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

Sandro Braun, Leander Kurscheidt

August 12, 2018

**Abstract**

## 1 Geometry of Deep Rectified Networks

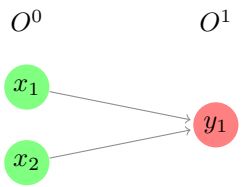
### 1.1 Lemma 2.1

**Lemma 1.1** (Structure in Gradients)**.**

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

#### 1.1.1 Example for Lemma 2.1 with single output

Take a simple feed forward neural network with only two inputs and one output. Remember that by definition $O^0 = x$. Illustrated:




The simple example can then be written with:

$$O^0 = \begin{bmatrix} O_1^0 \\ O_2^0 \end{bmatrix}, W^0 = \begin{bmatrix} W_1^0 \\ W_2^0 \end{bmatrix} \tag{2}$$

$$O^1 = \sigma({O^0}^T W^0) = \sigma(\underbrace{O_1^0 W_1^0 + O_2^0 W_2^0}_{z}). \tag{3}$$

Substituting the term in brackets will simplify notation.

Now calculate the partial derivatives of (2) with respect to $W_i^0$.

$$\frac{\partial O^1}{\partial W_1^0} = \sigma'(z) O_1^0$$
$$\frac{\partial O^1}{\partial W_2^0} = \sigma'(z) O_2^0$$

Summing up the $\frac{\partial O^1}{\partial W_j} W_j$ reveals,

$$\begin{aligned} \sum_{j=1}^{2} \frac{\partial O^1}{\partial W_j^0} W_j^0 &= \sigma'(z) O_1^0 W_1^0 + \sigma'(z) O_2^0 W_2^0 = \sigma'(z)(\underbrace{O_1^0 W_1^0 + O_2^0 W_2^0}_{\text{z}}) \\ &= \sigma'(z) z \end{aligned}$$

Note that this is equivalent to calculating $\langle \nabla_{W^0} O^1, W^0 \rangle$:

$$\nabla_{W^0} O^1 = \begin{bmatrix} \frac{\partial O^1}{\partial W_1^0} \\ \frac{\partial O^1}{\partial W_2^0} \end{bmatrix}$$

$$\langle \nabla_{W^0} O^1, W^0 \rangle = \sum_{j=1}^{2} \frac{\partial O^1}{\partial W_j^0} W_j^0 = \cdots = \sigma'(z)z$$

Using the relation $\sigma(z) = \sigma'(z)(z)$ reveals (2). Then reading left to right reveals (1) for $L = 0$, $\theta = W^0$ and $f_\theta(x) = O^1(x)$ which completes the example.

$$\sum_{t=0}^{L=0} \sum_{j=1}^{2} \frac{\partial O^1}{\partial W_j^t} W_j^t = \langle \nabla_{W^0} O^1, W^0 \rangle = \sigma'(z)z = \sigma(z) = O^1$$

### 1.1.2 Applying Lemma 2.1 to the network structure

Lemma 2.1 applied to the network notation leads to the following notation:

$$\text{Lemma 2.1} : \sigma(z) = \sigma'(z)z \tag{4}$$

$$\text{Network notation} : f_\theta(x) = \sigma_{L+1}(\sigma_L(\cdots \sigma_2(\sigma_1(x^T W^0)W^1)W^2)\cdots)W^L) \tag{5}$$

$$\text{combined} : f_\theta(x) = x^T W^0 D^1(x) W^1 D^2(x) \dots D^L W^L D^{L+1}(x) \tag{6}$$

### 1.1.3 Example for Lemma 2.1 with multiple outputs

Take again a simple neural network consisting of a single layer, hence $L = 0$. We will now start from (5) for $L = 0$, then use (4) to get (6). Now assume, w.l.o.g. the network has two inputs and two outputs. We then have

$$O^0 = \begin{bmatrix} O_1^0 \\ O_2^0 \end{bmatrix}, \; O^1 = \begin{bmatrix} O_1^1 \\ O_2^1 \end{bmatrix}, \; W^0 = \begin{bmatrix} W_{1,1}^0 & W_{1,2}^0 \\ W_{2,1}^0 & W_{2,2}^0 \end{bmatrix}, \tag{7}$$

$$O^1 = \sigma(O^{0^T} W^0) \tag{8}$$

$$= \sigma(\begin{bmatrix} O_1^0 W_{1,1}^0 + O_2^0 W_{1,2}^0 \\ O_1^0 W_{2,1}^0 + O_2^0 W_{2,2}^0 \end{bmatrix}) \tag{9}$$

$$= \begin{bmatrix} \sigma(O_1^0 W_{1,1}^0 + O_2^0 W_{1,2}^0) \\ \sigma(O_1^0 W_{2,1}^0 + O_2^0 W_{2,2}^0) \end{bmatrix} \tag{10}$$

$$= \begin{bmatrix} \sigma(N_1^1) \\ \sigma(N_2^1) \end{bmatrix} \tag{11}$$

$$\tag{12}$$

For simplicity, we can replace the sum within each vector element with $N_i^1$. Now remember (4) and apply it to (8).

$$O^1 = \begin{bmatrix} \sigma(N_1^1) \\ \sigma(N_2^1) \end{bmatrix} \tag{13}$$

$$= \begin{bmatrix} \sigma'(N_1^1)N_1^1 \\ \sigma'(N_2^1)N_2^1 \end{bmatrix} \tag{14}$$

$$= \begin{bmatrix} N_1^1 & N_2^1 \end{bmatrix} \underbrace{\begin{bmatrix} \sigma'(N_1^1) & 0 \\ 0 & \sigma'(N_2^1) \end{bmatrix}}_{D^1(x)} \tag{15}$$

$$= N^1 D^1(x) \tag{16}$$

The last step is to remember $N^{i+1} = O^{i^{\mathrm{T}}} W^1$, then we have

$$O^1 = O^{0^{\mathrm{T}}} W^0 D^1(x) \tag{17}$$

$$= x^{\mathrm{T}} W^0 D^1(x) \tag{18}$$

which is equivalent to inserting $L = 0$ to (6). This completes the example.

## 1.2 Corollary 2.1

### 1.2.1 Notes

1. *Proof.* We want to show $\frac{\partial l(f,Y)}{\partial f} = -y \Leftrightarrow yf < 1$. So,

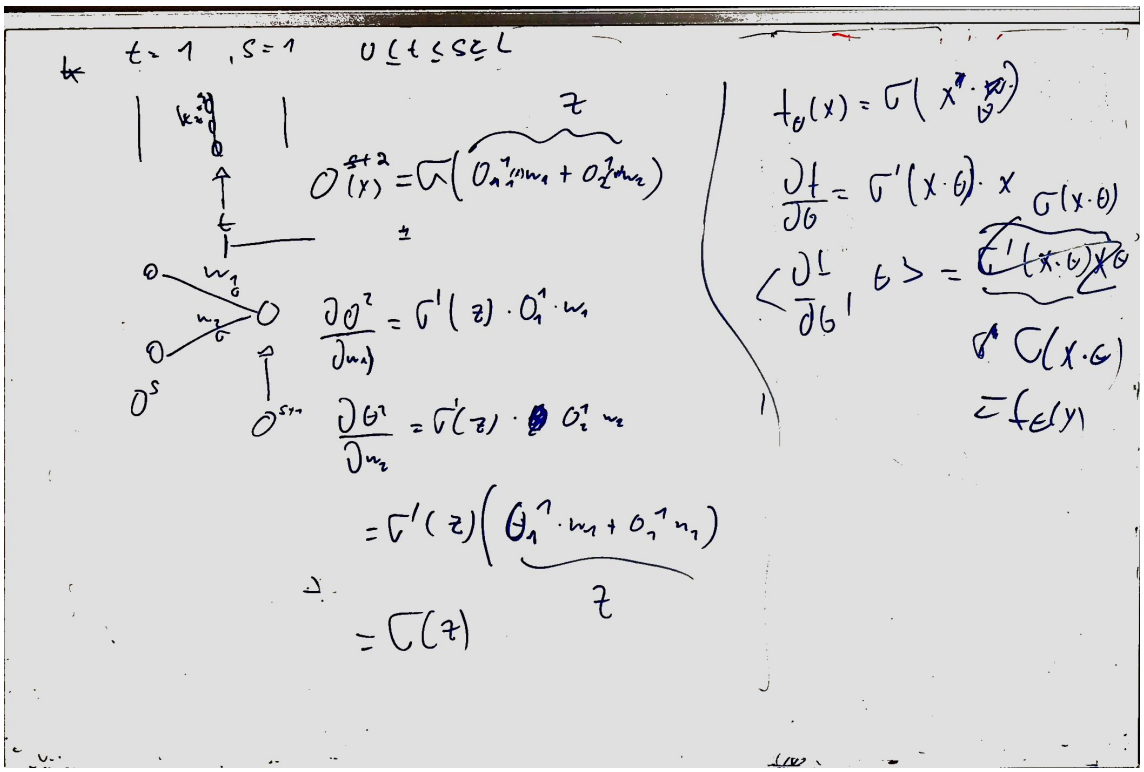
$$\begin{aligned} & 1 - y_i f_i > 0 \\ \Leftrightarrow \quad & l = 1 - y_i f_i \\ \Leftrightarrow \quad & \frac{\partial l}{\partial f} = -y_i \end{aligned}$$

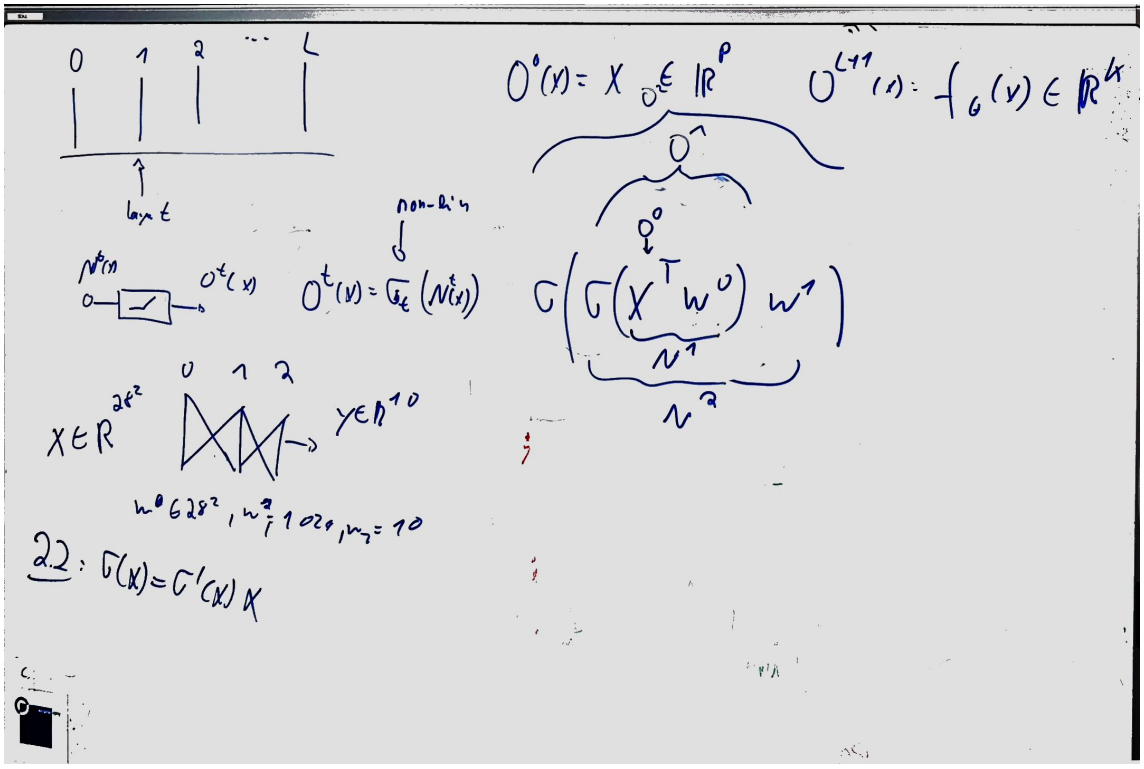
□

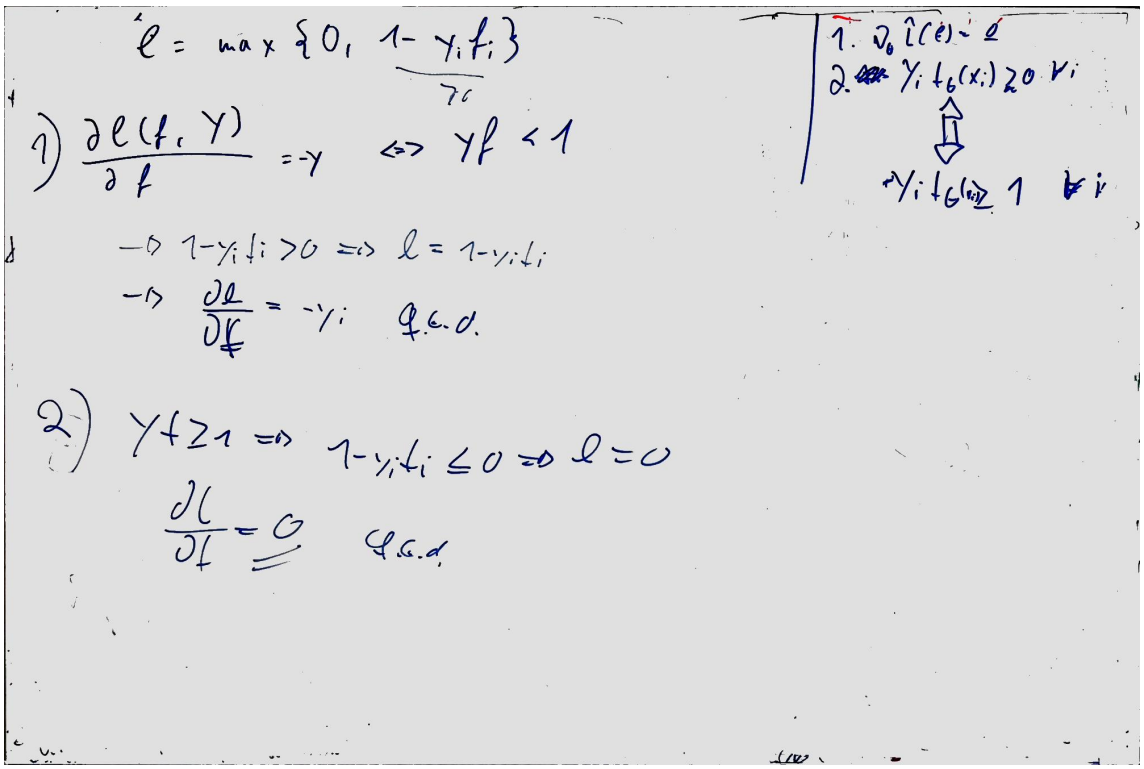
2. *Proof.* We want to show $\frac{\partial l(f,Y)}{\partial f} = 0 \Leftrightarrow yf > 1$. So,

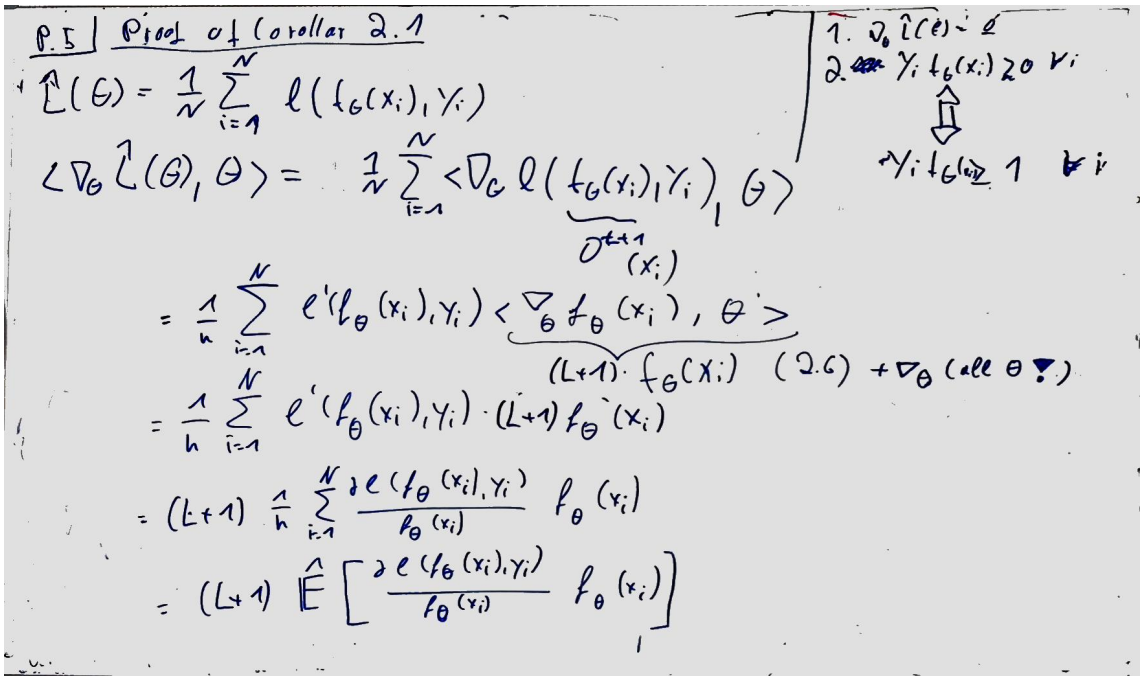
$$\begin{aligned} & 1 - y_i f_i < 0 \\ \Leftrightarrow \quad & l = 0 \\ \Leftrightarrow \quad & \frac{\partial l}{\partial f} = 0 \end{aligned}$$

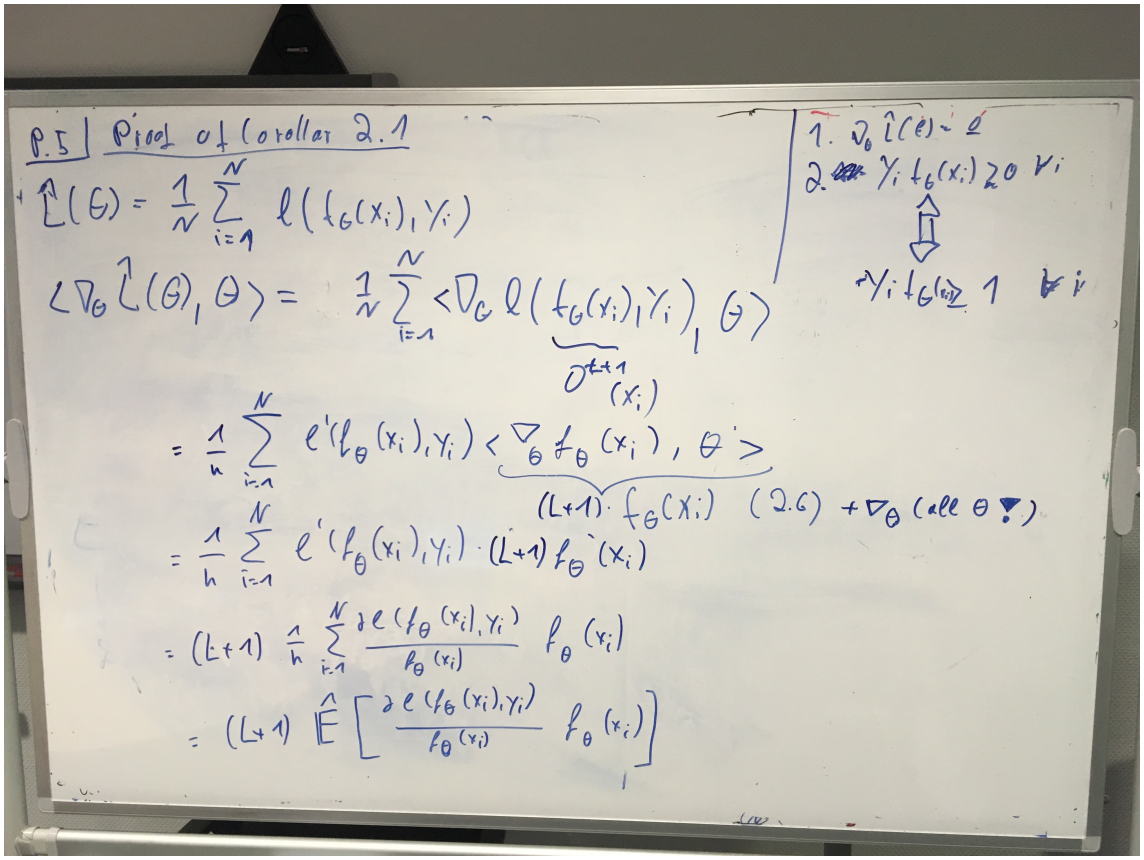
□




0 1 2 ... L
layer
non-lin
$O^0(x) = X \in \mathbb{R}^p$
$O^{L+1}(x) = f_G(x) \in \mathbb{R}^K$
$O^t(x) = \sigma_t(N^t(x))$
$\sigma(\sigma(X^T w^0) w^1)$
$X \in \mathbb{R}^{28^2}$
$y \in \mathbb{R}^{10}$
$w^0 \in 28^2, w^1 \in 1020, w^2 = 10$
2.2: $\sigma(x) = \sigma'(x) x$


$\ell = \max\{0, 1 - y_i f_i\}$
1) $\frac{\partial \ell(f, y)}{\partial f} = -y \iff yf < 1$
$\to 1 - y_i f_i > 0 \Rightarrow \ell = 1 - y_i f_i$
$\to \frac{\partial \ell}{\partial f} = -y_i$ 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. $\partial_\theta \hat{L}(\theta) = 0$
2. $y_i f_G(x_i) \geq 0 \ \forall i$
$y_i f_G(x_i) \geq 1 \ \forall i$


P.5 Proof of Corollar 2.1
$\hat{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)$
$\langle \nabla_\theta \hat{L}(\theta), \theta \rangle = \frac{1}{N}\sum_{i=1}^{N} \langle \nabla_\theta \ell(f_\theta(x_i), y_i), \theta \rangle$
$\mathcal{O}^{L+1}(x_i)$
$= \frac{1}{N}\sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \langle \nabla_\theta f_\theta(x_i), \theta \rangle$
$(L+1) \cdot f_\theta(x_i)$ (2.6) $+ \nabla_\theta$ (alle $\theta$ ?)
$= \frac{1}{N}\sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \cdot (L+1) f_\theta(x_i)$
$= (L+1) \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)$
$= (L+1) \hat{\mathbb{E}}\left[\frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)\right]$
1. $\nabla_\theta \hat{L}(\theta) = 0$
2. $y_i f_\theta(x_i) \geq 0 \ \forall i$
$\Leftrightarrow y_i f_\theta(x_i) \geq 1 \ \forall i$


P.5 Proof of Corollar 2.1
$\hat{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)$
$\langle \nabla_\theta \hat{L}(\theta), \theta \rangle = \frac{1}{N}\sum_{i=1}^{N} \langle \nabla_\theta \ell(f_\theta(x_i), y_i), \theta \rangle$
$\mathcal{O}^{L+1}(x_i)$
$= \frac{1}{N}\sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \langle \nabla_\theta f_\theta(x_i), \theta \rangle$
$(L+1) \cdot f_\theta(x_i)$ (2.6) $+ \nabla_\theta$ (alle $\theta$ ?)
$= \frac{1}{N}\sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \cdot (L+1) f_\theta(x_i)$
$= (L+1) \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)$
$= (L+1) \hat{\mathbb{E}}\left[\frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)\right]$
1. $\nabla_\theta \hat{L}(\theta) = 0$
2. $y_i f_\theta(x_i) \geq 0 \ \forall i$
$\Leftrightarrow y_i f_\theta(x_i) \geq 1 \ \forall i$

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

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

$\langle \theta, \nabla_\theta \ell(f_\theta(x,y)) \rangle^2 = \langle \theta, \nabla_\theta \ell(f_\theta(x,y)) \rangle \cdot \langle \theta, \nabla_\theta \ell(x,y) \rangle$ $\quad \langle x, y \rangle = x^T y$

$= \left[ \theta^T \nabla_\theta \ell(f_\theta(x,y)) \right] \cdot \left[ \nabla_\theta \ell(f_\theta(x), y)^T \theta \right]$

$= \theta^T \underbrace{\left( \nabla_\theta \ell(f_\theta(x,y)) \nabla_\theta \ell(f_\theta(x,y)) \right)}_{I} \theta$

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

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

$\langle \theta, \nabla_\theta \ell(f_\theta(x), y) \rangle$

$= \langle \left[ \nabla_\theta \left( \left( f_\theta(x), y \right) \right]^T, \theta \rangle$

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

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

$= x^T (\alpha y)$

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

(Chain) $= \left\langle \frac{\partial \ell(f_\theta(x), y)}{\partial f_\theta(x)} \nabla_\theta f_\theta(x), \theta \right\rangle$

$= \left\langle \nabla_\theta f_\theta(x) \frac{\partial \ell(f_\theta(x), y)}{\partial f_\theta(x)}, \theta \right\rangle$

$= \left\langle \frac{\partial \ell(f_\theta(x), y)}{\partial f_\theta(x)}, \nabla_\theta f_\theta(x)^T \theta \right\rangle$

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

$$E\left[ \left( \frac{\partial \ell(f_\theta(x), y)}{\partial f_\theta(y)}, (L+1) f_\theta(x) \right)^2 \right]$$

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

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

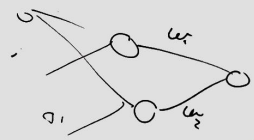
$\frac{(f_\theta(x)-y)^2}{2}$ $\quad G = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix}$ $\quad f_\theta^2(x) - y f_\theta(x)$

$\nabla_\theta O^{L+1} = \begin{pmatrix} \frac{\partial O^{L+1}}{\partial w_1} \\ \vdots \\ \partial w_N \end{pmatrix}$ $\quad \nabla\theta^T\theta =$ $\quad (\nabla\theta)^T G^{th}$

$\nabla_\theta O^{L+1}(x)^T \theta$

$= \sum_{i=0}^{|\theta|} \frac{\partial O^{L+1}(x)}{\partial w_i} w_i$

$= \sum_{t=0}^{L} \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial O^{L+1}(x)}{w_{ij}^t} w_{ij}^t$



$\|\theta\|_{tr}^2 = \langle \theta, I(G) G \rangle$

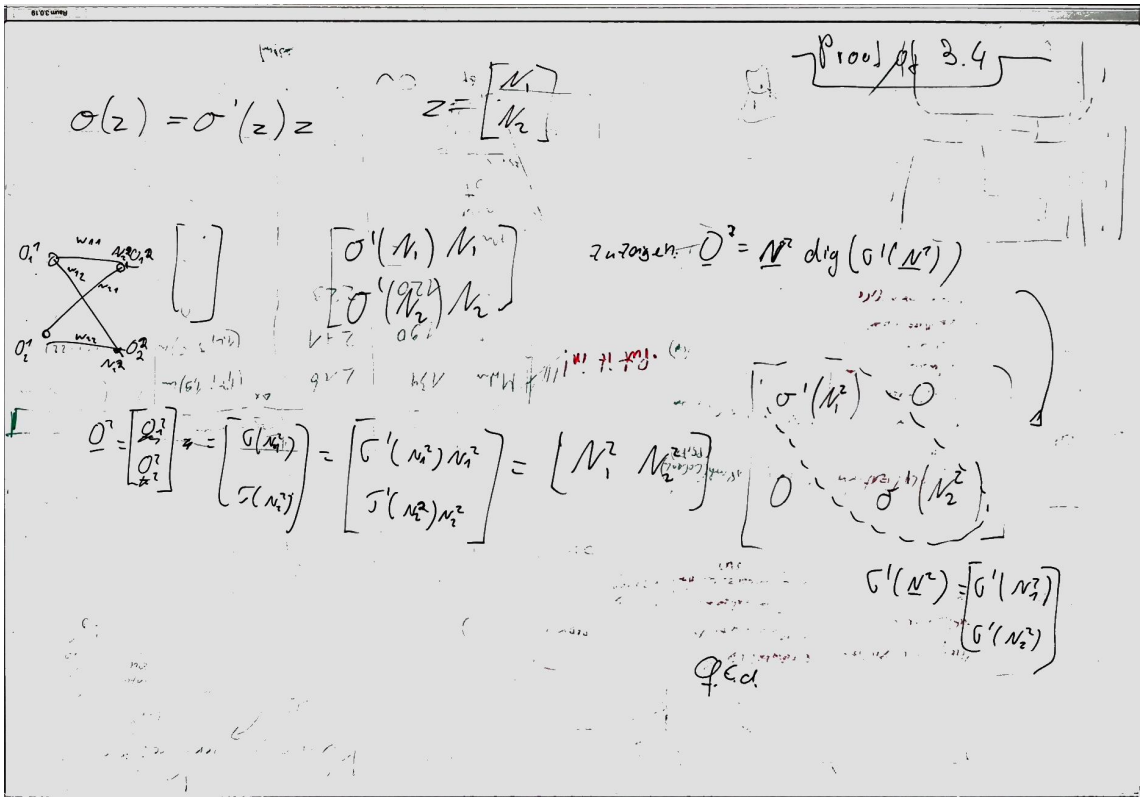
$\vdots$

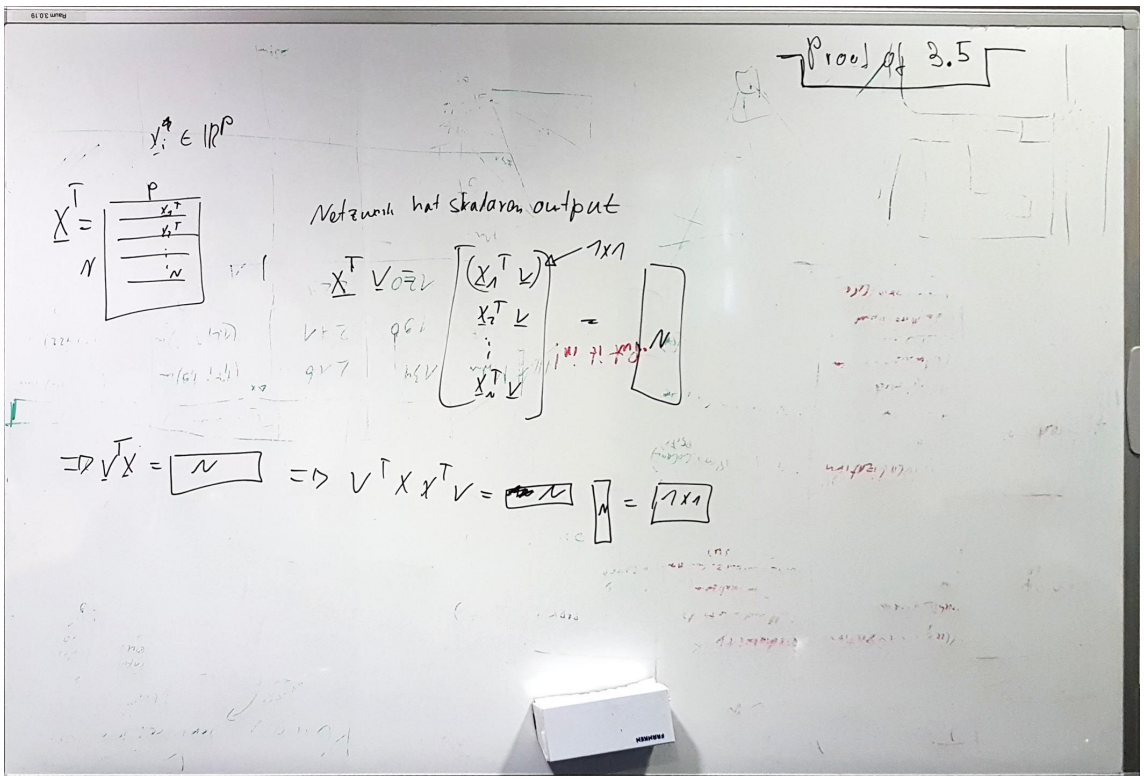
$= (L+1)^2 E\left( \left\langle \frac{\partial \ell(f_\theta(x), y)}{\partial f_\theta(x)}, f_\theta(x) \right\rangle^2 \right)$

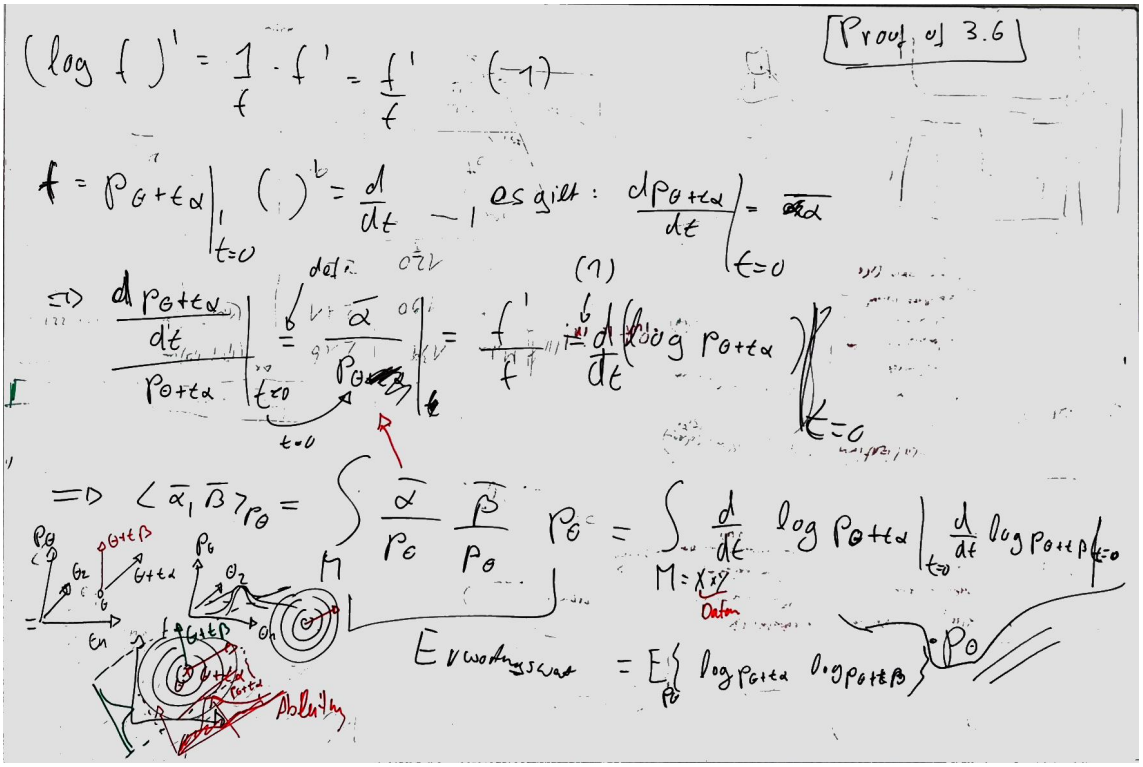
Norm von $\theta$ ist unabhängig von $G$

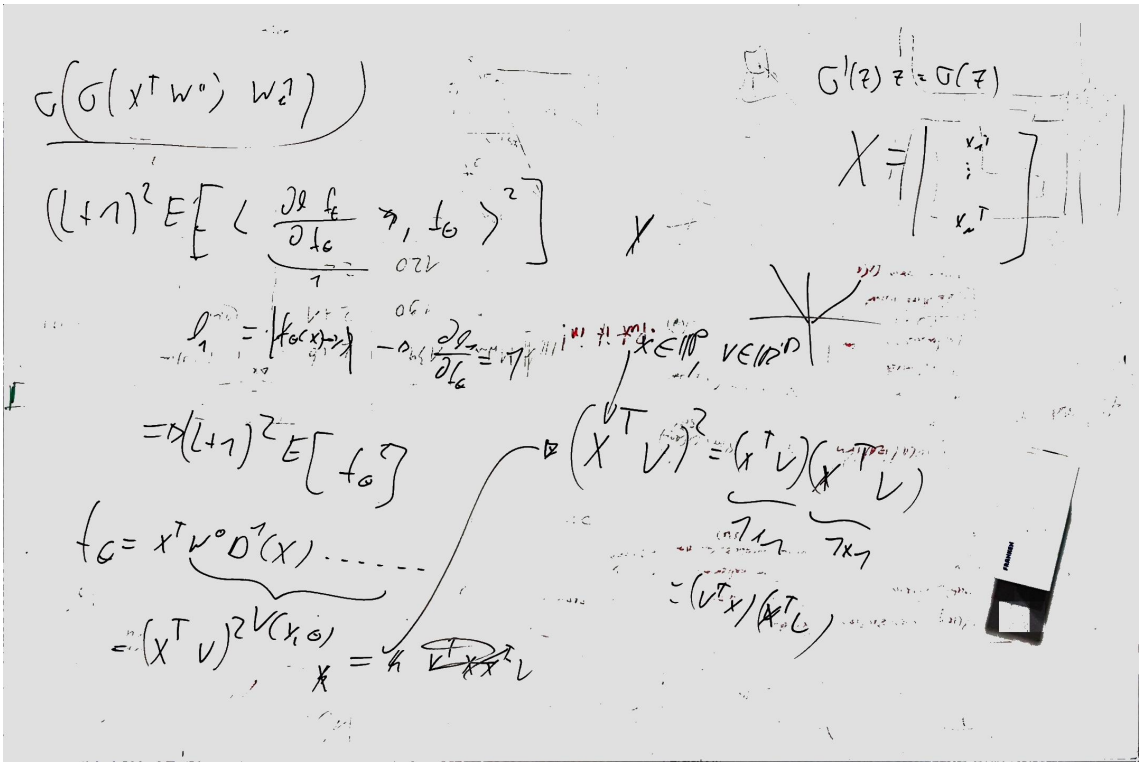
$(f - y)^2$

$f - y$


Proof of 3.4
$\sigma(z) = \sigma'(z) z$
$z = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}$
$\begin{bmatrix} \sigma'(N_1) N_1 \\ \sigma'(N_2) N_2 \end{bmatrix}$
$\underline{O}^2 = \underline{N}^2 \operatorname{diag}(\sigma'(\underline{N}^2))$
$\underline{O}^2 = \begin{bmatrix} O_1^2 \\ O_2^2 \end{bmatrix} = \begin{bmatrix} \sigma(N_1^2) \\ \sigma(N_2^2) \end{bmatrix} = \begin{bmatrix} \sigma'(N_1^2) N_1^2 \\ \sigma'(N_2^2) N_2^2 \end{bmatrix} = \begin{bmatrix} N_1^2 & N_2^2 \end{bmatrix}$
$\begin{bmatrix} \sigma'(N_1^2) & 0 \\ 0 & \sigma'(N_2^2) \end{bmatrix}$
$\sigma'(\underline{N}^2) = \begin{bmatrix} \sigma'(N_1^2) \\ \sigma'(N_2^2) \end{bmatrix}$
q.e.d.


Proof of 3.5
$y_i \in \mathbb{R}^d$
$\underline{X} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}$ N, d
Netzwerk hat skalaren output
$\underline{X}^T \underline{V} = \begin{bmatrix} \underline{x}_1^T \underline{V} \\ \underline{x}_2^T \underline{V} \\ \vdots \\ \underline{x}_N^T \underline{V} \end{bmatrix} = \underline{N}$
$\Rightarrow \underline{X}^T \underline{V} = \underline{N} \Rightarrow \underline{V}^T \underline{X} \underline{X}^T \underline{V} = \underline{N}^T \underline{N}$
$\underline{N} = \mathbb{R}^{1 \times 1}$


Proof of 3.6
Probability



3.6

Richtungsableitung $f(x,y/\theta) = p_\theta(x,y)$

$\langle \nabla_\theta p_\theta(x), \alpha \rangle = \frac{d\, p_{\theta+\epsilon\alpha}}{dt} = \bar{\alpha}$ $\quad \bar{\alpha} = \langle \nabla_\theta p_\theta(x), \alpha \rangle = (\nabla_\theta p_\theta(x))^T \alpha$

$\langle \bar{\alpha}, \bar{\beta} \rangle_{p_\theta} = E_{(x,y)\sim p_\theta}\left[ \frac{\bar{\alpha}}{p_\theta} \frac{\bar{\beta}}{p_\theta} \right] = \int_{(x,y)\sim p_\theta} \left[ \frac{(\nabla_\theta p_\theta(x))^T}{p_\theta(x)} \alpha \cdot \frac{(\nabla_\theta p_\theta(x))^T}{p_\theta(x)} \beta \right]$ (1×1, 1×1)

$\nabla_\theta \log p_\theta(x)$

$= E_{(x,y)\sim p_\theta}\left[ \alpha^T \nabla_\theta \log p_\theta(x) \nabla_\theta \log p_\theta(x)^T \beta \right]$

$\langle \alpha, I(\theta)\beta \rangle = \alpha^T E_{(x,y)\sim p_\theta}\left[ \nabla_\theta \log p_\theta(x) \nabla_\theta \log p_\theta(x)^T \right] \beta$

$\frac{d\, f(x+t\Delta x)}{dt}$ $\quad \int_{-\infty}^{\infty} x\, p_\theta(x)\, dx = E[x]$

$\mathbb{R}^n \times \mathbb{R}^4$

$\nabla f(x,\alpha)$

$\int x f(x) dx$

$= \frac{f'}{f} = (\log f)'$

$f = p_\theta(x) \Rightarrow \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} = \nabla_\theta \log p_\theta(x)$

$\int_{M(x\times y)} \left( \frac{\bar{\alpha}}{p_\theta} \cdot \frac{\bar{\beta}}{p_\theta} \right) p_\theta(x)\, d(x,y)$ $\quad E[\,]$

Def 3.)

$$\|\Theta\|_\sigma = \left[ \mathbb{E}\left( \|X\|^2 \prod_{t=1}^{L+1} \|D^t(x)\|_\sigma^2 \right] \right]^{1/2} \prod_{t=0}^{L} \|W^t\|_\sigma$$

$\|D^t(x)\|_0 \in \{0,1\}$ and (1)

$$= \sqrt{\mathbb{E}\left( \|X\|^2 \prod_{t=1}^{L+1} \|D^t(x)\|_\sigma^2 \prod_{t=0}^{L} \|W^t\|_\sigma^2 \right)}$$

$$= \sqrt{\mathbb{E}\left( \|X\|^2 \prod_{t=0}^{L} \|W^t \operatorname{diag}[\sigma'(N^{t+1}(x))] \|_\sigma^2 \right)}$$

$$\overset{(2)}{=} \sqrt{\mathbb{E}\left( \|X^T\|_\sigma^2 \prod_{t=0}^{L} \|W^t \operatorname{diag}[\sigma'(N^{t+1}(x)] \|_\sigma^2 \right)}$$

$$D^t(x) = \operatorname{diag}[\sigma'(N^t(x))] \qquad \sigma(x) = \sigma'(x)\, x$$

only 0 or 1

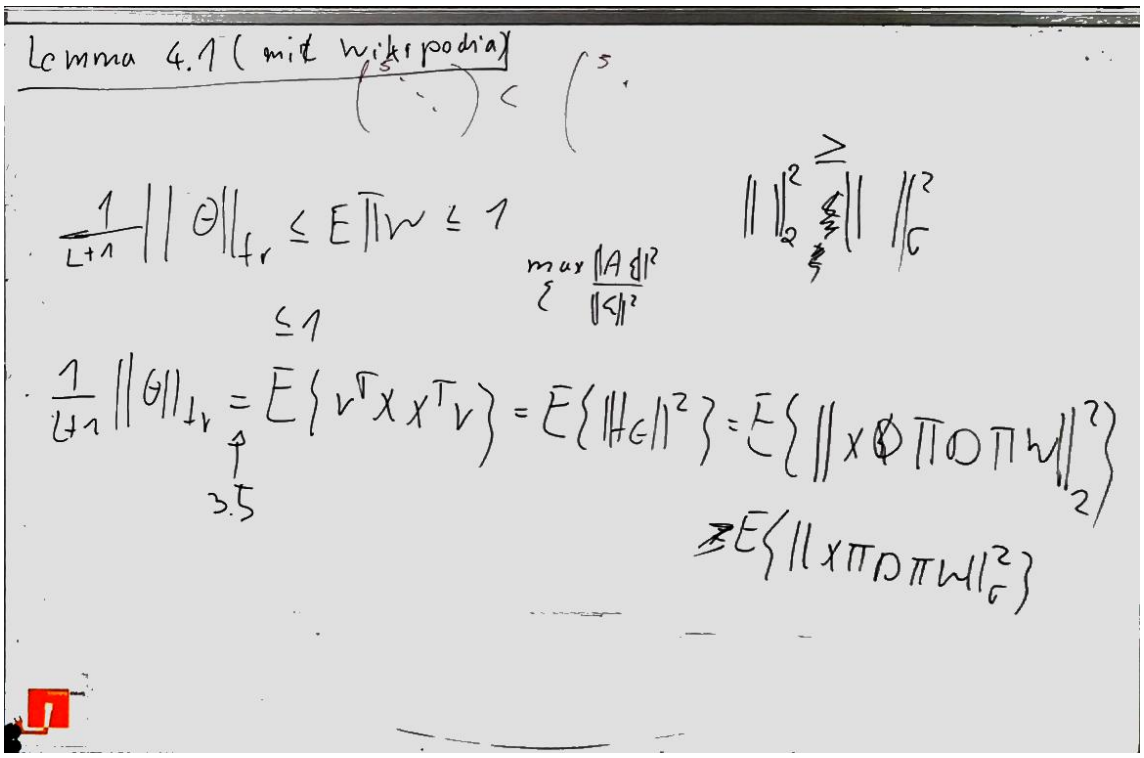
$$\|f\|_\sigma^2 = \|f\|^2 \quad (2)$$

$$\|XM\|^2 \le \|X\|^2 \|M\|_\sigma^2$$

$$\geq \sqrt{\mathbb{E}\left[ \|X^T\|_\sigma^2 \left\| \prod_{t=0}^{L} W^t D^{t+1}(x) \right\|_\sigma^2 \right]}$$

$$\geq \sqrt{\mathbb{E}\left[ \left\| X^T \prod_{t=0}^{L} W^t D^{t+1}(x) \right\|_\sigma^2 \right]}$$

$$= \sqrt{\mathbb{E}\left[ \|\nabla_\Theta(x)\|_\sigma^2 \right]}$$


Lemma 4.1 (mit Wikipedia)
$\frac{1}{L+1}\|\Theta\|_{Fr} \le E\,\mathrm{Tr}\,W \le 1$
$\le 1$
$\max \frac{\|Ax\|^2}{\|x\|^2}$
$\|\ \|_2^2 \geq \|\ \|_G^2$
$\frac{1}{L+1}\|\Theta\|_{Tr} = E\{v^T X X^T v\} = E\{\|H_G\|^2\} = E\{\|X \Phi \Pi \odot \Pi W\|_2^2\}$
3.5
$\geq E\{\|X \Pi \odot \Pi W\|_G^2\}$

# 2 Notes on Section 4

### 2.0.1 Explaining Lemma 4.1

From 3.5 in the paper, we had:

$$\frac{1}{(L+1)^2}\left\|\theta\right\|_{fr}^2 = \mathrm{E}\left\{v^{\mathrm{T}}XX^{\mathrm{T}}v\right\} = \mathrm{E}\left\{\left\|f_\theta\right\|^2\right\}. \tag{19}$$

For the Frobenius Norm of a Matrix $A$ and a vector $x$ the following holds.

$$\left\|A\right\|_F \geq \left\|A\right\|_\sigma \tag{20}$$
$$\left\|x\right\|_F = \left\|x\right\|_2 = \left\|x\right\|_\sigma \tag{21}$$
$$\left\|Ax\right\|_\sigma \leq \left\|A\right\|_\sigma \left\|x\right\|_2 \tag{22}$$
$$\text{for vectors x.} \tag{23}$$

It follows.

$$\mathrm{E}\left\{\left\|f_\theta\right\|_\sigma^2\right\} = \mathrm{E}\left\{\left\|f_\theta\right\|_\sigma^2\right\} \tag{24}$$
$$= \mathrm{E}\left\{\left\|x^T W^0 D^1(x) W^1 D^2(x) \dots D^L W^L D^{L+1}(x)\right\|_\sigma^2\right\} \tag{25}$$
$$\leq \mathrm{E}\left\{\left\|x\right\|_\sigma^2 \prod \left\|D^i(x)\right\|_\sigma^2 \prod \left\|W^i\right\|_\sigma^2\right\}. \tag{26}$$

Since $W^i$ is independent of the data $x$, it does not have to be inside the expectation.

$$\frac{1}{(L+1)^2}\left\|\theta\right\|_{fr}^2 = \mathrm{E}\left\{\left\|f_\theta\right\|_\sigma^2\right\} \tag{27}$$
$$\leq \mathrm{E}\left\{\left\|x\right\|_\sigma^2 \prod \left\|D^i(x)\right\|_\sigma^2 \prod \left\|W^i\right\|_\sigma^2\right\} \tag{28}$$
$$= \mathrm{E}\left\{\left\|x\right\|_2^2 \prod_{t=1}^{L+1} \left\|D^t(x)\right\|_\sigma^2\right\} \prod_{t=0}^{L} \left\|W^i\right\|_\sigma^2 \tag{29}$$

Now taking the root on both sides reveals Lemma 4.1 and concludes the explainiation.

## References