8. To iterate or not to iterate, that is the question.

```
iters = iters + 1
δ = |logProb − oldLogProb|
if(iters < minIters or δ > ε) then
    oldLogProb = logProb
    goto 3.
else
    return λ = (A, B, π)
end if
```

2.9 The Bottom Line

Hidden Markov models are powerful, efficient, and extremely useful in practice. Virtually no assumptions need to be made, yet the HMM process can extract significant statistical information from data. Thanks to efficient training and scoring algorithms, HMMs are practical, and they have proven useful in a wide range of applications. Even in cases where the underlying assumption of a (hidden) Markov process is questionable, HMMs are often applied with success. In Chapter 9 we consider selected applications of HMMs. Most of these applications are in the field of information security.

In subsequent chapters, we often compare and contrast other machine learning techniques to HMMs. Consequently, a clear understanding of the material in this chapter is crucial before proceeding with the remainder of the book. The homework problem should help the dedicated reader to clarify any remaining issues. And the applications in Chapter 9 are highly recommended, with the English text example in Section 9.2 being especially highly recommended.

2.10 Problems

When faced with a problem you do not understand,
do any part of it you do understand, then look at it again.
— Robert Heinlein

1. Suppose that we train an HMM and obtain the model $\lambda = (A, B, \pi)$ where

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}, \quad B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{pmatrix}, \quad \pi = \begin{pmatrix} 0.0 & 1.0 \end{pmatrix}.$$

Furthermore, suppose the hidden states correspond to H and C, respectively, while the observations are S, M, and L, which are mapped to 0, 1, and 2, respectively. In this problem, we consider the observation sequence $\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1, \mathcal{O}_2) = (M, S, L) = (1, 0, 2)$.

a) Directly compute $P(\mathcal{O} \mid \lambda)$. That is, compute

$$P(\mathcal{O} \mid \lambda) = \sum_X P(\mathcal{O}, X \mid \lambda)$$

using the probabilities in $\lambda = (A, B, \pi)$ for each of the following cases, based on the given observation sequence $\mathcal{O}$.

$$P(\mathcal{O}, X = HHH) = \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$
$$P(\mathcal{O}, X = HHC) = \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$
$$P(\mathcal{O}, X = HCH) = \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$
$$P(\mathcal{O}, X = HCC) = \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$
$$P(\mathcal{O}, X = CHH) = \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$
$$P(\mathcal{O}, X = CHC) = \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$
$$P(\mathcal{O}, X = CCH) = \underline{1.0} \cdot \underline{0.2} \cdot \underline{0.6} \cdot \underline{0.7} \cdot \underline{0.4} \cdot \underline{0.5} = \_\_\_\_\_$$
$$P(\mathcal{O}, X = CCC) = \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$

The desired probability is the sum of these eight probabilities.

b) Compute $P(\mathcal{O} \mid \lambda)$ using the α pass. That is, compute

$$\alpha_0(0) = \_\_\_ \cdot \_\_\_ = \_\_\_\_\_$$
$$\alpha_0(1) = \underline{1.0} \cdot \underline{0.2} = \_\_\_\_\_$$
$$\alpha_1(0) = (\_\_\_ \cdot \_\_\_ + \_\_\_ \cdot \_\_\_) \cdot \_\_\_ = \_\_\_\_\_$$
$$\alpha_1(1) = (\_\_\_ \cdot \_\_\_ + \_\_\_ \cdot \_\_\_) \cdot \_\_\_ = \_\_\_\_\_$$
$$\alpha_2(0) = (\_\_\_ \cdot \_\_\_ + \_\_\_ \cdot \_\_\_) \cdot \_\_\_ = \_\_\_\_\_$$
$$\alpha_2(1) = (\_\_\_ \cdot \_\_\_ + \_\_\_ \cdot \_\_\_) \cdot \_\_\_ = \_\_\_\_\_$$

where we initialize

$$\alpha_0(i) = \pi_i b_i(\mathcal{O}_0), \text{ for } i = 0, 1, \ldots, N-1$$

and the recurrence is

$$\alpha_t(i) = \left(\sum_{j=0}^{N-1} \alpha_{t-1}(j) a_{ji} \right) b_i(\mathcal{O}_t)$$

for $t = 1, 2, \ldots, T-1$ and $i = 0, 1, \ldots, N-1$. The desired probability is given by

$$P(\mathcal{O} \mid \lambda) = \sum_{i=0}^{N-1} \alpha_{T-1}(i).$$

c) In terms of N and T, and counting only multiplications, what is the work factor for the method in part a)? What is the work factor for the method in part b)?

2. For this problem, use the same model λ and observation sequence $\mathcal{O}$ given in Problem 1.

   a) Determine the best hidden state sequence (X_0, X_1, X_2) in the dynamic programming sense.

   b) Determine the best hidden state sequence (X_0, X_1, X_2) in the HMM sense.

3. Summing the numbers in the "probability" column of Table 2.2, we find $P(\mathcal{O} \mid \lambda) = 0.009629$ for $\mathcal{O} = (0, 1, 0, 2)$.

   a) By a similar direct calculation, compute $P(\mathcal{O} \mid \lambda)$ for each observation sequence of the form $\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3)$, where $\mathcal{O}_i \in \{0, 1, 2\}$. Verify that $\sum P(\mathcal{O} \mid \lambda) = 1$, where the sum is over the observation sequences of length four. Note that you will need to use the probabilities for A, B, and π given in equations (2.4), (2.5), and (2.6) in Section 2.2, respectively.

   b) Use the forward algorithm to compute $P(\mathcal{O} \mid \lambda)$ for the same observation sequences and model as in part a). Verify that you obtain the same results as in part a).

4. From equation (2.9) and the definition of $\alpha_t(i)$ in equation (2.10), it follows that

$$\alpha_t(i) = \sum_X \pi_{X_0} b_{X_0}(\mathcal{O}_0) a_{X_0,X_1} b_{X_1}(\mathcal{O}_1) \cdots a_{X_{t-2},X_{t-1}} b_{X_{t-1}}(\mathcal{O}_{t-1}) a_{X_{t-1},i} b_i(\mathcal{O}_t)$$

   where $X = (X_0, X_1, \ldots, X_{t-1})$. Use this expression for $\alpha_t(i)$ to directly verify the forward algorithm recurrence

$$\alpha_t(i) = \left(\sum_{j=0}^{N-1} \alpha_{t-1}(j) a_{ji} \right) b_i(\mathcal{O}_t).$$

5. As discussed in this chapter, the forward algorithm is used solve HMM Problem 1, while the forward algorithm and backward algorithm together are used to compute the gammas, which are then used to solve HMM Problem 2.

   a) Explain how you can solve HMM Problem 1 using the backward algorithm instead of the forward algorithm.

b) Using the model $\lambda = (A, B, \pi)$ and the observation sequence $\mathcal{O}$ in Problem 1, compute $P(\mathcal{O} \mid \lambda)$ using the backward algorithm, and verify that you obtain the same result as when using the forward algorithm.

6. This problem deals with the Baum-Welch re-estimation algorithm.

   a) Write the re-estimation formulae, as given in lines 3, 7, and 12 of Algorithm 2.3, directly in terms of the $\alpha_t(i)$ and $\beta_t(i)$.

   b) Using the re-estimation formulae obtained in part a), substitute the scaled values $\widehat{\alpha}_t(i)$ and $\widehat{\beta}_t(i)$ for $\alpha_t(i)$ and $\beta_t(i)$, respectively, and show that the resulting re-estimation formulae are exact.

7. Instead of using c_t to scale the $\beta_t(i)$, we can scale each $\beta_t(i)$ by

$$d_t = 1 \Big/ \sum_{j=0}^{N-1} \widetilde{\beta}_t(j)$$

   where the definition of $\widetilde{\beta}_t(i)$ is analogous to that of $\widetilde{\alpha}_t(i)$ as given in Algorithm 2.6.

   a) Using the scaling factors c_t and d_t show that the Baum-Welch re-estimation formulae in Algorithm 2.3 are exact with $\widehat{\alpha}$ and $\widehat{\beta}$ in place of α and β.

   b) Write $\log\big(P(\mathcal{O} \mid \lambda)\big)$ in terms of c_t and d_t.

8. When training, the elements of λ can be initialized to approximately uniform. That is, we let $\pi_i \approx 1/N$ and $a_{ij} \approx 1/N$ and $b_j(k) \approx 1/M$, subject to the row stochastic conditions. In Section 2.5.3, it is stated that it is a bad idea to initialize the values to exactly uniform, since the HMM would be stuck at a local maximum and hence it could not climb to an improved solution. Suppose that $\pi_i = 1/N$ and $a_{ij} = 1/N$ and $b_j(k) = 1/M$. Verify that the re-estimation process leaves all of these values unchanged.

9. In this problem, we consider generalizations of the HMM formulation discussed in this chapter.

   a) Consider an HMM where the state transition matrix is time dependent. Then for each t, there is an $N \times N$ row-stochastic $A_t = \{a_{ij}^t\}$ that is used in place of A in the HMM computations. For such an HMM, provide pseudo-code to solve HMM Problem 1.

   b) Consider an HMM of order two, that is, an HMM where the underlying Markov process is of order two. Then the state at time t depends on the states at time $t-1$ and $t-2$. For such an HMM, provide pseudo-code to solve HMM Problem 1.

10. Write an HMM program for the English text problem in Section 9.2 of Chapter 9. Test your program on each of the following cases.

    a) There are $N = 2$ hidden states. Explain your results.

    b) There are $N = 3$ hidden states. Explain your results.

    c) There are $N = 4$ hidden states. Explain your results.

    d) There are $N = 26$ hidden states. Explain your results.

11. In this problem, you will use an HMM to break a simple substitution ciphertext message. For each HMM, train using 200 iterations of the Baum-Welch re-estimation algorithm.

    a) Obtain an English plaintext message of 50,000 plaintext characters, where the characters consist only of lower case `a` through `z` (i.e., remove all punctuation, special characters, and spaces, and convert all upper case to lower case). Encrypt this plaintext using a randomly generated shift of the alphabet. Remember the key.

    b) Train an HMM with $N = 2$ and $M = 26$ on your ciphetext from part a). From the final B matrix, determine the ciphertext letters that correspond to consonants and vowels.

    c) Generate a digraph frequency matrix A for English text, where a_{ij} is the count of the number of times that letter i is followed by letter j. Here, we assume that `a` is letter 0, `b` is letter 1, `c` is letter 2, and so on. This matrix must be based on 1,000,000 characters where, as above, only the 26 letters of the alphabet are used. Next, add five to each element in your 26×26 matrix A. Finally, normalize your matrix A by dividing each element by its row sum. The resulting matrix A will be row stochastic, and it will not contain any 0 probabilities.

    d) Train an HMM with $N = M = 26$, using the first 1000 characters of ciphertext you generated in part a), where the A matrix is initialized with your A matrix from part c). Also, in your HMM, do not re-estimate A. Use the final B matrix to determine a putative key and give the fraction of putative key elements that match the actual key (as a decimal, to four places). For example, if 22 of the 26 key positions are correct, then your answer would be $22/26 = 0.8462$.

12. Write an HMM program to solve the problem discussed in Section 9.2, replacing English text with the following.

    a) French text.

    b) Russian text.

    c) Chinese text.

13. Perform an HMM analysis similar to that discussed in Section 9.2, replacing English with "Hamptonese," the mysterious writing system developed by James Hampton. For information on Hamptonese, see

    http://www.cs.sjsu.edu/faculty/stamp/Hampton/hampton.html

14. Since HMM training is a hill climb, we are only assured of reaching a local maximum. And, as with any hill climb, the specific local maximum that we find will depend on our choice of initial values. Therefore, by training a hidden Markov model multiple times with different initial values, we would expect to obtain better results than when training only once.

    In the paper [16], the authors use an expectation maximization (EM) approach with multiple random restarts as a means of attacking homophonic substitution ciphers. An analogous HMM-based technique is analyzed in the report [158], where the effectiveness of multiple random restarts on simple substitution cryptanalysis is explored in detail. Multiple random restarts are especially helpful in the most challenging cases, that is, when little data (i.e., ciphertext) is available. However, the tradeoff is that the work factor can be high, since the number of restarts required may be very large (millions of random restarts are required in some cases).

    a) Obtain an English plaintext message consisting of 1000 plaintext characters, consisting only of lower case `a` through `z` (i.e., remove all punctuation, special characters, and spaces, and convert all upper case letters to lower case). Encrypt this plaintext using a randomly selected shift of the alphabet. Remember the key. Also generate a digraph frequency matrix A, as discussed in part c) of Problem 11.

    b) Train n HMMs, for each of $n = 1$, $n = 10$, $n = 100$, and $n = 1000$, following the same process as in Problem 11, part d), but using the $T = 1000$ observations generated in part a) of this problem. For a given n select the best result based on the model scores and give the fraction of the putative key that is correct, calculated as in Problem 11, part d).

    c) Repeat part b), but only use the first $T = 400$ observations.

    d) Repeat part c), but only use the first $T = 300$ observations.

15. The Zodiac Killer murdered at least five people in the San Francisco Bay Area in the late 1960s and early 1970s. Although police had a prime suspect, no arrest was ever made and the murders remain officially unsolved. The killer sent several messages to the police and to local newspapers, taunting police for their failure to catch him. One of these

messages contained a homophonic substitution consisting of 408 strange symbols.[7] Not surprisingly, this cipher is known as the Zodiac 408. Within days of its release, the Zodiac 408 was broken by Donald and Bettye Harden, who were schoolteachers from Salinas, California. The Zodiac 408 ciphertext is given below on the left, while the corresponding plaintext appears on the right.

```
I L I K E K I L L I N G P E O P L
E B E C A U S E I T I S S O M U C
H F U N I T I S M O R E F U N T H
A N K I L L I N G W I L D G A M E
I N T H E F O R R E S T B E C A U
S E M A N I S T H E M O S T D A N
G E R O U E A N A M A L O F A L L
T O K I L L S O M E T H I N G G I
V E S M E T H E M O S T T H R I L
L I N G E X P E R E N C E I T I S
E V E N B E T T E R T H A N G E T
T I N G Y O U R R O C K S O F F W
I T H A G I R L T H E B E S T P A
R T O F I T I S T H A E W H E N I
D I E I W I L L B E R E B O R N I
N P A R A D I C E A N D A L L T H
E I H A V E K I L L E D W I L L B
E C O M E M Y S L A V E S I W I L
L N O T G I V E Y O U M Y N A M E
B E C A U S E Y O U W I L L T R Y
T O S L O I D O W N O R A T O P M
Y C O L L E C T I O G O F S L A V
E S F O R M Y A F T E R L I F E E
B E O R I E T E M E T H H P I T I
```

Note the (apparently intentional) misspellings in the plaintext, including "`FORREST`", "`ANAMAL`", and so on. Also, the final 18 characters (underlined in the plaintext above) appear to be random filler.

a) Solve the Zodiac 408 cipher using the HMM approach discussed in Section 9.4. Initialize the A matrix as in part c) of Problem 11, and do not re-estimate A. Use 1000 random restarts of the HMM, and 200 iterations of Baum-Welch re-estimation in each case. Give your answer as the percentage of characters of the actual plaintext that are recovered correctly.

b) Repeat part a), but use 10,000 random restarts.

c) Repeat part b), but use 100,000 random restarts.

d) Repeat part c), but use 1,000,000 random restarts.

[7] The Zodiac 408 ciphertext was actually sent in three parts to local newspapers. Here, we give the complete message, where the three parts have been combined into one. Also, a homophonic substitution is like a simple substitution, except that the mapping is many-to-one, that is, multiple ciphertext symbols can map to one plaintext symbol.

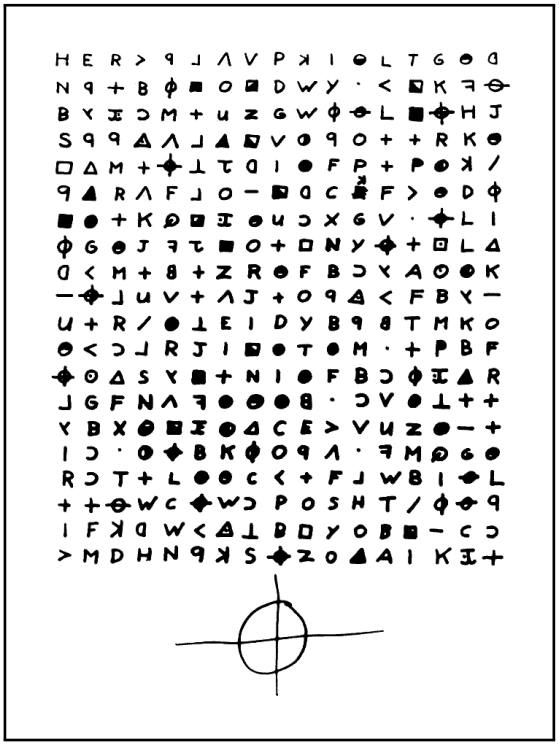
e) Repeat part a), except also re-estimate the A matrix.

f) Repeat part b), except also re-estimate the A matrix.

g) Repeat part c), except also re-estimate the A matrix.

h) Repeat part d), except also re-estimate the A matrix.

16. In addition to the Zodiac 408 cipher, the Zodiac Killer (see Problem 15) released a similar-looking cipher with 340 symbols. This cipher is known as the Zodiac 340 and remains unsolved to this day.[8] The ciphertext is given below.



a) Repeat Problem 15, parts a) through d), using the Zodiac 340 in place of the Zodiac 408. Since the plaintext is unknown, in each case, simply print the decryption obtained from your highest scoring model.

b) Repeat part a) of this problem, except use parts e) through h) of Problem 15.

[8]It is possible that the Zodiac 340 is not a cipher at all, but instead just a random collection of symbols designed to frustrate would-be cryptanalysts. If that's the case, your easily frustrated author can confirm that the "cipher" has been wildly successful.