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

Sandro Braun, Leander Kurscheidt

July 17, 2018

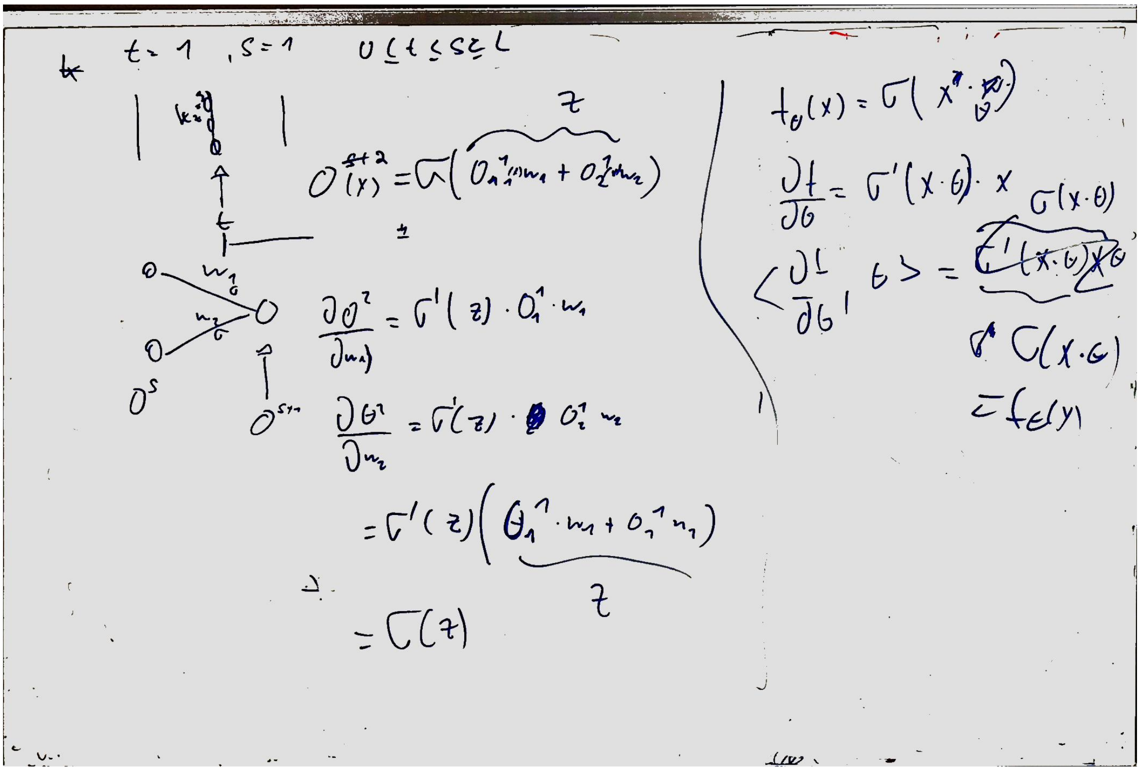
Abstract

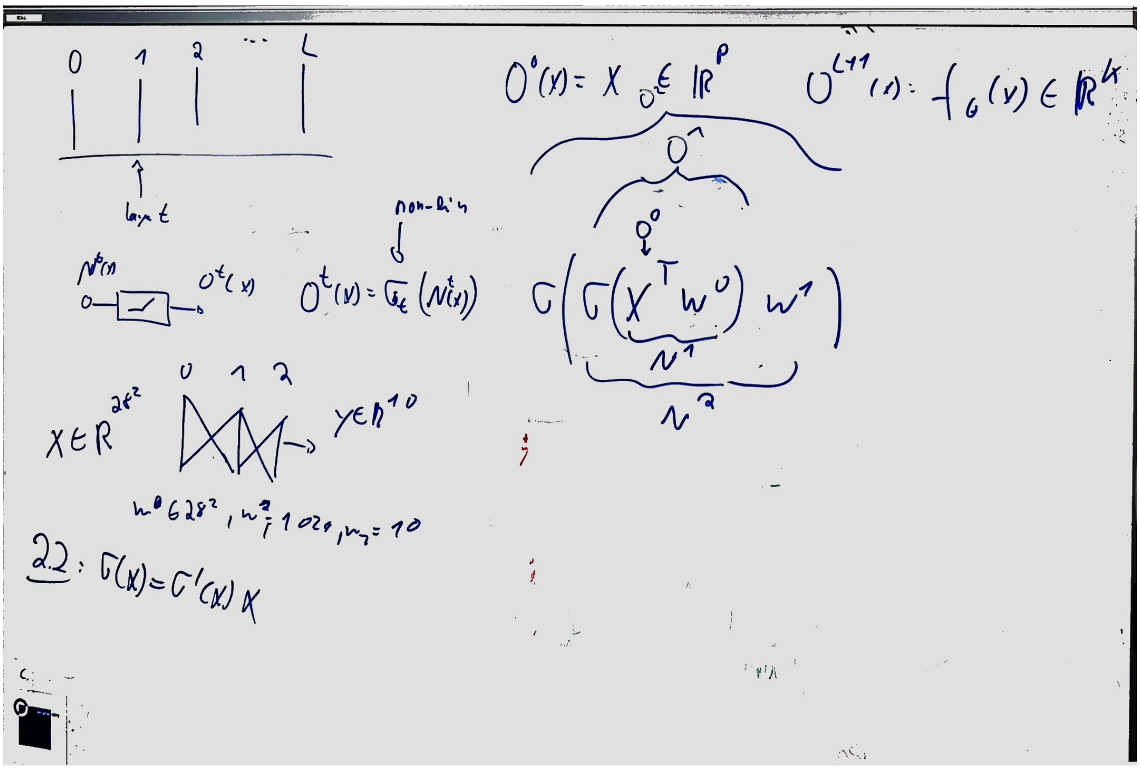
1 Geometry of Deep Rectified Networks

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   0 ≤ t ≤ s ≤ L
O^{t+1}(x) = σ(O_1^1 w_1 + O_2^1 w_2)
z
w_1
w_2
O^s
O^{s+1}
∂O^2/∂w_1 = σ'(z) · O_1^1 · w_1
∂O^2/∂w_2 = σ'(z) · O_2^1 w_2
= σ'(z)(O_1^1 · w_1 + O_2^1 w_2)
z
= σ(z)
f_θ(x) = σ(x^T · θ)
∂f/∂θ = σ'(x · θ) · x
σ(x · θ)
⟨∂f/∂θ, θ⟩ =
σ(x · θ)
= f_θ(x)


0 1 2 ... L
layer t
O^0(x) = x ∈ ℝ^p
O^{L+1}(x) = f_θ(x) ∈ ℝ^k
O^1
O^0
N^t(x)
O^t(x)
non-lin
O^t(x) = σ_t(N^t(x))
σ(σ(x^T w^0) w^1)
N^1
N^2
0 1 2
x ∈ ℝ^{28^2}
y ∈ ℝ^{10}
w^0 628^2, w^1 1024, w_2 = 10
2.2: σ(x) = σ'(x) x

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

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

$\Updownarrow$

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

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

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

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

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

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

P.5 | Proof of Corollar 2.1

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

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

$\Updownarrow$

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

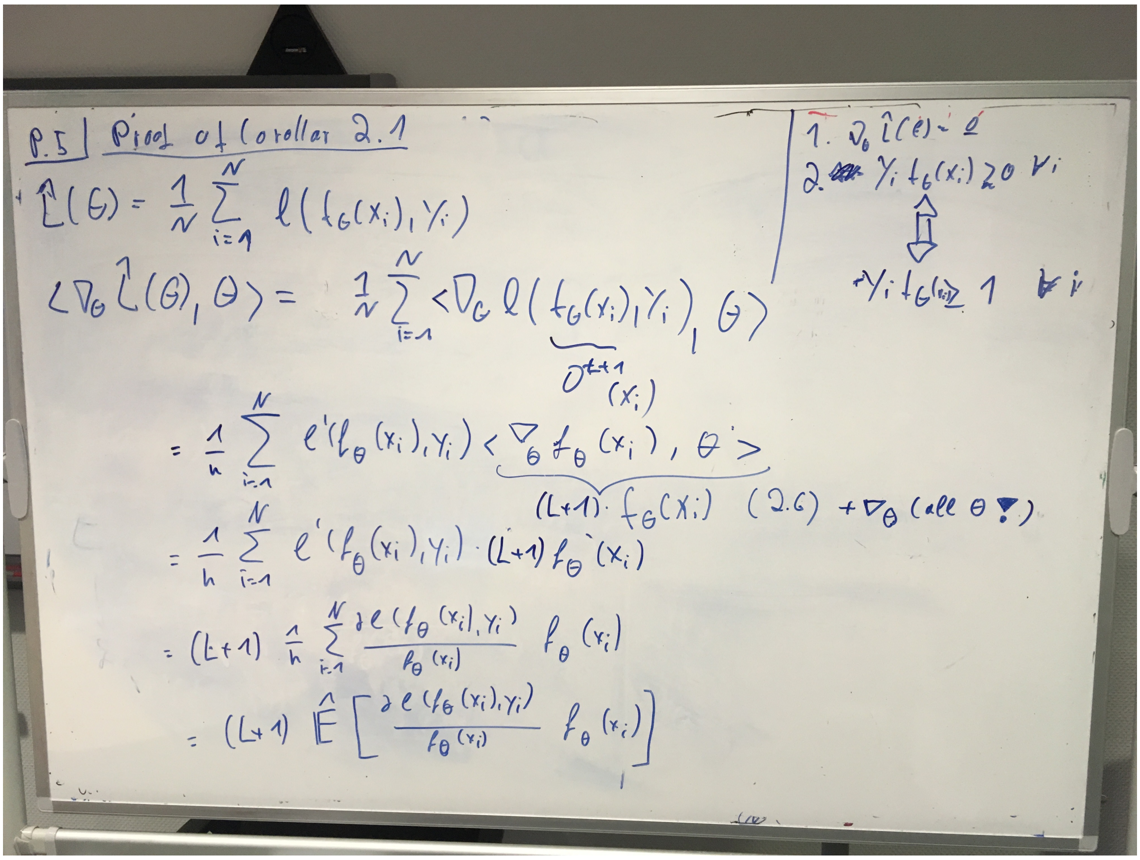
$$\langle \nabla_\theta \hat{L}(\theta), \theta \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle \nabla_\theta \ell(\underbrace{f_\theta(x_i)}_{o^{L+1}(x_i)}, y_i), \theta \rangle$$

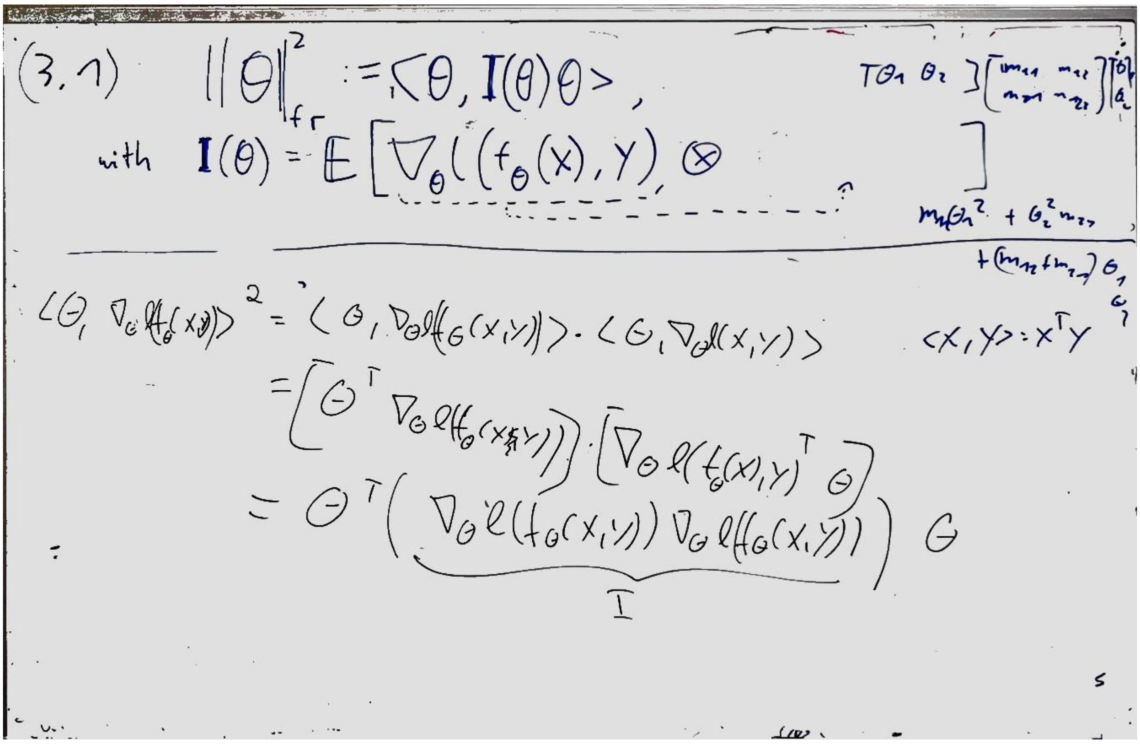
$$= \frac{1}{n} \sum_{i=1}^{N} \ell'(f_\theta(x_i), y_i) \underbrace{\langle \nabla_\theta f_\theta(x_i), \theta \rangle}_{(L+1) \cdot f_\theta(x_i)} \quad (2.6) + \nabla_\theta \text{ (alle } \theta \text{ ?)}$$

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

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

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


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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

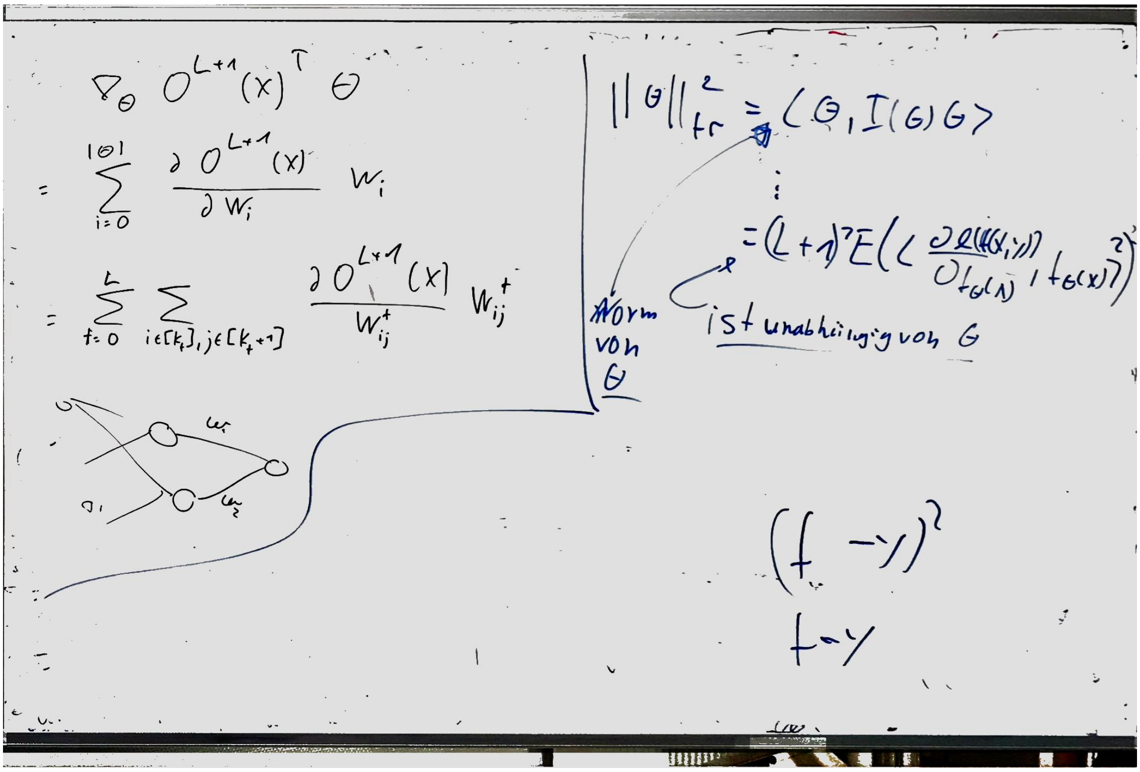
$l = \frac{(\ )^2}{2} \to \frac{\partial l}{\partial f} = |\ \ | \Rightarrow E\left[\langle f_\theta(x) - y, f_\theta(x) \rangle^2\right]$

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

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

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

$(\nabla\theta)^T G$


$\nabla_\Theta O^{L+1}(x)^T \Theta$
$= \sum_{i=0}^{|\Theta|} \frac{\partial O^{L+1}(x)}{\partial w_i} w_i$
$= \sum_{t=0}^{L} \sum_{i\in[k_t], j\in[k_{t+1}]} \frac{\partial O^{L+1}(x)}{w_{ij}^t} w_{ij}^t$
w_1
w_2
$\|\Theta\|_{fr}^2 = \langle \Theta, I(\Theta)\Theta \rangle$
$\vdots$
$= (L+1)^2 E(\langle \frac{\partial l(f(x),y)}{\partial f_\Theta(x)}, f_\Theta(x)\rangle^2)$
Norm von Θ
ist unabhängig von Θ
$(f-y)^2$
$f-y$