initial parameters are then computed as

$$\pi_m = N_m / N_{\text{tot}}$$

$$\mu_m = N_m^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T^{(n)}} o_t^{(n)} \, h_m(o_t^{(n)}) \quad (15)$$

t=0?

$$\sigma_m^2 = N_m^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T^{(n)}} (o_t^{(n)} - \mu_m)^2 \, h_m(o_t^{(n)}) \quad (16)$$

t=0?

This approximation is then improved ~~upon~~ by utilizing the expectation-maximization procedure described by Bilmes [22].

$$\pi'_m = N_{\text{tot}}^{-1} \sum_{n=1}^{N} \sum_{t=0}^{T^{(n)}} \chi_m(o_t^{(n)}, \mathbf{E}, \boldsymbol{\pi})$$

$$\mu'_m = (\pi_m N_{\text{tot}})^{-1} \sum_{n=1}^{N} \sum_{t=0}^{T^{(n)}} o_t^{(n)} \, \chi_m(o_t^{(n)}, \mathbf{E}, \boldsymbol{\pi})$$

$$\sigma'^2_m = (\pi_m N_{\text{tot}})^{-1} \sum_{n=1}^{N} \sum_{t=0}^{T^{(n)}} (o_t^{(n)} - \mu'_m)^2 \, \chi_m(o_t^{(n)}, \mathbf{E}, \boldsymbol{\pi}) \quad (17)$$

Is this right? Should these be χ_m instead of π_m?

like notation better!

where the function $\overset{P?}{p}(m|o, \mathbf{E}, \boldsymbol{\pi})$ is given by the fuzzy membership function:

$$\chi_m(o, \mathbf{E}, \boldsymbol{\pi}) = \frac{\pi_m \, \varphi(o|\mathbf{e}_m)}{\sum_{l=1}^{M} \pi_l \, \varphi(o|\mathbf{e}_l)} \quad (18)$$

This iterative procedure is terminated when the change in the parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$ falls below a certain relative threshold, such as $||\boldsymbol{\pi}^{(n)} - \boldsymbol{\pi}^{(n-1)}|| / ||\boldsymbol{\pi}^{(n)}|| < 10^{-4}$.

index (i) so not to confuse w/ data trace?

2. *Transition matrix estimation*

Once initial state observable emission parameters $\mathbf{E}$ are determined, an initial transition matrix is estimated using an iterative likelihood maximization approach that enforces detailed balance [23]. First, a matrix of fractional transition counts $\mathbf{C} \equiv (c_{ij})$ is estimated using the membership function:

$$c_{ij} = \sum_{n=1}^{N} \sum_{t=1}^{T^{(n)}} \chi_i(o_{t-1}^{(n)}, \mathbf{E}, \overset{\text{bold}}{\pi}) \, \chi_j(o_t^{(n)}, \mathbf{E}, \boldsymbol{\pi}) \quad (19)$$

A symmetric $M \times M$ matrix $\mathbf{X} \equiv (x_{ij})$ is initialized by

$$x_{ij} = x_{ji} = c_{ij} + c_{ji} \quad (20)$$

and a vector of row sums

$$x_i = \sum_{j=1}^{M} x_{ij} \quad (21)$$

Then, the iterative procedure described in Algorithm 1 of [23] is applied. For each update iteration, we first update the diagonal elements of $\mathbf{X}$:

$$x'_{ii} = \frac{c_{ii}(x_i - x_{ii})}{c_i - c_{ii}}$$

$$\cancel{x'_i = \sum_{j=1}^{M} x'_{ij}} \quad (22)$$

could we use the notation $x_{i\cdot}$ and $c_{i\cdot}$ to indicate $\sum_j x_{ij}$? You shouldn't need to update them then.

followed by the off-diagonal elements:

$$x'_{ij} = x'_{ji} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

$$\cancel{x'_i = \sum_{j=1}^{M} x'_{ij}} \quad (23)$$

where the quantities a, b, and c are computed from $\mathbf{X}$ and $\mathbf{C}$ as

$$a \equiv c_i - c_{ij} + c_j - c_{ji}$$
$$b \equiv c_i(x_j - x_{ji}) + c_j(x_i - x_{ij}) - (c_{ij} + c_{ji})(x_i - x_{ij} + x_j - x_{ji})$$
$$c \equiv -(c_{ij} + c_{ji})(x_i - x_{ij})(x_j - x_{ji}) \quad (24)$$

[JDC: Can we merge the updates for x'_{ii} and x'_{ij} to simplify this description?] Once a sufficient number of iterations have been completed to compute a stable estimate of $\mathbf{X}$ (such as the relative convergence criteria $||\mathbf{X}^{(n)} - \mathbf{X}^{(n-1)}|| / ||\mathbf{X}^{(n)}|| < 10^{-4}$, the maximum likelihood transition matrix estimate $\mathbf{T}$ is computed as

$$T_{ij} = \frac{x_{ij}}{x_i}. \quad (25)$$

Note that the equilibrium probability vector $\boldsymbol{\pi}$ computed during the Gaussian mixture model fitting is not respected during this step.

B. Fitting a maximum likelihood HMM

The HMM model parameters $\boldsymbol{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$ are fit to the observed data $\mathbf{O}$ through use of the expectation-maximization (EM) algorithm [24]. This is an iterative procedure, where the model parameters are subsequently refined through successive iterations.

During each iteration, the Baum-Welch algorithm [12] is used to compute, for each trace n, $\boldsymbol{\Xi}^{(n)} \equiv (\xi_{tij}^{(n)})$, which represents the probability that the system transitions from hidden state i at time $t-1$ to hidden state j at time t, and $\gamma_{ti}^{(n)}$, the probability that the system ~~was~~ is in state i at time t. This is accomplished by first executing the *forward algorithm*, which proceeds (suppressing the superscript (n)) as

$$\alpha_{tj} = \begin{cases} \overset{\rho_j?}{\rho_i} \varphi(o_0|\mathbf{e}_j) & t = 0 \\ \varphi(o_t|\mathbf{e}_j) \sum_{i=1}^{M} \alpha_{(t-1)i} T_{ij} & t = 1, \ldots, \cancel{T_n}^{(n)?} \end{cases} \quad (26)$$

or suppress?

followed by the *backward algorithm*,

$$\beta_{ti} = \begin{cases} 1 & t = T_n \\ \sum_{j=1}^{M} T_{ij} \varphi(o_{t+1}|\mathbf{e}_j) \beta_{(t+1)j} & t = T_n - 1, \ldots, 0 \end{cases} \quad (27)$$

The ~~$M \times M \times T_n$~~ $T_n \times M \times M$ matrix $\boldsymbol{\Xi}$ is then computed for $t = 0, \ldots, (T_n - 1)$ as

or $T_n - 1$?

$$\xi_{tij} = \alpha_{ti} \varphi(o_{t+1}|\underset{e_j?}{\mathbf{e}_i}) T_{ij} \beta_{(t+1)j} / \sum_{i=1}^{M} \alpha_{Ti} \quad (28)$$

$$\gamma_{ti} = \sum_{j=1}^{M} \xi_{tij} \quad (29)$$

In practice, the logarithms of these quantities are computed instead to avoid numerical underflow. [JDC: Not sure what happened to γ_{ti} for $t = T_n$. Check this?]