1 Question 1

1.1 Relationship between t and $\|y_t\|_2$ for $t = 0, \ldots, 15$ for your 10 sampled x0's

Solution:

Findings

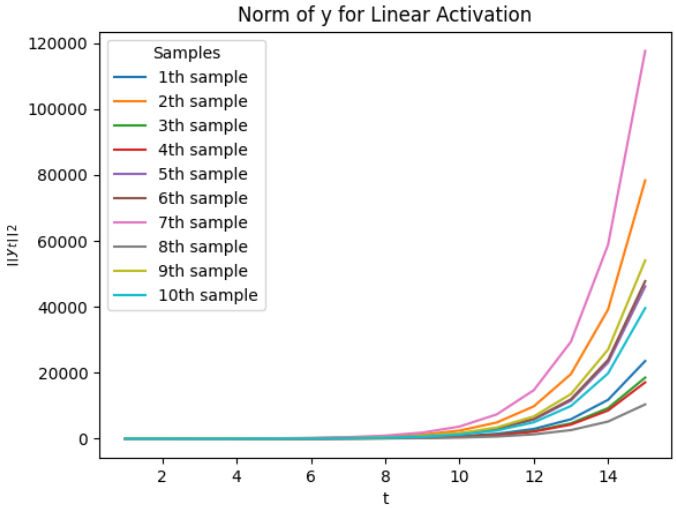

Figure 1: Linear Activation

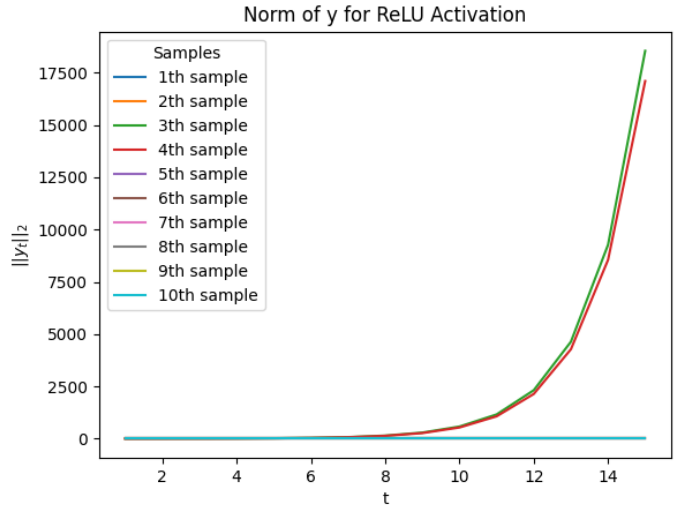

Figure 2: ReLU Activation

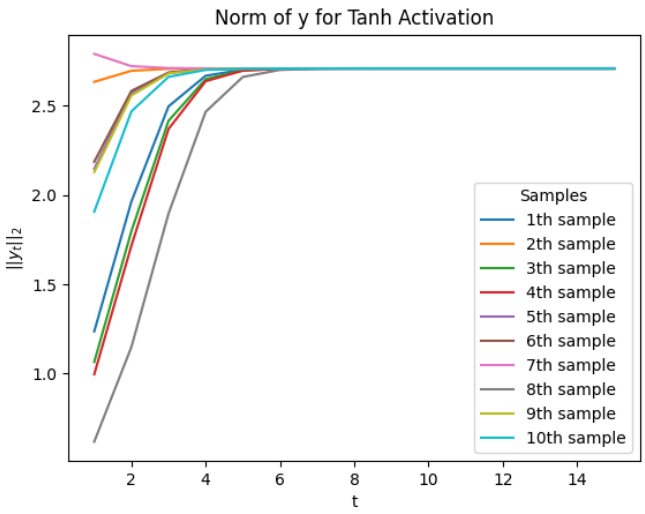

Figure 3: Tanh Activation

- As we see in figure 1, for linear activation, we see that $\|y_t\|_2$ has a wide range. It goes from values that are very close to 0, to values that are growing extremely large. This is seen in the pink and orange samples.
  We note that this can lead to numerical overflow/ underflow issues. We also see that this can lead to the problem of exploding gradients in the forward and backward passes, making the training unstable.
- As we see in Figure 2, the same linear behavior is seen for ReLU activation, which makes sense given that ReLU is linear in the positive half of the reals. However, we note that only 3/10 samples are shown in the plot, because the ReLU function squashes all the samples with negatives in them. This leads to 0 values for most samples and exponentially growing values for the remaining samples.
- As we see in figure 3, TanH is the most stable of the three activations explored here. We see that the values converge to one value for all samples as time increases.

□

1.2 Relationship between t and $\|y_t\|_2$ for $t = 0, \ldots, 15$ for x_0, x_1

where

$$x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Solution:

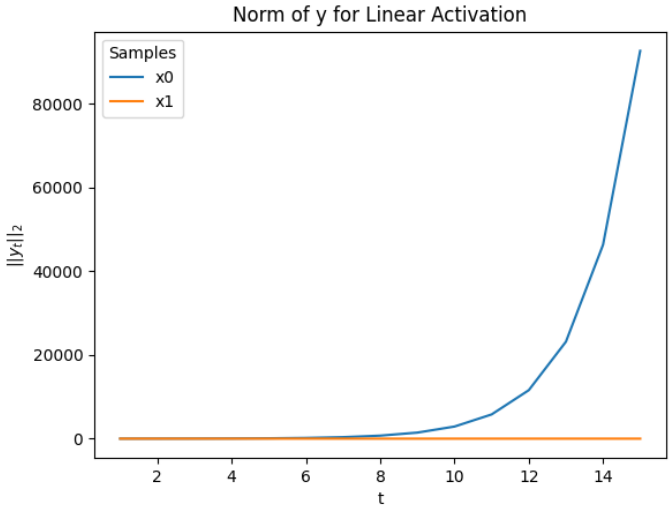

Figure 4: Linear Activation

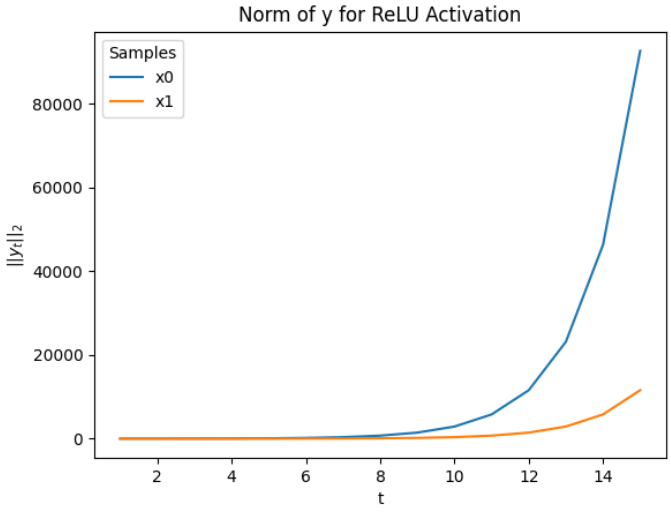

Figure 5: ReLU Activation

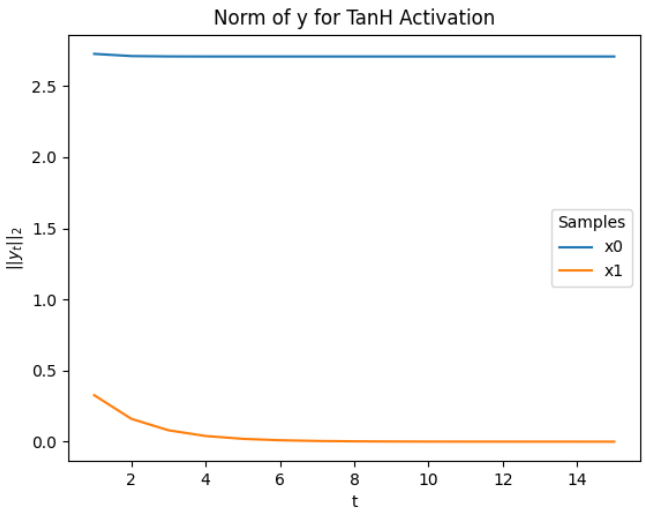

Figure 6: Tanh Activation

- Here in the figures 4, 5; we see the same trend as we saw for the previous section. The values are exploding for x_0 given both values are positive. We see a similar trend for x_1 albeit much slower growth than x_0. Thus, all the concerns we showed for the linear and ReLU activations (exponential increase hence numerical overflow and exploding gradient problems).
- TanH as seen in figure 6, gives values that are within a reasonable range of 0 to 3 and does not face the same issues as Linear and ReLU activations.

Thus, from the two studies, for different x values, we can conclude that

1. RNNs are very sensitive to the choice of activations, some of which can render to various issues like numerical over/underflow and exploding/vanishing gradients
2. TanH being the popular choice as the activation function for RNN makes sense given it is bounded and does not suffer the same issues as Linear and ReLU functions.

□

2 Question 2

1. Suppose we want the model to perform single-word sentence completion (e.g. without context, I ask a model to fill in the blank of the following sentence with a single word: "I like to eat apple after dinner."). Which evaluation metric (accuracy, PPL, or BLEU) would be appropriate and why?
2. Suppose we want the model to summarize an article. Which evaluation metric (accuracy, PPL, or BLEU) would be appropriate and why?
3. Suppose we want the model to perform sentiment analysis, the task of determining whether a sentence has a positive or negative sentiment (e.g. "This is Sam's favorite movie." has positive sentiment). Which evaluation metric (accuracy, PPL, or BLEU) would be appropriate and why?
4. Suppose we want the model to perform multiple choice question answering. Which evaluation metric (accuracy, PPL, or BLEU) would be appropriate and why?
5. What is a weakness of BLEU scores? Hint: How would BLEU and a human evaluate the similarity of these sentences: "She received the highest score." (Target) vs. "No one got more points than her." (output)

Solution:

1. For the first task, accuracy is the best method. It will directly tell us how good our single-word prediction is in comparison to the target word (gold/true label); we are thinking of this as a classification problem over the vocabulary. Neither the Blue Score nor the PPL is designed to do that.
2. BLEU score would be the best metric here because it tells us a way to evaluate the quality of machine-generated text by comparing it with the reference summaries written by humans.
3. We can think of this as a classification problem and accuracy will be the best metric for this, where our true label will be the correct sentiment.
4. For the multiple choice question answering, accuracy is the most suitable metric. We can use the correct answer to benchmark how our model is performing
5. Weakness of BLEU scores: If we were to consider the hint given: "She received the highest score." (Target) vs. "No one got more points than her." (output)
   A human can clearly tell that they mean the same thing. However, if we were to consider the exact match of the n-grams, we would get a BLEU score of 0.
   Therefore, we see that the weakness of the BLEU score is that it does not account for semantic similarity or paraphrasing that conveys the same meaning.

□

3 Question 3

1. To check your understanding of the model, please answer the following in your report:

   - (a) How are words encoded as vectors?

     Solution: They are encoded as one-hot vectors (one at the index of the word, 0 elsewhere). □

   - (b) What is the maximum sentence length that is used to train this model?

     Solution: `MAX_LENGTH = 10` Thus, the maximum sentence length that is used to train this model is 10. □

   - (c) What do we feed into the encoder?

     Solution: French sentences □

   - (d) What is produced by the decoder?

     Solution: English Sentences
     It takes in the context vector from the encoder and outputs at each time step. □

   - (e) How many samples are in the original and trimmed datasets?

     Solution: 135842 sentence pairs in the original and 11445 sentence pairs in the trimmed □

   - (f) How are sentences processed before feeding them into our BLEU-k score implemenation?

     Solution:
     - convert input sentence and the target sentence tensor using the `tensorFromSentence`
     - pass the input tensor to Encoder
     - pass the output and the hidden states of the encoder to the decoder
     - Get the topk= 1 (most probable) token and go over till EOS (end of sentence) token is hit.
     - Feed the concatenated input to our blue score function calculation from question 2. □

2. Use your BLEU score implementation to evaluate the performance of the model on the given evaluation data set. Include in your report the BLEU-k scores for k = 1, 2, 3, 4, 5 rounded to 4 decimal places. Do you notice a trend?

Solution:

BLEU-1 score	0.96045
BLEU-2 score	0.92505
BLEU-3 score	0.89651
BLEU-4 score	0.76388
BLEU-5 score	0.56437

Table 1: Blue-k score for k from 1 to 5

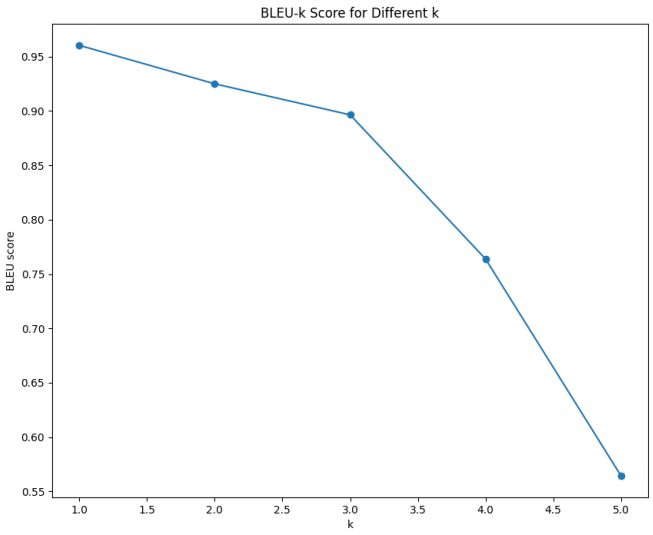

Figure 7: Blue-k Scores with respect to different k values

We see a clear trend that the Bleu score decreases as the k values increase. This is because as k increases, the length of sequences that we are matching (exact matching in the case of bleu score) is increasing. The longer the sequence, the probability of them matching decreases, and hence the Bleu score decreases. This is why we see the trend we see in Figure 9 and Table 1 □

3. Include in your report 5 example triples of inputs, targets, and model outputs.

Solution:

3.1.
```
> elles vont le tenter
= they re going to try
< they re going to try <EOS>
```

3.2.
```
> je suis desolee mais je ne comprends pas
= i m sorry but i don t understand
< i m sorry but i don t understand <EOS>
```

3.3. `> nous sommes la pour jouer au basket`
`= we re here to play basketball`
`< you re here to play basketball <EOS>`

3.4. `> je suis prete si tu l es`
`= i m ready if you are`
`< i m ready if you are <EOS>`

3.5. `> tu es fort raffinee`
`= you re very sophisticated`
`< you re very sophisticated <EOS>`

□