maximize the values of the paths that are used a lot (while still respecting the stochastic constraints). We then repeat this process, hoping to converge on optimal values for the model parameters μ.

The reestimation formulas are as follows:

$$\begin{aligned} \hat{\pi}_i &= \text{expected frequency in state } i \text{ at time } t = 1 \\ &= \gamma_i(1) \end{aligned} \tag{9.17}$$

$$\begin{aligned} \hat{a}_{ij} &= \frac{\text{expected number of transitions from state } i \text{ to } j}{\text{expected number of transitions from state } i} \\ &= \frac{\sum_{t=1}^{T} p_t(i,j)}{\sum_{t=1}^{T} \gamma_i(t)} \end{aligned} \tag{9.18}$$

$$\begin{aligned} \hat{b}_{ijk} &= \frac{\text{expected number of transitions from } i \text{ to } j \text{ with } k \text{ observed}}{\text{expected number of transitions from } i \text{ to } j} \\ &= \frac{\sum_{\{t:o_t=k, 1\le t\le T\}} p_t(i,j)}{\sum_{t=1}^{T} p_t(i,j)} \end{aligned} \tag{9.19}$$

Thus, from $\mu = (A, B, \Pi)$, we derive $\hat{\mu} = (\hat{A}, \hat{B}, \hat{\Pi})$. Further, as proved by Baum, we have that:

$$P(O|\hat{\mu}) \geq P(O|\mu)$$

This is a general property of the EM algorithm (see section 14.2.2). Therefore, iterating through a number of rounds of parameter reestimation will improve our model. Normally one continues reestimating the parameters until results are no longer improving significantly. This process of parameter reestimation does not guarantee that we will find the best model, however, because the reestimation process may get stuck in a *local maximum* (or even possibly just at a saddle point). In most problems of interest, the likelihood function is a complex nonlinear surface and there are many local maxima. Nevertheless, Baum-Welch reestimation is usually effective for HMMs.

LOCAL MAXIMUM

To end this section, let us consider reestimating the parameters of the crazy soft drink machine HMM using the Baum-Welch algorithm. If we let the initial model be the model that we have been using so far, then training on the observation sequence (lem, ice_t, cola) will yield the following values for $p_t(i,j)$:

(9.20)

		Time (and j)								
		1			2			3		
		CP	IP	γ_1	CP	IP	γ_2	CP	IP	γ_3
i	CP	0.3	0.7	1.0	0.28	0.02	0.3	0.616	0.264	0.88
	IP	0.0	0.0	0.0	0.6	0.1	0.7	0.06	0.06	0.12

and so the parameters will be reestimated as follows:

		Original			Reestimated		
Π	CP	1.0			1.0		
	IP	0.0			0.0		
		CP	IP		CP	IP	
A	CP	0.7	0.3		0.5486	0.4514	
	IP	0.5	0.5		0.8049	0.1951	
		cola	ice_t	lem	cola	ice_t	lem
B	CP	0.6	0.1	0.3	0.4037	0.1376	0.4587
	IP	0.1	0.7	0.2	0.1363	0.8537	0.0

Exercise 9.4 [⋆]

If one continued running the Baum-Welch algorithm on this HMM and this training sequence, what value would each parameter reach in the limit? Why?

The reason why the Baum-Welch algorithm is performing so strangely here should be apparent: the training sequence is far too short to accurately represent the behavior of the crazy soft drink machine.

Exercise 9.5 [⋆]

Note that the parameter that is zero in Π stays zero. Is that a chance occurrence? What would be the value of the parameter that becomes zero in B if we did another iteration of Baum-Welch reestimation? What generalization can one make about Baum-Welch reestimation of zero parameters?

9.4 HMMs: Implementation, Properties, and Variants

9.4.1 Implementation

Beyond the theory discussed above, there are a number of practical issues in the implementation of HMMs. Care has to be taken to make the implementation of HMM tagging efficient and accurate. The most obvious issue is that the probabilities we are calculating consist of keeping on multiplying together very small numbers. Such calculations will rapidly