Institut for Theoretical Physics
University Leipzig

Prof. Dr. B. Rosenow
A. Afanah , M. Staats

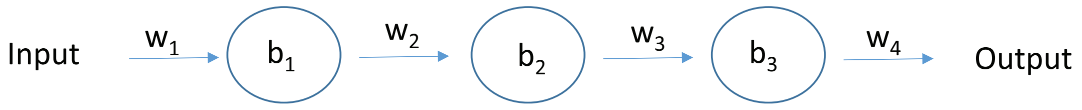
Advanced Statistical Physics - Sheet 1

Summer Terms 2022

Hand in: Thursday 14.04 at 15:15, please hand in only the tasks marked with * to mailbox no. (43) inside ITP room 105b. Due to Easter holiday, solutions to the problem set will be discussed Wednesday 20.04 at 9:15 (German tutorial) and at 15:15 (English tutorial).

1. The vanishing gradient problem (Bonus) * *2+1+4+3 Points*

Consider a simple neural network with one unit per hidden layer as depicted in the figure below. The output of the network is denoted by $f_w(x) = w_4 g(z_3)$ with the sigmoidal activation function $g(z) = \dfrac{1}{1+e^{-z}}$. The preactivations z_i are computed recursively in the following way: $z_1 = w_1 x + b_1$, and $z_i = w_i g(z_{i-1}) + b_i$. The deviation between the desired output $a(x)$ and the actual network output is quantified by the quadratic cost function $C = \frac{1}{2}[f_w(x) - a(x)]^2$.



Deep neural networks are known to suffer from the so-called vanishing gradient problem. The goal of this exercise is to invistigate the cause of this problem, and then to explore suitable restrictions to be imposed on the range of weights and activations to avoid this problem.

(a) Find the derivative $\dfrac{\partial C}{\partial b_1}$, and comment on the value of the derivative in the case $g'(z) < 1$.

(b) Consider the product $|wg'(wz+b)|$ and assume that $|wg'(wz+b)| \geq 1$. Argue that this can only ever occur if $|w| \geq 4$.

(c) Supposing that $|w| \geq 4$, consider the set of preactivations $\{z\}$ for which $|wg'(wz+b)| \geq 1$. Show that the set of $\{z\}$ satisfying that constraint can range over an interval no greater in width than

$$\frac{2}{|w|} \ln \left(\frac{|w|(1+\sqrt{1-4/|w|})}{2} - 1 \right)$$

(d) Show numerically that the above expression bounding the width of the range is greatest at $|w| \approx 6.9$, where it takes a value ≈ 0.45. So given that everything lines up just perfectly, we still have a fairly narrow range of input activations which can avoid the vanishing gradient problem.

2. Training with Tensorflow *8 Points*

Use the tensorflow function tf.random.set_seed(x) to make your results reproducible in all of the exercises.

(a) Build a neural network with one hidden layer and vary the size of the hidden layer (10,20,40,50,100,300). Train for 5 Epochs on the MNIST dataset and store the final validation accuracy. Then plot the accuracy as a function of the hidden layer size.

(b) Use the same network with fixed hidden layer size of 30 to estimate the impact of random initialisation. Run the network 10 times with different weight initialisation. Compute standard deviation, mean and plot the data such that the X-Axis is the number of the run and the Y-Axis is the vallidation accuracy.