FYS5429, FEB 1, 2023

Basics of NNs

Architecture (Model)
- # hidden layers
- # nodes/neurons/units
- # activation function

Cost function, optimization + regularization

- Cost function

  Regression $MSE = \frac{1}{n} ||(t-y)||_2^2$

  $||x||_2 = \sqrt{\sum_i x_i^2}$

  target

  model output

  $y = y(\theta; x)$ our Model

  $x =$ input data

  $t = f(x)$

  $\theta =$ unknown parameter of our model

  cost function $C(\theta)$

  $$\hat{\theta} = \underset{\theta \in \mathbb{R}^m}{\arg\min}\, C(\theta)$$

- Optimization
  - Gradient descent (GD)
  - GD with momentum

- stochastic GD (SGD)

- algorithms for learning rate
  - Adagrad
  - RMS prop
  - ADAM
  - ...

- Regularization

$$\ell_1 \sim \lambda \|\theta\|_1$$

$$\ell_2 \sim \lambda \|\theta\|_2$$

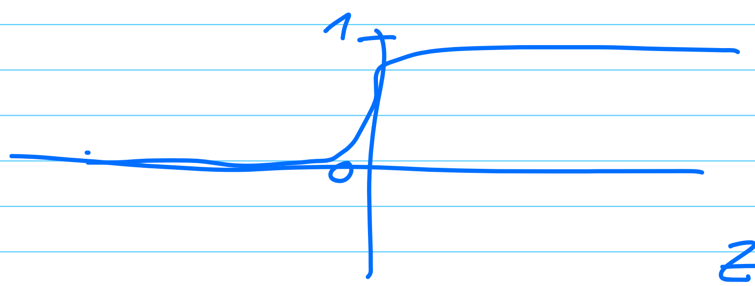
Universal approx theorem

Cybenko, 1989

Let σ be any continuous sigmoidal function

$$\sigma(z) \to \begin{cases} 1 & \text{as } z \to \infty \\ 0 & \text{as } z \to -\infty \end{cases}$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



Given a function $F \in C[0,1]^d$ and $\varepsilon > 0$, there is a one-hidden layer network $f(x;\theta)$

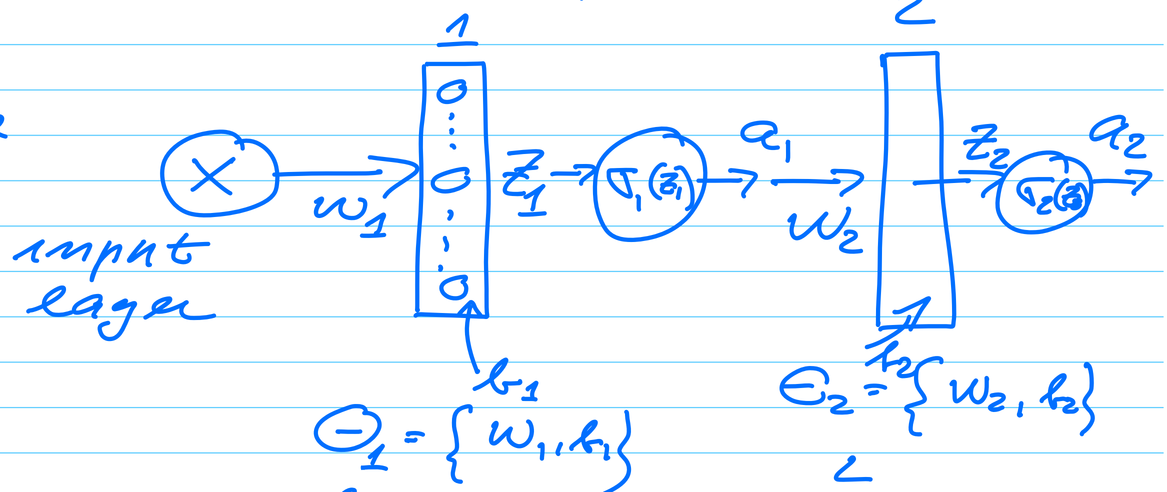
$$\Theta = \{W, b\}$$

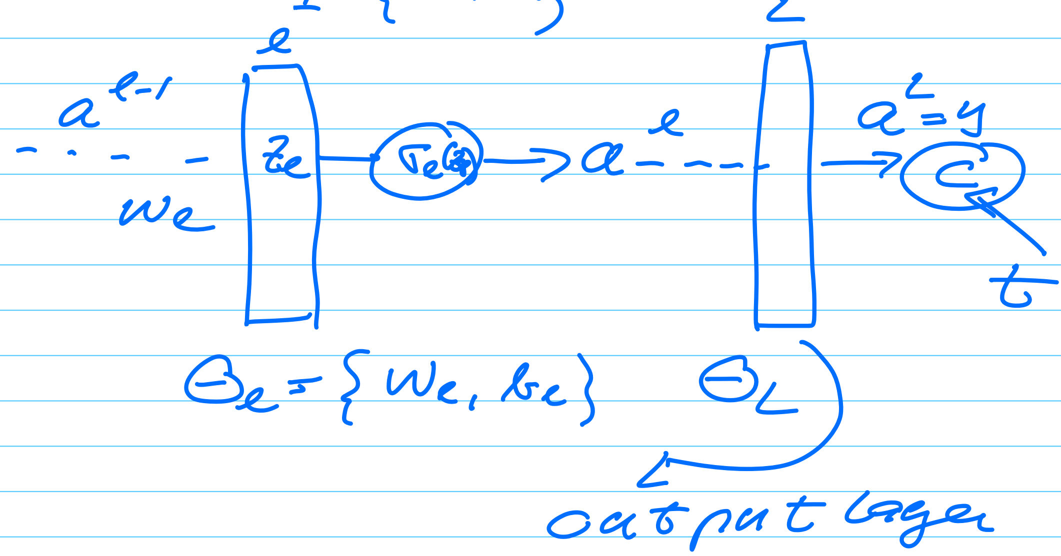
$$|f(x;\Theta) - F(x)| < \varepsilon$$

for all $x \in C[0,1]^d$

Hornik (1991) extended the theorem to apply to any by letting any non-constant bounded activation function to be included.

NN structure





$$\hat{\Theta} = \arg\min_{\Theta} C(\Theta)$$

$$C(\Theta) = \frac{1}{2}\left(t - a^L(x;\Theta_L)\right)^2$$

$$= \frac{1}{2}\left(t - y(x;\Theta_L)\right)^2 \quad (MSE)$$

Back propagation algo :

$$\frac{\partial C(\Theta)}{\partial \Theta_L} = 0 =$$

$$- \left(t - y(x;\Theta_L)\right)\frac{\partial y}{\partial \Theta_L}$$

$$= -\left(t - a_L(x;\Theta_L)\right)\frac{\partial a_L}{\partial \Theta_L}$$

Simple example

$$x \xrightarrow{w_1} \underset{\text{hidden}}{\overset{\sigma_1(x;w_1)}{\bigcirc}} \longrightarrow a_1 \xrightarrow{w_2} \overset{\sigma_2(a_1;w_2)}{\bigcirc} \xrightarrow{y=a_2} C \leftarrow t$$

$$\frac{\partial C}{\partial w_2} = 0$$

$$a_1 = \sigma_1(x;w_1)$$

$$y = a_2 = \sigma_2(a_1;w_2) =$$

$$\sigma_2(\sigma_1(x; w_1), w_2) = a_2$$

$$C = \frac{1}{2}(t - y)^2 = \frac{1}{2}(t - a_2)^2$$

$$\frac{\partial C}{\partial w_1} = ?$$

$$\frac{\partial C}{\partial w_2} = ?$$

activation functions:

$$\sigma_1(x; w_1) = x \cdot w_1 = a_1$$

$$\sigma_2(a_1; w_2) = a_1 w_2 =$$

$$w_2\, \sigma_1(x; w_1)$$

$$= y = a_2$$

$$\frac{\partial C}{\partial w_1} = -(t-y) \frac{dy}{da_1} \frac{\partial a_1}{\partial w_1} = \frac{\partial C}{\partial a_2} \frac{\partial a_2}{\partial a_1} \frac{\partial a_1}{\partial w_1}$$

$$\frac{\partial C}{\partial w_2} = -(t-y) \frac{\partial y}{\partial w_2} = -(t-y) \frac{\partial a_2}{\partial w_2}$$

$$\frac{\partial C}{\partial w_1} = -(t-y) \times w_2$$

$$\frac{\partial C}{\partial w_2} = -(t-y)\, w_1 x$$

$$w_1^{(k+1)} \Leftarrow w_1^{(k)} - \eta \frac{\partial C}{\partial w_1}\bigg|_{w_1 = w_1^{(k)}}$$

$$w_2^{(k+1)} \Leftarrow w_2^{(k)} - \eta \frac{\partial C}{\partial w_2}\bigg|_{w_2 = w_2^{(k)}}$$

$$\frac{\partial C}{\partial \theta_L} = \frac{\partial}{\partial \theta_L}\left(C\left(t, a_L(\theta_L; x)\right)\right)$$

$$\text{if } C = MSE = \frac{1}{2} \| t - a_L \|_2^2$$

$$\frac{\partial C}{\partial \theta_L} = -(t - a_L) \frac{\partial a_L}{\partial \theta_L}$$

$$\frac{\partial C}{\partial \theta_{L-1}} = \frac{\partial C}{\partial a_L} \frac{\partial a_L}{\partial \theta_{L-1}}$$

$$\frac{\partial C}{\partial \theta_{L-2}} = \frac{\partial C}{\partial a_L} \frac{\partial a^L}{\partial a_{L-1}} \frac{\partial a_{L-1}}{\partial \theta_{L-2}}$$

$$\frac{\partial C}{\partial \theta_\ell} = \frac{\partial C}{\partial a_L} \frac{\partial a_L}{\partial a_{L-1}} \frac{\partial a_{L-1}}{\partial a_{L-2}} \cdots$$

$$\cdots \frac{\partial a_{\ell+2}}{\partial a_{\ell+1}} \frac{\partial a_{\ell+1}}{\partial \theta_\ell}$$

Auto matic Diff — in python import autograd

JAX

Example

$$f(x) = \sqrt{x^2 + \exp(x^2)}$$

$$\frac{df}{dx} = \frac{x\,(1 + \exp(x^2))}{\sqrt{x^2 + \exp(x^2)}}$$

$f(x)$ in a brute force way

$x^2 = x \cdot x = 1$ FLOP

$\exp(x^2) = \exp(x \cdot x) = 2$ FLOPS

$x^2 + \exp(x^2) = 1$ FLOP

sqrt = — " —

5 FLOP.

$a = x^2$ $\exp(a)$ =>

4 FLOP.

Derivative

Numerator : 4 FLOPS

Denominator : 5 FLOPS

+ Division : 1 FLOP

10 FLOPS

$$\frac{df}{dx} \simeq \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

Automatic diff:

$$f(x) = \sqrt{x^2 + \exp(x^2)}$$

- Forward mode
- Reverse mode (AKA Back prop)

$a = x^2 \qquad b = \exp(x^2) = \exp(a)$

$c = a + b$

$d = \sqrt{c} = f(x)$

$$\begin{tikzcd}
 & & b \arrow[dr] & & & \\
x \arrow[r, "{[\cdot]^2}"] & a \arrow[ur, "{\exp[\cdot]}"] \arrow[rr] & & \boxed{+} \arrow[r] & c \arrow[r, "\sqrt{[\cdot]}"] & d
\end{tikzcd}$$

$\rightarrow b$

$$\frac{df}{dx} = ?$$

$$\frac{da}{dx} = 2x \qquad \frac{db}{dx} = \frac{db}{da}\frac{da}{dx}$$

$$= 2x \exp(x^2)$$