**Bioinformatics 529 Homework 3**

**Amelia Lauth**

**03/16/22**

**Resources used (Including websites, partners in class, etc.):**

Partners in Class: Crystal Wen, Elysia Chou, Catherine Barnier, Jianhui Gong, Matthew Pun, Brad Crone

<http://www.cs.cmu.edu/~guestrin/Class/10701-S07/Slides/baumwelch.pdf>

<http://www.adeveloperdiary.com/data-science/machine-learning/derivation-and-implementation-of-baum-welch-algorithm-for-hidden-markov-model/>

<http://web.mit.edu/6.047/book-2012/Lecture08_HMMSII/Lecture08_HMMSII_standalone.pdf>

<https://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm#Update>

<https://www.delftstack.com/howto/python/one-line-for-loop-python/>

<https://www.analyticsvidhya.com/blog/2021/05/complete-guide-to-expectation-maximization-algorithm/>

<https://people.cs.umass.edu/~mccallum/courses/inlp2004a/lect10-hmm2.pdf>

<https://datagy.io/python-get-dictionary-key-with-max-value/>

<https://www.youtube.com/watch?v=NeC1ZC0Nwvw>

[**https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization\_algorithm**](https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm)

[**https://docs.python.org/3/library/copy.html**](https://docs.python.org/3/library/copy.html)

[**https://www.geeksforgeeks.org/accessing-attributes-methods-python/**](https://www.geeksforgeeks.org/accessing-attributes-methods-python/)

**Statement of Objective (What was the purpose of this homework assignment):**

The purpose of homework #3 is to build the Baum-Welch algorithm that will learn the probabilities (state initiation, transition, and emission probabilities) that were generated from prior knowledge and used in the Viterbi, Forward, Backward, and Forward-Backward algorithms. The Baum-Welch algorithm is powerful because it will learn these probabilities for our model in an unsupervised manner, and can be used when we have little training data!

**Procedure (Explain in general terms how you went about implementing the homework assignment):**

I followed the series of steps laid out by Alan: initialization, expectation, maximization, and termination. Since it is an expectation-maximization algorithm, I knew I had to loop between the expectation and maximization steps until convergence was met. I followed the math equations to the best of my ability and had to keep track of all the indices.

**Difficulties and Roadblocks (What were the pain points in the implementation of this homework assignment):**

I had difficulty the summation of the (1/P(xj)). I later realized that P(xj) served as the probability of the sequence, so I had to sum up that probability over all the sequences and use that in the rest of the equation in the expectation step. Still getting used with classes and objects, so I decided to initialize the matrices respectively and set them equal to the pseudocount to ensure that I would not get a zero in the denominator. Another difficulty was keeping track of all the indices/summations and making sure I was indenting code correctly under certain for loops to include the summation. Other times, I had to indent to run the math equation correctly with all the summations set equal to variables outside of specific for loops. I also struggled with grabbing the values within the nested dictionaries for the maximization step, so I researched for loops coded on one line. More efficient but hard to understand at first. It was tricky to know when to divide the Pxj. I first did it while under the “for seq in sequences” loop but got large probability outputs. So, I chose to break the equation apart and just divide by Pxj outside of the for loop and it worked. Still confused by that. Lastly, the instructions for convergence were confusing with checking if no change occurred. With limited time, I chose an arbitrary threshold of iterations and ran a for loop through the range at the very beginning.