Labs
Machine Learning Course
Fall 2019

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
www.epfl.ch/labs/mlo/machine-learning-cs-433

Problem Set 11, Nov 28, 2019 (Solution to Theory Question)

1 Vanishing Gradient

Note that the overall function $f(\boldsymbol{x}^{(0)})$ is a composition of $(L+1)$ functions, where the first L functions correspond to the L layers of the neural network and the last one corresponds to the output layer. So we have

$$f(\boldsymbol{x}^{(0)}) = f^{(L+1)} \circ \cdots \circ f^{(2)} \circ f^{(1)}(\boldsymbol{x}^{(0)}).$$

where

$$\boldsymbol{x}^{(l)} = f^{(l)}(\boldsymbol{x}^{(l-1)}) = \phi((\boldsymbol{W}^{(l)})^\top \boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l)}). \tag{1}$$

As written in the statement of the problem, the partial derivative $\frac{\partial f}{\partial w_{1,1}^{1}}$ is the product of the derivatives of these $L+1$ functions. Now note that our activation functions are sigmoids and those have a maximal derivative of $\frac{1}{4}$, i.e.,

$$\max_x \left(\frac{1}{1+e^{-x}}\right)' = \max \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{4}.$$

Therefore, for each of the L layers we will get a factor of $\frac{1}{4}$ or smaller. It remains to bound the inner derivative for each such function. Note that by assumption each weight has magnitude at most 1 and we assumed that we have $K = 3$, i.e., we have only three nodes per layer. Therefore, we get at most a factor 3 from the inner derivative. This proves the claim.