CS 480/680

2/12/26

Lecture 11

. How a nn represents an f

. Analysis & practical issues

Q1 ✓

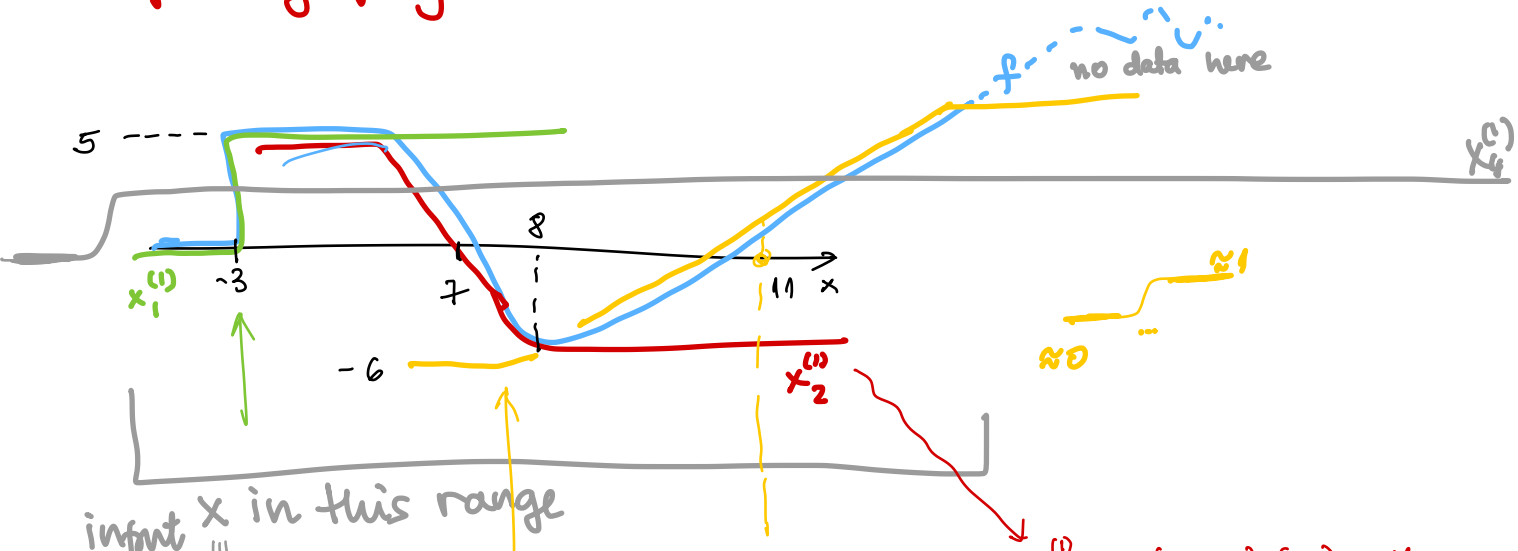
HW5 Wed 2/25

train nn!

Lec notes

TB Posted

Representing f by a nn.



$x^{(0)}$

$(x-3)\,1000 = \underbrace{1000}_{w_{11}^{(1)}}x \underbrace{-3000}_{w_{10}^{(1)}} = z_1^{(i)}$

$(x-11)\,0.5 = 0.5x - 5.5 = z_2^{(i)}$

$x_2^{(i)} = \varphi(z_2^{(i)})$

$z_2^{(i)} = (x-7)(-1) = 7-x$

$x_2^{(i)} = \varphi(7-x)$

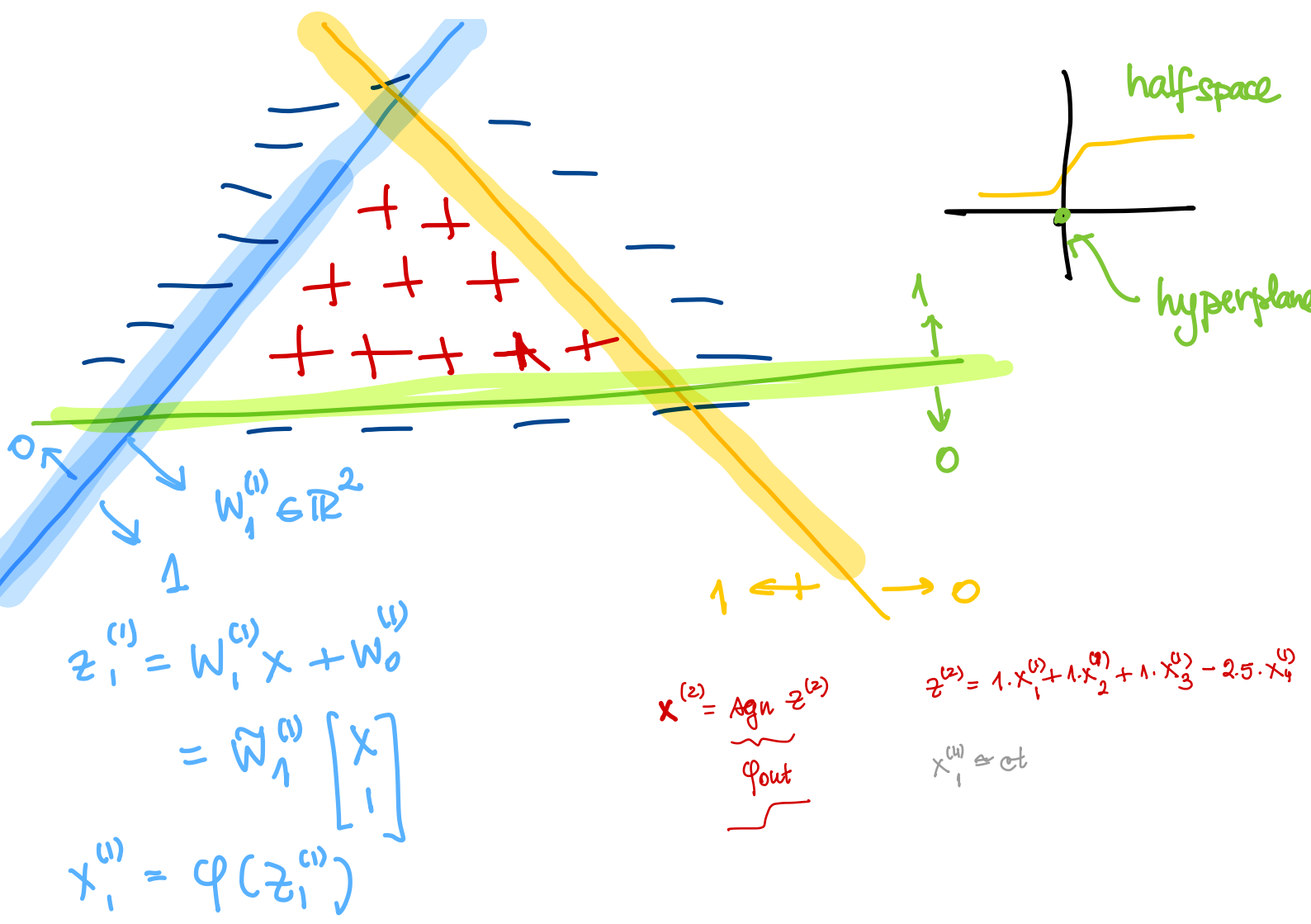
adding a constant

OR $z_4^{(i)} = x + 100000$

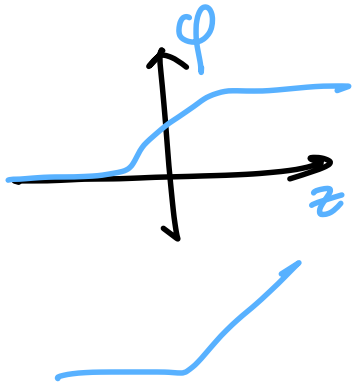
$x_4^{(i)} = \varphi(z^{(4)})$

Layer 2 $x^{(2)} = \varphi_{out}(z^{(2)})$ $\quad w_2^{(2)}$ $\quad w_3^{(2)}$ $\quad \varphi_{out}(z_2) = z$

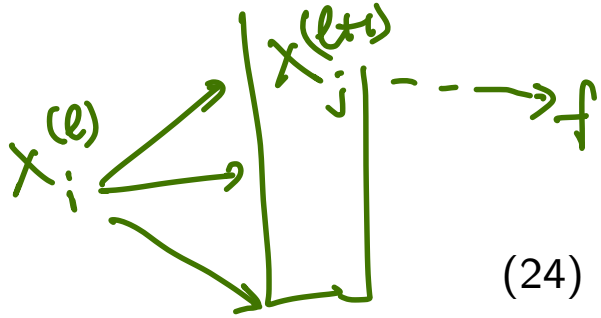
$z^{(2)} = 5 \cdot x_1^{(i)} + 11 \cdot x_2^{(i)} + 12 \cdot x_3^{(i)} - 11 \cdot$ $\quad w_0^{(2)}$


halfspace
hyperplane
1
0
0
1
0
$W_1^{(1)} \in \mathbb{R}^2$
1
$z_1^{(1)} = W_1^{(1)} x + W_0^{(1)}$
$= W_1^{(1)} \begin{bmatrix} x \\ 1 \end{bmatrix}$
$x_1^{(1)} = \varphi(z_1^{(1)})$
$x^{(2)} = \text{sgn}\, z^{(2)}$
φ_{out}
$z^{(2)} = 1 \cdot x_1^{(1)} + 1 \cdot x_2^{(1)} + 1 \cdot x_3^{(1)} - 2.5 \cdot x_4^{(1)}$
$x_1^{(1)} = \text{const}$

From 2 layers to L layers



$|z|$ large $\Rightarrow \varphi' \approx 0$ "saturation"



$$\frac{\partial x_i^{(l)}}{\partial z_i^{(l)}} = \phi'(z_i^{(l)}) \tag{24}$$

$$\frac{\partial x_i^{(l)}}{\partial W_i^{(l)}} = \phi'(z_i^{(l)}) x^{(l-1)} \tag{25}$$

$z_i^{(l)} = W_i^{(l)} x^{(l-1)}$

used for $\frac{\partial \mathcal{L}}{\partial W^{(l-1)}}$

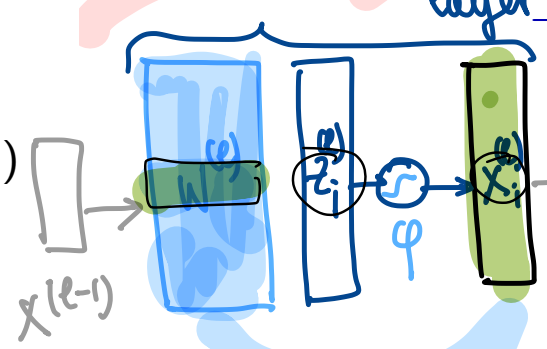
$$\frac{\partial x_i^{(l)}}{\partial x_i^{(l-1)}} = \phi'(z_i^{(l)}) W_i^{(l)} \tag{26}$$

$$\frac{\partial f}{\partial x_i^{(l)}} = \sum_{j=1}^{m_{l+1}} \frac{\partial f}{\partial x_j^{(l+1)}} \frac{\partial x_j^{(l+1)}}{\partial x_i^{(l)}} \tag{27}$$

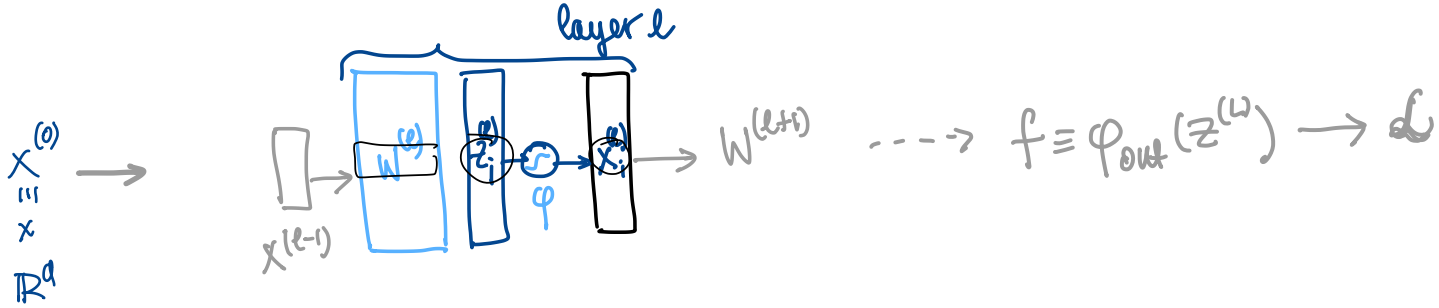
$$\frac{\partial f}{\partial W_i^{(l)}} = \frac{\partial f}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial W_i^{(l)}} \tag{28}$$

$$\frac{\partial \mathcal{L}}{\partial W_i^{(l)}} = (y_* - \phi_{\text{out}}) \frac{\partial f}{\partial W_i^{(l)}} \tag{29}$$

forward



$W^{(l+1)}$ ----> $f \equiv \varphi_{\text{out}}(z^{(L)}) \rightarrow \mathcal{L}$



for $t = 1, 2, \dots$

for $k = 1:n$ (examples)

FORWARD (Prediction)

for $l = 1:L$

compute $x^{(l)}$, $z^{(l)}$ from $x^{(l-1)}$, $W^{(l)}$ ← store!

$f = \varphi_{out}(z^{(L)})$, $\mathcal{L}^{train} \mathrel{+}= \frac{1}{n}\mathcal{L}(y^i, f(x^i))$ } → used for stopping not $\frac{\partial \mathcal{L}}{\partial W}$

BACK PROP.

for $l = L:-1:1$

compute $\frac{\partial \mathcal{L}}{\partial W_i^{(l)}}$ from $x^{(l)}$, $\frac{\partial f}{\partial x^{(l)}}$, $W^{(l)}$, $\frac{\partial x^{(l+1)}}{\partial x^{(l)}}$, $\frac{\partial \mathcal{L}}{\partial f}$

$i = 1:m_l$

$\frac{\partial x^{(l)}}{\partial x^{(l-1)}}$ → for $\frac{\partial \mathcal{L}}{\partial W^{(l-1)}}$

Issues with Back prop

ops/
1 iteration = $O(pn)$

Practical

p = #parameters

I. ops / iteration

T. Convergence $T = ?$ #steps

S. Saturation $\varphi' \approx 0$

L. Local minima

O. Overfitting

$$W = \{ W^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}} \}_{\ell = 1:L}$$

Ex1) $L = 2$

$d = 10^6 = m_0$

$m_1 = 1000$ hidden

$m_2 = 10$ classification

$p = 10^9 + \cdots$

Ex2) $L = 100$

$m_0 = m_1 = \cdots m_{L-1} = 100$

$p \simeq 100^3 \simeq 10^6$

Convergence (T = # iteration)

Theory GD is "order 1"

$err^t = \|w^t - w^0\|$ Wanted $\leq tol$

$\frac{err^{t+1}}{err^t} \leq \beta < 1$

$err^t \sim \beta^t \xRightarrow{Ex} T = \log tol$

Practice

$\mathcal{L}$ not convex $\Rightarrow \frac{\partial \mathcal{L}}{\partial w} \approx 0$ but w not converged !!

$w_{ij}^{(l)}$

$H_{(w)} = \left[\frac{\partial \mathcal{L}}{\partial w_i \partial w_j}\right]_{p \times p}$