4705-Homework 4

Xinyi Zhao

November 3, 2023

1456

Problem 1

For the function $f(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ **show that** $\frac{df}{dz}(z) = 1 - tanh^2(z)$**.**

i: The derivative of f(z) with respect to z is:

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

Consider a vector $z = (z_1, \ldots, z_K)$ **and the softmax of this vector** $a = softmax(z)$ **where** $a_j = \frac{e^{z_j}}{\sum_{i=1}^{K} e^{z_i}}$**. Find an expression for** $\frac{da_j}{dz_j}$ **and prove that it is** $\frac{\sum_{i=1, j\neq i}^{K} e^{z_i + z_j}}{(\sum_{i=1}^{K} e^{z_i})^2}$

2: Given the softmax function:

$$a_j = \frac{e^{z_j}}{\sum_{i=1}^{K} e^{z_i}}$$

The derivative of a_j with respect to z_j is:

$$\begin{aligned}
\frac{da_j}{dz_j} &= \frac{\sum_{i=1}^{K} e^{z_i} e^{z_j} - e^{z_j} \cdot e^{z_j}}{\left(\sum_{i=1}^{K} e^{z_i}\right)^2} \\
&= \frac{e^{z_j}\left(\sum_{i=1}^{K} e^{z_i} - e^{z_j}\right)}{\left(\sum_{i=1}^{K} e^{z_i}\right)^2} \\
&= \frac{e^{z_j} \sum_{i=1, j\neq i}^{K} e^{z_i}}{\left(\sum_{i=1}^{K} e^{z_i}\right)^2} \\
&= \frac{\sum_{i=1, j\neq i}^{K} e^{z_i + z_j}}{\left(\sum_{i=1}^{K} e^{z_i}\right)^2}
\end{aligned}$$

Proved.

Show that for $\sigma(z) = \frac{1}{1+e^{-z}}$ **we have** $\frac{d\sigma}{dz}(z) = (1-\sigma(z))\sigma(z)$**.**

3:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

The derivative of $\sigma(z)$ is:

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

proved.

Show that $tanh(z) = 2\sigma(2z) - 1$**.**

4: Given:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

we can express $2\sigma(2z) - 1$ as exponentials:

$$2\sigma(2z) - 1 = \frac{2}{1+e^{-2z}} - 1 = \frac{1-e^{-2z}}{1+e^{-2z}}$$

Because $e^z > 0$, we can multiply by e^z:

$$2\sigma(2z) - 1 = \frac{e^z - e^{-z}}{e^z + e^{-z}} = tanh(z)$$

Proved.

Problem 2

Review the XOR example from class. In the XOR example with a neural network, we picked γ and ν to be specific values. Suppose we use σ and not $ReLU$. Can you find γ and ν that work? Prove this. Do this by smart guess and check. You can write a small Python program to get you the values you need.

Response Given the neural network for the XOR problem, the computation process can be described:

$$\begin{aligned}
\mathbf{x} &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
\mathbf{z}[1] &= \beta_1 \mathbf{x} + \alpha_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} \\
\mathbf{a}[1] &= \sigma(\mathbf{z}[1]) \\
\mathbf{z}[2] &= \beta_2 \mathbf{a}[1] + \alpha_2 \\
\mathbf{a}[2] &= \text{softmax}(\mathbf{z}[2])
\end{aligned}$$

We set $\beta[1]$ and $\alpha[1]$ to:

$$\beta_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \alpha_1 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

where:

- β_1 and β_2 are the weight matrices for the first and second layers, respectively.
- α_1 and α_2 are the bias vectors for the first and second layers, respectively.
- σ is the sigmoid activation function applied element-wise.

We denote γ and ν as follows:

$$\gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \quad \nu = \nu$$

And the conditions to be satisfied are:

1. $\gamma^T \begin{bmatrix} \sigma(1) \\ \sigma(0) \end{bmatrix} + \nu > 0$

2. $\gamma^T \begin{bmatrix} \sigma(1) \\ \sigma(0) \end{bmatrix} + \nu > 0$ (Same as condition 1)

3. $\gamma^T \begin{bmatrix} \sigma(2) \\ \sigma(1) \end{bmatrix} + \nu < 0$

4. $\gamma^T \begin{bmatrix} \sigma(0) \\ \sigma(-1) \end{bmatrix} + \nu < 0$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

To find the values of γ and ν that satisfy the XOR problem, we employed a Python script to perform a linear space search. The search space for γ and ν was set to (-10, 10) with 21 steps.

The results from the script found the following values that satisfy:

- $\gamma = [7, -6], \nu = -2$
- $\gamma = [9, -7], \nu = -3$
- $\gamma = [10, -8], \nu = -3$

These values were found to satisfy the conditions for the XOR problem as defined in the computational graph and the problem statement.

Problem 3

Suppose we have a neural network as in class and the output of the layer $a^{[1]}$ is $(x_1, x_2, x_1x_2, x_1^2, x_2^2)$ where $x = (x_1, x_2)$ is the input. Recall that for XOR we have (x_1, x_2) maps to y via $y = x_1 + x_2 - 2x_1x_2$ so that $(0, 0)$ maps to 0 and $(1, 0)$ maps to 1 (see Lecture). Consider $z^{[2]} = \beta^{[2]}a^{[1]} + \alpha^{[2]}$ and

$a^{[2]} = \sigma(z^{[2]})$ and how we want $a^{[2]} > 1/2$ if $y = 1$ and $1 - a^{[2]} > 1/2$ if $y = 0$. Can you specify $\beta^{[2]}$ and $\alpha^{[2]}$ that make this happen? Notice $\beta^{[2]} \in \mathbb{R}^5$ and $\alpha^{[2]} \in \mathbb{R}$. Do this by smart guess and check. You can write a small Python program to get you the values you need.

Response:

Given the XOR problem and a neural network with an extended feature set in the first layer, the architecture of the neural network is as follows:

$$
\begin{aligned}
\mathbf{x} &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
\mathbf{a}^{[1]} &= \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix} \\
z^{[2]} &= \beta^{[2]} \mathbf{a}^{[1]} + \alpha^{[2]} \\
a^{[2]} &= \sigma(z^{[2]})
\end{aligned}
$$

where σ is the sigmoid activation function.

For the XOR problem, we want:

- $a^{[2]} > \frac{1}{2}$ if $y = 1$ for inputs (1,0) and (0,1)
- $a^{[2]} < \frac{1}{2}$ if $y = 0$ for inputs (0,0) and (1,1)

Then consider the $a[1]$ for each input X:

- for x = (0,0), a[1]= (0,0,0,0,0)
- for x = (1,0), a[1] = (1,0,0,1,0) or
- for x= (0,1), a[1]=(0,1,0,0,1)
- for x = (1,1), a[1] = (1,1,1,1,1)

Let's denote $\beta[2] = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

For inputs (1,0) and (0,1), we want $a[2] > 1/2$: $\beta_1 + \beta_4 > 0$, $\beta_2 + \beta_5 > 0$

Using a Python script to perform a search over possible values of $\beta^{[2]}$ and $\alpha^{[2]}$, we found one example $\beta[2]$ and $\alpha[2]$ that satisfy the conditions is:

$$\beta^{[2]} = [-8, -6, -8, 10, 10]$$
$$\alpha^{[2]} = -1$$

With these values, the neural network correctly classifies the XOR inputs according to the specified conditions.

Problem 4

Suppose we use a *ReLU* so that the recursions are $a^{[0]} = x$, $z^{[1]} = \beta^{[1]}a^{[0]} + \alpha^{[1]}$, $a^{[1]} = ReLU(z^{[1]})$, $z^{[2]} = \beta^{[2]}a^{[1]} + \alpha^{[2]}$ and then finally $a^{[2]} = \sigma(z^{[2]})$ and $\ell = \log(a^{[2]})$ (i.e. we assume $y = 1$). What are the derivatives of ℓ with respect to $\beta^{[1],[2]}$ and $\alpha^{[1],[2]}$. For each variable, when will they be zero? Give some sufficient conditions in terms of the $z^{[1]}$ or $z^{[2]}$ variables.

Response to problem 4:

The derivative of l with respect to $a^{[2]}$ is :

$$\frac{dl}{da^{[2]}} = \frac{1}{a^{[2]}}$$

Using the chain rule, the derivative of l with respect to $z^{[2]}$ is:

$$\frac{dl}{dz^{[2]}} = \frac{dl}{da^{[2]}}\frac{da^{[2]}}{dz^{[2]}} = \frac{1}{a^{[2]}}a^{[2]}(1 - a^{[2]}) = 1 - \sigma(z^{[2]})$$

So the derivative of l with respect to $a^{[1]}$ is:

$$\frac{dl}{da^{[1]}} = \frac{dl}{dz^{[2]}}\frac{dz^{[2]}}{da^{[1]}} = (1 - \sigma(z^{[2]}))\beta^{[2]}$$

also,

$$\frac{dl}{d\beta^{[2]}} = \frac{dl}{dz^{[2]}}\frac{dz^{[2]}}{d\beta^{[2]}} = (1 - \sigma z^{[2]})a^{[1]}$$

$$\frac{dl}{d\alpha^{[2]}} = \frac{dl}{dz^{[2]}}\frac{dz^{[2]}}{d\alpha^{[2]}} = 1 - \sigma z^{[2]}$$

For the relu function:

$$\frac{da^{[1]}}{dz^{[1]}} = \begin{cases} 1 \text{ if } z^{[1]} > 0, \\ 0 \text{ otherwise} \end{cases}$$

Then compute derivative of l with respect to $z^{[1]}$:

$$\frac{dl}{dz^{[1]}} = \frac{dl}{da^{[1]}}\frac{da^{[1]}}{dz^{[1]}} = \begin{cases} (1-\sigma z^{[2]})\beta^{[2]} \text{ if } z^{[1]} > 0, \\ 0 \text{ otherwise} \end{cases}$$

$$\frac{dl}{d\beta^{[1]}} = \frac{dl}{dz^{[1]}}\frac{dz^{[1]}}{d\beta^{[1]}} = \begin{cases} (1-\sigma z^{[2]})\beta^{[2]}a^{[0]} \text{ if } z^{[1]} > 0, \\ 0 \text{ otherwise} \end{cases}$$

$$\frac{dl}{d\alpha^{[1]}} = \frac{dl}{dz^{[1]}} = \begin{cases} (1-\sigma z^{[2]})\beta^{[2]} \text{ if } z^{[1]} > 0, \\ 0 \text{ otherwise} \end{cases}$$

When will they be zero?

$\frac{dl}{d\beta^{[2]}}$ will be zero if $a^{[2]} = 1$ or $a^{[1]} = 0$ in terms of z: $z^{[1]} \leq 0$ or $z^{[2]} \to \infty$.

$\frac{dl}{d\alpha^{[2]}}$ will be zero if $a^{[2]} = 1$. that in terms of z:$z^{[2]} \to \infty$

$\frac{dl}{d\beta^{[1]}}$ will be zero if $a^{[2]} = 1$ or $\beta^{[2]} = 0$ or $a^{[0]} = 0$ or $z^{[1]} \leq 0$. In terms of z: $z^{[1]} \leq 0$ or $z^{[2]} \to \infty$.

$\frac{dl}{d\alpha^{[1]}}$ will be zero if $a^{[2]} = 1$ or $\beta^{[2]} = 0$ or $z^{[1]} \leq 0$. That equals $z^{[1]} \leq 0$ or $z^{[2]} \to \infty$.

Problem 5

Suppose we have $a^{[0]} = x$, $z^{[1]} = \beta^{[1]}a^{[0]} + \alpha^{[1]}$, $a^{[1]} = \sigma(z^{[1]})$, $z^{[2]} = z^{[1]} + \beta^{[2]}a^{[1]} + \alpha^{[2]}$ and then finally $a^{[2]} = \sigma(z^{[2]})$ and again $\ell = \log(a^{[2]})$. What are the derivatives of ℓ with respect to $\beta^{[1],[2]}$ and $\alpha^{[1],[2]}$. For each variable, when will they be zero? Give some sufficient conditions in terms of the $z^{[1]}$ or $z^{[2]}$ variables. Also, draw the computational graph.

Response Given the functions

$$\begin{aligned} a[0] &= x \\ z[1] &= \beta[1]a[0] + \alpha[1] \\ a[1] &= \sigma(z[1]) \\ z[2] &= z[1] + \beta[2]a[1] + \alpha[2] \\ a[2] &= \sigma(z[2]) \\ l &= log(a[2]) \end{aligned}$$

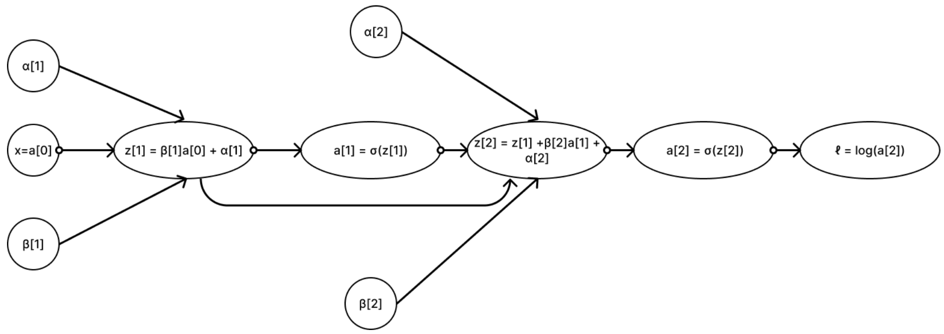

Figure 1: The computational graph of problem 5

Then we can find the derivatives:

$$\frac{\partial l}{\partial z[2]} = (1 - z[2])$$

$$\frac{\partial z[2]}{\partial \alpha[2]} = 1$$

$$\frac{\partial z[2]}{\partial \beta[2]} = a[1]$$

$$\frac{\partial z[1]}{\partial \alpha[1]} = 1$$

$$\frac{\partial a[1]}{\partial z[1]} = \sigma(z[1])(1 - \sigma(z[1]))$$

$$\frac{\partial z[1]}{\partial \beta[1]} = a[0]$$

Based on the above, we can calculate the derivatives of l:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta[1]} &= \frac{\partial \ell}{\partial a[2]} \cdot \frac{\partial a[2]}{\partial z[2]} \cdot \frac{\partial z[2]}{\partial z[1]} \cdot \frac{\partial z[1]}{\partial \beta[1]} + \frac{\partial \ell}{\partial z[2]} \cdot \frac{\partial z[2]}{\partial a[1]} \cdot \frac{\partial a[1]}{\partial z[1]} \cdot \frac{\partial z[1]}{\partial \beta[1]} \\ &= (1 - \sigma z[2])a[0] + (1 - \sigma z[2])\beta[2]\sigma(z[1])(1 - \sigma(z[1]))a[0]\end{aligned}$$

$$\frac{\partial \ell}{\partial \beta[2]} = \frac{\partial \ell}{\partial a[2]} \cdot \frac{\partial a[2]}{\partial z[2]} \cdot \frac{\partial z[2]}{\partial \beta[2]}$$
$$= (1 - \sigma z[2])\sigma z[1]$$

$$\frac{\partial \ell}{\partial \alpha[1]} = \frac{\partial \ell}{\partial a[2]} \cdot \frac{\partial a[2]}{\partial z[2]} \cdot \frac{\partial z[2]}{\partial z[1]} \cdot \frac{\partial z[1]}{\partial \alpha[1]} + \frac{\partial \ell}{\partial z[2]} \cdot \frac{\partial z[2]}{\partial a[1]} \cdot \frac{\partial a[1]}{\partial z[1]} \frac{\partial z[1]}{\partial \alpha[1]}$$
$$= (1 - \sigma(z[2])) + (1 - \sigma(z[2])) \cdot \beta[2]\sigma z[1](1 - \sigma(z[1]))$$

$$\frac{\partial \ell}{\partial \alpha[2]} = \frac{\partial \ell}{\partial a[2]} \cdot \frac{\partial a[2]}{\partial z[2]} \cdot \frac{\partial z[2]}{\partial \alpha[2]}$$
$$= (1 - \sigma z[2])$$

When will it be zero?
$\frac{\partial \ell}{\partial \beta[1]}$ will be zero if $z[2] \to \infty$ or $1 + \beta[2]\sigma(z[1])(1 - \sigma(z[1])) = 0$
$\frac{\partial \ell}{\partial \beta[2]}$ will be zero if $z[1] \to -\infty$ or $z[2] \to \infty$.
$\frac{\partial \ell}{\partial \alpha[1]}$ will be zero if $z[2] \to \infty$ or $1 + \beta[2]\sigma(z[1])(1 - \sigma(z[1])) = 0$
$\frac{\partial \ell}{\partial \alpha[2]}$ will be zero if $z[2] \to \infty$.

Problem 6

Suppose we have equations $u^{[3]} = u^{[1]} \times u^{[2]}$, $u^{[4]} = u^{[1]} + u^{[2]}$ and $u^{[5]} = 2 \times u^{[3]} \times u^{[4]}$. Suppose $\mathcal{L} = (u^{[5]} - u^{[1]} - u^{[2]} - u^{[3]} - u^{[4]})^2$ is what we'd like to minimize. Draw the computational graph. Suppose $u^{[1]} = 3$ and $u^{[2]} = 4$, find the partials $\{\frac{\partial \mathcal{L}}{\partial u^{[i]}}\}_{i=1}^{2}$. Write expressions for these partials in terms of local partials. To get the numerical answer: if you want, you can submit PyTorch code with this which sets up the above as tensors and then gets the gradients.

Response Given the equations:

$$u^{[3]} = u^{[1]} \times u^{[2]}$$
$$u^{[4]} = u^{[1]} + u^{[2]}$$
$$u^{[5]} = 2 \times u^{[3]} \times u^{[4]}$$
$$\mathcal{L} = (u^{[5]} - u^{[1]} - u^{[2]} - u^{[3]} - u^{[4]})^2$$

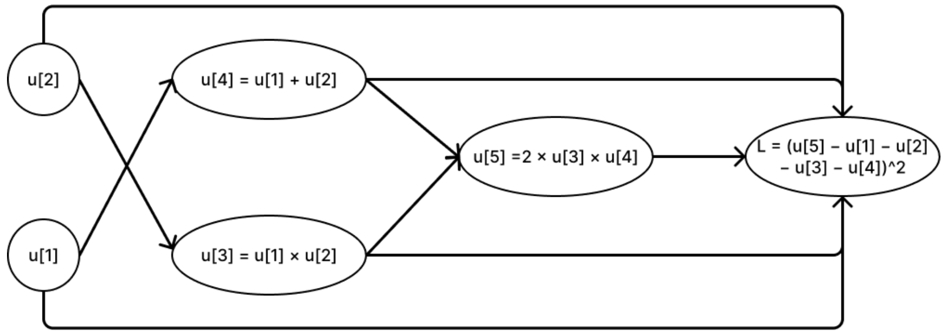

Figure 2: The computational graph for problem 6

Using the chain rule, the partial derivatives are expressed as:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial u^{[1]}} &= \frac{\partial(2u_1u_2(u_1+u_2) - u_1 - u_2 - u_1u_2 - u_1 - u_2)^2}{\partial u_1} \\ &= 2(2u_1^2u_2 + 2u_1u_2^2 - 2u_1 - 2u_2 - u_1u_2)(4u_1u_2 + 2u_2^2 - 2 - u_2) \\ &= 21016\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial u^{[2]}} &= \frac{\partial(2u_1u_2(u_1+u_2) - u_1 - u_2 - u_1u_2 - u_1 - u_2)^2}{\partial u_1} \\ &= 2(2u_1^2u_2 + 2u_1u_2^2 - 2u_1 - 2u_2 - u_1u_2)(4u_1u_2 + 2u_1^2 - 2 - u_1) \\ &= 17324\end{aligned}$$