**Discrete Hidden Markov Model Implementation**

工海三 柯哲邦 B05505053

* **Environment:**

Macbook pro 2016 late

CPU - 2 GHz Intel Core i5

RAM – 8 GB 1867 MHz LPDDR3

g++ compiler – Apple LLVM version 10.0.0 (clang-1000.11.45.2)

* **Compile:**

$ make : to compile *test\_hmm.cpp* & *train\_hmm.cpp*

$ make clean : to remove the executable files test & train

* **Execute:**

$ ./train #iter *model\_init.txt* *seq\_model\_0?.txt* *model\_0?.txt*

(#iter: the iteration number to train, ‘?’ = 1-5)

$ ./test *modellist.txt* *testing\_data?.txt* *result?.txt*

(‘?’ = 1-2)

* **Result:**

因為不同的iteration次數產生的model不同，也會因此影響到test的accuracy。所以本次實作觀察accuracy如何隨#iteration 而改變。(另寫了一個檔案去算accuracy)

數據顯示，當#iteration = 10時，accuracy會是最低點，之後就會回升，然後在#iteration = 880之後都趨於穩定。也在#iteration 880 達到最高峰

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #iteration | 1 | 10 | 50 | 100 | 300 | 500 | 700 | 850 | 880 | 1000 |
| accuracy | 0.766 | 0.540 | 0.822 | 0.810 | 0.848 | 0.856 | 0.8656 | 0.8692 | 0.8696 | 0.8696 |

Best\_Model:

#iteration = 880

Initial model = model\_init.txt

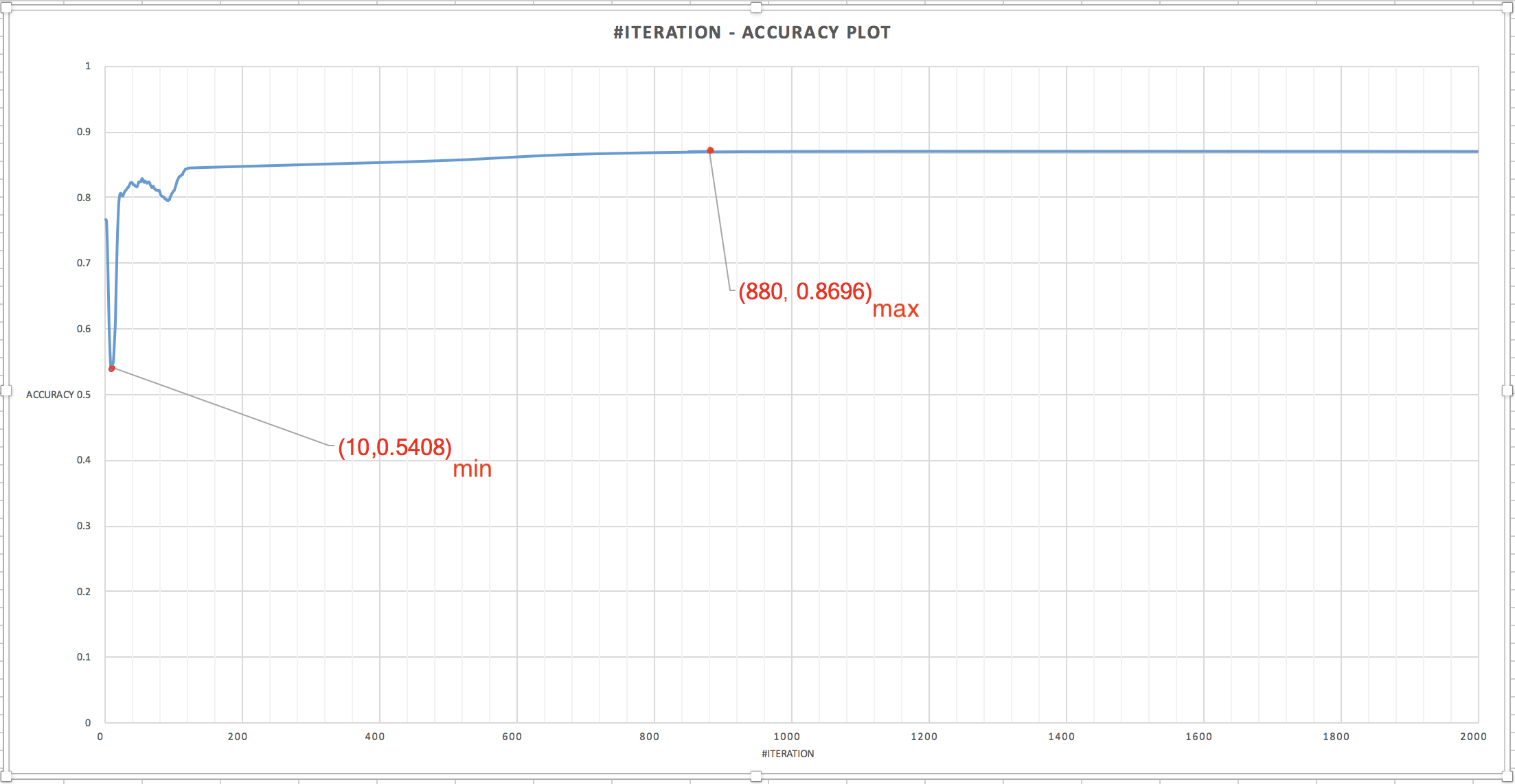
Accuracy of result1 to testing\_data1.txt = 0.8696

Worst\_Model:

#iteration = 10

Initial model = model\_init.txt

Accuracy of result1 to testing\_data1.txt = 0.5400



* **Additional experiment:**

考慮到iteration的次數會影響到model的大小，就會想到那如果我不取整個seq\_model，而是只取片段的seq\_model來train，比較完整和片段的結果。

Batch size: 1000 (only take 1000 sequences of the seq\_model\_0?.txt) and #iteration = 50:

Accuracy = 0.8153

* **Conclusion:**

#iteration的變化相較batch的size更能影響機率。可以看到，#iteration的不同，影響的機率最大可以達0.3，而batch size取的不同影響非常些微。

* **Problems to discuss:**

1. The best way to calculate P(O|lambda).
2. Auto adjust the HMM parameters to avoid overfitting.
3. Compare the performance of the EM version and convex optimization version of Baum-Welch.
4. Is using small batch update still coverage?