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

Sandro Braun, Leander Kurscheidt

July 29, 2018

Abstract

1 Geometry of Deep Rectified Networks

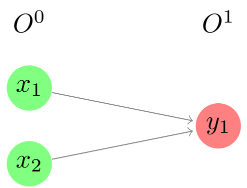
1.1 Lemma 2.1

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 2.1

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:

$$O^0 = \begin{bmatrix} O^0_1 \\ O^0_2 \end{bmatrix}, W^0 = \begin{bmatrix} W^0_1 \\ W^0_2 \end{bmatrix} \tag{2}$$

$$O^1 = \sigma({O^0}^T W^0) = \sigma(\underbrace{O^0_1 W^0_1 + O^0_2 W^0_2}_{z}). \tag{3}$$

Substituting the term in brackets will simplify notation.

Now calculate the partial derivatives of (2) with respect to W^0_i.

$$\frac{\partial O^1}{\partial W^0_1} = \sigma'(z) O^0_1$$
$$\frac{\partial O^1}{\partial W^0_2} = \sigma'(z) O^0_2$$

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

$$\sum_{j=1}^{2} \frac{\partial O^1}{\partial W^0_j} W^0_j = \sigma'(z) O^0_1 W^0_1 + \sigma'(z) O^0_2 W^0_2 = \sigma'(z)(\underbrace{O^0_1 W^0_1 + O^0_2 W^0_2}_{\text{z}})$$
$$= \sigma'(z) z$$

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.

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

1.2 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,

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

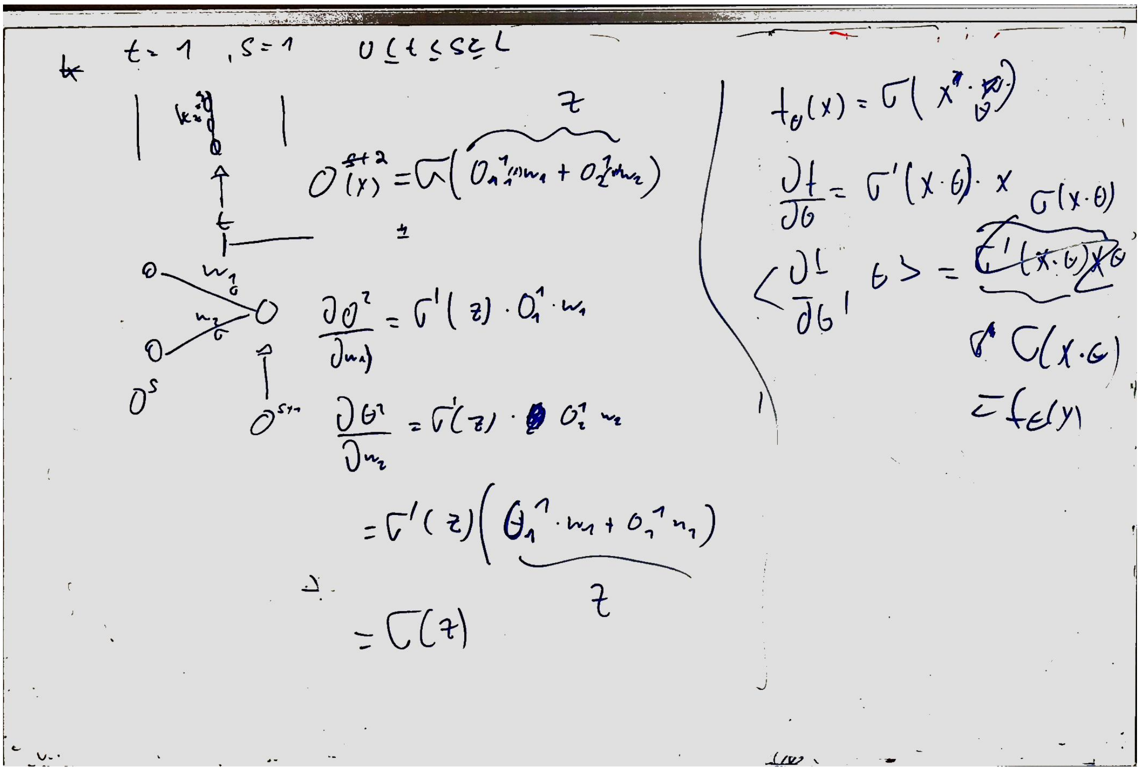
$\square$

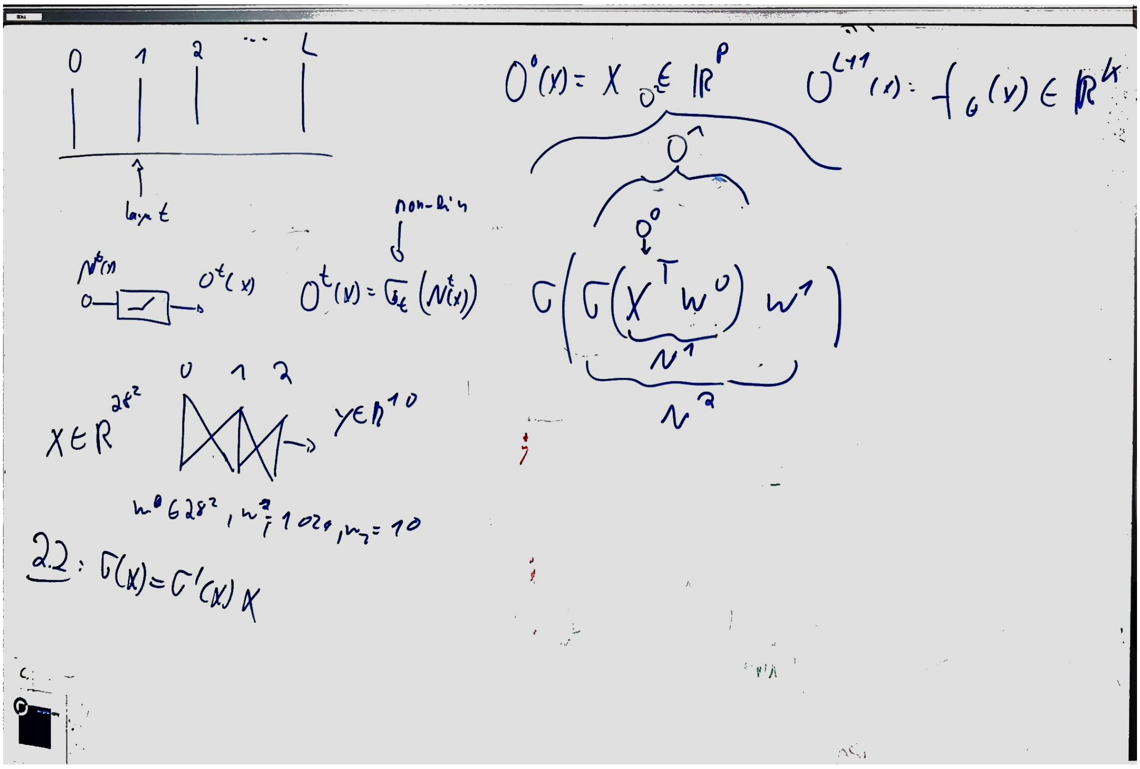
2. *Proof.* We want to show $\frac{\partial l(f,Y)}{\partial f} = 0 \Leftrightarrow yf > 1$. So,

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

$\square$

References


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


0 1 2 ... L
layer t
$N^t(x)$ $O^t(x)$
$O^t(x) = \sigma_t(N^t(x))$
non-lin
$O^0(x) = X \in \mathbb{R}^p$
$O^{L+1}(x) = f_\theta(x) \in \mathbb{R}^k$
$\overbrace{\sigma(\sigma(\underbrace{X^T W^0}_{N^1}) W^1)}^{O^1}$
N^2
$X \in \mathbb{R}^{28^2}$
0 1 2
$Y \in \mathbb{R}^{10}$
$w^0\ 628^2,\ w^1\ 1024,\ w_2 = 10$
2.2: $\sigma(x) = \sigma'(x) x$

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

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

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

$$\to \frac{\partial \ell}{\partial f} = -y_i \quad \text{q.e.d.}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $I(\theta) = E\left[\nabla_\theta \ell(f_\theta(x), y) \otimes \quad \right]$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

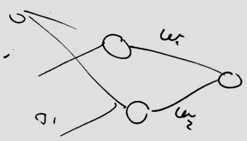
$\nabla_\theta \theta^{L+1} = \begin{bmatrix} \frac{\partial \theta^{L+1}}{\partial w_1} \\ \vdots \\ \partial w_N \end{bmatrix}$

$\cancel{\nabla\theta^T\theta =}$ $\quad (\nabla\theta)^T G$

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

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

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



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

$$\vdots$$

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

Norm von Θ

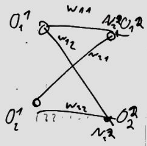
ist unabhängig von Θ

$$(f - y)^2$$

$$f = y$$

Proof of 3.4

$$\sigma(z) = \sigma'(z)\, z \qquad z = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}$$



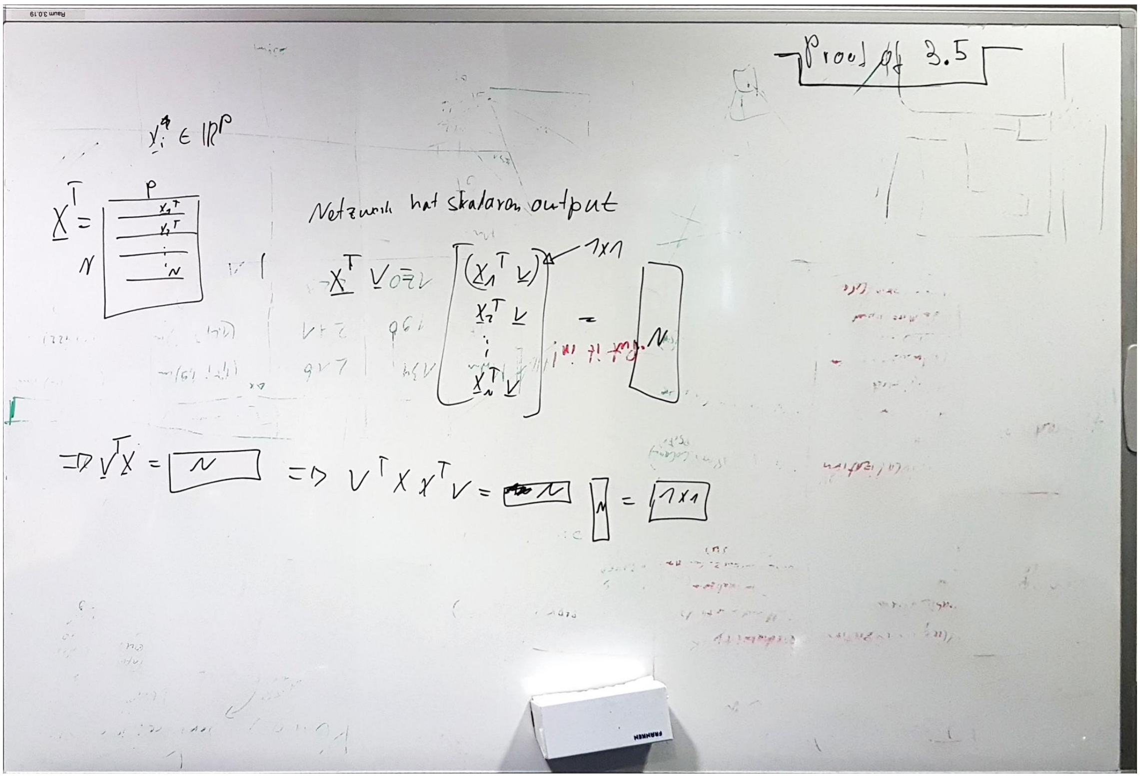
$$\begin{bmatrix} \sigma'(N_1)\, N_1 \\ \sigma'(N_2)\, N_2 \end{bmatrix}$$

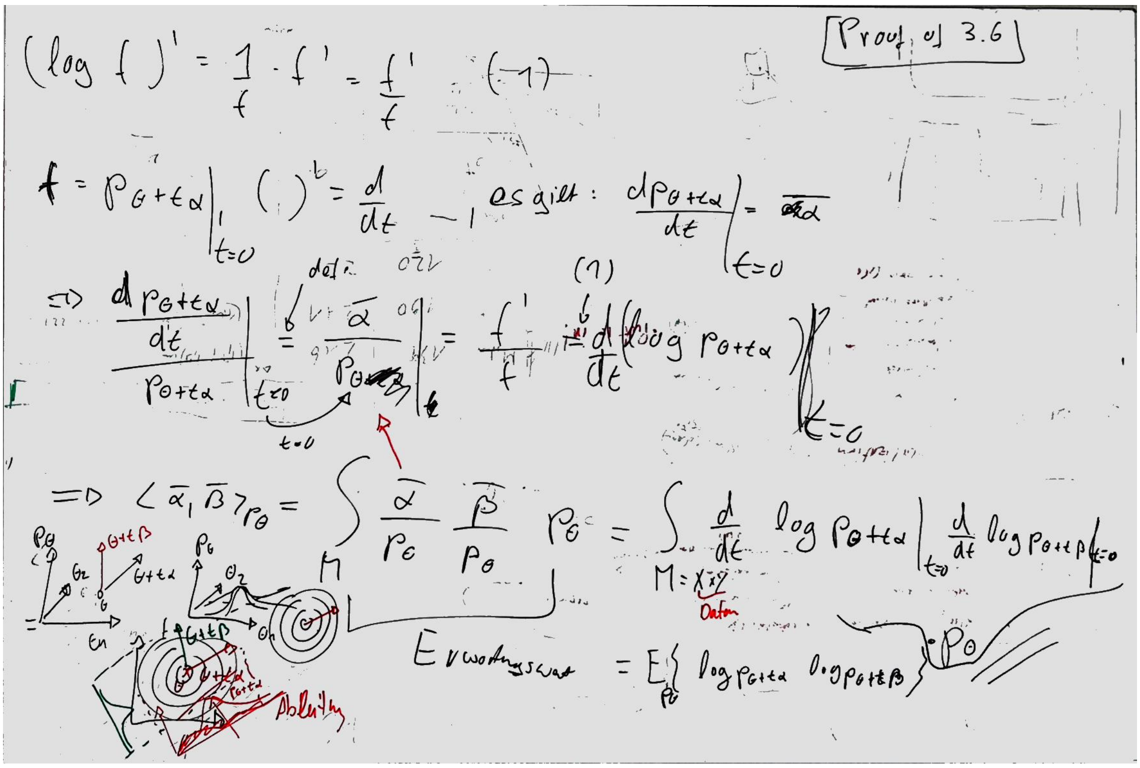
zu zeigen: $\underline{O}^2 = \underline{N}^2 \operatorname{diag}(\sigma'(\underline{N}^2))$

$$\underline{O}^2 = \begin{bmatrix} O_1^2 \\ O_2^2 \end{bmatrix} = \begin{bmatrix} \sigma(N_1^2) \\ \sigma(N_2^2) \end{bmatrix} = \begin{bmatrix} \sigma'(N_1^2)\, N_1^2 \\ \sigma'(N_2^2)\, N_2^2 \end{bmatrix} = \begin{bmatrix} N_1^2 & N_2^2 \end{bmatrix} \begin{bmatrix} \sigma'(N_1^2) & 0 \\ 0 & \sigma'(N_2^2) \end{bmatrix}$$

$$\sigma'(\underline{N}^2) = \begin{bmatrix} \sigma'(N_1^2) \\ \sigma'(N_2^2) \end{bmatrix}$$

q.e.d.


Proof of 3.5
Netzwerk hat skalaren output
1x1
N
1x1


Proof of 3.6
es gilt :
Daten
Erwartungswert
Ableitung

$$\sigma\left(\sigma(x^T w^0)\, w_1\right)$$

$$\sigma'(z)\, z = \sigma(z)$$

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

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$$l_1 = |f_\theta(x) - y| \qquad \frac{\partial l_1}{\partial f_\theta} = 1$$

$$x \in \mathbb{R}^D, \quad v \in \mathbb{R}^D$$

$$= \alpha (L+1)^2 E\left[f_\theta^2 \right]$$

$$f_\theta = \underbrace{x^T w^0 D^1(x)}\ \ldots\ldots$$

$$= (x^T v)^2\, V(x,\theta)$$

$$(x^T v)^2 = \underbrace{(x^T v)}_{1\times 1}\underbrace{(x^T v)}_{1\times 1}$$

$$= (v^T x)(x^T v)$$

$$= v^T x x^T v$$

Richtungsableitung

$$f(x,y \mid \theta) = p_\theta(x,y) \qquad \boxed{3.6}$$

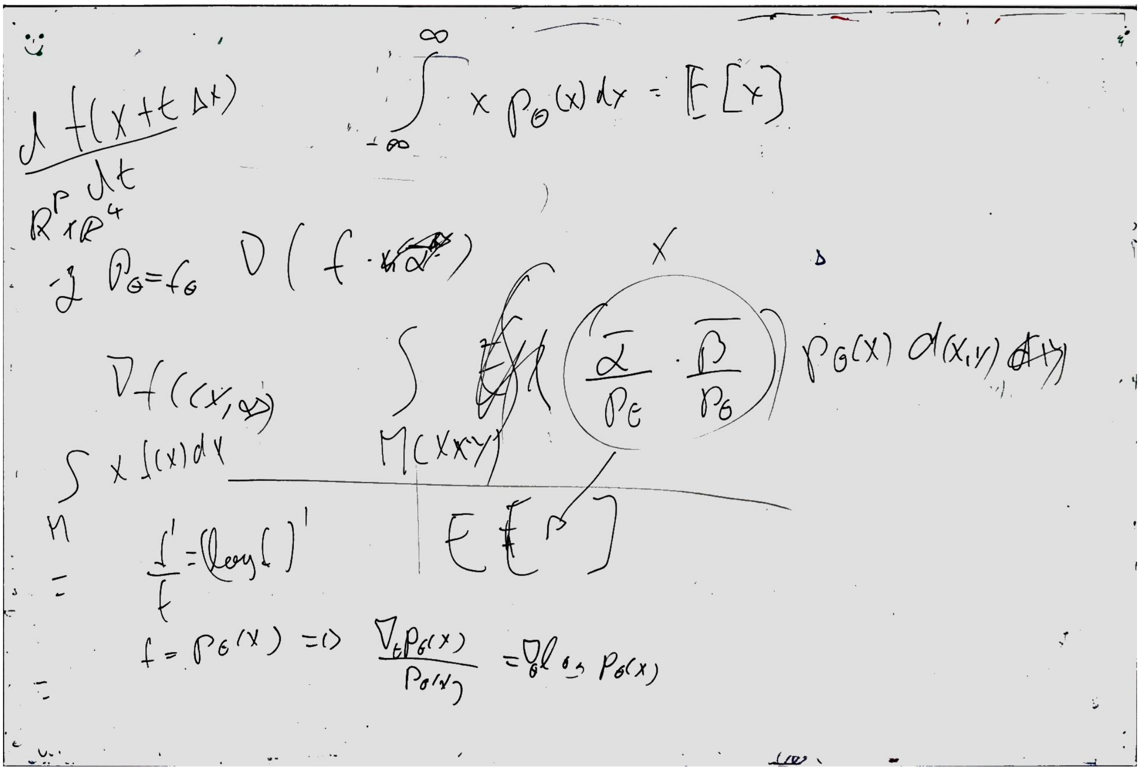
$$\langle \nabla_\theta p_\theta(x), \alpha \rangle = \frac{d\, p_{\theta + t\alpha}}{dt} = \bar{\alpha}$$

$$\bar{\alpha} = \langle \nabla_\theta p_\theta(x), \alpha \rangle = (\nabla_\theta p_\theta(x))^T \alpha$$

$$\langle \bar{\alpha}, \bar{\beta} \rangle_{p_\theta} = E_{(x,y)\sim p_\theta}\left[\frac{\bar{\alpha}}{p_\theta} \frac{\bar{\beta}}{p_\theta} \right] = E_{(x,y)\sim p_\theta}\left[\overbrace{\underbrace{\frac{(\nabla_\theta p_\theta(x))^T}{p_\theta(x)}}_{\nabla_\theta \log p_\theta(x)} \alpha}^{1\times 1}\ \overbrace{\frac{(\nabla_\theta p_\theta(x))^T \beta}{p_\theta(x)}}^{1\times 1} \right]$$

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

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


$\int_{-\infty}^{\infty} x\, p_\theta(x)\, dx = \mathbb{E}[x]$
$\frac{d\, f(x + t\Delta x)}{dt}$
$\frac{f'}{f} = (\log f)'$
$f = p_\theta(x) \Rightarrow \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} = \nabla_\theta \log p_\theta(x)$