Cheatsheet for Fisher-Rao Metric, Geometry, and Complexity of Neural Networks

Sandro Braun, Leander Kurscheidt

July 19, 2018

Abstract

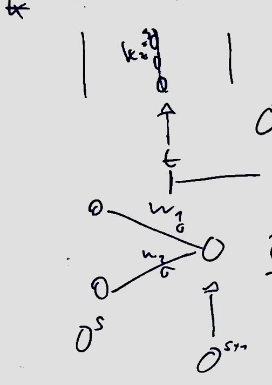
1 Geometry of Deep Rectified Networks

Lemma 2.1 (Structure in Gradient)

$$\sum_{t=0}^{L} \sum_{i\in[k_t],j\in[k_{t+1}]} \frac{\partial O^{L+1}}{\partial W^{t_{ij}}} W_{ij}^{t} = (L+1)O^{L+1}(x) = \langle \nabla_\theta f_\theta(x), \theta \rangle \tag{1}$$

References

$t = 1, \ s = 1 \qquad 0 \le t \le s \le L$



$$O^{s+2}(x) = \sigma\big(\overbrace{O_1^1 w_1 + O_2^1 w_2}^{z}\big)$$

$$\frac{\partial O^2}{\partial w_1} = \sigma'(z) \cdot O_1^1 \cdot w_1$$

$$\frac{\partial O^2}{\partial w_2} = \sigma'(z) \cdot O_2^1 \, w_2$$

$$= \sigma'(z)\big(\underbrace{O_1^1 \cdot w_1 + O_2^1 w_2}_{z}\big)$$

$$= \sigma(z)$$

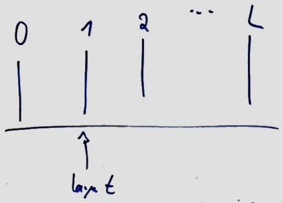
$$f_\theta(x) = \sigma(x^T \cdot \theta)$$

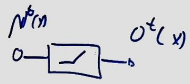
$$\frac{\partial f}{\partial \theta} = \sigma'(x \cdot \theta) \cdot x \qquad \sigma(x \cdot \theta)$$

$$\left\langle \frac{\partial f}{\partial \theta}, \theta \right\rangle = \cancel{\sigma'(x \cdot \theta)\, x\theta}$$

$$\sigma(x \cdot \theta)$$

$$= f_\theta(x)$$





non-lin

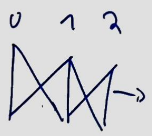
$$O^t(x) = \sigma_t(N^t(x))$$

$$O^0(x) = X = O^0 \in \mathbb{R}^p \qquad O^{L+1}(x) = f_\theta(x) \in \mathbb{R}^k$$

$$\overbrace{\sigma\Big(\overbrace{\underbrace{\sigma\big(\underbrace{X^T W^0}_{N^1}\big)}_{O^1 \text{ via } O^0} W^1}^{}\Big)}^{O^?}$$

$$\underbrace{\sigma(X^T W^0)\, W^1}_{N^2}$$

$X \in \mathbb{R}^{28^2}$



$Y \in \mathbb{R}^{10}$

$w^0\ 628^2,\ w^1\ 1024,\ w_2 = 10$

2.2: $\sigma(x) = \sigma'(x)\, x$

$$\ell = \max\{0, \underbrace{1 - y_i f_i}_{>0}\}$$

1) $\frac{\partial \ell(f, y)}{\partial f} = -y \iff yf < 1$

$\rightarrow 1 - y_i f_i > 0 \Rightarrow \ell = 1 - y_i f_i$

$\rightarrow \frac{\partial \ell}{\partial f} = -y_i$ q.e.d.

2) $yf \geq 1 \Rightarrow 1 - y_i f_i \leq 0 \Rightarrow \ell = 0$

$\frac{\partial \ell}{\partial f} = 0$ q.e.d.

1. $\nabla_\theta \hat{\mathcal{L}}(\theta) = 0$
2. $y_i f_\theta(x_i) \geq 0 \ \forall i$

$\Updownarrow$

$y_i f_\theta(x_i) \geq 1 \ \forall i$

P.5 | Proof of Corollar 2.1

$$\hat{\mathcal{L}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)$$

$$\langle \nabla_\theta \hat{\mathcal{L}}(\theta), \theta \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle \nabla_\theta \ell(\underbrace{f_\theta(x_i)}_{\theta^{L+1}(x_i)}, y_i), \theta \rangle$$

$$= \frac{1}{n} \sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \underbrace{\langle \nabla_\theta f_\theta(x_i), \theta \rangle}_{(L+1) \cdot f_\theta(x_i)} \quad (2.6) + \nabla_\theta \text{ (all } \theta \text{?)}$$

$$= \frac{1}{n} \sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \cdot (L+1) f_\theta(x_i)$$

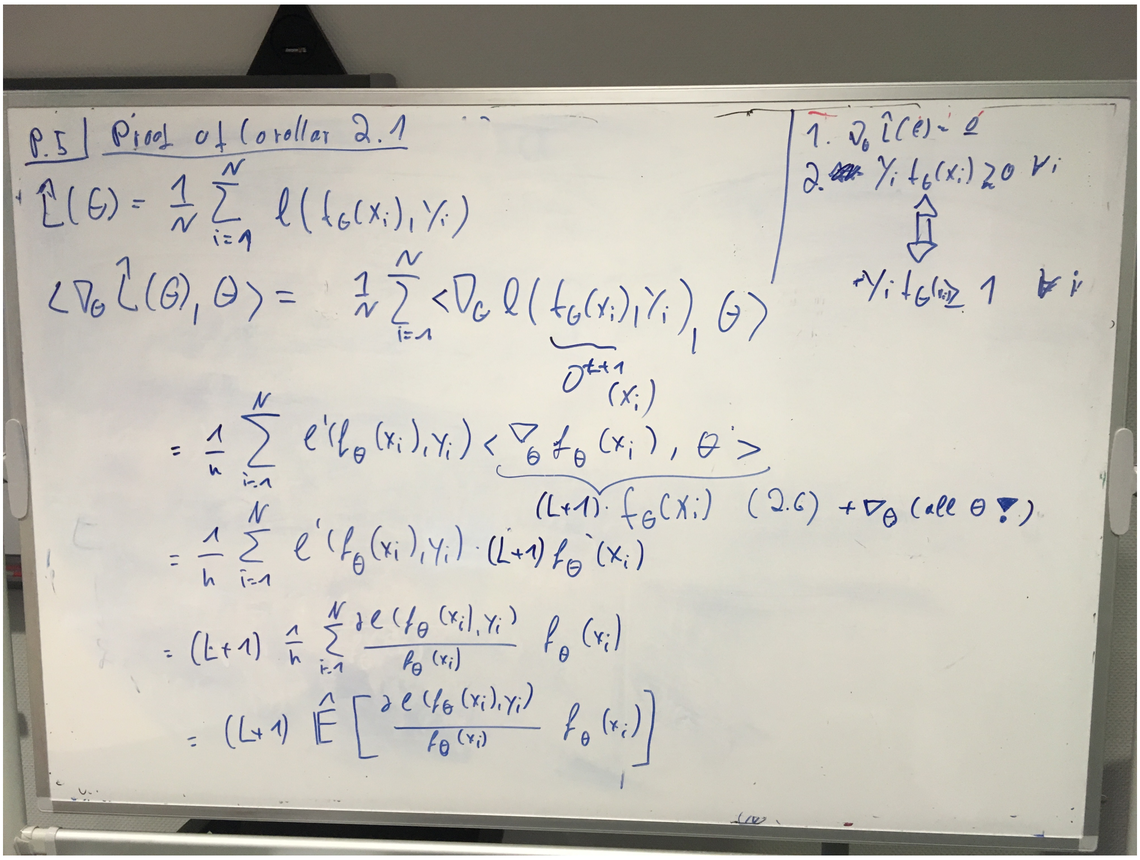
$$= (L+1) \frac{1}{n} \sum_{i=1}^{N} \frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)$$

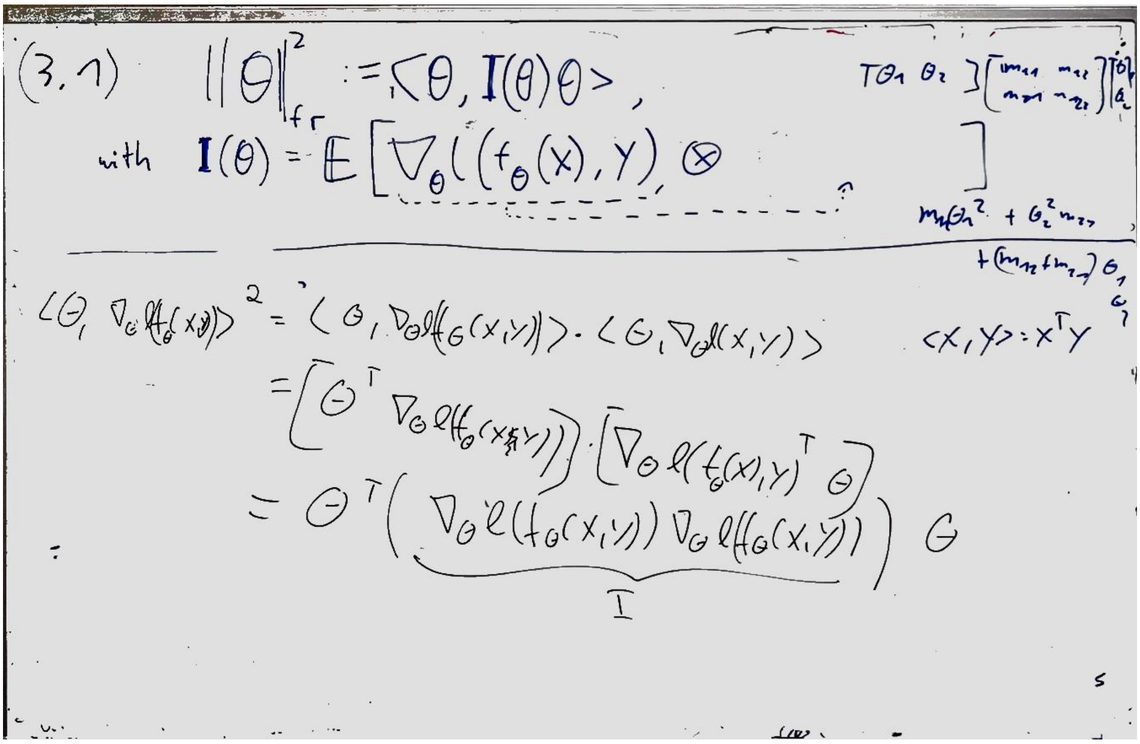
$$= (L+1) \hat{\mathbb{E}}\left[\frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)\right]$$

1. $\nabla_\theta \hat{\mathcal{L}}(\theta) = 0$
2. $y_i f_\theta(x_i) \geq 0 \ \forall i$

$\Updownarrow$

$y_i f_\theta(x_i) \geq 1 \ \forall i$


P.5 Proof of Corollar 2.1
$\hat{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)$
$\langle \nabla_\theta \hat{L}(\theta), \theta \rangle = \frac{1}{N}\sum_{i=1}^{N} \langle \nabla_\theta \ell(\underbrace{f_\theta(x_i)}_{\sigma^{L+1}(x_i)}, y_i), \theta \rangle$
$= \frac{1}{n}\sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \underbrace{\langle \nabla_\theta f_\theta(x_i), \theta \rangle}_{(L+1)\cdot f_\theta(x_i)}$ (2.6) + ∇_θ (all θ ?)
$= \frac{1}{n}\sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \cdot (L+1) f_\theta(x_i)$
$= (L+1) \frac{1}{n}\sum_{i=1}^{N} \frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)$
$= (L+1) \hat{E}\left[\frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)\right]$
1. $\partial_\theta \hat{L}(\theta) = 0$
2. $y_i f_\theta(x_i) \geq 0 \ \forall i$
$\Updownarrow$
$y_i f_\theta(x_i) \geq 1 \ \forall i$


(3.1) $\|\theta\|_{fr}^2 := \langle \theta, I(\theta)\theta \rangle$,
with $I(\theta) = E[\nabla_\theta \ell(f_\theta(x), y) \otimes \quad]$
$[\theta_1 \ \theta_2] \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$
$m_{11}\theta_1^2 + \theta_2^2 m_{22} + (m_{12}+m_{21})\theta_1\theta_2$
$\langle \theta, \nabla_\theta \ell(f_\theta(x,y)) \rangle^2 = \langle \theta, \nabla_\theta \ell(f_\theta(x,y)) \rangle \cdot \langle \theta, \nabla_\theta \ell(x,y) \rangle$
$\langle x, y \rangle = x^T y$
$= [\theta^T \nabla_\theta \ell(f_\theta(x,y))] \cdot [\nabla_\theta \ell(f_\theta(x), y)^T \theta]$
$= \theta^T \underbrace{(\nabla_\theta \ell(f_\theta(x,y)) \nabla_\theta \ell(f_\theta(x,y)))}_{I} \theta$

(3.1) $\|\theta\|^2_{fr} := \langle \theta, \mathbb{I}(\theta)\theta \rangle$,

with $\mathbb{I}(\theta) = \mathbb{E}\left[\nabla_\theta l(f_\theta(X),Y) \otimes \qquad \right]$

$T\theta_1\ \theta_2] \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$

$m_{11}\theta_1^2 + \theta_2^2 m_{22}$

$+ (m_{12} + m_{21})\theta_1 \theta_2$

$\langle \theta, \nabla_\theta l(f_\theta(X),Y) \rangle$

$= \langle \left[\nabla_\theta l(f_\theta(X),Y)\right], \theta \rangle$

Chain

$= \langle \frac{\partial l(f_\theta(X),Y)}{\partial f_\theta(X)} \nabla_\theta f_\theta(X), \theta \rangle$

$= \langle \nabla_\theta f_\theta(X) \partial \dots, \theta \rangle$

$= \langle \frac{\partial l(f_\theta(X),Y)}{\partial f_\theta(X)}, \nabla_\theta f_\theta(X)^T \theta \rangle$

$\langle X, y \rangle = x^T y$

$\langle \alpha x, y \rangle = \alpha(x^T y)$

$= x^T(\alpha y)$

$\langle \alpha, x^T y \rangle =$

Es war: $\langle \partial f_\theta(x), \theta \rangle = (L+1)\theta^{L+1}(x) = (L+1) f_\theta(x)$

$E\left[\langle \frac{\partial l(f_\theta(X,Y))}{\partial f_\theta(Y)}, (L+1) f_\theta(x) \rangle^2\right]$

$= (L+1)^2 E\left[\langle \frac{\partial l(f_\theta(x,y))}{\partial f_\theta(x)}, f_\theta(x) \rangle^2\right] = \|\theta\|^2_{fr}$

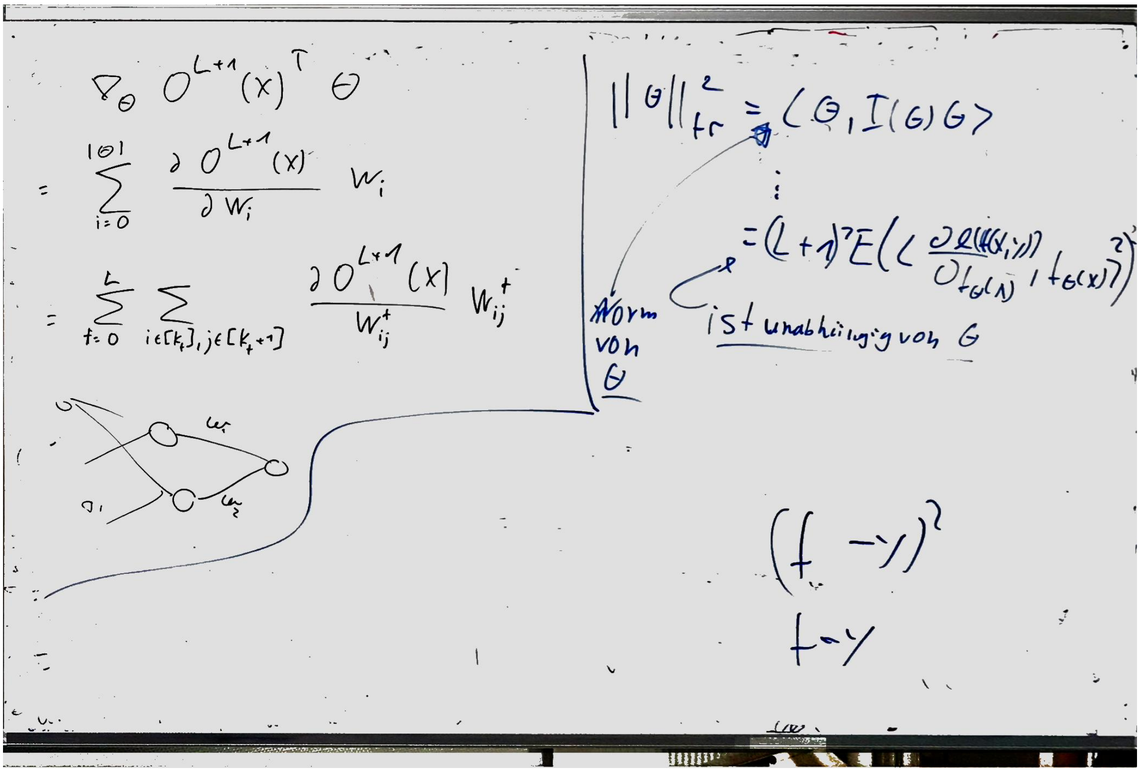
$l = \frac{(\ \ )^2}{2} \to \frac{\partial l}{\partial f} = |\ \ | \Rightarrow E\left[\langle f_\theta(x) - y, f_\theta(x) \rangle^2\right]$

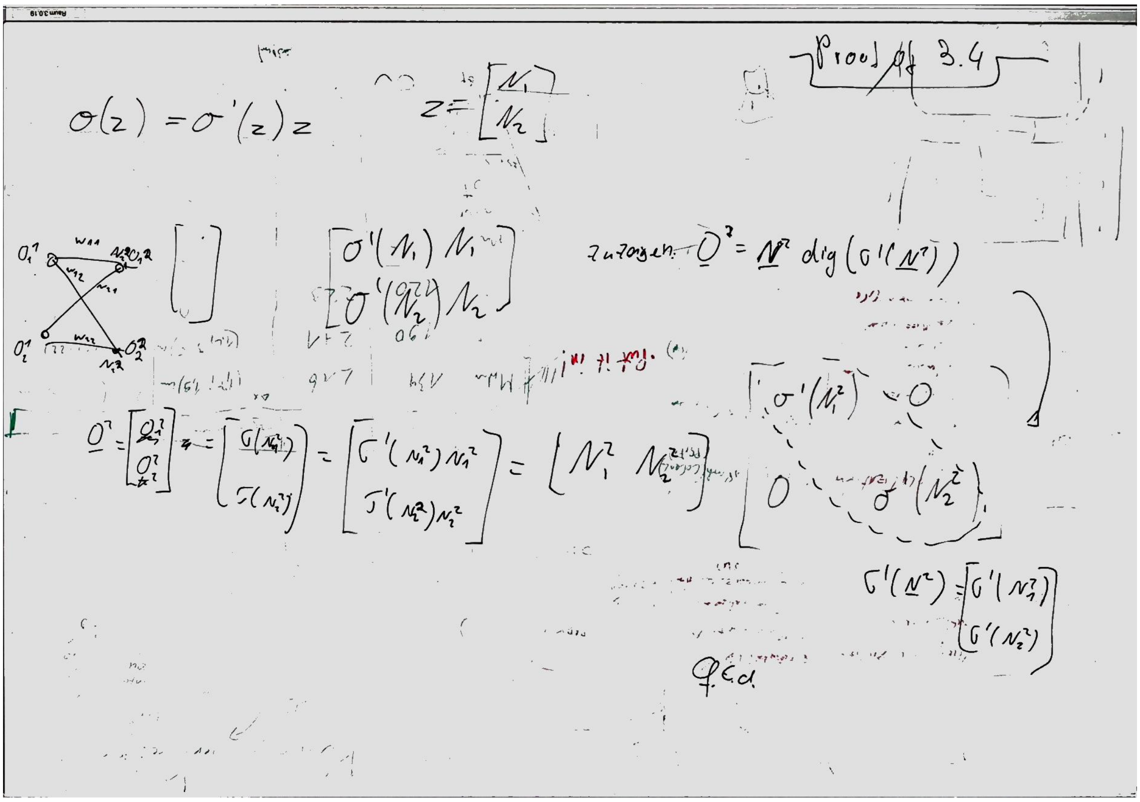
$\frac{(f_\theta(x) - y)^2}{2}$ $\quad G = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix}$ $\quad f_\theta^2(x) - y f_\theta(x)$

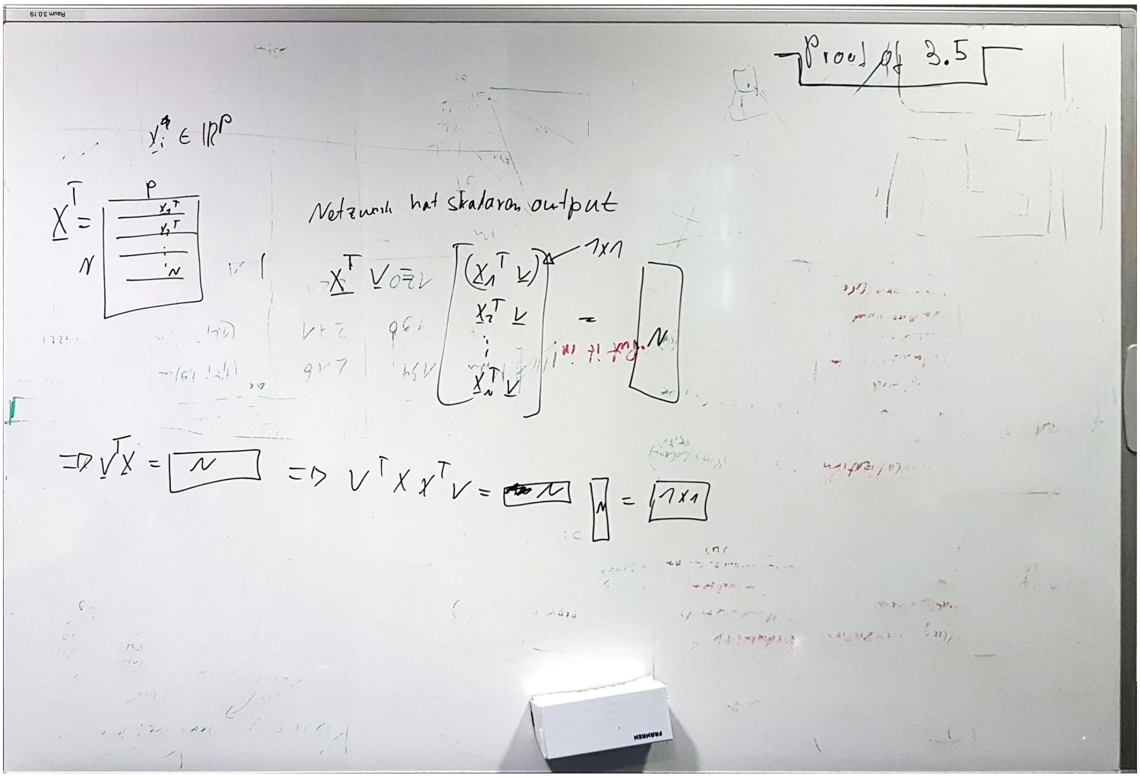
$\nabla_\theta \theta^{L+1} = \begin{bmatrix} \frac{\partial \theta^{L+1}}{\partial w_1} \\ \vdots \\ \partial w_N \end{bmatrix}$

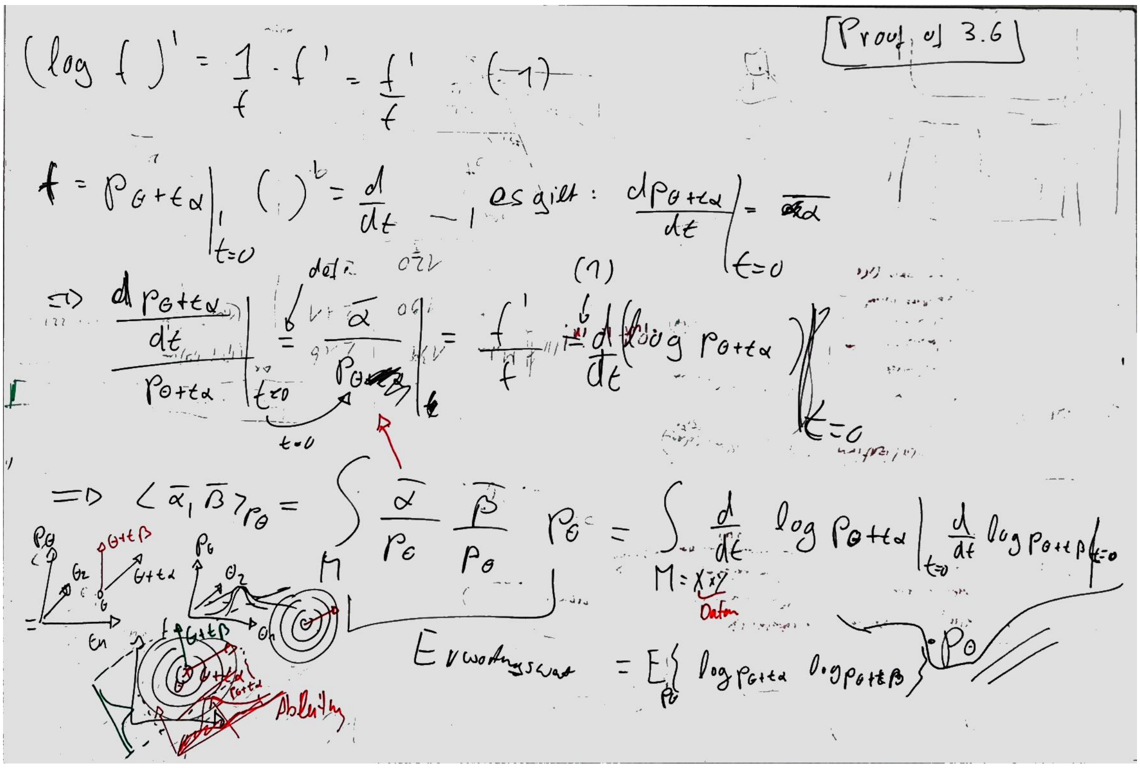
$\cancel{\nabla\theta^T\theta =}$

$(\nabla\theta)^T G$


$\nabla_\Theta O^{L+1}(x)^T \Theta$
$= \sum_{i=0}^{|\Theta|} \frac{\partial O^{L+1}(x)}{\partial W_i} W_i$
$= \sum_{t=0}^{L} \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O^{L+1}(x)}{W_{ij}^t} W_{ij}^t$
$\|\Theta\|_{tr}^2 = \langle \Theta, I(\Theta)\Theta \rangle$
Norm von Θ
ist unabhängig von Θ
$(f - y)^2$
$f = y$


Proof of 3.4
$\sigma(z) = \sigma'(z)z$
$z = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}$
$\begin{bmatrix} \sigma'(N_1) N_1 \\ \sigma'(N_2) N_2 \end{bmatrix}$
zuzeigen: $\underline{O}^2 = \underline{N}^2 \, dig(\sigma'(\underline{N}^2))$
$\underline{O}^2 = \begin{bmatrix} O_1^2 \\ O_2^2 \end{bmatrix} = \begin{bmatrix} \sigma(N_1^2) \\ \sigma(N_2^2) \end{bmatrix} = \begin{bmatrix} \sigma'(N_1^2) N_1^2 \\ \sigma'(N_2^2) N_2^2 \end{bmatrix} = [N_1^2 \; N_2^2]$
$\begin{bmatrix} \sigma'(N_1^2) & 0 \\ 0 & \sigma'(N_2^2) \end{bmatrix}$
$\sigma'(\underline{N}^2) = \begin{bmatrix} \sigma'(N_1^2) \\ \sigma'(N_2^2) \end{bmatrix}$
q.e.d.


Proof of 3.5
Netzwerk hat skalaren output
1x1
1x1


Proof of 3.6
es gilt :
t=0
Erwartungswert
Daten
Ableitung