Neural Networks

Regression

Samples: $(x_1, y_1), (x_2, y_2), \_\_\_, (x_n, y_n)$

$x_i \in \mathbb{R}^d \quad , \quad y_i \in \mathbb{R}$

Find the best fit:

$$f(x_i ; \underline{\theta}) \approx y_i$$

parameters.

Loss function:

example

$$\mathcal{L}(\underline{\theta}) = \sum_{i=1}^{n} (y_i - f(x_i ; \underline{\theta}))^2$$

__Goal:__ Find $\underline{\theta}$ that minimise $\mathcal{L}(\underline{\theta})$

↳ Solve: $\nabla \mathcal{L}(\underline{\theta}) = 0 \longrightarrow$ not always solvable!

Gradient Descent

* "Searching" for the minimum
  - Choose $\underline{\theta}^{(0)}$
  - $\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} - \beta \nabla \mathcal{L}(\underline{\theta})$

Neural Networks:

- "Complicated" function f
- $\underline{\theta} \in \mathbb{R}^D \qquad D$ - very large. (many params.)

Feedforward networks:

$\underline{x}^{(0)} \in \mathbb{R}^d$

activation function

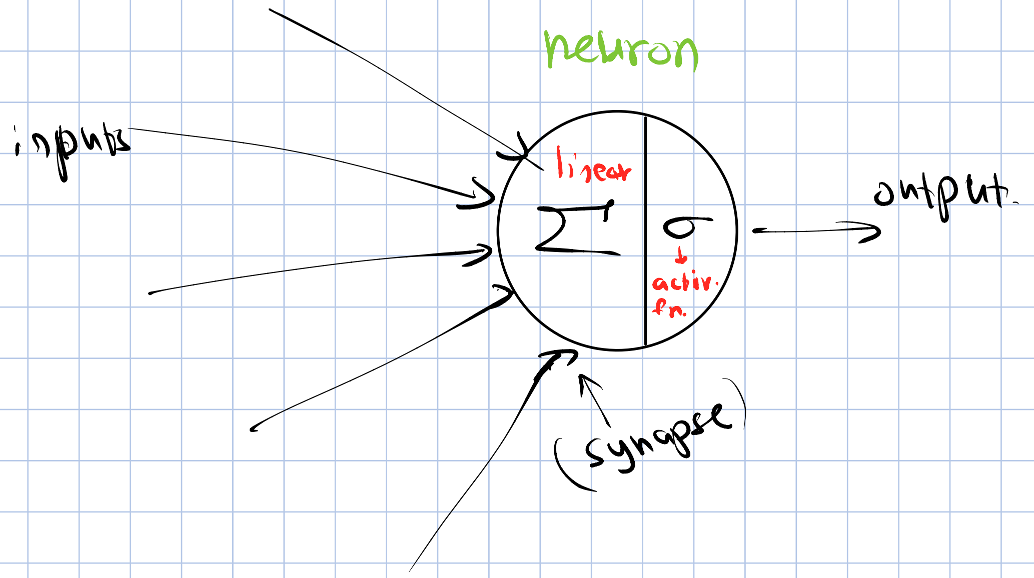
$$\underline{x}^{(1)} = \sigma^{(1)}(W^{(1)} \underline{x}^{(0)} + b^{(1)})$$

$$\underline{x}^{(2)} = \sigma^{(2)}(W^{(2)} \underline{x}^{(1)} + b^{(2)})$$

$$\vdots$$

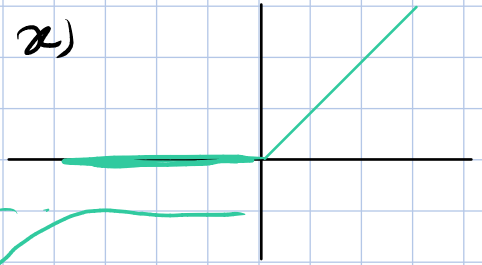
$$\underline{x}^{(L)} = \sigma^{(L)}(W^{(L)} \underline{x}^{(L-1)} + b^{(L)}) \longrightarrow f(\underline{x}^{(0)}, \underline{\theta})$$

$\underline{\theta} = (W^{(1)}, \_\_, W^{(L)}, b^{(1)}, \_\_ b^{(L)})$

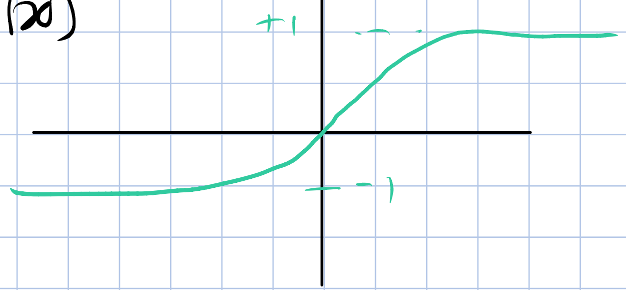


ACTIVATION FUNCTIONS:

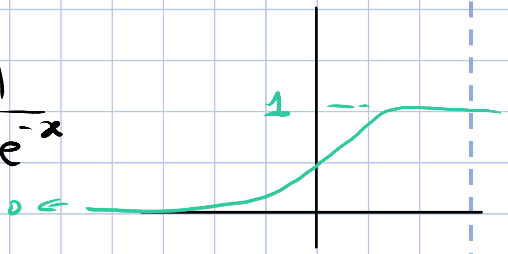
- ReLU$(x) = \max(0, x)$

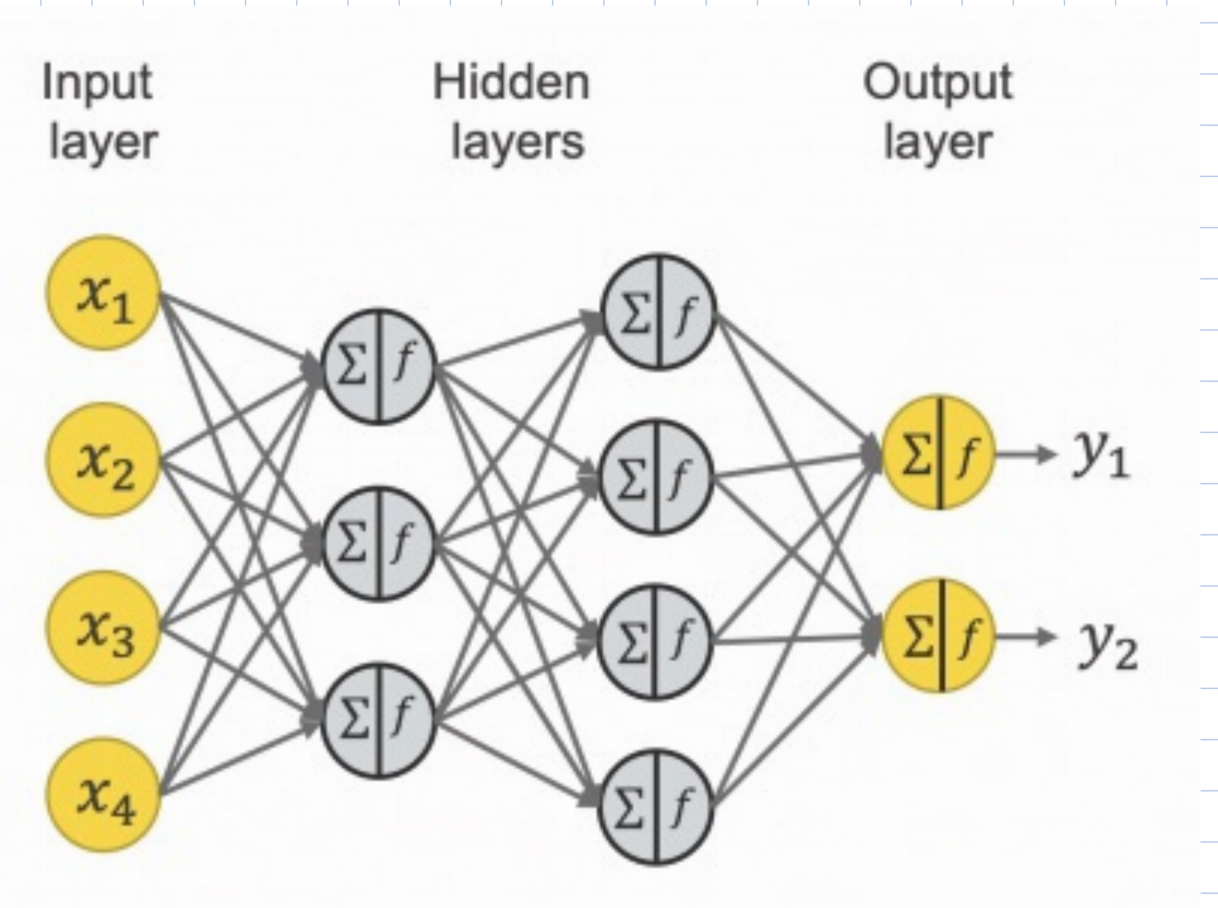


- $\tanh(x)$



- logistic function $= \dfrac{1}{1+e^{-x}}$





TRAINING A NN:

Training set $(x_1, y_1), \_\_\_, (x_n, y_n)$

Initialise $\underline{\theta}^{(0)}$ (possibly random)

Iterate:

(1) Feed forward:

$$x_1, \_\_\_, x_n \longrightarrow f(x_1, \underline{\theta}^{(k)}), \_\_\_, f(x_n, \underline{\theta}^{(k)})$$

(2) Compute loss:

$$\mathcal{L}(\underline{\theta}) = \sum_{i=1}^{n} (y_i - f(x_i ; \underline{\theta}^{(k)}))^2$$

efficient way to compute gradients for NN.

(3) Back-propagation: $\nabla \mathcal{L}(\underline{\theta}^{(k)})$

(4) Update parameters:

$$\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} - \beta \nabla \mathcal{L}(\underline{\theta}^{(k)})$$

step-size

STOCHASTIC GRADIENT DESCENT

$$\mathcal{L}(\underline{\theta}) = \mathcal{L}(\underline{\theta}; X) = \sum_{i=1}^{n} (y_i - f(x_i ; \underline{\theta}))^2 \quad (MSE)$$

$$\nabla \mathcal{L}(\underline{\theta}) = \sum_{i=1}^{n} \nabla \mathcal{L}(\underline{\theta} ; x_i)$$

n-large ⇒ slow

Gradient descent:

$$\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} - \beta \nabla \mathcal{L}(\underline{\theta} ; X)$$

Stochastic gradient descent:

- Chose $1 \leq i \leq n$ at random
- Update: $\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} - \beta \nabla \mathcal{L}(\underline{\theta} ; x_i)$

Mini-batch Gradient Descent:

- Set m (commonly 32-256)

- Take a random subset $X^{(m)} = \{x_{i_1}, \_\_, x_{i_m}\}$

- Update:

$$\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} - \beta \nabla J(\underline{\theta}; X^{(m)})$$

Common practice:

- Shuffle X.
- Split X into $X_1^{(m)}, X_2^{(m)}, \_\_\_, X_{\frac{n}{m}}^{(m)}$
- Update $\underline{\theta}$ using $X_1^{(m)}$

  " " " $X_2^{(m)}$

  ⋮

  " " " $X_{\frac{n}{m}}^{(m)}$

  epoch
- repeat

Back-Propagation

$$x \longrightarrow h_1(\cdot;\theta_1) \xrightarrow{z_1} h_2(\cdot;\theta_2) \xrightarrow{z_2} h_3(\cdot;\theta_3) \longrightarrow y$$

$$y = y(x;\theta_1,\theta_2,\theta_3) \quad (\text{ex. } y = \text{loss})$$

Q: What is $\dfrac{\partial y}{\partial \theta_1}$?

$$y = h_3(z_2;\theta_3) = h_3(h_2(z_1;\theta_2);\theta_3)$$

$$= h_3(h_2(h_1(x;\theta_1);\theta_2);\theta_3)$$

$$\frac{\partial y}{\partial \theta_1} = \frac{\partial y}{\partial z_2}\cdot\frac{\partial z_2}{\partial \theta_1} = \frac{\partial h_3}{\partial z_2}\cdot\frac{\partial z_2}{\partial \theta_1}$$

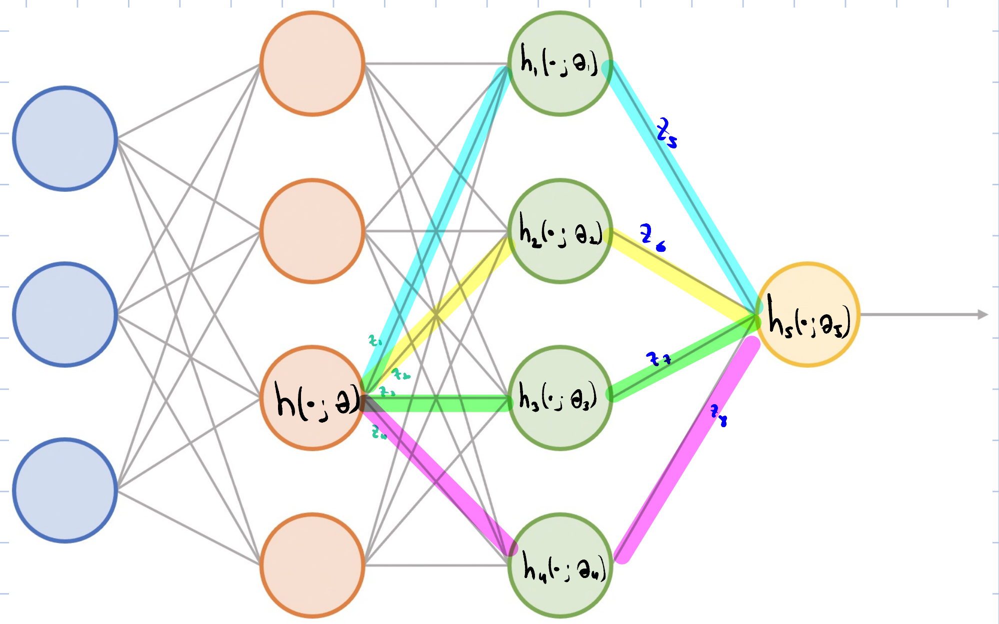
$$= \frac{\partial h_3}{\partial z_2}\cdot\frac{\partial h_2}{\partial z_1}\cdot\frac{\partial z_1}{\partial \theta_1}$$

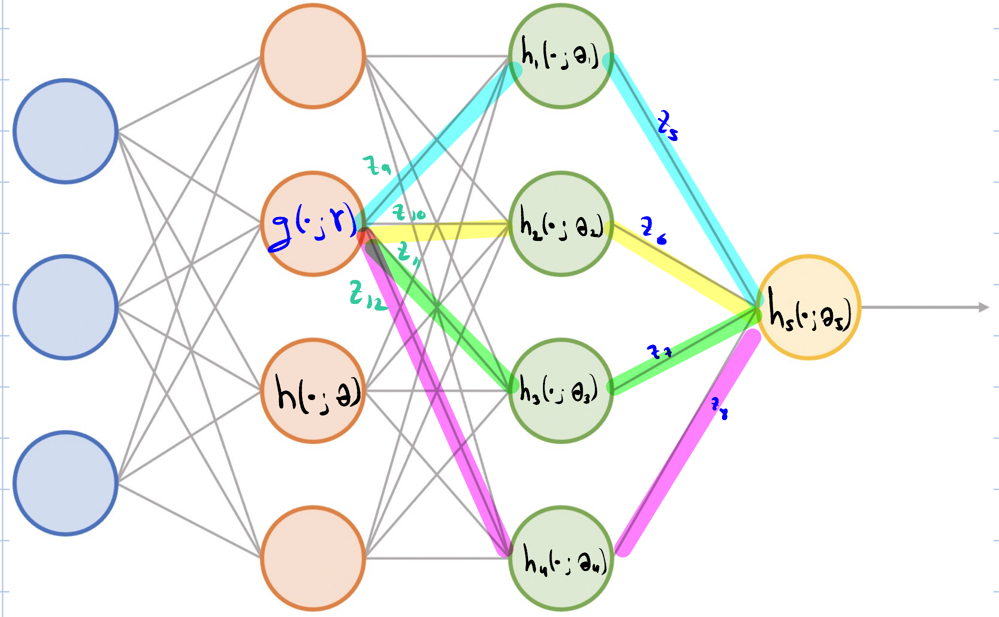
$$\frac{\partial y}{\partial \theta_1} = \frac{\partial h_3}{\partial z_2}\cdot\frac{\partial h_2}{\partial z_1}\cdot\frac{\partial h_1}{\partial \theta_1}$$

Feed-forward:

Take $x \rightarrow$ compute z_1, z_2, y.

$$\frac{\partial h_3}{\partial z_2}(z_2;\theta_3)\cdot\frac{\partial h_2}{\partial z_1}(z_1;\theta_2)\cdot\frac{\partial h_1}{\partial \theta_1}(x;\theta_1)$$





$$\frac{\partial y}{\partial \theta} = \frac{\partial h_5}{\partial z_5}\cdot\frac{\partial h_1}{\partial z_1}\cdot\frac{\partial h}{\partial \theta} + \frac{\partial h_6}{\partial z_6}\cdot\frac{\partial h_2}{\partial z_2}\cdot\frac{\partial h}{\partial \theta} + \frac{\partial h_7}{\partial z_7}\cdot\frac{\partial h_3}{\partial z_3}\cdot\frac{\partial h}{\partial \theta} + \frac{\partial h_8}{\partial z_8}\cdot\frac{\partial h_4}{\partial z_4}\cdot\frac{\partial h}{\partial \theta}$$

same values computed once

$$\frac{\partial y}{\partial \theta} = \frac{\partial h_5}{\partial z_5}\cdot\frac{\partial h_1}{\partial z_9}\cdot\frac{\partial g}{\partial \gamma} + \frac{\partial h_6}{\partial z_6}\cdot\frac{\partial h_2}{\partial z_{10}}\cdot\frac{\partial g}{\partial \gamma} + \frac{\partial h_7}{\partial z_7}\cdot\frac{\partial h_3}{\partial z_{11}}\cdot\frac{\partial g}{\partial \gamma} + \frac{\partial h_8}{\partial z_8}\cdot\frac{\partial h_4}{\partial z_{12}}\cdot\frac{\partial g}{\partial \gamma}$$