

Idea: Adaption of theory to networks with biases

Leander Kurscheidt

August 3, 2018

Abstract

1 Affine Transformations

Since our NN "somewhat" resembles linear functions following from the rule $\sigma(x) = \sigma'(x)x$ (i see this rule as separating the linear from the non-linear part), i am curious to see whether ideas of affine spaces and how to integrate them into linear-spaces translate.

I think they do.

2 Idea

Affine subspaces are usually define via $A = v + U_V$, where U_V is a subspace of V and v is a vector of V (see [\(german\) wikipedia](#)). But you can work in them using [homogenous coordinates](#).

Every affine transformation can be turned into a linear transformation with a constant 1-input using the [Augmented Matrix](#) trick.

For example the intercept in linear models in statistics is often modeled this way.

We can adapt this trick with only one additional constant input using this matrix:

$$W' = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} & bias_1 \\ w_{21} & w_{22} & w_{23} & \dots & w_{2n} & bias_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_{d1} & w_{d2} & w_{d3} & \dots & w_{dn} & bias_n \\ 0 & \dots & \dots & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} W & b \\ 0 & \dots & 0 & 1 \end{bmatrix}$$

the last line is the difference to the usual construction and should be omitted in the last layer.

This results in: $f'_\theta(x) = f_{\theta'}\left(\begin{bmatrix} x \\ 1 \end{bmatrix}\right)^1$,

where θ' is the parameter adapted in the above schema.

Usually proofs get a lot simple when you just have to worry about keeping one input constant. I think a lot of the proofs from the paper should still hold, for example Lemma 2.1 doesn't change. I hope the other proofs only need some tweaking here and there.

For this to work, we need to add the additional requirement that $\sigma(1) = 1$

¹slight abuse of notation. We define $\begin{bmatrix} x \\ 1 \end{bmatrix} := \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ 1 \end{bmatrix}$

3 Proof

In the following section, neural networks with biases as defined as:

$$f_\theta(x) = \sigma_{L+1}(\sigma_L(\dots \sigma_2(\sigma_1(xW^0 + b^0)W^1 + b^1)W^2 + b^2) \dots W^L + b^L)$$

We call b^i bias and additionally need the following constraint on the activation function σ_i :

$$\forall i \in L+1, \dots, 1. \exists c_i. \sigma_i(c_i) = 1$$

For f_θ with $\theta = (W^0, W^1, \dots, W^L)$ we define:

$$f_{0,\theta^0}(z) := \sigma_1(zW'^0), \text{ where } \theta^0 := (W'^0)$$

$$f_{L,\theta^L}, \text{ so that}$$

$$f_\theta = f_{L,\theta^L} \circ f_{0,\theta^0} \text{ and}$$

$$\theta^L := (W^1, \dots, W^L)$$

Definition 3.1. Homogenous Coordinates for Neural Networks

Let f_θ be a neural network with biases. Then augmented neural network $f_{\theta'}$ is a neural network as defined in Definition 1, (2.1) in [TODO: ref], where:

$$\theta' = (W'^0, \dots, W'^L)$$

and

$$W'^i = \begin{bmatrix} w_{11}^i & w_{12}^i & w_{13}^i & \dots & w_{1k_i}^i & b_1^i \\ w_{21}^i & w_{22}^i & w_{23}^i & \dots & w_{2k_i}^i & b_2^i \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_{k_{i+1}1}^i & w_{k_{i+1}2}^i & w_{k_{i+1}3}^i & \dots & w_{k_{i+1}k_i}^i & b_{k_{i+1}}^i \\ 0 & \dots & \dots & \dots & 0 & c_i \end{bmatrix} =: \begin{bmatrix} W^i & & & b^i \\ 0 & \dots & 0 & c_i \end{bmatrix} \in \mathbb{R}^{k_i+1, k_{i+1}+1} \forall i \in \{0, \dots, (L-1)\}$$

$$W'^i = \begin{bmatrix} w_{11}^i & w_{12}^i & w_{13}^i & \dots & w_{1k_i}^i & b_1^i \\ w_{21}^i & w_{22}^i & w_{23}^i & \dots & w_{2k_i}^i & b_2^i \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_{k_{i+1}1}^i & w_{k_{i+1}2}^i & w_{k_{i+1}3}^i & \dots & w_{k_{i+1}k_i}^i & b_{k_{i+1}}^i \end{bmatrix} =: [W^i \quad b^i] \in \mathbb{R}^{k_i+1, K} \text{ with } i = L$$

Theorem 3.1. Equality of the augmented NN

For all NN with Bias f_θ :

$$f'_\theta(x) = f_{\theta'}\left(\begin{bmatrix} x \\ 1 \end{bmatrix}\right)$$

Proof. via Induction over L.

1. For $L = 0$:

$$\Theta' = (W'^0) \tag{1}$$

$$W'^0 = [W^0 \quad b^0] \text{ because } 0 = L \tag{2}$$

$$f_{\theta'}\left(\begin{bmatrix} x \\ 1 \end{bmatrix}\right) = \sigma_1\left(\begin{bmatrix} x \\ 1 \end{bmatrix} W'^0\right) \tag{3}$$

$$= \sigma_1\left(\begin{bmatrix} x \\ 1 \end{bmatrix} [W^0 \quad b^0]\right) \tag{4}$$

$$= \sigma_1(xW^0 + b^0) \tag{5}$$

$$= f_\theta(x) \tag{6}$$

2. For $L \rightarrow (L+1)$ with $f_{L,\theta'} \rightarrow f_{L,\theta'} \circ f_{0,\theta'^0}$ holds Theorem 3.1

3. $L \rightarrow (L + 1)$:

$$\theta' = (W'^0, W'^1, \dots, W'^{(L+1)}) \quad (7)$$

$$W'^0 = \begin{bmatrix} W^0 & & b^0 \\ 0 & \dots & 0 & c_0 \end{bmatrix} \text{ because } 0 \neq (L + 1) \quad (8)$$

$$(9)$$

$$f_{0,\theta^0}\left(\begin{bmatrix} x \\ 1 \end{bmatrix}\right) = \sigma_1\left(\begin{bmatrix} x \\ 1 \end{bmatrix} W'^0\right) \quad (10)$$

$$= \begin{bmatrix} \sigma_1\left(\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} W^0 & b^0 \end{bmatrix}\right) \\ \sigma_1(1 * c_0) \end{bmatrix} \quad (11)$$

$$= \begin{bmatrix} f_{0,\theta^0}(x) \\ 1 \end{bmatrix} \quad (12)$$

$$f_{\theta'}\left(\begin{bmatrix} x \\ 1 \end{bmatrix}\right) = (f_{L,\theta'^L} \circ f_{0,\theta^0})\left(\begin{bmatrix} x \\ 1 \end{bmatrix}\right) \quad (13)$$

$$= f_{\theta'^L}\left(\begin{bmatrix} f_{\theta^0}(x) \\ 1 \end{bmatrix}\right) \quad \text{using (12)} \quad (14)$$

$$= (f_{L,\theta^L} \circ f_{0,\theta^0})(x) \quad \text{induction hypothesis} \quad (15)$$

$$= f_{\theta}(x) \quad (16)$$

□

Let $g(x) := \begin{bmatrix} x \\ 1 \end{bmatrix}$ (which is continous). Then, following Theorem 3.1:

For all NN with Bias $f_{\theta}:f'_{\theta} = f_{\theta'} \circ g$

Theorem 3.2. *Transformation of the Data-Distribution*

Let P_{data} be Data-Distribution, then $P'_{data} = g(P_{data})$ is the transformed Data-Distribution. Then:
 $E_{x \sim P_{data}} f_{\theta}(x) = E_{x \sim P'_{data}} f_{\theta'}(x)$

Proof.

$$E_{x \sim P_{data}} f_{\theta}(x) = E_{x \sim P_{data}} (f_{\theta'} \circ g)(x) \quad (17)$$

$$= E_{x \sim P'_{data}} f_{\theta'}(x) \quad (18)$$

(18) follows from *Law of the unconscious statistician*. □

The Rest of the proofs from the "Fisher-Rao Metric, Geometry ..." -paper should now follow from Theorem 3.1 and 3.2.

References