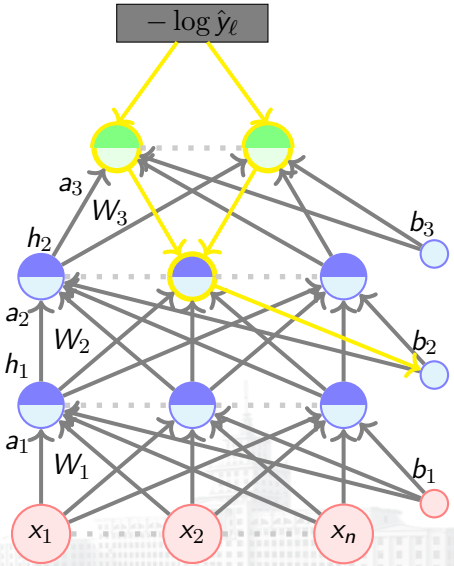


最后，看看对偏置的梯度

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

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

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



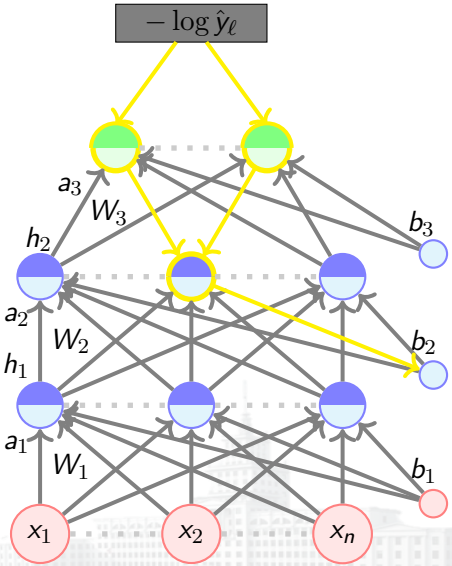
最后，看看对偏置的梯度

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

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

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

重写 gradient w.r.t. the vector b_k



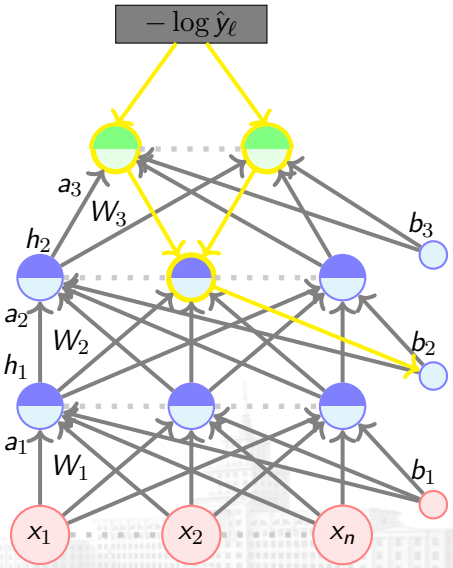
最后，看看对偏置的梯度

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

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

重写 gradient w.r.t. the vector b_k

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



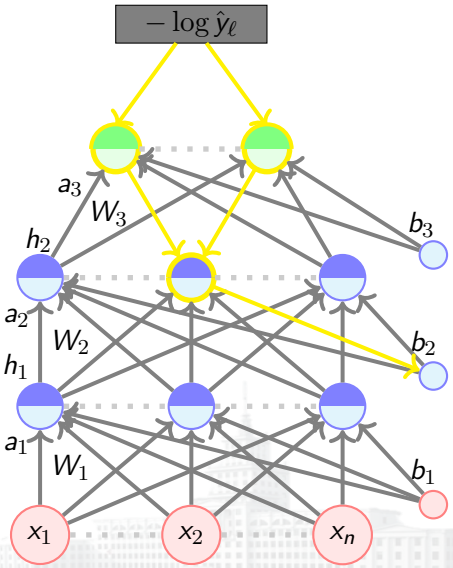
最后，看看对偏置的梯度

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

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

重写 gradient w.r.t. the vector b_k

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



反向传播: Pseudo code

所有需要的梯度已经计算出来

所有需要的梯度已经计算出来

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. output layer)}$$

所有需要的梯度已经计算出来

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. output layer)}$$

$$\nabla_{\mathbf{h_k}} \mathscr{L}(\theta), \nabla_{\mathbf{a_k}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. hidden layers, } 1 \leq k < L\text{)}$$

所有需要的梯度已经计算出来

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. output layer)}$$

$$\nabla_{\mathbf{h_k}} \mathscr{L}(\theta), \nabla_{\mathbf{a_k}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. hidden layers, } 1 \leq k < L\text{)}$$

$$\nabla_{W_k} \mathscr{L}(\theta), \nabla_{\mathbf{b_k}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. weights and biases, } 1 \leq k \leq L\text{)}$$

所有需要的梯度已经计算出来

$$\nabla_{\mathbf{a_L}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. output layer)}$$

$$\nabla_{\mathbf{h_k}} \mathscr{L}(\theta), \nabla_{\mathbf{a_k}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. hidden layers, } 1 \leq k < L\text{)}$$

$$\nabla_{W_k} \mathscr{L}(\theta), \nabla_{\mathbf{b_k}} \mathscr{L}(\theta) \quad \text{(gradient w.r.t. weights and biases, } 1 \leq k \leq L\text{)}$$

可以得出整个学习算法

Algorithm: gradient_descent()

$t \leftarrow 0$;
max_iterations $\leftarrow 1000$;
Initialize $\theta_0 = [W_1^0, ..., W_L^0, b_1^0, ..., b_L^0]$;

Algorithm: gradient_descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize $\theta_0 = [W_1^0, ..., W_L^0, b_1^0, ..., b_L^0]$;
while $t{+}{+} < max\_iterations$ **do**

end

Algorithm: gradient_descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize $\theta_0 = [W_1^0, ..., W_L^0, b_1^0, ..., b_L^0]$;
while $t{+}{+} < max\_iterations$ **do**
    $h_1, h_2, ..., h_{L-1}, a_1, a_2, ..., a_L, \hat{y} = forward\_propagation(\theta_t)$;
end

Algorithm: gradient_descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize $\theta_0 = [W_1^0, ..., W_L^0, b_1^0, ..., b_L^0]$;
while $t{+}{+} < max\_iterations$ **do**
    $h_1, h_2, ..., h_{L-1}, a_1, a_2, ..., a_L, \hat{y} = forward\_propagation(\theta_t)$;
    $\nabla\theta_t = backward\_propagation(h_1, h_2, ..., h_{L-1}, a_1, a_2, ..., a_L, y, \hat{y})$;
end

Algorithm: gradient_descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize $\theta_0 = [W_1^0, ..., W_L^0, b_1^0, ..., b_L^0]$;
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)

Algorithm: forward_propagation(θ)

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

end

Algorithm: forward_propagation(θ)

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

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

end

Algorithm: forward_propagation(θ)

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

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

$h_k = g(a_k)$;

end

Algorithm: forward_propagation(θ)

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

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

$h_k = g(a_k)$;

end

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

Algorithm: forward_propagation(θ)

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

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

$h_k = g(a_k)$;

end

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

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

执行前向传播，计算所有的 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 ;

执行前向传播，计算所有的 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 ;
$\nabla_{a_L} \mathscr{L}(\theta) = -(e(y) - \hat{y})$;

执行前向传播，计算所有的 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**

end

执行前向传播，计算所有的 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 ;

end

执行前向传播，计算所有的 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 ;
    $\nabla_{W_k} \mathscr{L}(\theta) = \nabla_{a_k} \mathscr{L}(\theta) h_{k-1}^T$;

end

执行前向传播，计算所有的 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 ;

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

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

end

执行前向传播，计算所有的 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 ;

end

执行前向传播，计算所有的 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 ;

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

end

执行前向传播，计算所有的 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);
end

执行前向传播，计算所有的 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

激活函数的导数

最后需要解决的事情是如何计算 g'

最后需要解决的事情是如何计算 g'

Logistic function

$$\begin{aligned} g(z) &= \sigma(z) \\ &= \frac{1}{1+e^{-z}} \end{aligned}$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z})
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z})
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z}) \\
&= \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right)
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z}) \\
&= \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z}) \\
&= \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}
$$

tanh

$$
\begin{aligned}
g(z) &= \tanh(z) \\
&= \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z}) \\
&= \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}
$$

tanh

$$
\begin{aligned}
g(z) &= \tanh(z) \\
&= \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \\
g'(z) &= \frac{\left(\begin{array}{l}(e^{z}+e^{-z})\frac{d}{dz}(e^{z}-e^{-z}) \\ \quad -(e^{z}-e^{-z})\frac{d}{dz}(e^{z}+e^{-z})\end{array}\right)}{(e^{z}+e^{-z})^2}
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z}) \\
&= \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}
$$

tanh

$$
\begin{aligned}
g(z) &= \tanh(z) \\
&= \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \\
g'(z) &= \frac{\left(\begin{array}{l}(e^{z}+e^{-z})\frac{d}{dz}(e^{z}-e^{-z}) \\ \quad -(e^{z}-e^{-z})\frac{d}{dz}(e^{z}+e^{-z})\end{array}\right)}{(e^{z}+e^{-z})^2} \\
&= \frac{(e^{z}+e^{-z})^2-(e^{z}-e^{-z})^2}{(e^{z}+e^{-z})^2}
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z}) \\
&= \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}
$$

tanh

$$
\begin{aligned}
g(z) &= \tanh(z) \\
&= \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \\
g'(z) &= \frac{\left(\begin{array}{l}(e^{z}+e^{-z})\frac{d}{dz}(e^{z}-e^{-z}) \\ \quad -(e^{z}-e^{-z})\frac{d}{dz}(e^{z}+e^{-z})\end{array}\right)}{(e^{z}+e^{-z})^2} \\
&= \frac{(e^{z}+e^{-z})^2-(e^{z}-e^{-z})^2}{(e^{z}+e^{-z})^2} \\
&= 1-\frac{(e^{z}-e^{-z})^2}{(e^{z}+e^{-z})^2}
\end{aligned}
$$

最后需要解决的事情是如何计算 g'

Logistic function

$$
\begin{aligned}
g(z) &= \sigma(z) \\
&= \frac{1}{1+e^{-z}} \\
g'(z) &= (-1)\frac{1}{(1+e^{-z})^2}\frac{d}{dz}(1+e^{-z}) \\
&= (-1)\frac{1}{(1+e^{-z})^2}(-e^{-z}) \\
&= \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}
$$

tanh

$$
\begin{aligned}
g(z) &= \tanh(z) \\
&= \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \\
g'(z) &= \frac{\left(\begin{array}{l}(e^{z}+e^{-z})\frac{d}{dz}(e^{z}-e^{-z}) \\ \quad -(e^{z}-e^{-z})\frac{d}{dz}(e^{z}+e^{-z})\end{array}\right)}{(e^{z}+e^{-z})^2} \\
&= \frac{(e^{z}+e^{-z})^2-(e^{z}-e^{-z})^2}{(e^{z}+e^{-z})^2} \\
&= 1-\frac{(e^{z}-e^{-z})^2}{(e^{z}+e^{-z})^2} \\
&= 1-(g(z))^2
\end{aligned}
$$