

## 1                    **S1 Appendix: Technical details of Baum-Welch**

2                    The Baum-Welch algorithm [17] is a form of Expectation Maximization [21],  
3                    which will iterate over solutions to reach a maximum likelihood estimate. While only  
4                    convergence to a local maximum likelihood can be guaranteed, it has been found that for  
5                    most applications Expectation Maximization is nevertheless highly effective. Like all  
6                    Expectation Maximization techniques, in Baum-Welch, we assume a parameterization to  
7                    start, which we use to compute probabilities of intermediate hidden values, which we then  
8                    use to produce a new and better estimate of the parameters. For Baum-Welch, this means  
9                    iteratively improving the transition and emission probabilities.

10                  The Baum-Welch algorithm is built partly upon the forward-backward algorithm.  
11                  For any HMM we have a transition matrix and an emission matrix which define the model.  
12                  Let  $X_{1:T}$  represent the random variables which take one of  $N$  possible values over  $T$  time  
13                  steps. Let  $Y_{1:T}$  represent the random variables representing the emission distribution. Let  $s$   
14                  denote the step in our Baum-Welch iteration process, and  $\theta^{(s)}$  represent our choice of  
15                  parameters for step  $s$ , which includes our transition probabilities, emission distributions,  
16                  and our initial conditions. Our transition matrix is given as:

$$17 \qquad T_{ji}^{(s)} = p(X_t = j | X_{t-1} = i, \theta^{(s)}) \qquad (4)$$

18                  In most introductory descriptions of Baum-Welch, the emission values are from a  
19                  discrete set, and an emission matrix is given as the probability of seeing a particular indexed  
20                  output given a particular state. In this application, our emissions are floating point values,  
21                  and we find it more clarifying to instead have a different emission matrix for every time  
22                  step, and to think of it as a diagonal matrix where the values represent the probability

23 density of the known observation at that time step. Letting  $t$  denote the time step, and  $y_t$   
 24 denote the true observed value at time  $t$ . We define our emission matrix  $O^{(s,t)}$  as follows:

$$25 \quad O_{ii}^{(s,t)} = p(Y_t = y_t | X_t = i, \theta^{(s)}) \quad (5)$$

26 We need to run the forward algorithm. We define our cumulative forward  
 27 probabilities as:

$$28 \quad f_i^{(s,t)} = p(Y_{1:t} = y_{1:t}, X_t = i | \theta^{(s)}) \quad (6)$$

29 We can use dynamic programming to determine  $f_i^{(s,t)}$  for every value of  $i$  and  $t$ , if  
 30 given initial conditions  $f_i^{(s,0)}$  for all  $i$ . The equation is:

$$31 \quad f^{(s,t)} = O^{(s,t)} T^{(s)} f^{(s,t-1)} \quad (7)$$

32 Or equivalently:

$$33 \quad p(Y_{1:t} = y_{1:t}, X_t = i | \theta^{(s)}) \\
 34 \quad = p(Y_t = y_t | X_t = i, \theta^{(s)}) \sum_j p(X_t = i | X_{t-1} = j, \theta^{(s)}) p(Y_{1:t-1} = y_{1:t-1}, X_{t-1} = j | \theta^{(s)}) \\
 35 \quad (8)$$

36 We also need to run the backward algorithm. We define the probabilities as:

$$37 \quad b_i^{(s,t)} = p(Y_{t+1:T} = y_{t+1:T} | X_t = i, \theta^{(s)}) \quad (9)$$

38 We can use dynamic programming to determine  $b_i^{(s,t)}$  for every value of  $i$  and  $t$ .

39 We initialize it with:

$$40 \quad b_i^{(s,T)} = 1 \quad (10)$$

41 And then compute:

$$42 \quad b^{(s,t)} = T^{(s)\top} O^{(s,t+1)} b^{(s,t+1)} \quad (11)$$

43 Or equivalently:

$$44 \quad p(Y_{t+1:T} = y_{t+1:T} | X_t = j, \theta^{(s)}) \\
 45 \quad = \sum_i p(X_t = i | X_{t-1} = j, \theta^{(s)}) p(Y_{t+1} = y_{t+1} | X_{t+1} = i, \theta^{(s)}) p(Y_{t+2:T} = y_{t+2:T} | X_{t+1} = i, \theta^{(s)})$$

46 (12)

47 This gives us the capability to compute two sets of key intermediate values.

48 The first is the same as what is computed in the forward-backward algorithm and  
 49 represents the probability of being in a state at a given time given the entire sequence of  
 50 observations. It's defined as:

51 
$$\gamma_i^{(s,t)} = p(X_t = i | Y_{1:T} = y_{1:T}, \theta^{(s)})$$
 (13)

52 And computed with the formula:

53 
$$\gamma_i^{(s,t)} = \frac{f_i^{(s,t)} b_i^{(s,t)}}{f^{(s,t)\top} b^{(s,t)}}$$
 (14)

54 Or equivalently:

55 
$$p(X_t = i | Y_{1:T} = y_{1:T}, \theta^{(s)})$$
  
 56 
$$= \frac{p(Y_{1:t} = y_{1:t}, X_t = i | \theta^{(s)}) p(Y_{t+1:T} = y_{t+1:T} | X_t = i, \theta^{(s)})}{\sum_j p(Y_{1:t} = y_{1:t}, X_t = j | \theta^{(s)}) p(Y_{t+1:T} = y_{t+1:T} | X_t = j, \theta^{(s)})}$$
 (15)

57 The next set of values represents the probability of a particular transition taking  
 58 place between time step  $t$  and time step  $t + 1$ . We define it as:

59 
$$\xi_{ij}^{(s,t)} = p(X_t = i, X_{t+1} = j | Y_{1:T} = y_{1:T}, \theta^{(s)})$$
 (16)

60 We can compute this with:

61 
$$\xi_{ij}^{(s,t)} = \frac{f_i^{(s,t)} T_{ij}^{(s)} O_{jj}^{(s,t+1)} b_j^{(s,t+1)}}{f^{(s,t)\top} T^{(s)} O^{(s,t+1)} b^{(s,t+1)}}$$
 (17)

62 Or equivalently:

63 
$$p(X_t = i, X_{t+1} = j | Y_{1:T} = y_{1:T}, \theta^{(s)})$$
  
 64 
$$= \frac{p(Y_{1:t} = y_{1:t}, X_t = i | \theta^{(s)}) p(X_{t+1} = j | X_t = i, \theta^{(s)}) p(Y_{t+1} = y_{t+1} | X_{t+1} = i, \theta^{(s)}) p(Y_{t+2:T} = y_{t+2:T} | X_t = i, X_{t+1} = j, \theta^{(s)})}{\sum_{\hat{i}, \hat{j}} p(Y_{1:t} = y_{1:t}, X_t = \hat{i} | \theta^{(s)}) p(X_{t+1} = \hat{j} | X_t = \hat{i}, \theta^{(s)}) p(Y_{t+1} = y_{t+1} | X_{t+1} = \hat{i}, \theta^{(s)}) p(Y_{t+2:T} = y_{t+2:T} | X_t = \hat{i}, X_{t+1} = \hat{j}, \theta^{(s)})}$$
  
 65 (18)

66 We then update our transition probabilities according to:

67 
$$T_{ij}^{(s+1)} = \frac{\sum_t \xi_{ij}^{(s,t)}}{\sum_t \gamma_i^{(s,t)}} \quad (19)$$

68 Or the equivalent expression:

69 
$$p(X_t = j | X_{t-1} = i, \theta^{(s+1)}) = \frac{\sum_t p(X_t = i, X_{t+1} = j | Y_{1:T} = y_{1:T}, \theta^{(s)})}{\sum_t p(X_t = i | Y_{1:T} = y_{1:T}, \theta^{(s)})} \quad (20)$$

70 We also need to determine our new initial conditions. We solve for these as:

71 
$$f_i^{(s+1,0)} = \gamma_i^{(s,0)} \quad (21)$$

72 To update our emission probabilities, we perform a weighted maximum likelihood  
 73 estimate, using it to determine the appropriate weights. The form of the maximum  
 74 likelihood estimate depends on the assumed shape of the output distribution, but if we  
 75 assume each output distribution is normally distributed, where  $Y_t \sim N(\mu_i^{(s)}, \sigma_i^{(s)})$

76 when  $X_t = i$ , we can compute  $\mu^{(s+1)}$  and  $\sigma^{(s+1)}$  as:

77 
$$\mu_i^{(s+1)} = \frac{\sum_t \gamma_i^{(s,t)} Y_t}{\sum_t \gamma_i^{(s,t)}} \quad (22)$$

78 
$$\sigma_i^{(s+1)} = \sqrt{\frac{\sum_t \gamma_i^{(s,t)} (Y_t - \mu_i^{(s+1)})^2}{\sum_t \gamma_i^{(s,t)}}} \quad (23)$$

79 Or the equivalent expressions:

80 
$$\mu_i^{(s+1)} = \frac{\sum_t p(X_t = i | Y_{1:T} = y_{1:T}, \theta^{(s)}) Y_t}{\sum_t p(X_t = i | Y_{1:T} = y_{1:T}, \theta^{(s)})} \quad (24)$$

81 
$$\sigma_i^{(s+1)} = \sqrt{\frac{\sum_t p(X_t = i | Y_{1:T} = y_{1:T}, \theta^{(s)}) (Y_t - \mu_i^{(s+1)})^2}{\sum_t p(X_t = i | Y_{1:T} = y_{1:T}, \theta^{(s)})}} \quad (25)$$

82 We can then compute a new emission matrix using these new parameter estimates:

83 
$$O_{ii}^{(s+1,t)} = N(y_t; \mu_i^{(s+1)}, \sigma_i^{(s+1)}) \quad (26)$$

84 This can be trivially generalized to multiple sequences. Let  $r$  denote the index of a  
 85 sequence, ranging from 1 to  $R$ . Then:

86 
$$T_{ij}^{(s+1)} = \frac{\sum_r \sum_t \xi_{ij}^{(s,r,t)}}{\sum_r \sum_t \gamma_i^{(s,r,t)}} \quad (27)$$

87 
$$f_i^{(s+1,0)} = \sum_r \frac{\gamma_i^{(s,r,0)}}{R} \quad (28)$$

88 
$$\mu_i^{(s+1)} = \frac{\sum_r \sum_t \gamma_i^{(s,r,t)} Y_t}{\sum_r \sum_t \gamma_i^{(s,r,t)}} \quad (29)$$

89 
$$\sigma_i^{(s+1)} = \sqrt{\frac{\sum_r \sum_t \gamma_i^{(s,r,t)} (Y_t - \mu_i^{(s+1)})^2}{\sum_r \sum_t \gamma_i^{(s,r,t)}}} \quad (30)$$

90

