Homework 3
600.482/682 Deep Learning
Spring 2020

March 1, 2020

Due Sun. 03/01/2020 11:59:00pm.
Please submit a latex generated PDF
to Gradescope with entry code 9G83Y7

1. We have talked about backpropagation in class. And here is a supplementary material for calculating the gradient for backpropagation (https://piazza.com/class_profile/get_resource/jxcftju833c25t/k0labsf3cny4qw). Please study this material carefully before you start this exercise. Suppose $P = WX$ and $L = f(P)$ which is a loss function.

   (a) Please show that $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial P} X^T$. Show each step of your derivation.

   Answer: $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial P}\frac{\partial P}{\partial W}$.$P = WX$, $\frac{\partial L}{\partial P} = \begin{pmatrix} \frac{\partial L}{\partial p_{11}} & \cdots & \frac{\partial L}{\partial p_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial p_{n1}} & \cdots & \frac{\partial L}{\partial p_{nm}} \end{pmatrix}$, $\frac{\partial P}{\partial W} = \begin{pmatrix} \frac{\partial P}{\partial w_{11}} & \cdots & \frac{\partial P}{\partial w_{1k}} \\ \vdots & \ddots & \vdots \\ \frac{\partial P}{\partial w_{n1}} & \cdots & \frac{\partial P}{\partial w_{nk}} \end{pmatrix}$

$$P = WX = \begin{pmatrix} w_{11}x_{11} + ... + w_{1k}x_{k1} & \cdots & w_{11}x_{1m} + ... + w_{1k}x_{km} \\ \vdots & \ddots & \vdots \\ w_{n1}x_{11} + ... + w_{nk}x_{k1} & \cdots & w_{n1}x_{1m} + ... + w_{nk}x_{km} \end{pmatrix}$$

$$\frac{\partial P}{\partial w_{ij}} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ x_{j1} & \cdots & x_{jm} \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial P}\frac{\partial P}{\partial w_{ij}} = \sum_{a=1}^{n}\sum_{b=1}^{m} \frac{\partial L}{\partial p_{ab}}\frac{\partial p_{ab}}{\partial w_{ij}} = \frac{\partial L}{\partial p_{i1}}x_{j1} + ... + \frac{\partial L}{\partial p_{im}}x_{jm}$$

$$\frac{\partial L}{\partial W} = \begin{pmatrix} \frac{\partial L}{\partial p_{11}}x_{11} + ... + \frac{\partial L}{\partial p_{1m}}x_{1m} & \cdots & \frac{\partial L}{\partial p_{11}}x_{k1} + ... + \frac{\partial L}{\partial p_{1m}}x_{km} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial p_{n1}}x_{11} + ... + \frac{\partial L}{\partial p_{nm}}x_{1m} & \cdots & \frac{\partial L}{\partial p_{n1}}x_{k1} + ... + \frac{\partial L}{\partial p_{nm}}x_{km} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial L}{\partial p_{11}} & \cdots & \frac{\partial L}{\partial p_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial p_{n1}} & \cdots & \frac{\partial L}{\partial p_{nm}} \end{pmatrix} \cdot \begin{pmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & \ddots & \vdots \\ x_{1m} & \cdots & x_{km} \end{pmatrix}$$

$$\text{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{km} \end{pmatrix}$$

$$So,\ \frac{\partial L}{\partial W} = \frac{\partial L}{\partial P} X^T$$

(b) Suppose the loss function is L2 loss. L2 loss is defined as $L(y, \hat{y}) = \|y - \hat{y}\|^2$ where y is the groundtruth; $\hat{y}$ is the prediction. Given the following initialization of W and X, please calculate the updated W after one iteration. (step size = 0.1)

$$W = \begin{pmatrix} 0.3 & 0.5 \\ -0.2 & 0.4 \end{pmatrix}, X = (\mathbf{x_1}, \mathbf{x_2}) = \begin{pmatrix} 0 & 2 \\ 3 & 1 \end{pmatrix}, Y = (\mathbf{y_1}, \mathbf{y_2}) = \begin{pmatrix} 0.5 & 1 \\ 1 & -1.5 \end{pmatrix}$$

Answer: $\hat{Y} = WX = \begin{pmatrix} 0.3 & 0.5 \\ -0.2 & 0.4 \end{pmatrix} \begin{pmatrix} 0 & 2 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 1.5 & 1.1 \\ 1.2 & 0 \end{pmatrix}$, $L(y, \hat{y}) = \|y - \hat{y}\|^2$, so $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial W} = -2(Y - \hat{Y})X^T = -2 \begin{pmatrix} -1 & -0.1 \\ -0.2 & -1.5 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 0.4 & 6.1 \\ 6 & 4.2 \end{pmatrix}$

$$\text{W} = \text{W} + \lambda \frac{\partial L}{\partial W} = \begin{pmatrix} 0.3 & 0.5 \\ -0.2 & 0.4 \end{pmatrix} + 0.1 \begin{pmatrix} 0.4 & 6.1 \\ 6 & 4.2 \end{pmatrix} = \begin{pmatrix} 0.26 & -0.12 \\ -0.8 & 0.02 \end{pmatrix}$$

2. In this exercise, we will explore how vanishing and exploding gradients affect the learning process. Consider a simple, 1-dimensional, 3 layer network with data $x \in \mathbb{R}$, prediction $\hat{y} \in [0, 1]$, true label $y \in \{0, 1\}$, and weights $w_1, w_2, w_3 \in \mathbb{R}$, where weights are initialized randomly via $\sim \mathcal{N}(0, 1)$. We will use the sigmoid activation function σ between all layers, and the cross entropy loss function $L(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$. This network can be represented as: $\hat{y} = \sigma(w_3 \cdot \sigma(w_2 \cdot \sigma(w_1 \cdot x)))$. Note that for this problem, we are not including a bias term.

(a) Compute the derivative for a sigmoid. What are the values of the extrema of this derivative, and when are they reached?

Answer: $f(x) = sigmoid(x) = \frac{1}{1 + e^{-x}}$, $f'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{1 + e^{-x}} - \frac{1}{(1 + e^{-x})^2}$,
$f''(x) = \frac{e^{-x}}{(1 + e^{-x})^2} - \frac{2e^{-x}}{(1 + e^{-x})^3} = \frac{e^{-2x} - e^{-x}}{(1 + e^{-x})^3}$
when x < 0, f"(x) > 0 and when x > 0, f"(x) < 0. so when x < 0, the f'(x) increase with x increase and when x > 0, the f'(x) decrease with x increase. when x=0, f'(x) reach the minimum, the minimum is $\frac{1}{4}$.

(b) Consider a random initialization of $w_1 = 0.25, w_2 = -0.11, w_3 = 0.78$, and a sample from the data set $(x = 0.63, y = 1)$. Using backpropagation, compute the gradients for each weight. What have you noticed about the magnitude of the gradient?

Answer: $\hat{y} = \sigma(w_3 \cdot \sigma(w_2 \cdot \sigma(w_1 \cdot x)) = \sigma(0.78 \cdot \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63))))$. Suppose $s_1 = \sigma(0.25 \cdot 0.63)) = 0.5393, s_2 = \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63)) = 0.4852, s_3 = \sigma(0.78 \cdot \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63)))) = 0.5935$.

$L(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$. $So\ \frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} = -\frac{1}{\hat{y}} = -\frac{1}{s_3} = -1.6849$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_3} = -1.6849 \cdot \frac{e^{-w_3 \cdot s_2}}{(1 + e^{-w_3 \cdot s_2})^2} \cdot s_2 = -0.1972$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s_2} \frac{\partial s_2}{\partial w_2} = -1.6849 \cdot \frac{e^{-w_3 \cdot s_2}}{(1 + e^{-w_3 \cdot s_2})^2} \cdot w_3 \frac{e^{-w_2 \cdot s_1}}{(1 + e^{-w_2 \cdot s_1})^2} \cdot s_1 = -0.0427$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s_2} \frac{\partial s_2}{\partial s_1} \frac{\partial s_1}{\partial w_1} = -1.6849 \cdot \frac{e^{-w_3 \cdot s_2}}{(1 + e^{-w_3 \cdot s_2})^2} \cdot w_3 \frac{e^{-w_2 \cdot s_1}}{(1 + e^{-w_2 \cdot s_1})^2} \cdot w_2 \frac{e^{-w_1 \cdot x}}{(1 + e^{-w_1 \cdot x})^2} \cdot$$
$w_1 = 0.00136$

I noticed that after backpropagation the absolute value of gradient will decrease. So after 3 times the gradient become extremely small.

Now consider that we want to switch to a regression task and use a similar network structure as we did above: we remove the final sigmoid activation, so our new network is defined as $\hat{y} = w_3 \cdot \sigma(w_2 \cdot \sigma(w_1 \cdot x))$, where predictions $\hat{y} \in \mathcal{R}$ and targets $y \in \mathcal{R}$; we use the L2 loss function instead of cross entropy: $L(y, \hat{y}) = (y - \hat{y})^2$. Derive the gradient of the loss function with respect to each of the weights w_1, w_2, w_3.

Answer: $\hat{y} = w_3 \cdot \sigma(w_2 \cdot \sigma(w_1 \cdot x)) = 0.78 \cdot \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63)))$. Suppose $s_1 = \sigma(0.25 \cdot 0.63)) = 0.5393, s_2 = \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63)) = 0.4852, s_3 = 0.78 \cdot \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63))) = 0.3784$.

$L(y, \hat{y}) = (y - \hat{y})^2$. *So* $\frac{\partial L}{\partial \hat{y}} = -2 \cdot (y - \hat{y}) = -2 \cdot (1 - \hat{y}) = -2 \cdot (1 - s_3) = -1.2431$

$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_3} = -1.2431 \cdot s_2 = -0.6031$

$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s_2} \frac{\partial s_2}{\partial w_2} = -1.2431 \cdot w_3 \frac{e^{-w_2 \cdot s_1}}{(1 + e^{-w_2 \cdot s_1})^2} \cdot s_1 = -0.1306$

$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s_2} \frac{\partial s_2}{\partial s_1} \frac{\partial s_1}{\partial w_1} = -1.2431 \cdot w_3 \frac{e^{-w_2 \cdot s_1}}{(1 + e^{-w_2 \cdot s_1})^2} \cdot w_2 \frac{e^{-w_1 \cdot x}}{(1 + e^{-w_1 \cdot x})^2} \cdot w_1 = 0.00417$

(c) Consider again the random initialization of $w_1 = 0.25, w_2 = -0.11, w_3 = 0.78$, and a sample from the data set $(x = 0.63, y = 128)$. Using backpropagation, compute the gradients for each weight. What have you noticed about the magnitude of the gradient?

Answer: $\hat{y} = w_3 \cdot \sigma(w_2 \cdot \sigma(w_1 \cdot x)) = 0.78 \cdot \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63)))$. Suppose $s_1 = \sigma(0.25 \cdot 0.63)) = 0.5393, s_2 = \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63)) = 0.4852, s_3 = 0.78 \cdot \sigma(-0.11 \cdot \sigma(0.25 \cdot 0.63))) = 0.3784$.

$L(y, \hat{y}) = (y - \hat{y})^2$. *So* $\frac{\partial L}{\partial \hat{y}} = -2 \cdot (y - \hat{y}) = -2 \cdot (128 - \hat{y}) = -2 \cdot (128 - s_3) = -255.2431$

$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_3} = -255.2431 \cdot s_2 = -123.8373$

$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s_2} \frac{\partial s_2}{\partial w_2} = -255.2431 \cdot w_3 \frac{e^{-w_2 \cdot s_1}}{(1 + e^{-w_2 \cdot s_1})^2} \cdot s_1 = -26.8184$

$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial s_2} \frac{\partial s_2}{\partial s_1} \frac{\partial s_1}{\partial w_1} = -255.2431 \cdot w_3 \frac{e^{-w_2 \cdot s_1}}{(1 + e^{-w_2 \cdot s_1})^2} \cdot w_2 \frac{e^{-w_1 \cdot x}}{(1 + e^{-w_1 \cdot x})^2} \cdot w_1 = 0.8562$

I noticed that even the y is very big, after backpropagation the gradient become extremely small.