

Note: All coding problems to be submitted with Github Link. Do not Upload the files/folder. Use git commands only.

Note: this is the distribution of questions:

- (a) Question 1 to Question 2: Required for everyone.
- (b) Question 3 and Question 4: Required by Graduate Students and Bonus for Undergrads
- (c) Question 5 to Question 6: Bonus question for both Graduate Students and Undergraduate Students

### Problem 1 (10 points)

We can represent the words in a vocabulary with binary vectors that have dimension of the number of words in the vocabulary and all values set to zero except the one value that corresponds to the index of the given word in the sorted version of this vocabulary. This is the so called one-in-K or one-hot encoding.

- (a) Describe a representation of a document with a vector. (Think of a representation that is based on the one-hot encoding of the words in that document and has the same dimension as a single word (size of the vocabulary).)
- (b) Explain why this representation is problematic:
  - (i) Simple sentence or two
  - (ii) Examples of the problem(s)
- (c) Provide atleast two more options to fix this problem.

### Problem 2 (30 points)

A recurrent network in Figure 3 takes a sequence of integers as an input and at the end of the sequence, on the last element produces a number between 0 and 1. What does a 0 mean? What does a 1 mean? Describe which function this network is computing (what is the meaning of this function). Assume all biases are 0, and make sure the hidden state is initialized to 0 as well. Note, the inputs, the weights, and the hidden state are just scalars in this RNN.

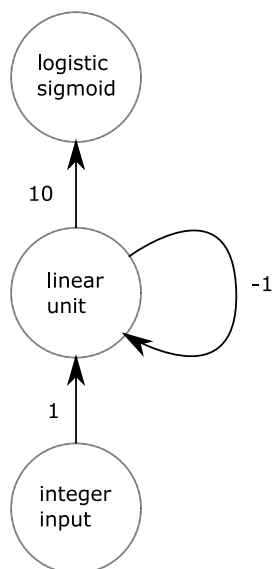


Figure 1: The RNN for Problem 2

---

Bonus for undergraduates beyond this line.

### Problem 3 (20 points)

In this problem, you will implement a recurrent neural network which implements binary addition. The inputs are given as binary sequences, starting with the least significant binary digit. (It is easier to start from the least significant bit, just like how you did addition in grade school.) The sequences will be padded with at least one zero on the end. For instance, the problem

$$100111 + 110010 = 1011001 \quad (1)$$

would be represented as :

- (a) Input 1: 1, 1, 1, 0, 0, 1, 0
- (b) Input 2: 0, 1, 0, 0, 1, 1, 0
- (c) Correct output: 1, 0, 0, 1, 1, 0, 1

There are two input units corresponding to the two inputs, and one output unit. Therefore, the pattern of inputs and outputs for this example would be:

Design the weights and biases for an RNN which has two input units, three hidden units, and one output unit, which implements binary addition. All of the units use the hard threshold activation function. In particular, specify weight

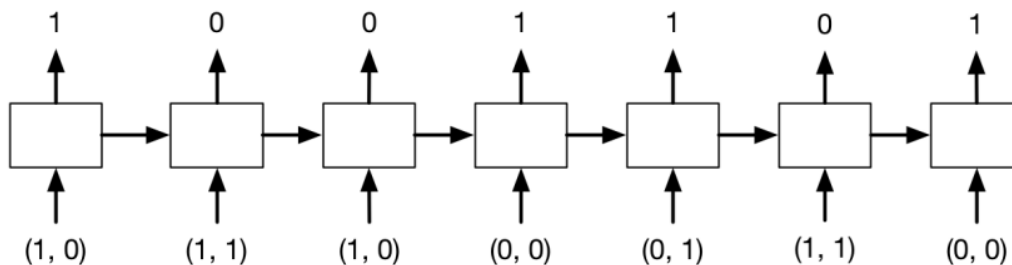


Figure 2: The RNN Binary for Problem 3

matrices  $U$ ,  $V$ , and  $W$ , bias vector  $b_h$ , and scalar bias  $b_y$  for the following architecture:

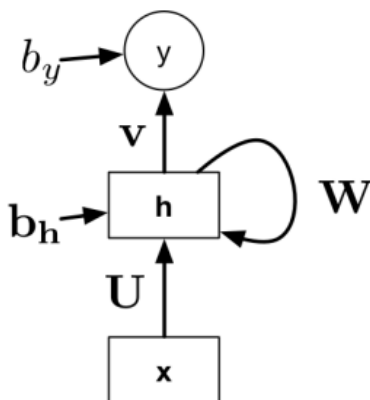


Figure 3: The RNN Architecture for Problem 3

Hint: In the grade school algorithm, you add up the values in each column, including the carry. Have one of your hidden units activate if the sum is at least 1, the second one if it is at least 2, and the third one if it is 3.

#### Problem 4 (20 points)

We have learned about regularization in image processing. How does regularization help in the context of Recurrent Neural Networks?

---

Bonus for both undergraduates and graduates beyond this line.

**Problem 5 (40 points)**

How is teacher forcing ratio more accurate than the model output for a sequence of inputs? How can we use teacher process to parallelize the computation?

**Problem 6 (40 points)**

Write a report on one of the following topics:

- (a) Attention Is All You Need {<https://arxiv.org/pdf/1706.03762.pdf>}
- (b) Transformers: {<https://arxiv.org/pdf/1910.03771v5.pdf>}