

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

Sandro Braun, Leander Kurscheidt

August 22, 2018

Abstract

1 Geometry of Deep Rectified Networks

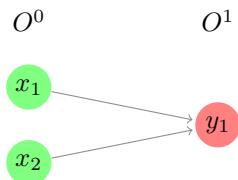
1.1 Lemma 2.1

Lemma 1.1 (Structure in Gradients).

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

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:

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

$$O^1 = \sigma(O^{0T} W^0) = \sigma(\underbrace{O_1^0 W_1^0 + O_2^0 W_2^0}_z) \quad (3)$$

Substituting the term in brackets will simplify notation.

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

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

Summing up the $\frac{\partial O^1}{\partial W_j^0} W_j$ reveals,

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

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

$$\begin{aligned}\nabla_{W^0} O^1 &= \begin{bmatrix} \frac{\partial O^1}{\partial W_1^0} \\ \frac{\partial O^1}{\partial W_2^0} \end{bmatrix} \\ \langle \nabla_{W^0} O^1, W^0 \rangle &= \sum_{j=1}^2 \frac{\partial O^1}{\partial W_j^0} W_j^0 = \dots = \sigma'(z)z\end{aligned}$$

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:

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

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

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

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

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

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

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

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

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

$$(12)$$

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

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

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

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

$$= N^1 D^1(x) \quad (16)$$

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

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

$$= x^T W^0 D^1(x) \quad (18)$$

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,

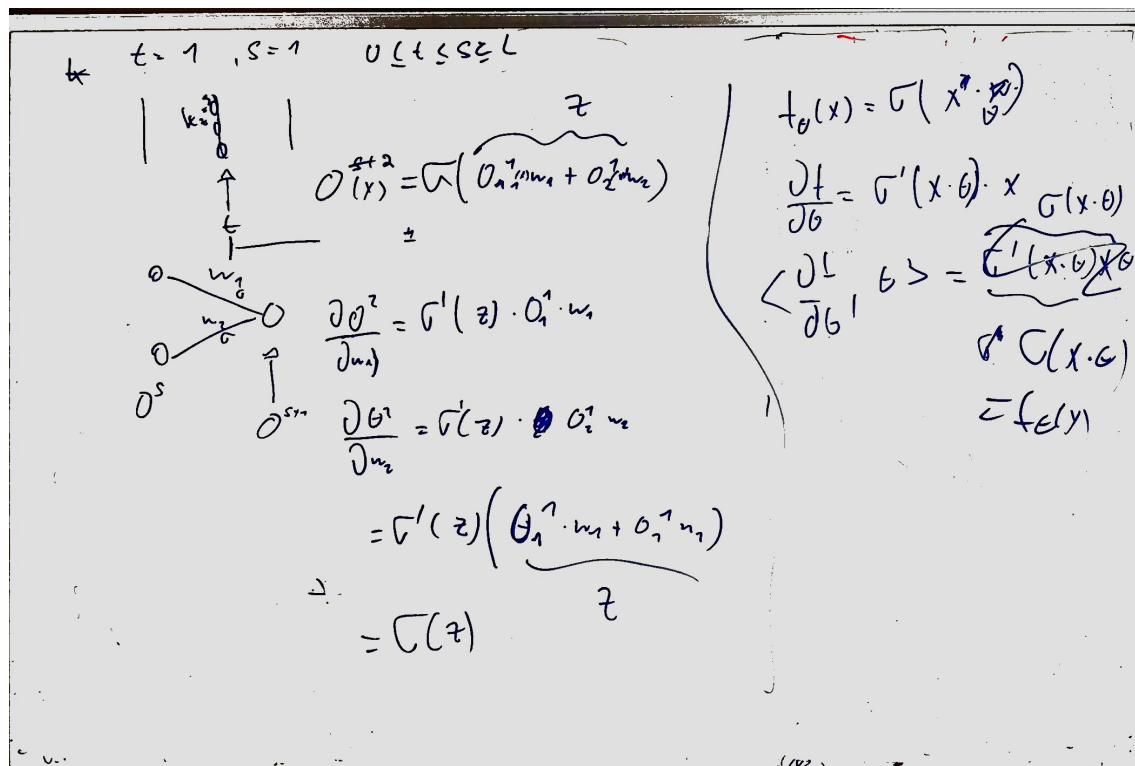
$$\begin{aligned} 1 - y_i f_i &> 0 \\ \Leftrightarrow l &= 1 - y_i f_i \\ \Leftrightarrow \frac{\partial l}{\partial f} &= -y_i \end{aligned}$$

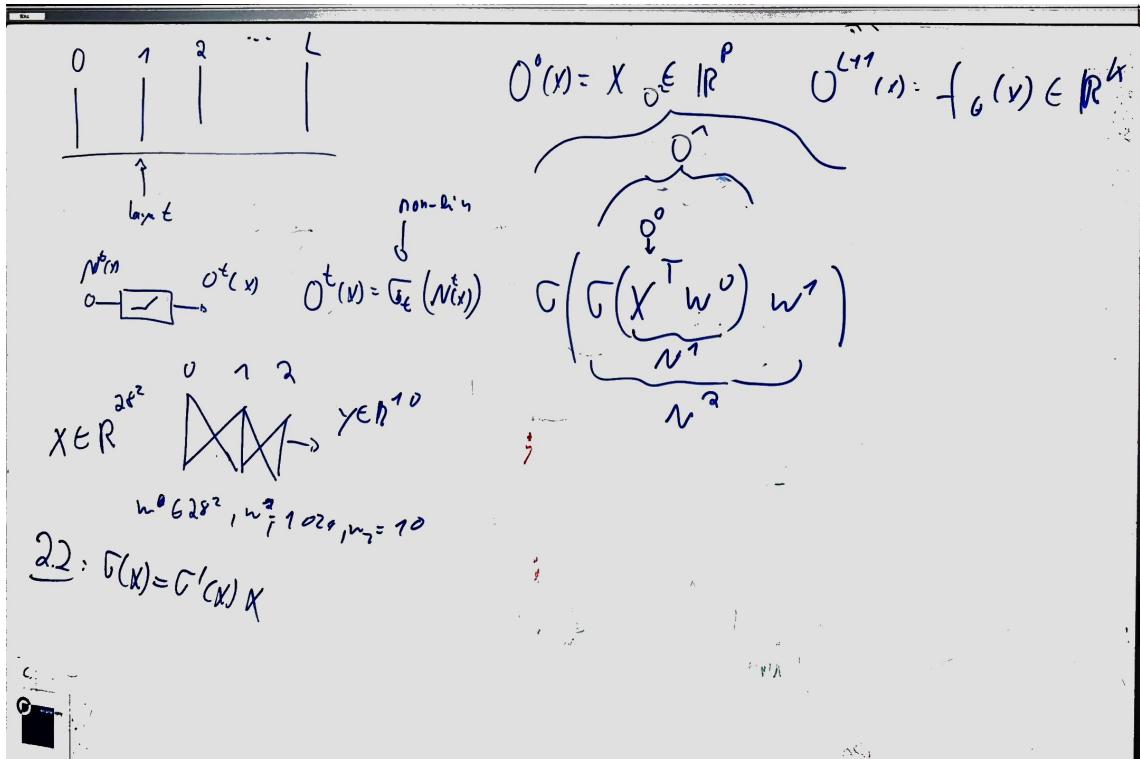
□

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

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

□





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

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

$\rightarrow 1 - y_i f_i > 0 \Rightarrow \ell = 1 - y_i f_i$
 $\rightarrow \frac{\partial \ell}{\partial f} = -y_i \text{ q.e.d.}$

$2) yf \geq 1 \Rightarrow 1 - y_i f_i \leq 0 \Rightarrow \ell = 0$
 $\frac{\partial \ell}{\partial f} = 0 \text{ q.e.d.}$

1. $\nabla_\theta \ell(\theta) = 0$
 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

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

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

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

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

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

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

1. $\nabla_{\theta} l(c) = 0$
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

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

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

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

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

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

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

1. $\nabla_{\theta} l(c) = 0$
2. $y_i f_\theta(x_i) \geq 0 \forall i$

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

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

with $I(\theta) = E \left[\nabla_\theta (\langle f_\theta(x), y \rangle, \otimes) \right]$

$$T\theta_1 \theta_2 \left[\begin{smallmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{smallmatrix} \right] \theta_1$$

$$m_1 G_1^2 + G_2^2 m_2,$$

$$+ (m_{12} + m_{21}) G_1 G_2$$

$$\begin{aligned} \langle \theta, \nabla_\theta l(f_\theta(x)) \rangle^2 &= \langle \theta, \nabla_\theta l(f_\theta(x), y) \rangle \cdot \langle \theta, \nabla_\theta l(x, y) \rangle \quad \langle x, y \rangle = x^T y \\ &= \left[\theta^T \nabla_\theta l(f_\theta(x), y) \right] \cdot \left[\nabla_\theta l(f_\theta(x), y)^T \theta \right] \\ &= \theta^T \underbrace{\left(\nabla_\theta l(f_\theta(x), y) \nabla_\theta l(f_\theta(x), y)^T \theta \right)}_I \theta \end{aligned}$$

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

with $I(\theta) = E \left[\nabla_\theta (\langle f_\theta(x), y \rangle, \otimes) \right]$

$$T\theta_1 \theta_2 \left[\begin{smallmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{smallmatrix} \right] \theta_1$$

$$m_1 G_1^2 + G_2^2 m_2,$$

$$+ (m_{12} + m_{21}) G_1 G_2$$

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

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

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

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

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

$$\langle x, x' \rangle = x^T (x')$$

$$= \langle \nabla_\theta f_\theta(x), \theta \rangle$$

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

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

$$\text{Es war: } \langle \nabla_{\theta} \mathcal{L}_{\theta}(x), \theta \rangle = (l+1) \mathcal{O}^{L+1}(x) \widehat{= (l+1)} f_{\theta}(x)$$

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

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

$$l = \frac{1}{2} - \frac{\partial \ell}{\partial l} = 1 \quad \Rightarrow E \left[\langle f_{\theta}(x) - y, f_{\theta}(x) \rangle^2 \right]$$

$$\nabla_{\theta} \mathcal{L}_{\theta} = \begin{pmatrix} \frac{\partial \ell(f_{\theta}(x), y)}{\partial w_1} \\ \vdots \\ \frac{\partial \ell(f_{\theta}(x), y)}{\partial w_n} \end{pmatrix}$$

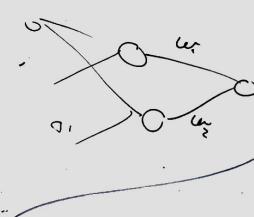
$$f_{\theta}(x) - y$$

$$(\nabla_{\theta} \mathcal{L}_{\theta})^T G \nabla_{\theta} \mathcal{L}_{\theta}$$

$$\begin{aligned} & \nabla_{\theta} \mathcal{L}_{\theta}^T \theta \\ &= \sum_{i=0}^{l+1} \frac{\partial \mathcal{L}_{\theta}^T}{\partial w_i} w_i \\ &= \sum_{t=0}^L \sum_{i \in [k_t], j \in [k_{t+1}]} \frac{\partial \mathcal{L}_{\theta}^T}{\partial w_{ij}} w_{ij} \end{aligned}$$

$$\|\theta\|_F^2 = \langle \theta, I(G) \theta \rangle$$

$$\begin{aligned} &= (l+1)^2 E \left(\left(\frac{\partial \ell(f_{\theta}(x), y)}{\partial f_{\theta}(x)}, f_{\theta}(x) \right)^2 \right) \\ &\text{Norm von } G \\ &\text{ist unabhängig von } G \end{aligned}$$



$$\begin{aligned} & (f, -y)^2 \\ & f = y \end{aligned}$$

$\sigma(z) = \sigma'(z) z$ $z = \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}$ $\rightarrow P_{100} \text{ of } 3.4$

$O^2 = \begin{bmatrix} O_1^2 \\ O_2^2 \end{bmatrix} = \begin{bmatrix} G(n_1) \\ G'(n_2)n_2 \end{bmatrix} = \begin{bmatrix} N_1 & N_2 \end{bmatrix} \begin{bmatrix} O_1^2 \\ G'(N_2)N_2 \end{bmatrix}$

$\text{zurgehen: } O^2 = N^2 \text{ diag}(G'(N^2))$

$O^3 = \begin{bmatrix} O_1^3 \\ O_2^3 \end{bmatrix} = \begin{bmatrix} G(O_1^2) \\ G'(O_2^2)O_2^2 \end{bmatrix} = \begin{bmatrix} N_1 & N_2 \end{bmatrix} \begin{bmatrix} O_1^3 \\ G'(O_2^2)O_2^2 \end{bmatrix}$

$G'(N^2) = \begin{bmatrix} G'(N_1) \\ G'(N_2) \end{bmatrix}$
 Q.E.d.

$y_i \in \mathbb{R}^p$ $\rightarrow P_{100} \text{ of } 3.5$

$X^T = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$ Note: we have scalar output

$X^T V \in \mathbb{R}^{n \times 1} \quad \begin{bmatrix} (X^T V)^T \\ X_1^T V \\ \vdots \\ X_n^T V \end{bmatrix} = \begin{bmatrix} n \end{bmatrix}$

$\Rightarrow V^T X = \boxed{n} \Rightarrow V^T X X^T V = \boxed{n} = \boxed{1 \times 1}$

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

[Proof of 3.6]

$$f = P_{\theta+\epsilon\alpha} \Big|_{t=0}, \quad (\cdot)^b = \frac{d}{dt} \quad \text{as diff: } \frac{dP_{\theta+\epsilon\alpha}}{dt} \Big|_{t=0} = \alpha$$

$$\Rightarrow \frac{dP_{\theta+\epsilon\alpha}}{dt} \Big|_{t=0} = \frac{\nu + \alpha}{P_\theta} \Big|_{t=0} = \frac{f'}{f} \Big|_{t=0} = \frac{d(\log P_{\theta+\epsilon\alpha})}{dt} \Big|_{t=0}$$

$$\Rightarrow \langle \bar{\alpha}, \bar{\beta} \rangle_{P_\theta} = \int \frac{\bar{\alpha}}{P_\theta} \frac{\bar{\beta}}{P_\theta} P_\theta = \int \frac{d}{dt} \log P_{\theta+t\alpha} \Big|_{t=0} \frac{d}{dt} \log P_{\theta+t\beta} \Big|_{t=0}$$

Erwartungswert = $E \left\{ \log P_{\theta+t\alpha} \log P_{\theta+t\beta} \right\}_{P_\theta}$

$$G(G(x^\top w^*) w_2)$$

$$(l+\gamma)^2 E \left[\left\langle \frac{\partial f}{\partial t_0} \Big|_{t=0}, f_0 \right\rangle^2 \right]$$

$$f_0 = \langle v^\top x \rangle - \alpha \frac{\partial f}{\partial t_0} \Big|_{t=0} \quad x \in \mathbb{R}^n, v \in \mathbb{R}^n$$

$$= (l+\gamma)^2 E[f_0^2]$$

$$f_0 = x^\top v^* \sigma(X) \dots$$

$$= (x^\top v)^2 V(x, \sigma)$$

$$G'(z) = G(z)$$

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

$$(X^\top V)^2 = \underbrace{(x^\top v)}_{\sigma(X)} \underbrace{(x^\top v)}_{\sigma(X)}$$

$$= (v^\top x) / (\sigma(X))$$

Richtungsableitung

$$f(x, y | \theta) = P_\theta(x, y)$$

[3.6]

$$\langle \nabla_{\theta} P_\theta(x), \alpha \rangle = \frac{d P_{\theta+\epsilon \alpha}}{d \epsilon} \Big|_{\epsilon=0} = \alpha^T \nabla_{\theta} P_\theta(x) = (\nabla_{\theta} P_\theta(x))^T \alpha$$

$$\begin{aligned} \langle \bar{\alpha}, \bar{\beta} \rangle &= E_{(x, y) \sim P_\theta} \left[\frac{\bar{\alpha}^T \bar{\beta}}{P_\theta} \right] = E_{(x, y) \sim P_\theta} \left[\frac{(\nabla_{\theta} P_\theta(x))^T \alpha}{P_\theta(x)} \cdot \frac{(\nabla_{\theta} P_\theta(x))^T \beta}{P_\theta(x)} \right] \\ &= E_{(x, y) \sim P_\theta} \left[\alpha^T (\nabla_{\theta} \log P_\theta(x)) (\nabla_{\theta} \log P_\theta(x))^T \beta \right] \end{aligned}$$

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

$$\int f(x + t \Delta t) x P_\theta(x) dx = E[x]$$

$$\text{if } P_\theta = f_\theta \quad D(f \cdot \bar{\alpha} \bar{\beta})$$

$$\begin{aligned} &\int x f(x, y) \frac{d(x) dy}{M(x, y)} = E[\bar{x}] \\ &= \frac{\int f'(x) dx}{E[\bar{x}]} \end{aligned}$$

$$f = P_\theta(x) \Rightarrow \frac{\nabla_x P_\theta(x)}{P_\theta(x)} = \nabla_x \log P_\theta(x)$$

Def 3.)

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

$\underbrace{\quad}_{0 \leq D^t \leq \infty} \quad \underbrace{\quad}_{D^t = 0} \quad (1)$

$$\begin{aligned} & \hookrightarrow \|D^+(x)\|_{\sigma} \in \{0, \infty\} \quad \Leftrightarrow \quad D^t = 0 \\ & \stackrel{(1)}{=} \sqrt{\mathbb{E} \left(\|X\|^2 \prod_{t=0}^L \|D^t(x)\|_{\sigma}^2 \prod_{t=0}^L \|W^t\|_{\sigma}^2 \right)} \\ & \text{and } (1) = \sqrt{\mathbb{E} \left(\|X\|^2 \prod_{t=0}^L \|D^{t+1}(x) W^t\|_{\sigma}^2 \right)} \\ & = \sqrt{\mathbb{E} \left(\|X\|^2 \prod_{t=0}^L \|W^t \text{diag} [\sigma'(N^{t+1}(x))]^{-1}\|_{\sigma}^2 \right)} \\ & \stackrel{(2)}{=} \sqrt{\mathbb{E} \left(\|X^T\|_{\sigma}^2 \prod_{t=0}^L \|W^t \text{diag} [\sigma'(N^{t+1}(x))]^{-1}\|_{\sigma}^2 \right)} \end{aligned}$$

$$D^+(x) = \text{diag} [\sigma'(N^t(x))]$$

$$\sigma(x) = \sigma'(x) x$$

! only 0 or 1 !

$$\|f\|_2^2 = \|f\|^2 \quad (2)$$

$$\|X M\|^2 \leq \|X\|^2 \|M\|_{\sigma}^2$$

$$\begin{aligned} & \geq \sqrt{\mathbb{E}[\|X^T\|_{\sigma}^2 \prod_{t=0}^L \|W^t D^{t+1}(x)\|_{\sigma}^2]} \\ & \geq \sqrt{\mathbb{E}[\|X^T \prod_{t=0}^L W^t D^{t+1}(x)\|_{\sigma}^2]} \\ & = \sqrt{\mathbb{E}[\|h_{\theta}(x)\|_{\sigma}^2]} \end{aligned}$$

Lemma 4.1 (mit Wikipedia)

$$\left(\begin{array}{c} \vdots \\ \vdots \end{array} \right) < \left(\begin{array}{c} \vdots \\ \vdots \end{array} \right)$$

$$\frac{1}{L+1} \|\theta\|_{fr} \leq E\|w\| \leq 1$$

$$\leq 1 \quad \max_{\zeta} \frac{\|A\zeta\|^2}{\|\zeta\|^2}$$

$$\|\cdot\|_2^2 \geq \|\cdot\|_C^2$$

$$\frac{1}{L+1} \|\theta\|_{fr} = E\left\{ v^T X X^T v \right\} = E\{\|h_G\|^2\} = E\left\{ \|X \Phi \Pi D \Pi w\|_2^2 \right\}$$

$$\geq E\left\{ \|X \Pi_D \Pi w\|_G^2 \right\}$$

2 Notes on Section 4

2.0.1 Explaining Lemma 4.1

A proof of Lemma 4.1 can also be found in the appendix of the original paper. However, the relation between the norms is missing in the original paper - that is why we state it here again. From 3.5 in the paper, we had:

$$\frac{1}{(L+1)^2} \|\theta\|_{fr}^2 = E\{v^T X X^T v\} = E\{\|f_\theta\|^2\}. \quad (19)$$

Remember that the output of the network, $f_\theta(x)$, is a vector in \mathbb{R}^k of probabilities for each of the k classes. For the Frobenius Norm of a Matrix A and a vector x the following holds.

$$\|A\|_F \geq \|A\|_\sigma \quad (20)$$

$$\|x\|_F = \|x\|_2 = \|x\|_\sigma \quad (21)$$

$$\|Ax\|_\sigma \leq \|A\|_\sigma \cdot \|x\|_2 \quad (22)$$

$$\|f_\theta\|_\sigma^2 \leq \|X\|_\sigma^2 \cdot \|v\|_2^2 \quad (23)$$

It follows.

$$E\{\|f_\theta\|_\sigma^2\} = E\{\|f_\theta\|_\sigma^2\} \quad (24)$$

$$= E\left\{\|x^T W^0 D^1(x) W^1 D^2(x) \dots D^L W^L D^{L+1}(x)\|_\sigma^2\right\} \quad (25)$$

$$\leq E\left\{\|x\|_\sigma^2 \prod \|D^i(x)\|_\sigma^2 \prod \|W^i\|_\sigma^2\right\}. \quad (26)$$

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

$$\frac{1}{(L+1)^2} \|\theta\|_{fr}^2 = E\{\|f_\theta\|_\sigma^2\} \quad (27)$$

$$\leq E\left\{\|x\|_\sigma^2 \prod \|D^i(x)\|_\sigma^2 \prod \|W^i\|_\sigma^2\right\} \quad (28)$$

$$= E\left\{\|x\|_2^2 \prod_{t=1}^{L+1} \|D^t(x)\|_\sigma^2\right\} \prod_{t=0}^L \|W^t\|_\sigma^2 \quad (29)$$

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

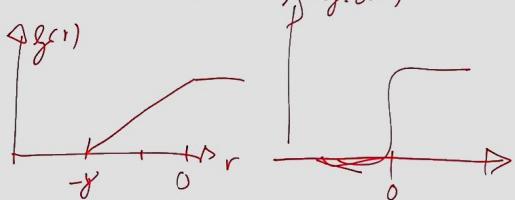
$$\text{margin} = f(x)_Y - \max_{i \neq Y} f(x)_i$$

$$W = \max \{d_1, \dots, d_L\}, \quad d_i = \text{input}$$

$N(V, \varepsilon, \| \cdot \|)$ = least card. of any subset $V \subseteq U$
that covers V with margin ε and
norm $\| \cdot \|$

$$R_f = \mathbb{E}(\ell_f(-M(f(x), y)))$$

$$\ell_f(r) = \begin{cases} 0 & r < -\gamma \\ 1-\gamma & r \in [-\gamma, 0] \\ 1 & r > 0 \end{cases} \quad \begin{matrix} \text{"correct"} \\ \text{"no decision margin"} \\ \text{"incorrect"} \end{matrix}$$



$$M(x_i, y) = y - \max_{i \neq y} f(x_i) \quad (\text{margin}) \quad f(x_i) = \max_{i \neq y} f(x_i)$$

$$F_\delta = \left\{ (x, y) \rightarrow \ell_y(-M(f(x), y)) : f \in \mathcal{F} \right\} \quad \begin{matrix} \text{"input to error with margins"} \\ \text{"mappings"} \end{matrix}$$

$$\mathcal{F} \subseteq \{f : f : \mathbb{R}^d \rightarrow \mathbb{R}^k\}$$

correct with $1 - \delta$

$$P[\text{anyway, } f(x)_i + y] \leq \hat{R}_f(f) + 2R_a(H_{1S}) + 3\sqrt{\frac{\ln(1/\delta)}{2n}}$$

\uparrow
false classification
(in test)
 \uparrow
exp. "fraction of
error with margin"

$$R_a(H_{1S}) = \mathbb{E} \sup_{h \in H} \sum_{i=1}^n \epsilon_i h(x_i, y_i) \quad S = ((x_1, y_1), \dots, (x_n, y_n))$$

$\mathcal{H}_X = \left\{ F_A(X^T) : A = (A_1, \dots, A_L), \|A_i\|_\sigma \leq s_i, \|A_i^T - M_i^T\|_{2,n} \leq b_i \right\}$

max dim cat axis = n
 every layer is bound by some spectral norm
 distance to ref. matrix is $\leq b_i$

NN F_A applied to X^T

3.2 Step III

$\ln \mathcal{N}(\mathcal{H}_X, \varepsilon, \|\cdot\|_2) \leq \frac{\|X\|_2^2 \ln(2W)}{\varepsilon^2} \left(\prod_{j=1}^L s_j p_j^2 \right) \left(\sum_{i=1}^L \left(\frac{b_i}{s_i} \right)^{\frac{2}{3}} \right)^3$

$\Rightarrow \begin{bmatrix} T & \frac{1}{\varepsilon} \end{bmatrix}_{\alpha}^{\sqrt{n}} = -\frac{T}{\frac{1}{\varepsilon}} + \frac{T}{\alpha}$

$R((F_x)_S) \leq \inf_{\alpha > 0} \left(\frac{6\alpha}{r_n} + \ln(r_n/\alpha) \frac{12\sqrt{R}}{n} \right)$

$\alpha = \frac{3\sqrt{R}}{r_n} \frac{12\sqrt{R}}{n} + \ln\left(\frac{r_n \cdot v_n}{3\sqrt{R}}\right) \frac{12\sqrt{R}}{n}$

$= \frac{12\sqrt{R}}{n} \left(1 + \ln\left(\frac{n}{3\sqrt{R}}\right) \right)$

$= \frac{12\sqrt{R}}{n} \left(R_n e + \ln\left(\frac{n}{3\sqrt{R}}\right) \right) = \frac{12\sqrt{R}}{n} \ln\left(\frac{e R_n}{3\sqrt{R}}\right) = \frac{\sqrt{R}}{n} \ln\left(\frac{e R^{1/n}}{3}\right)$

also move stuff into ln

References