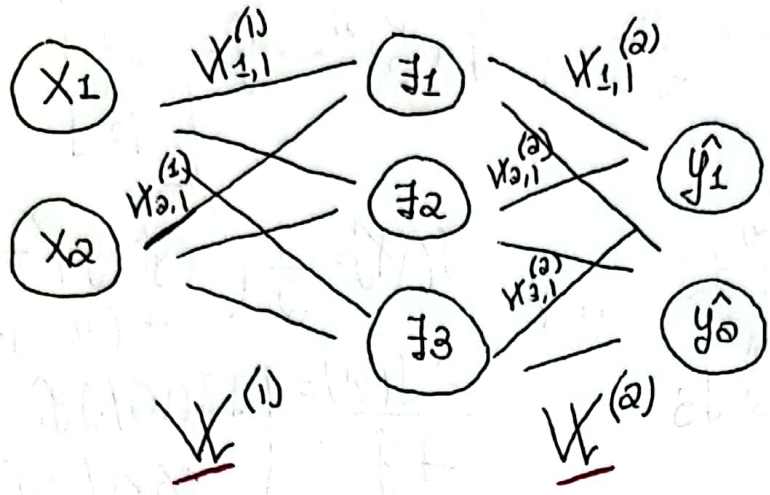
Anar Rzayev — MA8473: Homework Set 4 — ID: 20190188

4)



Fully-connected network

$$z_j = \sigma\left(\sum_{i=1}^{2} W_{i,j}^{(1)} x_i + b_j\right)$$

$$\hat{y}_k = \sum_{i=1}^{2} W_{i,k}^{(2)} z_i + b_k$$

$$J(W) = \frac{1}{2}\|y-\hat{y}\|_2^2$$

Vectorizing the above implementation, we can write the followings:

$$J(W) = \frac{1}{2}\|\hat{y}-y\|_2^2$$

$$z = \sigma\left(W^{(1)}x + b^{(1)}\right)$$

$$\hat{y} = W^{(2)}z + b^{(2)}$$

Now, it's time to proceed with backward propagation using correspondent gradients

$$\frac{\partial J(W)}{\partial W^{(2)}} = \frac{\partial J(W)}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial W^{(2)}}$$

Since $J(W) = \frac{1}{2}\|\hat{y}-y\|_2^2 \Rightarrow \boxed{\frac{\partial J(W)}{\partial \hat{y}} = \hat{y}-y}$ and $\hat{y} = W^{(2)}z + b^{(2)}$, it's easily seen that from matrix differentiation

$\boxed{\frac{\partial \hat{y}}{\partial W^{(2)}} = z^T}$ Thus, $\boxed{\frac{\partial J(W)}{\partial W^{(2)}} = (\hat{y}-y)z^T}$ ☆ Now, we proceed to evaluate gradient with respect to $b^{(2)}$. Using the chain Rule $\Rightarrow$

$$\frac{\partial J(W)}{\partial b^{(2)}} = \frac{\partial J(W)}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial b^{(2)}}$$

where $\frac{\partial J(W)}{\partial \hat{y}} = \hat{y}-y$ and $\frac{\partial \hat{y}}{\partial b^{(2)}} = 1^T$ using the formula

$\hat{y} = W^{(2)}z + b^{(2)}$ $\left(\frac{\partial \hat{y}}{\partial b^{(2)}} = \begin{bmatrix} \partial \hat{y}_1/\partial b_1^{(2)} \\ \partial \hat{y}_2/\partial b_2^{(2)} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)$ More specifically, if we evaluate $\frac{\partial \hat{y}}{\partial b^{(2)}}$ explicitly

$$\frac{\partial \hat{y}}{\partial b^{(2)}} = \begin{bmatrix} \partial \hat{y}_1/\partial b_1^{(2)} & \partial \hat{y}_1/\partial b_2^{(2)} \\ \partial \hat{y}_2/\partial b_1^{(2)} & \partial \hat{y}_2/\partial b_2^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I \Rightarrow \boxed{\frac{\partial J(W)}{\partial b^{(2)}} = \hat{y}-y}$$ ☆ Now

Scanned with CamScanner

Our aim is to evaluate gradient wrt vector-z $\Rightarrow$ Using the chain-rule,

$$\frac{\partial J(W)}{\partial z} = \frac{\partial J(W)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \quad \text{where } \hat{y} = W^{(2)} z + b^{(2)} \rightsquigarrow \hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}$$

$$\hat{y}_1 = W^{(2)}_{1,1} z_1 + W^{(2)}_{2,1} z_2 + W^{(2)}_{3,1} z_3$$

$$\hat{y}_2 = W^{(2)}_{1,2} z_1 + W^{(2)}_{2,2} z_2 + W^{(2)}_{3,2} z_3$$

and

$$J(W) = \frac{1}{2}\left((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 \right)$$

$$\frac{\partial J(W)}{\partial \hat{y}} = \begin{bmatrix} \partial J(W)/\partial \hat{y}_1 \\ \partial J(W)/\partial \hat{y}_2 \end{bmatrix}$$

$$\frac{\partial J(W)}{\partial \hat{y}} = \begin{bmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \end{bmatrix} \in \mathbb{R}^{2\times 1} \quad \text{and} \quad \hat{y} = \begin{bmatrix} W^{(2)}_{1,1} z_1 + W^{(2)}_{2,1} z_2 + W^{(2)}_{3,1} z_3 \\ W^{(2)}_{1,2} z_1 + W^{(2)}_{2,2} z_2 + W^{(2)}_{3,2} z_3 \end{bmatrix}$$

$$\frac{\partial \hat{y}}{\partial z} = \begin{bmatrix} \partial \hat{y}_1/\partial z_1 & \partial \hat{y}_1/\partial z_2 & \partial \hat{y}_1/\partial z_3 \\ \partial \hat{y}_2/\partial z_1 & \partial \hat{y}_2/\partial z_2 & \partial \hat{y}_2/\partial z_3 \end{bmatrix} = \begin{bmatrix} W^{(2)}_{1,1} & W^{(2)}_{2,1} & W^{(2)}_{3,1} \\ W^{(2)}_{1,2} & W^{(2)}_{2,2} & W^{(2)}_{3,2} \end{bmatrix}$$

where we know that

$$W^{(2)} = \begin{bmatrix} W^{(2)}_{1,1} & W^{(2)}_{1,2} \\ W^{(2)}_{2,1} & W^{(2)}_{2,2} \\ W^{(2)}_{3,1} & W^{(2)}_{3,2} \end{bmatrix}^T \rightsquigarrow \boxed{\frac{\partial \hat{y}}{\partial z} = W^{(2)}}$$

$W^{(2)} \in \mathbb{R}^{2\times 3}$, $\hat{y} - y \in \mathbb{R}^{2\times 1}$ — using the dimensions matching,

one can obtain that $\frac{\partial J(W)}{\partial z} = \frac{\partial J(W)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} = (W^{(2)})^T (\hat{y} - y) \in \mathbb{R}^{3\times 1}$

$$\boxed{\frac{\partial J(W)}{\partial z} = (W^{(2)})^T (\hat{y} - y)}$$

Now, we have to move with the gradient wrt $W^{(1)}$

$$\frac{\partial J(W)}{\partial W^{(1)}} = \frac{\partial J(W)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial W^{(1)}}$$

where we had

$$\frac{\partial J(W)}{\partial \hat{y}} = \hat{y} - y \in \mathbb{R}^{2\times 1}$$

$$\frac{\partial \hat{y}}{\partial z} = W^{(2)} \in \mathbb{R}^{2\times 3}$$

$z = \sigma(W^{(1)} x + b^{(1)})$ More equivalently, we can rewrite

$$\frac{\partial J(W)}{\partial W^{(1)}} = \frac{\partial J(W)}{\partial z} \cdot \frac{\partial z}{\partial W^{(1)}} \quad \text{with} \quad \frac{\partial J(W)}{\partial z} = (W^{(2)})^T (\hat{y} - y) \in \mathbb{R}^{3\times 1}$$

Scanned with CamScanner

Before tackling the gradient wrt $W^{(1)}$, it's important to define necessary terminologies so that we could obtain important results related to the gradient for $W^{(1)}$:

$\delta :=$ column-vector consisting of only zeros and ones in which

$$\delta_i = \begin{cases} 0, & \text{if } (W^{(1)}x + b^{(1)})_i \leq 0 \\ 1, & \text{otherwise} \end{cases}$$ (gradient of RELU giving effect of 1)

Leveraging the potential of those column-vectors, we'll find the gradient in which $z = \mathrm{ReLU}(\underbrace{W^{(1)}x + b^{(1)}}_{z'}) = \mathrm{ReLU}(z') \rightsquigarrow \frac{\partial J(W)}{\partial W^{(1)}} =$

$$= \frac{\partial J(W)}{\partial z} \cdot \frac{\partial z}{\partial W^{(1)}} = \frac{\partial J(W)}{\partial z} \frac{\partial z}{\partial z'} \frac{\partial z'}{\partial W^{(1)}}, \text{ considering } \frac{\partial J(W)}{\partial z} =$$

$$= (W^{(2)})^T(\hat{y} - y) \in \mathbb{R}^{3\times 1}, \quad \boxed{\frac{\partial z}{\partial z'} = \delta \in \mathbb{R}^{3\times 1}}$$ $\rightsquigarrow$ gradients wrt ReLU giving only 1's and 0's (aforementioned column-vector)

Since $z' = W^{(1)}x + b^{(1)} \Rightarrow \frac{\partial z'}{\partial W^{(1)}} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T = [x_1 \; x_2] \in \mathbb{R}^{1\times 2}$, therefore

$$\boxed{\frac{\partial z'}{\partial W^{(1)}} = x^T \in \mathbb{R}^{1\times 2}}$$

If we take into account the Hadamard operation (piecewise multiplication) for $(W^{(2)})^T(\hat{y} - y)$ and δ, where dimensions match as $\mathbb{R}^{3\times 1}$, we obtain

$$\boxed{\frac{\partial J(W)}{\partial W^{(1)}} = \left((W^{(2)})^T(\hat{y} - y) \odot \delta\right) x^T \in \mathbb{R}^{3\times 2}}$$

Using exactly similar gradients wrt $b^{(1)}$, chain-rule reveals that

$$\frac{\partial J(W)}{\partial b^{(1)}} = \frac{\partial J(W)}{\partial z} \frac{\partial z}{\partial b^{(1)}} = \frac{\partial J(W)}{\partial z} \frac{\partial z}{\partial z'} \frac{\partial z'}{\partial b^{(1)}} = (W^{(2)})^T(\hat{y} - y) \odot \delta$$

and since we also know considering $\frac{\partial z}{\partial z'} = \delta \in \mathbb{R}^{3\times 1}$, $(W^{(2)})^T(\hat{y} - y) \in \mathbb{R}^{3\times 1}$ (Hadamard operation giving correct calculation)

Scanned with CamScanner

$z^1 = W^{(1)}x + b^{(1)} \Rightarrow$ We conclude that gradient of $J(W)$ wrt $b^{(1)}$, is equal

$$\boxed{\frac{\partial J(W)}{\partial b^{(1)}} = (W^{(2)})^T(\hat{y}-y)\odot \delta}$$

In conclusion, the followings are final gradients to be calculated:

$$\frac{\partial J(W)}{\partial W^{(2)}} = (\hat{y}-y)z^T$$

$$\frac{\partial J(W)}{\partial W^{(1)}} = \left((W^{(2)})^T(\hat{y}-y)\odot\delta\right)x^T$$

$$\frac{\partial J(W)}{\partial b^{(2)}} = \hat{y}-y$$

$$\frac{\partial J(W)}{\partial b^{(1)}} = (W^{(2)})^T(\hat{y}-y)\odot\delta$$

Now, we will use the gradient descent approach with learning rate $\eta=1$

The gradient descent method illustrates the following updates (learning rate=1)

$$W^{(2)} := W^{(2)} - \frac{\partial J(W)}{\partial W^{(2)}} \qquad W^{(1)} := W^{(1)} - \frac{\partial J(W)}{\partial W^{(1)}}$$

$$b^{(2)} := b^{(2)} - \frac{\partial J(W)}{\partial b^{(2)}} \qquad b^{(1)} := b^{(1)} - \frac{\partial J(W)}{\partial b^{(1)}}$$

Given the dataset $D = \left\{ \begin{bmatrix} 2 \\ -3 \end{bmatrix}, \begin{bmatrix} -4 \\ 0 \end{bmatrix} \right\}$ with input $\begin{bmatrix} 2 \\ -3 \end{bmatrix}$ and target variable $\begin{bmatrix} -4 \\ 0 \end{bmatrix}$

our aim is to iterate forward-backward procedures for twice processes:

i) Plugging in $x = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$, $W^{(1)}x + b^{(1)} = \begin{bmatrix} 1 & -1 \\ 0 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ -3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} =$

$= \begin{bmatrix} 5 \\ 3 \\ -5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ -4 \end{bmatrix}$, applying ReLU $\rightsquigarrow \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix}$ and $W^{(2)}x' + b^{(2)}$

$= \begin{bmatrix} 0 & -2 & 1 \\ 1 & -1 & -1 \end{bmatrix}\begin{bmatrix} 5 \\ 3 \\ [illegible] \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -6 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -5 \\ 2 \end{bmatrix}$ where

we used $\delta = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ during the process of forward-propagation $\left(y = \begin{bmatrix} -4 \\ 0 \end{bmatrix}\right)$

Scanned with CamScanner

$\underline{1^{st}\text{ iteration:}}$ $W^{(2)} := W^{(2)} - \frac{\partial J(W)}{\partial W^{(2)}}$ where $\frac{\partial J(W)}{\partial W^{(2)}} = (\hat{y}-y)\, f^T =$

$$= \left(\begin{bmatrix} -5 \\ 2 \end{bmatrix} - \begin{bmatrix} -4 \\ 0 \end{bmatrix} \right) \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix}^T = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \begin{bmatrix} 5 & 3 & 0 \end{bmatrix} = \begin{bmatrix} -5 & -3 & 0 \\ 10 & 6 & 0 \end{bmatrix}$$

Thus, $W^{(2)} \rightsquigarrow \begin{bmatrix} 0 & -2 & 1 \\ 1 & -1 & -1 \end{bmatrix} - \begin{bmatrix} -5 & -3 & 0 \\ 10 & 6 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} 5 & 1 & 1 \\ -9 & -7 & -1 \end{bmatrix}}$

$b^{(2)} := b^{(2)} - \frac{\partial J(W)}{\partial b^{(2)}}$, where $\frac{\partial J(W)}{\partial b^{(2)}} = \hat{y} - y = \begin{bmatrix} -5 \\ 2 \end{bmatrix} - \begin{bmatrix} -4 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$

$b^{(2)} \rightsquigarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \underbrace{\begin{bmatrix} 2 \\ -2 \end{bmatrix}}$

$W^{(1)} := W^{(1)} - \frac{\partial J(W)}{\partial W^{(1)}}$ in which $\frac{\partial J(W)}{\partial W^{(1)}} = \left((W^{(2)})^T (\hat{y}-y) \odot f \right) x^T$

Since $W^{(1)}x + b^{(1)} = \begin{bmatrix} 5 \\ 3 \\ -4 \end{bmatrix} \Rightarrow f = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ and thus, $(W^{(2)})^T(\hat{y}-y) =$

$$= \begin{bmatrix} 0 & 1 \\ -2 & -1 \\ 1 & -1 \end{bmatrix} \left(\begin{bmatrix} -5 \\ 2 \end{bmatrix} - \begin{bmatrix} -4 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & 1 \\ -2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -3 \end{bmatrix}, \; f = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

So, $\frac{\partial J(W)}{\partial W^{(1)}} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 2 & -3 \end{bmatrix} = \begin{bmatrix} 4 & -6 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $W^{(1)} - \frac{\partial J(W)}{\partial W^{(1)}} =$

$$= \begin{bmatrix} 1 & -1 \\ 0 & -1 \\ -1 & 1 \end{bmatrix} - \begin{bmatrix} 4 & -6 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} -3 & 5 \\ 0 & -1 \\ -1 & 1 \end{bmatrix}} \rightsquigarrow W^{(1)} \text{ is updated as follows}$$

$b^{(1)} := b^{(1)} - \frac{\partial J(W)}{\partial b^{(1)}}$ where $\frac{\partial J(W)}{\partial b^{(1)}} = (W^{(2)})^T(\hat{y}-y) \odot f = \begin{bmatrix} 2 \\ 0 \\ -3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} =$

$= [2\ 0\ 0]^T$, thus $b^{(1)} \rightsquigarrow [0\ 0\ 1]^T - [2\ 0\ 0]^T = \underbrace{\begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}}$

Scanned with CamScanner

ii) Plugging in $x = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ with new updated values $W^{(2)}, b^{(2)}, W^{(1)}, b^{(1)}$

$$W^{(1)}x + b^{(1)} = \begin{bmatrix} -3 & 5 \\ 0 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -3 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -21 \\ 3 \\ -5 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} -23 \\ 3 \\ -4 \end{bmatrix} \text{ with ReLU} \rightsquigarrow \begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix} \text{ and } f = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ - gradient effect}$$

$$W^{(2)} \begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix} + b^{(2)} = \begin{bmatrix} 5 & 1 & 1 \\ -9 & -7 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix} + \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ -21 \end{bmatrix} + \begin{bmatrix} 2 \\ -2 \end{bmatrix} =$$

$$= \underbrace{\begin{bmatrix} 5 \\ -23 \end{bmatrix}}_{} \rightsquigarrow \hat{y} \text{ and } \begin{bmatrix} -4 \\ 0 \end{bmatrix} \rightsquigarrow y$$

<u>2nd iteration:</u> $W^{(2)} := W^{(2)} - \dfrac{\partial J(W)}{\partial W^{(2)}}$ with $\dfrac{\partial J(W)}{\partial W^{(2)}} = (\hat{y} - y) z^T =$

$$= \left(\begin{bmatrix} 5 \\ -23 \end{bmatrix} - \begin{bmatrix} -4 \\ 0 \end{bmatrix} \right) \begin{bmatrix} 0 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 9 \\ -23 \end{bmatrix} \begin{bmatrix} 0 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 27 & 0 \\ 0 & -69 & 0 \end{bmatrix}$$

$$W^{(2)} \rightsquigarrow \begin{bmatrix} 5 & 1 & 1 \\ -9 & -7 & -1 \end{bmatrix} - \begin{bmatrix} 0 & 27 & 0 \\ 0 & -69 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} 5 & -26 & 1 \\ -9 & 62 & -1 \end{bmatrix}}_{}$$

$b^{(2)} := b^{(2)} - \dfrac{\partial J(W)}{\partial b^{(2)}}$ in which $\dfrac{\partial J(W)}{\partial b^{(2)}} = \hat{y} - y = \begin{bmatrix} 5 \\ -23 \end{bmatrix} - \begin{bmatrix} -4 \\ 0 \end{bmatrix} = \begin{bmatrix} 9 \\ -23 \end{bmatrix}$

$$b^{(2)} \rightsquigarrow \begin{bmatrix} 2 \\ -2 \end{bmatrix} - \begin{bmatrix} 9 \\ -23 \end{bmatrix} = \underbrace{\begin{bmatrix} -7 \\ 21 \end{bmatrix}}_{}$$

$W^{(1)} := W^{(1)} - \dfrac{\partial J(W)}{\partial W^{(1)}}$ where $\dfrac{\partial J(W)}{\partial W^{(1)}} = \left((W^{(2)})^T (\hat{y} - y) \circ f \right) x^T$

$$(W^{(2)})^T (\hat{y} - y) = \begin{bmatrix} 5 & -9 \\ 1 & -7 \\ 1 & -1 \end{bmatrix} \left(\begin{bmatrix} 5 \\ -23 \end{bmatrix} - \begin{bmatrix} -4 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 5 & -9 \\ 1 & -7 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 9 \\ -23 \end{bmatrix}$$

Scanned with CamScanner

$$(W^{(2)})^T(\hat{y}-y) = \begin{bmatrix} 45+207 \\ 9+161 \\ 9+23 \end{bmatrix} = \begin{bmatrix} 252 \\ 170 \\ 32 \end{bmatrix}, \text{ since } f = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \Rightarrow$$

$$(W^{(2)})^T(\hat{y}-y) \circ f = \begin{bmatrix} 0 \\ 170 \\ 0 \end{bmatrix} \text{ with } \frac{\partial J(W)}{\partial W^{(1)}} = \begin{bmatrix} 0 \\ 170 \\ 0 \end{bmatrix} \begin{bmatrix} 2 & -3 \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & 0 \\ 340 & -510 \\ 0 & 0 \end{bmatrix}, \quad W^{(1)} \rightsquigarrow \begin{bmatrix} -3 & 5 \\ 0 & -1 \\ -1 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 340 & -510 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -3 & 5 \\ -340 & 509 \\ -1 & 1 \end{bmatrix}$$

$$b^{(1)} := b^{(1)} - \frac{\partial J(W)}{\partial b^{(1)}} \text{ where } \frac{\partial J(W)}{\partial b^{(1)}} = (W^{(2)})^T(\hat{y}-y) \circ f = \begin{bmatrix} 252 \\ 170 \\ 32 \end{bmatrix} \circ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 170 \\ 0 \end{bmatrix}, \quad b^{(1)} \rightsquigarrow \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 170 \\ 0 \end{bmatrix} = \begin{bmatrix} -2 \\ -170 \\ 1 \end{bmatrix}$$

In conclusion, after two iterations of optimization during the training data, we obtained (H) parameters $\longrightarrow$

$$W^{(1)} = \begin{bmatrix} -3 & 5 \\ -340 & 509 \\ -1 & 1 \end{bmatrix} \quad \boxed{V} \quad \boxed{\star}$$

$$b^{(1)} = \begin{bmatrix} -2 \\ -170 \\ 1 \end{bmatrix} \quad \boxed{\star}$$

$$\oplus$$

$$W^{(2)} = \begin{bmatrix} 5 & -26 & 1 \\ -9 & 62 & -1 \end{bmatrix} \quad \boxed{\star}$$

and

$$b^{(2)} = \begin{bmatrix} -7 \\ 21 \end{bmatrix} \quad \boxed{\star}$$

Scanned with CamScanner

1) a) Setting the initial state as $X_0=1$, we have to first construct Markov chains for <u>proposal distributions of q</u>. Since $X_0=1$, there exist two possible cases for $X_1 \rightsquigarrow$ <u>$X_1=2,4$</u>

Taking into account $\alpha_{ij}=\min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}$ measurement, we can find that

$q_{41}=q_{14}=\frac{1}{2}$ with $\pi_4=3$, $\pi_1=4 \Rightarrow \alpha_{14}=\min\left\{1, \frac{\pi_4 q_{41}}{\pi_1 q_{14}}\right\}=$

$=\min\left\{1, \frac{3}{4}\right\}=\frac{3}{4}$, $\boxed{\alpha_{14}=\frac{3}{4}}$ Similarly, $q_{12}=q_{21}=\frac{1}{2}$ where $\pi_1=4$, $\pi_2=2$

$\alpha_{12}=\min\left\{\frac{\pi_2 q_{21}}{\pi_1 q_{12}}, 1\right\}=\min\left\{1, \frac{2}{4}\right\}=\frac{1}{2}$, $\boxed{\alpha_{12}=\frac{1}{2}}$ It is well-known that the transition probability p_{ij} from state i to state j of the Markov chain is again

$p_{ij}=q_{ij}\alpha_{ij}$, $p_{12}=q_{12}\alpha_{12}=\frac{1}{2}\alpha_{12}=\frac{1}{4}$ and $p_{14}=q_{14}\alpha_{14}=\frac{1}{2}\frac{3}{4}=\frac{3}{8}$

$\boxed{p_{12}=\frac{1}{4}}$ $\boxed{p_{14}=\frac{3}{8}}$ Taking into account that the row sums should be 1, $\boxed{p_{13}=0}$ since $q_{13}=0$ and thus, $p_{11}=1-\frac{1}{4}-\frac{3}{8}=$

$=1-\frac{5}{8}=\frac{3}{8}$, $\boxed{p_{11}=\frac{3}{8},\ p_{12}=\frac{1}{4},\ p_{13}=0,\ p_{14}=\frac{3}{8}}$ ★ Now, moving to the second X_1

<u>$X_1=4$ will yield $X_2=3,1$</u> From the properties written above, it's clear

$\alpha_{43}=\min\left\{\frac{\pi_3 q_{34}}{\pi_4 q_{43}}, 1\right\}$ where $q_{34}=q_{43}=\frac{1}{2}$ and $\pi_3=\pi_4=3$

Thus, $\boxed{\alpha_{43}=1}$ and $p_{43}=q_{43}\alpha_{43}=\frac{1}{2}$

$\boxed{p_{43}=\frac{1}{2}}$ $\pi_1=4$, $\pi_4=3$, and $q_{14}=q_{41}=\frac{1}{2} \Rightarrow \frac{\pi_1 q_{14}}{\pi_4 q_{41}}=\frac{4}{3}>1$, so

$\alpha_{41}=1$ with $p_{41}=q_{41}\alpha_{41}=\frac{1}{2} \Rightarrow \boxed{p_{41}=\frac{1}{2}}$ Since $p_{42}=q_{42}\alpha_{42}=0$,

Scanned with CamScanner

$\boxed{p_{42}=0}$ Sum of rows are equal to 1, thus $p_{41}+p_{42}+p_{43}+p_{44}=1$ and

$p_{44}=1-\frac{1}{2}-\frac{1}{2}=0$, $\boxed{p_{41}=\frac{1}{2},\ p_{42}=0,\ p_{43}=\frac{1}{2},\ p_{44}=0}$ ☆ In the sub-sequent phase, it's clear that

by choosing $X_2=3 \Rightarrow$ we can declare $X_3=2$ and 4

$\pi_3=\pi_4=3$, $q_{34}=q_{43}=\frac{1}{2} \Rightarrow \pi_4 q_{43}=\pi_3 q_{34}$ meaning that $\boxed{\alpha_{34}=1}$

$q_{23}=q_{32}=\frac{1}{2}$, $\pi_2=2$, $\pi_3=3 \Rightarrow \frac{\pi_2 q_{23}}{\pi_3 q_{32}}=\frac{2}{3}<1$, thus $\boxed{\alpha_{32}=\frac{2}{3}}$

$\alpha_{34} q_{34}=q_{34}=\boxed{\frac{1}{2}=p_{34}}$ $\alpha_{32} q_{32}=\frac{2}{3}\frac{1}{2}=\frac{1}{3} \Rightarrow \boxed{p_{32}=\frac{1}{3}}$ Since we know $p_{31}=\alpha_{31}q_{31}=0$

$\boxed{p_{31}=0}$ Summation of row elements $=1 \Rightarrow p_{33}=1-\frac{1}{2}-\frac{1}{3}=\frac{1}{6}$, yielding that

$\boxed{p_{31}=0,\ p_{32}=\frac{1}{3},\ p_{33}=\frac{1}{6},\ p_{34}=\frac{1}{2}}$ ☆ In the last phase, choosing $X_3=2$

$\pi_1=4$, $\pi_2=2$, $q_{12}=q_{21}=\frac{1}{2} \Rightarrow$

$\frac{\pi_1 q_{12}}{\pi_2 q_{21}}=2>1$, so $\boxed{\alpha_{21}=1}$ $\pi_2=2$, $\pi_3=3$, $q_{23}=q_{32}=\frac{1}{2} \Rightarrow$

$p_{21}=\alpha_{21}q_{21}=q_{21}=\frac{1}{2}$

$\frac{\pi_3 q_{32}}{\pi_2 q_{23}}=\frac{3}{2}>1 \Rightarrow \boxed{\alpha_{23}=1}$ $p_{23}=\alpha_{23}q_{23}=\frac{1}{2}$ $\Big\}$ $\boxed{p_{21}=p_{23}=\frac{1}{2}}$

$p_{24}=\alpha_{24}q_{24}=$ should be 1

$=0$, therefore $p_{22}=1-p_{21}-p_{23}-p_{24}=0$ since sum of row elements

$\boxed{p_{21}=\frac{1}{2},\ p_{22}=0,\ p_{23}=\frac{1}{2},\ p_{24}=0}$ ☆ In conclusion, the desired transition matrix for q will be

$$P=\begin{bmatrix} \frac{3}{8} & \frac{1}{4} & 0 & \frac{3}{8} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$ ☆ ☆ $\boxed{Y}$

Scanned with CamScanner

Now, setting the initial state as $X_0=1$, we will construct Markov chains for proposal distribution of $\tilde{q}$. Since $X_0=1$, we will set $X_1=2$ and

$\pi_4=3$, $\pi_1=4$, $\tilde{q}_{14}=\frac{1}{3}$, $\tilde{q}_{41}=\frac{2}{3} \Rightarrow \frac{\pi_4 \tilde{q}_{41}}{\pi_1 \tilde{q}_{14}} = \frac{2}{4/3} = \frac{3}{2} > 1$ with

$\alpha_{14}=\min\left\{1, \frac{\pi_4 \tilde{q}_{41}}{\pi_1 \tilde{q}_{14}}\right\}=1$ where $p_{14}=\alpha_{14}\tilde{q}_{14}=\frac{1}{3} \Rightarrow \boxed{p_{14}=\frac{1}{3}}$

$\pi_2=2$, $\pi_1=4$, $\tilde{q}_{12}=\frac{2}{3}$, $\tilde{q}_{21}=\frac{1}{3} \Rightarrow \frac{\pi_2 \tilde{q}_{21}}{\pi_1 \tilde{q}_{12}} = \frac{2/3}{8/3} = \frac{1}{4} < 1$, so

$\alpha_{12}=\min\left\{1, \frac{\pi_2 \tilde{q}_{21}}{\pi_1 \tilde{q}_{12}}\right\}=\frac{1}{4}$, where $p_{12}=\alpha_{12}\tilde{q}_{12}=\frac{1}{4}\cdot\frac{2}{3}=\frac{1}{6}$

in the row should be equal to 1

$\boxed{p_{12}=\frac{1}{6}}$ $p_{13}=\alpha_{13}\tilde{q}_{13}=0 \Rightarrow p_{11}=1-\frac{1}{3}-\frac{1}{6}=1-\frac{1}{2}=\frac{1}{2}$ since sum of elems

$\boxed{p_{11}=\frac{1}{2},\ p_{12}=\frac{1}{6},\ p_{13}=0,\ p_{14}=\frac{1}{3}}$ ☆ Subsequently, having $X_1=4$, we set $X_2=3$ and 1:

$\frac{\pi_3 \tilde{q}_{34}}{\pi_4 \tilde{q}_{43}} = \frac{3\times\frac{2}{3}}{3\times\frac{1}{3}} = 2 > 1 \Rightarrow \alpha_{43}=\min\left\{1, \frac{\pi_3 \tilde{q}_{34}}{\pi_4 \tilde{q}_{43}}\right\}=1$ and

similarly, $p_{43}=\alpha_{43}\tilde{q}_{43}=1\times\frac{1}{3}=\frac{1}{3}$ $\boxed{p_{43}=\frac{1}{3}}$

$\frac{\pi_1 \tilde{q}_{14}}{\pi_4 \tilde{q}_{41}} = \frac{4\times\frac{1}{3}}{3\times\frac{2}{3}} = \frac{2}{3} < 1 \Rightarrow \alpha_{41}=\frac{2}{3}$ and $p_{41}=\alpha_{41}\tilde{q}_{41}=\frac{2}{3}\times\frac{2}{3}=\frac{4}{9}$

$\boxed{p_{41}=\frac{4}{9}}$

Since $p_{42}=\alpha_{42}\tilde{q}_{42}=0$, and sum of elem's in the row should be 1, therefore

$p_{44}=1-\frac{4}{9}-\frac{1}{3}=1-\frac{7}{9}=\frac{2}{9}$, $\boxed{p_{41}=\frac{4}{9},\ p_{42}=0,\ p_{43}=\frac{1}{3},\ p_{44}=\frac{2}{9}}$ ☆ Now, after

setting $X_2=3 \Rightarrow X_3=2$ and 4 $\frac{\pi_2 \tilde{q}_{23}}{\pi_3 \tilde{q}_{32}} = \frac{2\times 2/3}{3\times 1/3} = \frac{4}{3} > 1$, meaning

Scanned with CamScanner

$\alpha_{32}=1$ and $p_{32}=\alpha_{32}\tilde{q}_{32}=\tilde{q}_{32}=\frac{1}{3} \Rightarrow \boxed{p_{32}=\frac{1}{3}}$ $p_{34}=\alpha_{34}\tilde{q}_{34}=$

$$=\min\left\{\frac{\pi_4\tilde{q}_{43}}{\pi_3\tilde{q}_{34}},1\right\}\frac{2}{3}=\min\left\{1,\frac{\tilde{q}_{43}}{\tilde{q}_{34}}\right\}\frac{2}{3}=\min\left\{1,\frac{1/3}{2/3}\right\}\frac{2}{3}=$$

$=\frac{1}{\cancel{2}}\times\frac{\cancel{2}}{3}=\frac{1}{3}$, so $\boxed{p_{34}=\frac{1}{3}}$ $p_{31}=\alpha_{31}\tilde{q}_{31}=0$, and since the summation of row elements $=1$, we ensure

$p_{33}=1-\frac{1}{3}-\frac{1}{3}=\frac{1}{3} \Rightarrow \boxed{p_{31}=0,\ p_{32}=\frac{1}{3},\ p_{33}=\frac{1}{3},\ p_{34}=\frac{1}{3}}$ ☆ In the final phase, setting $X_3=2$ yields

$\frac{\pi_1\tilde{q}_{12}}{\pi_2\tilde{q}_{21}}=\frac{4\times\frac{2}{3}}{2\times\frac{1}{3}}=4>1$, which shows $\alpha_{21}=1$ and $p_{21}=\alpha_{21}\tilde{q}_{21}=$ $=\frac{1}{3}$, $\boxed{p_{21}=\frac{1}{3}}$ Similarly,

$\frac{\pi_3\tilde{q}_{32}}{\pi_2\tilde{q}_{23}}=\frac{3\times 1/3}{2\times 2/3}=\frac{3}{4}<1$, yielding $\alpha_{23}=\frac{3}{4}$ and $p_{23}=\alpha_{23}\tilde{q}_{23}=$ $=\frac{3}{4}\times\frac{2}{3}=\frac{1}{2}$

$\boxed{p_{23}=\frac{1}{2}}$ Since $p_{24}=\alpha_{24}\tilde{q}_{24}=0$ where summation of row elements is equal to 1, $p_{22}=1-\frac{1}{2}-\frac{1}{3}=\frac{1}{6}$. $\boxed{p_{21}=\frac{1}{3},\ p_{22}=\frac{1}{6},\ p_{23}=\frac{1}{2},\ p_{24}=0}$ ☆

In conclusion, transition matrix for $\tilde{q}$ will be

$$\tilde{P}=\begin{bmatrix}\frac{1}{2} & \frac{1}{6} & 0 & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} & 0\\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3}\\ \frac{4}{9} & 0 & \frac{1}{3} & \frac{2}{9}\end{bmatrix}$$ ☆ ☆

6) Now, we'll calculate the exact $E[f(X)] \Rightarrow$ Since from the given statements and definition, $E[f(X)]=$

$$=E[X^2-3]=\sum_{j=0}^{\infty}(j^2-3)P(X=j)=$$ since given is PMF prop problem

$$\Rightarrow E[f(X)]=(1^2-3)P(X=1)+(2^2-3)P(X=2)+(3^2-3)P(X=3)+(4^2-3)P(X=4)$$

Since we have to normalize probabilities given the fact that $\tilde{p}(X=1)=4$, $\tilde{p}(X=2)=2$, $\tilde{p}(X=3)=3$, $\tilde{p}(X=4)=3 \Rightarrow$ the followings are normalized:

Scanned with CamScanner

$$P(X=1) = \frac{4}{4+2+3+3} = \frac{4}{12} = \frac{1}{3}$$

$$P(X=2) = \frac{2}{4+2+3+3} = \frac{2}{12} = \frac{1}{6}$$

$$P(X=3) = \frac{3}{4+2+3+3} = \frac{3}{12} = \frac{1}{4}$$

$$P(X=4) = \frac{3}{4+2+3+3} = \frac{3}{12} = \frac{1}{4}$$

$$E[f(x)] = (-2)\frac{1}{3} + \frac{1}{6} + 6 \times \frac{1}{4} + 13 \times \frac{1}{4} = \frac{-2}{3} + \frac{1}{6} + \frac{3}{2} + \frac{13}{4} = -\frac{4}{6} + \frac{1}{6} + \frac{3}{2} + \frac{13}{4} = -\frac{1}{2} + \frac{3}{2} + \frac{13}{4} = \frac{17}{4}$$

Thus, $\boxed{E[f(x)] = \frac{17}{4}}$ ☆

Theoreticoppy, q and $\tilde{q}$ will definitely converge the expected value $\frac{17}{4}$ after some iterations, but it should be noted after plotting the errors along during programming $h=1, \dots, 1000$, $\tilde{q}$-proposal distribution will converge faster in the expected sense ☑ ⊕ as it could be observed over there (for $\tilde{q}$-proposed distribution, less # of iterations could be enough to converge)

Scanned with CamScanner

2) a) Since we know that $P(a_0|s_0) = \frac{1}{2}$ ①

$P(a_1|s_1) = \frac{1}{2}$ ②

$P(a_2|s_2) = \frac{1}{2}$ ③

$P(a_3|s_3) = \frac{3}{10}$ ④

are satisfied, meanwhile it's also clear that according to given statements

$P(\overbrace{(s_0+2)\%5}^{s_1} | s_0, a^1) = \frac{1}{3}$ ①

$P((s_1+3)\%5 | s_1, a^2) = \frac{1}{2}$ ②

$P((s_2+3)\%5 | s_2, a^2) = \frac{1}{2}$ ③

$P(\underbrace{(s_3+1)\%5}_{s_4} | s_3, a^2) = \frac{1}{2}$ ④

and moreover, we have to take into account that $P(r_1 = 1) = \frac{1}{2}$ holds true

Therefore, using all these results, we can measure the probability that trajectory is sampled:

$$P(\tau) = P(\tau = (s_0=0, a_0=a^1, r_0=0, s_1=2, a_1=a^2, r_1=1, s_2=0, a_2=a^2, r_2=0, s_3=3, a_3=a^2, r_3=0, s_4=4))$$

$$= \frac{1}{2} \times 1 \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times 1 \times \frac{1}{2} \times \frac{3}{10} \times 1 \times \frac{1}{2} =$$

$$= \frac{\frac{1}{24} \times \frac{3}{80}}{2} = \frac{1}{640} \cdot \frac{1}{2} = \boxed{\frac{1}{1280}}$$

b) We can use the Bellman Equation iteratively starting with a value function zero everywhere. This method converges usually very quickly. But since we have a very small state and action space, the Bellman Equation is only a small set of linear equations that can be obtained in an explicit manner. After that, we can solve them using matrix methods (basic LA)

Scanned with CamScanner

The Bellman equation for this problem is given as a system of $|S|$ equations: $V^{\pi}(s) = \sum_{\text{actions } a} \pi(a|s) \sum_{\text{states } s'} P(s'|s,a)\left(R(s',a,s) + \gamma \cdot V^{\pi}(s')\right)$

The trick is to rewrite this in matrix form and solve the achieved equation wrt V^{π}:

$V^{\pi} = \pi^1 P_1 (R_1 \mathbf{1}^T + \gamma V^{\pi}) + \pi^2 P_2 (R_2 \mathbf{1}^T + \gamma V^{\pi})$, or just

$V^{\pi} = \pi^1 P_1 (R_1 \mathbf{1}^T + \gamma V^{\pi}) + \pi^2 P_2 (R_2 \mathbf{1}^T + \gamma V^{\pi})$ where

$\pi^i = (\pi(a^i|0), \dots, \pi(a^i|4))$

$$R_i = \begin{pmatrix} R(0,a^i,0) & \dots & R(4,a^i,0) \\ \vdots & \ddots & \vdots \\ R(0,a^i,4) & \dots & R(4,a^i,4) \end{pmatrix}$$

$\mathbf{1} = (1, \dots, 1)$

Then, we can regroup and obtain the following:

$V^{\pi} = (\pi^1 P_1 R_1 + \pi^2 P_2 R_2)\mathbf{1}^T + (\pi^1 P_1 + \pi^2 P_2) V^{\pi}$

which finally results in

$V^{\pi} = (Id - \gamma(\pi^1 P_1 + \pi^2 P_2))^{-1} (\pi^1 P_1 R_1 + \pi^2 P_2 R_2)\mathbf{1}^T$

It should be noted that the inverse exists only for $\gamma < 1$, where the rewards are randomized and depend on state transition, too.

The below recursion also indicates how the computations are implemented correctly:

Scanned with CamScanner

We set initially $V^{\pi}(0)=(0,0,0,0,0)$ and

$$V^{\pi}(k+1,s) \leftarrow \sum_{\text{actions } a} \pi(a|s) \sum_{\text{states } s'} P(s'|s,a)\left(R(s',a,s) + \gamma \cdot V^{\pi}(k,s')\right)$$

c) <u>Policy Iteration:</u>

Setting $n=0$ and selecting an arbitrary policy $d_0 \in D^{MD}$ (1)

(policy evaluation) (2)

→ obtaining $v^{(n)}$ by solving $(I-\lambda P_{d_n})v^{(n)} = r_{d_n}$

(policy improvement) (3)

→ Choosing d_{n+1} to satisfy

$$d_{n+1} \in \operatorname{argmax}_{d \in D^{MD}} \{ r_d + \lambda P_d v^{(n)} \}$$

setting $d_{n+1} = d_n$ if possible

If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise, increasing n by 1 and returning to step 2 (4)

$$d_0 = [a^1\ a^1\ a^1\ a^1\ a^1]^T$$

$$P_a^2 = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 \end{bmatrix}$$

$$P_{d_0} = P_a^1 = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \end{bmatrix} \Rightarrow r_{d_0} = \begin{bmatrix} 0 \\ 1/3 \\ 0 \\ 1/3 \\ 1/3 \end{bmatrix} \cdot \frac{1}{2} = \begin{bmatrix} 0 \\ 1/6 \\ 0 \\ 1/6 \\ 1/6 \end{bmatrix}$$

Taking into account the equation $(I-\lambda P_{d_0})v^{(0)} = r_{d_0}$, and plugging in those matrices (and λ-variable) with solving in numpy algorithms will yield $v^{(0)} = (I-\lambda P_{d_0})^{-1} r_{d_0} =$

$$= \begin{bmatrix} 0.913182 \\ 1.036418 \\ 0.943111 \\ 1.042877 \\ 1.064410 \end{bmatrix}$$

Now, if we take into account policy improvement stage in which $d_1 = \operatorname{argmax}\{ r_{a_1} + \lambda P_{a_1} v^{(0)}, r_{a_2} + \lambda P_{a_2} v^{(0)} \}$

$r_{a_1} \to 1/2$ of the 1st column of P_{a_1} → similar for a_2

Scanned with CamScanner