FYS-STK 3155/4155, Oct 13, 2022

Important components of NNs

NN architecture (model)

- # hidden layers
- # hidden nodes in a layer
- Fully connected
- activation functions (sigmoid, tanh, ReLU, ELU, ...)

Cost function & optimization

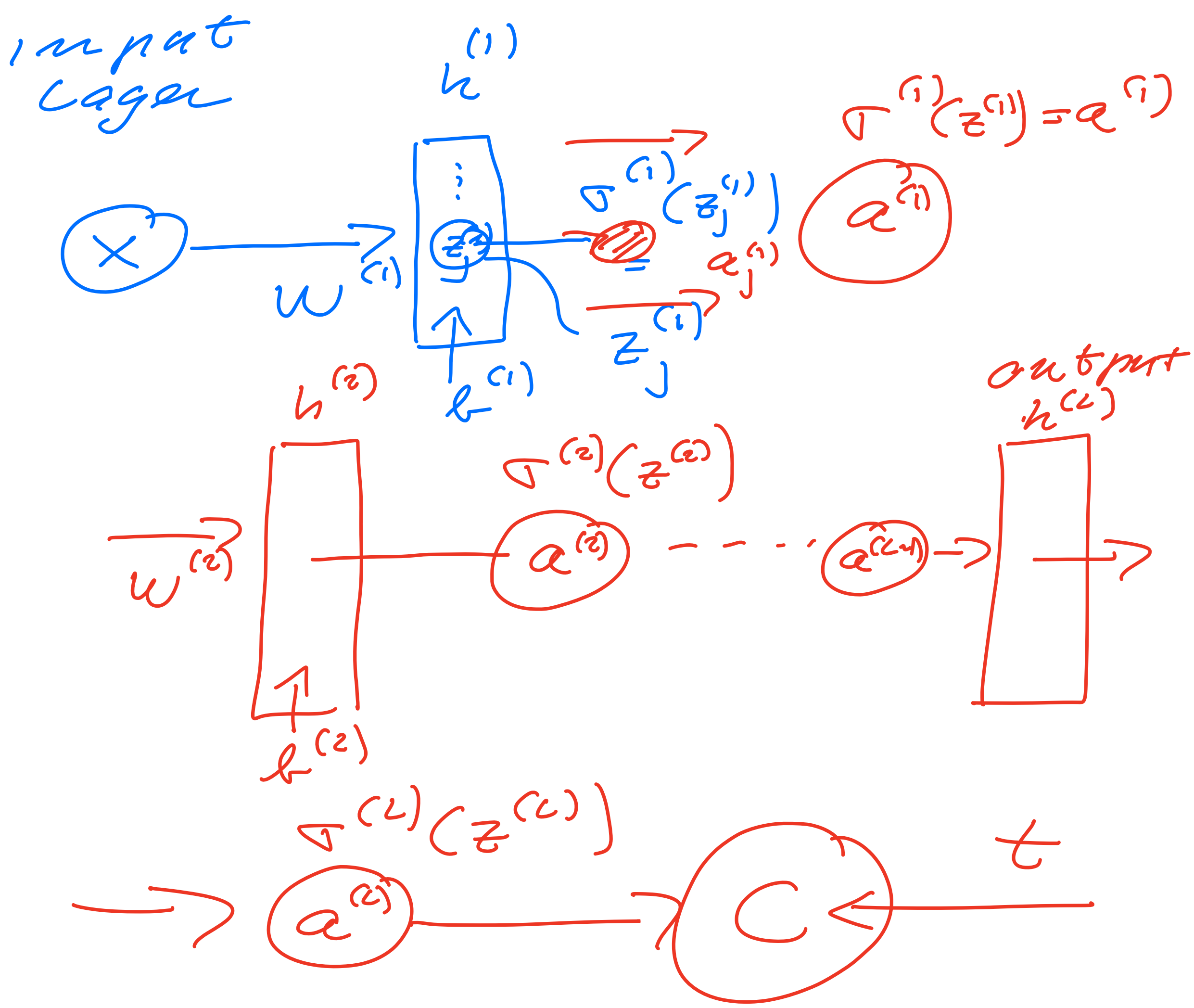
- Type of cost/loss function
- Regularization
- Gradient descent method
- SGD batch and step size (learning)

optimization choices

other specifics: Pre-processing initialization, dropout, batch normalization etc

Basic set up of an NN

– Feed-Forward stage



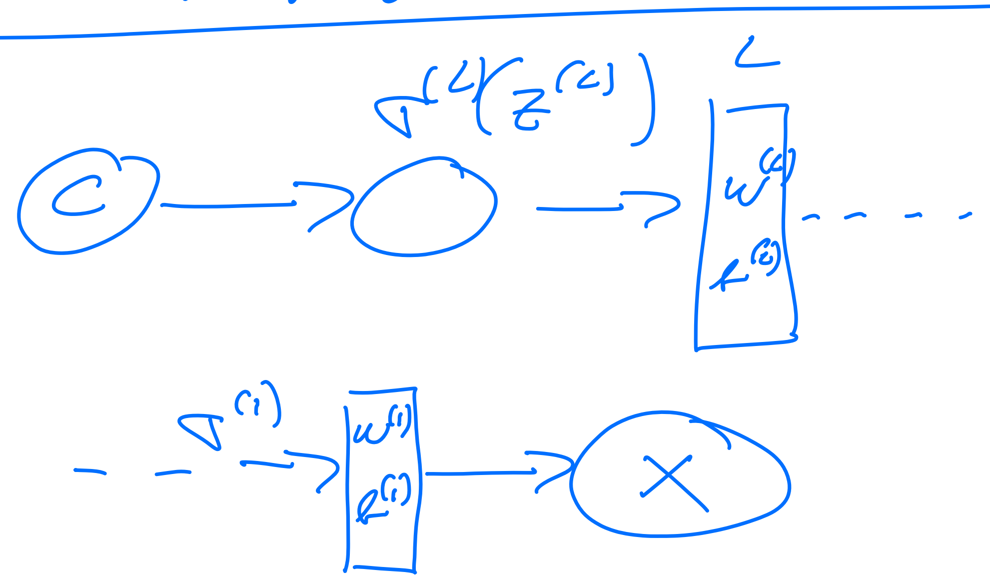
$$\tilde{y} = a^{(L)}(\Theta, x)$$

$$\Theta = \{ w^{(1)}, b^{(1)}, \ldots, w^{(L)}, b^{(L)} \}$$

$$= \sigma^{(L)}(\sigma^{(L-1)}(\sigma^{(L-2)} \cdots \sigma^{(1)}(xw^{(1)} + b^{(1)}) \cdots)$$

$$C(\Theta) = \frac{1}{n} \| (t - a^{L}(\Theta, x) \|_2^2$$

Back propagation algorithm



Back propagation

$$\frac{\partial C}{\partial \theta^{L-1}} = \frac{\partial C}{\partial \sigma^{(L)}} \frac{\partial \sigma^{(L)}}{\partial \theta^{L-1}}$$

$$= \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial \theta^{L-1}}$$

$$\frac{\partial C}{\partial \theta^{L-2}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial \theta^{(L-2)}}$$

$$\vdots$$

$$\frac{\partial C}{\partial \theta^{l}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial a^{(L-1)}} \cdots \frac{\partial a^{(l+2)}}{\partial a^{(l+1)}} \frac{\partial a^{(l+1)}}{\partial \theta^{(l)}}$$

Automatic differentiation

Example

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

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

Def $\quad a = x^2$

$\qquad\quad b = \exp(a) = f(x)$

$$x \longrightarrow \overset{x^2}{a} \longrightarrow \overset{\exp(x^2)}{b} \longrightarrow f$$

$$\frac{df}{dx} = \left[\frac{df}{db}\,\frac{db}{da}\right]\frac{da}{dx}$$

Reverse mode

$$= \frac{df}{db}\left[\frac{db}{da}\,\frac{da}{dx}\right]$$

Forward mode

Example 2

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

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

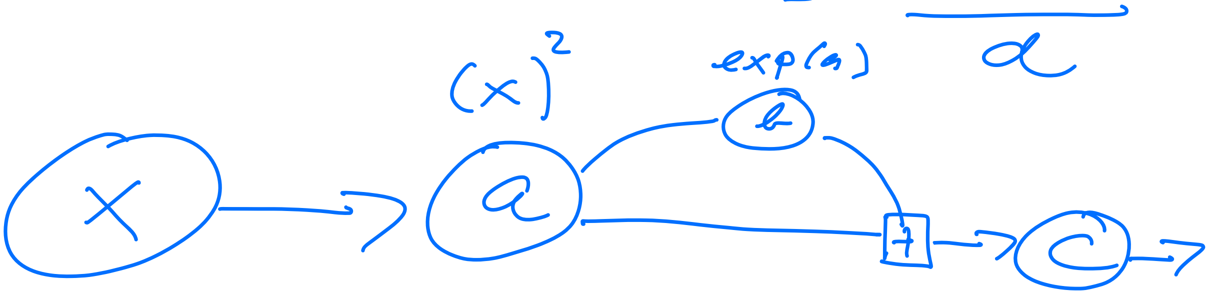
$c = a + b$

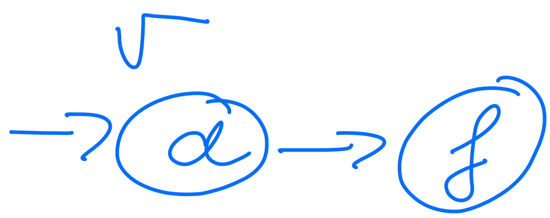
$d = \sqrt{c} = f(x)$

$$\frac{df}{dx} = \frac{x\left(1+\exp(x^2)\right)}{\sqrt{x^2+\exp(x^2)}} \quad \frac{5}{\;}$$

$$= \frac{x(1+b)}{\sqrt{a+b}} = \frac{x(1+b)}{\sqrt{c}}$$

$$= \frac{x(1+b)}{d}$$





$$f(x) = d$$

$$\frac{\partial a}{\partial x} = 2x \qquad \frac{\partial b}{\partial x} = \frac{\partial b}{\partial a}\frac{\partial a}{\partial x}$$

$$\frac{\partial c}{\partial x} = \left[\frac{\partial c}{\partial a}\frac{\partial a}{\partial x} + \underbrace{\frac{\partial c}{\partial b}\frac{\partial b}{\partial x}}\right]$$

$$\frac{\partial c}{\partial b}\frac{\partial b}{\partial a}\frac{\partial a}{\partial x}$$

$$\frac{\partial d}{\partial c} = \frac{1}{2\sqrt{c}}$$

$$\frac{\partial d}{\partial x} = \frac{\partial d}{\partial c}\frac{\partial c}{\partial x} = \frac{\partial f}{\partial x}$$

Computation in reverse mode:

$$\frac{\partial f}{\partial c} = \frac{\partial f}{\partial d}\frac{\partial d}{\partial c} = \frac{1}{2\sqrt{c}}$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} = \frac{1}{2\sqrt{c}}$$

$$\frac{\partial f}{\partial a} = \underbrace{\frac{\partial f}{\partial b}}_{=\frac{1}{2\sqrt{c}}}\underbrace{\frac{\partial b}{\partial a}}_{=\exp(a)} + \underbrace{\frac{\partial f}{\partial c}}_{=\frac{1}{2\sqrt{c}}}\frac{\partial c}{\partial a}$$

$$= \frac{1}{2\sqrt{c}} \left(1 + \exp(a) \right)$$

$$= \frac{1}{2\sqrt{c}} (1 + b)$$

$$= \frac{1}{2 \cdot d} (1 + b)$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial d} \frac{\partial a}{\partial x} = \frac{2x}{2d} (1+b)$$

$$= \frac{x(1+b)}{d}$$

Back propagation eqs,

def $\delta_j^l = \sigma'(z_j^l) \dfrac{\partial C}{\partial a_j^l}$

$$\frac{\partial C}{\partial w_{jk}^l} = \delta_j^l a_k^{l-1}$$

$$\delta_j^L = \frac{\partial C}{\partial b_j^L}$$

$$\delta_j^l = \sum_k \delta_k^{l+1} w_k^{l+1} \sigma'(z_j^l)$$

- Then for each layer we have

$l = L-1, L-2 \ldots 2, 1$

δ_j^l

update

$$w_{jk}^l \leftarrow w_{jk}^l - \eta \, \delta_j^l a_k^{l-1}$$

$$b_j^l \leftarrow b_j^l - \eta \, \delta_j^l$$

1 iteration consists of a feed forward pass and a Back prop pass

continue till cost

function converges.