Appendix I. Baum-Welch

(1) Setup

Consider a discrete HMM of length T. The space of observation $X = \{1, 2, \dots N\}$, and the space of underlying states $Z = \{1, 2 \dots M\}$. An HMM $\theta = (\pi, A, B)$ is parametrized by initial state matrix π, the state transition matrix A, and the emission matrix B.

The HMM training process includes learning the parametrization of θ from a dataset of D observations. Let $X = (X^{(1)}, \dots X^{(D)})$, where each $X^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_T^{(i)})$. Each observation is drawn iid.

Baum-Welch repeats the following two steps until convergence:

① Compute $Q(\theta, \theta^s) = \sum_{z \in Z} \log [P(X, z; \theta)] P(z | X; \theta^s)$;

② Set $\theta^{s+1} = \arg\max_\theta Q(\theta, \theta^s)$.

(2) Derivation.

Noting that $P(z, X) = P(X) P(z | X)$, and $P(X)$ is not affected by θ, we have

$$\arg\max_\theta \sum_{z \in Z} \log [P(X, z; \theta)] P(z | X; \theta^s) = \arg\max_\theta \sum_{z \in Z} \log [P(X, z; \theta)] P(z, X; \theta^s) = \arg\max_\theta \hat{Q}(\theta, \theta^s)$$

Now,

$$P(z, X; \theta) = \prod_{d=1}^{D} \left(\pi_{z_1^{(d)}} B_{z_1^{(d)}}(x_1^{(d)}) \prod_{t=2}^{T} A_{z_{t-1}^{(d)} z_t^{(d)}} B_{z_t^{(d)}}(x_t^{(d)}) \right)$$

Taking the log gives,

$$\log P(z, X; \theta) = \sum_{d=1}^{D} \left[\log \pi_{z_1^{(d)}} + \sum_{t=2}^{T} \log A_{z_{t-1}^{(d)} z_t^{(d)}} + \sum_{t=1}^{T} \log B_{z_t^{(d)}}(x_t^{(d)}) \right]$$

Plugging this to $\hat{Q}(\theta, \theta^s)$,

$$\hat{Q}(\theta, \theta^s) = \sum_{z \in Z} \sum_{d=1}^{D} \log \pi_{z_1^{(d)}} P(z, X; \theta^s) + \sum_{z \in Z} \sum_{d=1}^{D} \sum_{t=2}^{T} \log A_{z_{t-1}^{(d)} z_t^{(d)}} P(z, X; \theta^s) + \sum_{z \in Z} \sum_{d=1}^{D} \sum_{t=1}^{T} \log B_{z_t^{(d)}}(x_t^{(d)}) P(z, X; \theta^s)$$

Constrained by the validity of probability distribution π, A_i, $B_i(\cdot)$, we introduce Lagrange multipliers. Let $\hat{L}(\theta, \theta^s)$ be the Lagrangian,

$$\hat{L}(\theta, \theta^s) = \hat{Q}(\theta, \theta^s) - \lambda_\pi \left(\sum_{i=1}^{M} \pi_i - 1 \right) - \sum_{i=1}^{M} \lambda_{A_i} \left(\sum_{j=1}^{M} A_{ij} - 1 \right) - \sum_{i=1}^{M} \lambda_{B_i} \left(\sum_{j=1}^{N} B_i(j) - 1 \right)$$

Take the partial derivation of π_i's,

$$\partial \hat{L}(\theta, \theta^s) / \partial \pi_i = \frac{\partial}{\partial \pi_i} \left(\sum_{z \in Z} \sum_{d=1}^{D} \log_{z_1^{(d)}} P(z, X; \theta^s) \right) - \lambda_\pi = 0$$

$$\partial \hat{L}(\theta, \theta^s) / \partial \lambda_\pi = - \left(\sum_{i=1}^{M} \pi_i - 1 \right) = 0$$

Then it yields:

$$\pi_i = \frac{1}{D} \sum_{d=1}^{D} P(z_1^{(d)} = i | X^{(d)}; \theta^s) \quad (1)$$

Similarly for A_{ij}'s,

$$\partial \hat{L}(\theta,\theta^s)/\partial A_{ij} = \frac{\partial}{\partial A_{ij}}\left(\sum_{z\in Z}\sum_{d=1}^{D}\sum_{t=2}^{T}\log A_{z_{t-1}^{(d)}\, z_t^{(d)}} P(z,x;\theta^s)\right) - \lambda_{A_i} = 0$$

$$\partial \hat{L}(\theta,\theta^s)/\partial \lambda_{A_i} = -\left(\sum_{j=1}^{N} A_{ij} - 1\right) = 0$$

It yields $A_{ij} = \dfrac{\sum_{d=1}^{D}\sum_{t=2}^{T} P(z_{t-1}^{(d)} = i,\, z_t^{(d)} = j \mid x;\theta^s)}{\sum_{d=1}^{D}\sum_{t=2}^{T} P(z_{t-1}^{(d)} = i \mid X^{(d)};\theta^s)}$ (2)

The last thing is $B_i(j)$. Let $I(x)$ denote an indicator function, which is 1 if x is true, 0 otherwise.

$$\partial \hat{L}(\theta,\theta^s)/\partial B_i(j) = \frac{\partial}{\partial B_i(j)}\left(\sum_{z\in Z}\sum_{d=1}^{D}\sum_{t=1}^{T}\log B_{z_t^{(d)}}(x_t^{(d)}) P(z,x;\theta^s)\right) - \lambda_{B_i} = 0$$

$$\partial \hat{L}(\theta,\theta^s)/\partial \lambda_{B_i} = -\left(\sum_{j=1}^{N} B_i(j) - 1\right) = 0$$

It leads to,

$$B_i(j) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P(z_t^{(d)} = i \mid X^{(d)};\theta^s)\, I(x_t^{(d)} = j)}{\sum_{d=1}^{D}\sum_{t=1}^{T} P(z_t^{(d)} = i \mid X^{(d)};\theta^s)} \quad (3)$$

(1) ~ (3) give us an updated parameters of HMM.