Homework 3
600.482/682 Deep Learning
Fall 2020

18 September 2020

Due 2020 Fri. September 23 11:59pm.
Please submit a latex generated PDF
to Gradescope with entry code M63JEZ

1. We have presented a matrix back propagation example in class. In this exercise, you will follow the same logic we used in class to derive $\frac{\partial L}{\partial X} = W^T \frac{\partial L}{\partial Y}$.

   (a) Please derive $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y} X^T$ (please show each step of your computation).
   Solution: We derive the equation for arbitrary sized 2 dimensional matrices. Let $W \in \mathbb{R}^{a \times b}$, $X \in \mathbb{R}^{b \times c}$, $Y = WX \in \mathbb{R}^{a \times c}$, $L = f(Y) \in \mathbb{R}$.
   Let's write out the $WandX$ using subscripts:

$$W = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1b} \\ w_{21} & w_{22} & \dots & w_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ w_{a1} & w_{a2} & \dots & w_{ab} \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1c} \\ x_{21} & x_{22} & \dots & x_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ x_{b1} & x_{b2} & \dots & x_{bc} \end{bmatrix}$$

   Now we can write out Y using summation. The summations subscripts are left out to increase readability, but all the summations are over index i and go from $i = 1$ to $i = b$.

$$Y = \begin{bmatrix} \sum w_{1i} x_{i1} & \dots & \sum w_{1i} x_{ic} \\ \vdots & \ddots & \vdots \\ \sum w_{ai} x_{i1} & \dots & \sum w_{ai} x_{ic} \end{bmatrix}$$

   Now that we have the matrices written out, we can start calculating the gradients. We know $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y} \frac{\partial Y}{\partial W}$. We can expand each tensor:

$$\frac{\partial L}{\partial Y} = \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \dots & \frac{\partial L}{\partial y_{1c}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial y_{a1}} & \dots & \frac{\partial L}{\partial y_{ac}} \end{bmatrix} \qquad \frac{\partial Y}{\partial W} = \begin{bmatrix} \frac{\partial Y}{\partial w_{11}} & \dots & \frac{\partial Y}{\partial w_{1b}} \\ \vdots & \ddots & \vdots \\ \frac{\partial Y}{\partial w_{a1}} & \dots & \frac{\partial Y}{\partial w_{ab}} \end{bmatrix}$$

   We can further expand each of the terms inside of $\frac{\partial Y}{\partial W}$. Here we will expand the 4 terms shown for illustrative purposes:

$$\frac{\partial Y}{\partial w_{11}} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1c} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \qquad \frac{\partial Y}{\partial w_{1b}} = \begin{bmatrix} x_{b1} & x_{b2} & \dots & x_{bc} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\frac{\partial Y}{\partial w_{a1}} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \\ x_{11} & x_{12} & \dots & x_{1c} \end{bmatrix} \qquad \frac{\partial Y}{\partial w_{ab}} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \\ x_{b1} & x_{b2} & \dots & x_{bc} \end{bmatrix}$$

Now we can finally write out $\frac{\partial L}{\partial W}$ as a matrix after performing element wise dot products. For brevity, I excluded the summation indices, but all summations are over index i and go from $i = 0$ to $i = c$.

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \begin{bmatrix} \sum \frac{\partial L}{\partial y_{1i}} x_{1i} & \sum \frac{\partial L}{\partial y_{1i}} x_{2i} & \dots & \sum \frac{\partial L}{\partial y_{1i}} x_{bi} \\ \sum \frac{\partial L}{\partial y_{2i}} x_{1i} & \sum \frac{\partial L}{\partial y_{2i}} x_{2i} & \dots & \sum \frac{\partial L}{\partial y_{2i}} x_{bi} \\ \vdots & & \vdots & \ddots & \vdots \\ \sum \frac{\partial L}{\partial y_{ai}} x_{1i} & \sum \frac{\partial L}{\partial y_{ai}} x_{2i} & \dots & \sum \frac{\partial L}{\partial y_{ai}} x_{bi} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \dots & \frac{\partial L}{\partial y_{1c}} \\ \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \dots & \frac{\partial L}{\partial y_{2c}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial y_{a1}} & \frac{\partial L}{\partial y_{a2}} & \dots & \frac{\partial L}{\partial y_{ac}} \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} & \dots & x_{b1} \\ x_{12} & x_{22} & \dots & x_{b2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1c} & x_{2c} & \dots & x_{bc} \end{bmatrix} \\
&= \frac{\partial L}{\partial Y} X^T
\end{aligned}$$

(b) Suppose the loss function is L2 loss. Given the following initialization of W and X, please calculate the updated W after one iteration. (step size = 0.1)

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

Solution: Our computation should take the form: $W - \frac{\partial L}{\partial W}\Delta = W - \frac{\partial L}{\partial \hat{Y}} X^T \Delta$. First we need to compute $\hat{Y}$:

$$\hat{Y} = WX = \begin{bmatrix} 1.5 & 1.1 \\ 1.2 & 0 \end{bmatrix} \qquad Y = \begin{bmatrix} 0.5 & 0.2 \\ 1 & -1.5 \end{bmatrix}$$

Now we can calculate L:

$$L = \sum_{i=1}^{2} \sum_{j=1}^{2} (\hat{y}_{ij} - \hat{y}_{ij})^2 = 3.3$$

We can compute $\frac{\partial L}{\partial \hat{Y}}$ and $\frac{\partial L}{\partial W}$:

$$\frac{\partial L}{\partial \hat{Y}} = \begin{bmatrix} 2 & 0.2 \\ 0.4 & 3 \end{bmatrix} \qquad \frac{\partial L}{\partial W} = \frac{\partial L}{\partial \hat{Y}} X^T = \begin{bmatrix} 0.4 & 6.2 \\ 6 & 4.2 \end{bmatrix}$$

Finally, we can compute the updated weights:

$$W' = W - \Delta \frac{\partial L}{\partial W} = \begin{bmatrix} 0.3 & 0.5 \\ -0.2 & 0.4 \end{bmatrix} - \begin{bmatrix} 0.04 & 0.62 \\ 0.6 & 0.42 \end{bmatrix} = \begin{bmatrix} 0.26 & -0.12 \\ -0.8 & -0.02 \end{bmatrix}$$

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: We have calculated the derivative for the sigmoid several times already. To calculated the derivative apply the quotient rule:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad \sigma'(x) = \frac{e^x}{(1 + e^x)^2}$$

There is one extremum of the derivative, namely the maxima, $\sigma'(0) = 0.25$. The function is confined to the range $(0, 0.25]$.

(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 do you notice about the magnitude?

Solution: To aid in the calculations, I will break the network up into layers. Let $a = \sigma(w_1 x) = 0.539$, $b = \sigma(w_2 a) = 0.485$, $\hat{y} = \sigma(w_3 b) = 0.593$, $L = -log(\hat{y}) = -0.522$. We can also symbolically compute the gradients:

$$\frac{\partial L}{\partial \hat{y}} = -\frac{1}{\hat{y}} = -1.68$$

$$\frac{\partial L}{\partial b} = (-1.68) * \sigma'(w_3 b) w_3 = -0.317 \qquad \frac{\partial L}{\partial w_3} = (-1.68) * \sigma'(w_3 b) b = -0.197$$

$$\frac{\partial L}{\partial a} = (-0.317) * \sigma'(w_2 a) w_2 = 0.00871 \qquad \frac{\partial L}{\partial w_2} = (-0.317) * \sigma'(w_2 a) a = -0.0427$$

$$\frac{\partial L}{\partial x} = (0.00871) * \sigma'(w_1 x) w_1 = 0.000541 \qquad \frac{\partial L}{\partial w_1} = (0.00871) * \sigma'(w_1 x) x = 0.00136$$

We can see that the gradients of the loss function w.r.t. the weights decrease by an order of magnitude for each layer you back into the network.

Now consider that we want to switch to a regression task and keep a similar network structure. We will 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 will also 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.

(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 do you notice about the magnitude?

Solution: Similar to the last part, I will break the network up into layers. Let $a = \sigma(w_1 x) = 0.539$, $b = \sigma(w_2 a) = 0.485$, $\hat{y} = w_3 b = 0.378$, $L = (128 - \hat{y})^2 = 16287.26$. We can also symbolically compute the gradients:

$$\frac{\partial L}{\partial \hat{y}} = -2(128 - \hat{y}) = -255.24$$

$$\frac{\partial L}{\partial b} = (-255.24) * w_3 = -199.08 \qquad \frac{\partial L}{\partial w_3} = (-255.24) * b = -123.83$$

$$\frac{\partial L}{\partial a} = (-199.08) * \sigma'(w_2 a) w_2 = 5.470 \qquad \frac{\partial L}{\partial w_2} = (-199.08) * \sigma'(w_2 a) a = -26.82$$

$$\frac{\partial L}{\partial x} = (5.470) * \sigma'(w_2 a) w_1 = 0.340 \qquad \frac{\partial L}{\partial w_1} = (5.470) * \sigma'(w_1 x) x = 0.856$$

We can see that the gradients of the loss function w.r.t. the weights decrease by at least an order of magnitude for each layer you back into the network.