Homework 3
600.482/682 Deep Learning
Spring 2020

March 1, 2020

Mou Zhang
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.
   Soltion:
   Let's suppose that $P \in R^{N \times M}$, $W \in R^{N \times K}$, $X \in R^{K \times M}$

$$\text{Then } \frac{\partial L}{\partial W} = \frac{\partial L}{\partial P}\frac{\partial P}{\partial W}$$

$$\frac{\partial L}{\partial P} = \begin{pmatrix} \frac{\partial L}{\partial p_{11}} & \cdots & \frac{\partial L}{\partial p_{1M}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial p_{N1}} & \cdots & \frac{\partial L}{\partial p_{NM}} \end{pmatrix}$$

$$\frac{\partial P}{\partial W} = \begin{pmatrix} \frac{\partial P}{\partial w_{11}} & \cdots & \frac{\partial P}{\partial w_{1K}} \\ \vdots & \ddots & \vdots \\ \frac{\partial P}{\partial w_{N1}} & \cdots & \frac{\partial P}{\partial w_{NK}} \end{pmatrix},$$

$$\text{Since } P = WX = \begin{pmatrix} w_{11}x_{11} + \ldots + w_{1K}x_{K1} & \cdots & w_{11}x_{1M} + \ldots + w_{1K}x_{KM} \\ \vdots & \ddots & \vdots \\ w_{N1}x_{11} + \ldots + w_{NK}x_{K1} & \cdots & w_{N1}x_{1M} + \ldots + w_{NK}x_{KM} \end{pmatrix}$$

$$\text{We have } \frac{\partial P}{\partial w_{ij}} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \\ x_{j1} & \cdots & x_{jM} \\ 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

$$\text{So } \frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial P}\frac{\partial P}{\partial w_{ij}} = \begin{pmatrix} \frac{\partial L}{\partial p_{11}} & \cdots & \frac{\partial L}{\partial p_{1M}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial p_{N1}} & \cdots & \frac{\partial L}{\partial p_{NM}} \end{pmatrix} \cdot \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \\ x_{j1} & \cdots & x_{jM} \\ 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

$$= \frac{\partial L}{\partial p_{i1}} x_{j1} + \cdots + \frac{\partial L}{\partial p_{iM}} x_{jM}$$

So we have
$$\frac{\partial L}{\partial W} = \begin{pmatrix} \frac{\partial L}{\partial w_{11}} & \cdots & \frac{\partial L}{\partial w_{1K}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial w_{N1}} & \cdots & \frac{\partial L}{\partial w_{NK}} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{\partial L}{\partial p_{11}} x_{11} + \cdots + \frac{\partial L}{\partial p_{1M}} x_{1M} & \cdots & \frac{\partial L}{\partial p_{11}} x_{K1} + \cdots + \frac{\partial L}{\partial p_{1M}} x_{KM} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial p_{N1}} x_{11} + \cdots + \frac{\partial L}{\partial p_{NM}} x_{1M} & \cdots & \frac{\partial L}{\partial p_{N1}} x_{K1} + \cdots + \frac{\partial L}{\partial p_{NM}} x_{KM} \end{pmatrix}$$
$$= \frac{\partial L}{\partial P} X^T$$

(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)

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

Solution:

$$\begin{aligned} \frac{\partial L}{\partial W} &= \frac{\partial L}{\partial P} X^T \\ &= -2 \cdot (Y - \hat{Y}) \cdot X^T \\ &= -2 \cdot (Y - WX) \cdot X^T \\ &= -2 \cdot \left(\begin{pmatrix} 0.5 & 1 \\ 1 & -1.5 \end{pmatrix} - \begin{pmatrix} 0.3 & 0.5 \\ -0.2 & 0.4 \end{pmatrix} \begin{pmatrix} 0 & 2 \\ 3 & 1 \end{pmatrix} \right) \cdot \begin{pmatrix} 0 & 3 \\ 2 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.4 & 6.2 \\ 6 & 4.2 \end{pmatrix} \\ W &= W - step\_size * \frac{\partial L}{\partial W} \\ &= \begin{pmatrix} 0.3 & 0.5 \\ -0.2 & 0.4 \end{pmatrix} - 0.1 \times \begin{pmatrix} 0.4 & 6.2 \\ 6 & 4.2 \end{pmatrix} \\ &= \begin{pmatrix} 0.26 & -0.12 \\ -0.8 & -0.02 \end{pmatrix} \end{aligned}$$

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?
Solution:

$$\begin{aligned} S'(x) &= (\frac{1}{1 + e^{-x}})' \\ &= \frac{(1)' \cdot (1 + e^{-x}) - 1 \cdot (1 + e^{-x})'}{(1 + e^{-x})^2} \\ &= \frac{e^{-x}}{(1 + e^{-x})^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1+e^{-x}-1}{(1+e^{-x})^2} \\
&= \frac{1}{(1+e^{-x})}\left(1-\frac{1}{(1+e^{-x})}\right) \\
&= S(x)(1-s(x))
\end{aligned}
$$

The extrema of this derivative is 0.25. When x is equal to 0, we get the extrema 0.25.

(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?

Solution:

Suppose $a = S(w_1 \cdot x)$, $b = S(w_2 \cdot S(w_1 \cdot x)) = S(w_1 \cdot a)$, we have $\hat{y} = S(w_3 \cdot b)$
So we have a = 0.5393, b = 0.4852, $\hat{y} = 0.5935$
Then

$$
\begin{aligned}
\frac{\partial L}{\partial \hat{y}} &= -\left(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right) \\
&= -\frac{y}{\hat{y}} \text{ (Since y = 1)} = -1.6849 \\
\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial w_3} \\
&= -\frac{y}{\hat{y}} \cdot \hat{y} \cdot (1-\hat{y}) \cdot b = -0.1972 \\
\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial b}\frac{\partial b}{\partial w_2} \\
&= -\frac{y}{\hat{y}} \cdot \hat{y} \cdot (1-\hat{y}) \cdot w_3 \cdot b \cdot (1-b) \cdot a = -0.0427 \\
\frac{\partial L}{\partial w_1} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial b}\frac{\partial b}{\partial a}\frac{\partial a}{\partial w_1} \\
&= -\frac{y}{\hat{y}} \cdot \hat{y} \cdot (1-\hat{y}) \cdot w_3 \cdot b \cdot (1-b) \cdot w_2 \cdot a \cdot (1-a) \cdot x = 0.0014
\end{aligned}
$$

I noticed that after going through 3 sigmoid functions the magnitude of the gradient becomes extremely small. This is the gradient vanishing problem

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.

Solution:

Suppose $a = S(w_1 \cdot x)$, $b = S(w_2 \cdot S(w_1 \cdot x)) = S(w_1 \cdot a)$, we have $\hat{y} = w_3 \cdot b$
So we have a = 0.5393, b = 0.4852, $\hat{y} = 0.3784$
Then

$$
\begin{aligned}
\frac{\partial L}{\partial \hat{y}} &= -2(y-\hat{y}) = -1.2431 \\
\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial w_3} \\
&= -2(y-\hat{y}) \cdot b = -0.6031 \\
\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial b}\frac{\partial b}{\partial w_2} \\
&= -2(y-\hat{y}) \cdot w_3 \cdot b \cdot (1-b) \cdot a = -0.1306 \\
\frac{\partial L}{\partial w_1} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial b}\frac{\partial b}{\partial a}\frac{\partial a}{\partial w_1} \\
&= -2(y-\hat{y}) \cdot w_3 \cdot b \cdot (1-b) \cdot w_2 \cdot a \cdot (1-a) \cdot x = 0.0042
\end{aligned}
$$

(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?

Solution:

Suppose $a = S(w_1 \cdot x)$, $b = S(w_2 \cdot S(w_1 \cdot x)) = S(w_1 \cdot a)$, we have $\hat{y} = w_3 \cdot b$
So we have a = 0.5393, b = 0.4852, $\hat{y} = 0.3784$
Then

$$\begin{aligned}
\frac{\partial L}{\partial \hat{y}} &= -2(y - \hat{y}) = -255.2431 \\
\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_3} \\
&= -2(y - \hat{y}) \cdot b = -123.8372 \\
\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial b} \frac{\partial b}{\partial w_2} \\
&= -2(y - \hat{y}) \cdot w_3 \cdot b \cdot (1 - b) \cdot a = -26.81835 \\
\frac{\partial L}{\partial w_1} &= \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial b} \frac{\partial b}{\partial a} \frac{\partial a}{\partial w_1} \\
&= -2(y - \hat{y}) \cdot w_3 \cdot b \cdot (1 - b) \cdot w_2 \cdot a \cdot (1 - a) \cdot x = 0.8562
\end{aligned}$$

I noticed that after going through 2 sigmoid functions the magnitude of the gradient becomes much smaller than before, even though better than 3 sigmoid functions. This will cause the gradient vanishing problem.