CSE 527A: Assignment 2

Due: October 6 (Thursday), 2022

Notes:

- Please submit your homework via Gradescope.
- You can either submit a legibly handwritten or LaTeX generated pdf.
- Make sure you **specify the pages for each problem correctly**. You **will not get points** for problems that are not correctly connected to the corresponding pages.
- Homework is due **by one hour later after the due date**.
- Please keep in mind the collaboration policy as specified in the **Academic Integrity** section of the course syllabus.
- There are 4 problems in the written portion of this assignment and 3 for the coding half.

Problems:

1. (Eisenstein Ch. 18) You are given the following dataset of translations from "simple" to "difficult" English:

   a. Kids like cats
      Children adore felines

   b. Cats hats
      Felines fedoras

   Estimate a word-to-word statistical translation model from simple English (source) to difficult English (target), using the expectation-maximization. Compute two iterations of the algorithm by hand, starting from a uniform translation model, and using the simple alignment model $p(a_j|j, J^{(s)}, J^{(t)}) = \frac{1}{J^{(t)}}$. Hint: in the final M-step, you will want to switch from fractions to decimals.

2. (Eisenstein Ch. 18) Building on the problem 1, what will be the converged translation probability table? Can you state a general condition about the data, under which this translation model will fail in the way that it fails here?

3. (Eisenstein Ch. 18) Let $l_{j+1}^{(t)}$ represent the loss at word $j+1$ of the target, and let $h_n^{(s)}$ represent the hidden state at word n of the source. Write the expression for the derivative $\frac{\partial l_{j+1}^{(t)}}{\partial \boldsymbol{h}_n^{(s)}}$ in the sequence-to-sequence translation model expressed as:

   $$\boldsymbol{h}_j^{(s)} = \text{LSTM}\left(\boldsymbol{x}_j^{(s)}, \boldsymbol{h}_{j-1}^{(s)}\right)$$
   $$\boldsymbol{z} \triangleq \boldsymbol{h}_{J^{(s)}}^{(s)}$$

   where $\boldsymbol{x}_j^{(s)}$ is the embedding of source language word $w_j^{(s)}$. The encoding then provides the initial hidden state for the decoder LSTM:

   $$\boldsymbol{h}_0^{(t)} = \boldsymbol{z}$$
   $$\boldsymbol{h}_j^{(t)} = \text{LSTM}\left(\boldsymbol{x}_j^{(t)}, \boldsymbol{h}_{j-1}^{(t)}\right)$$

   where $\boldsymbol{x}_j^{(t)}$ is the embedding of the target language word $w_j^{(t)}$.

   You may assume that both the encoder and decoder are one-layer LSTMs. In general, how many terms are on the shortest backpropagation path from $l_{j+1}^{(t)}$ to $\boldsymbol{h}_n^{(s)}$?

4. (Eisenstein Ch. 18) Consider the neural attention model with sigmoid attention. The derivative $\frac{\partial l_{j+1}^{(t)}}{\partial z_n}$ is the sum of many paths through the computation graph; identify the shortest such path. You may assume that the initial state of the decoder recurrence $h_0^{(t)}$ is not tied to the final state of the encoder recurrence $h_{J^{(s)}}^{(s)}$.