The Austalian National University
School of Computing
Dylan Campbell

Semester 1, 2024
Tutorial 5

COMP/ENGN 4528/6528: Computer Vision

Question 1 Matrix Algebra

1. Let $\mathbf{A} = \begin{bmatrix} 2 & 6 \\ 2 & 0 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 0 & 1 \\ 2 & 8 \end{bmatrix}$, compute $\mathbf{AB}$.

   Solution.

$$\mathbf{AB} = \begin{bmatrix} 2 & 6 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 2 & 8 \end{bmatrix} = \begin{bmatrix} 12 & 50 \\ 0 & 2 \end{bmatrix}$$

2. Let $\mathbf{x} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 0 \\ 8 \end{bmatrix}$, compute $\|\mathbf{x} - \mathbf{y}\|_2$.

   Solution.

$$\|\mathbf{x} - \mathbf{y}\|_2 = \| \begin{bmatrix} 5 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 8 \end{bmatrix} \|_2 = \| \begin{bmatrix} 5 \\ -7 \end{bmatrix} \|_2 = \sqrt{5^2 + (-7)^2} = \sqrt{74} = 8.602$$

3. Let $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$ and $\mathcal{L} = \frac{1}{2}(\mathbf{w}^\top \mathbf{x} - 4)^2$, compute $\frac{\partial \mathcal{L}}{\partial \mathbf{w}}$ [1].

   Solution.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_1} &= \frac{\partial(\frac{1}{2}(8w_1 + 6w_2 - 4)^2)}{\partial w_1} \\ &= \frac{1}{2} \times 2 \times (8w_1 + 6w_2 - 4) \times 8 \\ &= 64w_1 + 48w_2 - 32 \end{aligned}$$

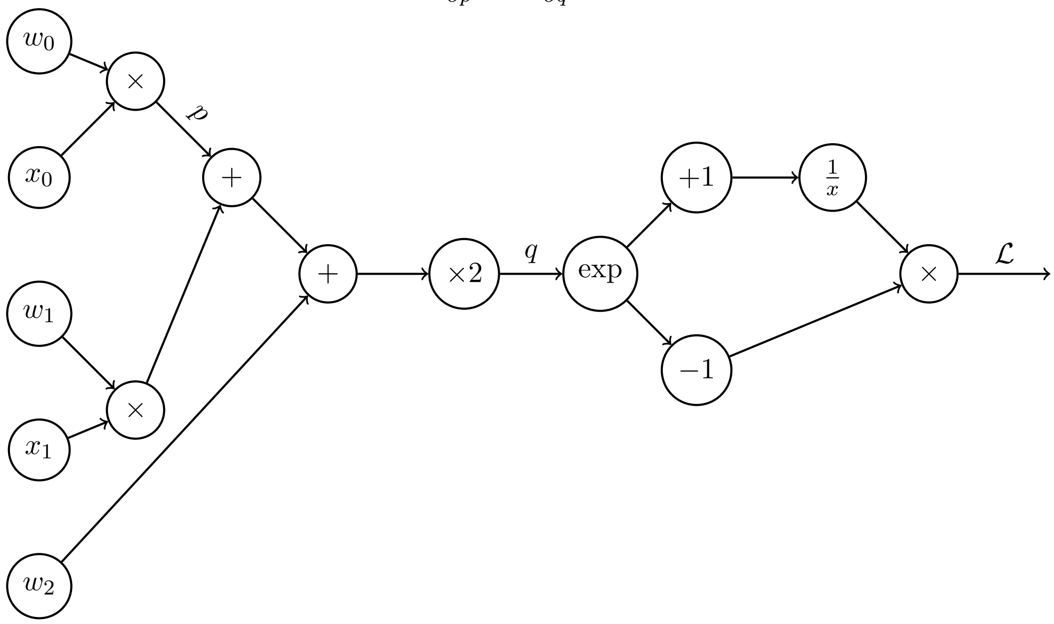
$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_2} &= \frac{\partial(\frac{1}{2}(8w_1 + 6w_2 - 4)^2)}{\partial w_2} \\ &= \frac{1}{2} \times 2 \times (8w_1 + 6w_2 - 4) \times 6 \\ &= 48w_1 + 36w_2 - 24 \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} &= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_1} & \frac{\partial \mathcal{L}}{\partial w_2} \end{bmatrix} \\ &= \begin{bmatrix} (64w_1 + 48w_2 - 32) & (48w_1 + 36w_2 - 24) \end{bmatrix} \end{aligned}$$

Question 2 Back Propagation

[1] read `https://en.wikipedia.org/wiki/Matrix_calculus` for more matrix calculus contents

1. Back propagation through the computational graph. The current values are $w_0 = 0.2$, $w_1 = 0.2$, $w_2 = 0.3$, $x_0 = 2$, $x_1 = 3$. p and q define the intermediate variables that are calculated during training, at the specified points in the computation graph. $\mathcal{L}$ is the output of the computational graph. Please provide the gradient $\frac{\partial \mathcal{L}}{\partial p}$ and $\frac{\partial \mathcal{L}}{\partial q}$ based on the back-propagated gradient calculation.



Solution. First, we can express $\mathcal{L}$ with q.

$$\begin{aligned}\mathcal{L} &= \frac{1}{e^q+1} \times (e^q - 1)\\ &= \frac{e^q-1}{e^q+1}\end{aligned}$$

Then, we can calculate the value of q.

$$\begin{aligned}q &= (w_0x_0 + w_1x_1 + w_2) \times 2\\ &= (0.2 \times 2 + 0.2 \times 3 + 0.3) \times 2\\ &= 2.6\end{aligned}$$

So, we can calculate $\frac{\partial \mathcal{L}}{\partial q}$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial q} &= \frac{\partial\left(\frac{e^q-1}{e^q+1}\right)}{\partial q}\\ &= \frac{\partial\left(\frac{e^q-1}{e^q+1}\right)}{\partial e^q} \cdot \frac{\partial e^q}{\partial q}\\ &= \frac{(e^q+1)-(e^q-1)}{(e^q+1)^2} \times e^q\\ &= \frac{2e^q}{(e^q+1)^2}\\ &= \frac{2e^{2.6}}{(e^{2.6}+1)^2} = 0.1287\end{aligned}$$

q can be expressed with p.

$$\begin{aligned}q &= (p + w_1x_1 + w_2) \times 2\\ &= 2p + 2w_1x_1 + 2w_2\end{aligned}$$

$\frac{\partial \mathcal{L}}{\partial p}$ can be calculated as follows.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p} &= \frac{\partial \mathcal{L}}{\partial q} \cdot \frac{\partial q}{\partial p} \\ &= \frac{2e^{2.6}}{(e^{2.6}+1)^2} \cdot \frac{\partial(2p + 2w_1x_2 + 2w_2)}{\partial p} \\ &= \frac{2e^{2.6}}{(e^{2.6}+1)^2} \times 2 = 0.2574\end{aligned}$$

2. Given the linear regression model $\hat{y} = \mathbf{w}^\top \mathbf{x} + b$ and the loss function is defined as $\mathcal{L}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$. The initial model weights are $\mathbf{w} = \begin{bmatrix} 6 \\ -4 \end{bmatrix}$ and $b = -8$. What is the new model weights after performing one gradient descent step with learning rate 0.01 and training data $\mathbf{x} = \begin{bmatrix} 8 \\ 1 \end{bmatrix}$, $y = 1$.

Solution. We first do the forward pass.

$$\begin{aligned}\hat{y} &= \mathbf{w}^\top \mathbf{x} + b \\ &= \begin{bmatrix} 6 & -4 \end{bmatrix} \begin{bmatrix} 8 \\ 1 \end{bmatrix} - 8 \\ &= 48 - 4 - 8 \\ &= 36\end{aligned}$$

Then, we calculate the gradient of the loss with respect to $\mathbf{w}$ and b.

$$\begin{aligned}\nabla_{\mathbf{w}} \mathcal{L} &= \left(\frac{\partial \mathcal{L}}{\partial \mathbf{w}}\right)^\top \\ &= \left(\frac{\partial(\frac{1}{2}(y - \hat{y})^2)}{\partial \mathbf{w}}\right)^\top \\ &= \left(\frac{\partial(\frac{1}{2}(y - \mathbf{w}^\top \mathbf{x} - b)^2)}{\partial \mathbf{w}}\right)^\top \\ &= \begin{bmatrix} \frac{\partial(\frac{1}{2}(y - w_0x_0 - w_1x_1 - b)^2)}{\partial w_0} & \frac{\partial(\frac{1}{2}(y - w_0x_0 - w_1x_1 - b)^2)}{\partial w_1} \end{bmatrix}^\top \\ &= \begin{bmatrix} -x_0(y - \hat{y}) \\ -x_1(y - \hat{y}) \end{bmatrix} \\ &= \begin{bmatrix} -8 \times (1 - 36) \\ -1 \times (1 - 36) \end{bmatrix} \\ &= \begin{bmatrix} 280 \\ 35 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\nabla_b \mathcal{L} &= \left(\frac{\partial \mathcal{L}}{\partial b}\right)^\top \\ &= \left(\frac{\partial(\frac{1}{2}(y - \hat{y})^2)}{\partial b}\right)^\top \\ &= \left(\frac{\partial(\frac{1}{2}(y - \mathbf{w}^\top \mathbf{x} - b)^2)}{\partial b}\right)^\top \\ &= -(y - \mathbf{w}^\top \mathbf{x} - b) \\ &= -(1 - 36) = 35\end{aligned}$$

After calculating gradients, we can perform gradient descent and obtain new model weights.

$$\begin{aligned}
\mathbf{w}' &= \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L} \\
&= \begin{bmatrix} 6 \\ -4 \end{bmatrix} - 0.01 \begin{bmatrix} 280 \\ 35 \end{bmatrix} \\
&= \begin{bmatrix} 6 - 2.8 \\ -4 - 0.35 \end{bmatrix} \\
&= \begin{bmatrix} 3.2 \\ -4.35 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
b' &= b - \eta \nabla_b \mathcal{L} \\
&= -8 - 0.01 \times 35 \\
&= -8.35
\end{aligned}$$

3. **!!** Given a logistic regression model Softmax($\mathbf{Wx} + \mathbf{b}$) where $\mathbf{W} = \begin{bmatrix} 2 & 0 & 4 \\ 2 & 2 & 6 \\ -1 & 1 & 4 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} -2 \\ -1 \\ -2 \end{bmatrix}$. The Softmax function is defined as $\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$. For the training data, you have $\mathbf{x} = \begin{bmatrix} -2 \\ -5 \\ 4 \end{bmatrix}$ and the ground truth label $\mathbf{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Calculate the gradient of the cross-entropy loss ($\mathcal{L} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$)[2] with respect to the bias vector $\mathbf{b}$.

Solution. We first calculate the forward pass.

$$\begin{aligned}
\mathbf{z} &= \text{Softmax}(\mathbf{Wx} + \mathbf{b}) \\
&= \text{Softmax}(\begin{bmatrix} 2 & 0 & 4 \\ 2 & 2 & 6 \\ -1 & 1 & 4 \end{bmatrix} \begin{bmatrix} -2 \\ -5 \\ 4 \end{bmatrix} + \begin{bmatrix} -2 \\ -1 \\ -2 \end{bmatrix}) \\
&= \text{Softmax}(\begin{bmatrix} 10 \\ 9 \\ 11 \end{bmatrix})
\end{aligned}$$

The loss $\mathcal{L}$ can be calculated as follows.

$$\begin{aligned}
\mathcal{L} &= -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \\
&= -(0 \times \log(z_1) + 0 \times \log(z_2) + 1 \times \log(z_3)) \\
&= -\log(z_3)
\end{aligned}$$

Then, we can start trying to calculate the gradient of $\mathcal{L}$ w.r.t $\mathbf{b}$.

$$\begin{aligned}
\nabla_{\mathbf{b}} \mathcal{L} &= (\frac{\partial \mathcal{L}}{\partial \mathbf{b}})^\top \\
&= (\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial b_1} & \frac{\partial \mathcal{L}}{\partial b_2} & \frac{\partial \mathcal{L}}{\partial b_3} \end{bmatrix})^\top
\end{aligned}$$

[2]The meaning of each variable is mentioned in `https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html#cross-entropy`

Note that $\frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial \mathcal{L}}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_1} + \frac{\partial \mathcal{L}}{\partial z_2} \cdot \frac{\partial z_2}{\partial b_1} + \frac{\partial \mathcal{L}}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_1}$. However, as z_1 and z_2 do not appear in $\mathcal{L}$, we can simplify the original expression $\frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial \mathcal{L}}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_1}$. Then, we can rewrite the original gradient expression.

$$\begin{aligned}
\nabla_{\mathbf{b}} \mathcal{L} &= (\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial b_1} & \frac{\partial \mathcal{L}}{\partial b_2} & \frac{\partial \mathcal{L}}{\partial b_3} \end{bmatrix})^\top \\
&= (\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_1} & \frac{\partial \mathcal{L}}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_2} & \frac{\partial \mathcal{L}}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_3} \end{bmatrix})^\top \\
&= (\begin{bmatrix} \frac{\partial(-\log(z_3))}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_1} & \frac{\partial(-\log(z_3))}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_2} & \frac{\partial(-\log(z_3))}{\partial z_3} \cdot \frac{\partial z_3}{\partial b_3} \end{bmatrix})^\top \\
&= (\begin{bmatrix} -\frac{1}{z_3} \cdot \frac{\partial z_3}{\partial b_1} & -\frac{1}{z_3} \cdot \frac{\partial z_3}{\partial b_2} & -\frac{1}{z_3} \cdot \frac{\partial z_3}{\partial b_3} \end{bmatrix})^\top
\end{aligned}$$

Then, we need to calculate the partial derivative of z_3 w.r.t b_1. b_2 and b_3.

$$\begin{aligned}
\frac{\partial z_3}{\partial b_1} &= \frac{\partial(\frac{\exp((\mathbf{Wx}+\mathbf{b})_3)}{\exp((\mathbf{Wx}+\mathbf{b})_1)+\exp((\mathbf{Wx}+\mathbf{b})_2)+\exp((\mathbf{Wx}+\mathbf{b})_3)})}{\partial b_1} \\
&= \frac{-\exp((\mathbf{Wx}+\mathbf{b})_3)\exp((\mathbf{Wx}+\mathbf{b})_1)}{(\exp((\mathbf{Wx}+\mathbf{b})_1)+\exp((\mathbf{Wx}+\mathbf{b})_2)+\exp((\mathbf{Wx}+\mathbf{b})_3))^2} \\
&= -\text{Softmax}(\mathbf{Wx}+\mathbf{b})_3 \times \text{Softmax}(\mathbf{Wx}+\mathbf{b})_1 \\
&= -z_3 z_1
\end{aligned}$$

$$\begin{aligned}
\frac{\partial z_3}{\partial b_2} &= \frac{\partial(\frac{\exp((\mathbf{Wx}+\mathbf{b})_3)}{\exp((\mathbf{Wx}+\mathbf{b})_1)+\exp((\mathbf{Wx}+\mathbf{b})_2)+\exp((\mathbf{Wx}+\mathbf{b})_3)})}{\partial b_2} \\
&= \frac{-\exp((\mathbf{Wx}+\mathbf{b})_3)\exp((\mathbf{Wx}+\mathbf{b})_2)}{(\exp((\mathbf{Wx}+\mathbf{b})_1)+\exp((\mathbf{Wx}+\mathbf{b})_2)+\exp((\mathbf{Wx}+\mathbf{b})_3))^2} \\
&= -\text{Softmax}(\mathbf{Wx}+\mathbf{b})_3 \times \text{Softmax}(\mathbf{Wx}+\mathbf{b})_2 \\
&= -z_3 z_2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial z_3}{\partial b_3} &= \frac{\partial(\frac{\exp((\mathbf{Wx}+\mathbf{b})_3)}{\exp((\mathbf{Wx}+\mathbf{b})_1)+\exp((\mathbf{Wx}+\mathbf{b})_2)+\exp((\mathbf{Wx}+\mathbf{b})_3)})}{\partial b_3} \\
&= \frac{\exp((\mathbf{Wx}+\mathbf{b})_3) \times \sum_{i=1}^{3}\exp((\mathbf{Wx}+\mathbf{b})_i) - (\exp((\mathbf{Wx}+\mathbf{b})_3))^2}{(\sum_{i=1}^{3}\exp((\mathbf{Wx}+\mathbf{b})_i))^2} \\
&= \frac{\exp((\mathbf{Wx}+\mathbf{b})_3)}{\sum_{i=1}^{3}\exp((\mathbf{Wx}+\mathbf{b})_i)} - \frac{(\exp((\mathbf{Wx}+\mathbf{b})_3))^2}{(\sum_{i=1}^{3}\exp((\mathbf{Wx}+\mathbf{b})_i))^2} \\
&= \text{Softmax}(\mathbf{Wx}+\mathbf{b})_3 - (\text{Softmax}(\mathbf{Wx}+\mathbf{b})_3)^2 \\
&= z_3 - z_3^2 \\
&= z_3(1-z_3)
\end{aligned}$$

Put these results back to the gradient of $\mathcal{L}$ w.r.t $\mathbf{b}$.

$$\begin{aligned}
\nabla_{\mathbf{b}}\mathcal{L} &= \left(\begin{bmatrix}-\frac{1}{z_3}\cdot\frac{\partial z_3}{\partial b_1} & -\frac{1}{z_3}\cdot\frac{\partial z_3}{\partial b_2} & -\frac{1}{z_3}\cdot\frac{\partial z_3}{\partial b_3}\end{bmatrix}\right)^\top \\
&= \left(\begin{bmatrix}-\frac{1}{z_3}\times(-z_3 z_1) & -\frac{1}{z_3}\times(-z_3 z_2) & -\frac{1}{z_3}\times(z_3\times(1-z_3))\end{bmatrix}\right)^\top \\
&= \left(\begin{bmatrix}z_1 & z_2 & (z_3-1)\end{bmatrix}\right)^\top \\
&= \begin{bmatrix}\text{Softmax}(\mathbf{Wx}+\mathbf{b})_1 \\ \text{Softmax}(\mathbf{Wx}+\mathbf{b})_2 \\ \text{Softmax}(\mathbf{Wx}+\mathbf{b})_3 - 1\end{bmatrix} \\
&= \begin{bmatrix}\frac{e^{10}}{e^{10}+e^9+e^{11}} \\ \frac{e^9}{e^{10}+e^9+e^{11}} \\ \frac{e^{11}}{e^{10}+e^9+e^{11}} - 1\end{bmatrix} = \begin{bmatrix}0.2447 \\ 0.0900 \\ -0.3348\end{bmatrix}
\end{aligned}$$