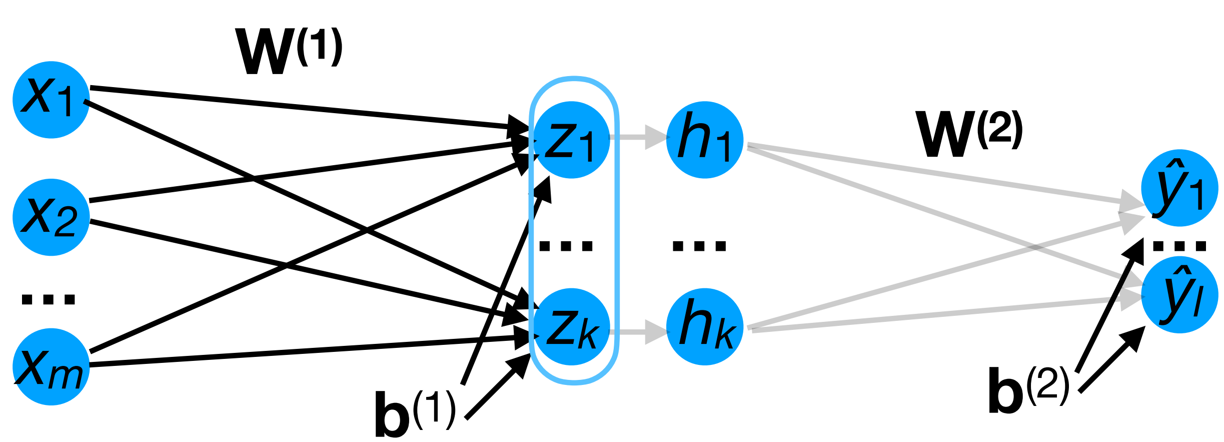
CS 4342: Class 21

Jacob Whitehill

Forwards and backwards propagation

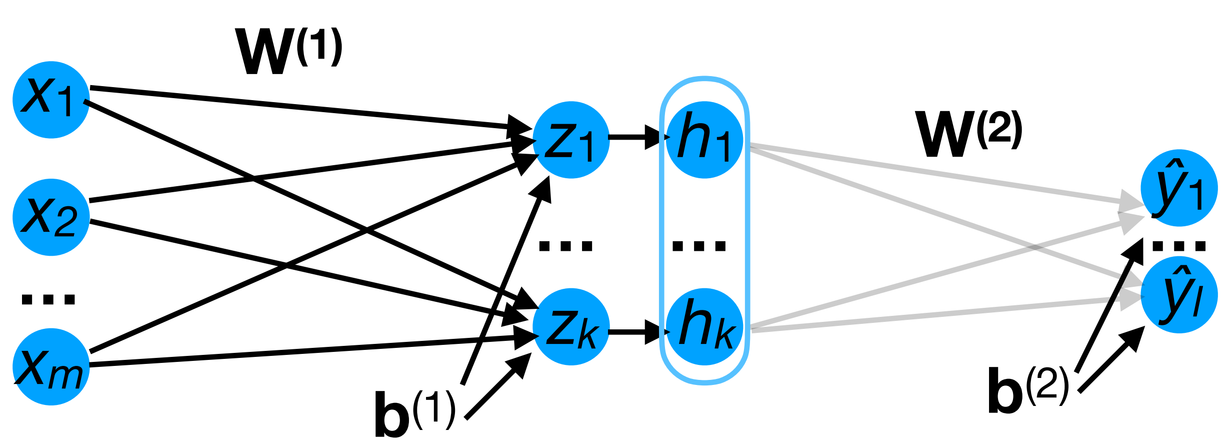
Computing the gradients

- Consider the 3-layer NN below:
  - From **x**, $\mathbf{W}^{(1)}$, and $\mathbf{b}^{(1)}$, we can compute **z**.



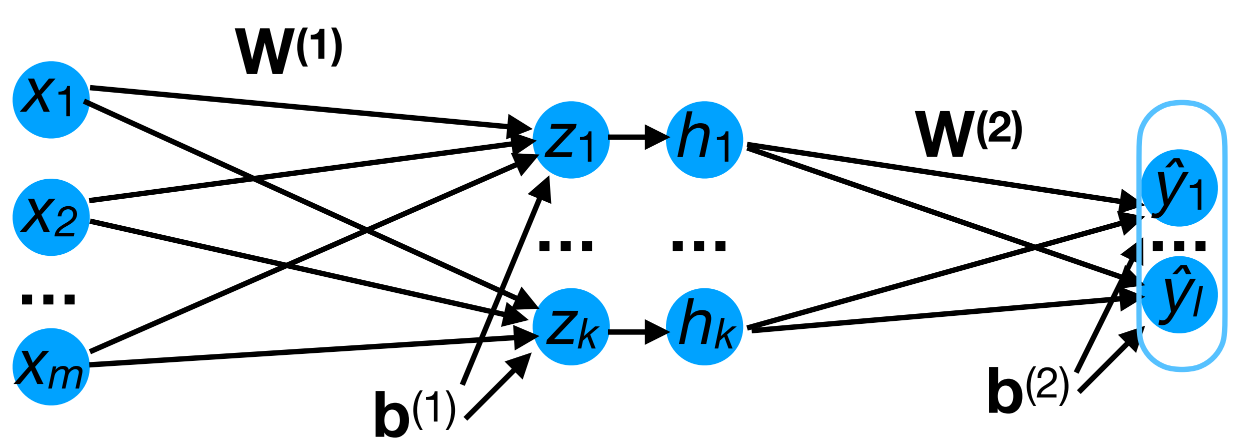
Computing the gradients

- Consider the 3-layer NN below:
  - From $\mathbf{x}$, $\mathbf{W}^{(1)}$, and $\mathbf{b}^{(1)}$, we can compute $\mathbf{z}$.
  - From $\mathbf{z}$ and σ, we can compute $\mathbf{h} = \sigma(\mathbf{z})$.



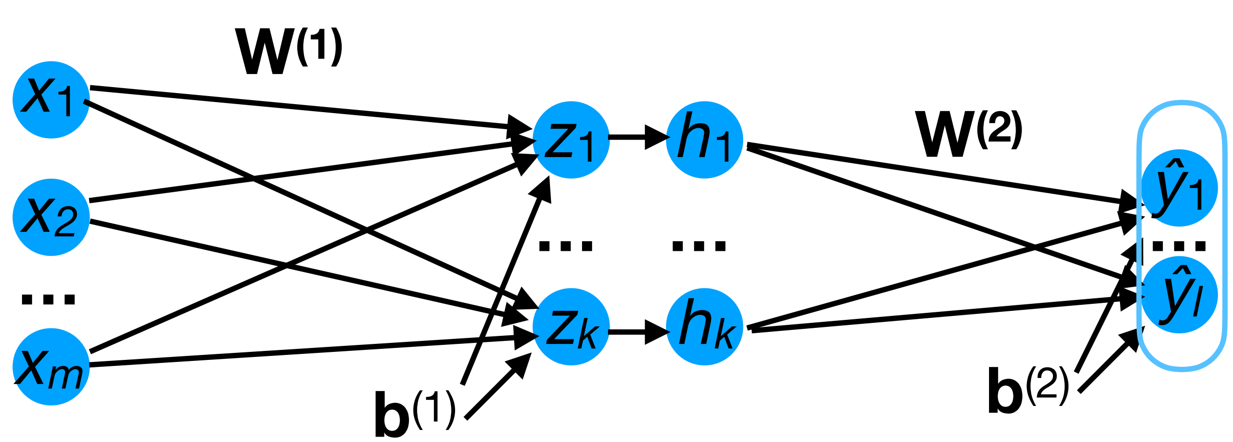
Computing the gradients

- Consider the 3-layer NN below:
  - From **x**, $\mathbf{W}^{(1)}$, and $\mathbf{b}^{(1)}$, we can compute **z**.
  - From **z** and σ, we can compute $\mathbf{h} = \sigma(\mathbf{z})$.
  - From **h**, $\mathbf{W}^{(2)}$, and $\mathbf{b}^{(2)}$, we can compute $\hat{\mathbf{y}}$.



Computing the gradients

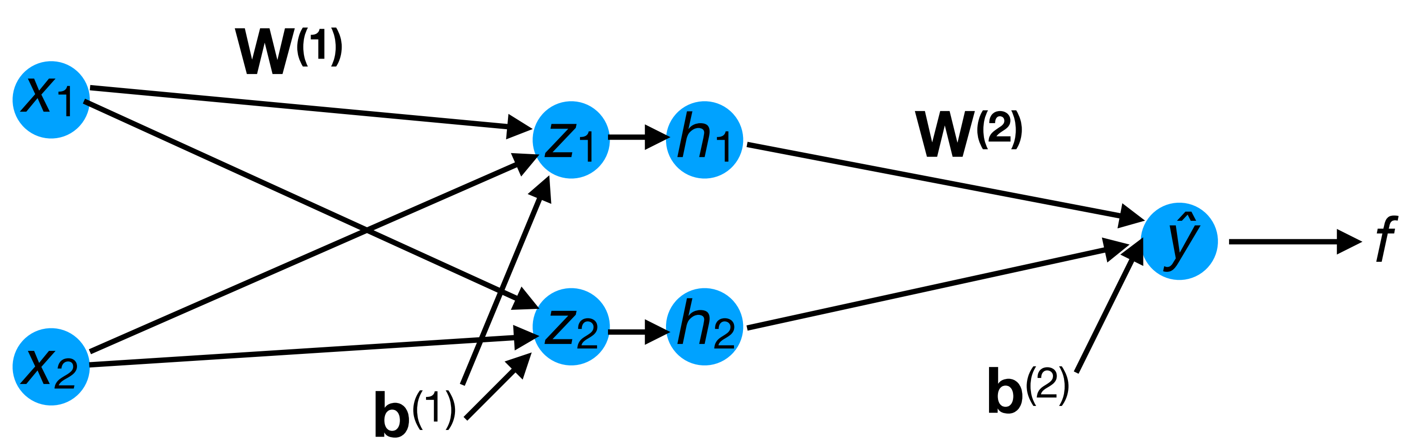
- This process is known as **forward propagation**.
  - It produces all the intermediary (**h**, **z**) and final ($\hat{\mathbf{y}}$) network outputs.



Computing the gradients

- Now, let's look at how to compute each gradient term:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(2)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}} \\
\frac{\partial f}{\partial \mathbf{b}^{(2)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}} \\
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
\frac{\partial f}{\partial \mathbf{b}^{(1)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{b}^{(1)}}
\end{aligned}
$$

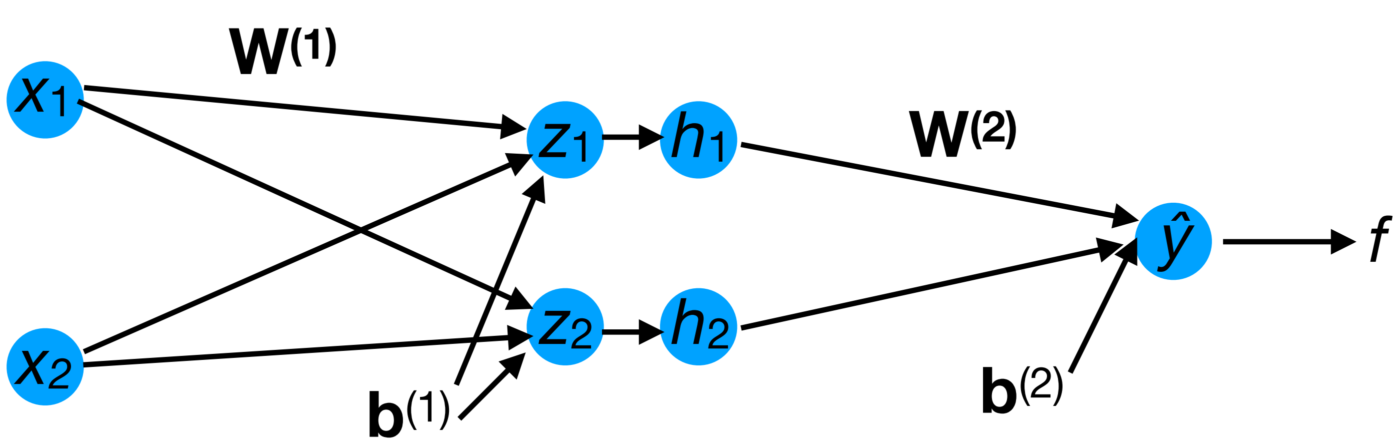


Computing the gradients

- Now, let's look at how to compute each gradient term:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(2)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}} \\
\frac{\partial f}{\partial \mathbf{b}^{(2)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}} \\
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
\frac{\partial f}{\partial \mathbf{b}^{(1)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{b}^{(1)}}
\end{aligned}
$$

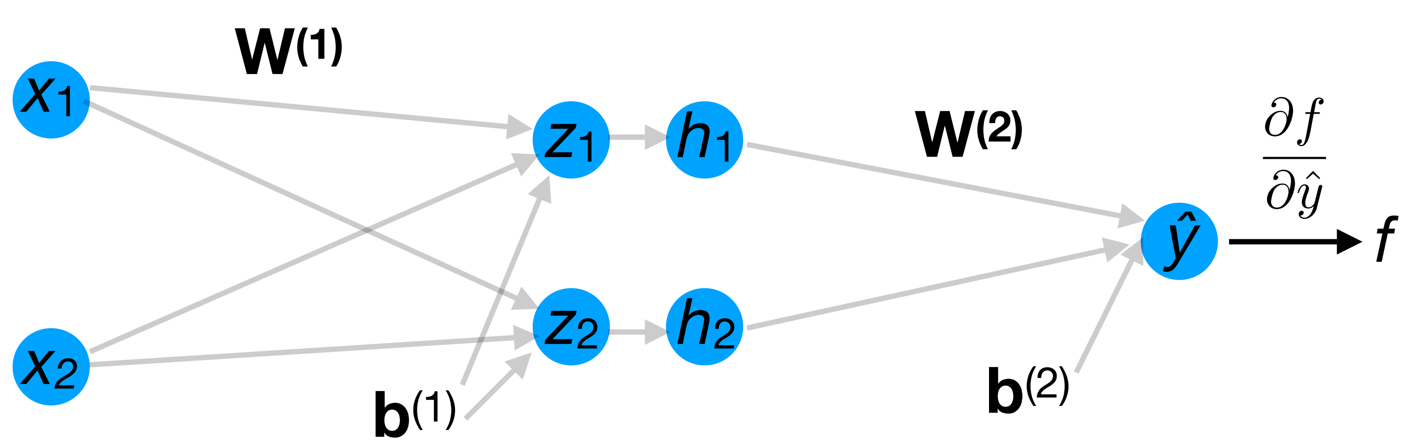
Redundant computation



Computing the gradients

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

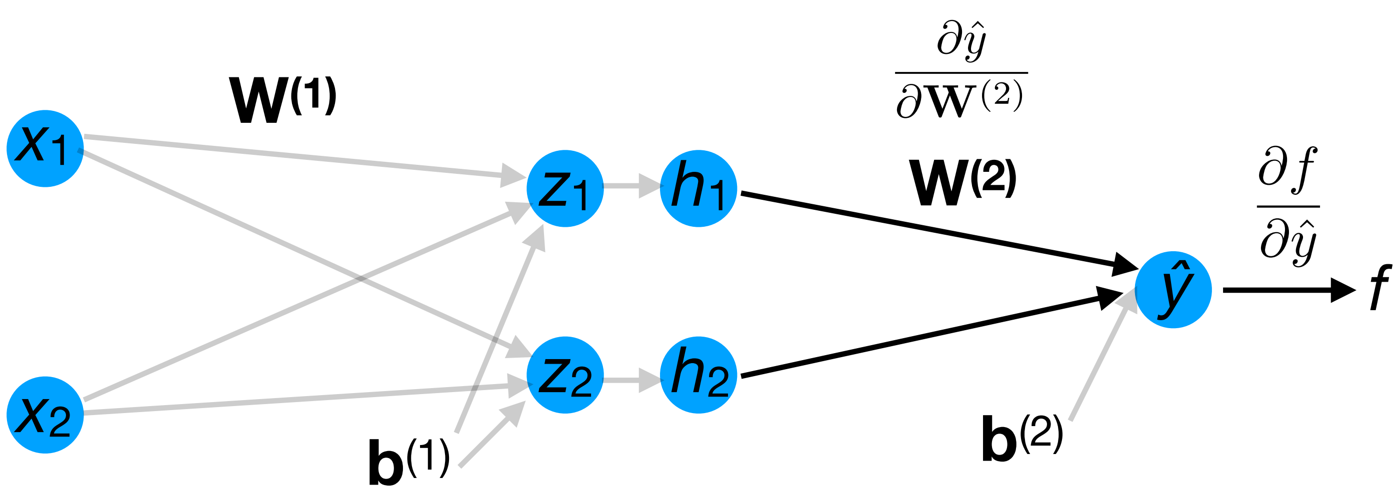
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}}$$



Computing the gradients

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

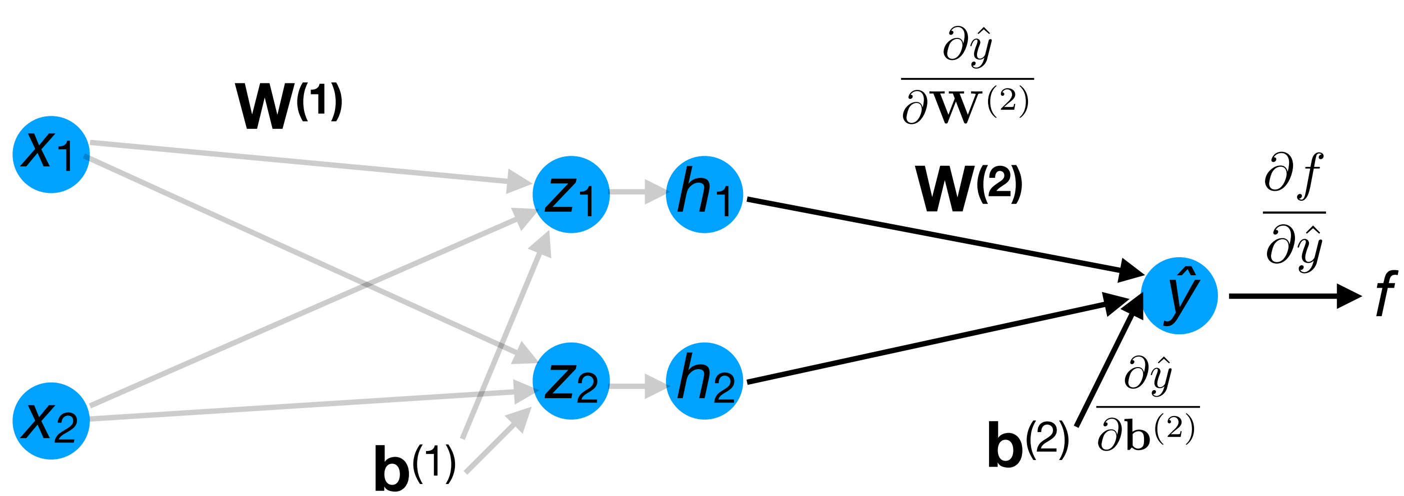
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$



Computing the gradients

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

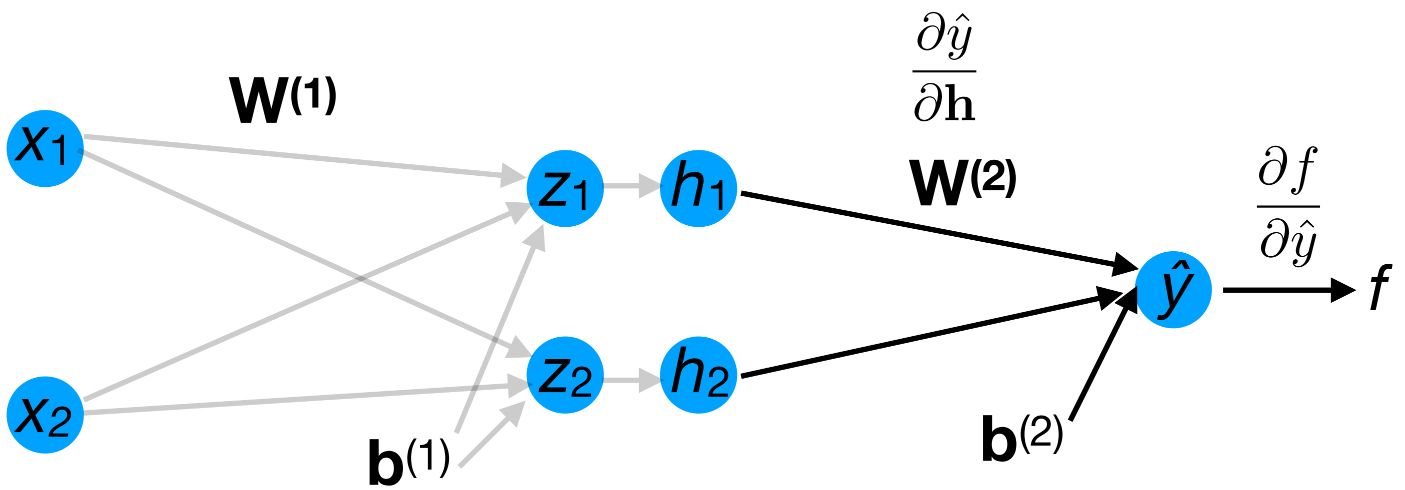
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$
$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$



Computing the gradients

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

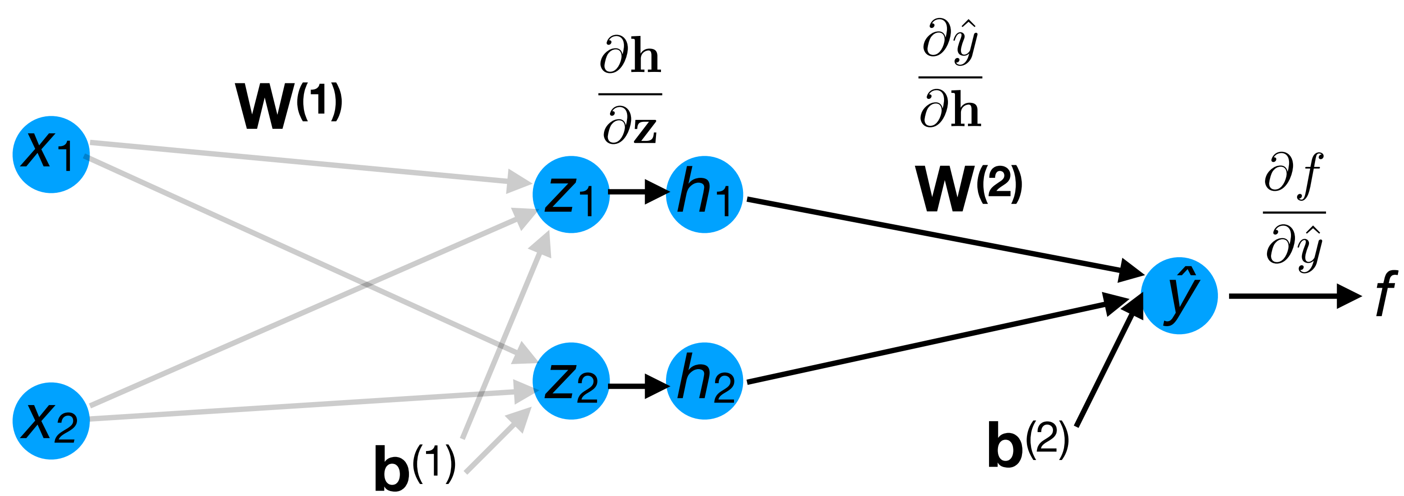
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$
$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$
$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}}$$



Computing the gradients

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

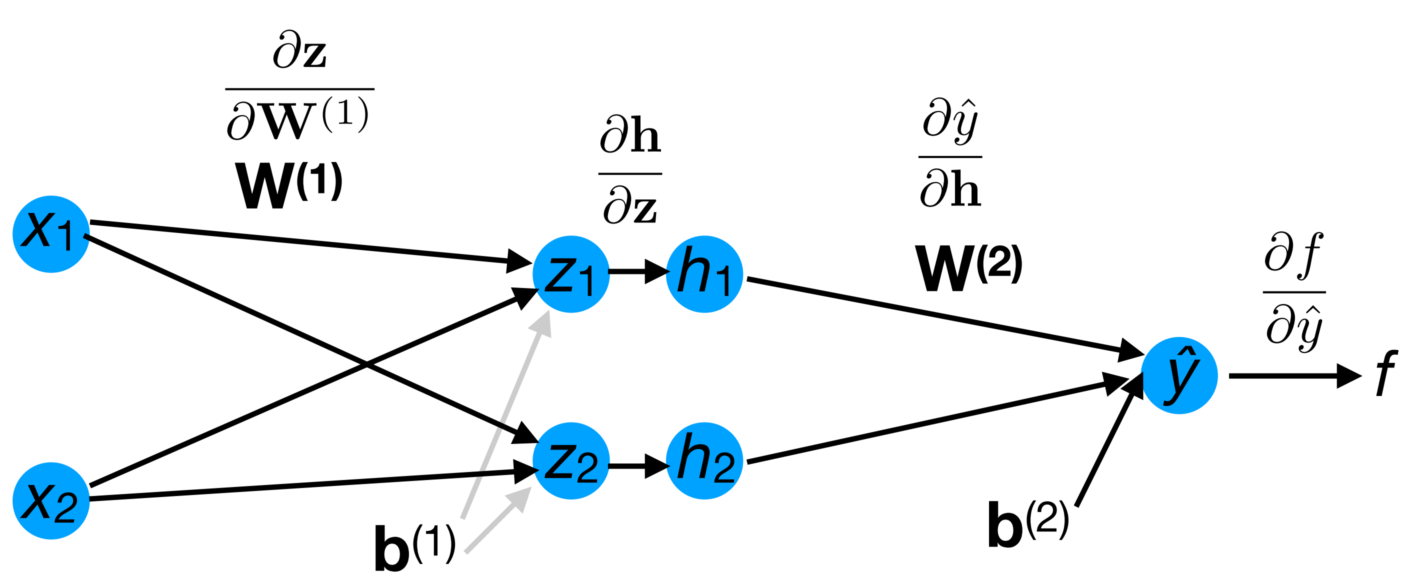
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$
$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$
$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}}$$



Computing the gradients

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

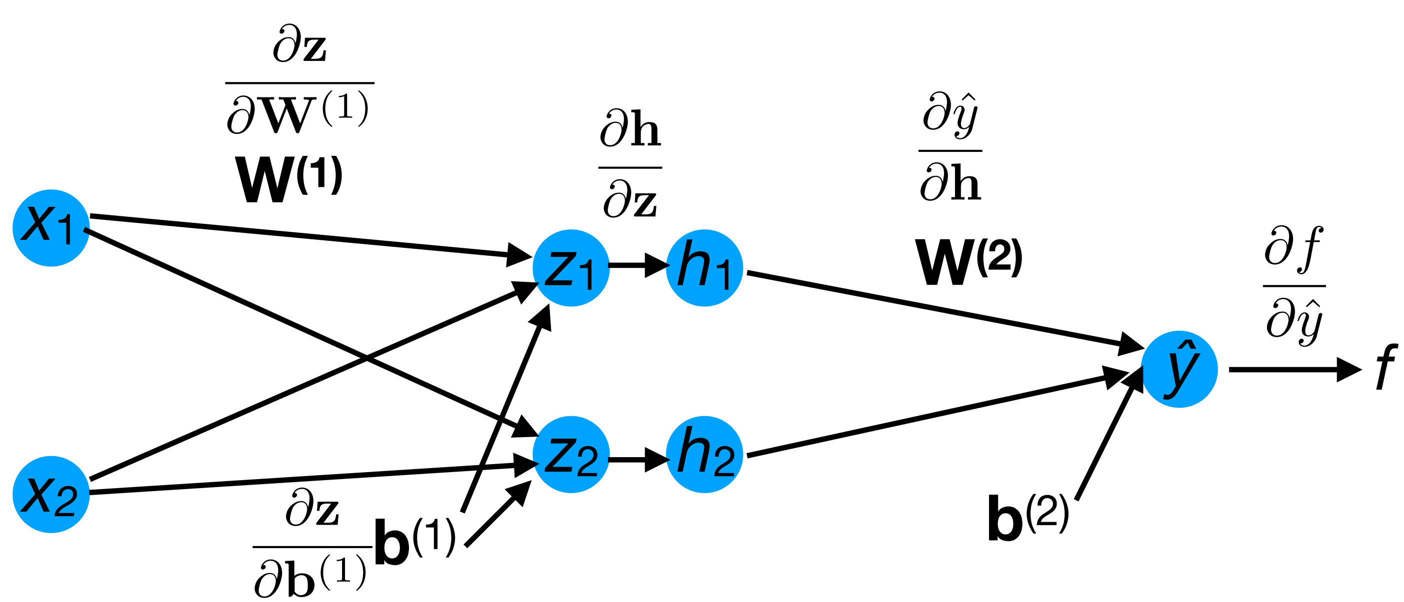
$$\frac{\partial f}{\partial \mathbf{W}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}}$$
$$\frac{\partial f}{\partial \mathbf{b}^{(2)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}}$$
$$\frac{\partial f}{\partial \mathbf{W}^{(1)}} = \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}}$$



Computing the gradients

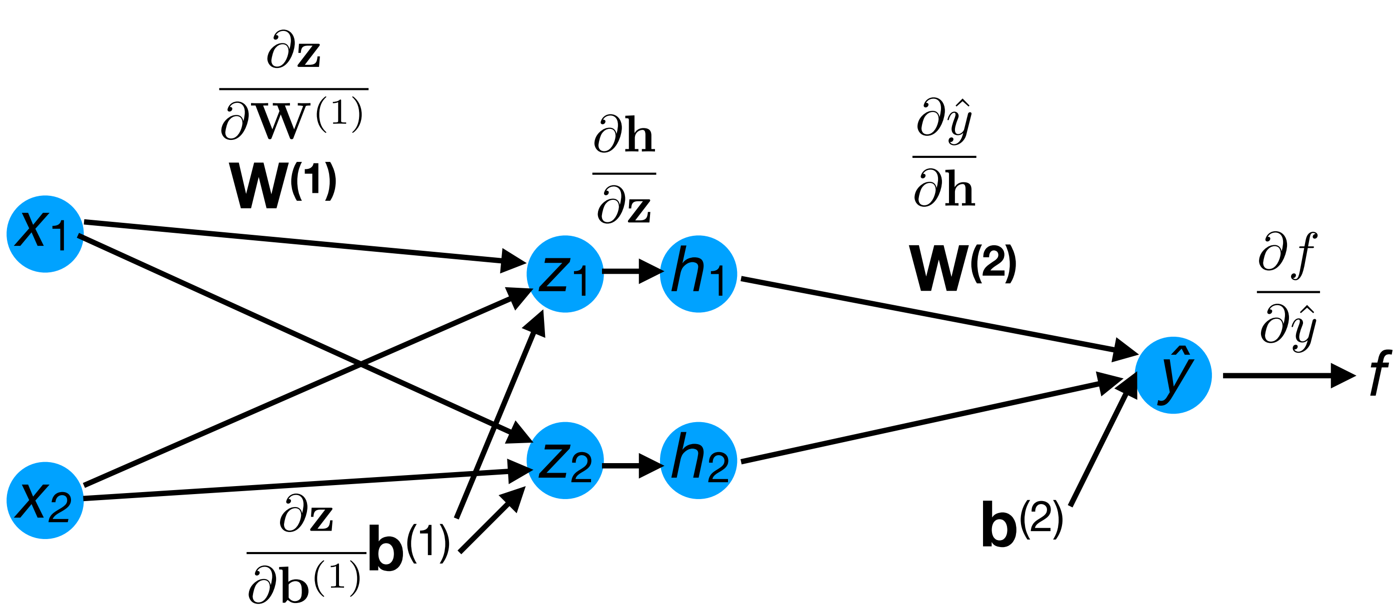
- Here's how we can compute all these *efficiently*:

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(2)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{W}^{(2)}} \\
\frac{\partial f}{\partial \mathbf{b}^{(2)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{b}^{(2)}} \\
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
\frac{\partial f}{\partial \mathbf{b}^{(1)}} &= \frac{\partial f}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{b}^{(1)}}
\end{aligned}
$$

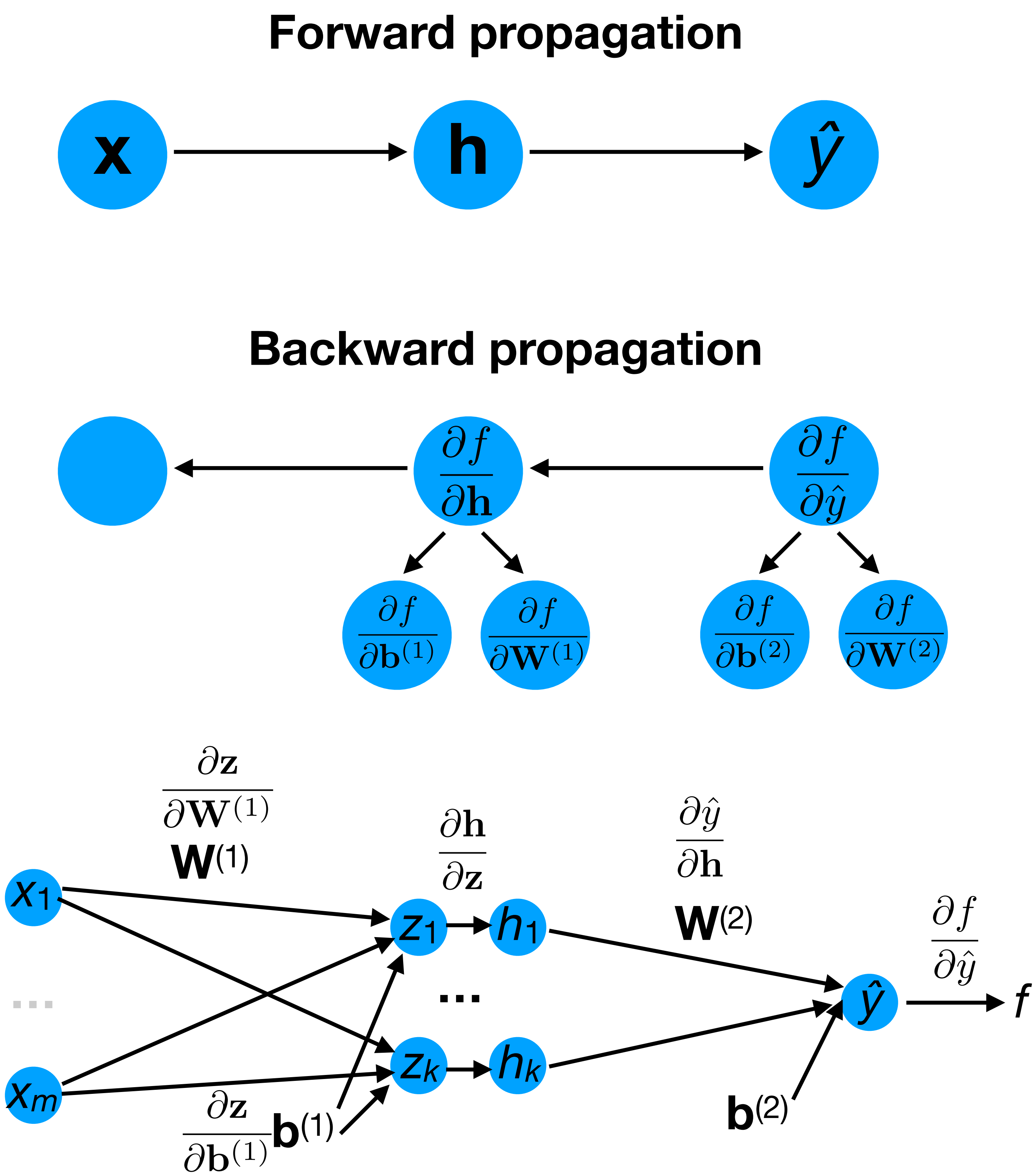


Computing the gradients

- 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



Jacob Whitehill, WPI

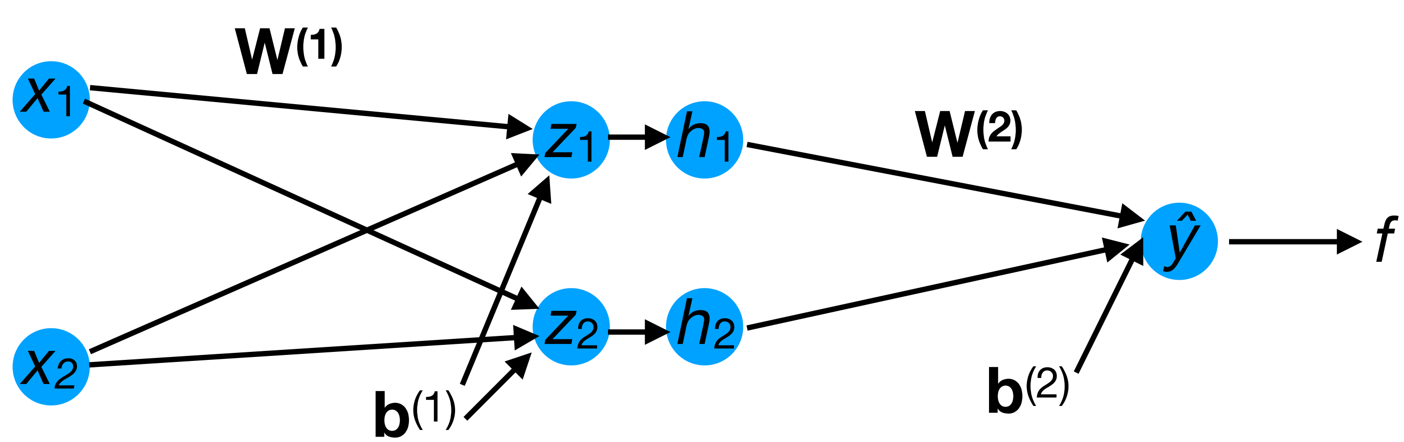
Computing the gradients

- 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

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



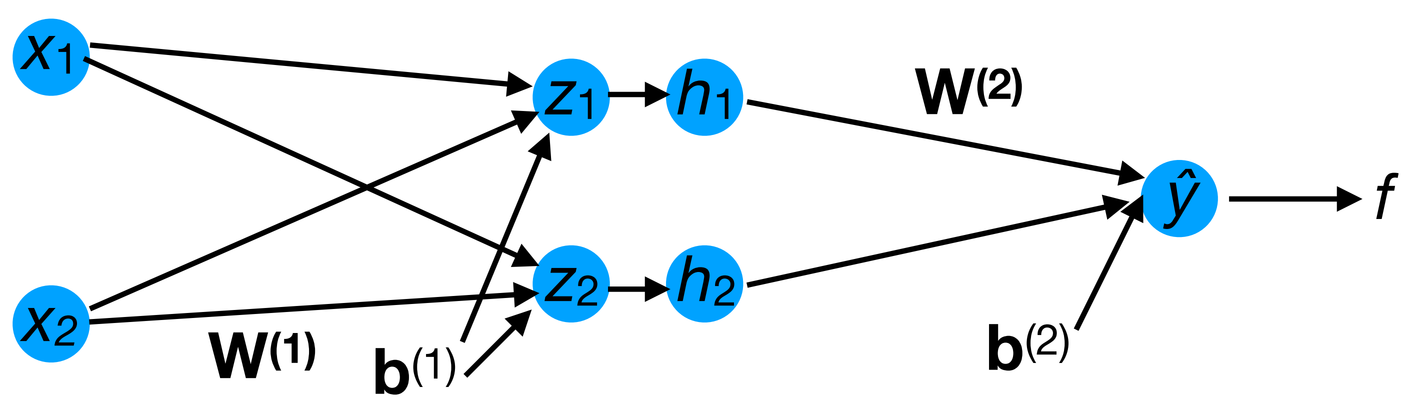
Computing the gradients

- Let's derive each gradient term in turn:

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

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

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



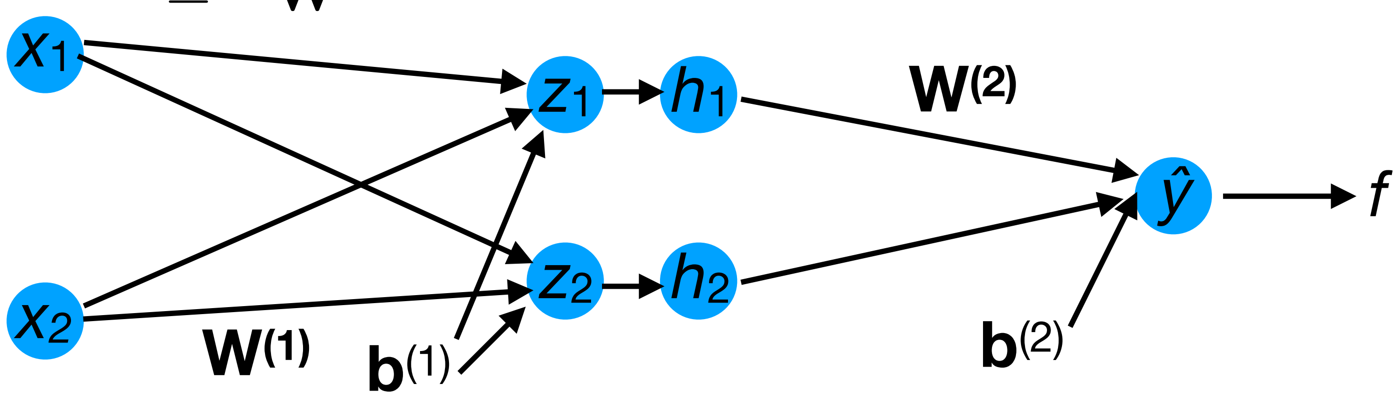
Computing the gradients

- Let's derive each gradient term in turn:

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

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

$$\begin{aligned}
\hat{y} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\
&= \mathbf{W}^{(2)}_1 \mathbf{h}_1 + \mathbf{W}^{(2)}_2 \mathbf{h}_2 + \mathbf{b}^{(2)} \\
\implies \frac{\partial \hat{y}}{\partial \mathbf{h}} &= \begin{bmatrix} \frac{\partial \hat{y}}{\partial \mathbf{h}_1} & \frac{\partial \hat{y}}{\partial \mathbf{h}_2} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{W}^{(2)}_1 & \mathbf{W}^{(2)}_2 \end{bmatrix} \\
&= \mathbf{W}^{(2)}
\end{aligned}$$



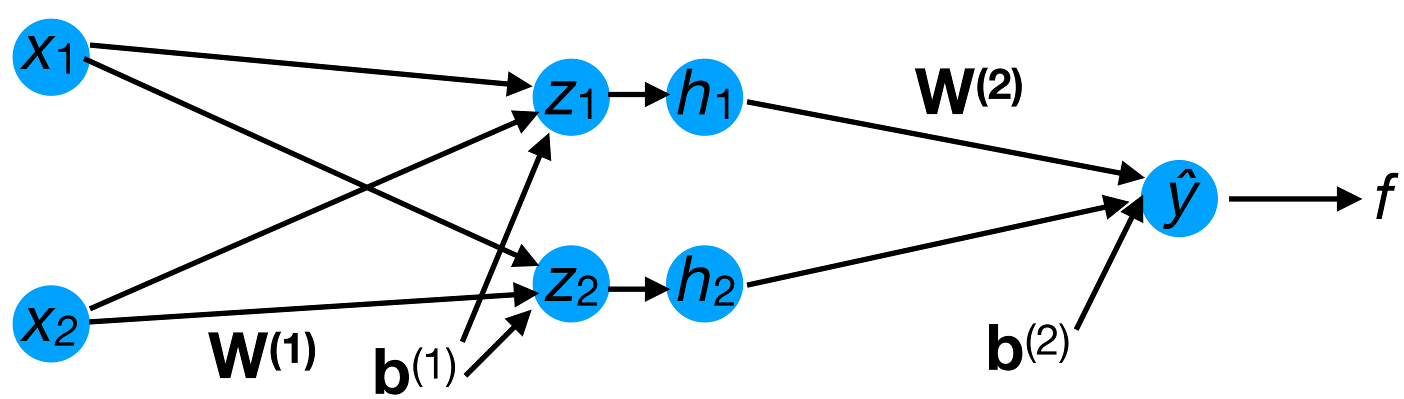
Computing the gradients

- Let's derive each gradient term in turn:

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

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

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



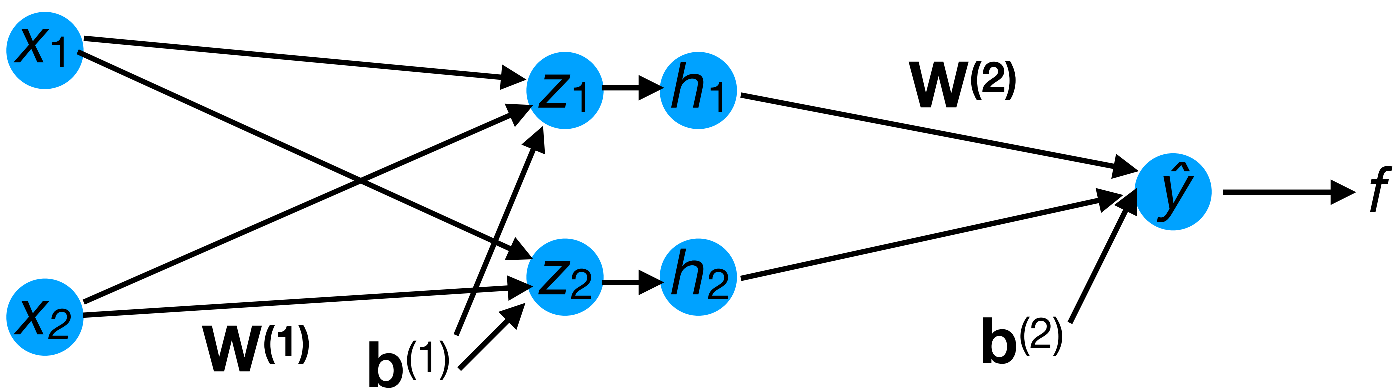
Computing the gradients

- Let's derive each gradient term in turn:

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

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

$$\begin{aligned} \mathbf{h} &= \begin{bmatrix} \mathrm{relu}(\mathbf{z}_1) \\ \mathrm{relu}(\mathbf{z}_2) \end{bmatrix} \\ \Longrightarrow \frac{\partial \mathbf{h}}{\partial \mathbf{z}} &= \begin{bmatrix} \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_2} \\ \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_2} \end{bmatrix} \end{aligned}$$



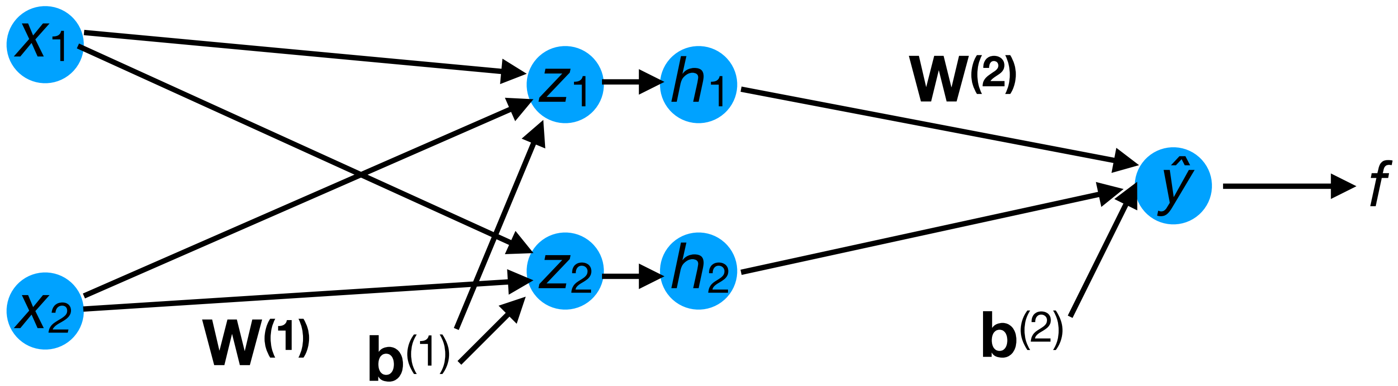
Computing the gradients

- Let's derive each gradient term in turn:

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

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

$$\begin{aligned}
\mathbf{h} &= \left[\begin{array}{c} \mathrm{relu}(\mathbf{z}_1) \\ \mathrm{relu}(\mathbf{z}_2) \end{array} \right] \\
\implies \frac{\partial \mathbf{h}}{\partial \mathbf{z}} &= \left[\begin{array}{cc} \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_1}{\partial \mathbf{z}_2} \\ \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_1} & \frac{\partial \mathbf{h}_2}{\partial \mathbf{z}_2} \end{array} \right] \\
&= \left[\begin{array}{cc} \mathrm{relu}'(\mathbf{z}_1) & 0 \\ 0 & \mathrm{relu}'(\mathbf{z}_2) \end{array} \right]
\end{aligned}$$



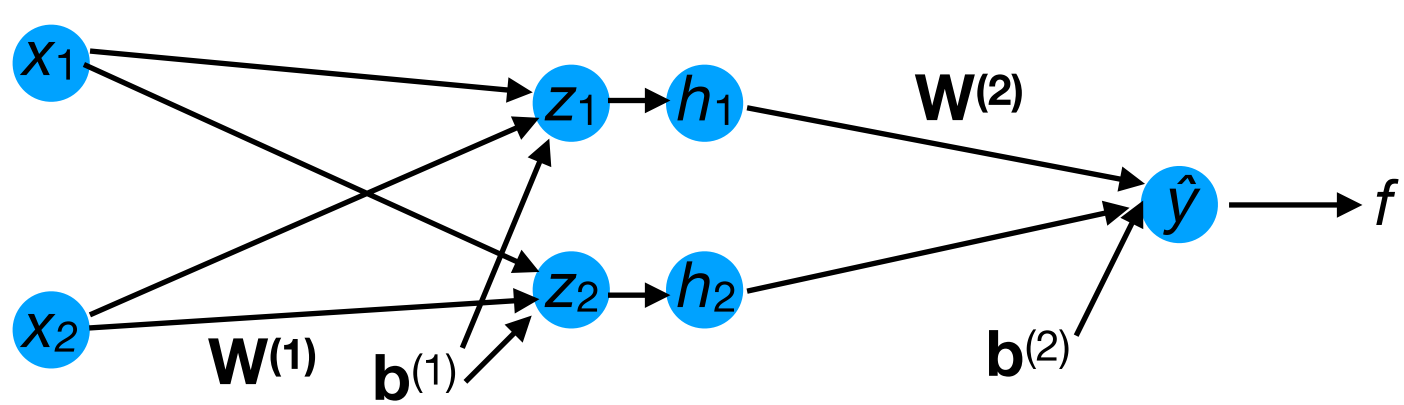
Computing the gradients

- Let's derive each gradient term in turn:

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

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

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



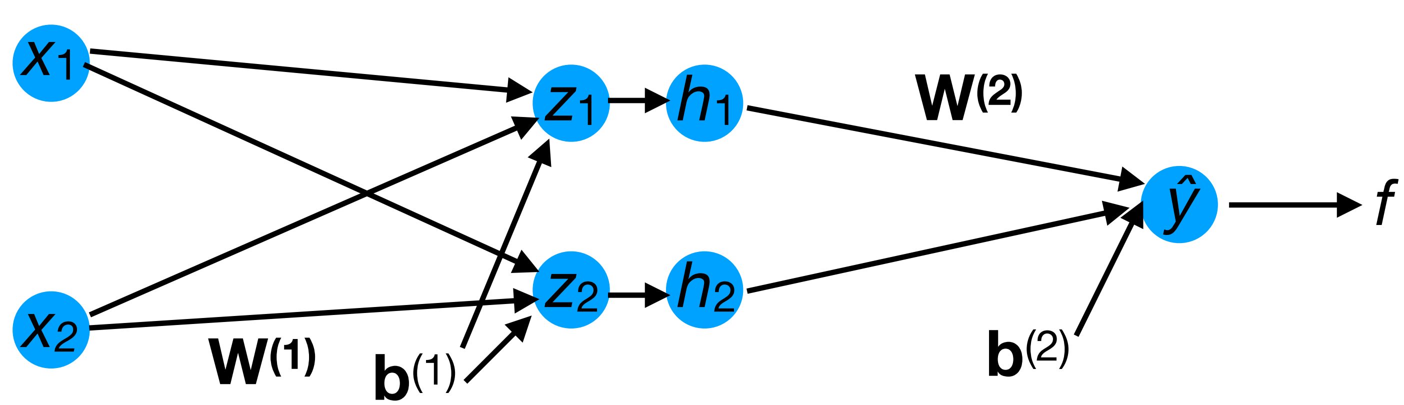
Computing the gradients

- Let's derive each gradient term in turn:

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

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

$$\begin{aligned} \mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\ \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{W}_1^{(1)} & \mathbf{W}_2^{(1)} \\ \mathbf{W}_3^{(1)} & \mathbf{W}_4^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{b}_1^{(1)} \\ \mathbf{b}_2^{(1)} \end{bmatrix} \end{aligned}$$



Computing the gradients

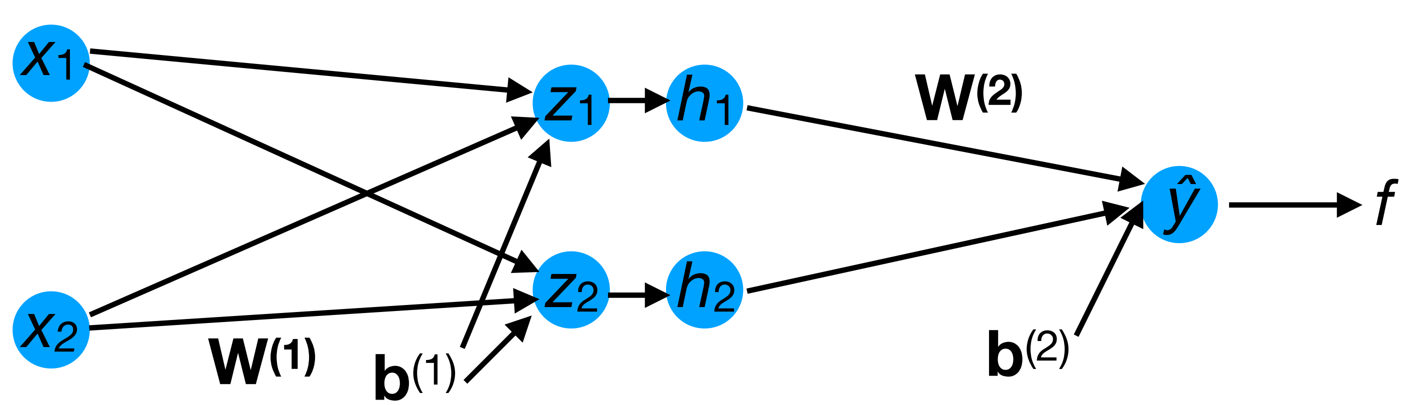
- Let's derive each gradient term in turn:

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

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

$$\begin{aligned}
\mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\
\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{W}^{(1)}_1 & \mathbf{W}^{(1)}_2 \\ \mathbf{W}^{(1)}_3 & \mathbf{W}^{(1)}_4 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{b}^{(1)}_1 \\ \mathbf{b}^{(1)}_2 \end{bmatrix} \\
\Longrightarrow \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} &= \begin{bmatrix} \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}^{(1)}_1} & \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}^{(1)}_2} & \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}^{(1)}_3} & \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}^{(1)}_4} \\ \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}^{(1)}_1} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}^{(1)}_2} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}^{(1)}_3} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}^{(1)}_4} \end{bmatrix}
\end{aligned}$$

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



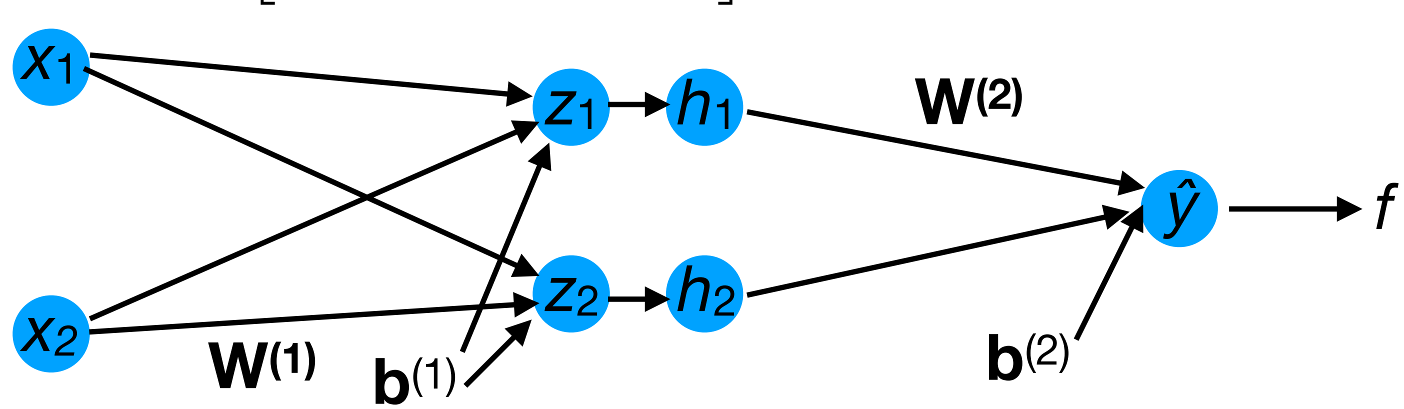
Computing the gradients

- Let's derive each gradient term in turn:

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

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

$$\begin{aligned}
\mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\
\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{W}_1^{(1)} & \mathbf{W}_2^{(1)} \\ \mathbf{W}_3^{(1)} & \mathbf{W}_4^{(1)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{b}_1^{(1)} \\ \mathbf{b}_2^{(1)} \end{bmatrix} \\
\implies \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} &= \begin{bmatrix} \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}_1^{(1)}} & \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}_2^{(1)}} & \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}_3^{(1)}} & \frac{\partial \mathbf{z}_1}{\partial \mathbf{W}_4^{(1)}} \\ \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}_1^{(1)}} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}_2^{(1)}} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}_3^{(1)}} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{W}_4^{(1)}} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix}
\end{aligned}$$



Analytical simplification

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

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

Analytical simplification

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

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{W}^{(1)}} &= \frac{\partial f}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^{(1)}} \\
&= (\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)} \left[\begin{array}{cc} \mathrm{relu}'(\mathbf{z}_1) & 0 \\ 0 & \mathrm{relu}'(\mathbf{z}_2) \end{array} \right] \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & 0 & 0 \\ 0 & 0 & \mathbf{x}_1 & \mathbf{x}_2 \end{array} \right]
\end{aligned}
$$

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 **W**$^{(1)}$:

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

Analytical simplification

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

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

Outer product