Learning Gaussian hidden Markov models from aggregate data

Rahul Singh and Yongxin Chen

***Abstract*—We consider system identification (learning) problems for Gaussian hidden Markov models (GHMMs). We propose an algorithm to tackle the cases where the data is recorded in aggregate (collective) form generated by a large population of individuals following a certain dynamics. Our parameter learning algorithm is built upon the expectation-maximization algorithm with a novel expectation step proposed recently known as the collective Gaussian forward-backward algorithm. The proposed learning algorithm generalizes the traditional Baum-Welch learning algorithm for GHMMs as it naturally reduces to the latter in case of individual observations.**

***Index Terms*—Markov process, identification, stochastic systems**

I. Introduction

LEARNING and inference from population-level data have gained great attention lately [1], [2], [3], [4]. In such settings due to the measurement cost or privacy reasons, the observations are collected in aggregate form such that the individual's association is unknown, as compared to more traditional setting where individual observations are recorded. Examples of aggregate data include human ensemble flow analysis and disease spread analysis [5], [6], among others. Estimating model parameters from such aggregate observations is an important problem and unavailability of individual's data association makes parameter learning more challenging.

In this work, we are concerned with the problem of parameter estimation or learning of continuous state hidden Markov models with Gaussian densities (also known as Gaussian hidden Markov models (GHMMs)) from aggregate observations. GHMMs [7], [4] are popular in modeling the temporal evolution of agents and has applications in many real-world problems such as optimal filtering [8], signal classification [7], and activity detection [9]. Moreover, the well known Kalman-filter is also for GHMMs. The problem of parameter estimation of GHMMs from individual observations have been studied in various works including [10], [11], [7], [12]. However, the existing techniques are not applicable to aggregate data settings due to loss of individual's association in recorded (noisy) observations.

The learning (system identification) and inference problems from aggregate data have been studied under the more general framework of collective graphical models (CGMs) [5]. The inference algorithms are fundamental to parameter learning and within the CGM framework, multiple inference algorithms have been proposed including Non-linear belief propagation [13] and Bethe-RDA [14]. However, these methods assume explicit observation model. A more recent work is Sinkhorn belief propagation [15] for general CGMs and it is under the name of collective forward-backward (CFB) algorithm [3] when specialized to discrete state HMMs. Based on the CFB algorithm, learning of discrete state HMMs has been studied in [16]. Most of the existing works on the inference of collective HMMs are focused on discrete states and can not be directly applied to GHMMs due to impracticability of discretization of continuous states for large dimensions. Recently an aggregate inference algorithm applicable to GHMMs known as the collective Gaussian forward-backward (CGFB) algorithm [4] has been proposed. The CGFB algorithm is a message passing type aggregate inference algorithm for GHMMs and exhibits convergence guarantees.

We propose an algorithm for parameter learning of GHMMs from aggregate observations. We employ the popular EM algorithm [17], [18] for this purpose. Based on the continuous aggregate observations, we use the CGFB algorithm [4] for estimating a function of the expected values of the latent variables (the E-step of EM algorithm). Then in the M-step, the maximum likelihood parameter estimates are computed. Our proposed learning algorithm also has local convergence guarantee. Note that the CGFB algorithm is focused on inference of GHMMs from aggregate data. In contrast, this work is focused on learning the GHMMs parameters from aggregate data.

The rest of the paper is organized as follows. Section II discusses the aggregate inference algorithms for GHMMs and Section III contains the proposed algorithm. In Section IV, we provide numerical experiments validating the proposed algorithm followed by the concluding remarks in Section V.

II. Background

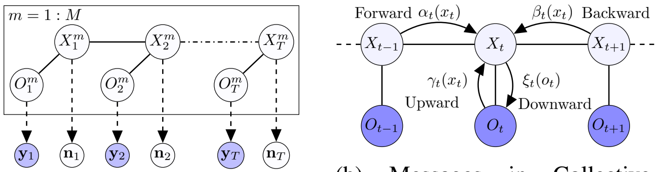
HMMs consist of a Markov process describing the evolution of hidden state over time and a corresponding observation process corrupted by noise. Let the state variables be $X_1, X_2, \dots$ and corresponding observation variables be $O_1, O_2, \dots$. An HMM is parameterized by initial distribution $p(X_1)$, the transition probabilities $p(X_{t+1} \mid X_t)$, and observation probabilities $p(O_t \mid X_t)$ for each time step. The joint distribution of a length T HMM is factorized as

$$p(\mathbf{x}, \mathbf{o}) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1} \mid x_t) \prod_{t=1}^{T} p(o_t \mid x_t), \qquad (1)$$

where $\mathbf{x} = \{x_1, x_2, \dots, x_T\}$ and $\mathbf{o} = \{o_1, o_2, \dots, o_T\}$ represent a particular assignment of hidden and observation variables, respectively. The state and observation variables can take either continuous or discrete values. In this paper, we assume that both state and observation variables take continuous values. There are two main problems concerning

This work was supported by NSF under grant 1942523 and 2008513.

R. Singh and Y. Chen are with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA; {rasingh, yongchen}@gatech.edu



(a) Collective HMM

(b) Messages in Collective HMM

Fig. 1: Collective HMMs and messages (shaded nodes represent aggregate noisy observations).

HMMs: inference of hidden states given the observations along with the underlying HMM parameters and learning (parameter estimation) of HMMs from the observations.

A. *Collective Hidden Markov Models*

Collective (aggregate) HMMs are generative models wherein a population of M individuals independently follow a certain Markov chain and the noisy observations are recorded in aggregate form such that the association to the individuals is unknown. Let $X_t^{(m)}$ be the random variable representing the state of m^{th} individual at time t and $O_t^{(m)}$ be the observation variable of m^{th} individual at time t. The observations are made in aggregate form $y_t(o_t)$ representing the distribution of the collective observations for each time step t. The goal of inference in collective HMMs is to estimate the aggregate state distributions $n_t(x_t)$ given all the aggregate observation distributions. A pictorial representation of a collective HMM is depicted in Figure 1a. Here, $n_t(x_t)$ is an estimate of the state distribution of the M agents at time step t.

Inference in collective HMMs aims to estimate the aggregate hidden distributions based on the indistinguishable aggregate measurements. Traditional inference algorithms such as forward-backward algorithms can not be used here due to data aggregation. The collective forward-backward algorithm [3] was recently proposed for aggregate inference in HMMs. It is a message passing type aggregate inference method employing four different types of messages over the underlying HMM as shown in Figure 1b, where $\alpha_t(x_t)$ are messages in the forward direction and $\beta_t(x_t)$ are messages in the backward direction. Moreover, $\gamma_t(x_t)$ denote the messages from corresponding observation nodes to hidden nodes and $\xi_t(o_t)$ are the messages from corresponding hidden nodes to observation nodes. The CFB algorithm was originally proposed for discrete state and discrete observation HMMs and its extension to discrete state and continuous observation settings was studied in [19]. When specialized to collective GHMMs, the CFB algorithm is termed the collective Gaussian forward-backward (CGFB) algorithm [4].

B. *Collective Gaussian Forward-Backward Algorithm*

A GHMM model is characterizated by

$$X_{t+1} = AX_t + W_t \tag{2a}$$

$$O_t = CX_t + V_t \tag{2b}$$

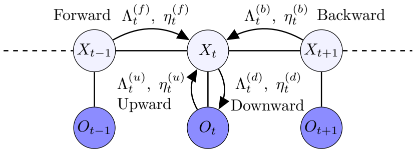

Forward $\Lambda_t^{(f)},\ \eta_t^{(f)}$
$\Lambda_t^{(b)},\ \eta_t^{(b)}$ Backward
X_{t-1}
X_t
X_{t+1}
$\Lambda_t^{(u)},\ \eta_t^{(u)}$
Upward
$\Lambda_t^{(d)},\ \eta_t^{(d)}$
Downward
O_{t-1}
O_t
O_{t+1}

Fig. 2: Messages in the CGFB algorithm.

with $W_t \sim \mathcal{N}(w_t; 0, Q)$, $V_t \sim \mathcal{N}(v_t; 0, R)$, and $X_1 \sim \mathcal{N}(x_1; \pi, \Pi)$. We assume the state and observation space dimensions to be d and s, respectively. Alternatively, GHMMs can be viewed as special cases of continuous state HMMs where the model densities take the form

$$p(X_{t+1}|X_t, \theta_X) = \mathcal{N}(x_{t+1}; Ax_t, Q) \tag{3a}$$

$$p(O_t|X_t, \theta_O) = \mathcal{N}(o_t; Cx_t, R) \tag{3b}$$

$$p(X_1|\theta_1) = \mathcal{N}(x_1; \pi, \Pi), \tag{3c}$$

where $\theta_X = \{A, Q\}$, $\theta_O = \{C, R\}$, and $\theta_1 = \{\pi, \Pi\}$ are the parameters characterizing model densities in the GHMM.

The aggregate observations constitute a total of M trajectories of continuous observations over a single GHMM characterized by (3). Let the recorded observations be $\{o_1^{(m)}, o_2^{(m)}, \ldots, o_t^{(m)}\}$, $\forall m = 1, 2, \ldots, M$ with $o_t^{(m)}$ being the continuous observations of the m^{th} trajectory at time t. The goal of the collective GHMM inference is to estimate the distributions $n_t(x_t)$, $\forall t$. It is assumed that the aggregate observations are approximated by Gaussian densities at each time step t, that is,

$$y_t(o_t) \sim \mathcal{N}(o_t; \hat{\mu}_t, \hat{P}_t), \tag{4}$$

where $\hat{\mu}_t$ and $\hat{P}_t$ are estimated from the observations. The messages in collective GHMMs are characterized by the following theorem.

Algorithm 1 CGFB Algorithm

```
Initialize all the message parameters
while not converged do
  Forward pass:
  for $t = 2, 3, \ldots, T$ do
    i) Update upward parameters $\Lambda_{t-1}^{(u)}$ and $\eta_{t-1}^{(u)}$
    ii) Update forward parameters $\Lambda_t^{(f)}$ and $\eta_t^{(f)}$
    ii) Update downward parameters $\Lambda_t^{(d)}$ and $\eta_t^{(d)}$
  end for
  Backward pass:
  for $t = T - 1, \ldots, 1$ do
    i) Update upward parameters $\Lambda_{t+1}^{(u)}$ and $\eta_{t+1}^{(u)}$
    ii) Update backward parameters $\Lambda_t^{(b)}$ and $\eta_t^{(b)}$
    ii) Update downward parameters $\Lambda_t^{(d)}$ and $\eta_t^{(d)}$
  end for
end while
Estimate required state density parameters $\mu_t$ and $P_t$
```

Theorem 1 ([4]). *The forward, backward, upward, and downward messages in collective GHMM are in the form:*

$$\alpha_t(x) \propto \exp\left(-\frac{1}{2}x^T\Lambda_t^{(f)}x + x^T\eta_t^{(f)}\right), \tag{5a}$$

$$\beta_t(x) \propto \exp\left(-\frac{1}{2}x^T\Lambda_t^{(b)}x + x^T\eta_t^{(b)}\right), \tag{5b}$$

$$\gamma_t(x) \propto \exp\left(-\frac{1}{2}x^T\Lambda_t^{(u)}x + x^T\eta_t^{(u)}\right), \tag{5c}$$

$$\xi_t(x) \propto \exp\left(-\frac{1}{2}x^T\Lambda_t^{(d)}x + x^T\eta_t^{(d)}\right), \tag{5d}$$

for $t = 1, 2, \ldots, T$. Here, the message parameters are the fixed points of the following recursive updates

$$\begin{aligned}
\Lambda_t^{(f)} &= Q^{-1} - Q^{-1}A(A^TQ^{-1}A + \Lambda_{t-1}^{(f)} + \Lambda_{t-1}^{(u)})^{-1}A^TQ^{-1}\\
\eta_t^{(f)} &= Q^{-1}A(A^TQ^{-1}A + \Lambda_{t-1}^{(f)} + \Lambda_{t-1}^{(u)})^{-1}(\eta_{t-1}^{(f)} + \eta_{t-1}^{(u)})\\
\Lambda_t^{(b)} &= A^TQ^{-1}(Q^{-1} + \Lambda_{t+1}^{(b)} + \Lambda_{t+1}^{(u)})^{-1}(\Lambda_{t+1}^{(b)} + \Lambda_{t+1}^{(u)})A\\
\eta_t^{(b)} &= A^TQ^{-1}(Q^{-1} + \Lambda_{t+1}^{(b)} + \Lambda_{t+1}^{(u)})^{-1}(\eta_{t+1}^{(b)} + \eta_{t+1}^{(u)})\\
\Lambda_t^{(d)} &= R^{-1} - R^{-1}C(C^TR^{-1}C + \Lambda_t^{(f)} + \Lambda_t^{(b)})^{-1}C^TR^{-1}\\
\eta_t^{(d)} &= R^{-1}C(C^TR^{-1}C + \Lambda_t^{(f)} + \Lambda_t^{(b)})^{-1}(\eta_t^{(f)} + \eta_t^{(b)})\\
\Lambda_t^{(u)} &= C^T(R + (\hat{P}_t^{-1} - \Lambda_t^{(d)})^{-1})^{-1}C\\
\eta_t^{(u)} &= C^TR^{-1}(R^{-1} + \hat{P}_t^{-1} - \Lambda_t^{(d)})^{-1}(\hat{P}_t^{-1}\hat{\mu}_t - \eta_t^{(d)}),
\end{aligned}$$

with boundary conditions

$$\Lambda_1^{(f)} = \Pi^{-1},\ \eta_1^{(f)} = \Pi^{-1}\pi, \quad \Lambda_T^{(b)} = 0,\ \eta_T^{(b)} = 0.$$

Moreover, the marginals can be computed as

$$n_t(x_t) \propto \alpha_t(x_t)\beta_t(x_t)\gamma_t(x_t) \propto \mathcal{N}(x_t; \mu_t, P_t),$$

where

$$P_t \;=\; (\Lambda_t^{(f)} + \Lambda_t^{(b)} + \Lambda_t^{(u)})^{-1} \tag{6a}$$

$$\mu_t \;=\; P_t(\eta_t^{(f)} + \eta_t^{(b)} + \eta_t^{(u)}). \tag{6b}$$

Based on Theorem 1, the CGFB algorithm (Algorithm 1) was proposed in [4] for aggregate inference in collective GHMMs. The four different messages involved in the algorithm are illustrated in Figure 2. Moreover, the joint densities are

$$\begin{aligned}
n_{t,t+1}(x_t, x_{t+1}) \propto\ & p(x_{t+1}|x_t)\alpha_t(x_t)\gamma_t(x_t)\\
& \beta_{t+1}(x_{t+1})\gamma_{t+1}(x_{t+1})
\end{aligned} \tag{7a}$$

$$n_{t,t}(x_t, o_t) \propto \frac{p(o_t|x_t)\alpha_t(x_t)\beta_t(x_t)y_t(o_t)}{\xi_t(o_t)}. \tag{7b}$$

III. Main Results

In this section, we present our GHMM learning algorithm based on aggregate measurements. We have aggregate observations $\{o_1^{(m)}, o_2^{(m)}, \ldots, o_T^{(m)}\},\ \forall m = 1, 2, \ldots, M$ following GHMM model in (3) such that the individuals association is unknown. We approximate the observations at each time step as Gaussian distributions given by (4). We are interested in estimation of the GHMM parameters $\theta = \{\pi, \Pi, A, Q, C, R\}$ from the aggregate observations. We employ the EM algorithm [18] for this purpose. The EM algorithm involves two operations: the Expectation-step (E-step) computes the log-likelihood of the observations given the current estimate of parameters, and the Maximization-step (M-step) maximizes the log-likelihood.

Denote the set of hidden distribution for all the time steps $t = 1, \ldots, T$ by $\mathbf{n}$ and the set of observation distributions by $\mathbf{y} = \{y_1(\cdot), y_2(\cdot), \ldots, y_T(\cdot)\}$. The E-step required inferring the conditional distribution of $\mathbf{n}$ given the aggregate observations $\mathbf{y}$, its application to collective settings is not straightforward since the aggregate data likelihood $p(\mathbf{n}, \mathbf{y}; \theta)$ does not have a tractable convex (concave) form. It turns out that the aggregate data log-likelihood can be approximated by Bethe free energy [3]

$$\log p(\mathbf{n}, \mathbf{y}; \theta) \propto -\mathcal{F}(\mathbf{n}, \mathbf{y}; \theta). \tag{8}$$

Thus, maximizing $\log p(\mathbf{n}, \mathbf{y}; \theta)$ is equivalent to minimizing Bethe energy $\mathcal{F}(\mathbf{n}, \mathbf{y}; \theta)$. For continuous state HMMs, the Bethe free energy takes the form

$$\begin{aligned}
\mathcal{F}(\mathbf{n}, \mathbf{y}; \theta) = & -\sum_{t=1}^{T}\int n_{t,t}(x_t, o_t)\log p(o_t|x_t)\ dx_t\ do_t\\
& -\sum_{t=1}^{T-1}\int n_{t,t+1}(x_t, x_{t+1})\log p(x_{t+1}|x_t)\ dx_t\ dx_{t+1}\\
& -\int n_1(x_1)\log p(x_1)\ dx_1 - \int_{x_1} n_1(x_1)\log n_1(x_1)\ dx_1\\
& -2\sum_{t=2}^{T-1}\int n_t(x_t)\log n_t(x_t)dx_t - \int n_T(x_T)\log n_T(x_T)dx_T\\
& +\sum_{t=1}^{T-1}\int n_{t,t+1}(x_t, x_{t+1})\log n_{t,t+1}(x_t, x_{t+1})\ dx_t\ dx_{t+1}\\
& +\sum_{t=1}^{T}\int n_{t,t}(x_t, o_t)\log n_{t,t}(x_t, o_t)dx_t\ do_t.
\end{aligned} \tag{9}$$

Based on the above approximation of aggregate data likelihood, we recently proposed CGFB algorithm [4] for inference in GHMMs. For learning the GHMM parameters, we use the CGFB algorithm in the E-step to infer hidden distributions $\mathbf{n}^*$ given the current estimate of the parameters and then update the model parameters based on the maximization of completed data likelihood in the M-step. The sequence of our learning method is listed in Algorithm 2.

Note that the E-step minimizes the free-energy $\mathcal{F}(\mathbf{n}, \mathbf{y}; \theta)$ given by (9) with respect to $\mathbf{n}$. The statistics in the E-step for GHMM with aggregate observations are characterized by the following proposition.

Proposition 1. *The statistics in the E-step for GHMM with* *aggregate observations are computed as*

$$K_{11} = \sum_{t=1}^{T-1} \mathbb{E}[x_t x_t^T] = \sum_{t=1}^{T-1} P_t + \mu_t \mu_t^T \tag{10a}$$

$$K_{22} = \sum_{t=1}^{T-1} \mathbb{E}[x_{t+1} x_{t+1}^T] = \sum_{t=1}^{T-1} P_{t+1} + \mu_{t+1} \mu_{t+1}^T \tag{10b}$$

$$K_{12} = K_{21}^T = \sum_{t=1}^{T-1} \mathbb{E}[x_t x_{t+1}^T] = \sum_{t=1}^{T-1} \Sigma_{12} + \mu_t \mu_{t+1}^T \tag{10c}$$

$$\bar{K}_{11} = \sum_{t=1}^{T} \mathbb{E}[x_t x_t^T] = \sum_{t=1}^{T} P_t + \mu_t \mu_t^T \tag{10d}$$

$$\bar{K}_{22} = \sum_{t=1}^{T} \mathbb{E}[o_t o_t^T] = \sum_{t=1}^{T} (\hat{P}_t + \hat{\mu}_t \hat{\mu}_t^T) \tag{10e}$$

$$\bar{K}_{12} = \bar{K}_{21}^T = \sum_{t=1}^{T} \mathbb{E}[x_t o_t^T] = \sum_{t=1}^{T} \bar{\Sigma}_{12} + \mu_t \hat{\mu}_t^T \tag{10f}$$

with

$$\Sigma_{12} = P_t A^T Q^{-1} (Q^{-1} + \Lambda_{t+1}^{(b)} + \Lambda_{t+1}^{(u)})^{-1} \tag{11a}$$

$$\bar{\Sigma}_{12} = P_t C^T R^{-1} (R^{-1} + \hat{P}_t^{-1} - \Lambda_t^{(d)})^{-1}. \tag{11b}$$

Algorithm 2 Approximate EM algorithm for GHMM parameter learning

```
Initialize model parameters $\theta^0 = \{\pi, \Pi, A, Q, C, R\}$
for $\ell = 1, 2, \dots$ do
    E-step: Obtain hidden densities $\mathbf{n}^*$ using the CGFB algorithm with parameters $\theta^{\ell-1}$
    M-step: Update $\theta^\ell = \text{argmin}_\theta \mathcal{F}(\mathbf{n}^*, \theta)$ using Equation (12)
end for
```

Proof. Note that (10a)-(10b) and (10d)-(10e) directly follow from definitions. We prove (10c) and (10e).

First, using (7a), the joint density $n_{t,t+1}(x_t, x_{t+1})$ equals

$$n_{t,t+1}(x_t, x_{t+1}) \propto \exp\left(-\frac{1}{2} x_t^T \Lambda_{11} x_t - \frac{1}{2} x_{t+1}^T \Lambda_{22} x_{t+1} + \frac{1}{2} x_{t+1}^T \Lambda_{21} x_t + \frac{1}{2} x_t^T \Lambda_{12} x_{t+1} \right)$$

with $\Lambda_{11} = A^T Q^{-1} A + \Lambda_t^{(f)} + \Lambda_t^{(u)}$, $\Lambda_{12} = -A^T Q^{-1}$, $\Lambda_{21} = -Q^{-1} A$, and $\Lambda_{22} = Q^{-1} + \Lambda_{t+1}^{(b)} + \Lambda_{t+1}^{(u)}$. Using these partitioned precision matrices, the covariance matrices can be computed as

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}^{-1}.$$

Considering concatenated vector $[x_t^T, x_{t+1}^T]^T$, and taking expectation over the joint distribution

$$\mathbb{E}[x_t x_{t+1}^T] = \Sigma_{12} + \mu_t \mu_{t+1}^T,$$

where Σ_{12} is the covariance between variables x_t and x_{t+1}. Next, using (7b), the joint density $n_{t,t}(x_t, o_t)$ becomes

$$n_{t,t}(x_t, o_t) \propto \exp\left(-\frac{1}{2} x_t^T \bar{\Lambda}_{11} x_t - \frac{1}{2} o_t^T \bar{\Lambda}_{22} o_t \quad \frac{1}{2} o_t^T \bar{\Lambda}_{21} x_t + \frac{1}{2} x_t^T \bar{\Lambda}_{12} o_t \right)$$

with $\bar{\Lambda}_{11} = C^T R^{-1} C + \Lambda_t^{(f)} + \Lambda_t^{(b)}$, $\bar{\Lambda}_{12} = -C^T R^{-1}$, $\bar{\Lambda}_{21} = -R^{-1} C$, and $\bar{\Lambda}_{22} = R^{-1} + \hat{P}_t^{-1} - \Lambda_t^{(d)}$. Using these partitioned precision matrices, the covariance matrices can be computed as

$$\begin{bmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \bar{\Lambda}_{11} & \bar{\Lambda}_{12} \\ \bar{\Lambda}_{21} & \bar{\Lambda}_{22} \end{bmatrix}^{-1}.$$

Considering concatenated vector $[x_t^T, o_t^T]^T$, and taking expectation over the joint distribution

$$\mathbb{E}[x_t o_t^T] = \bar{\Sigma}_{12} + \mu_t \hat{\mu}_t^T,$$

where $\bar{\Sigma}_{12}$ is the covariance between variables x_t and o_t. □

Based on the statistics computed in the E-step, the M-step updates in Algorithm 2 are characterized by the following theorem.

Theorem 2. *The M-step updates in GHMM learning from aggregate data are given by*

$$\pi = \mu_1, \quad \Pi = P_1 \tag{12a}$$

$$A = K_{21} (K_{11})^{-1} \tag{12b}$$

$$Q = \frac{1}{T-1} \left[K_{22} - K_{21} K_{11}^{-1} K_{12} \right] \tag{12c}$$

$$C = \bar{K}_{21} \bar{K}_{11}^{-1} \tag{12d}$$

$$R = \frac{1}{T} \left[\bar{K}_{22} - \bar{K}_{21} \bar{K}_{11}^{-1} \bar{K}_{12} \right] \tag{12e}$$

where K and $\bar{K}$ are computed in the E-step using Proposition 1.

Proof. In the expression of Bethe free energy (9), keeping only the model parameter terms, it can be decomposed into three terms as

$$\mathcal{F}(\mathbf{n}, \mathbf{y}; \theta) = F_1 + F_X + F_O, \tag{13}$$

where F_1 corresponds to initial density parameters, F_X represents transition density parameter terms, and F_O corresponds to observation density parameters.

The updates for initial density parameters $\theta_1 = \{\pi, \Pi\}$ follows directly by minimizing

$$\begin{aligned} F_1 &= -\int n_1(x_1) \log p(x_1) \, dx_1 \\ &= -\int \mathcal{N}(x_1; \mu_1, P_1) \log \left(\mathcal{N}(x_1; \pi, \Pi) \right) \, dx_1. \\ &= \frac{d}{2} \log 2\pi + \frac{1}{2} \log |\Pi| + \frac{1}{2} \left(\mathrm{Tr}\{\Pi^{-1}(P_1 + \mu_1 \mu_1^T)\} \right. \\ &\qquad \left. + \mu_1^T(-2\Pi^{-1}\pi) + \pi^T \Pi^{-1} \pi \right). \end{aligned} \tag{14}$$

Differentiating F_1 with respect to π and Π^{-1} and equating to zero, we get (12a).

The updates for transition density parameters $\theta_X = \{A, Q\}$ are obtained by minimizing

$$F_X = -\sum_{t=1}^{T-1} \int n_{t,t+1}(x_t, x_{t+1}) \log p(x_{t+1}|x_t) \ dx_t \ dx_{t+1}.$$

Expanding $-\log p(x_{t+1}|x_t)$ so that

$$F_X = \frac{1}{2} I + \sum_{t=1}^{T-1} \left(\frac{1}{2} \log |Q| \right). \tag{15}$$

In the above equation,

$$\begin{aligned} I = &\sum_{t=1}^{T-1} \int n_{t,t+1}(x_t, x_{t+1})(x_t^T A^T Q^{-1} A x_t + x_{t+1}^T Q^{-1} x_{t+1} \\ &\qquad - x_t^T A^T Q^{-1} x_{t+1} - x_{t+1}^T Q^{-1} A x_t) \ dx_t \ dx_{t+1} \\ = &\sum_{t=1}^{T-1} \big[\mathrm{Tr}\{A^T Q^{-1} A (P_t + \mu_t \mu_t^T)\} \\ &\qquad + \mathrm{Tr}\{Q^{-1}(P_{t+1} + \mu_{t+1}\mu_{t+1}^T)\} \\ &\qquad - \mathbb{E}[x_t^T A^T Q^{-1} x_{t+1} + x_{t+1}^T Q^{-1} A x_t] \big] \\ = &\mathrm{Tr}\{A^T Q^{-1} A K_{11}\} + \mathrm{Tr}\{Q^{-1} K_{22}\} \\ &\qquad - \mathrm{Tr}\{Q^{-1} A K_{12}\} - \mathrm{Tr}\{A^T Q^{-1} K_{21}\}. \end{aligned} \tag{16}$$

Clearly, F_X is convex in A and Q^{-1}. In the view of (16), differentiating (15) with respect to A and Q^{-1} and equating to zero, we arrive at (12b) and (12c).

The updates for observation density parameters $\theta_O = \{C, R\}$ are obtained by minimizing

$$F_O = -\sum_{t=1}^{T} \int n_{t,t}(x_t, o_t) \log p(o_t|x_t) \ dx_t \ do_t$$

Expanding $-\log p(o_t|x_t)$ so that

$$F_O = \frac{1}{2} J + \frac{1}{2} \sum_{t=1}^{T} \log |R|, \tag{17}$$

where

$$\begin{aligned} J = &\sum_{t=1}^{T} \int n_{t,t}(x_t, o_t)(x_t^T C^T R^{-1} C x_t + o_t^T R^{-1} o_t \\ &\qquad - x_t^T C^T R^{-1} o_t - o_t^T R^{-1} C x_t) \ dx_t \ do_t \\ = &\mathrm{Tr}\{C^T R^{-1} C \bar{K}_{11}\} + \mathrm{Tr}\{R^{-1} \bar{K}_{22}\} \\ &\qquad - \mathrm{Tr}\{R^{-1} C \bar{K}_{12}\} - \mathrm{Tr}\{C^T R^{-1} \bar{K}_{21}\}. \end{aligned} \tag{18}$$

Clearly, F_O is convex in C and R^{-1}. In the view of (18), differentiating (17) with respect to C and R^{-1} and equating to zero, we arrive at (12d) and (12e). □

Since the CGFB algorithm employed in the E-step of the algorithm is guaranteed to convergence [4], our algorithm also exhibits convergence guarantees at least locally. The convergence can be argued due to the convergence of coordinate descent. Since the E-step and M-step in Algorithm 2 are coordinate descent updates of free energy $\mathcal{F}(\mathbf{n}, \mathbf{y}; \theta)$ with respect to $\mathbf{n}$ and θ and thus it decreases monotonically. Moreover, the free energy $\mathcal{F}(\mathbf{n}, \mathbf{y}; \theta)$ equals the Kullback-Leibler divergence between the inferred distribution and the distribution induced by the prior HMM dynamics over the space of trajectories and is thus bounded below by 0. There two properties ensures the local convergence of our algorithm. Note that the parameters are not unique in terms of measurement data likelihood, i.e., there exist multiple sets of parameters which result in the same data likelihood.

Our algorithm scales well with the problem dimension. In particular, the complexity of each iteration increases linearly with the length T of the model. Moreover, the worst case complexity in each iteration is $\mathcal{O}(d^3)$ in terms of the state dimension d due to matrix inversion.

Remark 1. *In case of an ensemble of aggregate observation distributions $\{\mathbf{y}^j\}_{j=1}^J$, we find these J number of hidden distribution sets $\{\mathbf{n}^j\}_{j=1}^J$ in the E-step. Then the parameter updates in the M-step are given by*

$$\pi = \frac{1}{J} \mu_1^J, \quad \Pi = \frac{1}{J} P_1^J \tag{19a}$$

$$A = K_{21}^J \left(K_{11}^J\right)^{-1} \tag{19b}$$

$$Q = \frac{1}{J(T-1)} \left[K_{22}^J - K_{21}^J \left(K_{11}^J\right)^{-1} K_{12}^J \right] \tag{19c}$$

$$C = \bar{K}_{21}^J \left(\bar{K}_{11}^J\right)^{-1} \tag{19d}$$

$$R = \frac{1}{JT} \left[\bar{K}_{22}^J - \bar{K}_{21}^J \left(\bar{K}_{11}^J\right)^{-1} \bar{K}_{12}^J \right], \tag{19e}$$

where the terms with superscript $\mu_1^J = \sum_{j=1}^J \mu_1^j$, $P_1^J = \sum_{j=1}^J P_1^j$, and $K_{ab}^J = \sum_{j=1}^J K_{ab}^j$ with $a, b \in \{1, 2\}$.

Remark 2. *When the aggregate observations are in Dirac form, corresponding to individual observations, our learning algorithm naturally reduces to the standard Baum-Welch [17], [10] learning algorithm for GHMMs.*

IV. Numerical Examples

We perform multiple experiments to evaluate the performance of our proposed learning algorithm. First, we consider a system with true GHMM parameters:

$$\pi = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \Pi = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & \Delta t \\ -\Delta t & 1 - 0.5\Delta t \end{bmatrix}, \quad C = \begin{bmatrix} 0 & \Delta t \end{bmatrix}$$

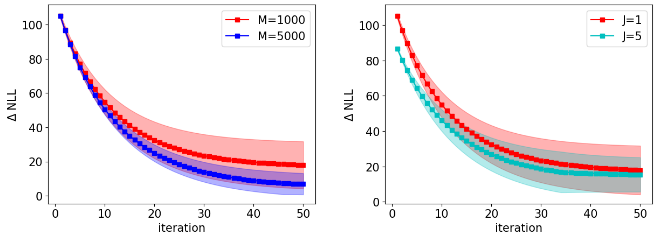
$$Q = \Delta t \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad R = \Delta t \begin{bmatrix} 0.7 \end{bmatrix},$$

where Δt is set to 0.05 for all our experiments. Based on the above GHMM parameters, M number of trajectories are generated and the observations are recorded in aggregate form $\{o_1^{(m)}, o_2^{(m)}, \ldots, o_T^{(m)}\}$, $\forall m = 1, 2, \ldots, M$.

For testing purpose, we generate another set of trajectories (of same length and same population size as in training) and record aggregate observations. We evaluate the performance in terms of difference in negative log likelihoods of test data based on the learned parameters and the ground truth:

$$\Delta\mathrm{NLL} = \mathrm{NLL}(\theta) - \mathrm{NLL}(\theta^*), \tag{20}$$

where θ is the set of learned parameters and θ^* represents ground truth parameters. We normalize ΔNLL by the HMM length. Figure 3(a) shows the performance of our algorithm for different population sizes. A better performance can be observed in case of large population. We further test the performance of our algorithm with different ensemble sizes. We divide the total population M into J number of ensembles and use our algorithm as mentioned in Remark 1. In Figure 3(b), we plot the behavior with different ensemble sizes. It can be observed that small amount of aggregation ($J = 5$) has better performance as compared with full aggregation of observations ($J = 1$).



(a)

(b)

Fig. 3: (a) Δ NLL with $T = 200$ and different population size M. (b) Δ NLL for an ensemble of observations with $T = 200$ and $M = 1000$. The results are averaged over 5 different random seeds.

In our next example, we consider a GHMM model with

$$\pi = \begin{bmatrix} \underline{0}_3 \\ \underline{1}_3 \end{bmatrix}, \; \Pi = \Delta t I_6,$$
$$A = \begin{bmatrix} I_3 & \Delta t I_3 \\ -\Delta t(I_3 + 0.1D) & (1 - 0.5\Delta t) I_3 \end{bmatrix},$$
$$C = \begin{bmatrix} \Delta t I_3 & 0_3 \end{bmatrix}, \; Q = 0.1\Delta t I_6, \; R = 0.2\Delta t I_3.$$

Here $\Delta t = 0.05$, I_k and 0_k respectively denote identity matrix and zero matrix of dimension k, and $\underline{0}_k$ and $\underline{1}_k$ represent zero vector and one vector of dimension k. The matrix D is randomly generated whose elements belong to the interval $[0, 1]$. The performance of our algorithm over this 6-dimensional system with varying M is shown in Figure 4, from which a convergent behavior is observed.

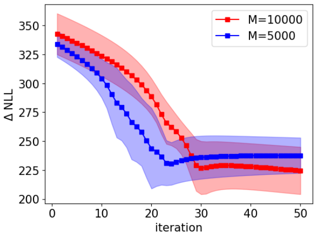


Fig. 4: Δ NLL with $T = 200$ and different M ($d = 6$).

V. Conclusion

In this paper, we proposed an algorithm for learning GHMMs. Our algorithms is based on the EM algorithm wherein we utilized collective Gaussian forward-backward algorithm for data completion in the E-step and derived the parameter updates for the M-step. To the best of our knowledge, our proposed learning algorithm is the first effort towards learning continuous state HMMs from aggregate observations. Our algorithm is restricted to GHMMs and only ensures local convergence. A future direction is to study learning parameters for general models such as mixed-mode GHMMs from aggregate observations [7]. Another interesting direction is to utilize the proposed algorithm to identify models of miniature systems such as bacterial or cell.

References

[1] D. Sheldon, T. Sun, A. Kumar, and T. Dietterich, "Approximate inference in collective graphical models," in *International Conference on Machine Learning*, 2013, pp. 1004–1012.

[2] G. Bernstein and D. Sheldon, "Consistently estimating Markov chains with noisy aggregate data," in *Artificial Intelligence and Statistics*, 2016, pp. 1142–1150.

[3] R. Singh, I. Haasler, Q. Zhang, J. Karlsson, and Y. Chen, "Inference with aggregate data in probabilistic graphical models: An optimal transport approach," *IEEE Transactions on Automatic Control*, vol. in press, 2022.

[4] R. Singh and Y. Chen, "Inference of collective Gaussian hidden Markov models," in *60th IEEE Conference on Decision and Control*, 2021.

[5] D. R. Sheldon and T. G. Dietterich, "Collective graphical models," in *Advances in Neural Information Processing Systems*, 2011, pp. 1161–1169.

[6] G. King, *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press, 2013.

[7] P. L. Ainsleigh, N. Kehtarnavaz, and R. L. Streit, "Hidden Gauss-Markov models for signal classification," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1355–1367, 2002.

[8] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Prentice Hall, 2000.

[9] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa, ""Shape Activity": a continuous-state HMM for moving/deforming shapes with application to abnormal activity detection," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1603–1616, 2005.

[10] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of time series analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[11] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," 1996.

[12] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.

[13] T. Sun, D. Sheldon, and A. Kumar, "Message passing for collective graphical models," in *International Conference on Machine Learning*, 2015, pp. 853–861.

[14] L. Vilnis, D. Belanger, D. Sheldon, and A. Mccallum, "Bethe projections for non-local inference," *Uncertainty in Artificial Intelligence - Proceedings of the 31st Conference, UAI 2015*, 03 2015.

[15] I. Haasler, R. Singh, Q. Zhang, J. Karlsson, and Y. Chen, "Multi-marginal optimal transport and probabilistic graphical models," *IEEE Transactions on Information Theory*, vol. 67, no. 7, pp. 4647–4668, 2021.

[16] R. Singh, Q. Zhang, and Y. Chen, "Learning hidden markov models from aggregate observations," *Automatica*, vol. 137, p. 110100, 2022.

[17] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[19] Q. Zhang, R. Singh, and Y. Chen, "Inference of aggregate hidden Markov models with continuous observations," *IEEE Control Systems Letters*, 2022.