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

Sandro Braun, Leander Kurscheidt

July 25, 2018

Abstract

1 Geometry of Deep Rectified Networks

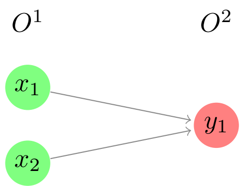
1.1 Lemma 2.1

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

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

1.1.1 Example for Lemma 2.2

$$\begin{aligned}
\frac{\partial O^2}{\partial W_1} &= \sigma'(z)O^1W_1 \\
\frac{\partial O^2}{\partial W_2} &= \sigma'(z)O^1W_2 \\
\text{therefore:} & \\
\sum_{t=1}^{2} \frac{\partial O^2}{\partial W_t} &= \sigma'(z)(\underbrace{O^1W_1 + O^1W_2}_{\text{z}}) \\
&= \sigma'(z)z
\end{aligned}$$



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,

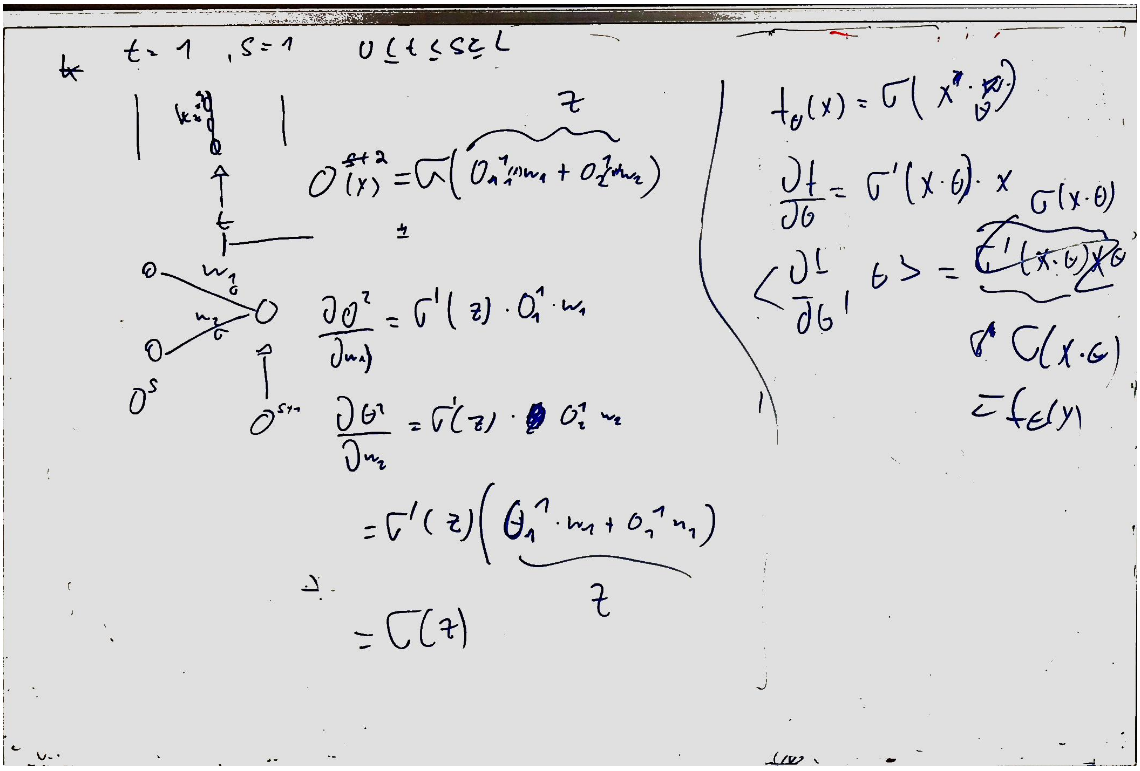
$$\begin{aligned}
& 1 - y_i f_i > 0 \\
\Leftrightarrow \quad & l = 1 - y_i f_i \\
\Leftrightarrow \quad & \frac{\partial l}{\partial f} = -y_i
\end{aligned}$$

□

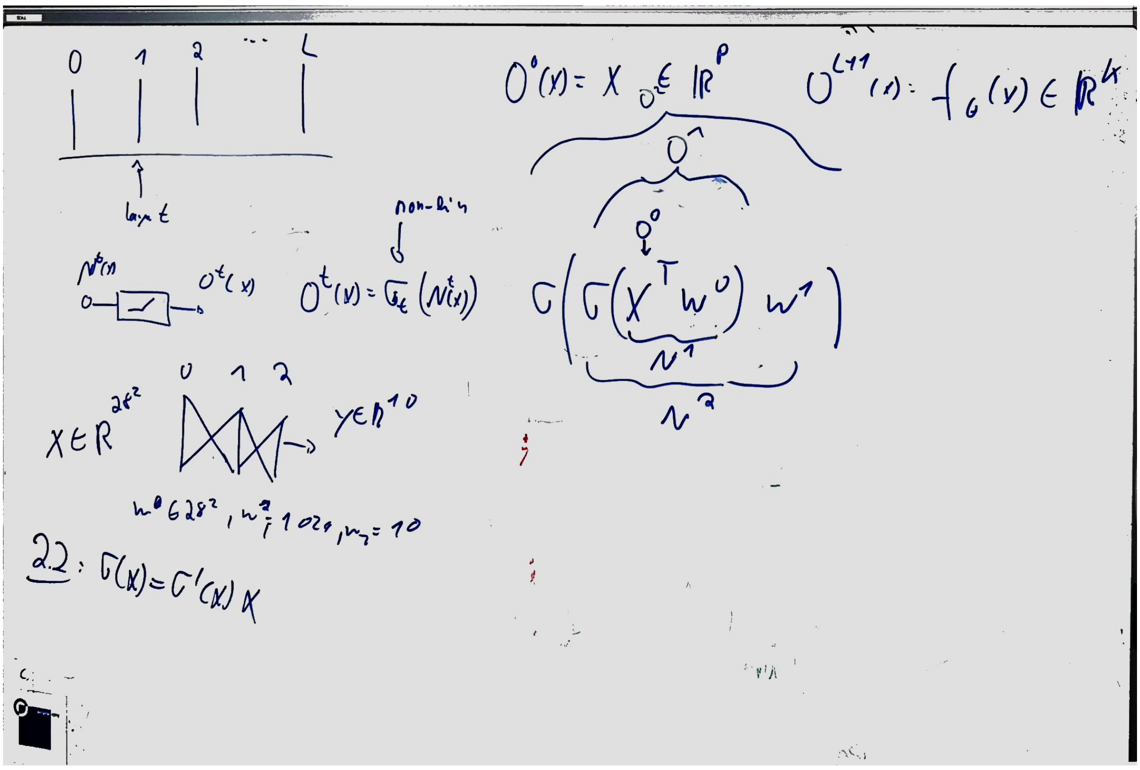
2. *Proof.* We want to show $\frac{\partial l(f,Y)}{\partial f} = 0 \Leftrightarrow yf > 1$. So,

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

□



t = 1 , s = 1 0 ≤ t ≤ s ≤ L
$f_\theta(x) = \sigma(x^T \cdot \theta)$
$\frac{\partial f}{\partial \theta} = \sigma'(x \cdot \theta) \cdot x$
$\frac{\partial O^2}{\partial w_1} = \sigma'(z) \cdot O_1^1 \cdot w_1$
$\frac{\partial O^2}{\partial w_2} = \sigma'(z) \cdot O_2^1 w_2$
$= \sigma'(z)(O_1^1 \cdot w_1 + O_2^1 w_2)$
$= \sigma(z)$
$= f_\theta(y)$



0 1 2 … L
layer t
non-lin
$O^0(x) = x \in \mathbb{R}^p$
$O^{L+1}(x) = f_\theta(x) \in \mathbb{R}^k$
$O^t(x) = \sigma_t(N^t(x))$
$\sigma(\sigma(x^T w^0) w^1)$
$X \in \mathbb{R}^{28^2}$
$Y \in \mathbb{R}^{10}$
$2.2: \sigma(x) = \sigma'(x)x$

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

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

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

$\to \dfrac{\partial \ell}{\partial f} = -y_i$ q.e.d.

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

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

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

$\Updownarrow$

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

P.5 | Proof of Corollar 2.1

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

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

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

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

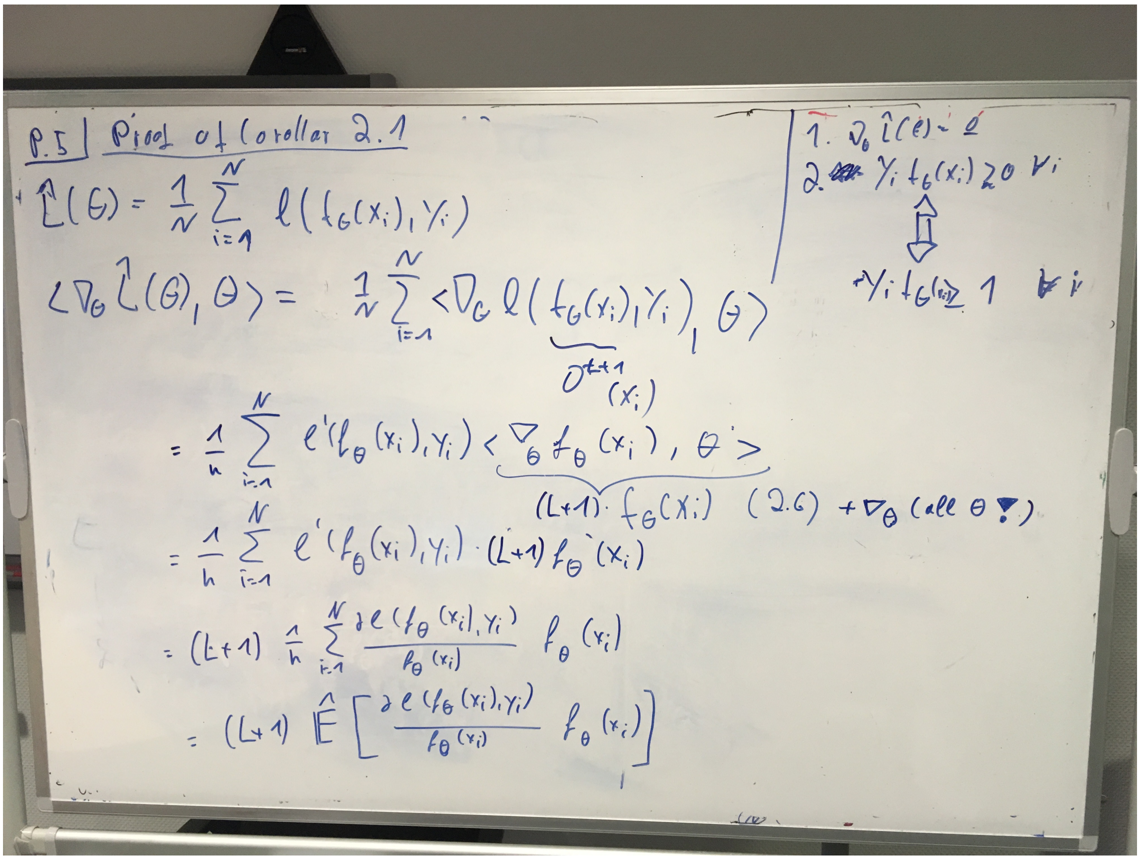
$$= (L+1) \frac{1}{n} \sum_{i=1}^{N} \frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)$$

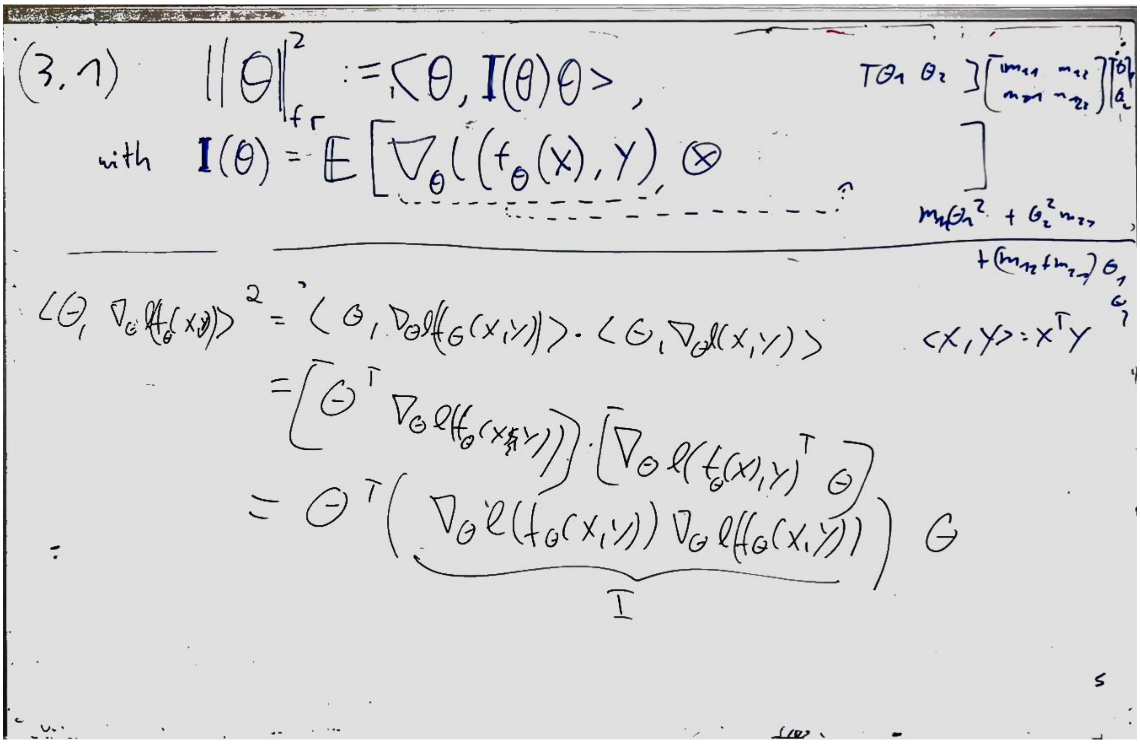
$$= (L+1) \hat{\mathbb{E}}\left[\frac{\partial \ell(f_\theta(x_i), y_i)}{f_\theta(x_i)} f_\theta(x_i)\right]$$

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

$\Updownarrow$

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


P.5 | Proof of Corollary 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) $+ \nabla_\theta$ (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. $\nabla_\theta \hat{L}(\theta) = 0$
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) $\|\theta\|^2_{fr} := \langle \theta, I(\theta)\theta \rangle$,

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

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

$m_{11}\theta_1^2 + \theta_2^2 m_{22} + (m_{12}+m_{21})\theta_1\theta_2$

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

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

chain

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

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

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

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

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

$= x^T(\alpha y)$

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

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

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

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

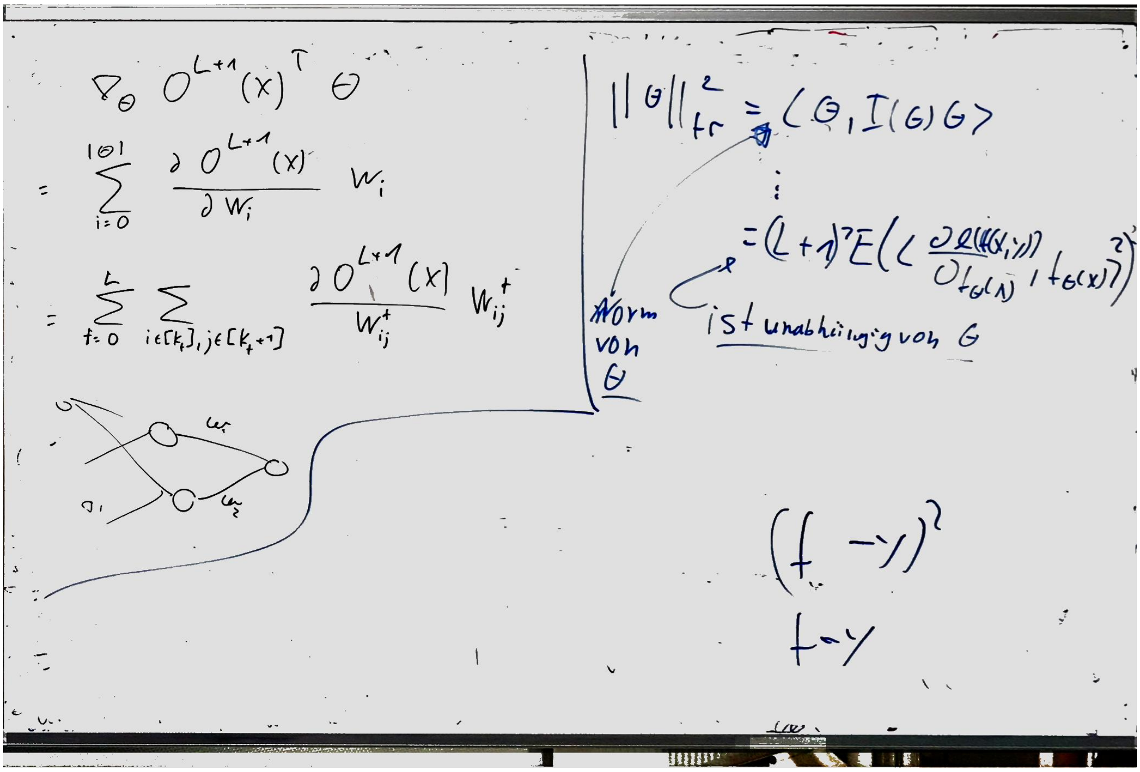
$l = \frac{(\;)^2}{2} \to \frac{\partial l}{\partial f} = |\quad| \Rightarrow E\left[\langle f_\theta(x) - y, f_\theta(x) \rangle^2\right]$

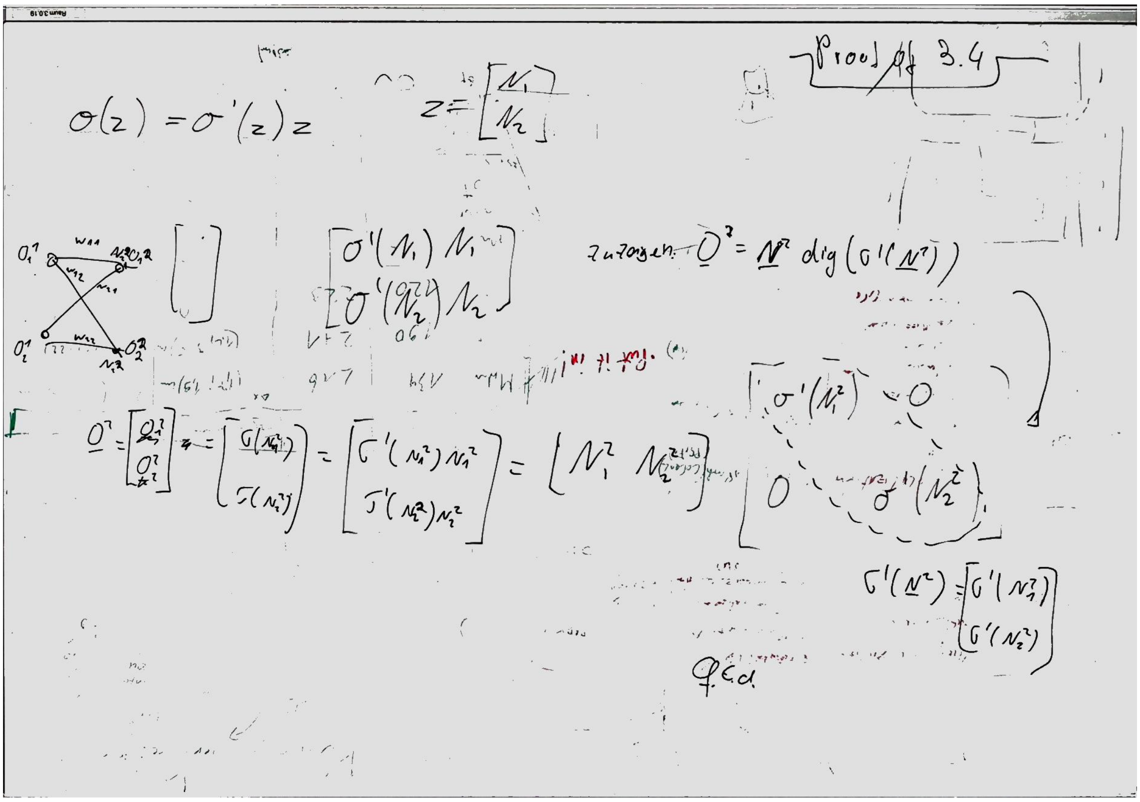
$\frac{(f_\theta(x)-y)^2}{2}$ $\quad G = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix}$ $\quad f_\theta^2(x) - y f_\theta(x)$

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

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

$(\nabla\theta)^T G$


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


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

Proof of 3.5

$x_i \in \mathbb{R}^p$

$$\underline{X}^T = \begin{array}{c|c} & p \\ \hline N & \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \end{array}$$

Netzwerk hat skalaren output

$$\underline{X}^T \underline{v} = \begin{bmatrix} \overset{1\times 1}{(\underline{x}_1^T \underline{v})} \\ \underline{x}_2^T \underline{v} \\ \vdots \\ \underline{x}_N^T \underline{v} \end{bmatrix} = \begin{bmatrix} \\ N \\ \\ \end{bmatrix}$$

$$\Rightarrow \underline{v}^T \underline{X} = \boxed{\quad N \quad} \Rightarrow v^T X X^T v = \boxed{\quad N \quad}\,\boxed{N} = \boxed{1\times 1}$$

Proof of 3.6

$$(\log f)' = \frac{1}{f} \cdot f' = \frac{f'}{f} \quad (1)$$

$$f = p_{\theta+t\alpha}\Big|_{t=0}, \quad (\;)' = \frac{d}{dt} \qquad \text{es gilt:} \quad \frac{dp_{\theta+t\alpha}}{dt}\Big|_{t=0} = \bar{\alpha}$$

$$\Rightarrow \frac{\frac{dp_{\theta+t\alpha}}{dt}}{p_{\theta+t\alpha}}\Bigg|_{t=0} \overset{\text{def}}{=} \frac{\bar{\alpha}}{p_\theta}\Bigg|_{t} = \frac{f'}{f} \overset{(1)}{=} \frac{d}{dt}\left(\log p_{\theta+t\alpha}\right)\Bigg|_{t=0}$$

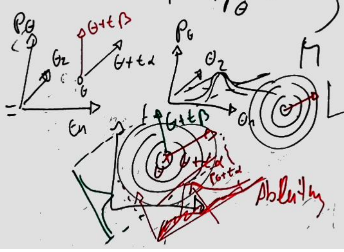
$t=0$

$$\Rightarrow \langle \bar{\alpha}, \bar{\beta} \rangle_{p_\theta} = \underbrace{\int_M \frac{\bar{\alpha}}{p_\theta} \frac{\bar{\beta}}{p_\theta}\, p_\theta}_{\text{Erwartungswert}} = \int_{M} \frac{d}{dt} \log p_{\theta+t\alpha}\Big|_{t=0} \frac{d}{dt} \log p_{\theta+t\beta}\Big|_{t=0}$$

$M = X \times Y$

Daten

$$\text{Erwartungswert} = E_{p_\theta}\{\log p_{\theta+t\alpha}\, \log p_{\theta+t\beta}\} \quad p_\theta$$



References