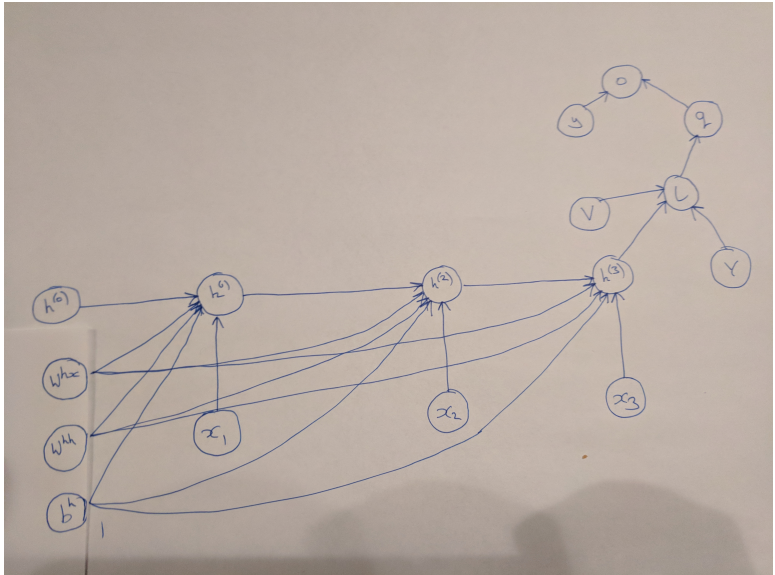
Question 1a



Question 1a (continued)

The three paths from W^{hx} to o are

$$W^{hx} \to h^{(3)} \to l \to q \to o$$

$$W^{hx} \to h^{(2)} \to h^{(3)} \to l \to q \to o$$

$$W^{hx} \to h^{(1)} \to h^{(2)} \to h^{(3)} \to l \to q \to o$$

Question 1b

$$\begin{aligned}\frac{\partial o}{\partial W^{hx}} &= A \times \frac{\partial h^{(3)}}{\partial W^{hx}} \\ &\quad + A \times B^2 \times \frac{\partial h^{(2)}}{\partial W^{hx}} \\ &\quad + A \times B^2 \times B^1 \times \frac{\partial h^{(1)}}{\partial W^{hx}}\end{aligned}$$

Question 1c

The values for all non-leaf variables—$h^{(1)}, h^{(2)}, h^{(3)}, l, q, o$—vary as x_1 varies. Because of this **all** Jacobians in the graph vary as x_1 varies (the Jacobians for any non-leaf variables depend on the value of the variable computed in the forward pass).

Question 2a

Inputs: A sequence $x_1 \dots x_n$ where each $x_j \in \mathbb{R}^d$. A label $y \in \{1 \dots K\}$ for position i.

Computational Graph:

- For $t = 1 \dots n$, $h^{(t)} = g(W^{hx} x^{(t)} + W^{hh} h^{(t-1)} + b^h)$
- For $t = n \dots 1$, $\eta^{(t)} = g(W^{bhx} x^{(t)} + W^{bhh} \eta^{(t+1)} + b^{bh})$
- For $t = 1 \dots n$,
  $h^{(2,t)} = g(W^{2hx} \times \mathsf{CONCAT}(h^{(t)}, \eta^{(t)}) + W^{2hh} h^{(2,t-1)} + b^{2h})$
- For $t = n \dots 1$,
  $\eta^{(2,t)} = g(W^{2bhx} \times \mathsf{CONCAT}(h^{(t)}, \eta^{(t)}) + W^{2bhh} \eta^{(2,t+1)} + b^{2bh})$
- $l = V \times \mathsf{CONCAT}(h^{(2,i)}, \eta^{(2,i)}) + \gamma, \quad q = \mathsf{LS}(l)$, $o = -q_y$

Question 2b

Inputs: A sequence $x_1 \dots x_n$ where each $x_j \in \mathbb{R}^d$. A label $y \in \{1 \dots K\}$ for position i. A sequence of tags $y_1 \dots y_{i-1}$.

Computational Graph:

- For $t = 1 \dots n$, $h^{(t)} = g(W^{hx} x^{(t)} + W^{hh} h^{(t-1)} + b^h)$
- For $t = n \dots 1$, $\eta^{(t)} = g(W^{bhx} x^{(t)} + W^{bhh} \eta^{(t+1)} + b^h)$
- For $j = 1 \dots (i-1)$, $\beta^{(j)} = g(W^{hy} y^{(j)} + W^{yhh} \beta^{(j-1)} + b^y)$
- $l = V \times \mathsf{CONCAT}(h^{(i)}, \eta^{(i)}, \beta^{(i-1)}) + \gamma, \quad q = \mathsf{LS}(l)$, $o = -q_y$

Question 3

- $z^{(t)}$ controls how much the new hidden state $h^{(t)}$ copies information across from $h^{(t-1)}$, and how much a new update is incorporated
- $r^{(t)}$ controls how much of $h^{(t-1)}$ is reset to zero in the update part of the network