Lecture 11

Q1 done ✓

Hw5 out Wed 2/25

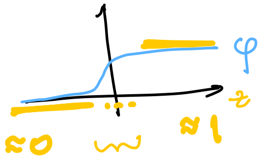
LIV - T.B. extended

1. How does a nn represent a function f ? ✓

2. Backprop. analysis ←

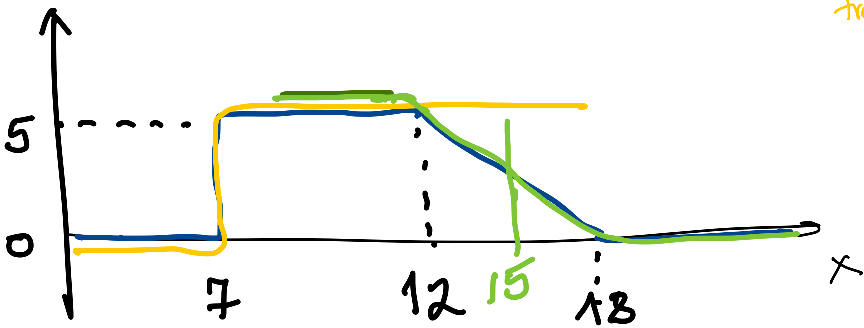
$f = \varphi(w^T x)$

$x^{(\ell)} = \varphi\left(w_i^{(\ell)} x^{(\ell-1)}\right)$



$z = w^T x \Rightarrow$ steeper slope for $\|w\|$ large $\Rightarrow$

"transition" smaller



$w_1^1 = \overset{1000}{\text{large}} > 0$

$w_0^1 = -7 \cdot 1000$

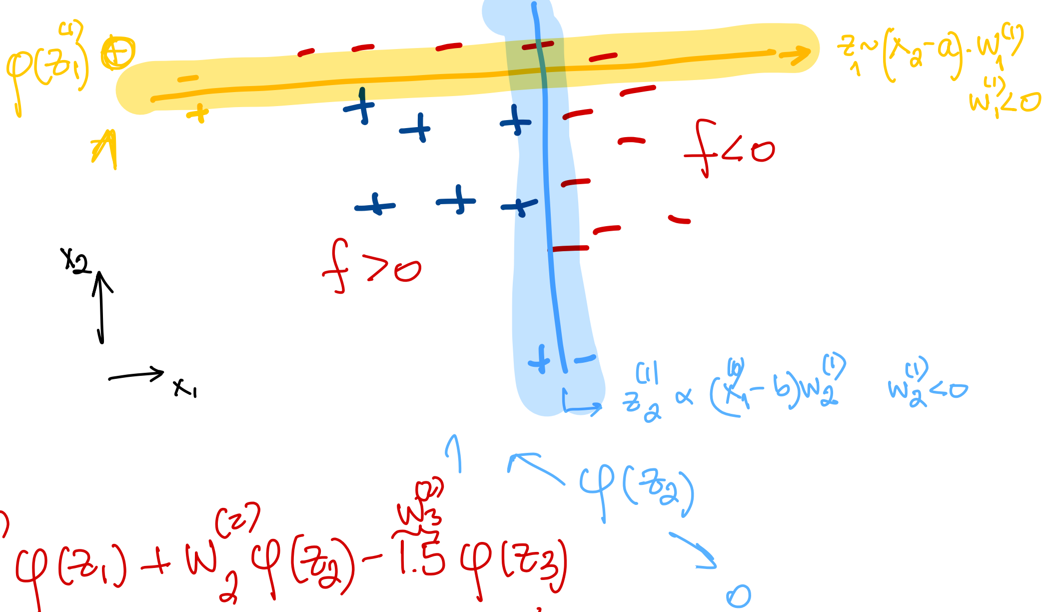
$z_1 = 1000(x-7)$

$w_1^2 = 1$

$w_0^2 = -15$

$z_2' = x - 15$

$f = \overset{w_1^3}{5} \cdot \varphi(z_1) - \underset{w_2^3}{5}\,\varphi(z_2)$



<- - - . .

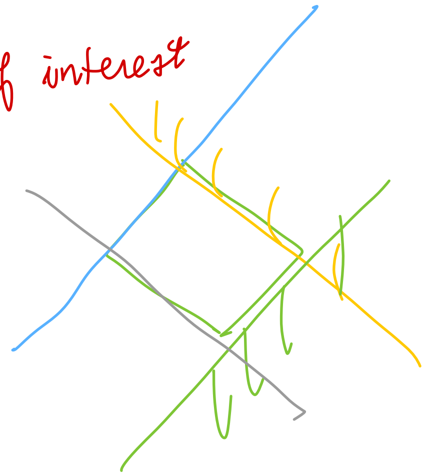
-10000

$$z^{(2)}_1 \leftarrow f = w^{(2)}_1 \varphi(z_1) + w^{(2)}_2 \varphi(z_2) - \overbrace{1.5}^{w^{(2)}_3} \underbrace{\varphi(z_3)}_{\approx 1}$$

$z^{(2)}_{1:m_2}$

$z_3 = x_1 + 10000 \Rightarrow \varphi(z_3) \approx 1$ on the region of interest

$$w^{(1)}_1 = w^{(1)}_2 = 1$$



Analysis of Backprop

$\mathcal{W} = \{ W^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}}, \ell = 1:L \}$ $O(p \cdot n)$ op/iter

I Computation / iteration ?

T How many ~~steps~~ to convergence $T \sim ?$

$\left[\sum_{\ell=1}^{L} m_\ell m_{\ell-1} \text{ parameters} \right] = p$

T.p.n for training

S Saturation $(\varphi'(z^{(\ell)}) \approx 0)$

1) $L = 2$ $m_1 = 1000$, $m_2 = 1$

$m_0 = d = 1M$

#params $= 10^9 + \cdots$

L Local optima → (minima)

O Overfitting

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

$L = 100$

#params $\approx 100^2 \cdot 100$ ~~10^6~~

"T" ← GD "slow" compared to 2nd order

saturation

Optimization "hierarchy"

Newton 2nd order

Grad 1st order $\|w^t - w^\infty\| \propto \beta^t$

$\beta < 1$

L layer network

$x^{(\ell-1)}$, $W^{(\ell)}$, $z^{(\ell)}$, φ (element wise), $x^{(\ell)}$

$$x \to \left[W^{(1)} \to z^{(1)} \to \varphi \to x^{(1)} \right] \to \left[W^{(2)} \to z^{(2)} \to \varphi^{out}(z^{(2)}) = f \right]$$

$d = m_0$

m_ℓ

layer ℓ layer $\ell+1$

Parameters

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

NEW

$$\frac{\partial f}{\partial x^{(\ell+1)}} =$$

$\} \times n$ data points (x^i, y^i)

$$\boxed{\frac{\partial f}{\partial x^{(\ell)}} = \frac{\partial f}{\partial x^{(\ell+1)}} \cdot \frac{\partial x^{(\ell+1)}}{\partial x^{(\ell)}}}$$

$$\frac{\partial x_j^{(\ell+1)}}{\partial x^{(\ell)}} = \varphi'(z_j^{(\ell+1)}) \cdot \underbrace{\frac{\partial z_j^{(\ell+1)}}{\partial x^{(\ell)}}}_{W_j^{(\ell)}}$$

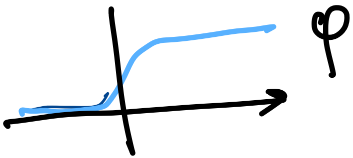
Chain Rule:

$$\frac{\partial f}{\partial z^{(2)}} = \varphi'_{out}(z^{(2)})$$

$$\frac{\partial z^{(2)}}{\partial W^{(2)}} = x^{(1)}$$

$z = W^T x$

$$\frac{\partial z^{(2)}}{\partial x^{(1)}} = W^{(2)}$$



$$z_j^{(\ell+1)} = W_j^{(\ell+1)T} \cdot x^{(\ell)}$$

$$\frac{\partial \mathcal{L}}{\partial W^{(2)}} = \frac{\partial \mathcal{L}}{\partial f} \cdot \frac{\partial f}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial W^{(2)}} = -(y_* - \varphi_{out}) \cdot \varphi'_{out}(z^{(2)}) \cdot x^{(1)}$$

$$\boxed{\frac{\partial f}{\partial x^{(\ell)}}} = \frac{\partial f}{\partial z^{(\ell+1)}} \cdot \underbrace{\frac{\partial z_j^{(\ell+1)}}{\partial x^{(\ell)}}}_{W_j^{(\ell)}} =$$

$$\frac{\partial x_k^{(1)}}{\partial z_k^{(1)}} = \varphi'(z_k^{(1)})$$

$$\frac{\partial x^{(1)}}{\partial z^{(1)}} = \varphi'(z^{(1)}) \in \mathbb{R}^m$$

$z_k^{(1)} = W_k^{(1)T} x^{(0)}$ row k of $W^{(1)}$

$$\frac{\partial z_k^{(1)}}{\partial x^{(0)}} = W_k^{(1)}$$

$$\frac{\partial z_k^{(1)}}{\partial W_k^{(1)}} = x^{(0)}$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial W_k^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f} \cdot \boxed{\frac{\partial f}{\partial x_k^{(\ell)}}} \cdot \underbrace{\frac{\partial x_k^{(\ell)}}{\partial z_k^{(\ell)}}}_{\varphi'} \cdot \underbrace{\frac{\partial z_k^{(\ell)}}{\partial W_k^{(\ell)}}}_{x^{(\ell)}} = -(y_* - \varphi_{out})\, \varphi'_{out}(z^{(2)}) \cdot W_k^{(2)} \varphi'(z_k^{(1)})\, x^{(\ell)}$$