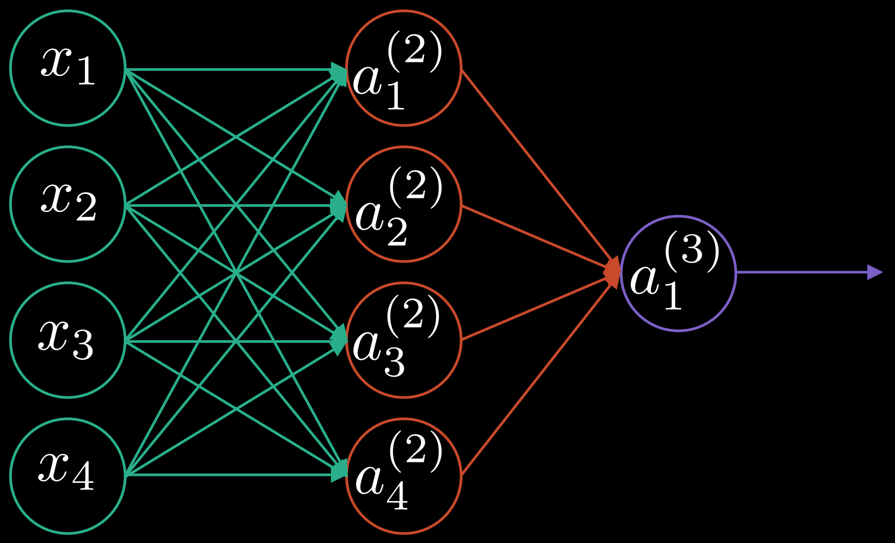
# Practical 2.1

Neural Networks – back propagation (training)

# Overview

- Review on forward pass
- Training data and labels
- Loss function
- Local error
  - Output error
  - Layer $\ell$ error (Jacobian of composite functions)
- Parameters' gradient
- Back propagation
- Stochastic and (mini-) batch gradient descent

# Neural network




$$\boldsymbol{a}^{(1)} = \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{s_1}$$

$$\boldsymbol{a}^{(\ell+1)} = \sigma\left(\underline{z}^{(\ell+1)}\right) = \sigma\left(\Theta^{(\ell)} \hat{\underline{a}}^{(\ell)}\right) \in \mathbb{R}^{s_{\ell+1}}, \quad \ell = 1, 2, \ldots, L-1$$

$$\hat{\boldsymbol{a}} = \begin{bmatrix} +1 \\ \underline{a} \end{bmatrix}$$

$$h_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{a}^{(L)} \in \mathbb{R}^{s_L} = \mathbb{R}^{K}$$

$$\boldsymbol{\Theta}^{(j)} \in \mathbb{R}^{s_{j+1} \times (s_j + 1)} \qquad \boldsymbol{\Theta} = \{\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \ldots, \boldsymbol{\Theta}^{(L-1)}\}$$

# Training data (I)

$$\boldsymbol{X} = \underbrace{\begin{bmatrix} — \underline{x}^{(1)} — \\ — \underline{x}^{(2)} — \\ — \underline{x}^{(3)} — \\ \vdots \\ — \underline{x}^{(m)} — \end{bmatrix}}_{n} \Big\} m \qquad \boldsymbol{Y} = \underbrace{\begin{bmatrix} — \underline{y}^{(1)} — \\ — \underline{y}^{(2)} — \\ — \underline{y}^{(3)} — \\ \vdots \\ — \underline{y}^{(m)} — \end{bmatrix}}_{K} \Big\} m$$

$$h_\theta(\underline{x}) \in \mathbb{R}^K$$

# Training data (II)

$$
\boldsymbol{X} \quad
\begin{array}{cccc}
\underline{x_1} & \underline{x_2} & \cdots & \underline{x_n} \\
\hline
x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\
x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\
\vdots & \vdots & & \vdots \\
x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)}
\end{array}
\qquad\qquad
\boldsymbol{Y} \quad
\begin{array}{cccc}
\underline{y_1} & \underline{y_2} & \cdots & \underline{y_k} \\
\hline
y_1^{(1)} & y_2^{(1)} & \cdots & y_k^{(1)} \\
y_1^{(2)} & y_2^{(2)} & \cdots & y_k^{(2)} \\
\vdots & \vdots & & \vdots \\
y_1^{(m)} & y_2^{(m)} & \cdots & y_k^{(m)}
\end{array}
$$

$$\text{if} \quad K = 1 \Rightarrow [Y] \equiv \underline{y_1}$$

# Loss function

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{m}\sum_{i=1}^{m} E^{(i)}$$

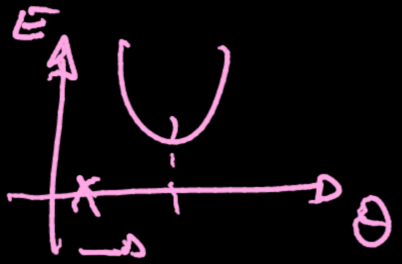
$$E = E(h_\theta(\underline{x}))$$

$\underline{x}$, $\underline{y}$ are fixed
$[\Theta]$ can vary

$$E^{(i)} = E(\underline{h_{\boldsymbol{\Theta}}(\boldsymbol{x})}) = \frac{1}{2}\left\| \underline{y}^{(i)} - \underbrace{h_\theta(\underline{x}^{(i)})}_{\underline{a}^{(L)}} \right\|^2 = \frac{1}{2}\sum_k \left(y_k^{(i)} - a_k^{(L)(i)}\right)^2$$

$$\boxed{\frac{\partial E}{\partial \Theta_{ij}^{(\ell)}}}$$

$$\underline{\boldsymbol{\Theta}} \to \boldsymbol{\Theta} - \eta \underline{\frac{\partial E}{\partial \boldsymbol{\Theta}}}$$



# Local error

$\underline{z_i^{(\ell)}} \to z_i^{(\ell)} + \Delta z_i^{(\ell)}$

$\underline{E} \to E + \frac{\partial E}{\partial z_i^{(\ell)}} \cdot \Delta z_i^{(\ell)}$

$\delta_i^{(\ell)} := \frac{\partial E}{\partial z_i^{(\ell)}}$

error @ i-th neuron
@ layer $\ell$

$\underline{\left| \frac{\partial E}{\partial z_i^{(\ell)}} \right|}$

$\gg 0 \Rightarrow \Delta z_i^{(\ell)}$ can influence $E$

$\simeq 0 \Rightarrow$ near optimal

# Output error

$$E = \frac{1}{2}\sum_k \left(y_k - a_k^{(L)}\right)^2$$

$$\Rightarrow \frac{1}{2}\cdot 2\left(a_i^{(L)} - y_i\right)$$

$$\delta_i^{(L)} = \frac{\partial E}{\partial a_i^{(L)}} \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} = \underbrace{\frac{\partial E}{\partial a_i^{(L)}}}_{a_i^{(L)} - y_i} \underbrace{\sigma'\left(z_i^{(L)}\right)}_{a_i^{(L)}\left(1 - a_i^{(L)}\right)}$$

$$\sigma(z) = \left(1 + \exp(-z)\right)^{-1}$$

$$\sigma'(z) = +\left(1+\exp(-z)\right)^{-2} \cdot \exp(-z)(+1)$$

$$= \left(1+\exp(-z)\right)^{-2} \cdot \left[\left(1+\exp(-z)\right) - 1\right]$$

$$= \left(1+\exp(-z)\right)^{-1} - \left(1+\exp(-z)\right)^{-2} =$$

$$= \sigma(z) - \sigma^2(z) = \sigma(z)\left(1 - \sigma(z)\right)$$

$$= a(1-a)$$

$$\delta^{(L)} = \underbrace{\nabla_a E}_{\underline{a}^{(L)} - \underline{y}} \odot \underbrace{\sigma'\left(\underline{z}^{(L)}\right)}_{\underline{a}^{(L)} \odot \left(1 - \underline{a}^{(L)}\right)}$$

# Layer $\ell$ error (I)

$$\boldsymbol{\delta}^{(\ell)} = \delta^{(\ell)}\left(\delta^{(\ell+1)}\right)$$

# Jacobian of composite function

$g : \mathbb{R} \to \mathbb{R}^n$ diff. $u_0$ , $f : \mathbb{R}^n \to \mathbb{R}$ diff. $\underline{x}^0 = g(u_0)$

$h = f \circ g : \mathbb{R} \to \mathbb{R}$ der. in $u_0$ $\Rightarrow$

$$h'(u_0) = (Jf)(\underline{x}^0)\,(Jg)(u_0) = \left[\frac{\partial f}{\partial x_1}(\underline{x}^0), \ldots, \frac{\partial f}{\partial x_n}(\underline{x}^0)\right] \begin{bmatrix} g_1'(u_0) \\ \vdots \\ g_n'(u_0) \end{bmatrix} = \langle \nabla f(\underline{x}^0), g'(u_0) \rangle$$

# Layer $\ell$ error (II)

$$\underline{\boldsymbol{\delta}^{(\ell)}} = \boldsymbol{\delta}^{(\ell)}(\underline{\boldsymbol{\delta}^{(\ell+1)}})$$

$$\delta_i^{(\ell)} = \frac{\partial E}{\partial z_i^{(\ell)}} \qquad\qquad \delta_j^{(\ell+1)} = \frac{\partial E}{\partial z_j^{(\ell+1)}}$$

$$= \sum_j \overbrace{\frac{\partial E}{\partial z_j^{(\ell+1)}}}^{\delta_j^{(\ell+1)}} \boxed{\frac{\partial z_j^{(\ell+1)}}{\partial z_i^{(\ell)}}}$$

# Layer $\ell$ error (III)

$$\frac{\partial z_j^{(\ell+1)}}{\partial z_i^{(\ell)}} = \Theta_{ji}^{(\ell)} \sigma'(z_i^{(\ell)})$$

$$z_j^{(\ell+1)} = \sum_k \Theta_{jk}^{(\ell)} \hat{a}_k^{(\ell)} = \sum_k \Theta_{jk}^{(\ell)} \sigma(z_k^{(\ell)})$$

$$\delta_i^{(\ell)} = \sum_j \delta_j^{(\ell+1)} \frac{\partial z_j^{(\ell+1)}}{\partial z_i^{(\ell)}} = \sum_j \Theta_{ji}^{(\ell)} \delta_j^{(\ell+1)} \sigma'(z_i^{(\ell)})$$

# Layer $\ell$ error (IV)

$$\delta_i^{(\ell)} = \sum_j \Theta_{ji}^{(\ell)} \delta_j^{(\ell+1)} \sigma'(z_i^{(\ell)})$$

$$\boldsymbol{\delta}^{(\ell)} = \left[ \left( \underline{\Theta}^{(\ell)} \right)^T \underline{\delta}^{(\ell+1)} \right] \odot \sigma' \left( \underline{z}^{(\ell)} \right)$$

# Parameters' gradient

$$\frac{\partial E}{\partial \Theta_{ij}^{(\ell)}} = \underbrace{\frac{\partial E}{\partial z_i^{(\ell+1)}}}_{\delta_i^{(\ell+1)}} \underbrace{\frac{\partial z_i^{(\ell+1)}}{\partial \Theta_{ij}^{(\ell)}}}_{a_j^{(\ell)}}, \quad \forall i, j$$

from back $\delta_i^{(\ell+1)}$ $a_j^{(\ell)}$ forward pass

$$\frac{\partial E}{\partial \boldsymbol{\Theta}^{(\ell)}} = \underline{a}^{(\ell)} \left(\underline{\delta}^{(\ell+1)}\right)^T$$

# Back propagation

$$i) \quad \boldsymbol{a}^{(1)} = \boldsymbol{x}, \hat{\boldsymbol{a}}^{(1)} = \begin{bmatrix} +1 \\ \boldsymbol{a}^{(1)} \end{bmatrix}$$

$$ii) \quad \boldsymbol{z}^{(\ell+1)} = \boldsymbol{\Theta}^{(\ell)} \hat{\boldsymbol{a}}^{(\ell)}, \boldsymbol{a}^{(\ell)} = \sigma(\boldsymbol{z}^{(\ell)}), \hat{\boldsymbol{a}}^{(\ell)} = \begin{bmatrix} +1 \\ \boldsymbol{a}^{(\ell)} \end{bmatrix}$$

$$iii) \quad \boldsymbol{\delta}^{(L)} = \nabla_h E \odot \sigma'(\boldsymbol{z}^{(L)})$$

$$iv) \quad \boldsymbol{\delta}^{(\ell)} = [(\boldsymbol{\Theta}^{(\ell)})^\top \boldsymbol{\delta}^{(\ell+1)}] \odot \sigma'(\boldsymbol{z}^{(\ell)})$$

$$v) \quad \frac{\partial E}{\partial \boldsymbol{\Theta}^{(\ell)}} = \boldsymbol{a}^{(\ell)} (\boldsymbol{\delta}^{(\ell+1)})^\top$$

# Weight update

GRADIENT DESCENT

- Stochastic SGD

$$\boldsymbol{\Theta}^{(\ell)} \to \Theta^{(e)} - \eta\, \underline{a}^{(e)}\, \underline{\delta}^{(e+1)}$$

- Batch (or mini-batch)

$$\boldsymbol{\Theta}^{(\ell)} \to \Theta^{(e)} - \eta \frac{1}{m} \sum_{i=1}^{m} \left( \underline{a}^{(e)}\, \underline{\delta}^{(e+1)} \right)^{(i)}$$