The immutability during the Baum–Welch updating

To prove that, if there are some zeroes in the transition matrix or in the emission matrix, during the learning updating, the zeros will never be updated. In other words, they will always be zeroes.

Here are the demonstrations.

For the a_{ij}

We suppose that $a_{xy} = 0$

So we got that $\tilde{\xi}_t(x,y) = \alpha_t(x) \cdot \underbrace{a_{xy}}_{=0} \cdot \beta_{t+1}(y) \cdot b_y(\mathcal{O}_{t+1}) = 0$

And $\xi_t(x,y) = \tilde{\xi}_t(x,y) \Big/ \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \tilde{\xi}_t(i,j) = 0$

Then in every updating, $a_{xy} = \sum_{t=0}^{T-1} \xi_t(x,y) \Big/ \sum_{t=0}^{T-1} \gamma_t(x) = 0$

For the $b_i(\mathcal{O}_t)$

We suppose that, $b_x(k) = 0$

That is to say, for all $\mathcal{O}_t = k$, we have $b_x(\mathcal{O}_t) = 0$

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

So if $\mathcal{O}_t = k$, we have $\alpha_t(x) = \left[\sum_{j=0}^{N-1} \alpha_{t-1}(j) \cdot a_{jx} \right] \cdot \underbrace{b_x(\mathcal{O}_t)}_{0} = 0.$

In other words, $\sum_{t=1,\cdots,T-1 | \mathcal{O}_t = k} \alpha_t(x) = \sum 0 = 0$

Due to $\gamma_t(i) = \alpha_t(i) \cdot \beta_t(i)$, we have $\sum_{t=1,\cdots,T-1 | \mathcal{O}_t = k} \gamma_t(x) = \sum_{t=1,\cdots,T-1 | \mathcal{O}_t = k} \alpha_t(x) \cdot \beta_t(x) = 0$

Then, in every updating, $b_x(k) = \sum_{t=1,\cdots,T-1 | \mathcal{O}_t = k} \gamma_t(x) \Big/ \sum_t \gamma_t(x) = 0$

The EM (Expectation–maximization) algorithm

What we want, is to get the Maximum Likelihood Estimation of the parameters of a statistical model by $\theta_{MLE} = \arg\max_{\theta} \mathscr{L}(\theta \mid x)$ where $\mathscr{L}(\theta \mid x) = P(X = x \mid \theta)$.

In a normal statistical model, we usually use the log-likelihood $\ell(\theta \mid x) = \ln \mathscr{L}(\theta \mid x)$, then solve the equation $\frac{\partial \ell(\theta \mid x)}{\partial \theta} = 0$ to get the MLE of the parameters θ.

But, for the Mixture Model, it's too hard to compte $\frac{\partial \ell(\theta \mid x)}{\partial \theta}$, and much more difficult to get the solution of the "partial derivative = 0" equation.

So the idea of EM algorithm is to calculate the θ_{MLE} by recursion.
We add a latent variable Z, where $\ell(\theta; x, z) = \ln P(x, z \mid \theta)$.
And the marginal probability of observed data is $P(x \mid \theta) = \int_z P(x, z \mid \theta) dz$.
The recursion runs by :

$$\theta^{(t+1)} = \arg\max_{\theta} \int_z \ln P(x, z \mid \theta) \cdot P(z \mid x, \theta^{(t)}) dz = \arg\max_{\theta} E_{Z \sim P(z \mid x, \theta^{(t)})}[\ell(\theta; x, z)]$$

Or we can say, $\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)})$ where $Q(\theta \mid \theta^{(t)}) = \int_z \ell(\theta; x, z) \cdot P(z \mid x, \theta^{(t)}) dz$

The convergence:

To prove the algorithm is convergent, we should prove that $P(x \mid \theta^{(t+1)}) \geq P(x \mid \theta^{(t)})$
We know that $\ln P(x \mid \theta) = \ln \frac{P(x, z \mid \theta)}{P(z \mid x, \theta)} = \ln P(x, z \mid \theta) - \ln P(z \mid x, \theta)$.
Then we compute the expectation of $Z \sim P(z \mid x, \theta^{(t)})$
For the left side:
$E[\text{Left}] = E_Z[\ln P(x \mid \theta)] = \ln P(x \mid \theta)$ Because X is independent of Z.
For the right side:

$$E[\text{Right}] = E_Z[\ln P(x, z \mid \theta) - \ln P(z \mid x, \theta] = \int_z P(z \mid x, \theta^{(t)}) \cdot \ln P(x, z \mid \theta) dz - \int_z P(z \mid x, \theta^{(t)}) \cdot \ln P(z \mid x, \theta) dz$$

If we define $H(\theta \mid \theta^{(t)}) = \int_z P(z \mid x, \theta^{(t)}) \cdot \ln P(z \mid x, \theta) dz$

We will get: $\ln P(x \mid \theta) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)})$
It's evident that $Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)})$, because $\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)})$.
As for $H(\theta \mid \theta^{(t)})$, we calculate like:

$$\begin{aligned} H(\theta^{(t)} \mid \theta^{(t)}) - H(\theta^{(t+1)} \mid \theta^{(t)}) &= \int_z P(z \mid x, \theta^{(t)}) \cdot \ln P(z \mid x, \theta^{(t)}) dz - \int_z P(z \mid x, \theta^{(t)}) \cdot \ln P(z \mid x, \theta^{(t+1)}) dz \\ &= \int_z P(z \mid x, \theta^{(t)}) \cdot \ln \frac{P(z \mid x, \theta^{(t)})}{P(z \mid x, \theta^{(t+1)})} dx \\ &= E_{Z \sim P(z \mid \theta^{(t)})}[\ln \frac{P(z \mid x, \theta^{(t)})}{P(z \mid x, \theta^{(t+1)})}] \\ &= E_{Z \sim P(z \mid \theta^{(t)})}[-\ln \frac{P(z \mid x, \theta^{(t+1)})}{P(z \mid x, \theta^{(t)})}] \end{aligned}$$

Because the function $f(x) = -\ln x$ is a convex function, we have the theorem that $E[f(X)] \geq f(E[X])$

Then we have:

$$\begin{aligned} H(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) &= E_{Z\sim P(z|\theta^{(t)})}[-\ln \frac{P(z|x,\theta^{(t+1)})}{P(z|x,\theta^{(t)})}] \\ &\geq -\ln E_{Z\sim P(z|\theta^{(t)})}[\frac{P(z|x,\theta^{(t+1)})}{P(z|x,\theta^{(t)})}] \\ &= -\ln \int_z P(z|x,\theta^{(t)}) \cdot \frac{P(z|x,\theta^{(t+1)})}{P(z|x,\theta^{(t)})} dx \\ &= -\ln \int_z P(z|x,\theta^{(t+1)}) dx \\ &= -\ln 1 \\ &= 0 \end{aligned}$$

That's to say, $H(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) \geq 0 \Rightarrow -H(\theta^{(t+1)}|\theta^{(t)}) \geq -H(\theta^{(t)}|\theta^{(t)})$

If we combine $\begin{cases} Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) \\ -H(\theta^{(t+1)}|\theta^{(t)}) \geq -H(\theta^{(t)}|\theta^{(t)}) \end{cases}$, we have

$$\begin{gathered} Q(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \\ \Downarrow \\ \ln P(x|\theta^{(t+1)}) \geq \ln P(x|\theta^{(t)}) \\ \Downarrow \\ P(x|\theta^{(t+1)}) \geq P(x|\theta^{(t)}) \end{gathered}$$

In conclusion, EM algorithm is convergent.

The application of EM in HMM

In HMM, we have the parameters $\lambda = \{\pi, A, B\}$, the latent variable X and the observed variable $\mathcal{O}$.

Then $\mathscr{L}(\lambda; \mathcal{O}, X) = \pi_{x_0} \cdot \prod_{t=1}^{T-1} a_{x_{t-1},x_t} \cdot \prod_{t=0}^{T-1} b_{x_t}(\mathcal{O}_t)$

And $\ell(\lambda; \mathcal{O}, X) = \ln \mathscr{L}(\lambda; \mathcal{O}, X) = \ln \pi_{x_0} + \sum_{t=1}^{T-1} \ln a_{x_{t-1},x_t} + \sum_{t=0}^{T-1} \ln b_{x_t}(\mathcal{O}_t)$

In order to avoid confusion the index of HMM time and the generation of parameters, we use g for the recursion of parameters.

The function Q will be $Q(\lambda|\lambda^{(g)}) = \sum_I \ell(\lambda; \mathcal{O}, X) \cdot P(X|\mathcal{O}, \lambda^{(g)})$

If we pay attention to $P(X|\mathcal{O}, \lambda^{(g)})$, we can find that $P(X|\mathcal{O}, \lambda^{(g)}) = \frac{P(X, \mathcal{O}|\lambda^{(g)})}{P(\mathcal{O}|\lambda^{(g)})}$

For the function Q, $\lambda^{(g)}$ is a constant, so the $P(\mathcal{O}|\lambda^{(g)})$ is also a constant, that's to say

$$\begin{aligned} \lambda^{(g+1)} = \arg\max_\lambda Q(\lambda|\lambda^{(g)}) &= \arg\max_\lambda \sum_I \ell(\lambda; \mathcal{O}, X) \cdot P(X|\mathcal{O}, \lambda^{(g)}) \\ &= \arg\max_\lambda \sum_I \ell(\lambda; \mathcal{O}, X) \cdot P(X, \mathcal{O}|\lambda^{(g)}) \end{aligned}$$

Then we unfold the I and ℓ, we have:

$$\lambda^{(g+1)} = \arg\max_\lambda \sum_{i_0=0}^{N-1} \sum_{i_1=0}^{N-1} \cdots \sum_{i_{T-1}=0}^{N-1} \left(\ln \pi_{i_0} + \sum_{t=1}^{T-1} \ln a_{i_{t-1},i_t} + \sum_{t=0}^{T-1} \ln b_{i_t}(\mathcal{O}_t) \right) \cdot P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O}|\lambda^{(g)})$$

Update π_i

$$\pi^{(g+1)} = \arg\max_{\pi} \sum_{i_0=0}^{N-1}\sum_{i_1=0}^{N-1}\cdots\sum_{i_{T-1}=0}^{N-1}\left(\ln \pi_{i_0} + \sum_{t=1}^{T-1}\ln a_{i_{t-1},i_t} + \sum_{t=0}^{T-1}\ln b_{i_t}(\mathcal{O}_t)\right)\cdot P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O}\,|\,\lambda^{(g)})$$

$$= \arg\max_{\pi} \sum_{i_0=0}^{N-1}\sum_{i_1=0}^{N-1}\cdots\sum_{i_{T-1}=0}^{N-1}\ln \pi_{i_0}\cdot P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O}\,|\,\lambda^{(g)})$$

$$= \arg\max_{\pi} \sum_{i_0=0}^{N-1}\ln \pi_{i_0}\sum_{i_1=0}^{N-1}\cdots\sum_{i_{T-1}=0}^{N-1} P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O}\,|\,\lambda^{(g)})$$

$$= \arg\max_{\pi} \sum_{i_0=0}^{N-1}\ln \pi_{i_0}\cdot P(X_0 = q_{i_0}, \mathcal{O}\,|\,\lambda^{(g)})$$

More succinctly, we use $\pi^{(g+1)} = \arg\max_{\pi} \sum_{i=0}^{N-1}\ln \pi_i \cdot P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)})$.

The rest of the work, is to maximise this, under the condition $\sum \pi_i = 1$

To avoid confusion of symbols, we use $\mathbf{LM}()$ for the function $\mathscr{L}()$, η for the extra variable λ.

Let the function $\mathbf{LM}_{\pi}(\pi, \eta) = \sum_{i=0}^{N-1}\ln \pi_i \cdot P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)}) - \eta(\sum_{i=0}^{N-1}\pi_i - 1)$

Then we calculate $\dfrac{\partial \mathbf{LM}_{\pi}(\pi,\eta)}{\partial \pi_i} = \dfrac{1}{\pi_i}\cdot P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)}) - \eta.$

Let the partial derivative equal to 0, we have

$$\frac{1}{\pi_i}\cdot P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)}) - \eta = 0$$

$$\Rightarrow P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)}) - \pi_i\cdot\eta = 0$$

$$\Rightarrow \sum_{i=0}^{N-1}\left(P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)}) - \pi_i\cdot\eta\right) = 0$$

$$\Rightarrow P(\mathcal{O}\,|\,\lambda^{(g)}) - \eta = 0$$

$$\Rightarrow \eta = P(\mathcal{O}\,|\,\lambda^{(g)})$$

Then we replace η by $P(\mathcal{O}\,|\,\lambda^{(g)})$, we will get

$$\pi_i^{(g+1)} = \frac{P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)})}{\eta} = \frac{P(X_0 = q_i, \mathcal{O}\,|\,\lambda^{(g)})}{P(\mathcal{O}\,|\,\lambda^{(g)})} = P(X_0 = q_i\,|\,\mathcal{O}, \lambda^{(g)})$$

Know that $\gamma_t(i) = P(X_t = q_i\,|\,\mathcal{O}, \lambda)$

So each π_i will be updated by $\gamma_0(i)$.

Update a_{ij}

The same as π_i, we get:

$$a^{(g+1)} = \arg\max_{a} \sum_{i_0=0}^{N-1}\sum_{i_1=0}^{N-1}\cdots\sum_{i_{T-1}=0}^{N-1}\sum_{t=1}^{T-1}\ln a_{i_{t-1},i_t}\cdot P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O}\,|\,\lambda^{(g)})$$

$$= \arg\max_{a} \sum_{i_{t-1}=0}^{N-1}\sum_{i_t=0}^{N-1}\sum_{t=1}^{T-1}\ln a_{i_{t-1},i_t}\sum_{i_1=0}^{N-1}\cdots\sum_{i_{T-1}=0}^{N-1} P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O}\,|\,\lambda^{(g)})$$

$$= \arg\max_{a} \sum_{i_{t-1}=0}^{N-1}\sum_{i_t=0}^{N-1}\sum_{t=1}^{T-1}\ln a_{i_{t-1},i_t}\cdot P(X_{t-1} = q_{i_{t-1}}, X_t = q_{i_t}, \mathcal{O}\,|\,\lambda^{(g)})$$

$$= \arg\max_{a} \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{t=1}^{T-1}\ln a_{ij}\cdot P(X_{t-1} = q_i, X_t = q_j, \mathcal{O}\,|\,\lambda^{(g)})$$

The constraint condition is: $\forall i \in [0, N-1], \sum_{j=0}^{N-1} a_{ij} = 1$

So we create the function like:

$$\mathbf{LM}_a(a, \eta) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \sum_{t=1}^{T-1} \ln a_{ij} \cdot P(X_{t-1} = q_i, X_t = q_j, \mathcal{O} \mid \lambda^{(g)}) - \sum_{i=0}^{N-1} \eta_i (\sum_{j=0}^{N-1} a_{ij} - 1)$$

So we have $\frac{\partial \mathbf{LM}_a(a, \eta)}{\partial a_{ij}} = \sum_{t=1}^{T-1} \frac{1}{a_{ij}} \cdot P(X_{t-1} = q_i, X_t = q_j, \mathcal{O} \mid \lambda^{(g)}) - \eta_i$

Let the partial derivative equal to 0, we have:

$$\sum_{t=1}^{T-1} \frac{1}{a_{ij}} \cdot P(X_{t-1} = q_i, X_t = q_j, \mathcal{O} \mid \lambda^{(g)}) - \eta_i = 0$$

$$\Rightarrow \sum_{t=1}^{T-1} P(X_{t-1} = q_i, X_t = q_j, \mathcal{O} \mid \lambda^{(g)}) - a_{ij} \cdot \eta_i = 0$$

$$\Rightarrow \sum_{j=0}^{N-1} \left(\sum_{t=1}^{T-1} P(X_{t-1} = q_i, X_t = q_j, \mathcal{O} \mid \lambda^{(g)}) - a_{ij} \cdot \eta_i \right) = 0$$

$$\Rightarrow \sum_{t=1}^{T-1} P(X_{t-1} = q_i, \mathcal{O} \mid \lambda^{(g)}) - 1 \cdot \eta_i = 0$$

$$\Rightarrow \eta_i = \sum_{t=1}^{T-1} P(X_{t-1} = q_i, \mathcal{O} \mid \lambda^{(g)})$$

Finally, we got

$$\begin{aligned} a_{ij}^{(g+1)} &= \sum_{t=1}^{T-1} P(X_{t-1} = q_i, X_t = q_j, \mathcal{O} \mid \lambda^{(g)}) \Big/ \eta_i \\ &= \sum_{t=1}^{T-1} P(X_{t-1} = q_i, X_t = q_j, \mathcal{O} \mid \lambda^{(g)}) \Big/ \sum_{t=1}^{T-1} P(X_{t-1} = q_i, \mathcal{O} \mid \lambda^{(g)}) \\ &= \sum_{t=0}^{T-2} P(X_t = q_i, X_{t+1} = q_j, \mathcal{O} \mid \lambda^{(g)}) \Big/ \sum_{t=0}^{T-2} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) \end{aligned}$$

Know that $\begin{cases} \gamma_t(i) = P(X_t = q_i \mid \mathcal{O}, \lambda) \\ \xi_t(i, j) = P(X_t = q_i, X_{t+1} = q_j \mid \mathcal{O}, \lambda) \end{cases}$

So each a_{ij} will be updated by $\sum_{t=0}^{T-2} \xi_t(i, j) \Big/ \sum_{t=0}^{T-2} \gamma_t(i)$

Update $b_i(\mathcal{O}t)$ (or $b_i(k)$)

Same:

$$\begin{aligned} b^{(g+1)} &= \arg\max_b \sum_{i_0=0}^{N-1} \sum_{i_1=0}^{N-1} \cdots \sum_{i_{T-1}=0}^{N-1} \sum_{t=0}^{T-1} \ln b_{i_t}(\mathcal{O}_t) \cdot P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O} \mid \lambda^{(g)}) \\ &= \arg\max_b \sum_{i_t=0}^{N-1} \sum_{t=0}^{T-1} \ln b_{i_t}(\mathcal{O}_t) \sum_{i_1=0}^{N-1} \cdots \sum_{i_{T-1}=0}^{N-1} P(X_0 = q_{i_0}, X_1 = q_{i_1}, \cdots, X_{T-1} = q_{i_{T-1}}, \mathcal{O} \mid \lambda^{(g)}) \\ &= \arg\max_b \sum_{i_t=0}^{N-1} \sum_{t=0}^{T-1} \ln b_{i_t}(\mathcal{O}_t) \cdot P(X_t = q_{i_t}, \mathcal{O} \mid \lambda^{(g)}) \\ &= \arg\max_b \sum_{i=0}^{N-1} \sum_{t=0}^{T-1} \ln b_i(\mathcal{O}_t) \cdot P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) \end{aligned}$$

For the discret observation

The constraint condition is: $\forall i \in [0, N-1], \sum_{k=0}^{M-1} b_i(k) = 1$

So we create the function like:

$$\mathbf{LM}_b(b, \eta) = \sum_{i=0}^{N-1} \sum_{t=0}^{T-1} \ln b_i(\mathcal{O}_t) \cdot P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) - \sum_{i=0}^{N-1} \eta_i \left(\sum_{k=0}^{M-1} b_i(k) - 1 \right)$$

Then the partial derivative $\frac{\partial \mathbf{LM}_b(b, \eta)}{\partial b_i(k)} = \sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} \frac{1}{b_i(\mathcal{O}_t)} \cdot P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) - \eta_i$

Let it equals 0, we will get:

$$\sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} \frac{1}{b_i(\mathcal{O}_t)} \cdot P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) - \eta_i = 0$$

$$\Rightarrow \sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) - b_i(\mathcal{O}_t)\Big|_{\mathcal{O}_t=k} \cdot \eta_i = 0$$

$$\Rightarrow \sum_{k=0}^{M-1} \left(\sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) - b_i(k) \cdot \eta_i \right) = 0$$

$$\Rightarrow \sum_{t=0}^{T-1} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) - 1 \cdot \eta_i = 0$$

$$\Rightarrow \eta_i = \sum_{t=0}^{T-1} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)})$$

Finally,

$$b_i(k)^{(g+1)} = \sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) \Big/ \eta_i$$

$$= \sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)}) \Big/ \sum_{t=0}^{T-1} P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)})$$

If we replace $P(X_t = q_i \mid \mathcal{O}, \lambda)$ by $\gamma_t(i)$, we will get that

Each $b_i(k)$ will be updated by $\sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} \gamma_t(i) \Big/ \sum_{t=0}^{T-1} \gamma_t(i)$.

For the continuous observation (gaussian emission)

We got $b_i(\mathcal{O}_t) = \frac{1}{\sigma_i \sqrt{2\pi}} \cdot exp(-\frac{(\mathcal{O}_t - \mu_i)^2}{2{\sigma_i}^2})$ where μ is the mean, and σ is the standard deviation.

Then we do the log, we got: $\ln b_i(\mathcal{O}_t) = -\left(\frac{1}{2} \ln(2\pi) + \ln(\sigma_i) + \frac{(\mathcal{O}_t - \mu_i)^2}{2{\sigma_i}^2} \right)$

We continue the $b^{(g+1)}$, we got:

$$b^{(g+1)} = \arg\max_b \sum_{i=0}^{N-1} \sum_{t=0}^{T-1} \ln b_i(\mathcal{O}_t) \cdot P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)})$$

$$= \arg\max_b - \sum_{i=0}^{N-1} \sum_{t=0}^{T-1} \left(\frac{1}{2} \ln(2\pi) + \ln(\sigma_i) + \frac{(\mathcal{O}_t - \mu_i)^2}{2{\sigma_i}^2} \right) \cdot P(X_t = q_i, \mathcal{O} \mid \lambda^{(g)})$$

We replace $P(X_t = q_i \mid \mathcal{O}, \lambda)$ by $\gamma_t(i)$, and we remove $\frac{1}{2} \ln(2\pi)$ because it's a constant, we got:

$$
\begin{aligned}
b^{(g+1)} &= \arg\max_{b} -\sum_{i=0}^{N-1}\sum_{t=0}^{T-1}\left(\frac{1}{2}\ln(2\pi)+\ln(\sigma_i)+\frac{(\mathcal{O}_t-\mu_i)^2}{2{\sigma_i}^2}\right)\cdot P(X_t=q_i,\mathcal{O}\,|\,\lambda^{(g)}) \\
&= \arg\min_{b} \sum_{i=0}^{N-1}\sum_{t=0}^{T-1}\left(\ln(\sigma_i)\cdot\gamma_t(i)+\frac{(\mathcal{O}_t-\mu_i)^2}{2{\sigma_i}^2}\cdot\gamma_t(i)\right) \\
&= \arg\min_{\sigma} \sum_{i=0}^{N-1}\sum_{t=0}^{T-1}\left(\ln(\sigma_i)\cdot\gamma_t(i)\right)+\arg\min_{\sigma,\mu} \sum_{i=0}^{N-1}\sum_{t=0}^{T-1}\left(\frac{(\mathcal{O}_t-\mu_i)^2}{2{\sigma_i}^2}\cdot\gamma_t(i)\right)
\end{aligned}
$$

For μ_i : $\dfrac{\partial b_i(\mathcal{O}_t)}{\partial\mu_i}=\sum_{t=0}^{T-1}\left(\dfrac{(\mu_i-\mathcal{O}_t)}{{\sigma_i}^2}\cdot\gamma_t(i)\right)$

Let it equal to 0, we have:

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\left(\frac{(\mu_i-\mathcal{O}_t)}{{\sigma_i}^2}\cdot\gamma_t(i)\right)=0 \\
\Rightarrow\ &\sum_{t=0}^{T-1}(\mu_i-\mathcal{O}_t)\cdot\gamma_t(i)=0 \\
\Rightarrow\ &\sum_{t=0}^{T-1}\mu_i\cdot\gamma_t(i)-\sum_{t=0}^{T-1}\mathcal{O}_t\cdot\gamma_t(i)=0 \\
\Rightarrow\ &\mu_i=\sum_{t=0}^{T-1}\mathcal{O}_t\cdot\gamma_t(i)\Big/\sum_{t=0}^{T-1}\gamma_t(i)
\end{aligned}
$$

For σ_i : $\dfrac{\partial b_i(\mathcal{O}_t)}{\partial\sigma_i}=\sum_{t=0}^{T-1}\left(\dfrac{1}{\sigma_i}\cdot\gamma_t(i)\right)+\sum_{t=0}^{T-1}\left(\dfrac{(\mu_i-\mathcal{O}_t)^2}{-{\sigma_i}^3}\cdot\gamma_t(i)\right)$

Let it equal to 0, we have:

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\left(\frac{1}{\sigma_i}\cdot\gamma_t(i)\right)-\sum_{t=0}^{T-1}\left(\frac{(\mu_i-\mathcal{O}_t)^2}{{\sigma_i}^3}\cdot\gamma_t(i)\right)=0 \\
\Rightarrow\ &\sum_{t=0}^{T-1}{\sigma_i}^2\cdot\gamma_t(i)-\sum_{t=0}^{T-1}(\mu_i-\mathcal{O}_t)^2\cdot\gamma_t(i)=0 \\
\Rightarrow\ &{\sigma_i}^2=\sum_{t=0}^{T-1}(\mu_i-\mathcal{O}_t)^2\cdot\gamma_t(i)\Big/\sum_{t=0}^{T-1}\gamma_t(i) \\
\Rightarrow\ &{\sigma_i}^2=\left[\sum_{t=0}^{T-1}{\mathcal{O}_t}^2\cdot\gamma_t(i)\Big/\sum_{t=0}^{T-1}\gamma_t(i)\right]-{\mu_i}^2
\end{aligned}
$$

Conclusion

If we got $\begin{cases} \gamma_t(i) = P(X_t = q_i \,|\, \mathcal{O}, \lambda) \\ \xi_t(i,j) = P(X_t = q_i, X_{t+1} = q_j \,|\, \mathcal{O}, \lambda) \end{cases}$

So the updates (note: $\mathbb{N}_m^n = \{m, m+1, \cdots, n\}$ where $n \geq m$):

$$\forall i \in \mathbb{N}_0^{N-1}, \qquad \pi_i{}^+ \leftarrow \gamma_0^-(i)$$

$$\forall i \in \mathbb{N}_0^{N-1}, \forall j \in \mathbb{N}_0^{N-1}, \qquad a_{ij}^+ \leftarrow \sum_{t=0}^{T-2} \xi_t^-(i,j) \Big/ \sum_{t=0}^{T-2} \gamma_t^-(i)$$

For the discret observation:

$$\forall i \in \mathbb{N}_0^{N-1}, \forall k \in \mathbb{N}_0^{M-1}, \qquad b_i^+(k) \leftarrow \sum_{t=0,\cdots,T-1/\mathcal{O}_t=k} \gamma_t^-(i) \Big/ \sum_{t=0}^{T-1} \gamma_t^-(i)$$

Or for the continuous observation (gaussian emission):

$$\forall i \in \mathbb{N}_0^{N-1} \qquad \mu_i^+ \leftarrow \sum_{t=0}^{T-1} \left[\gamma_t^-(i) \cdot \mathcal{O}_t \right] \Big/ \sum_{t=0}^{T-1} \gamma_t^-(i)$$

$$\begin{aligned} \forall i \in \mathbb{N}_0^{N-1} \qquad (\sigma_i^+)^2 &\leftarrow \sum_{t=0}^{T-1} \left[\gamma_t^-(i) \cdot (\mu_i^+ - \mathcal{O}_t)^2 \right] \Big/ \sum_{t=0}^{T-1} \gamma_t^-(i) \\ &= \left[\sum_{t=0}^{T-1} \left[\gamma_t^-(i) \cdot {\mathcal{O}_t}^2 \right] \Big/ \sum_{t=0}^{T-1} \gamma_t^-(i) \right] - (\mu_i^+)^2 \end{aligned}$$