Homework 3
600.482/682 Deep Learning
Fall 2019

Jason Zhang

September 27, 2019

Due 2019 Fri. 09/27 11:59pm.
Please submit a latex generated PDF
to Gradescope with entry code MKDPGK

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:
   Let $W = \mathbb{R}^{mxd}$ and $X = \mathbb{R}^{dxn}$. If we suppose that $P = WX$, then $P = \mathbb{R}^{mxn}$. Given that $L = f(P)$, we can calculate $\frac{\partial L}{\partial W}$.

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial P}\frac{\partial P}{\partial W} \quad \text{Using the chain rule}$$

$$= \begin{bmatrix} \frac{\partial L}{p_{0,0}} & \cdots & \frac{\partial L}{p_{0,n}} \\ \vdots & \ddots & \cdots \\ \frac{\partial L}{p_{m,0}} & \cdots & \frac{\partial L}{p_{m,n}} \end{bmatrix} \begin{bmatrix} \frac{\partial P}{w_{0,0}} & \cdots & \frac{\partial P}{w_{0,n}} \\ \vdots & \ddots & \cdots \\ \frac{\partial P}{w_{m,0}} & \cdots & \frac{\partial P}{w_{m,n}} \end{bmatrix} \quad \text{Substituting the matrix representations.}$$

We can now consider $P = WX$ where $W = \begin{bmatrix} w_{0,0} & \cdots & w_{0,d} \\ \vdots & \ddots & \vdots \\ w_{m,0} & \cdots & w_{m,d} \end{bmatrix}$, and $X = \begin{bmatrix} x_{0,0} & \cdots & x_{0,n} \\ \vdots & \ddots & \vdots \\ x_{d,0} & \cdots & x_{n,d} \end{bmatrix}$.

Thus $P = \begin{bmatrix} w_{0,0}x_{0,0} + ... + w_{0,d}x_{d,0} & \cdots & w_{0,0}x_{0,n} + ... + w_{0,d}x_{n,d} \\ \vdots & \ddots & \vdots \\ w_{m,0}x_{0,0} + ... + w_{m,d}x_{d,0} & \cdots & w_{m,0}x_{0,n} + ... + w_{m,d}x_{n,d} \end{bmatrix}$.

Therefore,

$$\frac{\partial P}{\partial W} = \begin{bmatrix} \begin{bmatrix} x_{0,0} & \cdots & x_{0,n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} & \cdots & \begin{bmatrix} x_{d,0} & \cdots & x_{n,d} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ x_{0,0} & \cdots & x_{0,n} \end{bmatrix} & \cdots & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ x_{d,0} & \cdots & x_{n,d} \end{bmatrix} \end{bmatrix}$$

To calculate $\frac{\partial L}{\partial W}$, we need to consider the derivative with respect to every $w_{i,j}$ in W.

Thus, $\frac{\partial L}{\partial W} = \begin{bmatrix} \frac{\partial L}{w_{0,0}} & \cdots & \frac{\partial L}{w_{0,n}} \\ \vdots & \ddots & \cdots \\ \frac{\partial L}{w_{m,0}} & \cdots & \frac{\partial L}{w_{m,n}} \end{bmatrix}$.

For any $i, j \leq m, n$ $\frac{\partial L}{w_{i,j}} = \frac{\partial L}{\partial P}(\frac{\partial P}{\partial W}_{i,j}) = \sum_{k=1}^{m}\sum_{l=1}^{n}\frac{\partial L}{\partial p_{k,l}}\frac{\partial p_{k,l}}{w_{i,j}}$(Calculated essentially as a dot product since we know L and $x_{i,j}$ are scalar values). Using this fact we get:

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \begin{bmatrix} \frac{\partial L}{w_{0,0}} & \cdots & \frac{\partial L}{w_{0,n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{w_{m,0}} & \cdots & \frac{\partial L}{w_{m,n}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial L}{p_{0,0}}x_{0,0} + ... + \frac{\partial L}{p_{0,n}}x_{0,n} & \cdots & \frac{\partial L}{p_{0,0}}x_{d,0} + ... + \frac{\partial L}{p_{0,n}}x_{n,d} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{p_{m,0}}x_{0,0} + ... + \frac{\partial L}{p_{m,n}}x_{0,n} & \cdots & \frac{\partial L}{p_{m,0}}x_{d,0} + ... + \frac{\partial L}{p_{m,n}}x_{n,d} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial L}{p_{0,0}} & \cdots & \frac{\partial L}{p_{0,n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{p_{m,0}} & \cdots & \frac{\partial L}{p_{m,n}} \end{bmatrix} \begin{bmatrix} x_{0,0} & \cdots & x_{d,0} \\ \vdots & \ddots & \vdots \\ x_{0,n} & \cdots & x_{n,d} \end{bmatrix} \\
&= \frac{\partial L}{\partial P}X^T
\end{aligned}$$

(b) Suppose the loss function is L2 loss. For n examples, L2 loss is defined as $L = \sum_{i=1}^{n}\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2$, where $\mathbf{y}_i$ is the groundtruth label for the i-th example; $\hat{\mathbf{y}}_i$ is the prediction for the corresponding 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:

To start, our predictions $\hat{Y} = WX = \begin{pmatrix} 1.5 & 1.1 \\ 1.2 & 0 \end{pmatrix}$. As extra set up we will now compute the gradient of L.

$$\frac{\partial L}{\partial \hat{y}_{i,j}} = \sum_{l=1}^{2}\sum_{k=1}^{2} -2(y_{k,l} - \hat{y}_{k,l}) = -2(y_{i,j} - \hat{y}_{i,j})$$

Therefore $\frac{\partial L}{\partial \hat{Y}} = \begin{pmatrix} 2.0 & 0.2 \\ 0.4 & 3 \end{pmatrix}$

Now we calculate:

$$\begin{aligned}
\frac{\partial \hat{Y}}{\partial W} &= \begin{pmatrix} \frac{\partial \hat{Y}}{\partial w_{0,0}} & \frac{\partial \hat{Y}}{\partial w_{0,1}} \\ \frac{\partial \hat{Y}}{\partial w_{1,0}} & \frac{\partial \hat{Y}}{\partial w_{1,1}} \end{pmatrix} \\
&= \begin{pmatrix} \begin{pmatrix} x_{0,0} & x_{0,1} \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} x_{1,0} & x_{1,1} \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \\ x_{0,0} & x_{0,1} \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ x_{1,0} & x_{1,1} \end{pmatrix} \end{pmatrix} \\
&= \begin{pmatrix} \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 3 & 1 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 3 & 1 \end{pmatrix} \end{pmatrix}
\end{aligned}$$

For any $i, j \leq m, n$ $\frac{\partial L}{w_{i,j}} = \frac{\partial L}{\partial \hat{Y}}(\frac{\partial \hat{Y}}{\partial W}_{i,j}) = \sum_{k=1}^{m}\sum_{l=1}^{n}\frac{\partial L}{\partial \hat{y}_{k,l}}\frac{\partial \hat{y}_{k,l}}{w_{i,j}}$ Using this fact we get:

$$\frac{\partial L}{\partial W} = \begin{pmatrix} 0.4 & 6.2 \\ 6 & 4.2 \end{pmatrix}$$

Updating W: $W_{new} = W - \alpha * \frac{\partial L}{\partial W}$, where $\alpha = 0.1$ (Step size).

$W_{new} = \begin{pmatrix} 0.3 & 0.5 \\ -0.2 & 0.4 \end{pmatrix} - \begin{pmatrix} 0.04 & 0.62 \\ 0.6 & 0.42 \end{pmatrix} = \begin{pmatrix} 0.26 & -0.12 \\ -0.8 & -0.02 \end{pmatrix}$

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:
   Let $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\begin{aligned}\frac{d(\sigma(x))}{dx} &= -1(1+e^{-x})^{-2}(-e^{-x}) \\ &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1+e^{-x}-1}{(1+e^{-x})^2} \\ &= \frac{1+e^{-x}}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2} \\ &= \sigma(x) - \sigma(x)^2 \\ &= \sigma(x)(1-\sigma(x))\end{aligned}$$

   To find the extrema of the derivative we start by examining the critical points of the derivative, thus we need to take the second derivative of $\sigma(x)$).

$$\begin{aligned}\frac{d^2(\sigma(x))}{dx^2} &= (\frac{1}{(1+e^{-x})} - \frac{1}{(1+e^{-x})^2})' \qquad \text{Starting from the first derivative} \\ &= (1+e^{-x})^{-2}(e^{-x})(1-2(1+e^{-x})^{-1}) \\ &= \sigma'(x)(1-2\sigma(x))\end{aligned}$$

   The critical points will only be apparent when $(1-2\sigma(x)) = 0$, since $\sigma'(x)$ is non zero for all x. Thus solving for x we get: $\sigma(x) = \frac{1}{2}$, which can only happen at $x = 0$ because $1/1 + e^0 = 1/2$. Thus the value of the extrema of the derivative of sigmoid is found at: $x = 0$ and $\sigma'(0) = \frac{1}{4}$

   (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:
   We start by doing a few computations. $w_1 \cdot x = 0.157$, $o_1 = \sigma(w_1 \cdot x) = 0.539$, $w_2 \cdot o_1 = -0.11 * 0.539 = -0.059$, $o_2 = \sigma(w_2 \cdot o_1) = 0.485$, $w_3 \cdot o_2 = 0.78 * 0.485 = 0.378$, and $\hat{y} = 0.593$.
   Utilizing chain rule we can compute the gradients for each weight:

$$\begin{aligned}\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial w_3} \\ &= (-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}})(\sigma(w_3 o_2)(1-\sigma(w_3 o_2))o_2 \\ &= -0.197\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_2}\frac{\partial o_2}{\partial w_2} \\ &= (-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}})(\sigma(w_3 o_2)(1-\sigma(w_3 o_2))w_3 * \\ &\quad (\sigma(w_2 o_1)(1-\sigma(w_2 o_1))o_1) \\ &= -0.0427\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial w_1} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_2}\frac{\partial o_2}{\partial o_1}\frac{\partial o_1}{\partial w_1} \\ &= (-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}})(\sigma(w_3 o_2)(1-\sigma(w_3 o_2))w_3 * \\ &(\sigma(w_2 o_1)(1-\sigma(w_2 o_1)))w_2(\sigma(w_1 x)(1-\sigma(w_1 x))x \\ &= 0.00136\end{aligned}$$

Examining the magnitude of the gradients, all three gradients have magnitudes less than or equal to 0.1. This suggests the vanishing gradient issue in the sense that each of these gradients are so small that an update will have relatively no change, which means the network is not making as much progress. This is especially evident in the gradient of w_1 since the value is 0.00136, which is an extremely small value to update the gradient by, virtually making no significant change to the weights.

(c) 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: $L(y, \hat{y}) = (y - \hat{y})^2$.

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:

We start by doing a few computations. $w_1 \cdot x = 0.157$, $o_1 = \sigma(w_1 \cdot x) = 0.539$, $w_2 \cdot o_1 = -0.11 * 0.539 = -0.059$, $o_2 = \sigma(w_2 \cdot o_1) = 0.485$, $w_3 \cdot o_2 = 0.78 * 0.485 = 0.378$, and $\hat{y} = 0.378$.

Utilizing chain rule we can compute the gradients for each weight:

$$\begin{aligned}\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial w_3} \\ &= (-2(y-\hat{y}))o_2 \\ &= -123.79\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_2}\frac{\partial o_2}{\partial w_2} \\ &= (-2(y-\hat{y}))w_3(\sigma(w_2 o_1)(1-\sigma(w_2 o_1))o_1) \\ &= -26.8\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial w_1} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial o_2}\frac{\partial o_2}{\partial o_1}\frac{\partial o_1}{\partial w_1} \\ &= (-2(y-\hat{y}))w_3(\sigma(w_2 o_1)(1-\sigma(w_2 o_1)))w_2(\sigma(w_1 x)(1-\sigma(w_1 x))x \\ &= 0.8563\end{aligned}$$

Examining each of the gradients, with this new regression task, the gradients in comparison to part b are dramatically larger. The vanishing gradient problem that b faced is now replaced by an exploding gradient problem that this problem faces. Looking at gradients for w_3 and w_2 the gradients are relatively large integer values, which means per update, our gradients will move us further from the optimal value and keep growing, essentially exploding.