STATISTICAL MACHINE LEARNING

COMP4670/8600

2025 Semester 1, Week 7, Lecture 2

Jing Jiang
School of Computing

Australian National University

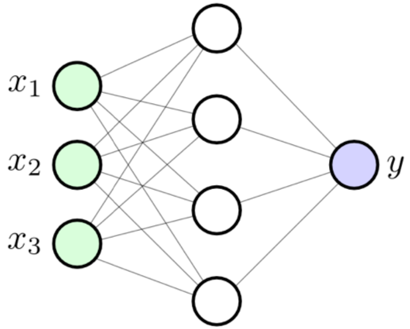
Administrative Matters

- Assignment 2 will be released at the end of Week 7
  - Due at the end of Week 12

Recap

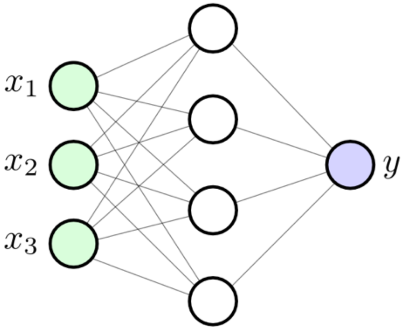
- Motivation
- Elements of a neural network
- Why deep neural networks?
- Forward pass → making predictions
- Backward pass or backpropagation → update model parameters

Terminologies



- It is useful to distinguish three types of units:
  - Input units (often denoted by X): input to the network
  - Hidden units (often denoted by Z): receive data from and send data to units within the network
  - Output units: the type of the output depends on the task (e.g., regression, binary or multiclass classification). In many cases, there is only one scalar output unit.
- Given inputs X, a neural network produces an output y

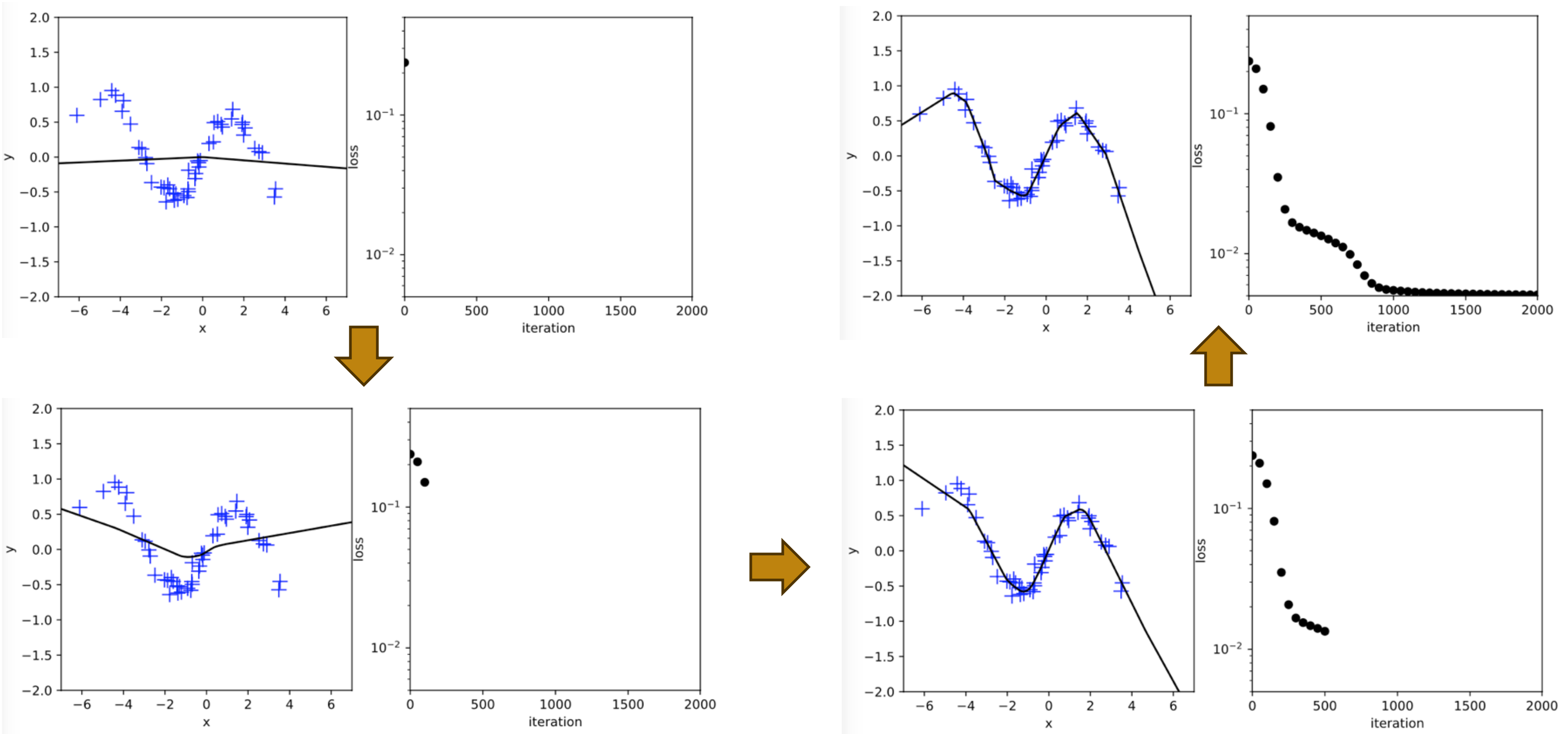
Terminologies



- A feedforward or fully-connected neural net includes the following:
  - A set of processing units (also called neurons or nodes)
  - L unit layers (one input layer, one output layer, and $L-2$ hidden layers), and $L-1$ weight layers
  - Weights $w_{i,k}^{l}$, which are connection strengths from unit i of the l-th layer to unit k of the $(l+1)$-th layer
  - A propagation rule that determines the total input or activation S_k of unit k, from the outputs in the previous layer that are connected to unit k
  - The output Z_k for each unit k, which is a function of the activation S_k, $Z_k = h_k(S_k)$, where h_k is an activation function

Example: using a feedforward NN for non-linear regression

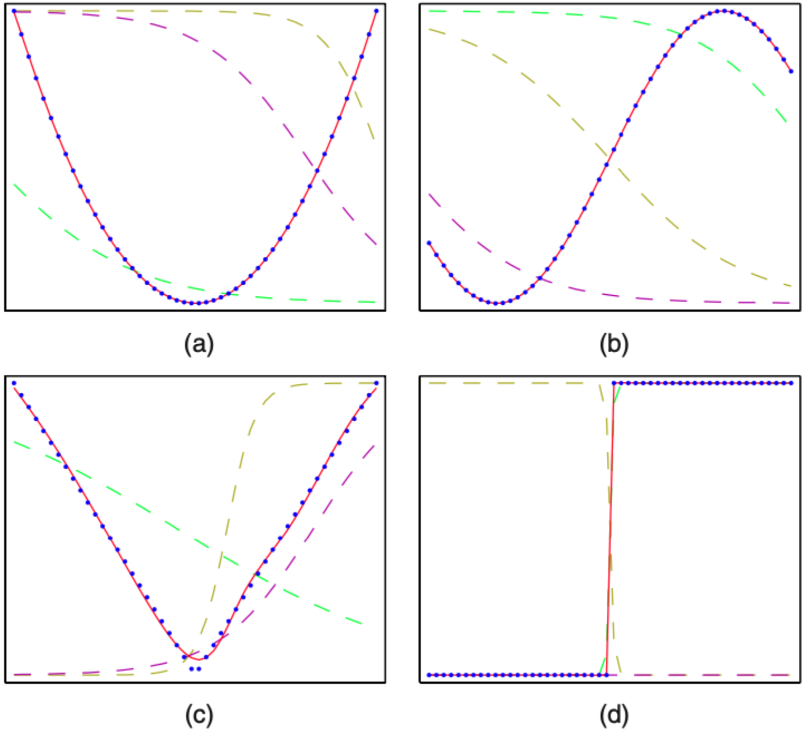
- Two hidden layers with 20 ReLU units



Capability of Feedforward Neural Networks

- Fitting the following functions:
  - (a) $f(x) = x^2$
  - (b) $f(x) = \sin x$
  - (c) $f(x) = |x|$
  - (d) $f(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$

    (Heaviside step function)
- A feedforward NN with one hidden layer of 3 hidden units is used
- The three dashed curves show the outputs of the hidden units



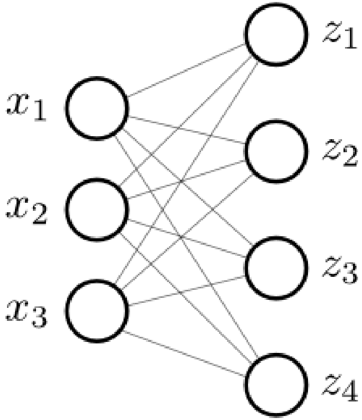
example from Bishop 5.1 (Figure 5.3)

Recap

- Motivation
- Elements of a neural network
- Why deep neural networks?
- **Forward pass → making predictions**
- Backward pass or backpropagation → update model parameters

Computation in a single layer: an example

Consider a single weight layer connecting two hidden layers in a network.



Computation in a single layer: an example

We can put things together:

$$[S_1 \quad S_2 \quad S_3 \quad S_4] = [x_1 \quad x_2 \quad x_3] \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \end{bmatrix} + [b_1 \quad b_2 \quad b_3 \quad b_4] \,.$$

or

$$\mathbf{S} = \mathbf{x}\mathbf{W} + \mathbf{b}$$

where $\mathbf{S}$, $\mathbf{b}$ are row vectors with 4 elements, $\mathbf{x}$ is a row vector with 3 elements, and W is a matrix of size 3×4.

Computation in a single layer: an example

We can put things together:

$$\begin{bmatrix} S_1 & S_2 & S_3 & S_4 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \end{bmatrix}.$$

or

$$\mathbf{S} = \mathbf{x}\mathbf{W} + \mathbf{b}$$

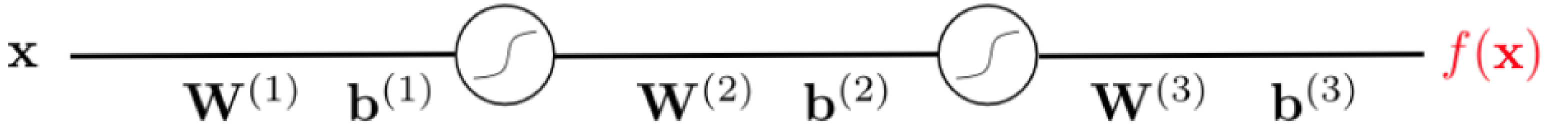
where $\mathbf{S}$, $\mathbf{b}$ are row vectors with 4 elements, $\mathbf{x}$ is a row vector with 3 elements, and W is a matrix of size 3×4.

We can then apply the activation function:

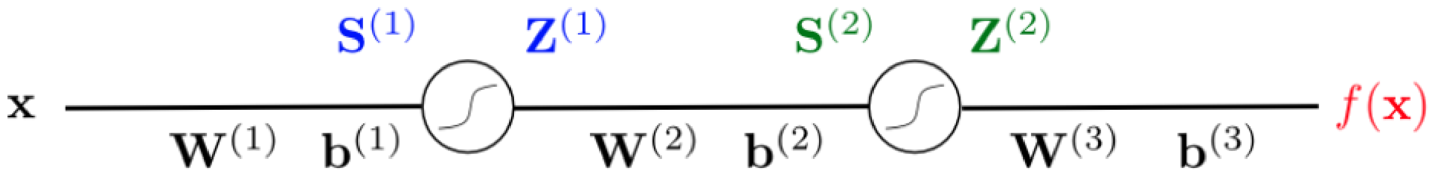
$$\mathbf{Z} = \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \end{bmatrix} = \begin{bmatrix} h(S_1) & h(S_2) & h(S_3) & h(S_4) \end{bmatrix} := h(\mathbf{S})$$

We can then use $\mathbf{Z} = \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \end{bmatrix}$ as the input to the next layer.

Computation with multiple layers: forward pass



Computation with multiple layers: forward pass



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \qquad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \qquad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

Note: We do not apply the activation function at the output layer.

Agenda

- Motivation
- Elements of a neural network
- Why deep neural networks?
- Forward pass → making predictions
- Backward pass or backpropagation → update model parameters

Network Training – Loss Function

Let's consider a **regression** problem with the following training data:

$$D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$$

Network Training – Loss Function

Let's consider a regression problem with the following training data:

$$D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$$

Let's assume that the output y follows a Gaussian distribution with an $\mathbf{x}$-dependent mean:

$$y_n = f_\theta(\mathbf{x}_n) + \epsilon_n$$

where $f_\theta(\mathbf{x}_n)$ is the value of the output node of the neural network, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is noise, and θ denotes all the weights in the neural network

Network Training – Loss Function

The likelihood function is as follows:

$$\prod_{n=1}^{N} p(y_n | \mathbf{x}_n, \theta, \sigma^2)$$

Network Training – Loss Function

The likelihood function is as follows:

$$\prod_{n=1}^{N} p(y_n | \mathbf{x}_n, \theta, \sigma^2)$$

Taking negative logarithm, we get the following loss function:

$$\frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(y_n - f_\theta(\mathbf{x}_n)\right)^2 + \frac{N}{2} \log(2\pi\sigma^2)$$

Network Training – Loss Function

The likelihood function is as follows:

$$\prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \theta, \sigma^2)$$

Taking negative logarithm, we get the following loss function:

$$\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n - f_\theta(\mathbf{x}_n)\right)^2 + \frac{N}{2}\log(2\pi\sigma^2)$$

If we adopt maximum likelihood estimation (MLE), we will need to minimise the loss function above.

Network Training – Loss Function

$$\frac{1}{2\sigma^2}\sum_{n=1}^{N}\big(y_n - f_\theta(\mathbf{x}_n)\big)^2 + \frac{N}{2}\log(2\pi\sigma^2)$$

Finding the optimal θ that minimises the above loss function is equivalent to finding the optimal θ that minimises the following *sum-of-squares* loss function:

$$\mathcal{L}(\theta) = \frac{1}{2}\sum_{n=1}^{N}\big(y_n - f_\theta(\mathbf{x}_n)\big)^2$$

Network Training – Loss Function

Let's consider a **binary classification** problem with the following training data:

$$D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$$

where y_n is either 0 or 1

Network Training – Loss Function

Let's consider a binary classification problem with the following training data:

$$D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$$

where y_n is either 0 or 1

Let's assume that

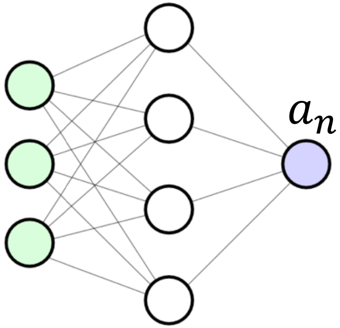
$$f_\theta(\mathbf{x}_n) = \sigma(a_n) = \frac{1}{1 + \exp(-a_n)}$$

where a_n is the activation of the output node, i.e., $f_\theta(\mathbf{x}_n)$ is the value of the output node after applying sigmoid as the activation function of the output node

Network Training – Loss Function

Let's consider a binary classification problem with the following training data:

$$D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$$



where y_n is either 0 or 1

Let's assume that

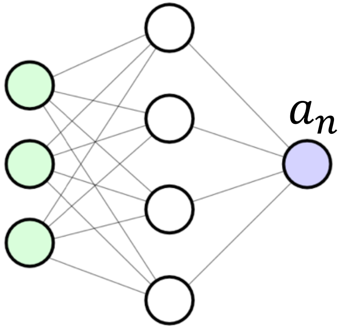
$$f_\theta(\mathbf{x}_n) = \sigma(a_n) = \frac{1}{1 + \exp(-a_n)}$$

where a_n is the activation of the output node, i.e., $f_\theta(\mathbf{x}_n)$ is the value of the output node after applying sigmoid as the activation function of the output node

Network Training – Loss Function

Let's consider a binary classification problem with the following training data:

$$D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$$



where y_n is either 0 or 1

Let's assume that

$$f_\theta(\mathbf{x}_n) = \sigma(a_n) = \frac{1}{1 + \exp(-a_n)}$$

where a_n is the activation of the output node, i.e., $f_\theta(\mathbf{x}_n)$ is the value of the output node after applying sigmoid as the activation function of the output node

Let's interpret $f_\theta(\mathbf{x}_n)$ as the conditional probability $p(y_n = 1|\mathbf{x}_n)$

Network Training – Loss Function

The likelihood function is as follows:

$$\prod_{n=1}^{N} f_\theta(\mathbf{x}_n)^{y_n}\left(1 - f_\theta(\mathbf{x}_n)\right)^{1-y_n}$$

Network Training – Loss Function

The likelihood function is as follows:

$$\prod_{n=1}^{N} f_\theta(\mathbf{x}_n)^{y_n} \left(1 - f_\theta(\mathbf{x}_n)\right)^{1-y_n}$$

Taking negative logarithm, we get the following *cross-entropy* loss function:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} \left(y_n \log f_\theta(\mathbf{x}_n) + (1 - y_n) \log\left(1 - f_\theta(\mathbf{x}_n)\right)\right)$$

Network Training – Loss Function

In summary

- For regression we use a linear output and a sum-of-squares loss function
- For binary classification we use a logistic sigmoid output and a cross-entropy loss function

Network Training – Gradient Descent

Let $\mathcal{L}(\theta)$ denote the loss function.

Our goal when training our neural network:

$$\theta^* \leftarrow \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta)$$

Network Training – Gradient Descent

Let $\mathcal{L}(\theta)$ denote the loss function.

Our goal when training our neural network:

$$\theta^* \longleftarrow \underset{\theta}{\operatorname{argmin}} \mathcal{L}(\theta)$$

Let us now consider gradient descent to find a local minimum of $\mathcal{L}(\theta)$.

Network Training – Gradient Descent

Let $\mathcal{L}(\theta)$ denote the loss function.

Our goal when training our neural network:

$$\theta^* \longleftarrow \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta)$$

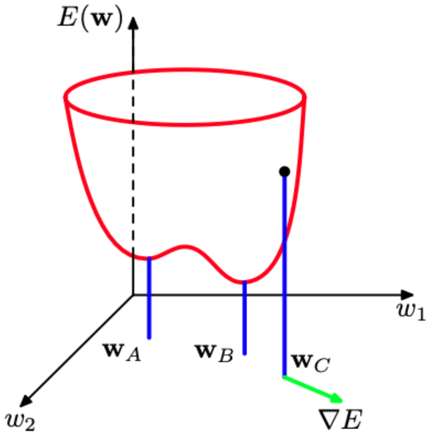


image from Bishop 5.2 (Figure 5.5)

Let us now consider gradient descent to find a local minimum of $\mathcal{L}(\theta)$.

Network Training – Gradient Descent

Let $\mathcal{L}(\theta)$ denote the loss function.

Our goal when training our neural network:

$$\theta^* \leftarrow \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta)$$

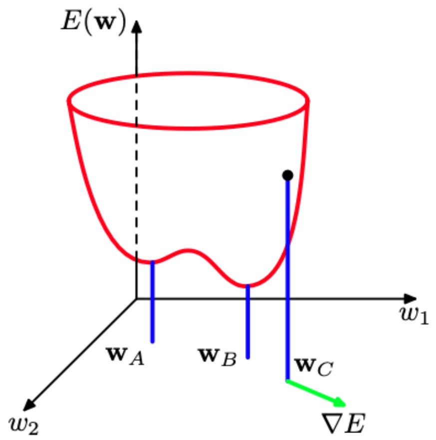


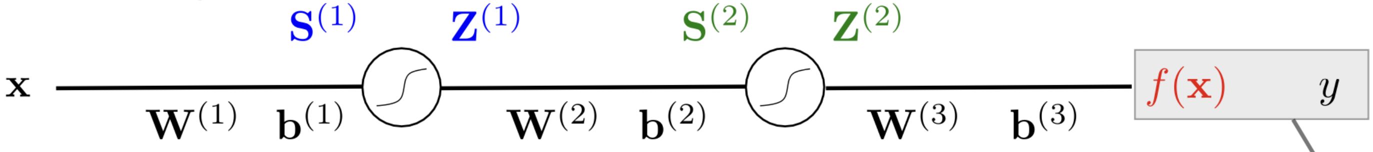
image from Bishop 5.2 (Figure 5.5)

Let us now consider gradient descent to find a local minimum of $\mathcal{L}(\theta)$.

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right)$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

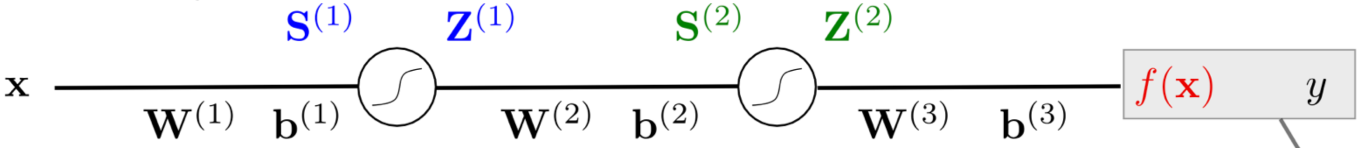
$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \quad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

Let's first consider loss for a single data point (x, y) (instead of the loss of the entire training set).

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \quad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

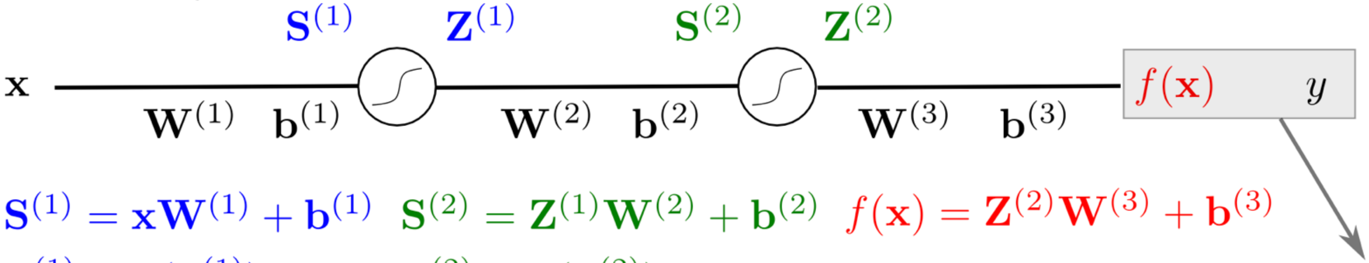
$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \qquad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \qquad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

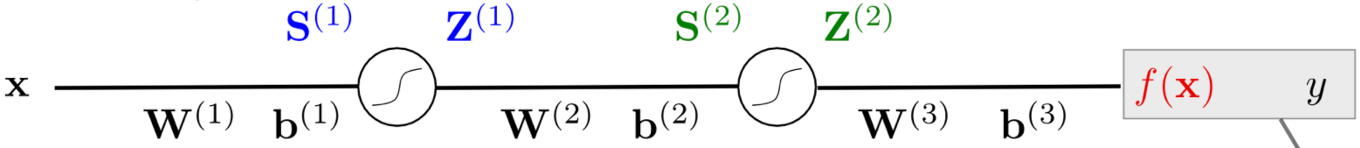
chain rule

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \quad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

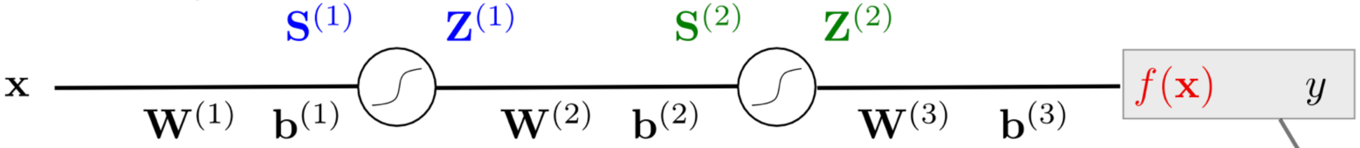
$$\frac{\partial f(x)}{\partial \mathbf{W}^{(3)}} = \frac{\partial\left(\mathbf{Z}^{(2)}\mathbf{W}^{(3)} + b^{(3)}\right)}{\partial \mathbf{W}^{(3)}} = \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \quad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

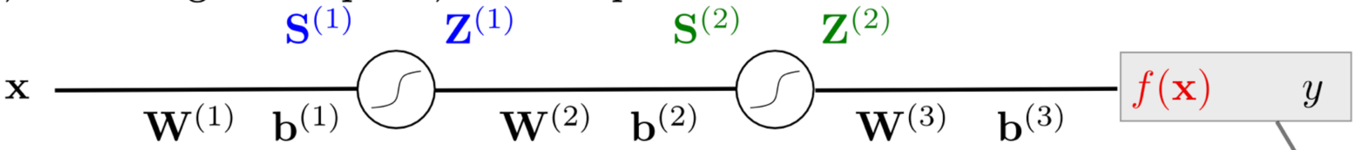
$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \quad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

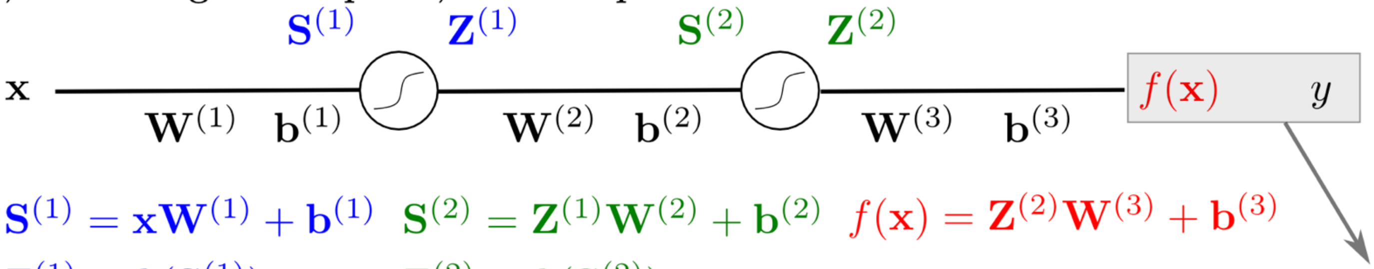
$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \qquad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \qquad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

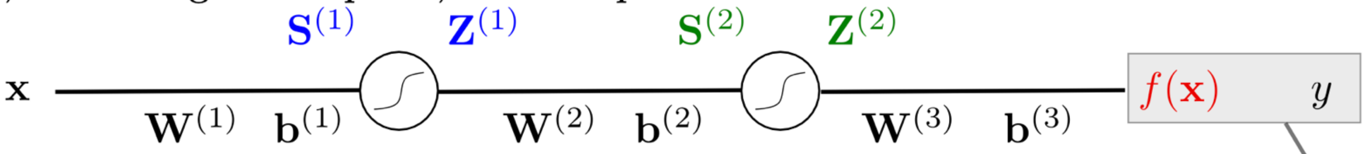
$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \quad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

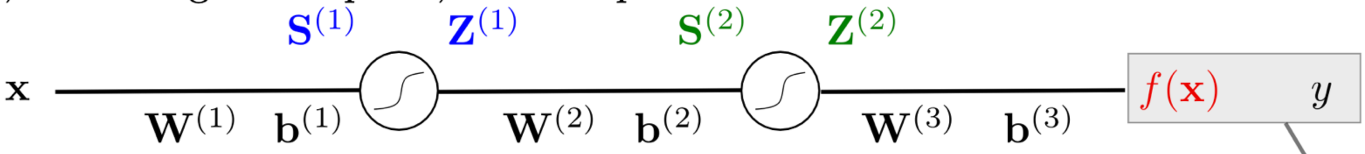
chain rule

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \qquad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \qquad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

similar

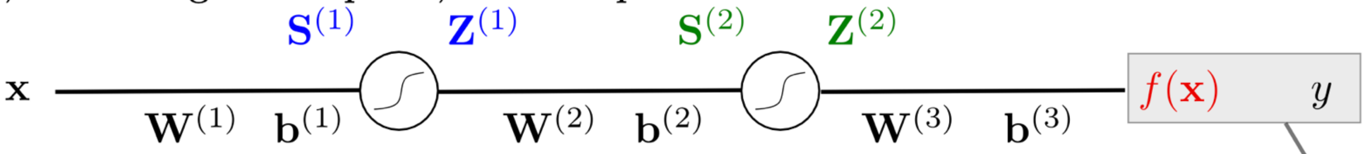
$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \quad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)}$$

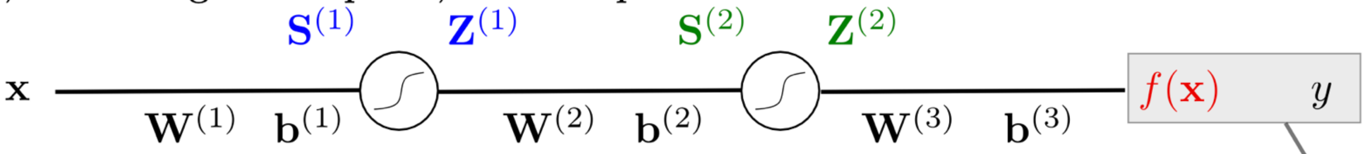
$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

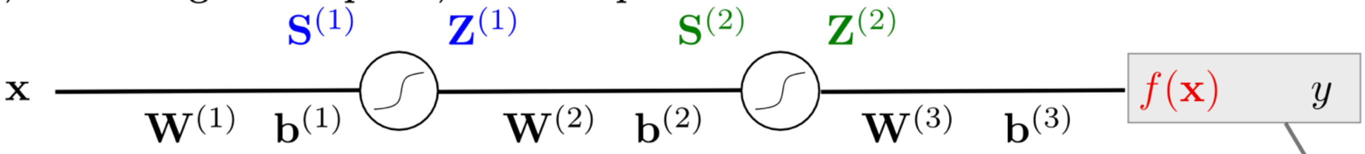
$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

chain rule

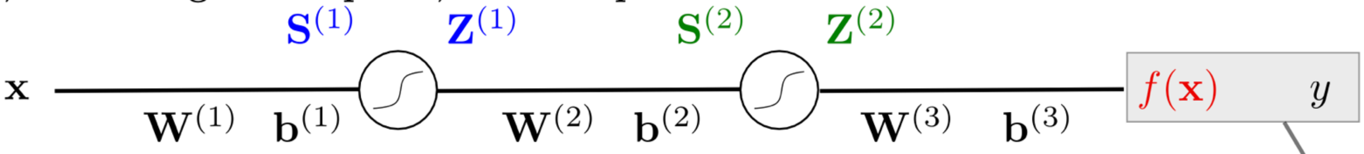
$$\frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} \frac{\partial \mathbf{Z}^{(2)}}{\partial \mathbf{S}^{(2)}} \qquad \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} \frac{\partial \mathbf{Z}^{(2)}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

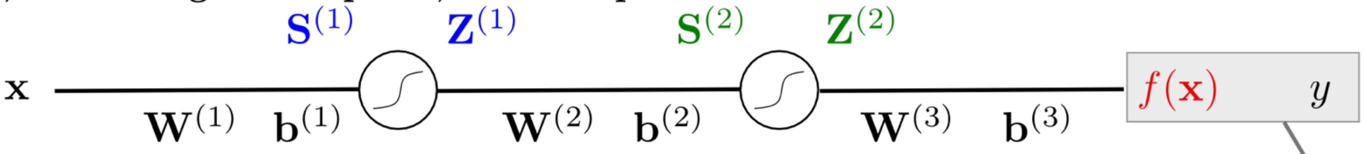
$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{W}^{(3)}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \qquad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \qquad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} \frac{\partial \mathbf{Z}^{(2)}}{\partial \mathbf{S}^{(2)}}$$

This depends on the activation function used.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

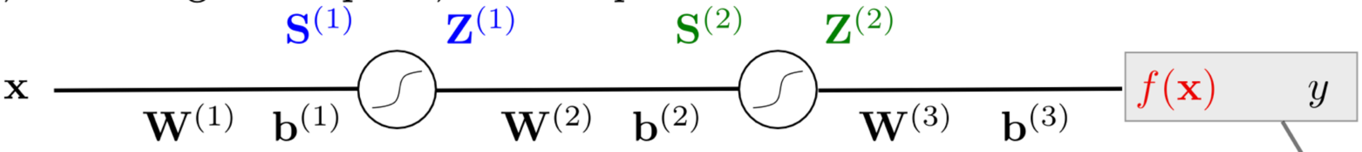
$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{W}^{(3)}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \qquad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \qquad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} \frac{\partial \mathbf{Z}^{(2)}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(1)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{Z}^{(1)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{W}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

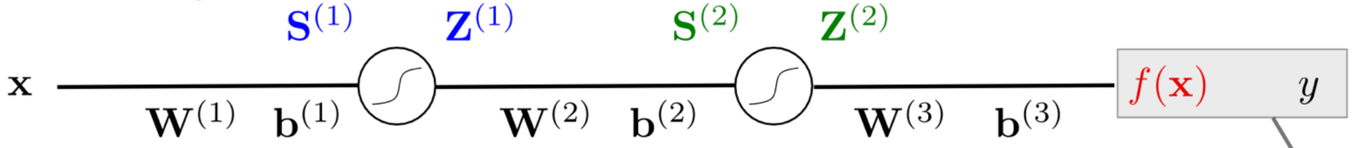
$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{W}^{(3)}$$

Backpropagation

For simplicity, consider a single input dimension and only one hidden unit in each hidden layer, and a single data point, aka. all quantities here are scalar.



$$\mathbf{S}^{(1)} = \mathbf{x}\mathbf{W}^{(1)} + \mathbf{b}^{(1)} \quad \mathbf{S}^{(2)} = \mathbf{Z}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)} \quad f(\mathbf{x}) = \mathbf{Z}^{(2)}\mathbf{W}^{(3)} + \mathbf{b}^{(3)}$$

$$\mathbf{Z}^{(1)} = h(\mathbf{S}^{(1)}) \qquad \mathbf{Z}^{(2)} = h(\mathbf{S}^{(2)})$$

$$\mathcal{L} = (y - f(\mathbf{x}))^2$$

Exercise

$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(1)}}, \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(1)}}$

$$\frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} = -2(y - f(\mathbf{x}))$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{b}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{Z}^{(2)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{W}^{(3)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(2)}} \frac{\partial \mathbf{Z}^{(2)}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{W}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{Z}^{(1)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{b}^{(2)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}^{(1)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \frac{\partial \mathbf{S}^{(2)}}{\partial \mathbf{Z}^{(1)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}^{(2)}} \mathbf{W}^{(2)}$$

Backpropagation – A closer look

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

Backpropagation – A closer look

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

- Recall that the gradient is used to update the weights:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right)$$

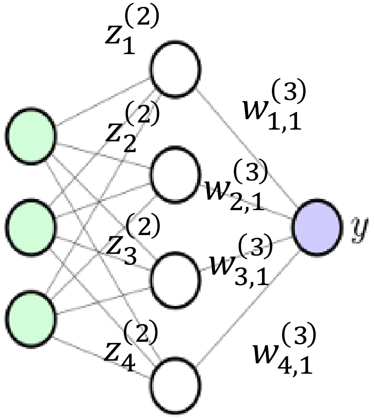
Backpropagation – A closer look

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

- Recall that the gradient is used to update the weights:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right)$$

- Assume that we have 4 hidden units in the previous layer



Backpropagation – A closer look

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

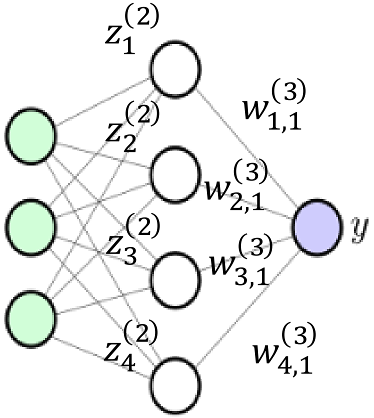
- Recall that the gradient is used to update the weights:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{\eta} \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right)$$

- Assume that we have 4 hidden units in the previous layer

$$\frac{\partial \mathcal{L}}{\partial w_{1,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_1^{(2)}, \qquad \frac{\partial \mathcal{L}}{\partial w_{2,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_2^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial w_{3,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_3^{(2)}, \qquad \frac{\partial \mathcal{L}}{\partial w_{4,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_4^{(2)}$$



Backpropagation – A closer look

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

- Recall that the gradient is used to update the weights:

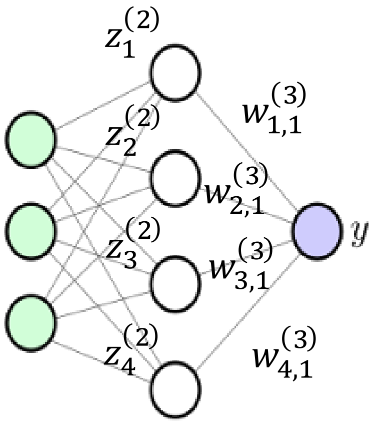
$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right)$$

- Assume that we have 4 hidden units in the previous layer

$$\frac{\partial \mathcal{L}}{\partial w_{1,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_1^{(2)}, \qquad \frac{\partial \mathcal{L}}{\partial w_{2,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_2^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial w_{3,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_3^{(2)}, \qquad \frac{\partial \mathcal{L}}{\partial w_{4,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_4^{(2)}$$

same value: $-2(y - f(x))$



Backpropagation – A closer look

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial \mathbf{W}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(\mathbf{x})} \mathbf{Z}^{(2)}$$

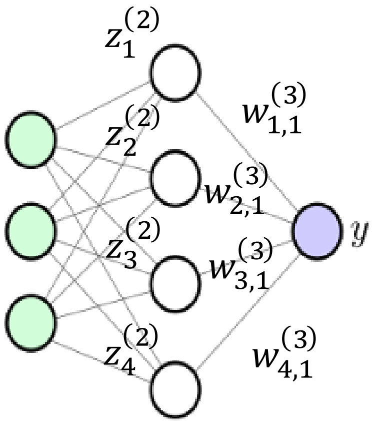
- Recall that the gradient is used to update the weights:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{\eta} \nabla \mathcal{L}\left(\boldsymbol{\theta}^{(t)}\right)$$

- Assume that we have 4 hidden units in the previous layer

$$\frac{\partial \mathcal{L}}{\partial w_{1,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_1^{(2)}, \quad \frac{\partial \mathcal{L}}{\partial w_{2,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_2^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial w_{3,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_3^{(2)}, \quad \frac{\partial \mathcal{L}}{\partial w_{4,1}^{(3)}} = \frac{\partial \mathcal{L}}{\partial f(x)} z_4^{(2)}$$

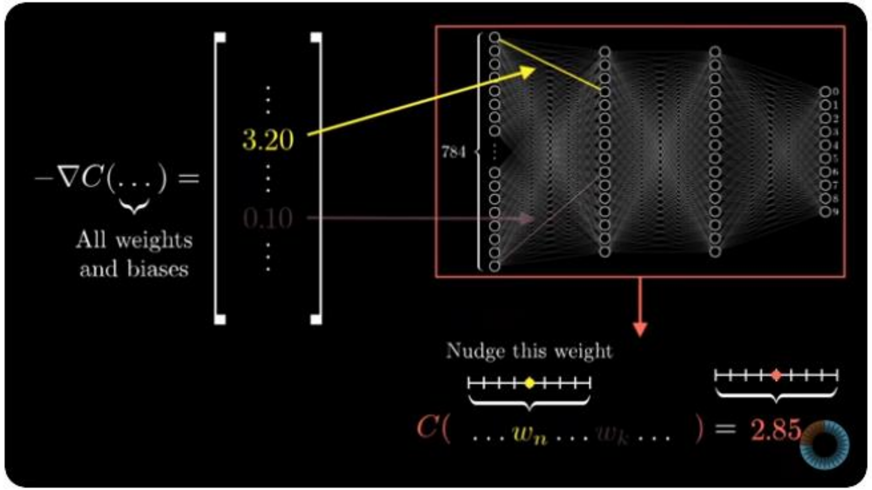


the output from the previous layer determines how much the corresponding weight needs to be adjusted

Backpropagation

- A nice video explaining backpropagation:

https://youtu.be/Ilg3gGewQ5U



Summary

- Recap of forward pass
- Choice of the activation function at the output layer and the error function
  - Regression
  - Classification
- Backpropagation

Reading

- Bishop 5.2, 5.3

Acknowledgments

- The slides are largely adopted from COMP4670/8600, 2024 Semester 1.