CS/DS 541: Class 6

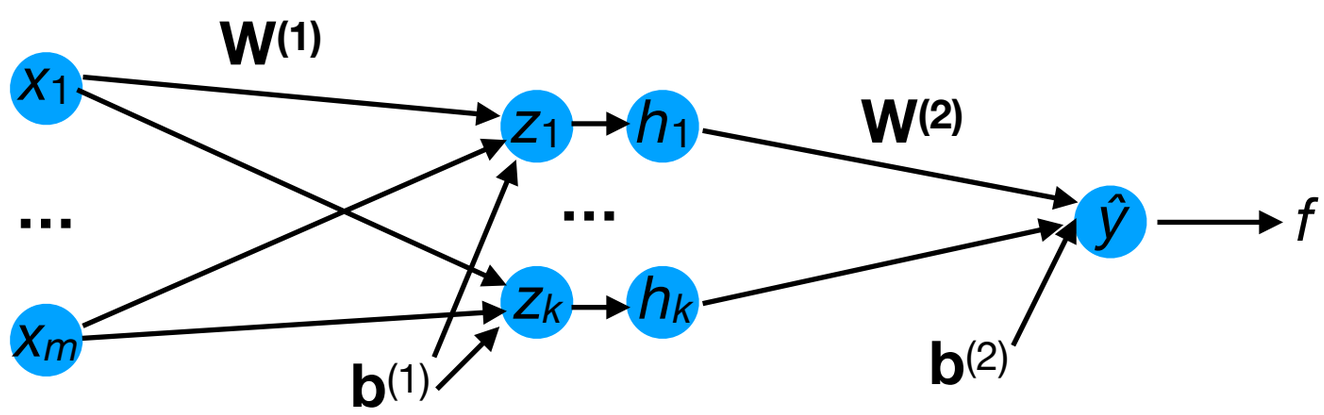
Jacob Whitehill

Training neural networks

Training neural networks

- To train arbitrarily deep NNs, we use the same strategy as we did for linear regression and softmax regression:
- For each parameter *p*, estimate how the loss function changes as *p* changes, i.e., compute:

$$\nabla_p f(\mathbf{X}, \mathbf{y}; \mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots)$$



Training neural networks

- To train arbitrarily deep NNs, we use the same strategy as we did for linear regression and softmax regression:
  - For each parameter *p*, estimate how the loss function changes as *p* changes, i.e., compute:

$$\nabla_p f(\mathbf{X}, \mathbf{y}; \mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots)$$

  - Update the parameter by moving it slightly opposite the gradient:

$$p^{\text{new}} \leftarrow p^{\text{old}} - \epsilon \nabla_p f(\mathbf{X}, \mathbf{y}; \ldots p^{\text{old}} \ldots)$$

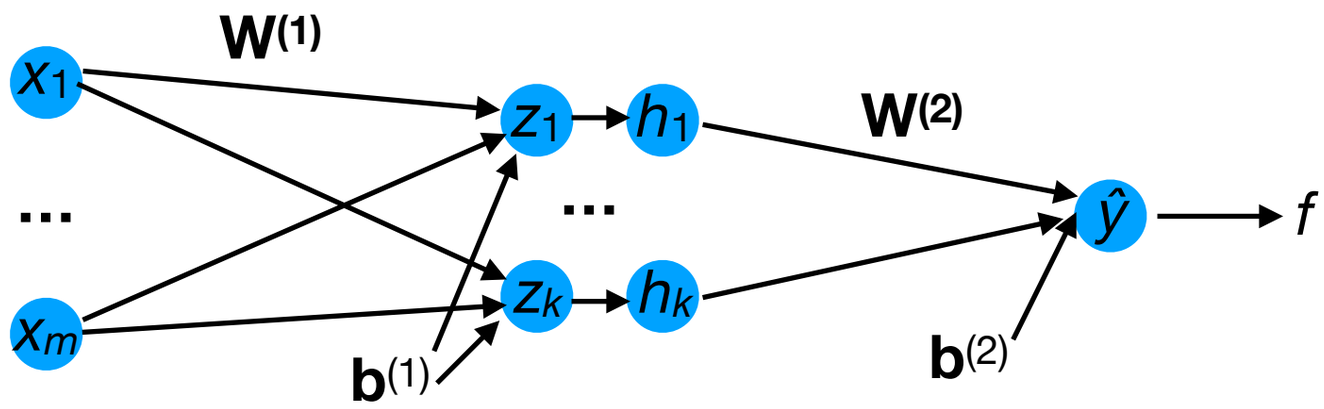
Training neural networks

- Neural networks represent deep compositions of functions.

- Hence, to compute each gradient, we use the chain rule of multivariate calculus, i.e.:

$$\frac{\partial(f \circ g \circ h)}{\partial \mathbf{x}} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial h}\frac{\partial h}{\partial \mathbf{x}}$$

- This means we must compute multiple Jacobian matrices and multiple them together.



Jacobian matrices

- For any function $f : \mathbb{R}^m \to \mathbb{R}^n$, we can define the **Jacobian matrix** of all partial derivatives:

Columns are the *inputs* to *f*.

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{x}_1} & \cdots & \frac{\partial f_1}{\partial \mathbf{x}_m} \\ & \ddots & \\ \frac{\partial f_n}{\partial \mathbf{x}_1} & \cdots & \frac{\partial f_n}{\partial \mathbf{x}_m} \end{bmatrix}$$

Row are the *outputs* of *f*.

Computing the gradients

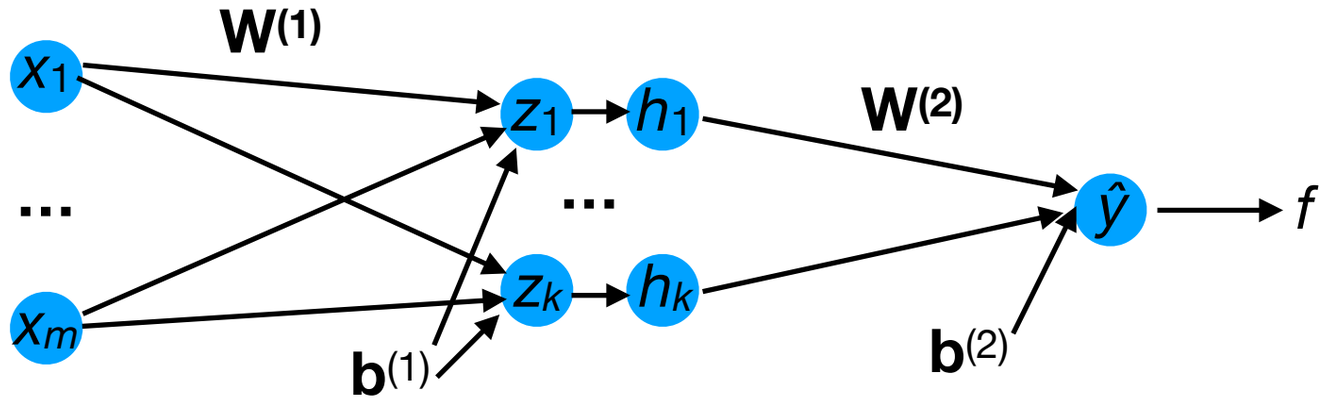
- To train the NN below, we need to compute the gradients:

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{b}^{(1)}}$$



Computing the gradients

- Note that the gradient terms share some computation:

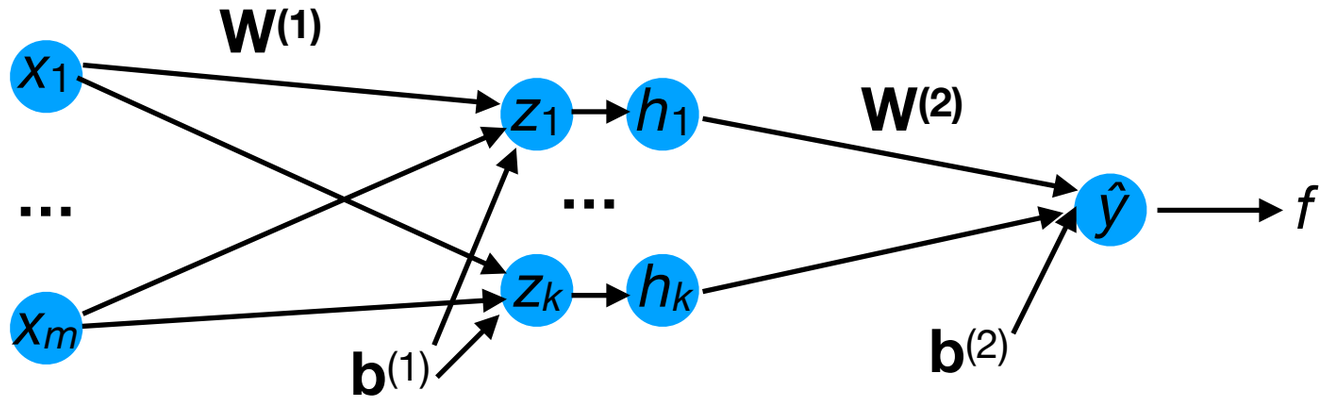
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$

Redundant computation

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

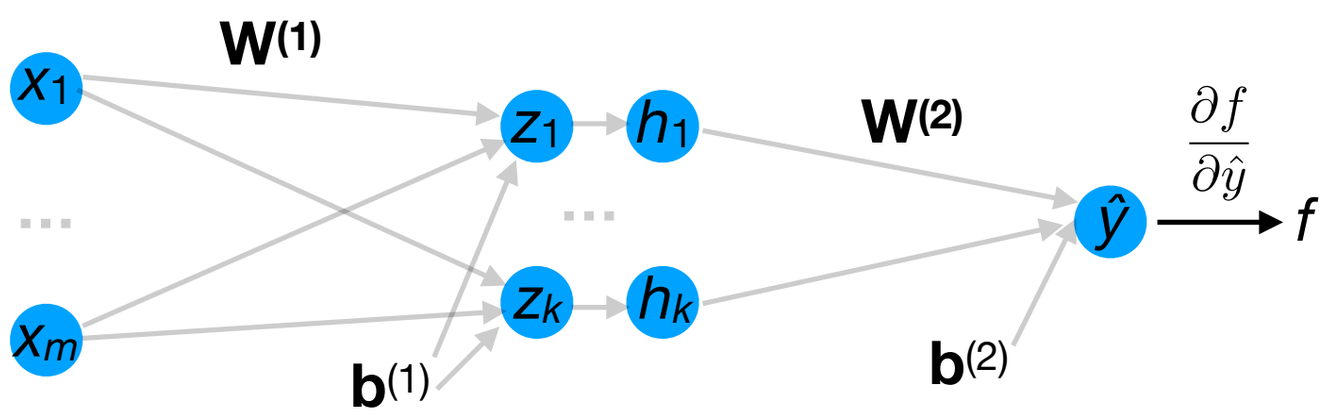
$$\frac{\partial f}{\partial \mathbf{b}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{b}^{(1)}}$$



Computing the gradients

- Here's how we can compute all these *efficiently*:

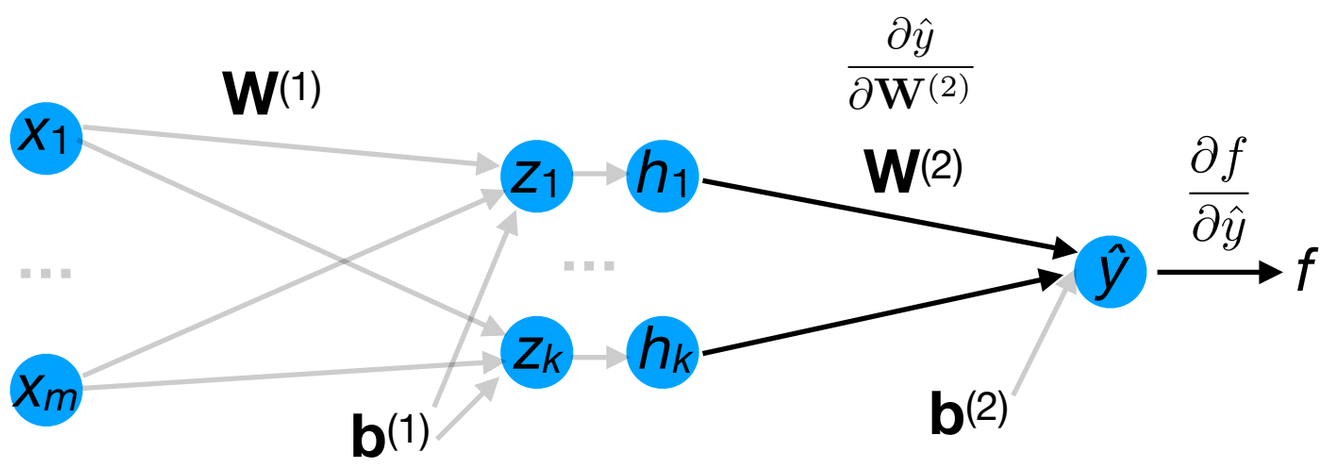
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}}$$



Computing the gradients

- Here's how we can compute all these *efficiently*:

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

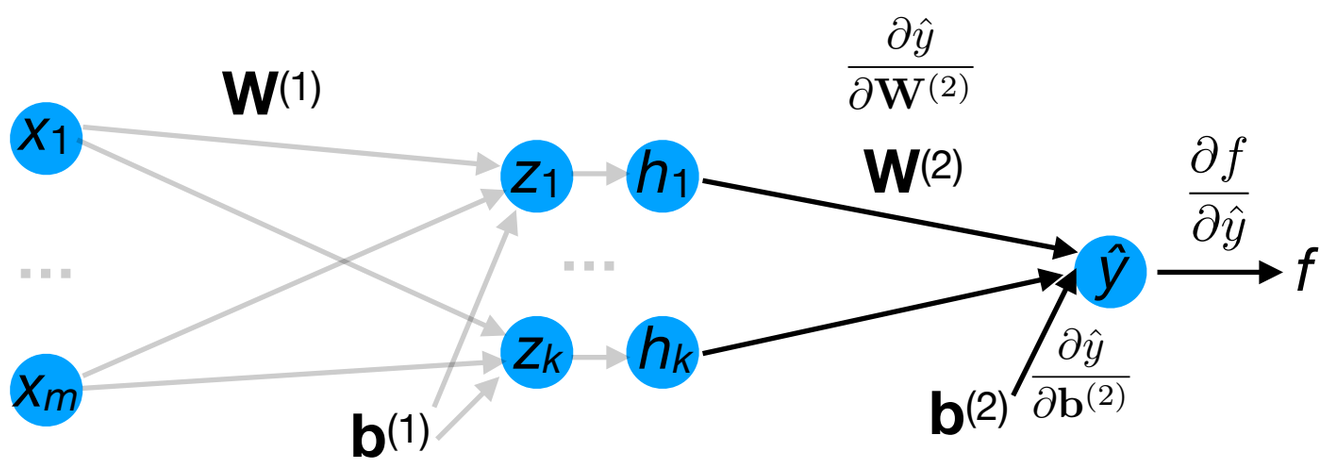


Computing the gradients

- Here's how we can compute all these *efficiently*:

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$



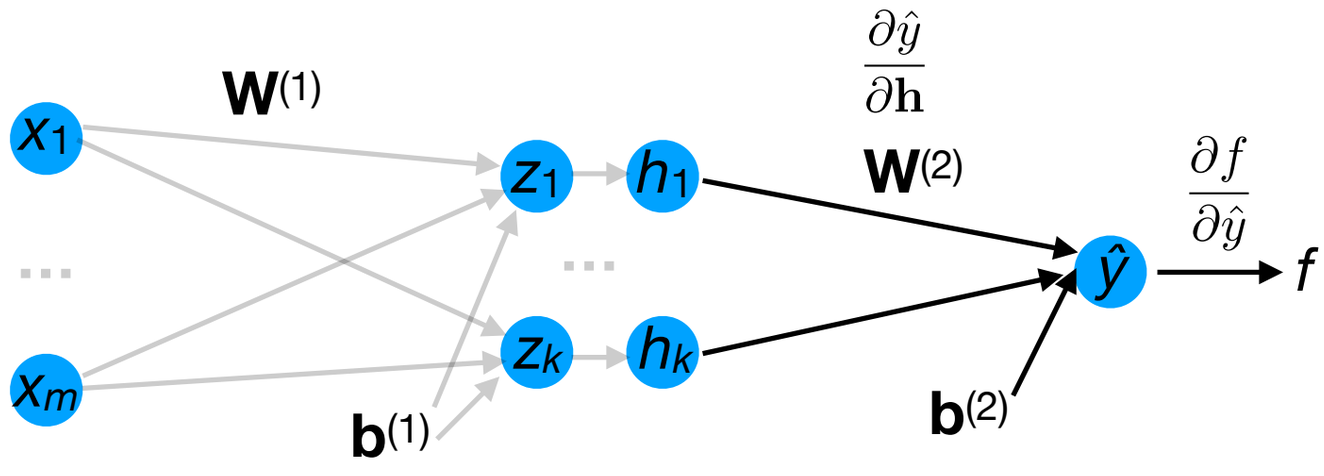
Computing the gradients

- Here's how we can compute all these *efficiently*:

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}}$$



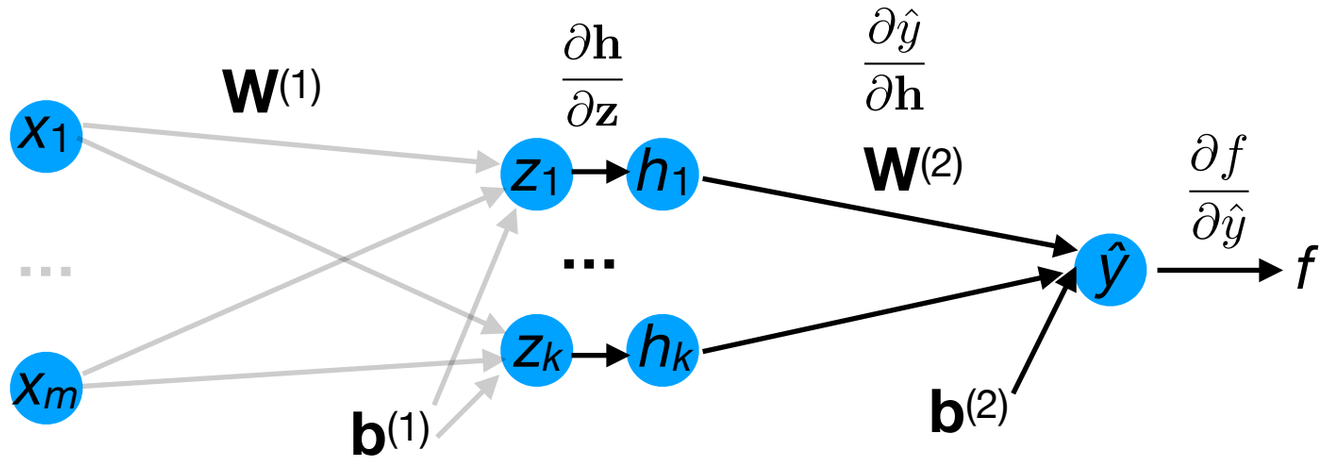
Computing the gradients

- Here's how we can compute all these *efficiently*:

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}}$$



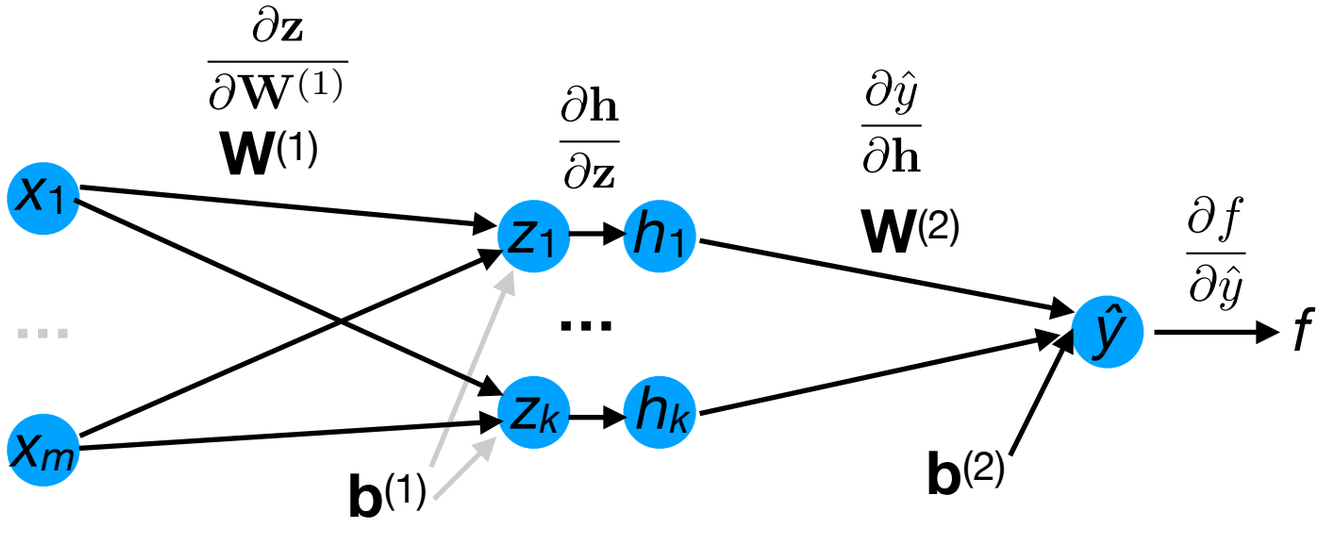
Computing the gradients

- Here's how we can compute all these *efficiently*:

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$



Computing the gradients

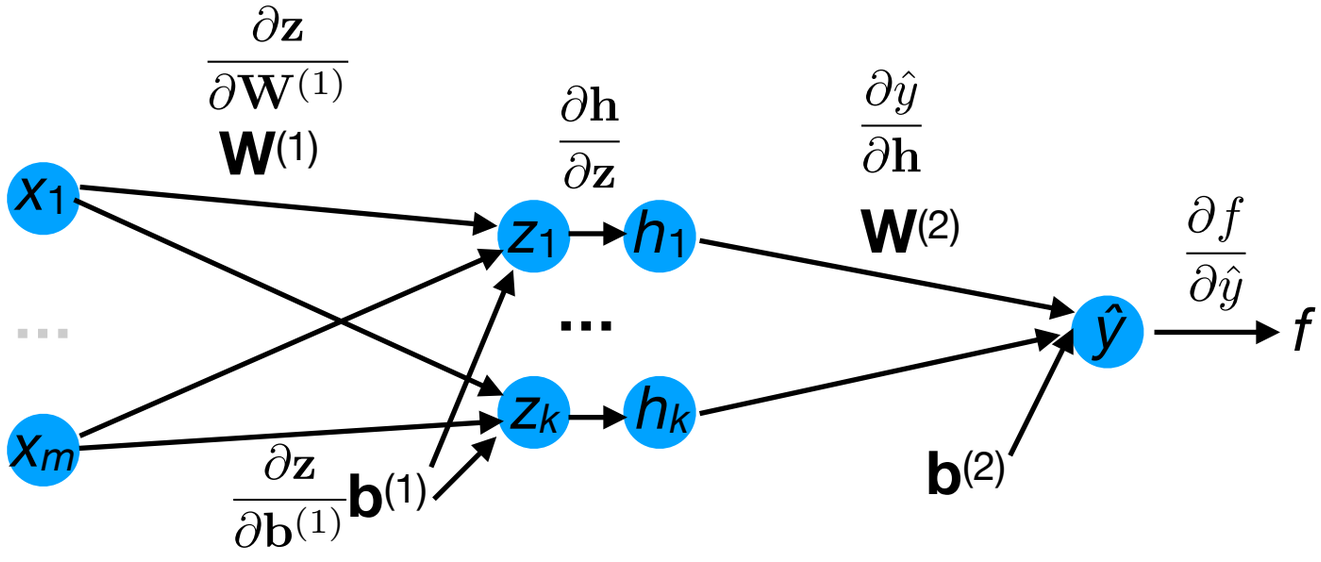
- Here's how we can compute all these *efficiently*:

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$

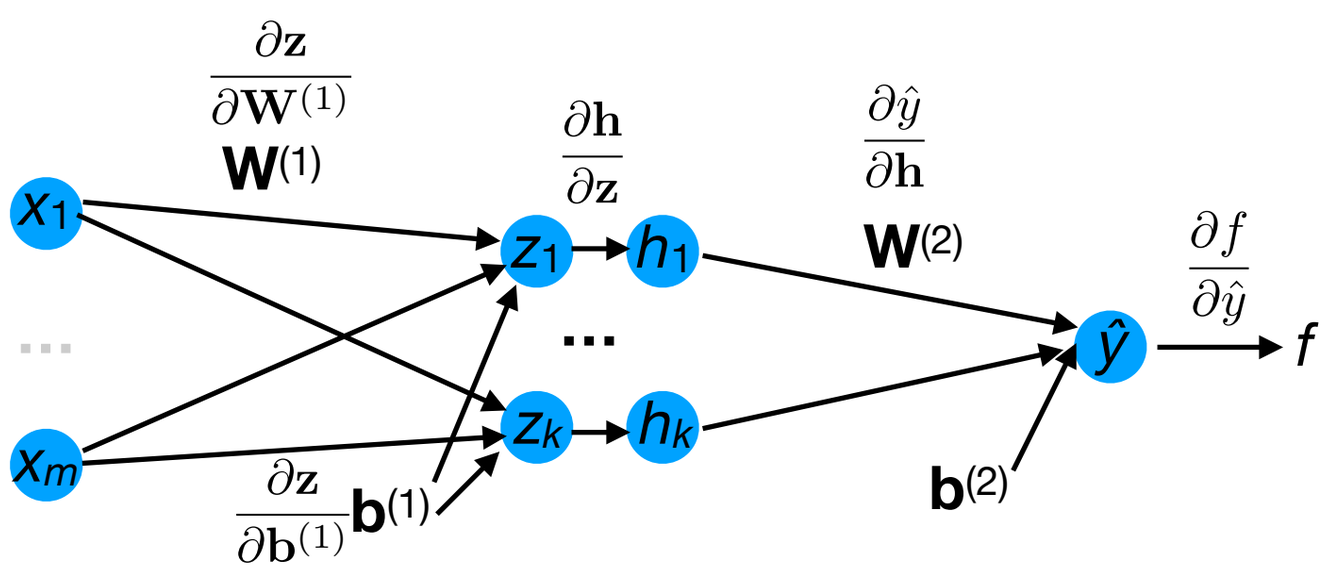
$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

$$\frac{\partial f}{\partial \mathbf{b}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{b}^{(1)}}$$



Computing the gradients

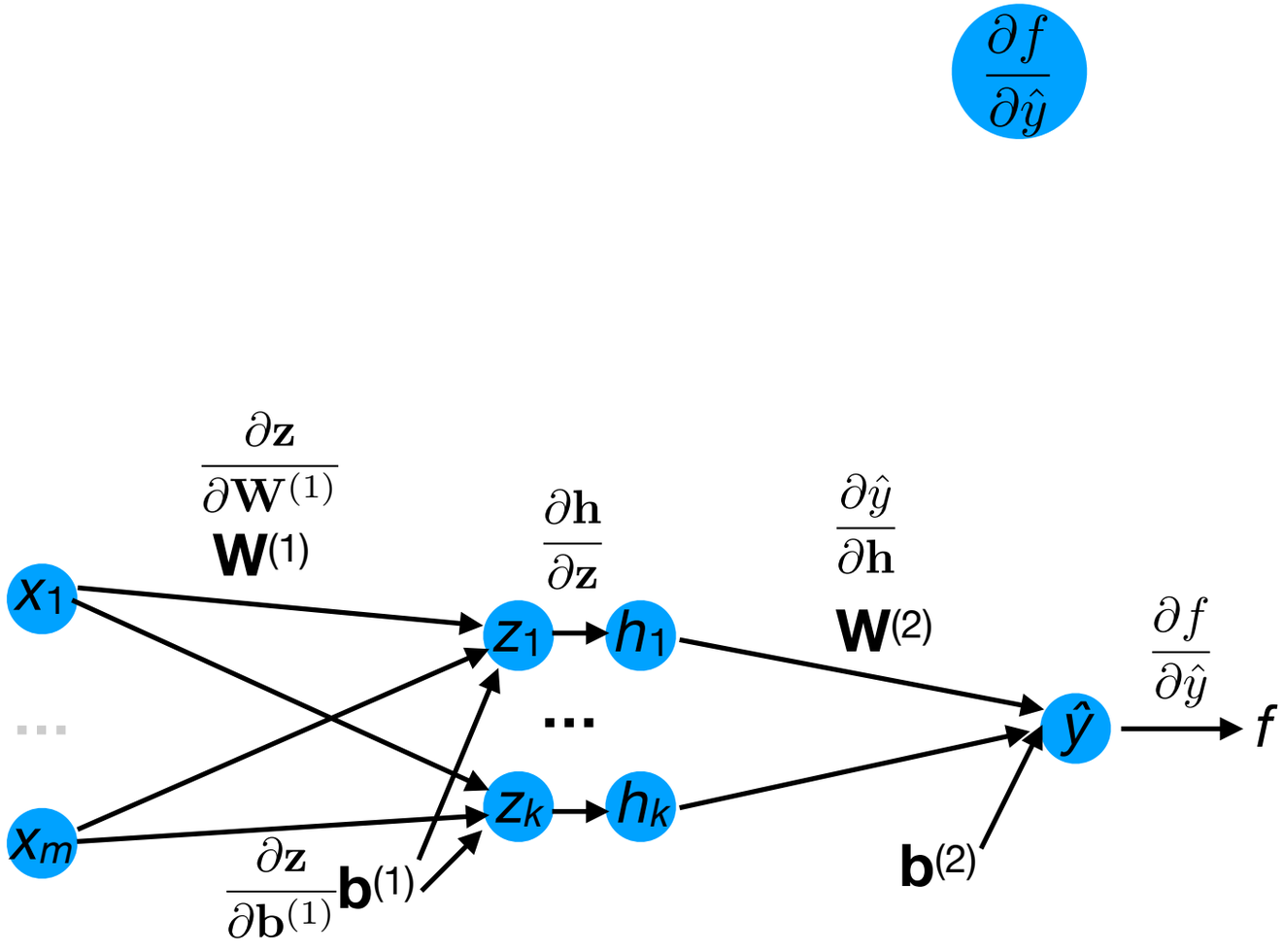
- This process is known as **backwards propagation ("backprop"):**
  - It produces the gradient terms of all the weight matrices and bias vectors.
  - It requires first conducting forward propagation.



Computing the gradients

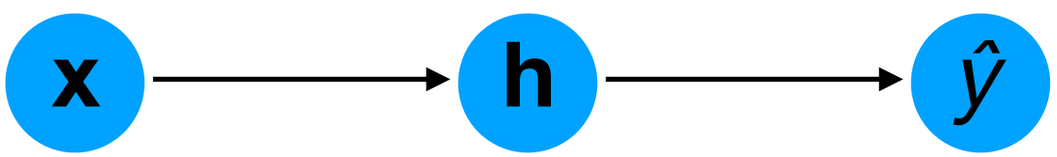
Forward propagation

Backward propagation

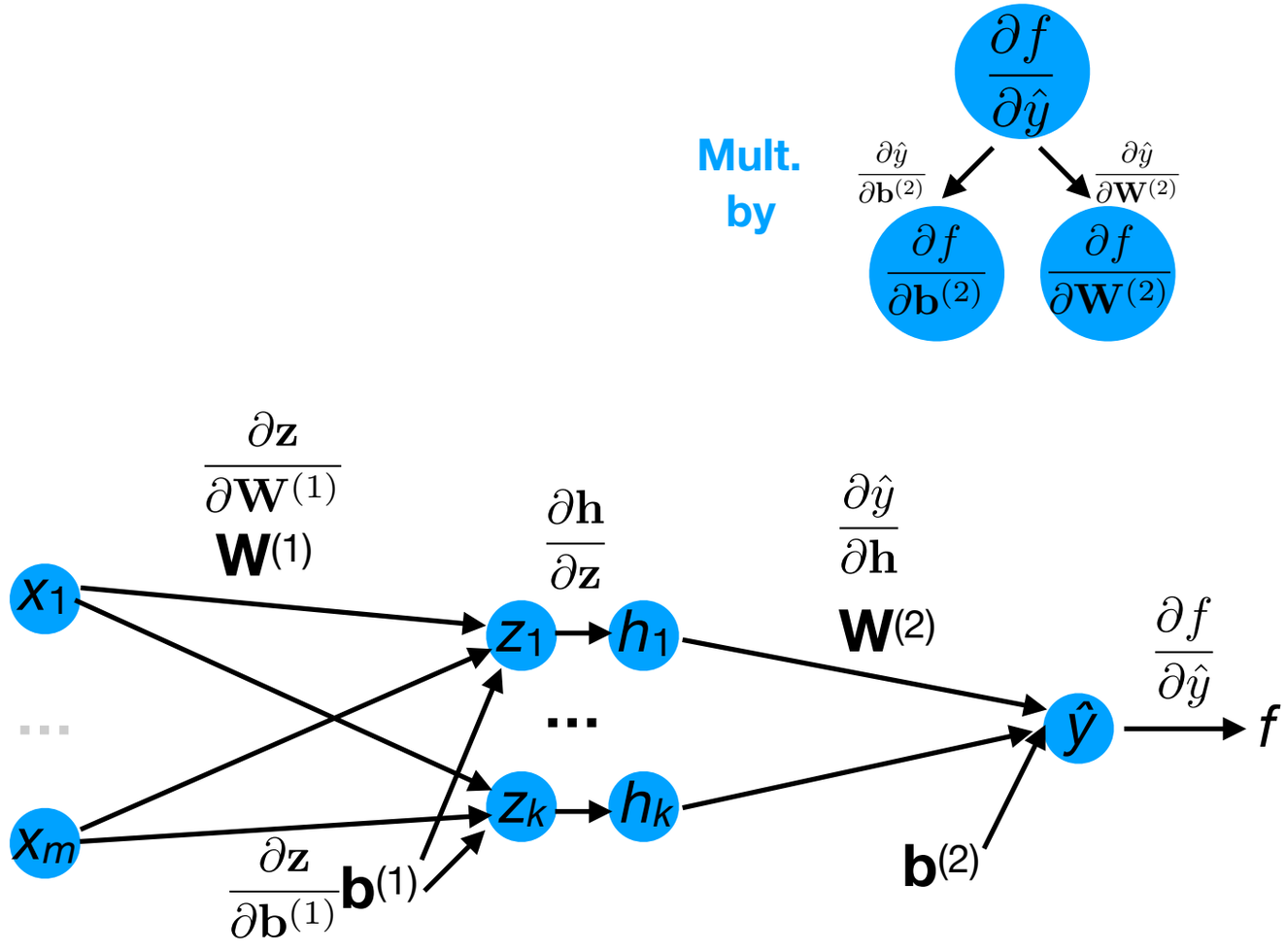


Computing the gradients

Forward propagation



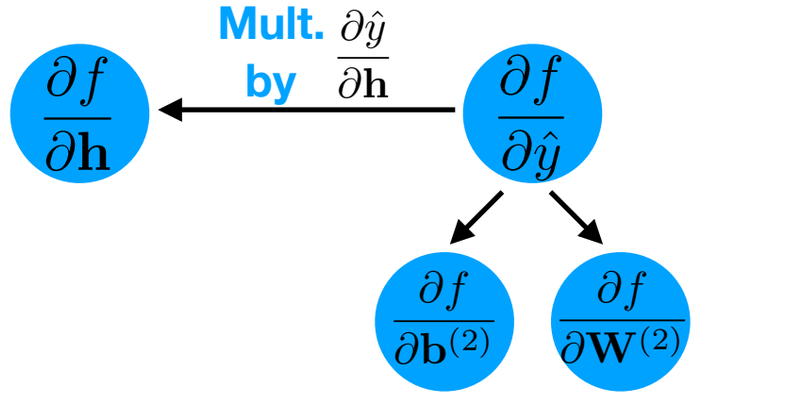
Backward propagation

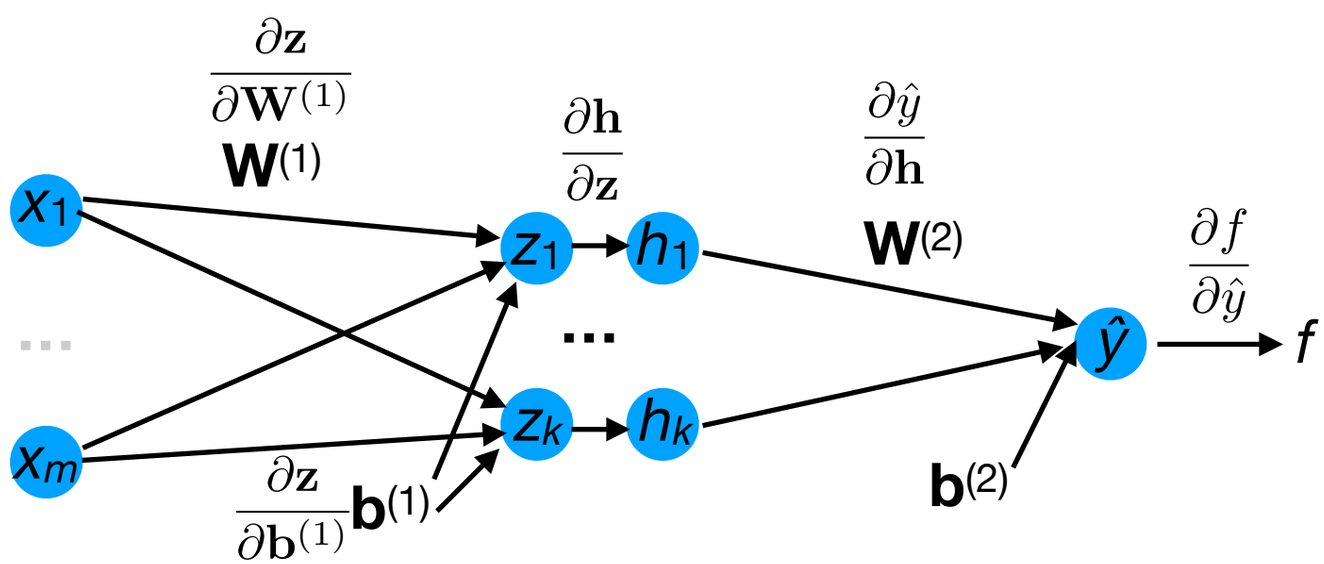


Computing the gradients

Forward propagation

Backward propagation

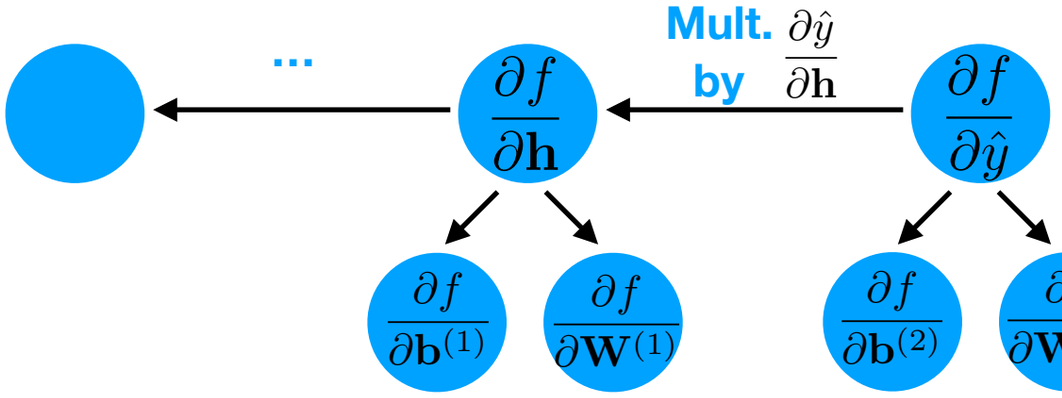


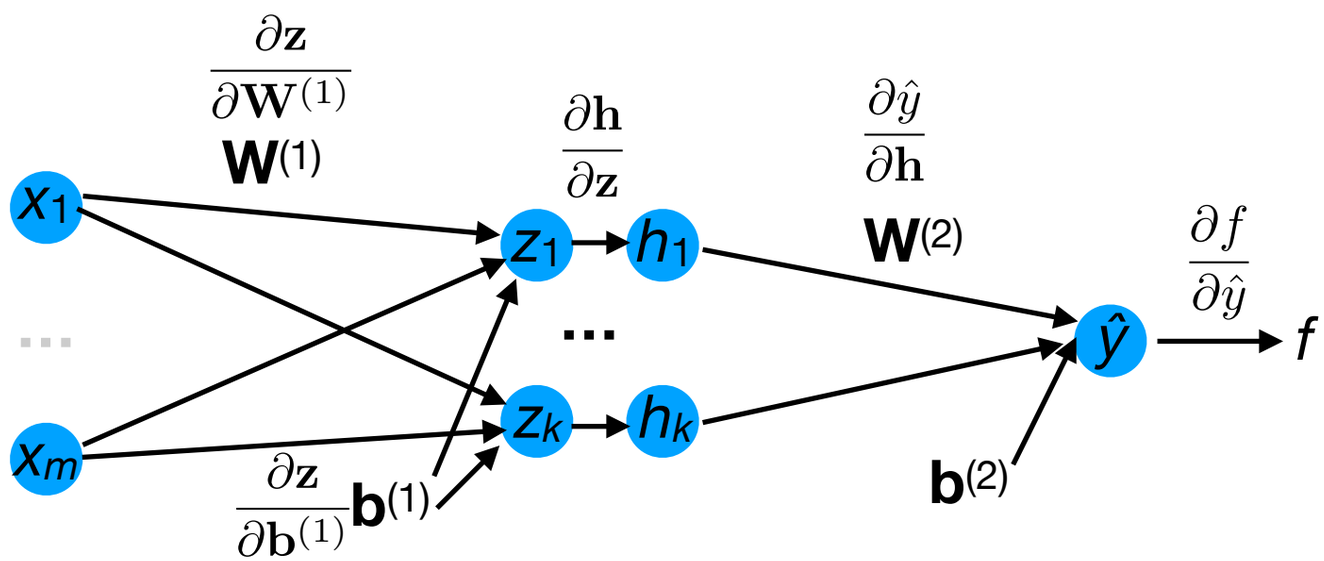


Computing the gradients

Forward propagation

Backward propagation





Computing the gradients

- Applying the chain rule is equivalent to multiplying a sequence of Jacobian matrices.

- However, for the vast majority of NN designs, it turns out that we can simplify this process analytically.

- After simplifying, we obtain the following algorithm for training arbitrarily deep FFNNs…

Computing the gradients

Algorithm 6.4 Backward computation for the deep neural network of algorithm 6.3, which uses, in addition to the input $\boldsymbol{x}$, a target $\boldsymbol{y}$. This computation yields the gradients on the activations $\boldsymbol{a}^{(k)}$ for each layer k, starting from the output layer and going backwards to the first hidden layer. From these gradients, which can be interpreted as an indication of how each layer's output should change to reduce error, one can obtain the gradient on the parameters of each layer. The gradients on weights and biases can be immediately used as part of a stochastic gradient update (performing the update right after the gradients have been computed) or used with other gradient-based optimization methods.

After the forward computation, compute the gradient on the output layer:

$\boldsymbol{g} \leftarrow \nabla_{\hat{\boldsymbol{y}}} J = \nabla_{\hat{\boldsymbol{y}}} L(\hat{\boldsymbol{y}}, \boldsymbol{y})$

for $k = l, l-1, \ldots, 1$ **do**

Convert the gradient on the layer's output into a gradient on the pre-nonlinearity activation (element-wise multiplication if f is element-wise):

$\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{a}^{(k)}} J = \boldsymbol{g} \odot f'(\boldsymbol{a}^{(k)})$

Compute gradients on weights and biases (including the regularization term, where needed):

$\nabla_{\boldsymbol{b}^{(k)}} J = \boldsymbol{g} + \lambda \nabla_{\boldsymbol{b}^{(k)}} \Omega(\theta)$

$\nabla_{\boldsymbol{W}^{(k)}} J = \boldsymbol{g}\, \boldsymbol{h}^{(k-1)\top} + \lambda \nabla_{\boldsymbol{W}^{(k)}} \Omega(\theta)$

Propagate the gradients w.r.t. the next lower-level hidden layer's activations:

$\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{h}^{(k-1)}} J = \boldsymbol{W}^{(k)\top} \boldsymbol{g}$

end for

http://www.deeplearningbook.org/contents/mlp.html

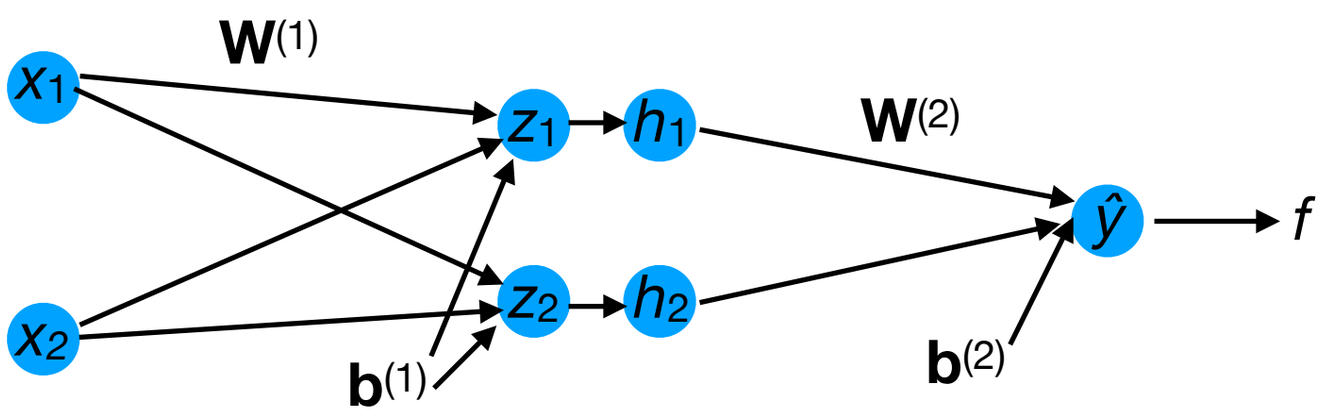
Computing the gradients

- Where do these come from?

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

where

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y}) \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top})
$$



Computing the gradients

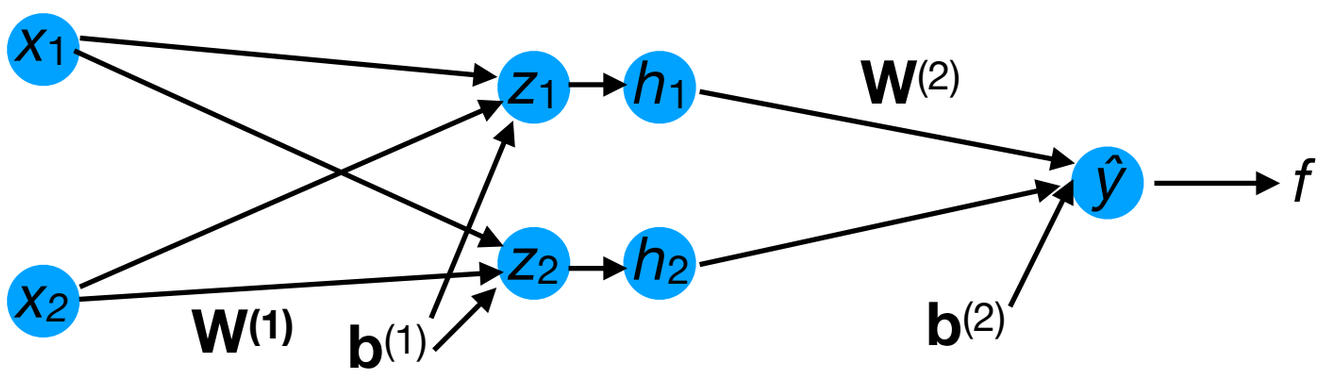
- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

- How does *f* depend on $\hat{y}$?

$$f(\mathbf{x}; \mathbf{W}^{(2)}) = \frac{1}{2}(\hat{y} - y)^2$$

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



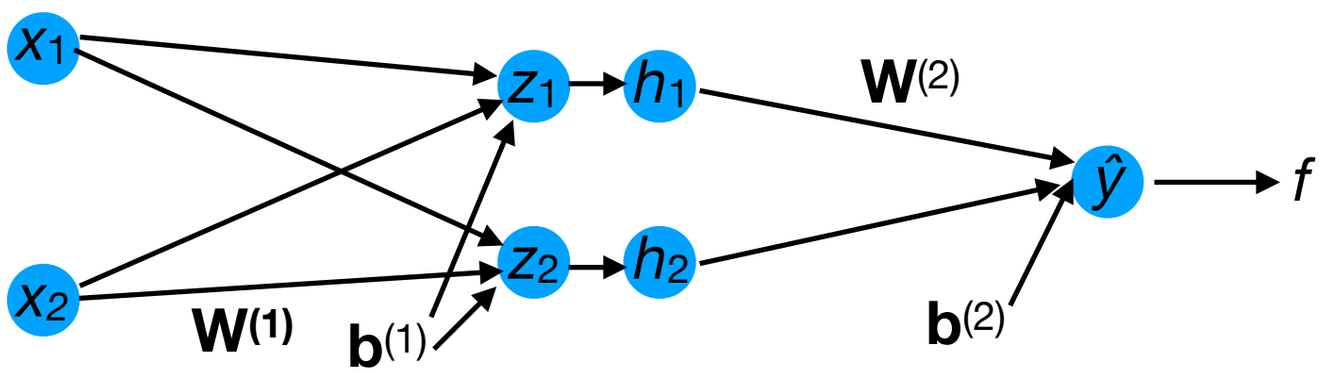
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$

- How does $\hat{y}$ depend on $\mathbf{W}^{(2)}$?

$$\begin{aligned} \hat{y} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\ &= W_1^{(2)} h_1 + W_2^{(2)} h_2 + b^{(2)} \\ \implies \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}} &= \begin{bmatrix} h_1 & h_2 \end{bmatrix} \\ &= \mathbf{h}^\top \end{aligned}$$



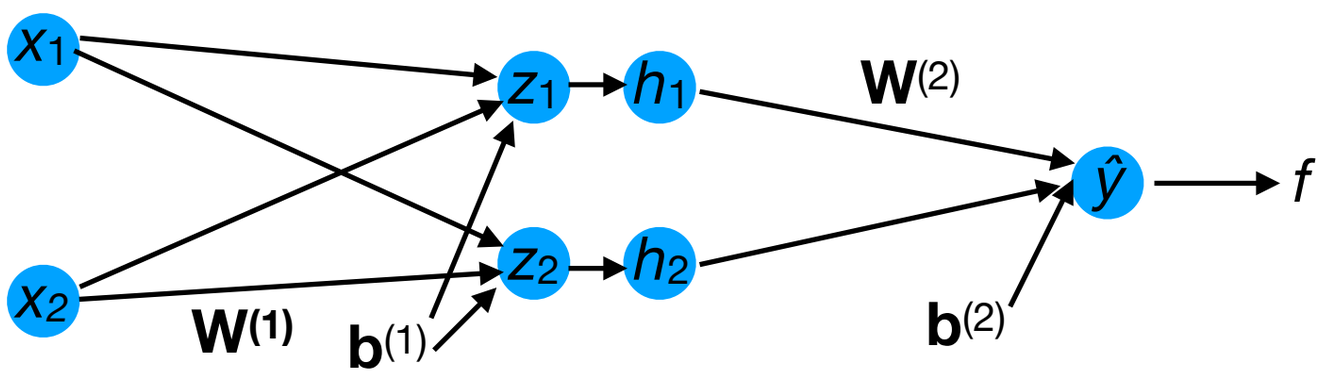
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}} = (\hat{y} - y)\mathbf{h}^\top$$

- How does $\hat{y}$ depend on $\mathbf{W}^{(2)}$?

$$\begin{aligned} \hat{y} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\ &= W_1^{(2)} h_1 + W_2^{(2)} h_2 + b^{(2)} \\ \implies \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}} &= \begin{bmatrix} h_1 & h_2 \end{bmatrix} \\ &= \mathbf{h}^\top \end{aligned}$$



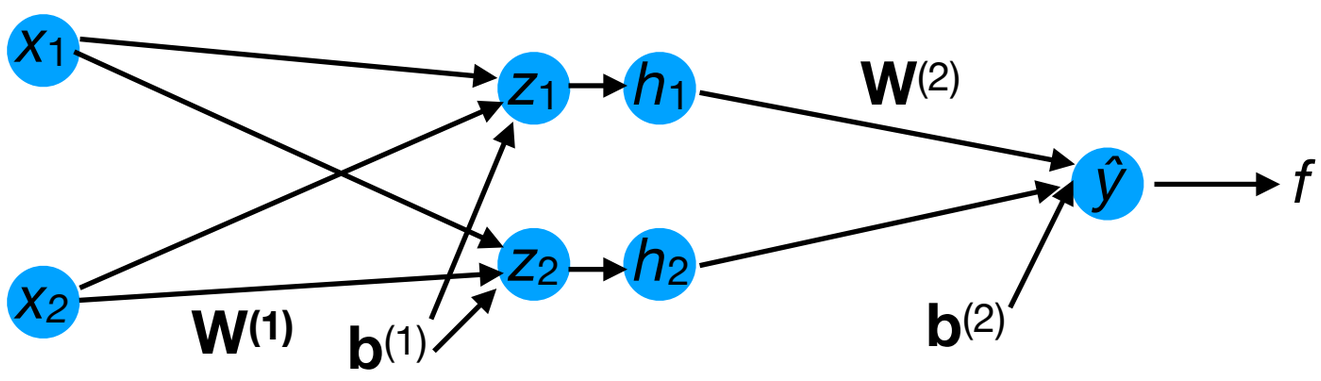
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$

- How does $\hat{y}$ depend on $\mathbf{b}^{(2)}$?

$$\begin{aligned} \hat{y} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\ &= W_1^{(2)} h_1 + W_2^{(2)} h_2 + b^{(2)} \\ \implies \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}} &= 1 \end{aligned}$$



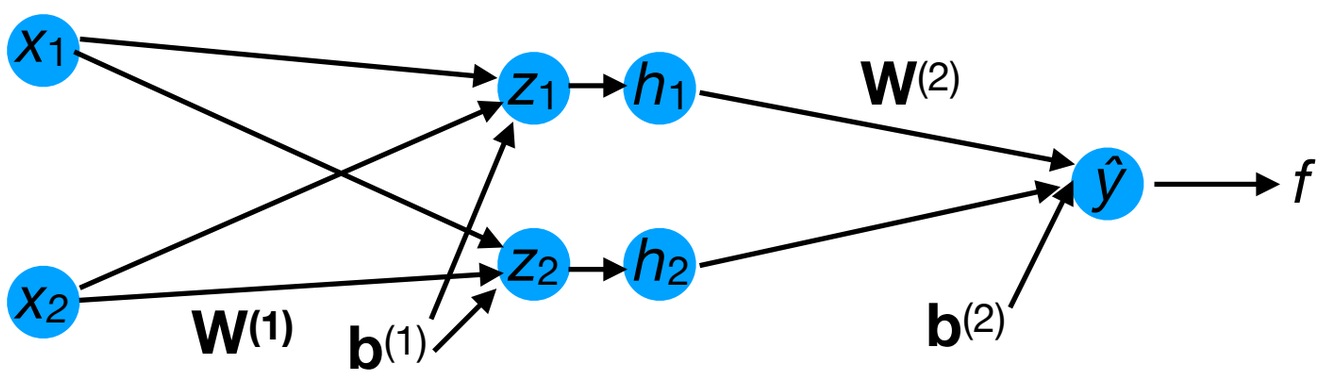
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}} = (\hat{y} - y)$$

- How does $\hat{y}$ depend on $\mathbf{b}^{(2)}$?

$$\begin{aligned} \hat{y} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\ &= W_1^{(2)} h_1 + W_2^{(2)} h_2 + b^{(2)} \\ \implies \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}} &= 1 \end{aligned}$$



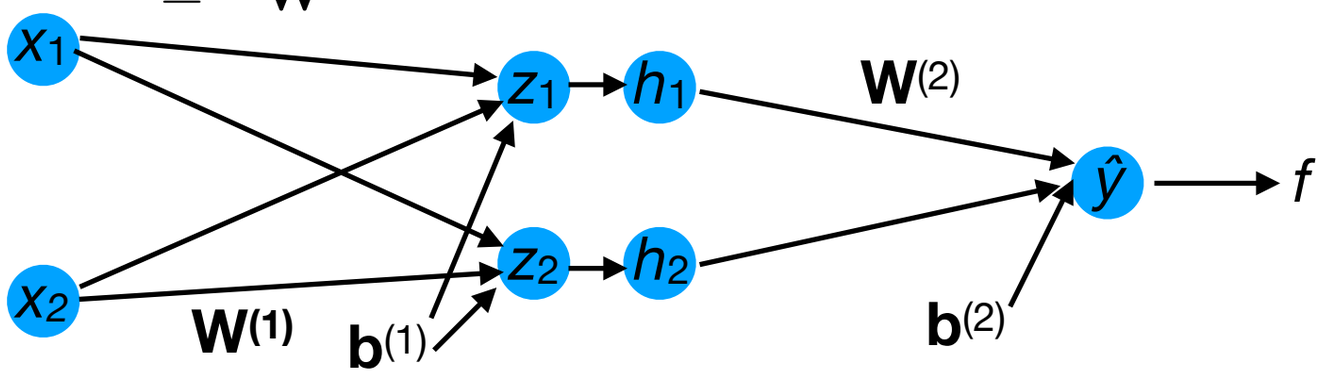
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

- How does $\hat{y}$ depend on **h**?

$$\begin{aligned} \hat{y} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\ &= W_1^{(2)} h_1 + W_2^{(2)} h_2 + b^{(2)} \\ \Longrightarrow \frac{\partial \hat{y}}{\partial \mathbf{h}} &= \begin{bmatrix} \frac{\partial \hat{y}}{\partial h_1} & \frac{\partial \hat{y}}{\partial h_2} \end{bmatrix} \\ &= \begin{bmatrix} W_1^{(2)} & W_2^{(2)} \end{bmatrix} \\ &= \mathbf{W}^{(2)} \end{aligned}$$



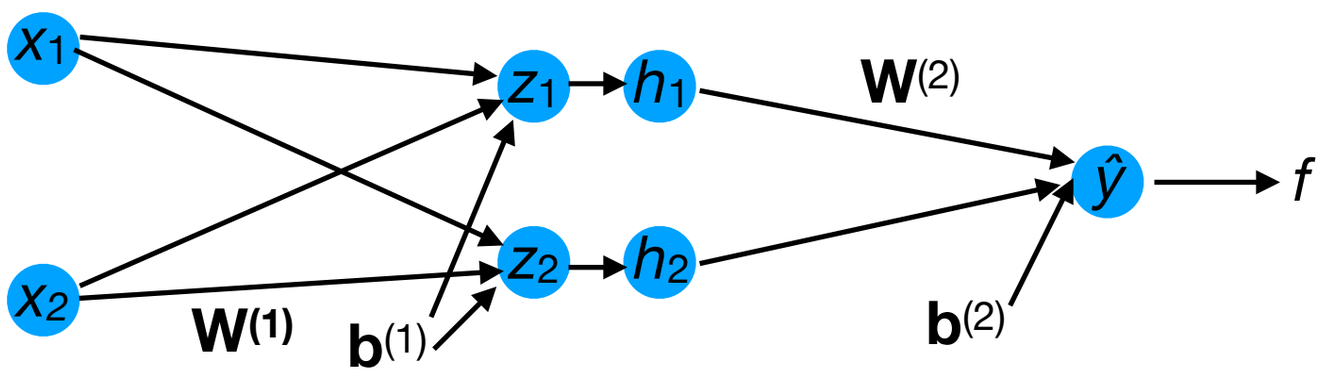
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

- How does **h** depend on **z**?

$$\mathbf{h} = \begin{bmatrix} \text{relu}(\mathbf{z}_1) \\ \text{relu}(\mathbf{z}_2) \end{bmatrix}$$



Computing the gradients

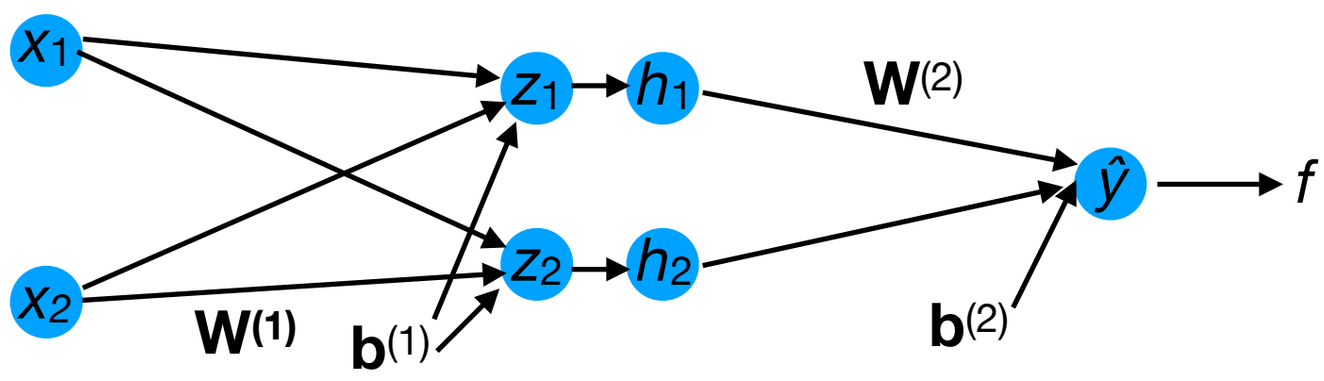
- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

- How does **h** depend on **z**?

$$\mathbf{h} = \begin{bmatrix} \text{relu}(\mathbf{z}_1) \\ \text{relu}(\mathbf{z}_2) \end{bmatrix}$$

$$\implies \frac{\partial \mathbf{h}}{\partial \mathbf{z}} = \begin{bmatrix} \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_2} \\ \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_2} \end{bmatrix}$$



Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

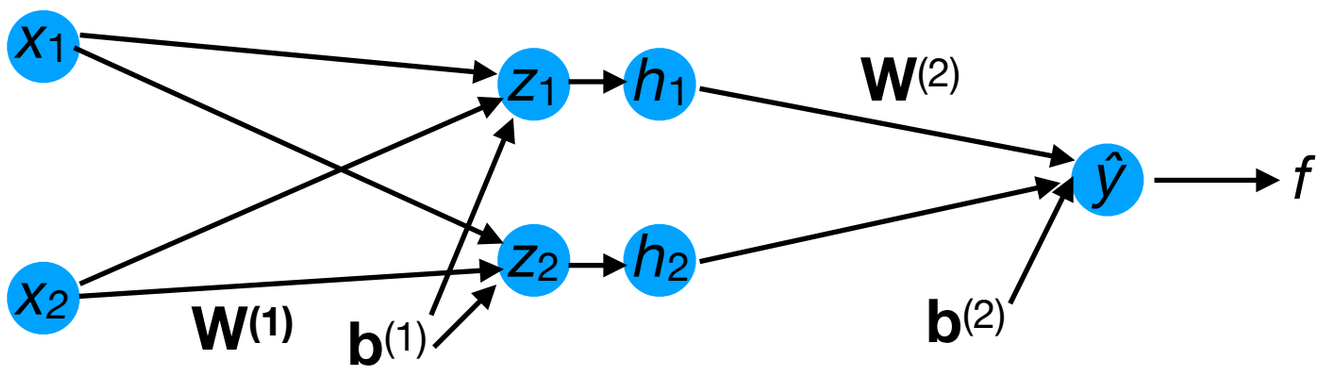
$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

- How does **h** depend on **z**?

$$\mathbf{h} = \begin{bmatrix} \text{relu}(\mathbf{z}_1) \\ \text{relu}(\mathbf{z}_2) \end{bmatrix}$$

$$\implies \frac{\partial \mathbf{h}}{\partial \mathbf{z}} = \begin{bmatrix} \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_2} \\ \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_2} \end{bmatrix}$$

$$= \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & 0 \\ 0 & \text{relu}'(\mathbf{z}_2) \end{bmatrix}$$



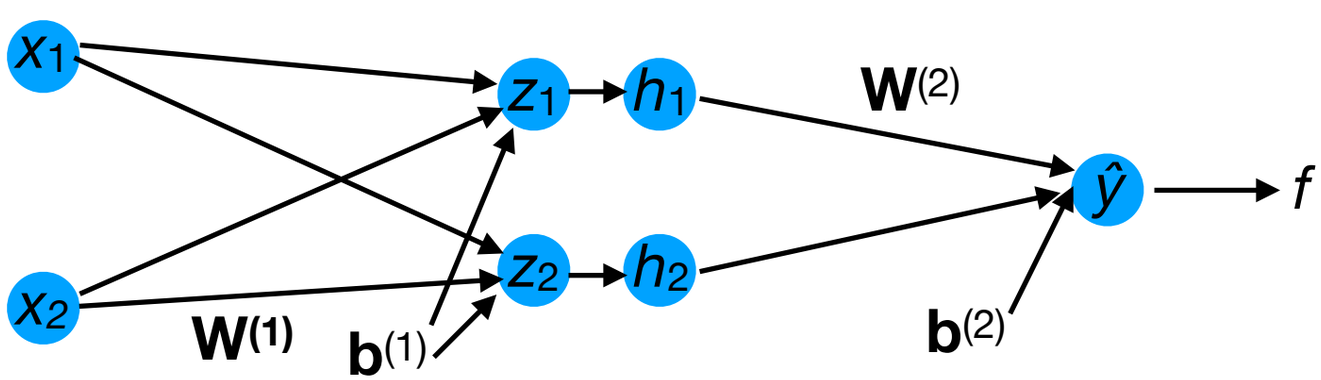
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

- How does **z** depend on $\mathbf{W}^{(1)}$?

$$\mathbf{z} = \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}$$



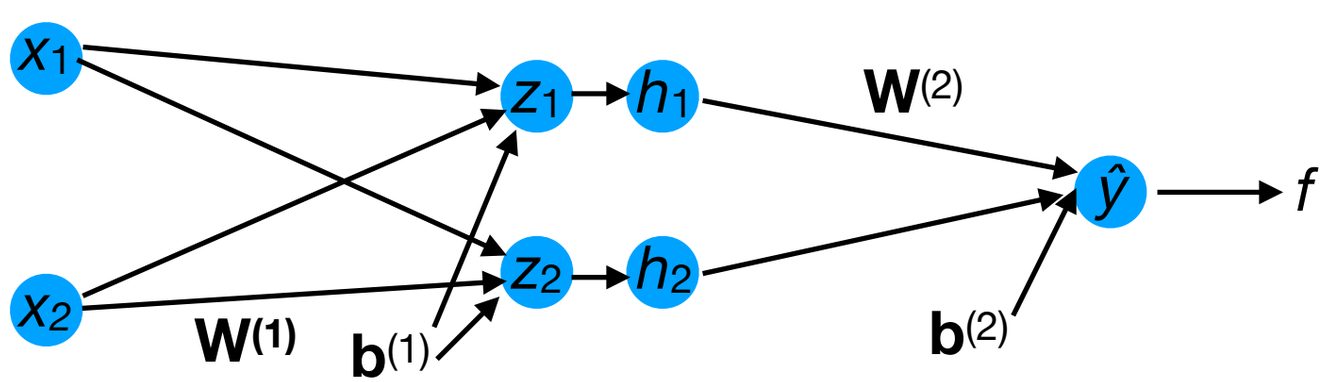
Computing the gradients

- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

- How does **z** depend on $\mathbf{W}^{(1)}$?

$$\begin{aligned} \mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\ \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} &= \begin{bmatrix} W_1^{(1)} & W_2^{(1)} \\ W_3^{(1)} & W_4^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix} \end{aligned}$$



Computing the gradients

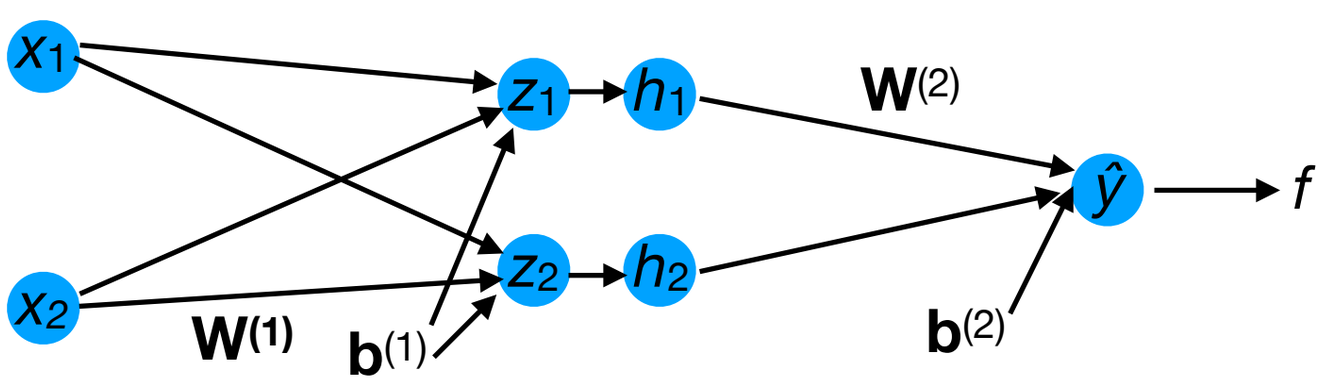
- Let's derive each gradient term in turn (for *n*=1):

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

- How does **z** depend on **W**(1)?

$$\begin{aligned}
\mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\
\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} &= \begin{bmatrix} W_1^{(1)} & W_2^{(1)} \\ W_3^{(1)} & W_4^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix} \\
\implies \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} &= \begin{bmatrix} \frac{\partial z_1}{\partial W_1^{(1)}} & \frac{\partial z_1}{\partial W_2^{(1)}} & \frac{\partial z_1}{\partial W_3^{(1)}} & \frac{\partial z_1}{\partial W_4^{(1)}} \\ \frac{\partial z_2}{\partial W_1^{(1)}} & \frac{\partial z_2}{\partial W_2^{(1)}} & \frac{\partial z_2}{\partial W_3^{(1)}} & \frac{\partial z_2}{\partial W_4^{(1)}} \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_1 & x_2 \end{bmatrix}
\end{aligned}$$

For Jacobian matrix, we have to treat W(1) as if it were a *vector*.



Analytical simplification

- We can now finally derive the gradient update for **W**(1):

$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$

Analytical simplification

- We can now finally derive the gradient update for $\mathbf{W}^{(1)}$:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
&= (\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & 0 \\ 0 & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix}
\end{aligned}
$$

Analytical simplification

- We can now finally derive the gradient update for $\mathbf{W}^{(1)}$:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
&= (\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & 0 \\ 0 & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \left(\left((\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \right) \odot \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \right) \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix}
\end{aligned}
$$

since multiplying by a diagonal matrix is equivalent to element-wise (Hadamard) product.

Analytical simplification

- We can now finally derive the gradient update for $\mathbf{W}^{(1)}$:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
&= (\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & 0 \\ 0 & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \left(\left((\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \right) \odot \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \right) \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix}
\end{aligned}
$$

To simplify notation, let's define a new vector that equals the first few terms.

Analytical simplification

- We can now finally derive the gradient update for $\mathbf{W}^{(1)}$:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
&= (\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & 0 \\ 0 & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \left(\left((\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \right) \odot \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \right) \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{g}_1 \mathbf{x}_1 & \mathbf{g}_1 \mathbf{x}_2 & \mathbf{g}_2 \mathbf{x}_1 & \mathbf{g}_2 \mathbf{x}_2 \end{bmatrix}
\end{aligned}
$$

Analytical simplification

- We can now finally derive the gradient update for $\mathbf{W}^{(1)}$:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
&= (\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & 0 \\ 0 & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \left(\left((\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \right) \odot \begin{bmatrix} \text{relu}'(\mathbf{z}_1) & \text{relu}'(\mathbf{z}_2) \end{bmatrix} \right) \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{g}_1 \mathbf{x}_1 & \mathbf{g}_1 \mathbf{x}_2 & \mathbf{g}_2 \mathbf{x}_1 & \mathbf{g}_2 \mathbf{x}_2 \end{bmatrix} \\
\implies \nabla_{\mathbf{W}^{(1)}} f &= \mathbf{g}\mathbf{x}^\top
\end{aligned}
$$

Outer product

Weight initialization

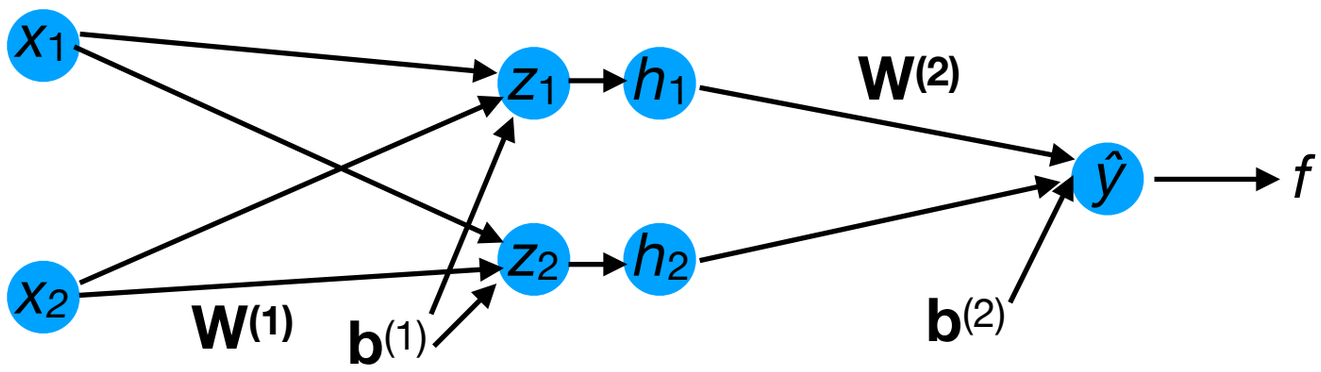
Weight initialization: example

- Suppose we initialize all the weights and bias terms of a 3-layer NN to be 0.

- What will happen during SGD?

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\text{CE}} &= (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\text{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\text{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\text{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \text{relu}'(\mathbf{z}^{(1)\top})
$$



Weight initialization: example

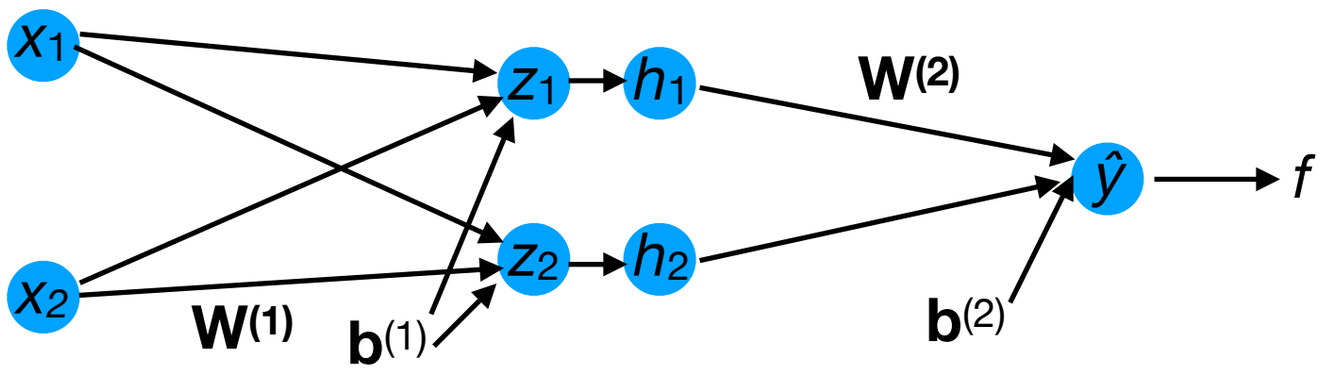
- Suppose we initialize all the weights and bias terms of a 3-layer NN to be 0.

- What will happen during SGD?

During forwards propagation, z and h will be 0. Hence, ŷ will also be 0.

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\text{CE}} &= (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\text{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\text{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\text{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \text{relu}'(\mathbf{z}^{(1)\top})
$$



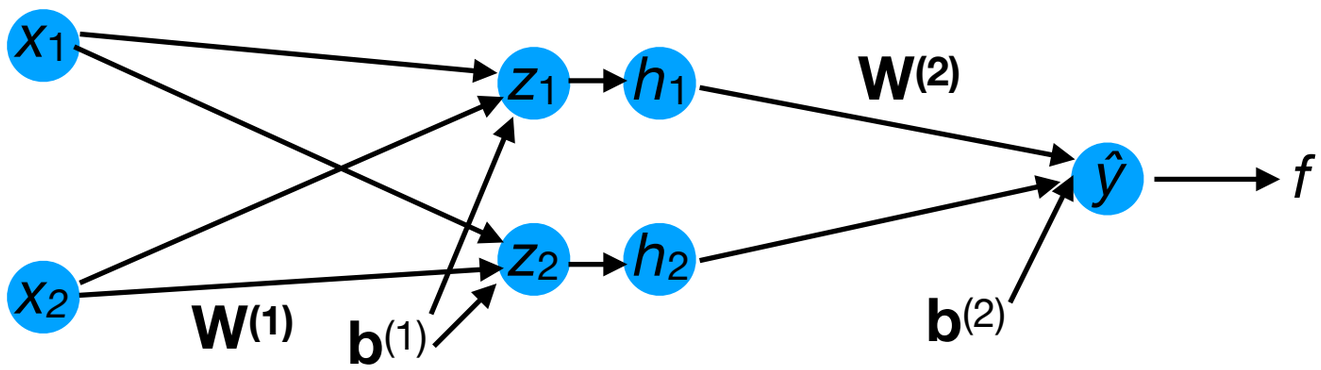
Weight initialization: example

- Suppose we initialize all the weights and bias terms of a 3-layer NN to be 0.

- What will happen during SGD?

During backwards propagation, we have:

$$\nabla_{\mathbf{W}^{(2)}} f_{\text{CE}} = (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^{(1)\top}$$
$$\nabla_{\mathbf{b}^{(2)}} f_{\text{CE}} = (\hat{\mathbf{y}} - \mathbf{y})$$
$$\nabla_{\mathbf{W}^{(1)}} f_{\text{CE}} = \mathbf{g}\mathbf{x}^{\top}$$
$$\nabla_{\mathbf{b}^{(1)}} f_{\text{CE}} = \mathbf{g}$$
$$\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top}\mathbf{W}^{(2)}\right) \odot \text{relu}'(\mathbf{z}^{(1)\top})$$



Weight initialization: example

- Suppose we initialize all the weights and bias terms of a 3-layer NN to be 0.

- What will happen during SGD?

During backwards propagation, we have:

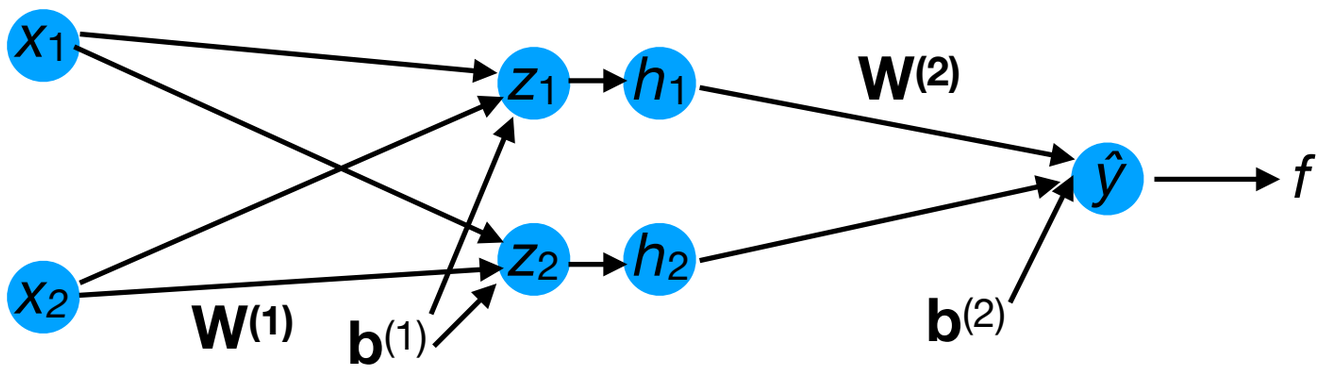
$$\nabla_{\mathbf{W}^{(2)}} f_{\text{CE}} = (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^{(1)\top} \; 0$$

$$\nabla_{\mathbf{b}^{(2)}} f_{\text{CE}} = (\hat{\mathbf{y}} - \mathbf{y})$$

$$\nabla_{\mathbf{W}^{(1)}} f_{\text{CE}} = \mathbf{g}\mathbf{x}^{\top} \; 0$$

$$\nabla_{\mathbf{b}^{(1)}} f_{\text{CE}} = \mathbf{g} \; 0$$

$$\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)}\right) \odot \text{relu}'(\mathbf{z}^{(1)\top}) \; 0$$



Weight initialization: example

- Because the gradients w.r.t. $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, and $\mathbf{b}^{(1)}$ are all 0, they will *never change*.

- Only $\mathbf{b}^{(2)}$ will change (to the mean of the target values *y*).

During backwards propagation, we have:

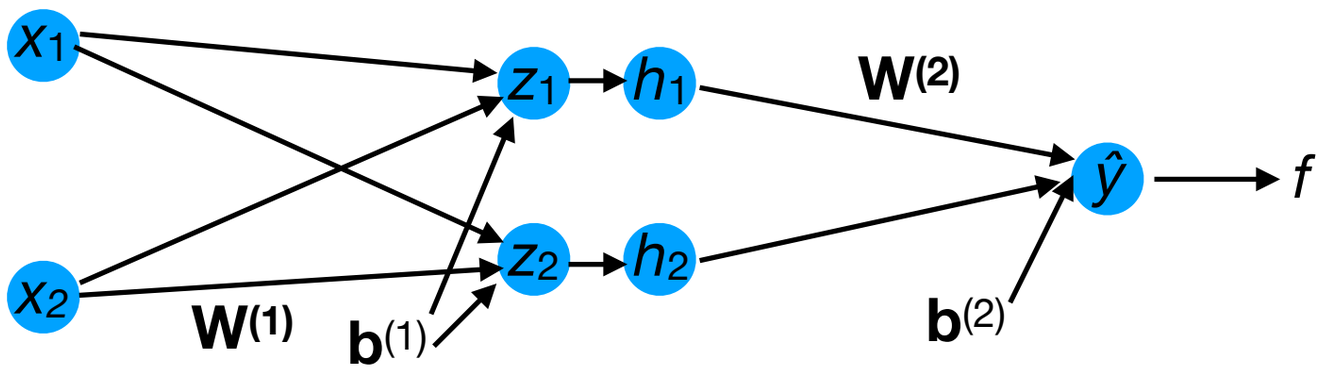
$$\nabla_{\mathbf{W}^{(2)}} f_{\text{CE}} = (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^{(1)\top} \; \mathbf{0}$$

$$\nabla_{\mathbf{b}^{(2)}} f_{\text{CE}} = (\hat{\mathbf{y}} - \mathbf{y})$$

$$\nabla_{\mathbf{W}^{(1)}} f_{\text{CE}} = \mathbf{g}\mathbf{x}^{\top} \; \mathbf{0}$$

$$\nabla_{\mathbf{b}^{(1)}} f_{\text{CE}} = \mathbf{g} \; \mathbf{0}$$

$$\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)}\right) \odot \text{relu}'(\mathbf{z}^{(1)\top}) \; \mathbf{0}$$

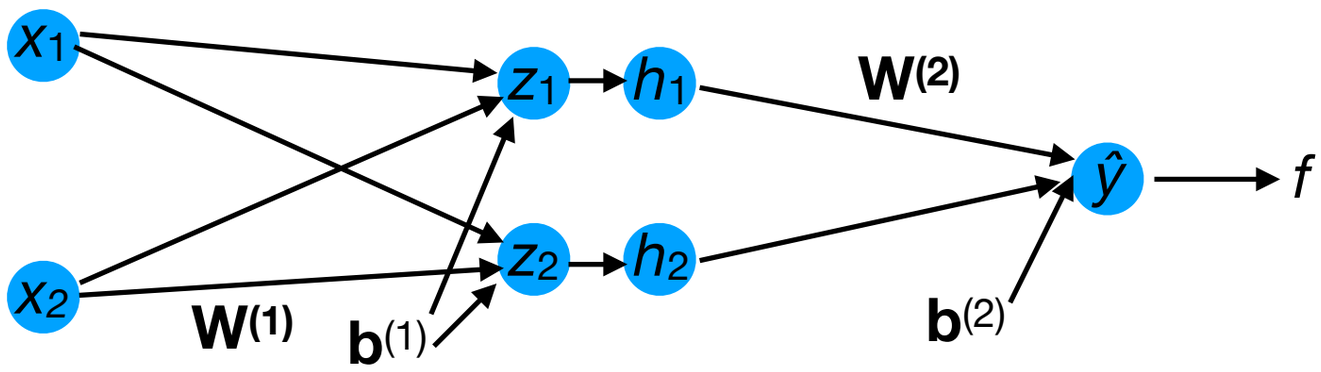


Weight initialization: exercise 1

- Suppose we initialize $\mathbf{W}^{(1)}=\mathbf{b}^{(1)}=0$, but $\mathbf{W}^{(2)}$, $\mathbf{b}^{(2)}$ are non-zero.
- Assume relu'(0)=0.
- What will happen during SGD?

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top})
$$

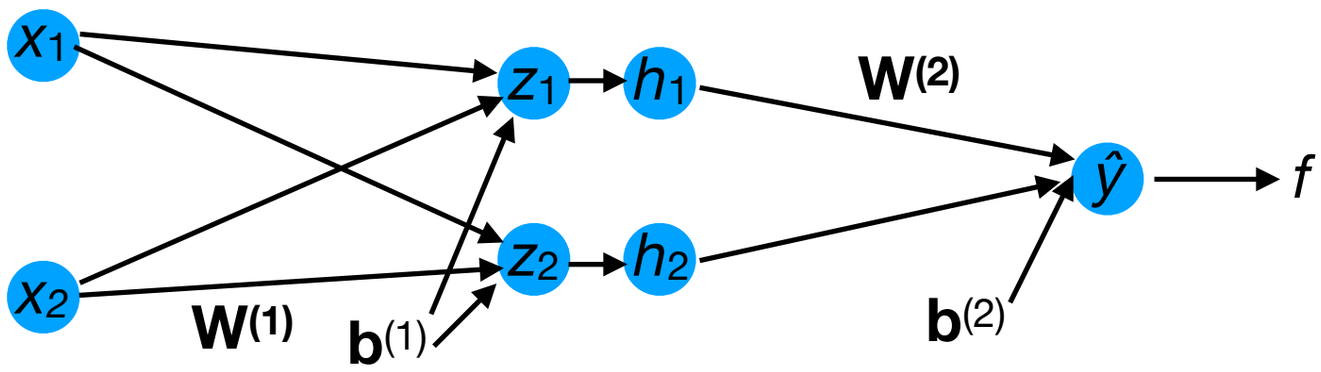


Weight initialization: exercise 1

- Since $\mathbf{W}^{(1)}=\mathbf{b}^{(1)}=0$, then $\mathbf{z}^{(1)}=\mathbf{h}^{(1)}=0$. Hence, $\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} = 0$.

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}){\mathbf{h}^{(1)}}^{\top} \; \mathbf{0} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'({\mathbf{z}^{(1)}}^{\top})
$$

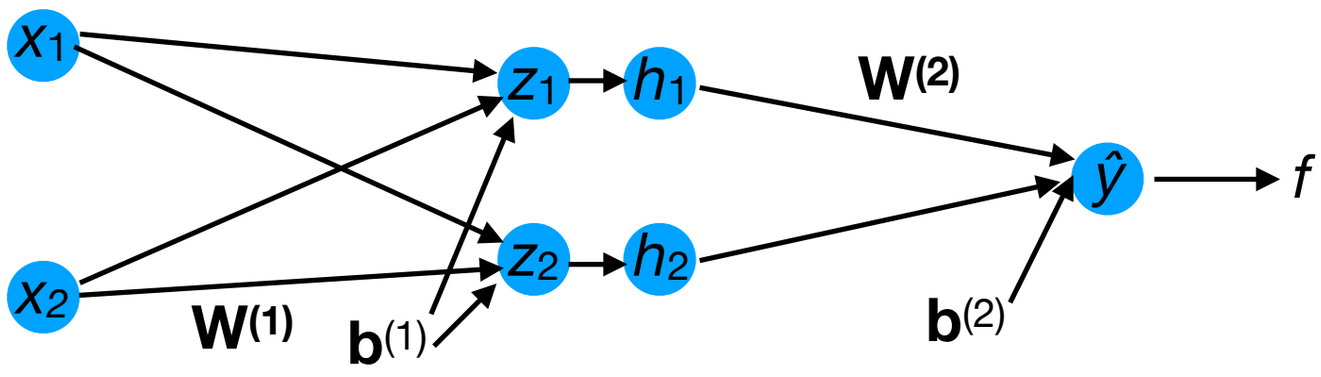


Weight initialization: exercise 1

- Since $\mathbf{W}^{(1)}$=$\mathbf{b}^{(1)}$=0, then $\mathbf{z}^{(1)}$=$\mathbf{h}^{(1)}$=0. Hence, $\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}}$= 0.
- Since relu'(0) = 0, then $\mathbf{g}$=0. Hence, gradients w.r.t. $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ are 0.
- Only $\mathbf{b}^{(2)}$ can change (so that $\hat{y}$ approaches mean of y).

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^{(1)\top} \; \mathbf{0} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}\mathbf{x}^{\top} \; \mathbf{0} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g} \; \mathbf{0}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top}) \; \mathbf{0}
$$

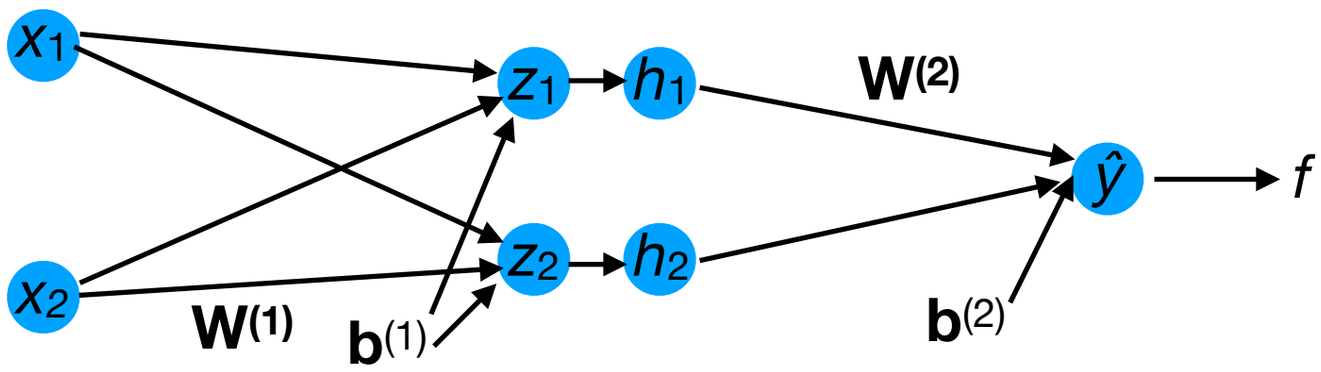


Weight initialization: exercise 2

- Suppose we initialize $\mathbf{W}^{(2)}=\mathbf{b}^{(2)}=0$, but $\mathbf{W}^{(1)}$, $\mathbf{b}^{(1)}$ are non-zero.
- What will happen during SGD?

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}}-\mathbf{y})\mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}}-\mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}}-\mathbf{y})^{\top}\mathbf{W}^{(2)}\right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top})
$$

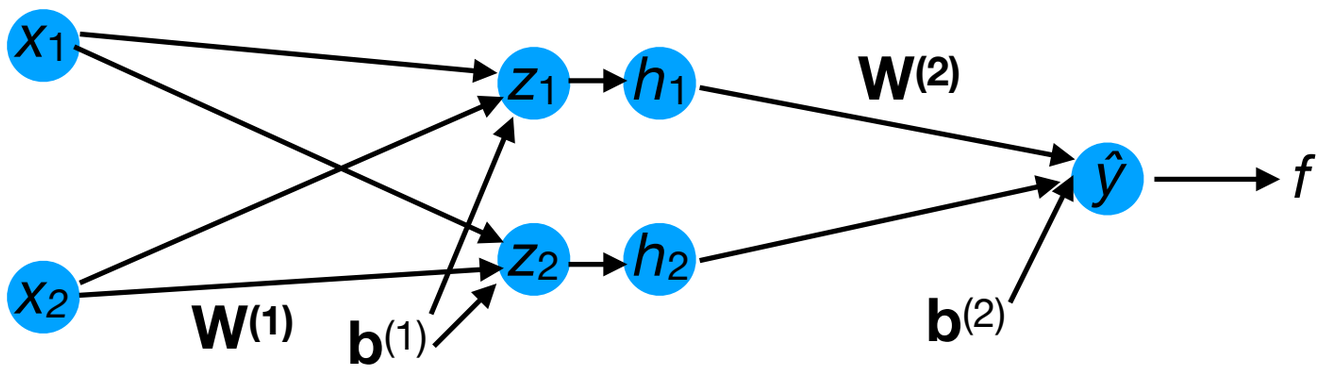


Weight initialization: exercise 2

- Since $\mathbf{W}^{(2)}=0$, then $\mathbf{g}=0$. Hence, $\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}}, \nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} = 0$.

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top})
$$

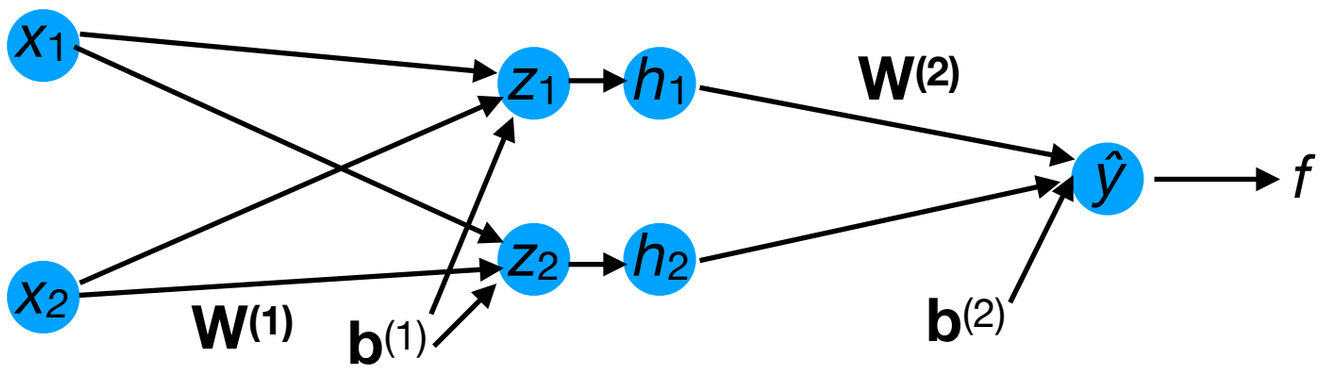


Weight initialization: exercise 2

- Since $\mathbf{W}^{(2)}$=0, then $\mathbf{g}$=0. Hence, $\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}}, \nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} = 0$.
- However, $\mathbf{h}$ is non-zero. Hence, $\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}}$ is nonzero => $\mathbf{W}^{(2)}$ will change.

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top})
$$

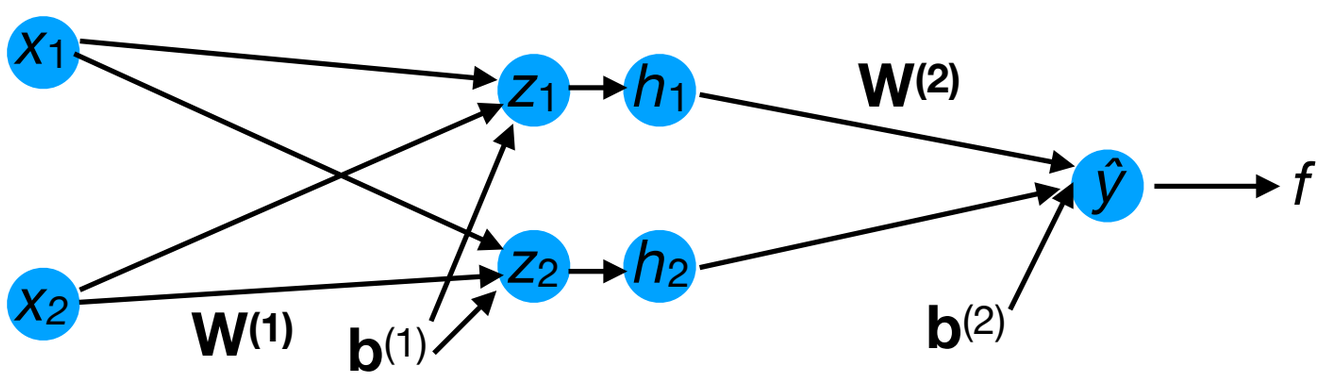


Weight initialization: exercise 2

- Since **W**(2)=0, then **g**=0. Hence, $\nabla_{\mathbf{W}^{(1)}} f_{\text{CE}}, \nabla_{\mathbf{b}^{(1)}} f_{\text{CE}} = 0$.
- However, **h** is non-zero. Hence, $\nabla_{\mathbf{W}^{(2)}} f_{\text{CE}}$ is nonzero => $\mathbf{W}^{(2)}$ will change.
- During the next gradient update, **g** is non-zero => **W**(1), **b**(1) will change.
- In summary: this initialization does not severely inhibit the network's performance (though initializing **W**(2), **b**(2) to 0 is still not recommended).

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\text{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\text{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\text{CE}} &= \mathbf{g}\mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\text{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \text{relu}'(\mathbf{z}^{(1)\top})
$$



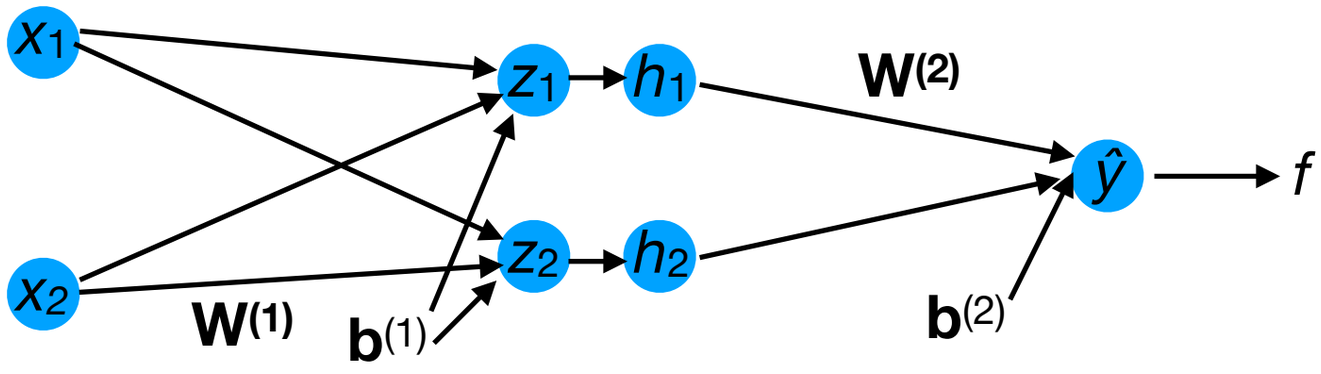
Weight initialization: exercise 3

- Suppose that each weight matrix & bias vector consists of the same row *repeated many times*.

- What will happen during SGD?

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g} \mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top})
$$

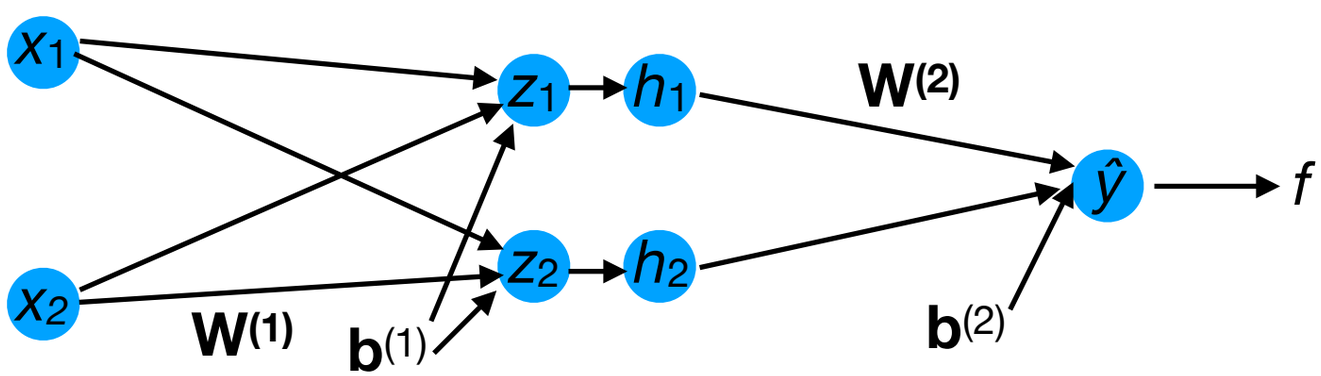


Weight initialization: exercise 3

- In this case, every node of **h** (and **z**) has the *same value*.
- Therefore, the gradient update to each row of $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ has the *same value*.
- The NN is performing redundant computation — although it has 2 hidden units, it might as well just have 1!

$$
\begin{aligned}
\nabla_{\mathbf{W}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \mathbf{h}^{(1)\top} \\
\nabla_{\mathbf{b}^{(2)}} f_{\mathrm{CE}} &= (\hat{\mathbf{y}} - \mathbf{y}) \\
\nabla_{\mathbf{W}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g} \mathbf{x}^{\top} \\
\nabla_{\mathbf{b}^{(1)}} f_{\mathrm{CE}} &= \mathbf{g}
\end{aligned}
$$

$$
\mathbf{g}^{\top} = \left((\hat{\mathbf{y}} - \mathbf{y})^{\top} \mathbf{W}^{(2)} \right) \odot \mathrm{relu}'(\mathbf{z}^{(1)\top})
$$



Weight initialization methods

- There are various different methods of initializing the weights of a neural network.

- One common approach:
  - For weight matrix $\mathbf{W}^{(j)}$, sample each component from a 0-mean Gaussian with deviation $1/\sqrt{\text{cols}(\mathbf{W}^{(j)})}$.
    - Within certain NNs, helps to ensure that the gradients are usually non-zero.

Weight initialization methods

- There are various different methods of initializing the weights of a neural network.

- One common approach:
  - For weight matrix $\mathbf{W}^{(j)}$, sample each component from a 0-mean Gaussian with deviation $1/\sqrt{\text{cols}(\mathbf{W}^{(j)})}$.
    - Within certain NNs, helps to ensure that the gradients are usually non-zero.
  - *Optional*: orthogonalize the rows of $\mathbf{W}^{(j)}$ to reduce correlation between different units of the pre-activation layer $\mathbf{z}^{(j)}$.

Why “deep” learning?

Why “deep” learning?

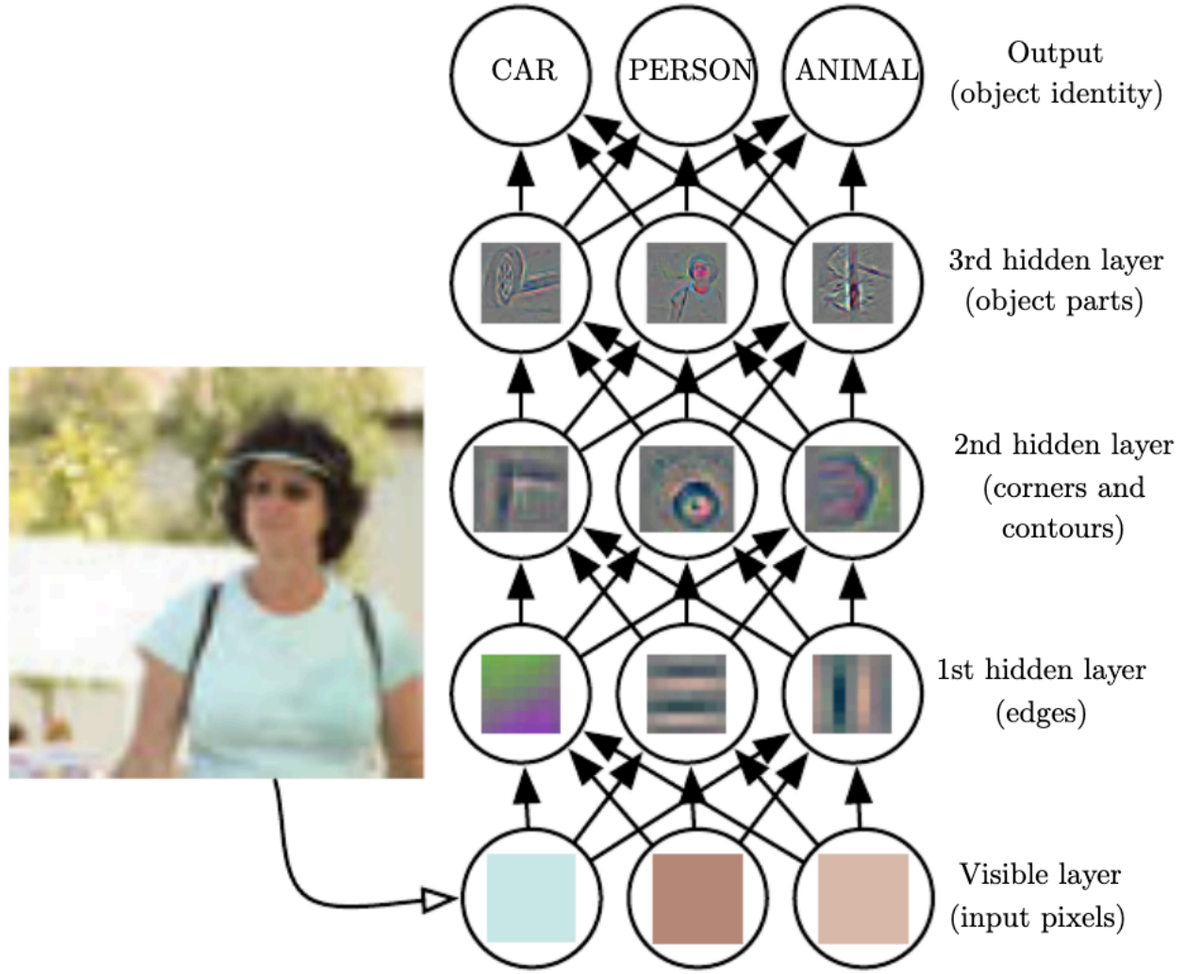
- One of the reasons for the resurgence of NNs around 2010 is that increasing the depth is very powerful.

- There is some theory, and an abundance of empirical results, that deeper networks are more powerful (can represent more complicated functions) and more accurate than shallow networks.

- Why?

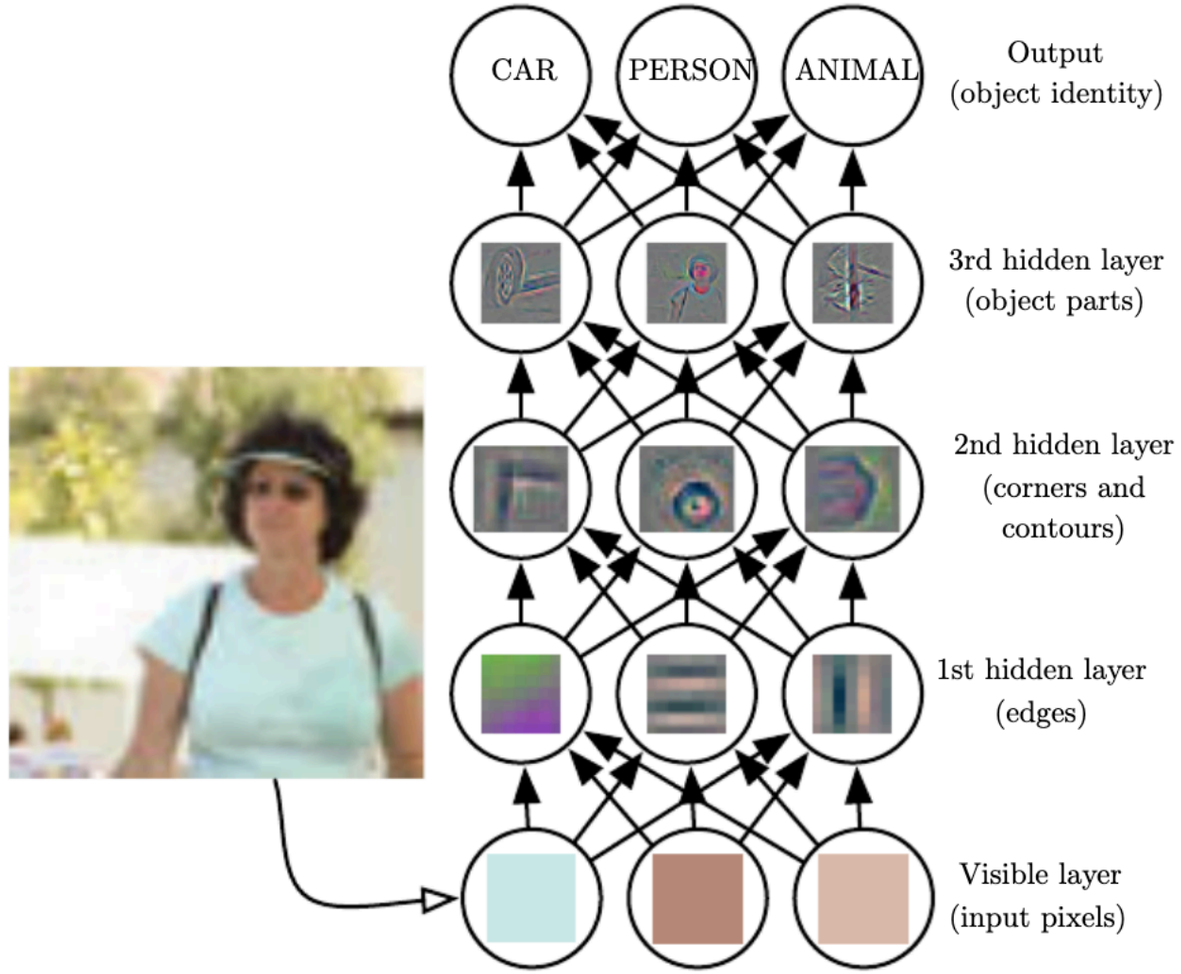
Why “deep” learning?

- One explanation for the success of deep NNs is that each layer can represent **increasingly abstract representations** of the input (from Zeiler & Fergus 2014):



Why “deep” learning?

- The hidden units can represent the content of the input in a compact way.



Demo

- As a simple illustration, let's consider two feed-forward NNs with the **same number of weights (60)**:
  - 3 layers with dimensions: (2, 20, 1)
  - 6 layers with dimensions: (2, 4, 4, 4, 4, 1)
- If we initialize their weights randomly, and use a non-linear activation function (`abs` in this case), then what kinds of functions can they represent?

Difficulty in deep learning

- Training very deep neural networks has traditionally been very difficult for two main reasons:
  - **Vanishing gradients**: gradients of weights tends to 0 as depth increases. => no learning
  - **Exploding gradient**: gradients of weights tends to infinity as depth increases. => requires small learning rate.

Difficulty in deep learning

- Since ~2008, DL researchers have:
  - Harnessed faster hardware to achieve good accuracy more quickly.
  - Found better architectures and training algorithms that prevent the gradients from behaving “badly”.

Practical suggestions

Tackle the problem based on your past experience

- Think of the machine learning problem you have solved that is *most similar* to the one you are tackling now, e.g.:
  - Same task (object detection, semantic segmentation, speaker diarization, etc).
  - Same dataset (MNIST, ImageNet, Faces in the Wild, etc.)
- How did you solve it?
  - NN design
  - Training procedures
  - Hyperparameters
- Start with the approach you used before, make sensible adjustments, and start training.

Start small

- Debug your code on a smaller version of the problem:
  - Subset of data (e.g., just 10K images instead of 50K)
  - Subset of classes (e.g., just 2 MNIST classes instead of 10)

- Advantages:
  - Less time for initialization.
  - Less time for training.
  - Fewer variables to examine during forward/back-propagation

Start simple

- Until you gain confidence & experience, train a simple model first:
    - They’re often faster to train and easier to debug than more powerful models.

Start simple

- Until you gain confidence & experience, train a simple model first:
    - They’re often faster to train and easier to debug than more powerful models.
- Make sure your model’s accuracy is above chance:
    - Take the prior class probabilities into account! (If the classes are 90/10, then the baseline rate for guessing the majority class is 90%.)
    - Make sure your model is not always predicting the dominant class.

But not too simple

- Sometimes a complex neural network is necessary to solve a problem with high accuracy.

- For instance, a 3-layer NN may not have enough representational power to analyze complex images.

- You might need to use a standard architecture (e.g., ResNet) — but use one of the simpler versions (e.g., 50 instead of 100 layers).

Start small & simple

- Try to find a model (and hyper-parameters) whose training loss decreases *smoothly*.

- Afterwards, increase the size of the training set and model complexity.

Regularization

- If there is a large divergence between training accuracy and testing accuracy (i.e., overfitting), then try regularizing the model:
  - Increasing L_1, L_2 regularization strength.
  - Adding/increasing dropout (for NNs).
  - Reducing number of training epochs (for NNs).
  - Synthesizing more training examples with label-preserving transformations (geometric & noise-based).

Hyper-parameter optimization

- Try a variety of hyper-parameters:
  - Find a few values manually that seem to work; use these as a guide to pick a reasonable range (e.g., for learning rate, 1e-5 to 1e0, spaced logarithmically).
  - Be disciplined about optimizing parameters on a validation set, not the test set!

Hyper-parameter optimization

- Try a variety of hyper-parameters:
  - When you have intuition, then it's sometimes worthwhile to watch the loss evolve over time.
  - When you do not have intuition, then automate the process and get on with your life — do not succumb to the temptation to watch learning curves like a movie.

Normalization

- It can be helpful to put every feature onto the same scale.

- In particular, the scale can interact with the L_2 regularization strength.

Normalization: example

- Suppose you are predicting tomorrow's temperature based on (1) today's temperature and (2) wind speed.

- Suppose we measure temperature in Kelvin and wind speed in km/h.

- Suppose the optimal weights w_1, w_2 for these two features, for L_2-regularized linear regression, are 1 and 2, i.e.:

  - $\hat{y} = w_1 t + w_2 s$ (t = today's temp, s = today's wind speed)
    $\hat{y} = 1*t + 2*s$

Normalization: example

- Now, suppose we change the units for wind speed from km/h to m/s.
  - E.g., 18 km/h = 5 m/s **Numerical values reduced by 3.6x**

- If we don't adjust our model weights w_1, w_2, then our predictions will be wrong:

  - $\hat{y} = 1^*t + 2^*s$
    $\hat{y}(4, 18) = 4 + 36 = 40$ **km/h**
    $\hat{y}(4, 5) = 4 + 10 = 14$ **m/s**

Normalization: example

- Because the numerical values of the wind speed were reduced by factor of 3.6, the corresponding weight w_2 must compensate by increasing by 3.6x, i.e.:
  - $\hat{y} = w_1 t + \tilde{w}_2 s$ (t = today's temp, s = today's wind speed)
    $\hat{y} = 1{*}t + 3.6{*}2{*}s$

- Without regularization, the training procedure (e.g., minimize f_{MSE}) will account for the change-of-scale seamlessly, i.e.:

$$\arg\min_{\tilde{w}_2} f_{\mathrm{MSE}}^{\mathrm{m/s}}(\cdot) \quad \textbf{= 3.6 *} \quad \arg\min_{w_2} f_{\mathrm{MSE}}^{\mathrm{km/h}}(\cdot)$$

Normalization: example

- But with L_2 regularization, the issue is more complicated:

$$\arg\min_{\tilde{w}_2} \left[f_{\text{MSE}}^{\text{km/h}}(\cdot) + \frac{1}{2} w_2^2 \right]$$

- The regularization term "discourages" w_2 from growing too big:
  - When we rescale from km/h to m/s, the L_2 term prevents the weight w_2 from compensating exactly.

Normalization: recommendations

- For features in a finite range, try rescaling to [0,1] or [-1,1].
- For features in infinite range, try subtracting the mean and dividing by standard deviation (so that the distribution has zero-mean and unit standard deviation).

Data augmentation

Data augmentation

- The more training data you have, the less is the risk of overfitting.
- Unfortunately, training data are often hard to find.
- Can we synthesize new training examples automatically?

Data augmentation

- **Data augmentation** is the creation of new examples based on existing ones.

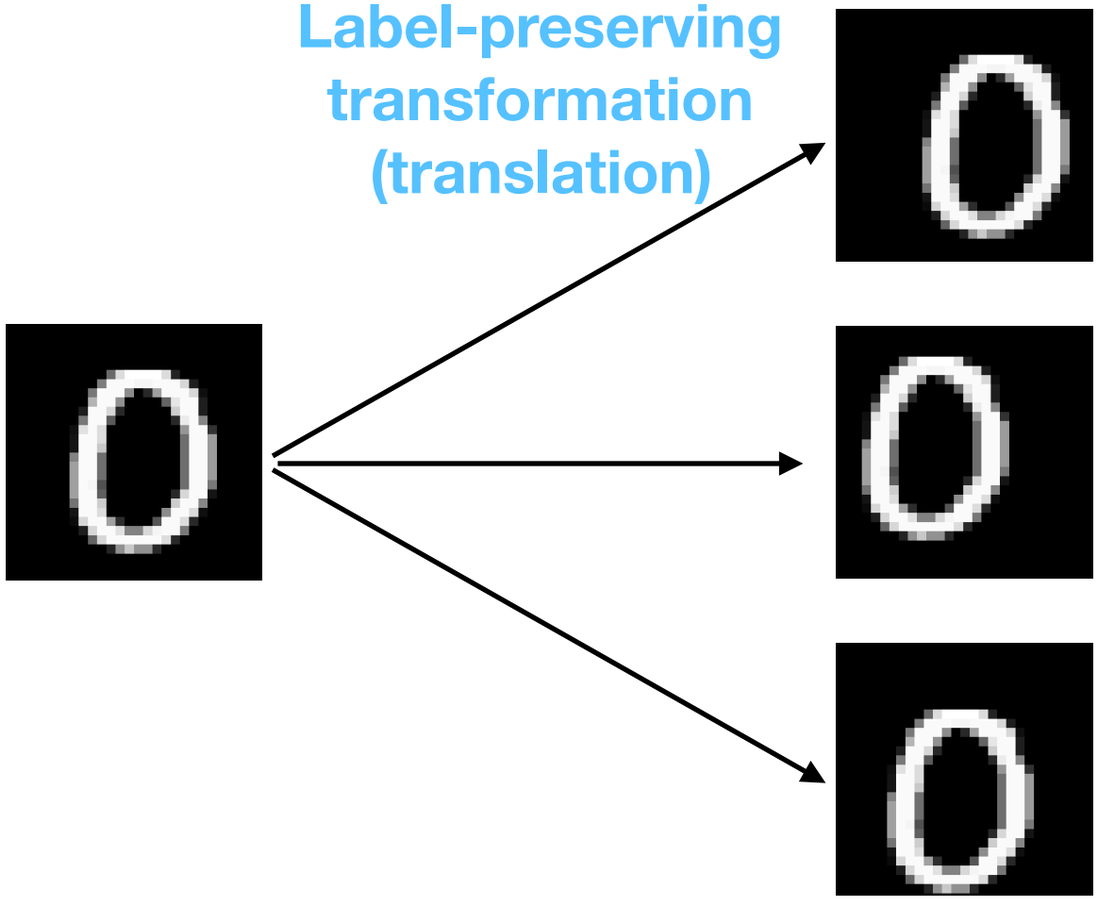
- If we can alter an existing training example without affecting its associated label, then we can generate many new training examples and train on them.

Data augmentation

- Several commonly used methods of data augmentation:
  - Adding noise to existing examples (e.g., Gaussian, Laplacian).
  - Geometric transformations (e.g., flip left/right, rotate, translate).

Example: translation

- From an existing MNIST image, translate all the pixels by some random amount (dx, dy).



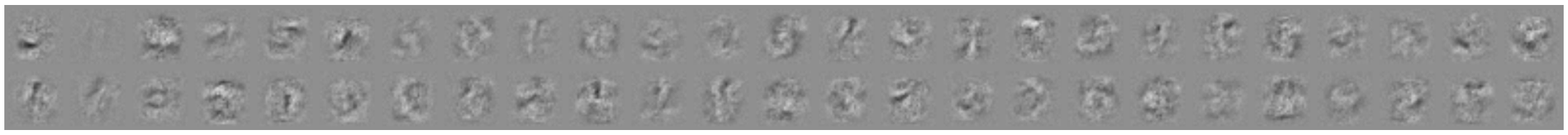
Example: translation

- Data augmentation via translation encourages the NN to learn **translation-invariant** features — they are useful for classification no matter where in the image they occur.

Example: translation

- Here are the weights $\mathbf{W}^{(1)}$ (transformed to 100x28x28) of a MNIST classification network **without** data augmentation:

Acc=98.07



Example: translation

- Here are the weights $\mathbf{W}^{(1)}$ (transformed to 100x28x28) of a MNIST classification network **with** data augmentation:

Acc=98.44

- Compared to the previously shown weights, these show visually more well-defined contours.

(Unsupervised)
pre-training

Feature representations

- One of the reasons why NNs are so powerful is that they can learn **feature representations** of the raw input data.

- Classifying/regressing the target variable $\hat{y}$ is often easier once the raw data have been transformed into a different feature space.

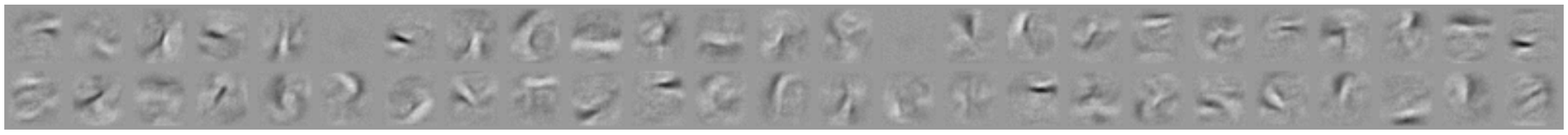
  - We saw this with XOR.

Feature representations

- One of the reasons why NNs are so powerful is that they can learn **feature representations** of the raw input data.

- Classifying/regressing the target variable $\hat{y}$ is often easier once the raw data have been transformed into a different feature space.

  - With MNIST, the NN seemed to recognize brush-strokes:



Unlabeled data

- In some application domains (e.g., object recognition/ detection in images), collecting *labeled* data is hard, but collecting *unlabeled* data is easy.

- How might abundant unlabeled data help us to train better ML models?

Learning good features

- We can harness unsupervised learning algorithms to learn good feature representations from unlabeled data.

  - **Unsupervised**: examples without labels.

Learning good features

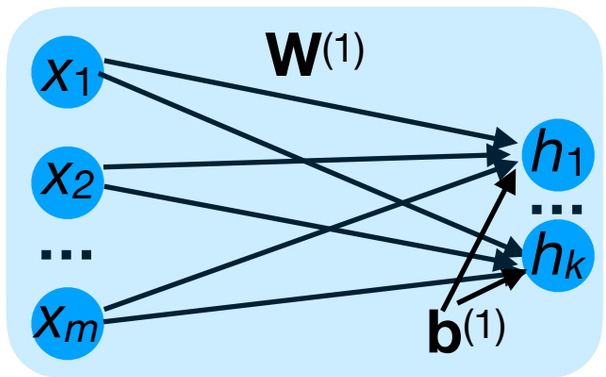
- We can harness unsupervised learning algorithms to learn good feature representations from unlabeled data.
    - **Unsupervised**: examples without labels.

- Key intuition: a good representation captures the essence of the raw input data.
    - We can “compress” the data into a smaller representation.
    - We can “uncompress” it to *reconstruct* the original data.

Auto-encoders

- Let **x** be the raw input data.
- Let **h** be the hidden representation that captures the "essence" of **x**. We say **h** has been **encoded** from **x**.
- We can compute **h** using a neural network (just 1 layer in this example, but could be deeper).

$$\mathbf{h} = \sigma^{(1)}\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right)$$

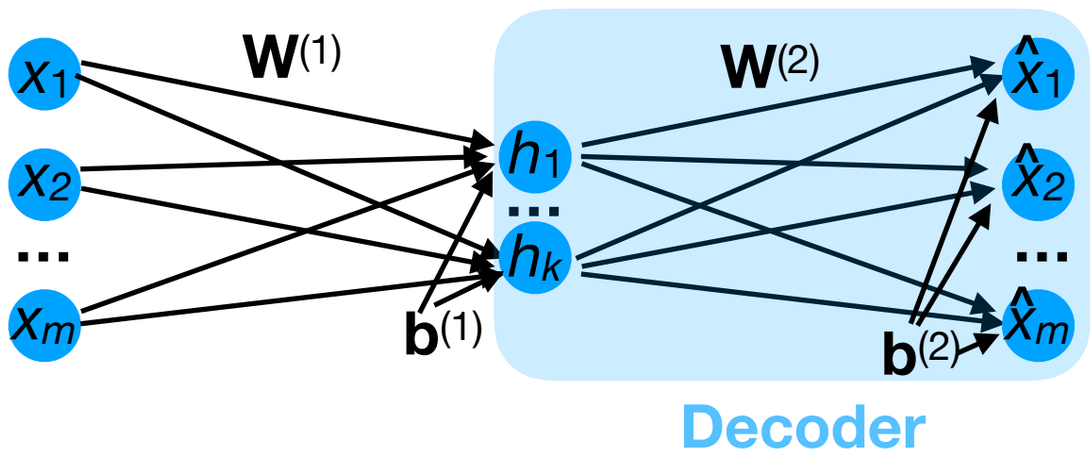


Auto-encoders

- If **h** contains the essential features of **x**, then we can use **h** to reconstruct (approximate) the original data **x**.

- Let $\hat{\mathbf{x}}$ denote our reconstruction of **x**. We say $\hat{\mathbf{x}}$ has been **decoded** from **h**.

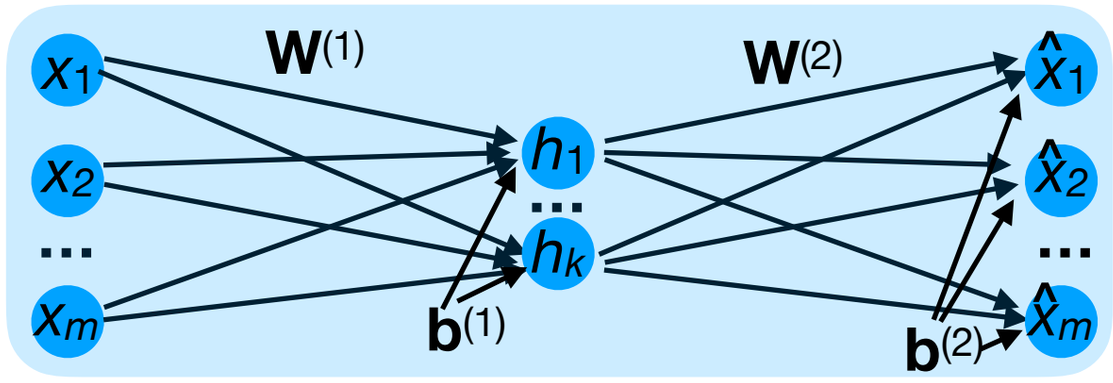
$$\hat{\mathbf{x}} = \sigma^{(2)}\left(\mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}\right)$$



Auto-encoders

- Putting the two components (encoder+decoder) together, we arrive at an **auto-encoder**.

$$\hat{\mathbf{x}} = \sigma^{(2)}\left(\mathbf{W}^{(2)}\sigma^{(1)}\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right) + \mathbf{b}^{(2)}\right)$$



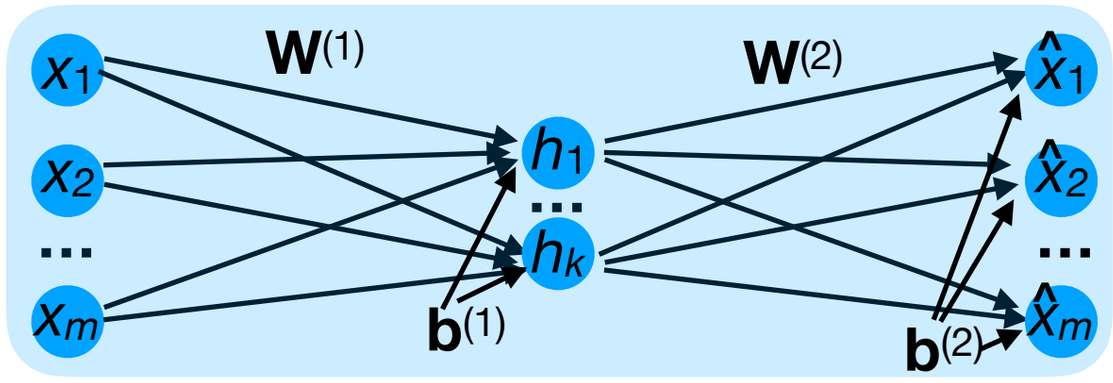
Auto-encoder

Auto-encoders: training loss function

- With auto-encoders, we optimize $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$ to make our reconstructions as accurate as possible, i.e., minimize:

$$f_{\text{MSE}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)})^\top(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)})$$



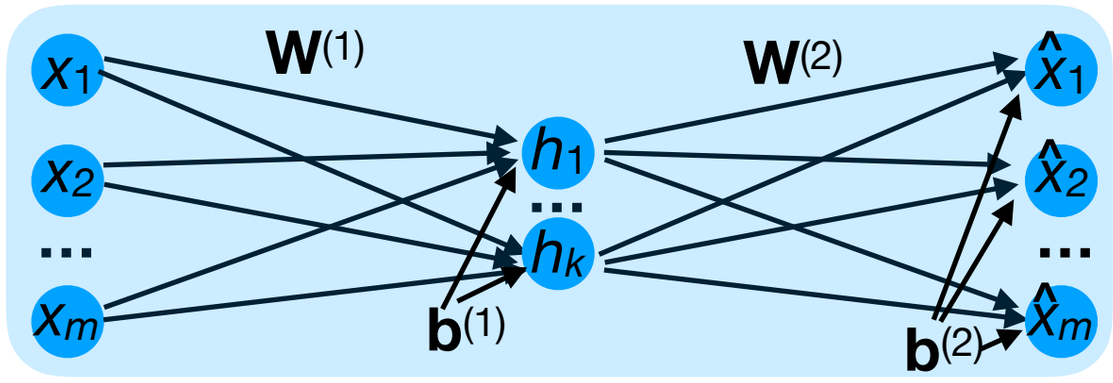
Auto-encoder

Auto-encoders: training loss function

- Notice that this loss function does not require any training labels — there is no mention of any *y*!

$$f_{\text{MSE}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)})^2$$

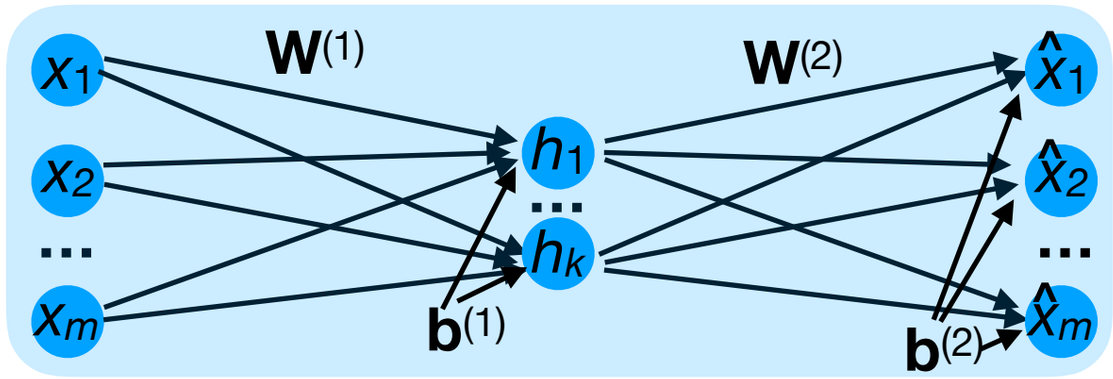
$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)})^\top(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)})$$



Auto-encoder

Bottleneck

- If we let $k=m$, then what is an easy (but useless) way to set the weights to give a perfect reconstruction (0 MSE)?

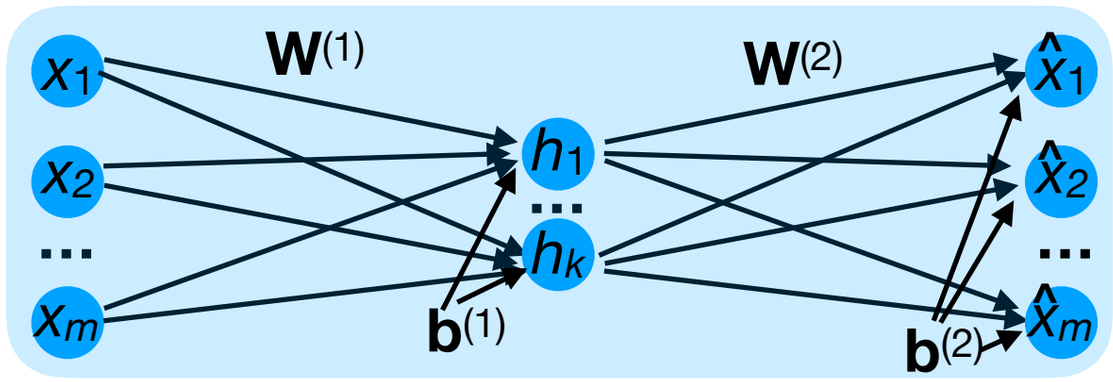


Auto-encoder

Bottleneck

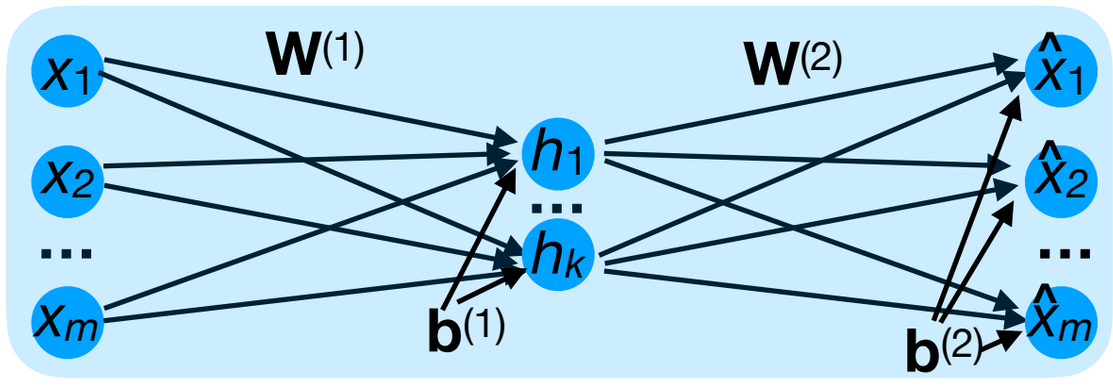
- If we let k=m, then what is an easy (but useless) way to set the weights to give a perfect reconstruction (0 MSE)?

- If $k = m$, then we can just set $\mathbf{W}^{(1)} = \mathbf{W}^{(2)} = \mathbf{I}$ (identity matrix). This gives 0 MSE but does not learn any interesting representation!



Bottleneck

- If we let $k=m$, then what is an easy (but useless) way to set the weights to give a perfect reconstruction (0 MSE)?
- For this reason, we usually set $k < m^*$; the hidden layer is then called a **bottleneck**.



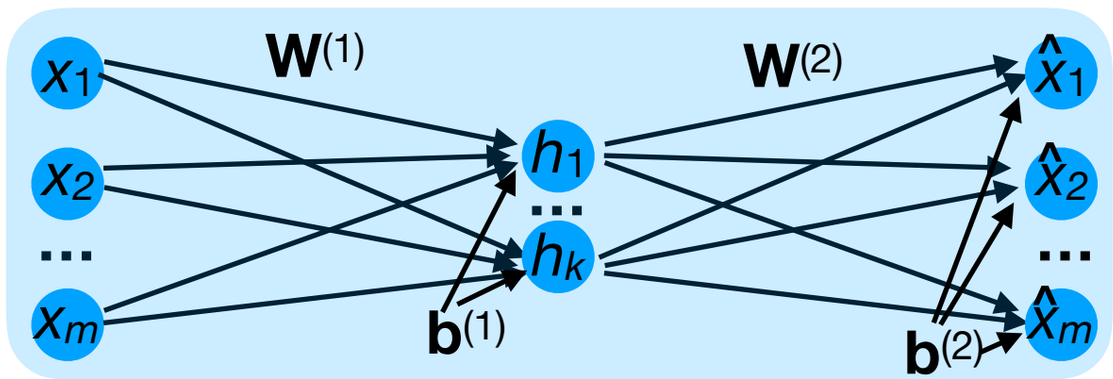
Auto-encoder

*An important exception are de-noising auto-encoders.

Autoencoders vs PCA

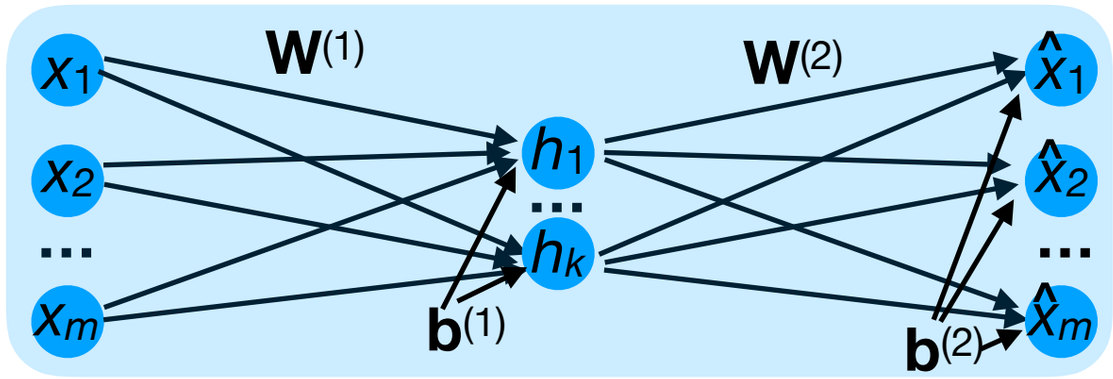
- When the activation functions are linear, and we constrain $\mathbf{W}^{(2)} = \mathbf{W}^{(1)\top}$, then the auto-encoder is almost identical to principal component analysis (with *k* PCs):
  - The span of the vectors in $\mathbf{W}^{(1)}$ is the same as the span of the *k* PCs.
- However, the vectors in $\mathbf{W}^{(1)}$ need not be orthogonal.

$$\hat{\mathbf{x}} = \sigma^{(2)}\left(\mathbf{W}^{(2)}\sigma^{(1)}\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right) + \mathbf{b}^{(2)}\right)$$



Auto-encoders for unsupervised pre-training

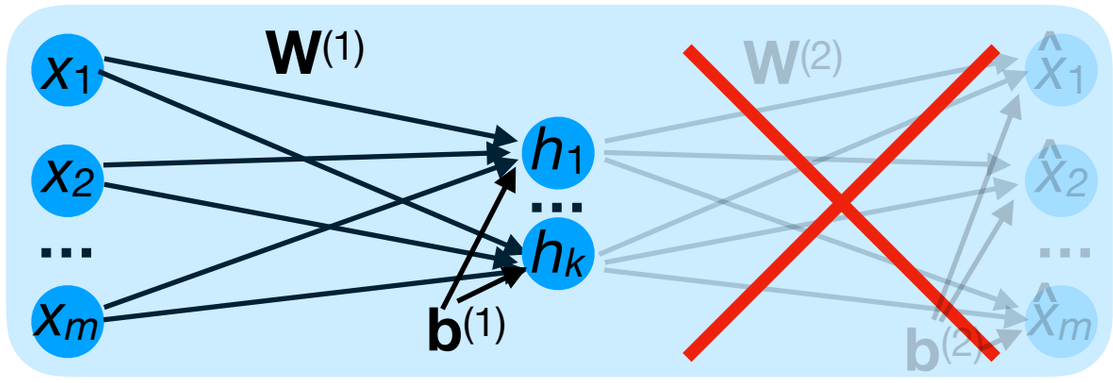
- After training the auto-encoder NN, $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ have hopefully learned to encode $\mathbf{x}$ into a representation that is useful for a variety of classification/regression problems.



Auto-encoder

Auto-encoders for unsupervised pre-training

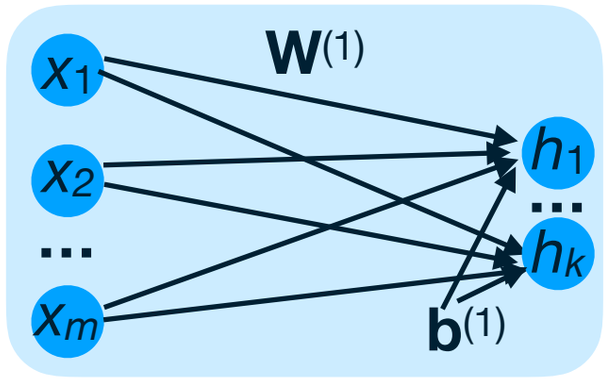
- After training the auto-encoder NN, $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ have hopefully learned to encode $\mathbf{x}$ into a representation that is useful for a variety of classification/regression problems.
- We can now just “chop off” the decoder layer(s)…



Auto-encoder

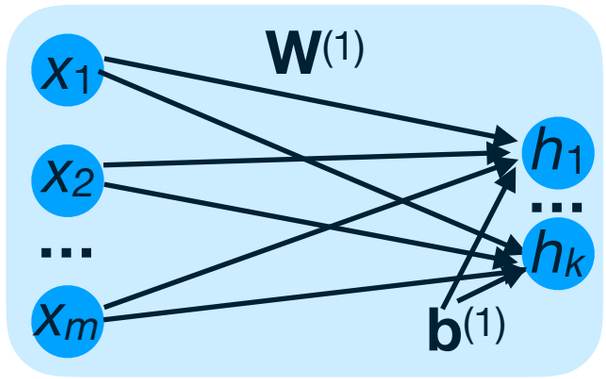
Auto-encoders for unsupervised pre-training

- After training the auto-encoder NN, $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ have hopefully learned to encode $\mathbf{x}$ into a representation that is useful for a variety of classification/regression problems.
- …and keep just the encoder layer(s).



Auto-encoders for unsupervised pre-training

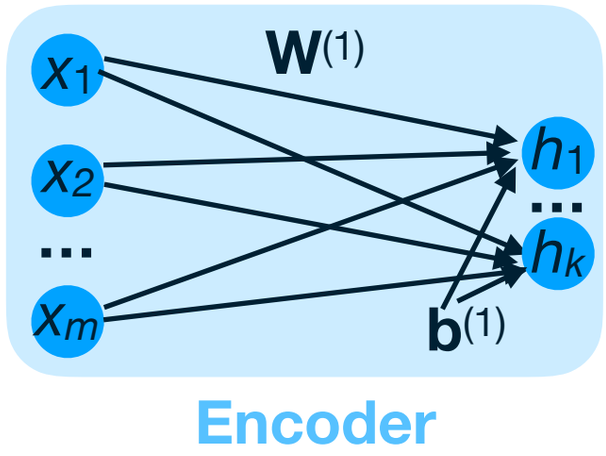
- We now have a trained encoder network that can "compress" every input example **x** into its "essence" **h**.



Encoder

Auto-encoders for unsupervised pre-training

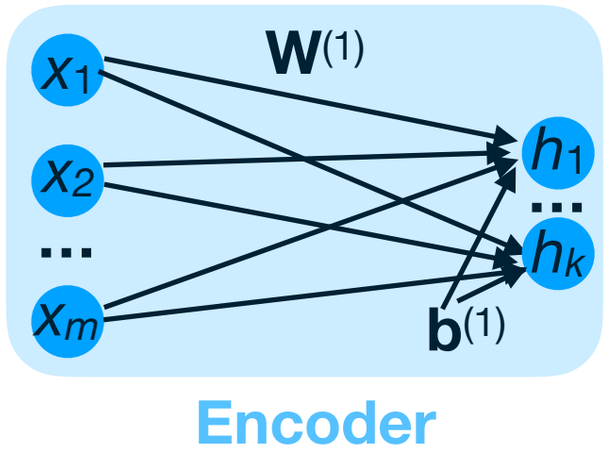
- Now, suppose we also have a (typically smaller) set of *labeled* examples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$.



Auto-encoders for unsupervised pre-training

- Now, suppose we also have a (typically smaller) set of *labeled* examples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$.

- We can use the encoder network to convert each **x** to **h** to obtain $\{(\mathbf{h}^{(i)}, y^{(i)})\}_{i=1}^{n}$.

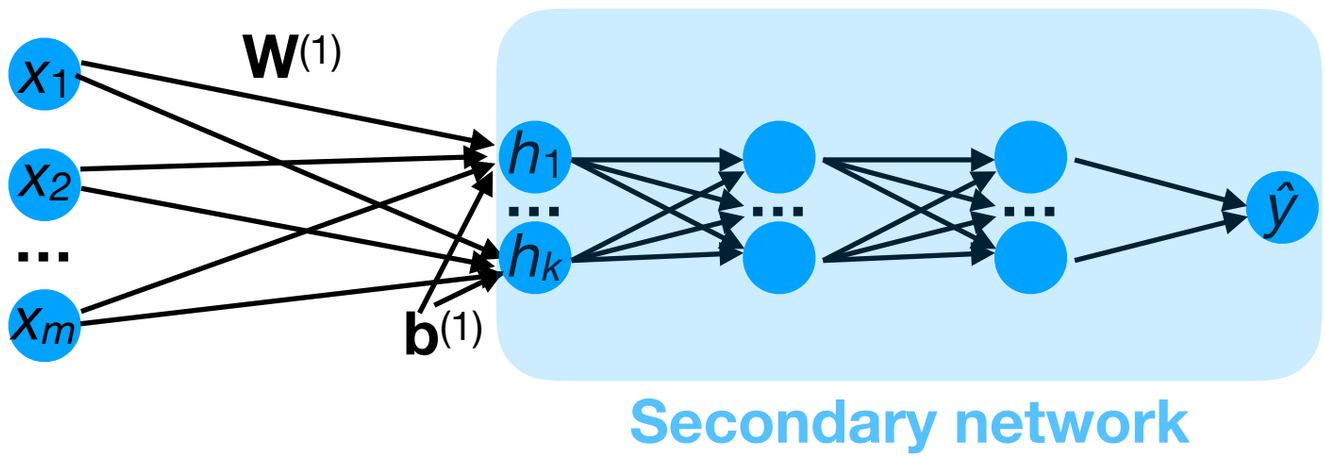


Auto-encoders for unsupervised pre-training

- Now, suppose we also have a (typically smaller) set of *labeled* examples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$.

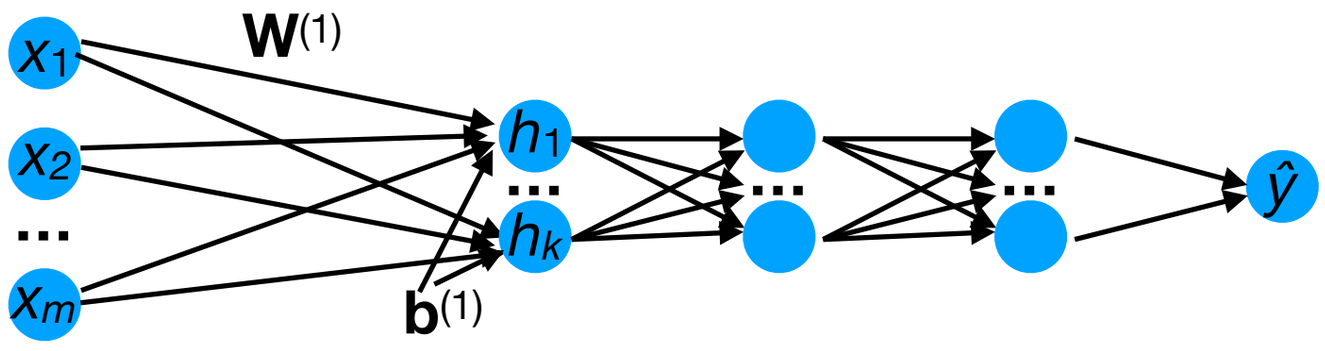
- We can use the encoder network to convert each **x** to **h** to obtain $\{(\mathbf{h}^{(i)}, y^{(i)})\}_{i=1}^{n}$.

- We then train a *secondary* NN (or any other ML model) to predict *y* from **h**.



Auto-encoders for unsupervised pre-training

- After training, the two networks (encoder + secondary) can be seen as a *single NN* that analyzes each input **x** to make a prediction $\hat{y}$.

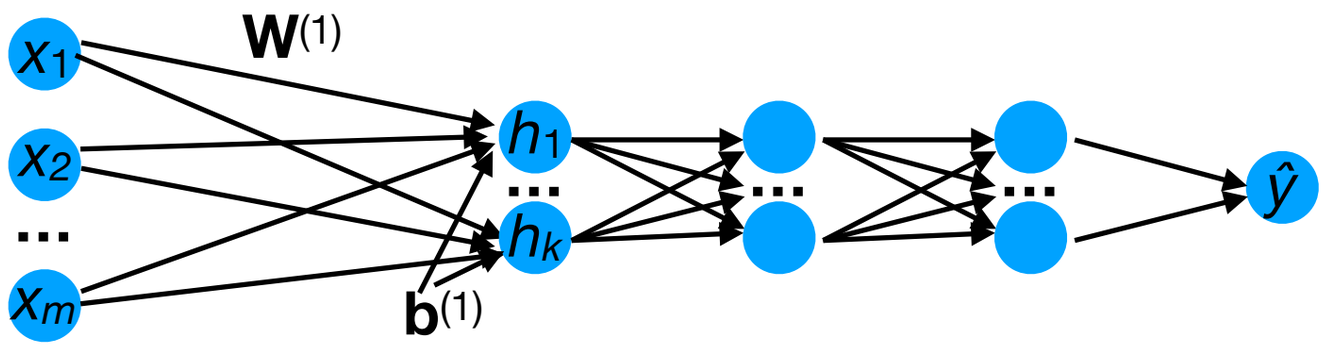


Auto-encoders for unsupervised pre-training

- Why does this help?

- The first layers of the overall network were trained on a large amount of data.

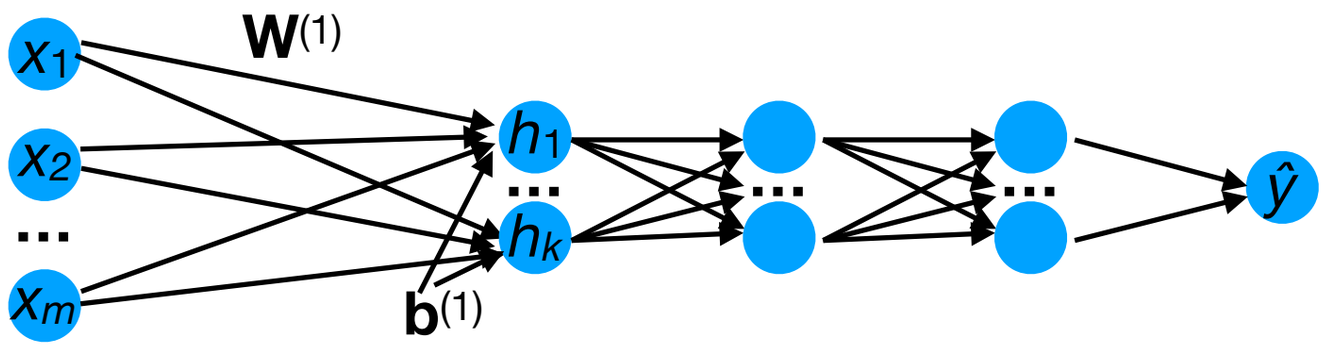
- Compressing **x** into **h** makes the secondary predictions (hopefully) easier.



Auto-encoders for unsupervised pre-training

- In addition to training the secondary NN, we can — optionally — adjust the parameters of the encoder network.

- Since the encoder was trained on a much larger (unlabeled) dataset, we don't want to "mess up" its weights too much based on just a small labeled dataset.

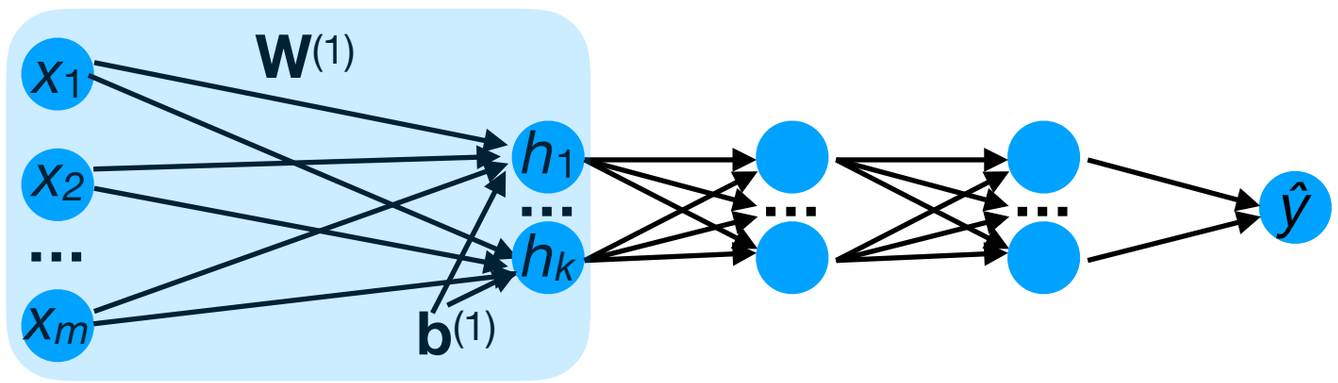


Auto-encoders for unsupervised pre-training

- In addition to training the secondary NN, we can — optionally — adjust the parameters of the encoder network.

- Since the encoder was trained on a much larger (unlabeled) dataset, we don't want to "mess up" its weights too much based on just a small labeled dataset.

- Hence, we often use a small learning rate ==> **fine-tuning**.



(Supervised)

pre-training

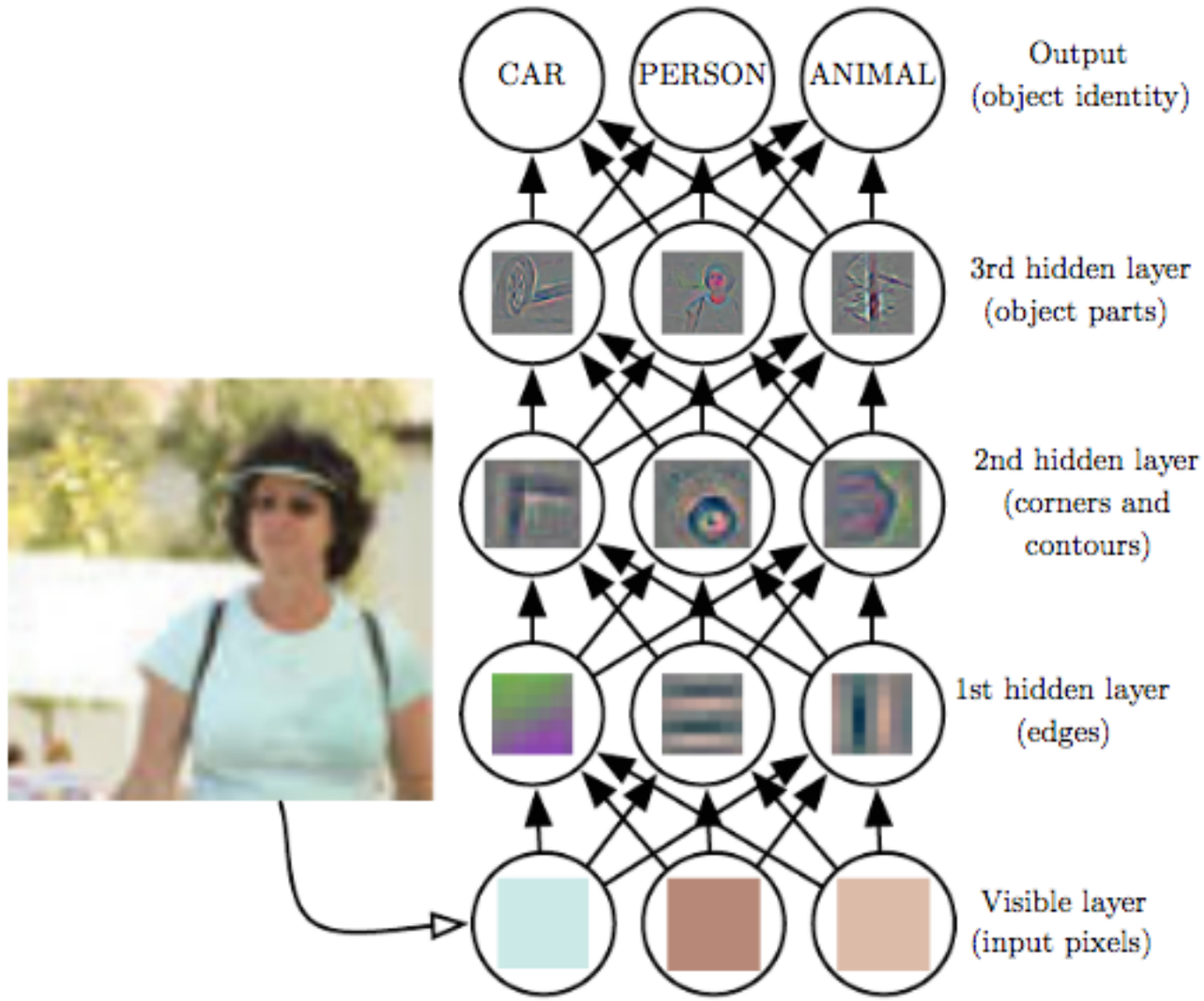
Supervised pre-training

- An alternative strategy to finding good feature representations is to borrow a NN from a related task.

- For instance, there now exist high-accuracy networks for recognizing 1000+ object categories from images (next slide).

- We can “borrow” the feature representation from one ML model and apply it to another application domain…

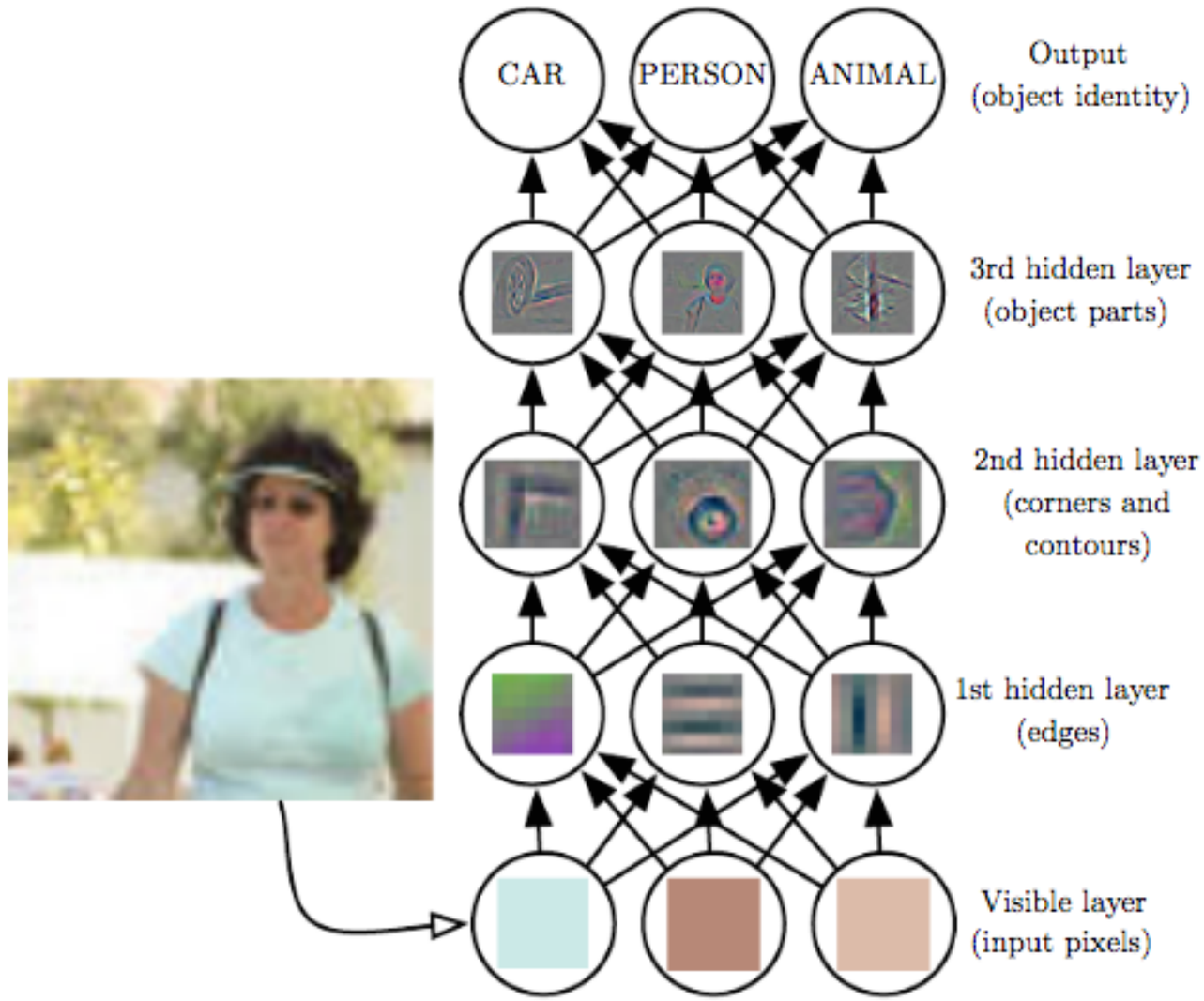
Learning representations



- The first feature representation looks vaguely like the representation learned by my MNIST network.

http://www.deeplearningbook.org/contents/intro.html

Learning representations



- Each layer of the network finds successively more abstract feature representations.
  - This was not “hard-coded” — it just turned out that these representations were useful for predicting the target labels.

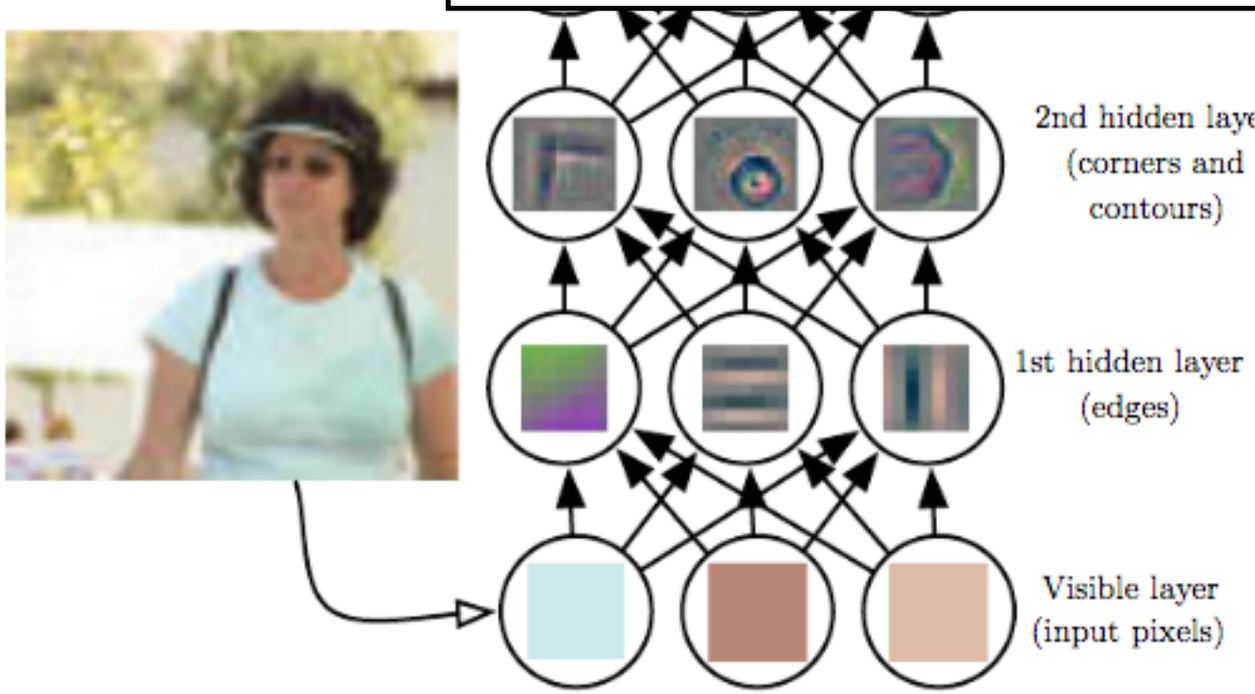
http://www.deeplearningbook.org/contents/intro.html

Supervised pre-training

- Might one (or more) of the feature representations from this NN do well on a different but related problem, e.g., smile detection or age estimation?
- Strategy:
  1. Pre-train a NN on a large dataset (e.g., ImageNet) for a general-purpose image recognition task.
  2. “Chop off” the final layer(s).
  3. Add a secondary network in place of the deleted layers, and train it for the new prediction task.
  4. Optional: fine-tune the rest of the NN.

Supervised pre-training

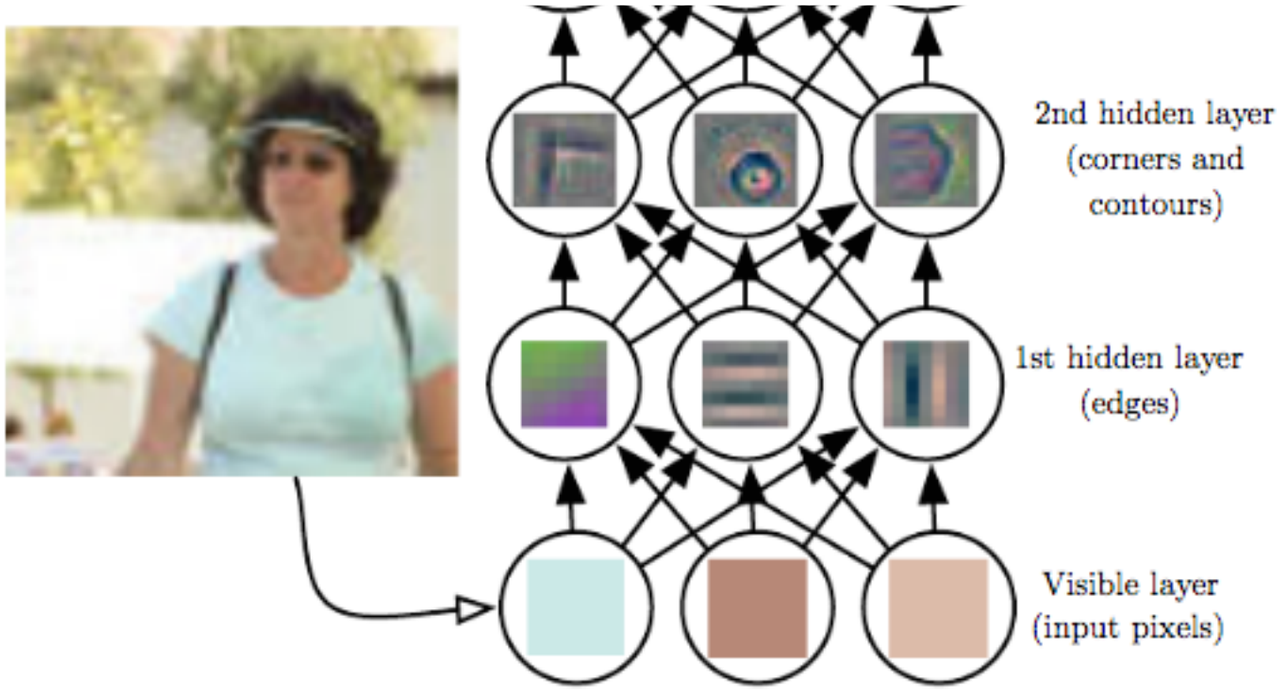
Replace with secondary network for new application domain.



http://www.deeplearningbook.org/contents/intro.html

Supervised pre-training

Chop off.



http://www.deeplearningbook.org/contents/intro.html

Supervised pre-training

- This strategy is known as **supervised pre-training** and can be highly effective for application domains for which only a small number of labeled data are available.

Multi-task learning (MTL)

Multi-task learning (MTL)

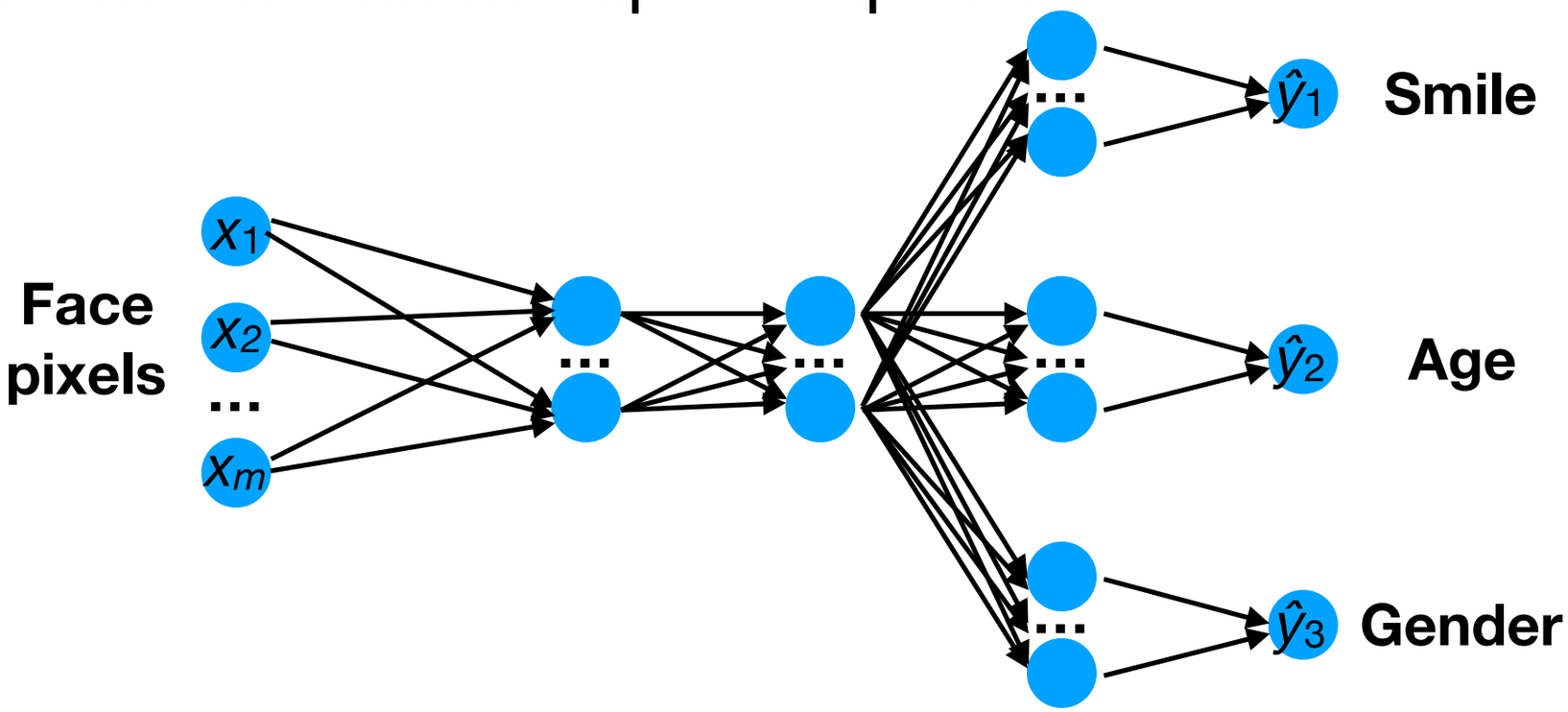
- A NN can generalize to unseen data when it computes a hidden representation that explains the data well.

- We can sometimes encourage the NN to learn a general hidden representation by training it to solve multiple related tasks.

- E.g., for automated face analysis:
  - Smile detection
  - Age estimation
  - Gender detection

Multi-task learning (MTL)

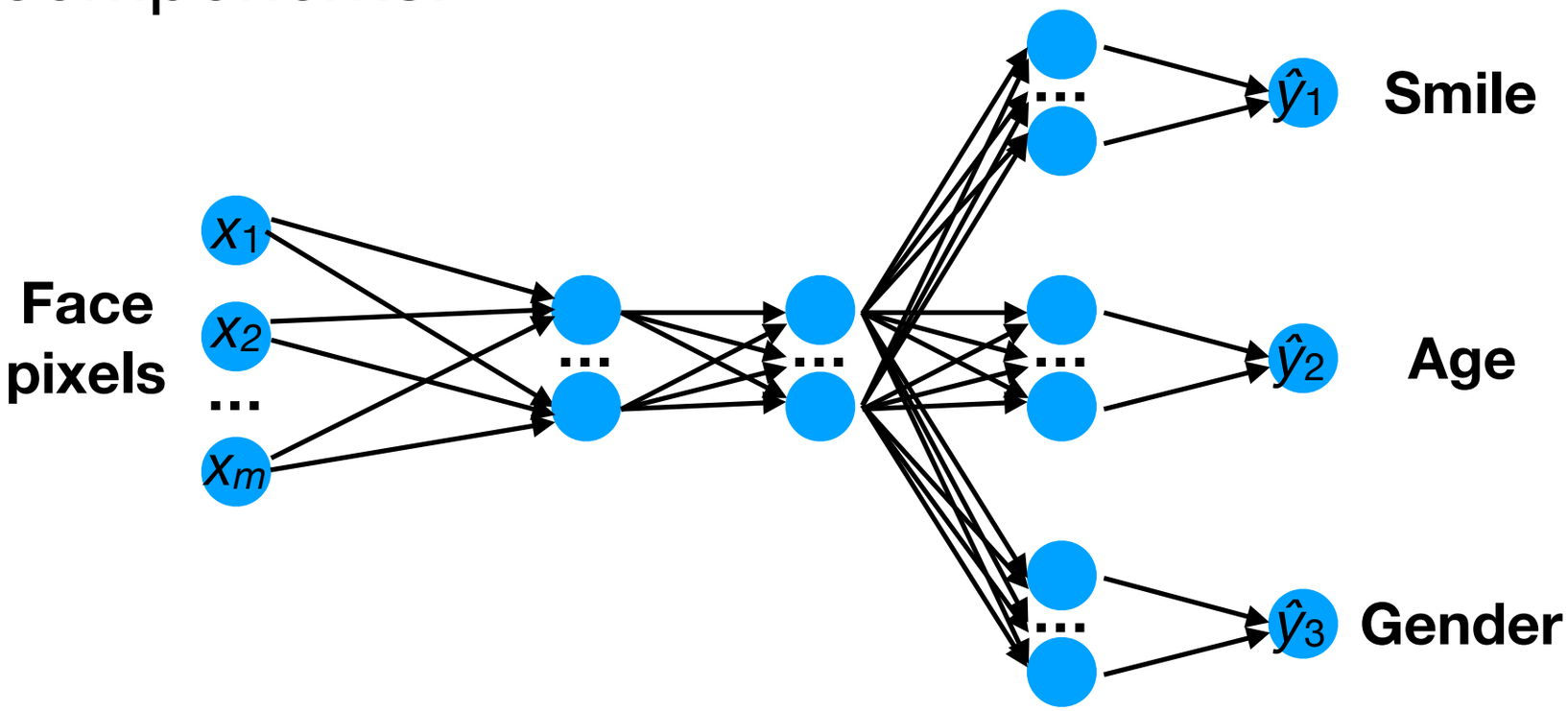
- If we have labeled data for all these tasks, we can train a network with multiple outputs:



- Note that the labels for the auxiliary tasks can be useful even if we only care about one task.

Multi-task learning (MTL)

- With MTL, our loss function consists of multiple components:



$$f_{\text{MTL}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots) = f_{\text{smile}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots) + f_{\text{age}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots) + f_{\text{gender}}(\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \ldots)$$