COMS 4705, Homework 4

Part #1 (10 points)

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

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

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:

Action	Stack	Buffer	Dependencies
	$[root_0]$	[John$_1$ saw$_2$ Mary$_3$]	$\{\}$
a_1 = SHIFT			
	$[root_0$ John$_1]$	[saw$_2$ Mary$_3$]	$\{\}$
a_2			
	$[root_0$ John$_1$ saw$_2]$	[Mary$_3$]	$\{\}$
a_3			
	$[root_0$ saw$_2]$	[Mary$_3$]	$\{2 \rightarrow^{nsubj} 1\}$
a_4			
	$[root_0$ saw$_2$ Mary$_3]$	[]	$\{2 \rightarrow^{nsubj} 1\}$
a_5			
	$[root_0$ saw$_2]$	[]	$\{2 \rightarrow^{nsubj} 1, 2 \rightarrow^{dobj} 3\}$
a_6			
	$[root_0]$	[]	$\{2 \rightarrow^{nsubj} 1, 2 \rightarrow^{dobj} 3, 0 \rightarrow^{root} 2\}$

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?

Part #2 — 30 points

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:

$$J^{1\to4}(u^1) = \frac{\partial f^4(u^1)}{\partial u^1} = 3$$

$$J^{2\to5}(u^2, u^3) = \frac{\partial f^5(u^2, u^3)}{\partial u^2} = u^3$$

$$J^{3\to5}(u^2, u^3) = \frac{\partial f^5(u^2, u^3)}{\partial u^3} =$$

$$J^{4\to6}(u^4, u^5) = \frac{\partial f^6(u^4, u^5)}{\partial u^4} =$$

$$J^{5\to 6}(u^4, u^5) = \frac{\partial f^6(u^4, u^5)}{\partial u^5} =$$

$$J^{3\to 7}(u^3, u^6) = \frac{\partial f^7(u^3, u^6)}{\partial u^3} =$$

$$J^{6\to 7}(u^3, u^6) = \frac{\partial f^7(u^3, u^6)}{\partial u^6} =$$

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

$$\left.\frac{\partial u^7}{\partial u^j}\right|^{h^7}_{u^1,u^2,u^3} = \frac{\partial h^7(u^1, u^2, u^3)}{\partial u^j}$$

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

$$\left.\frac{\partial u^7}{\partial u^1}\right|^{h^7}_{u^1,u^2,u^3} = J^{6\to 7}(u^3, u^6) \times J^{4\to 6}(u^4, u^5) \times J^{1\to 4}(u^1)$$

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

$$\left.\frac{\partial u^7}{\partial u^3}\right|^{h^7}_{u^1,u^2,u^3}$$

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.**

Part #3 — 10 points

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

$$W_1 = \langle 1, 1 \rangle$$
$$W_2 = \langle 1, 0 \rangle$$
$$W_3 = \langle 1, -1 \rangle$$

- 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

$$\text{RELU}(z) = \quad z \text{ if } z \geq 0, \text{ 0 otherwise}$$

- 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, and 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 parameters. 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?

Part #4 — 20 points

Consider a computational graph with the following definitions:

Inputs: A training example (x^i, y^i) where $x^i = (x_1^i, x_2^i, x_3^i)$ and x_1^i, x_2^i and x_3^i 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

$$\text{RELU}(z) = \; z \text{ if } z \geq 0, \, 0 \text{ otherwise}$$

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:

$$\begin{aligned}
x_1' \in \mathbb{R}^2 &= E \times \text{Onehot}(x_1^i, D) \\
x_2' \in \mathbb{R}^2 &= E \times \text{Onehot}(x_2^i, D) \\
x_3' \in \mathbb{R}^2 &= E \times \text{Onehot}(x_3^i, D) \\
u \in \mathbb{R}^2 &= x_1' + x_2' + x_3' \\
z \in \mathbb{R}^1 &= Wu + b \\
h \in \mathbb{R}^1 &= g(z) \\
l \in \mathbb{R}^K &= Vh + \gamma \\
q \in \mathbb{R}^K &= \text{Log-Softmax}(l) \\
o \in \mathbb{R} &= -q_{y^i}
\end{aligned}$$

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

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

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

$$E \times \text{Onehot}(x, D) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2}$$

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:

$$x_1^i = \text{the}, \quad x_2^i = \text{a}, \quad x_3^i = \text{this},$$

$u =$

$$x_1^i = \text{the}, \quad x_2^i = \text{a}, \quad x_3^i = \text{mouse},$$

$u =$

$$x_1^i = \text{the}, \quad x_2^i = \text{dog}, \quad x_3^i = \text{mouse},$$

$u =$

$$x_1^i = \text{cat}, \quad x_2^i = \text{dog}, \quad x_3^i = \text{mouse},$$

$u =$

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^i, x_2^i, x_3^i) do we have $h > 0$? **Make sure to explain your reasoning. Make sure to specify *all* values of x_1^i, x_2^i, x_3^i 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 $\mathcal{Y} = \{1, 2\}$. 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 $x_1^i \in \{\text{the, a, this}\}$ and $x_2^i \in \{\text{the, a, this}\}$ and $x_3^i \in \{\text{the, a, this}\}$,

$$p(1|x^i; W, b, V, \gamma) = 0.8, \quad p(2|x^i; W, b, V, \gamma) = 0.2$$

- 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, γ_1, γ_2 give this distribution? **Make sure to give justification for your answer.**