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

Problem Set 08, Nov 19, 2021 (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

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

Applying the chain rule to calculate $\frac{\partial f}{\partial W_{1,1}^{(1)}}$ we get:

$$\frac{\partial f}{\partial W_{1,1}^{(1)}} = f'_{L+1} \times f'_L \times \ldots \times f'_2 \times \frac{\partial f_1}{\partial W_{1,1}^{(1)}}$$

We are interested in showing that this value vanishes exponentially with L, i.e. $\left\|\frac{\partial f}{\partial W_{1,1}^{(1)}}\right\|_2 \leq O(\frac{3}{4}^L)$. We first remark the following definition:

Definition: The 2-operator norm of a matrix can be defined as $\|A\|_2^2 := \max_v \frac{\|Av\|_2^2}{\|v\|_2^2}$ where the maximum is taken over all vectors. Note that the norm in the left side is a opeator norm which is different from the norms in the right side corresponding to L2-norm defined in vector space which is also shown by the symbol $\|v\|_2$ but where v is a vector.

Applying $\|Av\|_2 \leq \|A\|_2 \|v\|_2$ (which follows from the definition of 2-operator norm) and $\|AB\|_2 \leq \|A\|_2 \|B\|_2$ (which can be seen by noting that $\frac{\|ABv\|_2^2}{\|v\|_2^2} = \frac{\|A(Bv)\|_2^2}{\|Bv\|_2^2} \cdot \frac{\|Bv\|_2^2}{\|v\|_2^2}$ and taking the max), we get

$$\left\|\frac{\partial f}{\partial W_{1,1}^{(1)}}\right\|_2 \leq \|f'_{L+1}\|_2 \cdot \|f'_L\|_2 \cdot \ldots \cdot \|f'_2\|_2 \times \left\|\frac{\partial f_1}{\partial W_{1,1}^{(1)}}\right\|_2 . \tag{2}$$

From (1) we can obtain

$$f'_l(\boldsymbol{x}^{(l-1)}) = (\boldsymbol{W}^{(l)})^\top \operatorname{diag}(\phi'((\boldsymbol{W}^{(l)})^\top \boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l)})$$

where $\operatorname{diag}(v)$ converts a vector to a diagonal matrix with the diagonal entries filled with elements of v. We can now bound the norm as

$$\left\|f'_l(\boldsymbol{x}^{(l-1)})\right\| \leq \left\|(\boldsymbol{W}^{(l)})^\top\right\|_2 \cdot \left\|\operatorname{diag}(\phi'((\boldsymbol{W}^{(l)})^\top \boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l)})\right\|_2 \leq \left\|(\boldsymbol{W}^{(l)})^\top\right\|_2 \cdot \max[\phi'((\boldsymbol{W}^{(l)})^\top \boldsymbol{x}^{(l-1)}] \tag{3}$$

where the last inequality follows from the second term being diagonal. 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, the sigmoid term in (3) is upper bounded by $\frac{1}{4}$. 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. Now note that for any vector v

$$(\boldsymbol{W}^{(l)} v)_i = \sum_{j=1}^{3} (\boldsymbol{W}^{(l)}_{i,j} v_j) \leq \sum_{j=1}^{3} |\boldsymbol{W}^{(l)}_{i,j}| \cdot |v_j| \leq \sum_{j=1}^{3} |v_j|$$

Using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ (which can be proven using Cauchy-Schwarz inequality), we can write for any vector v,

$$\frac{\left\|\boldsymbol{W}^{(l)} v\right\|_2^2}{\|v\|_2^2} \leq \frac{3(\sum_{j=1}^{3} |v_j|)^2}{(\sum_{j=1}^{3} |v_j|^2)} \leq 9\,.$$

Therefore, we get $\left\|\boldsymbol{W}^{(l)}\right\|_2 \leq 3$ which means the second term in (3) is bounded by 3. Therefore $\left\|f_l'(\boldsymbol{x}^{(l-1)})\right\|_2 \leq \frac{3}{4}$ which in combination with (2) proves our goal.