Notes on Analysis of Activation Functions

Ue-Yu Pen[*]
Chicago, IL
ueyupen@gmail.com

Vedant Sachdeva[†]
Chicago, IL 60615
sachdved@gmail.com

Abstract

Here, we explore the solutions to regression functions, with varying degrees of non-linearity and depth to explore how and when certain correlations are captured by MLPs.

1 Intro

Typically, multi-layer perceptron (MLP) networks are characterized by two parameters - depth and width. Depth refers to the number of layers in a network, while the width of a given layer typically refers to the number of nodes in said layer. Layers are parametrized by their weights, $W^{(j)}$, where j indicates that we are examining the jth layer. $W^{(j)}$ is typically drawn from $\mathbb{R}^{d^{(j)} \times d^{(j+1)}}$, where $d^{(j)}$ is the width of the jth layer. Previous works have shown that with non-linear activation functions and MLP architectures, a vast family of functions can be arbitrarily approximated. However, how such functions correlate to the moments of the data being leveraged is not clearly explained. This document will detail how non-linearities uncover the higher order correlations in data. For the purposes of this note, we will focus exclusively on mean-square-error objective functions.

2 Linear Activation Functions

In MLPs with linear activation functions, we observe that the output of the jth layer is simply given by

$$x^{(j+1)} = W^{(j)} x^{(j)} + b^{(j)}. \tag{1}$$

For simplicity, instead of including a bias term, we will append a node to each layer with value, 1, independent of input. This value will provide an effective bias to each layer. We can then express the output of the jth layer from the initial layer by writing:

$$x^{(j+1)} = \prod_{i=1}^{i=j} W^{(i)} x^{(0)} \tag{2}$$

Consequently, any network of arbitrary depth and width with linear activation functions can be re-written as a single layered MLP. Thus, when optimizing our objective function, we are simply optimizing:

$$\mathcal{L} = \min_{W} \sum_{i} ||y_i - W x_i||^2. \tag{3}$$

[*]No funding :(
[†]Also unfunded :(

Preprint. Under review.

We optimize this objective function by computing a derivative with respect to each element of W and equating to 0. This yields a set of simultaneous equations of the form:

$$\frac{\partial \mathcal{L}}{\partial W^{(k)}} = \sum_i (-2y_i x_i^{(k)} + \sum_{m \neq k} W^m x_i^m x_i^k + W^k x_i^k x_i^k) \tag{4}$$

Thus, the optimal weight vector can be expressed as:

$$W^k = \frac{\mathrm{Cov}(Y, X^k) - \sum_{m \neq k} \mathrm{Cov}(X^m, X^k)}{\mathrm{Cov}(X^k, X^k)} \tag{5}$$

As we can see, one outcome will be that if much of the covariance in Y with X_k can be explained by X_k covarying with another feature in X, the net weight will be constrained, but the individual weights can become unbounded. Further, such a model cannot consider further than second order moments in the dataset.

3 Quadratic Activation Functions with one layer

We will expand our activation function forward to consider quadratic terms. We are doing so with the intent to understand how the optimal solution for polynomial non-linearities involves inference of other statistical moments of the data. We will start here, before moving on to Taylor expansions of common activation functions in the neighborhoods of high curvature regions of the activation function. We will begin our analysis of quadratic activation functions on models with a one dimensional input and a one dimensional output. That simplifies to a model of the form

$$y = Wx + \frac{\alpha}{2}(W^2 x^2). \tag{6}$$

We have introduced a parameter, α, that serves as a weight on the quadratic term in our activation function.

Analytically solving for the optimal weight, as before, yields an equation of the form

$$\mathcal{L} = \min_W \sum_i ||y_i - Wx_i - \frac{\alpha}{2} W^2 x_i^2||. \tag{7}$$

$$\frac{d\mathcal{L}}{dW} = \sum_i (-2y_i x_i - 2\alpha W y_i x_i^2 + 2W x_i^2 + 3\alpha W^2 x_i^3 + \alpha^2 W^3 x_i^4) \tag{8}$$

The above is a cubic equation, and its roots can be expressed analytically in terms of the expectations of our random variables. Even without solving, we can immediately see that by including a quadratic term in the activation function, we obtain a dependency on the fourth order moment with only one parameter to fit.

We note that this is in contrast to if we were simply doing quadratic regression, in which case the objective would be:

$$\mathcal{L} = \min_{W_1, W_2} \sum_i ||y_i - W_1 x_i - \frac{\alpha}{2} W_2 x_i^2|| \tag{9}$$

$$\frac{d\mathcal{L}}{dW_1} = -2\mathrm{E}[XY] + 2W_1 \mathrm{E}[X^2] + \alpha W_2 \mathrm{E}[X^3] \tag{10}$$

$$\frac{d\mathcal{L}}{dW_2} = -2\alpha \mathrm{E}[X^2 Y] + \alpha W_1 \mathrm{E}[X^3] + \frac{\alpha^2}{2} W_2 \mathrm{E}[X^4] \tag{11}$$

Here, we see a similar dependence on fourth order moments, but we now have two parameters to fit. In some sense, we might argue that a quadratic activation function is compressing more moments into fewer parameters.

4 Quadratic Activation Functions with Multiple Layers

We now imagine the setting with a one dimensional input, a single one dimensional hidden layer, and a one dimensional output.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.