ITCS 6101/8101 Homework 4: Theory (50 points)

March 19, 2026

1 Transformer: PreNorm Equations (25 points)

In class, we have shown the *token-wise* computation graph equations for the `PreNorm` Transformer block on slides 33 (version 1) and 34 (version 2), and the *vectorized* and computation graph equations for the `PostNorm` Transformer block on slide 46. Show the analogous vectorized computation graph equations for the `PreNorm` Transformer block (version 2), and the token-wise computation graph equations for the `PostNorm` Transformer lock.

2 Transformer: Parameter Counts (25 points)

Consider the Transformer architecture with number of layers, number of attention heads, embedding size, and vocabulary size as specified by the default values from the implementation part of the homework assignment. Assume since/cosine positional endodings. Compute the total number of parameters in this Transformer, separating the embedding parameters from the rest of the parameters. Show your work in detail, by mapping the parameter counts to the computation graph equations shown at Section 1 above. Verify that your total counts are correct by using PyTorch to print the total number of parameters.

3 Submission

Submit your theory responses on Canvas as one file named `hw05-theory.pdf`. 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).