Consider a shift-reduce parser applied to the sentence *John saw Mary*, where as usual a parse configuration is a triple consisting of a stack, a buffer, and a set of dependencies.

We will assume that the set of dependency labels for the parser are \( \mathcal{D} = \{ \text{root, nsubj, dobj} \} \). The set of possible actions are as follows:

**SHIFT**

- \( \text{LEFT-ARC}(l) \) for any label \( l \in \mathcal{D} \)
- \( \text{RIGHT-ARC}(l) \) for any label \( l \in \mathcal{D} \)

The table below shows the sequence of configurations, with actions \( a_1, a_2, \ldots, a_6 \) mapping one parse configuration to the next configuration:

<table>
<thead>
<tr>
<th>Action</th>
<th>Stack</th>
<th>Buffer</th>
<th>Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 = \text{SHIFT} )</td>
<td>([\text{root}_0])</td>
<td>([\text{John}_1 \text{ saw}_2 \text{ Mary}_3])</td>
<td>({})</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>([\text{root}_0 \text{ John}_1])</td>
<td>([\text{saw}_2 \text{ Mary}_3])</td>
<td>({})</td>
</tr>
<tr>
<td>( a_3 )</td>
<td>([\text{root}_0 \text{ John}_1 \text{ saw}_2])</td>
<td>([\text{Mary}_3])</td>
<td>({})</td>
</tr>
<tr>
<td>( a_4 )</td>
<td>([\text{root}_0 \text{ saw}_2])</td>
<td>([\text{Mary}_3])</td>
<td>({2 \rightarrow \text{nsubj} 1})</td>
</tr>
<tr>
<td>( a_5 )</td>
<td>([\text{root}_0 \text{ saw}_2 \text{ Mary}_3])</td>
<td>([])</td>
<td>({2 \rightarrow \text{nsubj} 1})</td>
</tr>
<tr>
<td>( a_6 )</td>
<td>([\text{root}_0 \text{ saw}_2])</td>
<td>([])</td>
<td>({2 \rightarrow \text{nsubj} 1, 2 \rightarrow \text{dobj} 3})</td>
</tr>
<tr>
<td>( a_6 )</td>
<td>([\text{root}_0])</td>
<td>([])</td>
<td>({2 \rightarrow \text{nsubj} 1, 2 \rightarrow \text{dobj} 3, 0 \rightarrow \text{root} 2})</td>
</tr>
</tbody>
</table>

What values should the actions \( a_2, a_3, a_4, a_5, a_6 \) take to give the sequence of configurations given in the table?
Consider a computational graph with the following definitions:

- Number of vertices \( n = 7 \)
- Number of leaves \( l = 3 \)
- Edges \( E = \{(1, 4), (2, 5), (3, 5), (4, 6), (5, 6), (3, 7), (6, 7)\} \)
- Variables \( u^i \in \mathbb{R}^{d^i} \) for \( i = 1 \ldots 7 \), where \( d^i = 1 \) for all \( i \) (i.e., all variables in the graph are scalars).
- \( f^4 \ldots f^7 \) are defined as follows:
  
  \[
  f^4(u^1) = 3 \times u^1 \\
  f^5(u^2, u^3) = u^2 \times u^3 \\
  f^6(u^4, u^5) = (u^4)^2 + (u^5)^2 \\
  f^7(u^3, u^6) = u^3 + 10 \times u^6
  \]

**Question 1** (5 points) Draw the graph corresponding to the edges \( E \) and the vertices \( 1 \ldots 7 \).

**Question 2** (5 points) Assume we have leaf values \( u^1 = 1, u^2 = 2, \) and \( u^3 = 3 \). Write down the values for \( u^4, u^5, u^6, u^7 \) as calculated by the forward algorithm.

**Question 3** (10 points) Complete expressions for the Jacobian function associated with each edge in the graph:

\[
J^{1\to 4}(u^1) = \frac{\partial f^4(u^1)}{\partial u^1} = 3 \\
J^{2\to 5}(u^2, u^3) = \frac{\partial f^5(u^2, u^3)}{\partial u^2} = u^3 \\
J^{3\to 5}(u^2, u^3) = \frac{\partial f^5(u^2, u^3)}{\partial u^3} = \\
J^{4\to 6}(u^4, u^5) = \frac{\partial f^6(u^4, u^5)}{\partial u^4} = \\
J^{5\to 6}(u^4, u^5) = \frac{\partial f^6(u^4, u^5)}{\partial u^5} = \\
J^{6\to 7}(u^3, u^6) = \frac{\partial f^7(u^3, u^6)}{\partial u^3} = \\
J^{7\to 7}(u^3, u^6) = \frac{\partial f^7(u^3, u^6)}{\partial u^6} = \\
J^{1\to 4}(u^1) = \frac{\partial f^4(u^1)}{\partial u^1} = 3
\]
\[ J^{5\to6}(u^4, u^5) = \frac{\partial f^6(u^4, u^5)}{\partial u^5} = \]

\[ J^{3\to7}(u^3, u^6) = \frac{\partial f^7(u^3, u^6)}{\partial u^3} = \]

\[ J^{6\to7}(u^3, u^6) = \frac{\partial f^7(u^3, u^6)}{\partial u^6} = \]

**Question 4** (10 points) Define \( h^7 \) to be the global function that maps leaf values \( u^1, u^2, u^3 \) to the output value \( u^7 \) from the forward algorithm:

\[ u^7 = h^7(u^1, u^2, u^3) \]

Again assume we have leaf values \( u^1 = 1, u^2 = 2, \) and \( u^3 = 3 \). Recall that to calculate a partial derivative \( \frac{\partial u^7}{\partial u^j} \bigg|_{h^7} \) for any leaf value \( j \in \{1, 2, 3\} \), we need to sum over all directed paths from vertex \( j \) to vertex 7, taking a product of Jacobians over each path. It follows for example that

\[ \frac{\partial u^7}{\partial u^1} \bigg|_{h^7} = J^{6\to7}(u^3, u^6) \times J^{4\to6}(u^4, u^5) \times J^{1\to4}(u^1) \]

because there is a single directed path \((1, 4), (4, 6), (6, 7)\) from vertex 1 to vertex 7 in the graph.

Write down an expression for

\[ \frac{\partial u^7}{\partial u^3} \bigg|_{h^7} \]

in terms of the Jacobian functions, and calculate the value for \( \frac{\partial u^7}{\partial u^3} \) assuming leaf values \( u^1 = 1, u^2 = 2, \) and \( u^3 = 3 \). Make sure to show all your working.
Assume we have a feedforward neural network with the following definitions:

- The input dimension $d = 2$. Hence each input to the network $x$ is a vector in $\mathbb{R}^d$ with components $x_1$ and $x_2$.
- The number of hidden units $m = 3$.
- A parameter matrix $W \in \mathbb{R}^{m \times d}$. The $m$ rows of $W$ are defined as
  
  $W_1 = (1, 1)$
  $W_2 = (1, 0)$
  $W_3 = (1, -1)$

- The bias parameters are all 0, that is $b_1 = b_2 = b_3 = 0$
- The transfer function is $g(z) = \text{RELU}(z)$ where
  \[
  \text{RELU}(z) = \begin{cases} 
  z & \text{if } z \geq 0, \\
  0 & \text{otherwise}
  \end{cases}
  \]

- Given an input $x$, the outputs from the three neurons in the model are
  
  $h_1 = g(W_1 \cdot x + b_1)$
  $h_2 = g(W_2 \cdot x + b_2)$
  $h_3 = g(W_3 \cdot x + b_3)$

  We use $h$ to refer to the vector in $\mathbb{R}^3$ with components $h_1, h_2, h_3$.

- The set of output labels in the model are $\mathcal{Y} = \{-1, +1\}$. For each label $y \in \mathcal{Y}$ we define $v(y) \in \mathbb{R}^3$ to be a parameter vector associated with label $y$, and $\gamma_y \in \mathbb{R}$ to be a bias parameter. We then have
  \[
  p(y|x) = \frac{\exp\{v(y) \cdot h + \gamma_y\}}{\sum_{y'} \exp\{v(y') \cdot h + \gamma_{y'}\}}
  \]

**Question 5** (10 points) Assume the input to the network is a vector $x$ with $x_1 = 10, x_2 = -20$. What are the values for $h_1, h_2, h_3$ for this network with this input?
Consider a computational graph with the following definitions:

**Inputs:** A training example \((x^i, y^i)\) where \(x^i = (x^i_1, x^i_2, x^i_3)\) and \(x^i_1, x^i_2\) and \(x^i_3\) are words, and \(y^i \in \mathcal{Y}\) where \(\mathcal{Y}\) is a set of labels. A word dictionary \(D\) with size \(s(D)\). An embedding matrix \(E \in \mathbb{R}^{2 \times s(D)}\). A single-layer feedforward network with \(m = 1\) neurons, and a transfer function \(g(z) = \text{RELU}(z)\) where

\[
\text{RELU}(z) = \begin{cases} 
z & \text{if } z \geq 0 \\
0 & \text{otherwise}
\end{cases}
\]

The feedforward network has parameters \(W \in \mathbb{R}^{m \times 2}\), \(b \in \mathbb{R}^m\), \(V \in \mathbb{R}^{K \times m}\), and \(\gamma \in \mathbb{R}^K\), where \(K = |\mathcal{Y}|\).

**Computational Graph:**

\[
\begin{align*}
x'_1 & \in \mathbb{R}^2 = E \times \text{Onehot}(x^i_1, D) \\
x'_2 & \in \mathbb{R}^2 = E \times \text{Onehot}(x^i_2, D) \\
x'_3 & \in \mathbb{R}^2 = E \times \text{Onehot}(x^i_3, D) \\
u & \in \mathbb{R}^2 = x'_1 + x'_2 + x'_3 \\
z & \in \mathbb{R}^1 = Wu + b \\
h & \in \mathbb{R}^1 = g(z) \\
l & \in \mathbb{R}^K = Vh + \gamma \\
q & \in \mathbb{R}^K = \text{Log-Softmax}(l) \\
o & \in \mathbb{R} = -q_{y^i}
\end{align*}
\]

Note that \(u\) is calculated by summing the values for \(x'_1, x'_2, x'_3\), not by concatenating the three values.

Assume in addition that the set of possible words in the vocabulary is \{"the, a, this, dog, cat, mouse\} and furthermore for any word \(x\) in the set \{"the, a, this\} we have

\[
E \times \text{Onehot}(x, D) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}
\]

and for any word in the set \{"dog, cat, mouse\} we have

\[
E \times \text{Onehot}(x, D) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]
Question 6  (10 points) What is the value for \( u \) for each of the following inputs \( x_1^i, x_2^i, x_3^i \) given below? Write the value for \( u \) for set of input values for \( x_1^i, x_2^i, x_3^i \) shown below:

\[ x_1^i = \text{the, } x_2^i = a, \ x_3^i = \text{this}, \]

\[ u = \]

\[ x_1^i = \text{the, } x_2^i = a, \ x_3^i = \text{mouse}, \]

\[ u = \]

\[ x_1^i = \text{the, } x_2^i = \text{dog, } x_3^i = \text{mouse}, \]

\[ u = \]

\[ x_1^i = \text{cat, } x_2^i = \text{dog, } x_3^i = \text{mouse}, \]

\[ u = \]
**Question 7** (10 points) Now assume that the bias parameter $b = -2$, and assume

$$ W = [1, 0] $$

Note that from the computational graph given above,

$$ h = g(Wu + b) $$

where $g(z) = \text{RELU}(z)$.

For what values for the triple $(x_1, x_2, x_3)$ do we have $h > 0$? **Make sure to explain your reasoning.** Make sure to specify all values of $x_1, x_2, x_3$ that lead to $h > 0$, not just the example values given above.

**Question 8** (10 points) Again assume that the bias parameter $b = -2$, and assume

$$ W = [1, 0] $$

Note that from the computational graph given above,

$$ h = g(Wu + b) $$

where $g(z) = \text{RELU}(z)$.

Assume that the set of possible labels is $Y = \{1, 2\}$. It follows that there are parameters $V_1 \in \mathbb{R}^1$, $V_2 \in \mathbb{R}^1$, $\gamma_1 \in \mathbb{R}$, $\gamma_2 \in \mathbb{R}$.

Assume we would like the probability distribution under the model to be the following:

- If $x_1 \in \{\text{the, a, this}\}$ and $x_2 \in \{\text{the, a, this}\}$ and $x_3 \in \{\text{the, a, this}\}$,
  $$ p(1|x^i; W, b, V, \gamma) = 0.8, \quad p(2|x^i; W, b, V, \gamma) = 0.2 $$

- Otherwise
  $$ p(1|x^i; W, b, V, \gamma) = p(2|x^i; W, b, V, \gamma) = 0.5 $$

What values for the parameters $V_1$, $V_2$, $\gamma_1$, $\gamma_2$ give this distribution? **Make sure to give justification for your answer.**