**Quiz 6** **Name:** ______________________
**Week 7, Oct/21/2020** **On your left:** ______________________
**CS 280: Fall 2020** **On your right:** ______________________
**Instructor: Xuming He, Lan Xu**

**Instructions:**
Please answer the questions below. Show all your work. This is an open-book test. NO discussion/collaboration is allowed.

**Problem 1.** (10 points) *Parity-check network*. Note that the initial parity bit is 1, what's the relation between each input and the previous parity bit? Determine the relation between the parity and inputs and complete the parity bits($p_1, p_2, p_3, p_4$) and design and draw a RNN to predict parity. Hints: use the hard threshold strategy with corresponding weight value and basis value similar to Lecture12-Page17.

$$\begin{array}{rcccccccccccc} \textit{Parity bits}: & 0 & 0 & 0 & 1 & 0 & 1 & p_1 & p_2 & p_3 & p_4 & \rightarrow \\ \textit{Input}: & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & \end{array}$$

**Solution:**

1. (3') The relation between each input and the previous parity bit is XNOR. Given the truth table is also right:

| P(t-1) | X(t) | P(t) |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |

$X(t)$ and $P(t)$ refer to input bit and parity bit at time $t$ respectively

ps: Answer XNOR or draw the Truth table can get 3 point, 0 point if there is any mistake. BTW, there could be two correct truth tables(with four cases).

2. (3') If the answer of part 1 is XNOR, then $[p_1 \quad p_2 \quad p_3 \quad p_4]$ should be: $[1 \quad 1 \quad 0 \quad 1]$.

If the answer of part 1 is the correct truth table, as the table is not complete, there could be two correct answer: $[1 \quad 1 \quad 0 \quad 1]$ or $[0 \quad 0 \quad 1 \quad 0]$ according to the truth table.

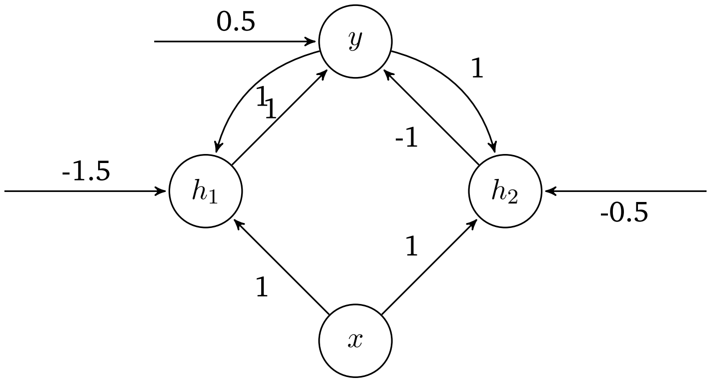
ps: If you answer does not fit your answer of part 1, you will get 0 point.

3. (4') We can design a RNN as the slice of Lecture12-page17, where $x$ and $y$ refer to input bits and parity bits respectively and $h_1, h_2$ are two hidden units. All units use hard threshold strategy:

$$\text{unit output} = \begin{cases} 1 & \text{if unit state} \geq 0 \\ 0 & \text{if unit state} < 0 \end{cases}$$

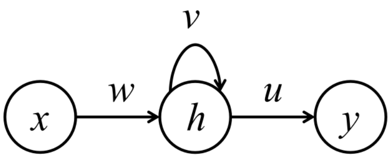
| y(t-1) | x(t) | h1(t) | h2(t) | y(t) |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

One example of the RNN network is as below:



ps: give the right hard threshold: 1 point;
draw correct units and connections of the RNN network: 2 points;
give correct weights and bias: 1 point.

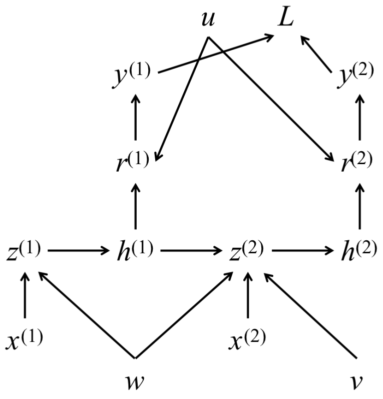
**Problem 2.** (10 points) We have a simple recurrent neural network with the following network structure, and for simplicity let us assume all the variables and operations in activation functions are linear:



- Draw the feed-forward network when unfolding the recurrent network through time for 2 time step $t = 1, 2$.
- Given a training sample $(\hat{y}_1, \hat{y}_2)$ as the ground truth for $(y_1, y_2)$, derive $\frac{\partial L}{\partial w}$ using back propagation based on the graph you have built.
- Please explain the two kinds of gradient problems in RNN training, using the above case as an example.

**Solution:**

1. (3')



2. (3') We can differentiate to the loss function at the each time step and sum all together:

$$\begin{aligned}
\frac{\partial L}{\partial w} &= \sum_{t=1}^{2} \frac{\partial L}{\partial w} \\
&= \frac{\partial L}{\partial y_1}\frac{\partial y_1}{\partial w} + \frac{\partial L}{\partial y_2}\frac{\partial y_2}{\partial w} \\
&= \frac{\partial L}{\partial y_1}\frac{\partial y_1}{\partial r_1}\frac{\partial r_1}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial w} + \frac{\partial L}{\partial y_2}\frac{\partial y_2}{\partial r_2}\frac{\partial r_2}{\partial h_2}\frac{\partial h_2}{\partial z_2}(\frac{\partial z_2}{\partial w} + \frac{\partial z_2}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial w})
\end{aligned}$$

In this scenario, since all the operations are linear, we obtain

$$\frac{\partial y_t}{\partial r_t} = 1, \quad \frac{\partial h_t}{\partial z_t} = 1$$

Thus,

$$\begin{aligned}\frac{\partial L}{\partial w} &= \frac{\partial L}{\partial y_1} \cdot 1 \cdot u \cdot 1 \cdot x_1 + \frac{\partial L}{\partial y_2} \cdot 1 \cdot u \cdot 1 \cdot (x_2 + v \cdot 1 \cdot x_1) \\ &= \frac{\partial L}{\partial y_1} \cdot u \cdot x_1 + \frac{\partial L}{\partial y_2} \cdot u \cdot (x_2 + v \cdot x_1)\end{aligned}$$

3. (a) Vanishing gradients.(1') In the back propagation process, we adjust our weight matrices with the use of a gradient. More explicitly, gradients are calculated by continuous multiplications of derivatives $\frac{\partial h_{j+1}}{\partial h_j}$ and the value of these derivatives may be too small or large. If we perform eigen decomposition on the Jacobian matrix $\frac{\partial h_{j+1}}{\partial h_j}$, when the largest eigenvalue $\lambda_1 < 1$, that these continuous multiplications may cause the gradient to practically "vanishes".(1')

   (b) Exploding gradients.(1') Similarly, if the dominant eigenvalue $\lambda_1$ is greater than 1, the gradient "explodes".(1')