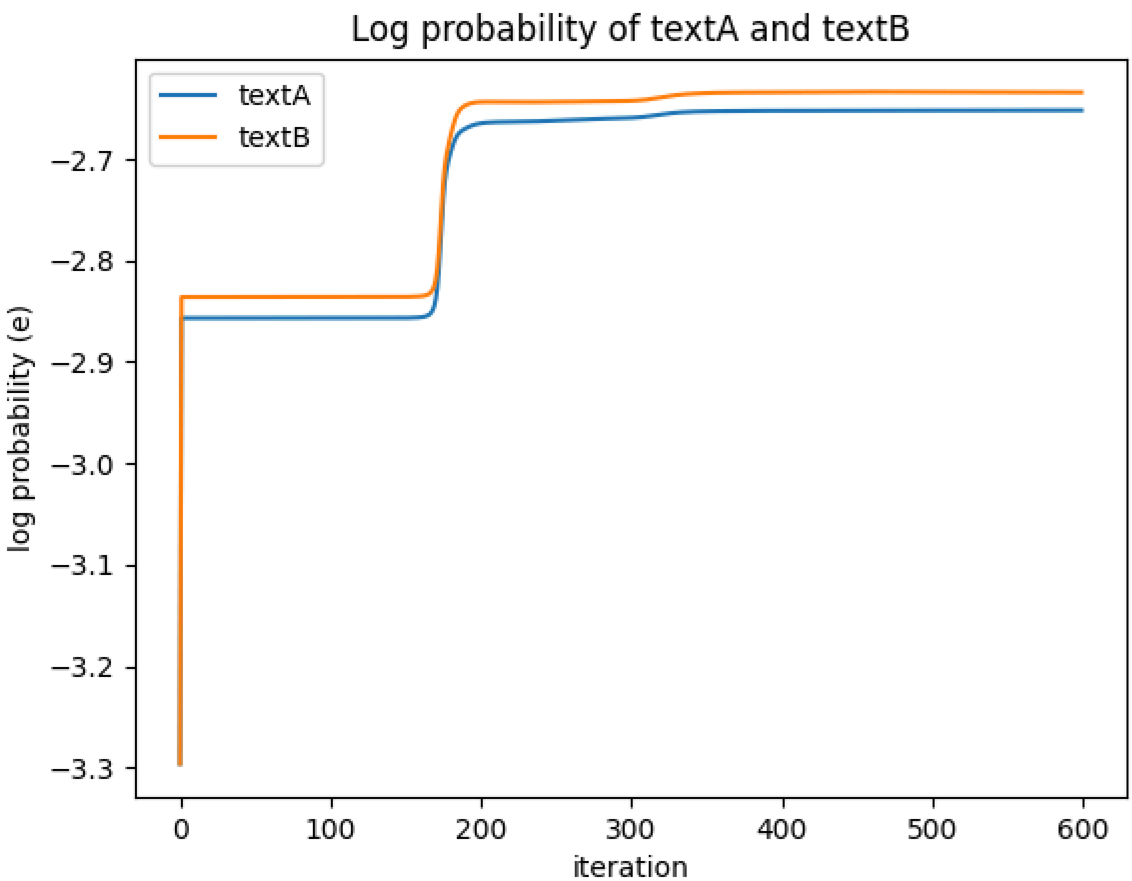
Project 1 Writeup

Problem 1

a

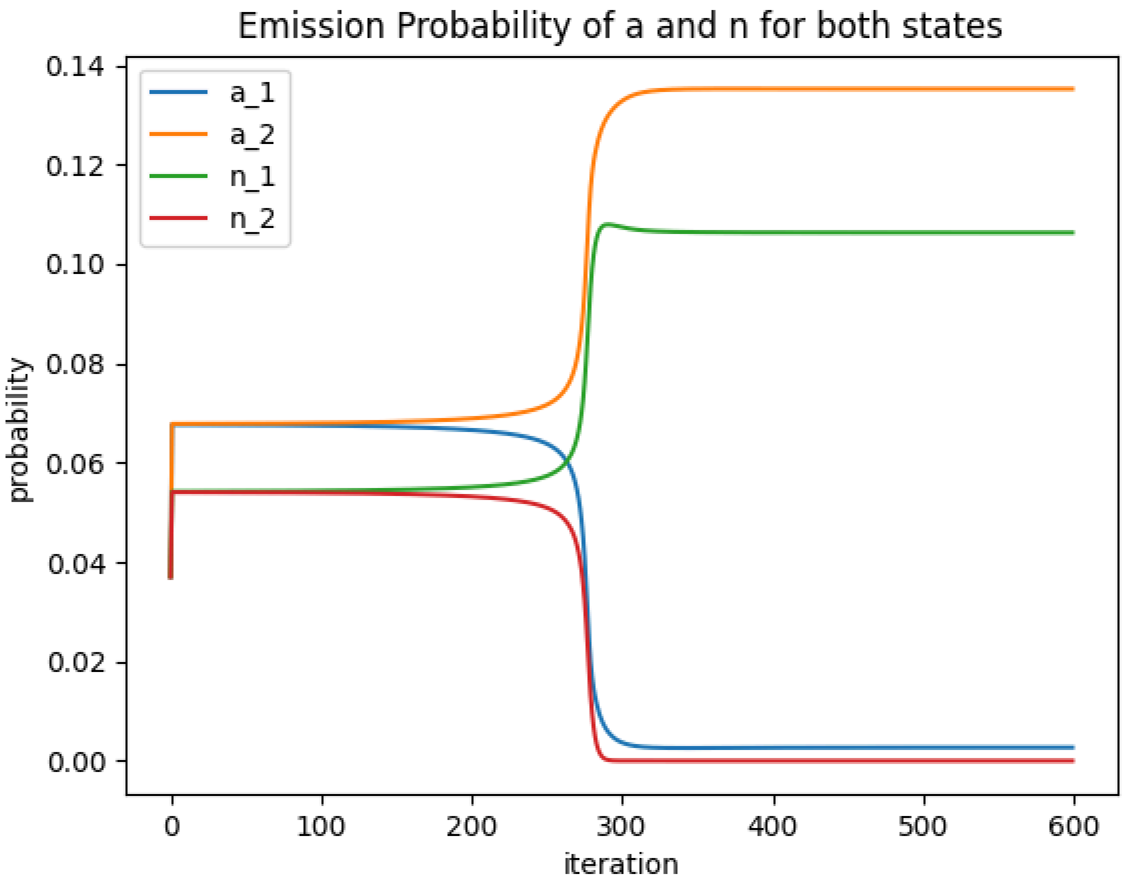
If we set probability to uniform, the state transition probability stays constant and the emission probability becomes unigram character probability of the corpus.

b



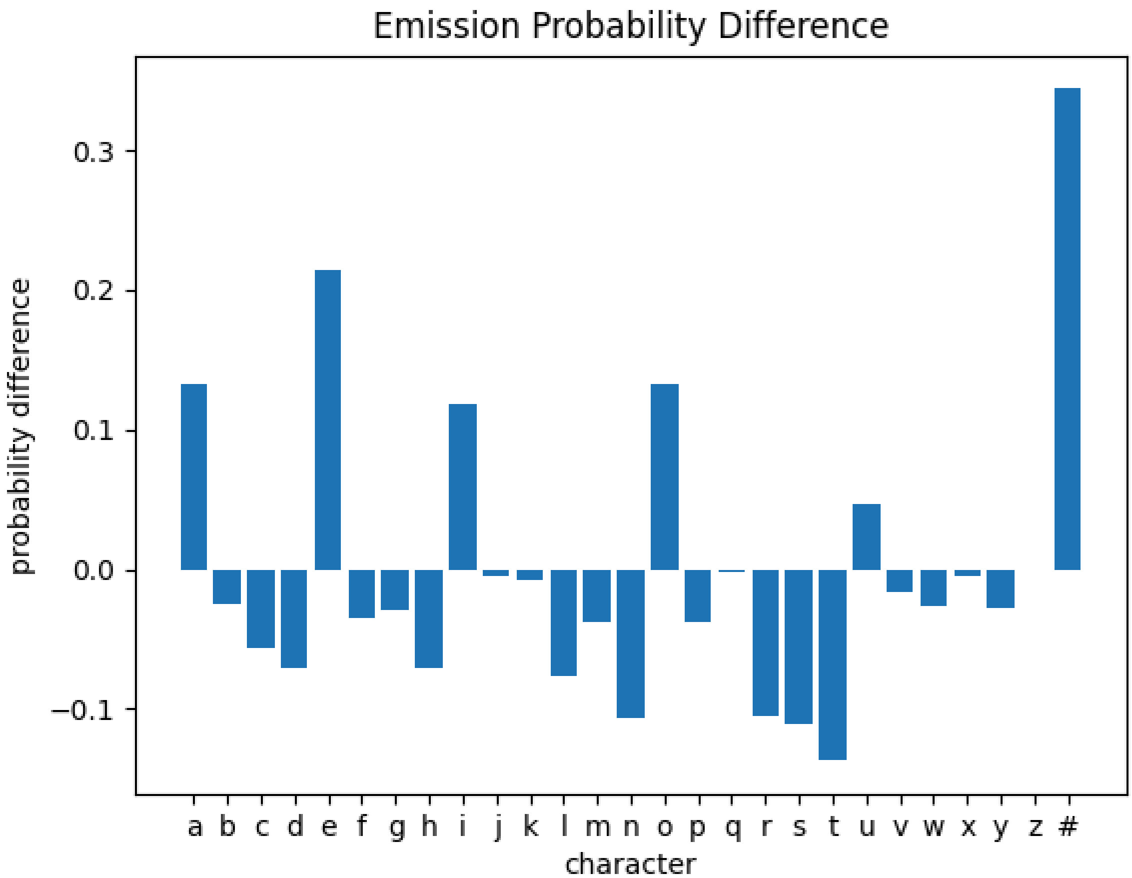
c

We see that probability of emission tend to go opposite for different states. a_1 goes to zero for



d

Below graph shows the emission probability difference between state 1 and state 2. (q(y | 2) - q(y|1))



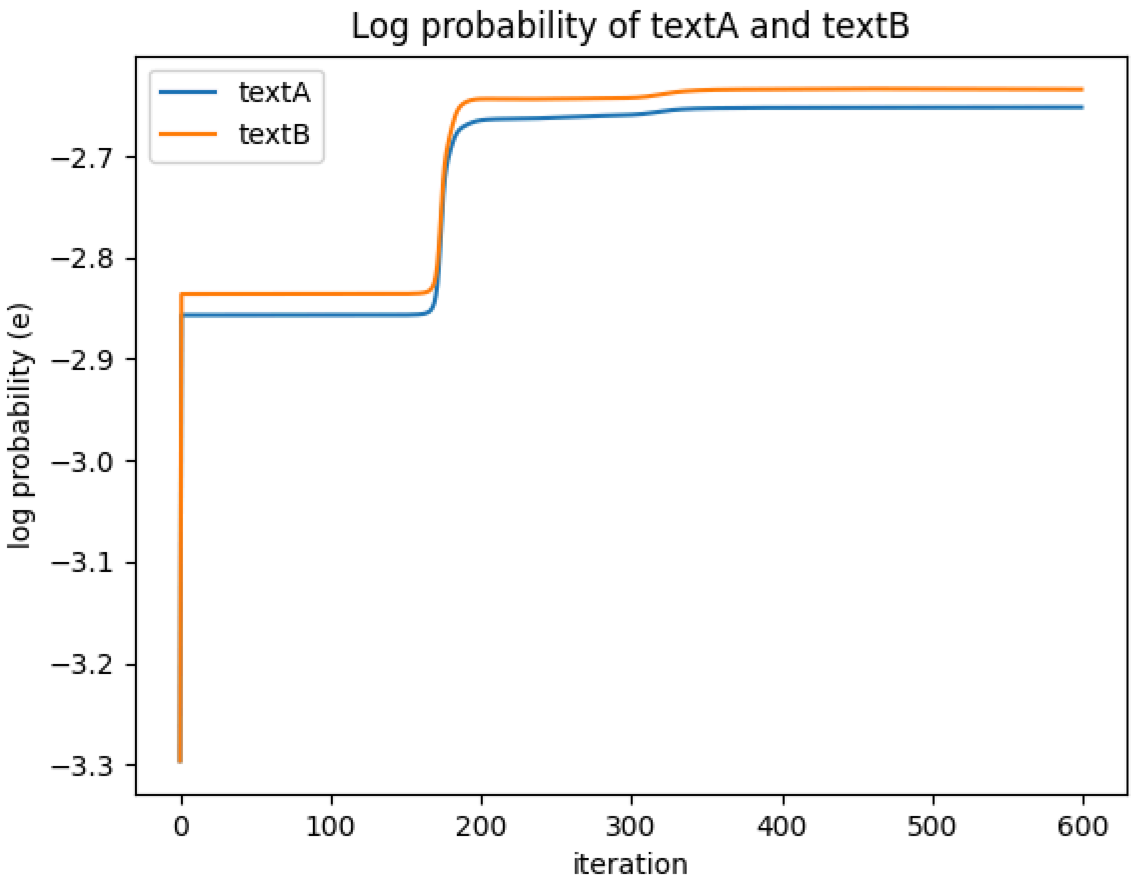
We see that emission probability for vowels and space (#) is higher for state 2 compared to state 1 while state 1 has higher emisssion probability for rest of the alphabet compared to state 2.

2

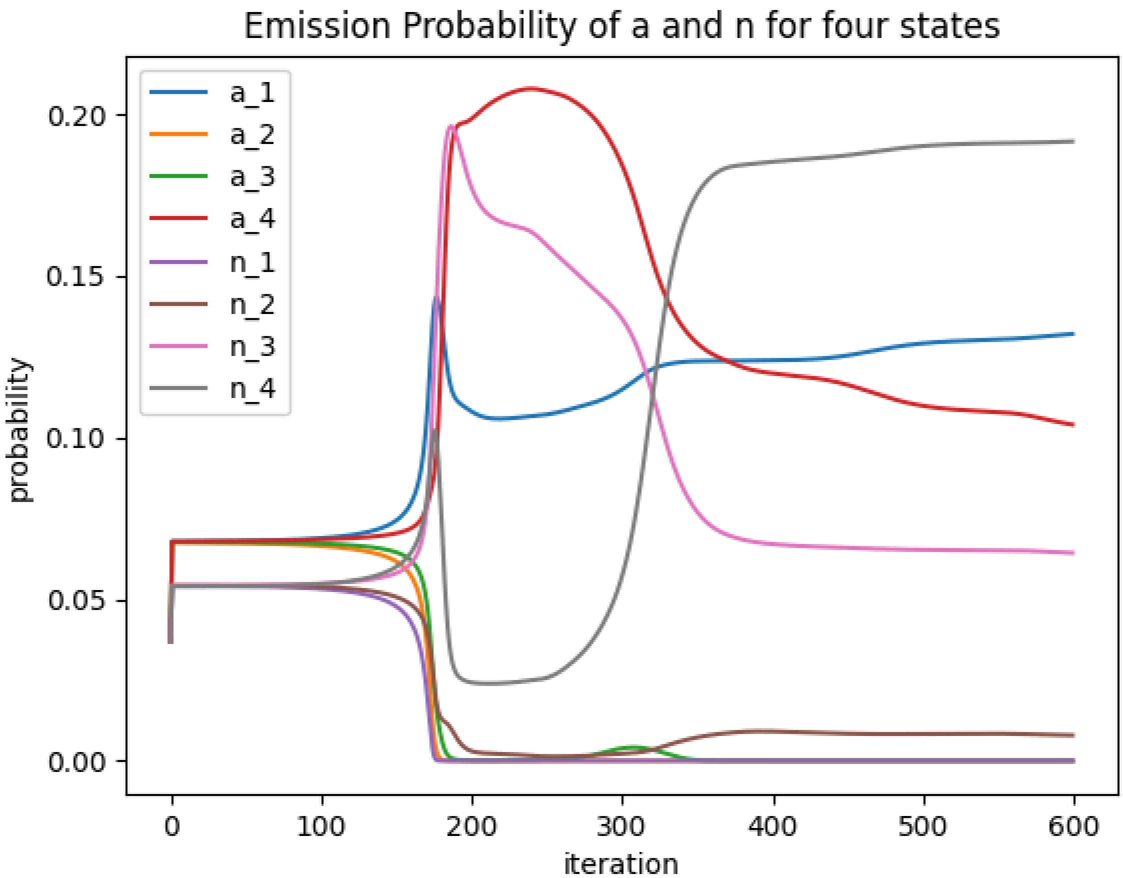
a

Similar to part 1, the transition probability remains constant and each emission probability, regardless of the state look the same if the probability is uniformly initialized.

b

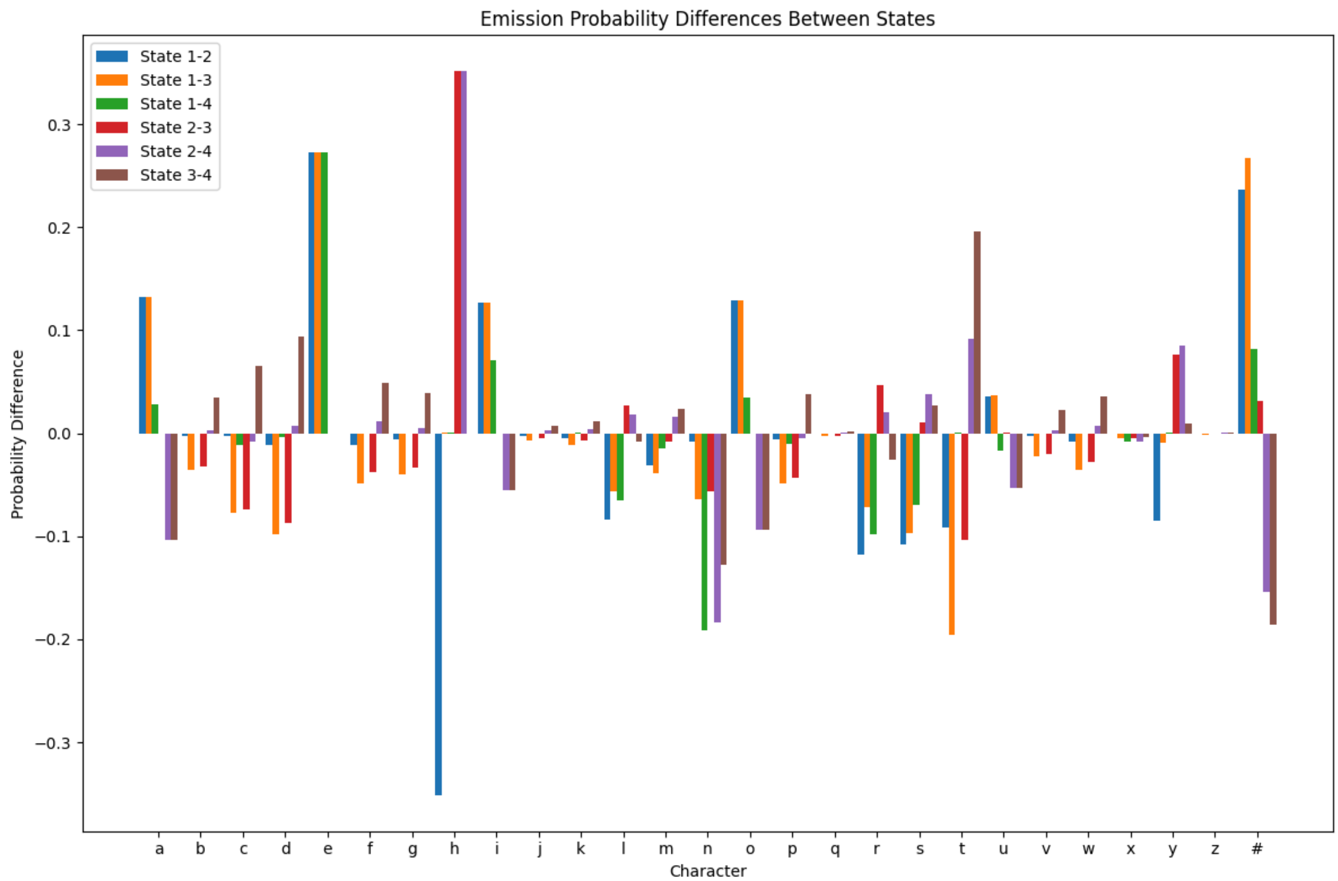

Log probability of textA and textB
textA
textB
−2.7
−2.8
−2.9
−3.0
−3.1
−3.2
−3.3
log probability (e)
0
100
200
300
400
500
600
iteration

c



Emission Probability of a and n for four states
a_1
a_2
a_3
a_4
n_1
n_2
n_3
n_4
probability
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
500
600
iteration

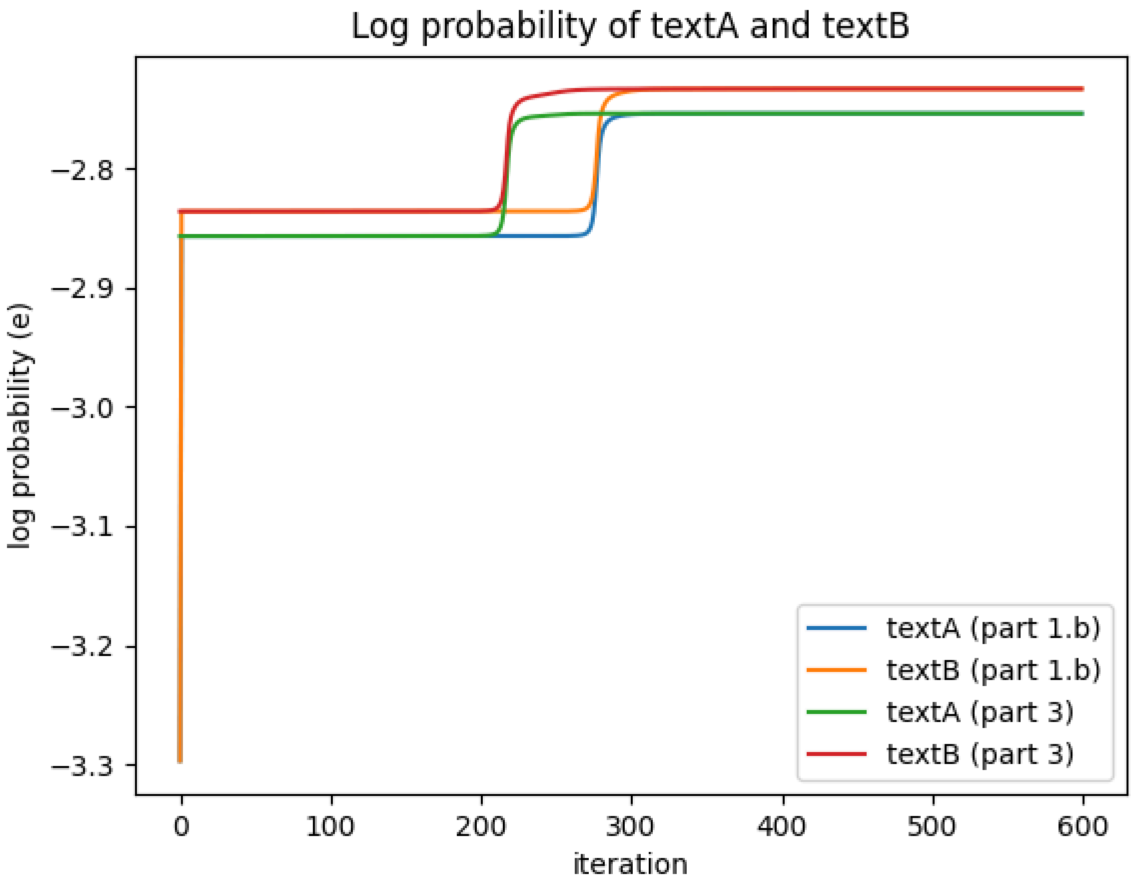
d



In here, legend "State a-b" means (p(y|a) - p(y|b))

Unlike 2-state hmms, we don't see a clear trend among the differences. We still see that vowels are more likely to be emitted on some nodes while consonants are more likely to be emitted on other nodes. The largest difference between the nodes still happens around the vowels.

Part 3



IN this graph, we see that the initialization technique used in part 3 results in a faster convergence compared to the initialization technique in part 1.

The final average log probs are as follows

alternative logprob textA: -2.753
alternative logprob textB: -2.733
part 1.b logprob textA: -2.753
part 1.b logprob textB: -2.733

The final average log probs are very close to equal.