ITCS 6101/8101 Homework 3: Theory (50 points)

February 19, 2026

1 Universal Approximation (30 points)

Let NN be a neural network with 2 input units, 1 hidden layer with h ReLU neurons, and 1 output unit using sigmoid as output function. Furthermore, consider a training set that contains the following 4 examples:

x_1	x_2	y
1	-1	0
1	0	1
2	-1	1
2	0	0

1. (10p)What is the minimum h for which there is a network with h hidden neurons that fits the training data? Show the network, its weights, and prove that it fits the dataset above.

2. (5p) Implement the dataset and the network with the manual weights above in PyTorch and show that it fits the dataset.

3. (5p) Consider the same neural network NN, but without the *ramp* activation function in the hidden layer. Answer the same questions as at item (1) above.

4. (10p): Train the neural network in PyTorch such that it fits the dataset above, starting from a random initialization of parameters. You may have to increase the number of neurons on the hidden layer to make it work (but try to keep it as small as possible).

2 Language Modeling and Sampling (20 points)

1. (10p) Consider the FCN approach to language modeling shown on slides 14 and 15 in lecture 6, also used in the homework assignment. Assume K words are used as input, with word embeddings of size E each. The FCN has H hidden layers with N neurons each, and the vocabulary has size V. What is the total number of parameters required by this network? Show your calculation.

2. (5p) Show that the perplexity of a language model on a test sequence of N tokens is equal with the exponential of the average cross-entropy computed on the same sequence of tokens.

3. (5p) Suppose you want to use an LM to generate the next token, but you want the LM to never generate a particular token, even when it has the highest logit score. Suppose the logits scores are computed in the array `logits` and that the index of this token

is `k`. What value should `logits[k]` be changed to such that when computing the token probabilities `p = softmax(logits)`, this token will never be generated (sampled), whether we use greedy or multinomial sampling.

3 Submission

Submit your theory responses on Canvas as one file named `hw03-theory.pdf` and the PyTorch solutions as one file named `hw03-pytorch.py`. It is important that you show clearly all the derivation steps. We recommend using an editor such as Overleaf for Latex, Word, or Jupyter-Notebook that allows proper formatting of equations. Alternatively, if you choose to write your solutions on paper, submit an electronic scan / photo of it exported to **PDF**. Make sure that your writing is **legible** and the scan has good quality (we will not grade solutions that we struggle to read).