Inference of Aggregate Hidden Markov Models with Continuous Observations

Qinsheng Zhang, Rahul Singh, and Yongxin Chen

***Abstract*—We consider a class of inference problems for large populations where each individual is modeled by the same hidden Markov model (HMM). We focus on aggregate inference problems in HMMs with discrete state space and continuous observation space. The continuous observations are aggregated in a way such that the individuals are indistinguishable from measurements. We propose an aggregate inference algorithm called continuous observation collective forward-backward algorithm. It extends the recently proposed collective forward-backward algorithm for aggregate inference in HMMs with discrete observations to the case of continuous observations.**

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

I. Introduction

HIDDEN Markov models (HMMs) [1] are widely used in the systems and control to model dynamical systems in areas such as robotics, navigation, and autonomy. An HMM has two major components, a Markov process that describes the evolution of the state of the system and a measurement process corrupted by noise. One critical task in HMMs is to reliably estimate the state using the available noisy measurements, known as inference (filtering or smoothing) [2]. Over the years, many filtering algorithms have been proposed. The most well-known one could be the Kalman Filter [3], which is optimal for HMMs with Gaussian transition probability and Gaussian measurement noise. Several other well-known algorithms include the forward-backward algorithm [1] for discrete time discrete state HMMs and the Wonham algorithm [4] for continuous time discrete state HMMs with Gaussian measurement noise.

Recently a new class of inference problems for large populations with aggregate measurements have attracted much attention [5], [6], [7]. In these inference problems, instead of individual measurements, only population-level data in the form of counts or contingency table is available [5]. Such scenarios may occur due to privacy or economic reasons. For instance, in the study of animal flocking, it is too expensive, if not impossible, to track each individual. Aggregate measurements from surveillance cameras or other sensors are instead used. Another timely example is the modeling of pandemic, where the goal is to use available testing data to predict the evolution of the pandemic. Since the test results are anonymized, they form aggregate measurements.

This work was supported by the NSF under grant 1901599, 1942523 and 2008513. Q. Zhang and R. Singh contribute equally to this work. Q. Zhang, R. Singh, and Y. Chen are with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. {qzhang419,rasingh,yongchen}@gatech.edu

Inference for large populations is a difficult task. The lack of individual measurement makes it even more challenging, Standard algorithms such as the forward-backward algorithm for HMMs are no longer applicable. A recent framework developed to address aggregate inference problems is the collective graphical models (CGMs) [5]. Within this framework, several algorithms have been proposed, including approximate MAP [8], non-linear belief propagation [9] and Bethe-RDA [10]. Several other recent works on inference with aggregate observations include [11], [12], [7], [13], [14].

It turns out that the aggregate inference problem for general graphical models can be reformulated as an entropy regularized multi-marginal optimal transport (MOT) problem [6]. Building on this connection between aggregate inference and MOT, we proposed, in [6], [7], a novel algorithm for aggregate inference termed Sinkhorn belief propagation [15], [16]. When used in aggregate inference problems for HMMs, it is called collective forward-backward (CFB) algorithm [7].

However, most of the existing methods for aggregate inference require both the state space and the observation space to be discrete. They are not directly applicable to the setting with continuous observation space. To use these methods in such setting, we need to first discretize the observation space to make it discrete. This step introduces discretization error. More importantly, the discretization becomes impractical when the dimension of the observation space is large. The goal of this work is to develop an efficient aggregate inference algorithm for aggregate HMMs with continuous observation spaces and discrete state spaces. Such HMMs with discrete hidden states and continuous observations are used in applications such as fault detection [4], [17], [18].

In this work, we extend the existing CFB algorithm, which is limited to HMMs with discrete observations, to continuous observation HMMs and call our algorithm as continuous observation collective forward-backward (CO-CFB) algorithm. The CFB algorithm can be formally extended to HMMs with continuous observation by replacing the summation by integration in some key steps. However, this extension does not lead to any practical algorithm since numerical integration by grid discretization is expensive. Rewriting this integration step in terms of expectation circumvents this difficulty, and we arrive at the CO-CFB that only uses samples from the aggregate observations. Moreover, just like CFB, the CO-CFB algorithm also exhibits convergence guarantee. To the best of our knowledge, CO-CFB is the first efficient algorithm for aggregate inference of HMMs with general continuous

observations. Note that the recent work [14] is limited to the special case of HMMs of Gaussian Markov processes. In contrast, our proposed algorithm CO-CFB is applicable to general continuous observations. Another related method is the probability hypothesis density (PHD) filter [19] which aims to estimate the state distribution of a time-varying population. This method is designed for filtering problems while our algorithm is for smoothing.

II. Background

An HMM is a Markov process accompanied by observation noise consisting of a sequence of unobservable states X_t (also known as hidden states) and corresponding observable states O_t for $t = 1, 2, \ldots$. Here X_t and O_t are random variables taking values from $\mathcal{X}$ and $\mathcal{O}$, respectively. In general, both $\mathcal{X}$ and $\mathcal{O}$ can be either finite or infinite sets. We assume $\mathcal{X}$ to be a set with d elements. An HMM is characterized[1] by the distribution $\pi(x_1)$ of the starting state X_1, the transition probability $p(x_{t+1}|x_t)$, and the observation model $p(o_t|x_t)$. The joint probability distribution over T steps reads

$$p(\mathcal{T}) = \pi(x_1) \prod_{t=1}^{T-1} p(x_{t+1}|x_t) \prod_{t=1}^{T} p(o_t|x_t), \tag{1}$$

where $x_t \in \mathcal{X}$ is a realization of X_t, $o_t \in \mathcal{O}$ is a realization of O_t, and $\mathcal{T} = \{x_1, o_1, \cdots, x_t, o_t, \cdots, x_T, o_T\}$ is a sample trajectory.

One of the most important problems in HMMs is Bayesian inference which aims to compute the posterior distributions of the hidden states X_t given a sequence of observations. This is also known as filtering or smoothing (inference) [1] in systems and control. A well-known and efficient algorithm for this task is the standard forward-backward algorithm [1].

A. Aggregate Hidden Markov Models

Aggregate HMMs [9], [7] deal with the problems of estimating the behavior of a collection of indistinguishable HMMs based on aggregate observations. Assume the observation space $\mathcal{O}$ is a finite set. Given M trajectories $\{\mathcal{T}^{(1)}, \cdots, \mathcal{T}^{(M)}\}$ with $\mathcal{T}^{(m)} = \{x_1^{(m)}, o_1^{(m)}, \cdots, x_T^{(m)}, o_T^{(m)}\}$, sampled from the same HMM, the associated aggregate HMM is concerned with the aggregate quantities $\mathbf{n} = \{\mathbf{n}_t, \tilde{\mathbf{n}}_t, \mathbf{n}_{tt}, \tilde{\mathbf{n}}_{tt}\}$ over the entire population defined by, for every $x, x' \in \mathcal{X}, o \in \mathcal{O}$,

$$\tilde{n}_{tt}(x, o) = \sum_{m=1}^{M} \mathbb{I}[x_t^{(m)} = x, o_t^{(m)} = o], t \in \{1, \cdots, T\} \tag{2a}$$

$$n_{tt}(x, x') = \sum_{m=1}^{M} \mathbb{I}[x_t^{(m)} = x, x_{t+1}^{(m)} = x'], t \in \{1, \cdots, T-1\} \tag{2b}$$

$$n_t(x) = \sum_{m=1}^{M} \mathbb{I}[x_t^{(m)} = x], \ t \in \{1, \cdots, T\} \tag{2c}$$

$$\tilde{n}_t(o) = \sum_{m=1}^{M} \mathbb{I}[o_t^{(m)} = o], \ t \in \{1, \cdots, T\}, \tag{2d}$$

[1]We assume that the HMM is time-homogeneous, that is, transition and observation probabilities are time-invariant. This is for notational convenience. All the results in this paper apply to time-varying HMMs.

where $\mathbb{I}$ denotes the indicator function. Note that we have adopted the bold symbol to represent the vectors, e.g., $\mathbf{n}_t$ is a vector with entry $n_t(x)$ for each $x \in \mathcal{X}$. These quantities represent the counts of M realizations of HMM taking specific values; they are histograms. Clearly, $\mathbf{n} \in \mathbb{L}_M^{\mathbb{Z}}$, the *integer-valued scaled local polytope* [9], that is, the entries of $\mathbf{n}$ are all integers and they satisfy the constraints

$$\sum_{o \in \mathcal{O}} \tilde{n}_t(o) = M, \ \sum_{x \in \mathcal{X}} n_t(x) = M, \ \forall t \in \{1, \cdots, T\} \tag{3a}$$

$$\sum_{x \in \mathcal{X}} n_{tt}(x, x_{t+1}) = n_{t+1}(x_{t+1}), \sum_{x \in \mathcal{X}} n_{tt}(x_t, x) = n_t(x_t), \quad \forall t \in \{1, \cdots, T-1\} \tag{3b}$$

$$\sum_{o \in \mathcal{O}} \tilde{n}_{tt}(x, o) = n_t(x), \quad \forall t \in \{1, \cdots, T\} \tag{3c}$$

$$\sum_{x \in \mathcal{X}} \tilde{n}_{tt}(x, o) = \tilde{n}_t(o), \quad \forall t \in \{1, \cdots, T\}. \tag{3d}$$

The HMMs are aggregate in the sense that they are indistinguishable to each other. In this setting, the observations[2] are $\mathbf{y}_t$ with $y_t(o)$ denoting the number of trajectories such that $o_t^{(m)} = o$, imposing the constraints $\tilde{\mathbf{n}}_t = \mathbf{y}_t$ for $t = 1, \ldots, T$.

The goal of inference in aggregate HMMs is to estimate the most likely $\mathbf{n}$ given the aggregate measurements $\{\mathbf{y}_t\}$. The exact inference is proved to be computationally infeasible [5] for problems with large T and M. It is proposed in [7] that this aggregate inference can be approximately achieved by solving a free energy minimization problem. Moreover, the approximation error vanishes as the size M of the population goes to infinity. For the sake of simplicity and without loss of generality, we normalize the observation $\mathbf{y}_t$ and the statistics $\mathbf{n}$ by population size M, yielding a modification on (3a)

$$\sum_{o \in \mathcal{O}} \tilde{n}_t(o) = 1, \quad \sum_{x \in \mathcal{X}} n_t(x) = 1 \quad \forall t \in \{1, \cdots, T\}. \tag{4}$$

Denote the local polytope (without integer constraints) described by Equations (4)-(3b)-(3c)-(3d) by $\mathbb{M}$ [20], and define the free energy

$$\begin{aligned} \mathcal{F}(\mathbf{n}) = & -\sum_{t=1}^{T} \sum_{x_t, o_t} \tilde{n}_{tt}(x_t, o_t) \log(p(o_t|x_t)/\tilde{n}_{tt}(x_t, o_t)) \\ & - \sum_{t=1}^{T-1} \sum_{x_t, x_{t+1}} n_{tt}(x_t, x_{t+1}) \log(p(x_{t+1}|x_t)/n_{tt}(x_t, x_{t+1})) \\ & - \sum_{x_1} n_1(x_1) \log(\pi(x_1) n_1(x_1)) \\ & - 2 \sum_{t=2}^{T-1} \sum_{x_t} n_t(x_t) \log n_t(x_t) - \sum_{x_T} n_T(x_T) \log n_T(x_T), \end{aligned} \tag{5}$$

then the aggregate inference problem is equivalent [7] to the following convex optimization problem.

Problem 1.

$$\min_{\mathbf{n} \in \mathbb{M}} \quad \mathcal{F}(\mathbf{n}) \tag{6a}$$

$$\textit{subject to} \quad \tilde{\mathbf{n}}_t = \mathbf{y}_t, \quad \forall t \in \{1, \cdots, T\}. \tag{6b}$$

[2]A slightly different observation has been considered in [9].

B. Collective Forward-Backward Algorithm

The collective forward-backward (CFB) algorithm (Algorithm 1) was proposed in [7] to address Problem 1. It leverages an elegant connection between MOT [21], [22], [23] and this inference problem with fixed marginal constraints. We refer the reader to [7] for more details on the CFB algorithm and its connections to multi-marginal optimal transport. The solution to Problem 1 is characterized by the following.

Theorem 1 ([7, Corollary 1])**.** *The solution to the inference problem (Problem 1) for aggregate HMM is*

$$n_t(x_t) \propto \alpha_t(x_t)\beta_t(x_t)\gamma_t(x_t), \quad \forall t = 1, 2 \ldots, T \tag{7}$$

where $\alpha_t(x_t), \beta_t(x_t)$, $\gamma_t(x_t)$, and $\xi_t(o_t)$ correspond to the fixed point of the following updates

$$\alpha_t(x_t) \propto \sum_{x_{t-1}} p(x_t|x_{t-1})\alpha_{t-1}(x_{t-1})\gamma_{t-1}(x_{t-1}), \tag{8a}$$

$$\beta_t(x_t) \propto \sum_{x_{t+1}} p(x_{t+1}|x_t)\beta_{t+1}(x_{t+1})\gamma_{t+1}(x_{t+1}), \tag{8b}$$

$$\gamma_t(x_t) \propto \sum_{o_t} p(o_t|x_t)\frac{y_t(o_t)}{\xi_t(o_t)}, \tag{8c}$$

$$\xi_t(o_t) \propto \sum_{x_t} p(o_t|x_t)\alpha_t(x_t)\beta_t(x_t), \tag{8d}$$

with boundary conditions $\alpha_1(x_1) = \pi(x_1)$ and $\beta_T(x_T) = 1$.

Algorithm 1 Collective Forward-Backward Algorithm

Initialize all the messages $\alpha_t(x_t), \beta_t(x_t), \gamma_t(x_t), \xi_t(o_t)$
while not converged **do**
  Forward pass:
  for $t = 2, 3, \ldots, T$ **do**
    i) Update $\gamma_{t-1}(x_{t-1})$ using (8c)
    ii) Update $\alpha_t(x_t), \xi_t(o_t)$ according to (8a), (8d)
  end for
  Backward pass:
  for $t = T-1, \ldots, 1$ **do**
    i) Update $\gamma_{t+1}(x_{t+1})$ using (8c)
    ii) Update $\beta_t(x_t), \xi_t(o_t)$ according to (8b),(8d)
  end for
end while

Figure 1 illustrates $\alpha, \beta, \gamma, \xi$ in Theorem 1. The Forward-backward algorithm [1] for standard HMM inference is a special case of the CFB algorithm when all the measurements $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t\}$ are Dirac distributions [7].

III. Main Results

We consider smoothing (inference) problems for aggregate HMMs with continuous observations. Specifically, we consider an inference problem similar to Problem 1 but assume the observation space of the underlying HMM to be a subset of the Euclidean space, i.e., $\mathcal{O} \subset \mathbb{R}^\ell$. An important instance of this observation model corresponds to the Gaussian measurement noise, that is, $p(o_t|x_t) = \mathcal{N}(\mu(x_t), \Sigma(x_t))$ with $\mathcal{N}(\mu, \Sigma)$ denoting the probability density with mean μ and covariance Σ.

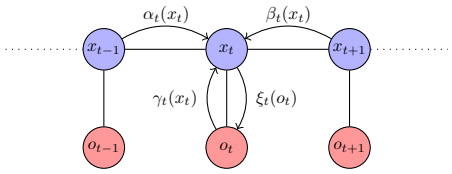


Fig. 1: Illustration of the CFB algorithm.

A. Aggregate Inference with Continuous Observations

We start with an ideal setting when the aggregate observations are given in closed-form. More precisely, we assume the observations are the probability distributions over $\mathcal{O}$, that is, $y_t(\cdot)$ maps $o \in \mathcal{O}$ to $\mathbb{R}_+$ and $\int_{o\in\mathcal{O}} y_t(o)\, do = 1$. In this scenario, the aggregate inference can again be formulated by Problem 1 with two modifications due to the continuous observations. First, the constraints (4) and (3c) become

$$\int_{o\in\mathcal{O}} \tilde{n}_t(o)do = 1, \ \sum_{x\in\mathcal{X}} n_t(x) = 1 \ \forall t \in \{1, \cdots, T\} \tag{9a}$$

$$\int_{o\in\mathcal{O}} \tilde{n}_{tt}(x, o)do = n_t(x) \quad \forall t \in \{1, \cdots, T\}. \tag{9b}$$

Thus, the constraint set $\mathbb{M}$ is described by (9a)-(3b)-(9b)-(3d). Second, the objective function is free energy in similar form but the summation over $\mathcal{O}$ needs to be replaced by integration. For convenience, we overload the bold symbol notation so that $\tilde{\mathbf{n}}_t, \mathbf{y}_t$ are functions over $\mathcal{O}$ instead of vectors, and $\tilde{\mathbf{n}}_{tt}$ is a function over $(\mathcal{X}, \mathcal{O})$.

It turns out that the solution to this aggregate inference problem with continuous observation has a similar characterization as in the discrete setting as in the following theorem.

Theorem 2. *The solution to inference problem (Problem 1) in an aggregate HMM with continuous observation is*

$$n_t(x_t) \propto \alpha_t(x_t)\beta_t(x_t)\gamma_t(x_t), \quad \forall t = 1, \ldots, T \tag{10}$$

where $\alpha_t(x_t), \beta_t(x_t), \gamma_t(x_t), \xi_t(o_t)$ correspond to the fixed point of the following updates

$$\alpha_t(x_t) \propto \sum_{x_{t-1}} p(x_t|x_{t-1})\alpha_{t-1}(x_{t-1})\gamma_{t-1}(x_{t-1}) \tag{11a}$$

$$\beta_t(x_t) \propto \sum_{x_{t+1}} p(x_{t+1}|x_t)\beta_{t+1}(x_{t+1})\gamma_{t+1}(x_{t+1}) \tag{11b}$$

$$\gamma_t(x_t) \propto \int_{\mathcal{O}} p(o_t|x_t)\frac{y_t(o_t)}{\xi_t(o_t)}do_t \tag{11c}$$

$$\xi_t(o_t) \propto \sum_{x_t} p(o_t|x_t)\alpha_t(x_t)\beta_t(x_t) \tag{11d}$$

with boundary conditions $\alpha_1(x_1) = \pi(x_1)$ and $\beta_T(x_T) = 1$.

Proof. Introducing Lagrange multipliers $a_t(x_t), b_t(o_t), c_t(x_t), d_t(x_t), e_t(o_t), f_{1t}, f_{2t}$ for constraints (9b), (3d), (3b), (9a), we arrive at the Lagrangian

$$\mathcal{L} = \mathcal{F}(\mathbf{n}) + \sum_{t=1}^{T}\sum_{x_t} a_t(x_t)[\int_{\mathcal{O}} \tilde{n}_{tt}(x_t, o_t)do_t - n_t(x_t)]$$

$$
\begin{aligned}
&+\sum_{t=1}\int_{\mathcal{O}} b_t(o_t)[\sum_{x_t}\tilde{n}_{tt}(x_t,o_t)-\tilde{n}_t(o_t)]do_t\\
&+\sum_{t=1}^{T-1}\sum_{x_t} c_{t+1}(x_{t+1})[\sum_{x_t} n_{tt}(x_t,x_{t+1})-n_{t+1}(x_{t+1})]\\
&+\sum_{t=1}^{T-1}\sum_{x_t} d_t(x_t)[\sum_{x_{t+1}} n_{tt}(x_t,x_{t+1})-n_t(x_t)]\\
&+\sum_{t=1}^{T}\int_{\mathcal{O}} e_t(o_t)[\tilde{n}_t(o_t)-y_t(o_t)]do_t\\
&+\sum_{t=1}^{T}\{f_{1t}[\sum_{x_t} n_t(x_t)-1]+f_{2t}[\int_{\mathcal{O}}\tilde{n}_t(o_t)do_t-1]\}.
\end{aligned}
$$

Setting the derivatives of the Lagrangian with respect to the $\{\mathbf{n}_t, \tilde{\mathbf{n}}_t, \mathbf{n}_{tt}, \tilde{\mathbf{n}}_{tt}\}$ to zero, we obtain the optimality conditions

$$
\begin{cases}
-1-\log n_t(x_t)-a_t(x_t)-d_t(x_t)+f_{1t}=0, t=1\\
-1-\log n_t(x_t)-a_t(x_t)-c_t(x_t)+f_{1t}=0, t=T\\
-2-2\log n_t(x_t)-a_t(x_t)-c_t(x_t)-d_t(x_t)+f_{1t}=0\\
\qquad t=2\cdots T-1
\end{cases} \tag{12a}
$$

$$
-b_t(o_t)+e_t(o_t)+f_{2t}=0 \tag{12b}
$$

$$
\tilde{n}_{tt}(x_t,o_t)=p(o_t|x_t)e^{-a(x_t)-b(o_t)} \tag{12c}
$$

$$
n_{tt}(x_t,x_{t+1})=p(x_{t+1}|x_t)e^{-c(x_{t+1})-d(x_t)}. \tag{12d}
$$

Define, for all $t=1,2,\cdots,T$,

$$
\alpha_t(x_t)=\sum_{x_{t-1}} p(x_t|x_{t-1})e^{-d_t(x_{t-1})} \tag{13a}
$$

$$
\beta_t(x_t)=\sum_{x_{t+1}} p(x_{t+1}|x_t)e^{-c_t(x_{t+1})} \tag{13b}
$$

$$
\xi_t(o_t)=\sum_{x_t} p(o_t|x_t)e^{-a_t(x_t)} \tag{13c}
$$

$$
\gamma_t(x_t)=\int_{\mathcal{O}} p(o_t|x_t)e^{-b_t(o_t)}do_t. \tag{13d}
$$

Next we present the case with $t=2,\cdots,T-1$ in (12a); the other two cases with $t=1$ or $t=T$ can be analyzed similarly. It follows directly from (12a) that

$$
n_t(x_t)\propto e^{-\frac{a_t(x_t)+c_t(x_t)+d_t(x_t)}{2}}. \tag{14}
$$

Since $\mathbf{n}_t, \mathbf{n}_{tt}$ follow the constraints (3b), we obtain

$$
n_t(x_t)n_t(x_t)=[\sum_{x_{t-1}} n_{t-1,t}(x_{t-1},x_t)][\sum_{x_{t+1}} n_{tt}(x_t,x_{t+1})].
$$

Plugging (14) for $\mathbf{n}_t$ and (12d) and (13) for $\mathbf{n}_{t-1,t-1}, \mathbf{n}_{t,t}$ into the above equation, we arrive at

$$
e^{-a_t(x_t)}\propto \alpha_t(x_t)\beta_t(x_t). \tag{15}
$$

Since $\mathbf{n}_t, \tilde{\mathbf{n}}_{tt}$ also follow the constraints (3c), we reach

$$
n_t(x_t)n_t(x_t)=[\sum_{x_{t+1}} n_{tt}(x_t,x_{t+1})][\int_{\mathcal{O}}\tilde{n}_{tt}(x_t,o_t)do_t].
$$

Plugging (14) (for $\mathbf{n}_t$), (13b) (13c) and (12d) (for $\mathbf{n}_{tt}, \tilde{\mathbf{n}}_{tt}$) into above equation, we get

$$
e^{-c_t(x_t)}\propto \beta_t(x_t)\gamma_t(x_t). \tag{16}
$$

Similarly, we can obtain

$$
e^{-d_t(x_t)}\propto \alpha_t(x_t)\gamma_t(x_t). \tag{17}
$$

Clearly, (10) is a consequence of (14)-(15)-(16)-(17) by

$$
n_t(x_t)\propto e^{\frac{-a_t(x_t)-c_t(x_t)-d_t(x_t)}{2}}\propto \alpha_t(x_t)\beta_t(x_t)\gamma_t(x_t).
$$

Moreover, (11a) follows by combining (17) and (13a), (11b) follows by combining (16) and (13b), and (11d) follows by combining (15) and (13c). Finally, by (12c) and in view of (13c)

$$
e^{-b_t(o_t)}=\frac{\sum_{x_t}\tilde{n}_{tt}(x_t,o_t)}{\sum_{x_t} p(o_t|x_t)e^{-a_t(x_t)}}\propto\frac{y_t(o_t)}{\xi_t(o_t)}, \tag{18}
$$

where in the last step we have utilized the constraint $\sum_{x_t}\tilde{n}_{tt}(x_t,o_t)=y_t(o_t)$. The update (11c) then follows by plugging (18) into (13d). □

B. Continuous Observation Collective Forward-Backward Algorithm

In Section III-A, we assumed that the full distributions of the observations are given, which is hardly the case in real applications. Very often only samples from these marginal distributions are accessible. Of course, one can always estimate a probability density based on these samples and then run the updates (11). The performance of this approach highly depends on that of the density estimators. Another reason that makes this approach undesirable is that the step (11c) requires numerical integration which is expensive for high-dimensional observation space $\mathcal{O}$. A better approach is to rewrite the step (11c) as

$$
\gamma_t(x_t)=\mathbb{E}_{o_t\sim y_t}\left[\frac{p(o_t|x_t)}{\xi_t(o_t)}\right] \tag{19}
$$

where $\mathbb{E}_{o_t\sim y_t}$ stands for expectation with respect to the distribution $y_t(\cdot)$. Let $\mathbf{o}_t=\{o_t^{(1)},\cdots,o_t^{(M)}\}$ be the aggregate observation at time t, then this update formula for $\gamma_t(x_t)$ can be estimated by the empirical average

$$
\gamma_t(x_t)\approx\frac{1}{M}\sum_{o\in\mathbf{o}_t}\frac{p(o|x_t)}{\xi_t(o)}. \tag{20}
$$

Note that the empirical (sample) average (20) converges to the expected value (19) as $M\to\infty$ due to the law of large numbers. With this update in hand, we propose continuous observation collective forward-backward (CO-CFB) algorithm (Algorithm 2) for solving aggregate inference in aggregate HMMs with continuous observations.

Algorithm 2 can be viewed as a counterpart of the Collective forward-backward algorithm (Algorithm 1) for aggregate HMMs with continuous observations. It resembles the latter except for the difference in the update step for γ_t, replacing (8c) by (20). For a given observation $\mathbf{o}_t=\{o_t^{(1)},\cdots,o_t^{(M)}\}$, if we restrict the observation space to the set $\mathbf{o}_t$, then the aggregate HMM with continuous observation reduces to an aggregate HMM with discrete observation over the set $\mathbf{o}_t$. Since all the samples in $\mathbf{o}_t$ are equally important, the corresponding probability vector measurement $y_t(o_t)$ should be uniform over $\mathbf{o}_t$. Thus, to some extent, CO-CFB and CFB are equivalent.

Remark 1. *Since CO-CFB and CFB are effectively equivalent in implementation, the CO-CFB (Algorithm 2) inherits the convergence properties of CFB and has global convergence guarantee with linear rate [7] under the assumption that the underlying Markov chain is ergodic.*

Algorithm 2 CO-CFB Algorithm

Initialize all the messages $\alpha_t(x_t), \beta_t(x_t), \gamma_t(x_t), \xi_t(o_t)$
while not converged **do**
  Forward pass:
  for $t = 2, 3, \ldots, T$ **do**
    i) Update $\gamma_{t-1}(x_{t-1})$ using (20)
    ii) Update $\alpha_t(x_t), \xi_t(o_t)$ according to (11a), (11d)
  end for
  Backward pass:
  for $t = T-1, \ldots, 1$ **do**
    i) Update $\gamma_{t+1}(x_{t+1})$ using (20)
    ii) Update $\beta_t(x_t), \xi_t(o_t)$ according to (11b), (11d)
  end for
end while

C. Connections to Bayesian Inference in HMMs

As discussed in Section II-B, for HMMs with discrete observations, the aggregate inference reduces to standard Bayesian inference when the measurements are Dirac distributions, and the CFB algorithm (Algorithm 1) reduces to the standard forward-backward algorithm. It turns out that this equivalent relation remains for HMMs with continuous observations. We remark that the forward-backward algorithm for continous-time HMMs with continuous observation is also known as Wonham smoothing [4].

Theorem 3. *When the measurements are Dirac distributions, the CO-CFB algorithm (Algorithm 2) reduces to the standard forward-backward algorithm.*

Proof. When the observation $\mathbf{o}_t = \{o_t^{(1)}, \cdots, o_t^{(M)}\}$ is a Dirac, $o_t^{(1)} = \cdots = o_t^{(M)}$. The update (20) for γ becomes

$$\gamma_t(x_t) = \frac{p(o_t^{(1)}|x_t)}{\xi_t(o_t^{(1)})} \propto p(o_t^{(1)}|x_t).$$

With this γ_t, the marginal distribution $\mathbf{n}_t$ becomes

$$n_t(x_t) \propto \gamma_t(x_t)\alpha_t(x_t)\beta_t(x_t) = p(o_t^{(1)}|x_t)\alpha_t(x_t)\beta_t(x_t), \quad (21)$$

which is same as the forward-backward algorithm [1]. □

IV. Numerical Examples

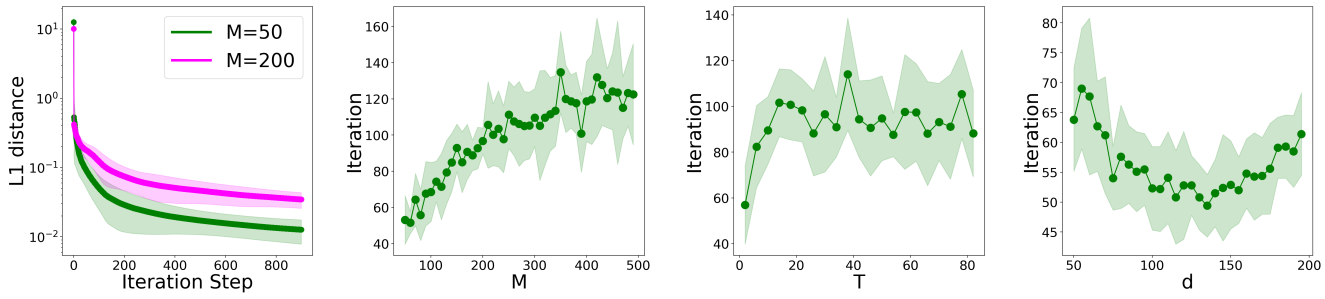
We conduct several experiments to evaluate the performance of the CO-CFB algorithm. The first one is on synthetic data. We consider HMMs with Gaussian observations. The initial state probability π is sampled uniformly over the probability simplex. The transition matrix is generated from a random permutation of a perturbed Identity matrix $\mathcal{I} + 0.05 \times \sqrt{d} \times \exp(U[-1, 1])$, where $\mathcal{I}$ denotes an identity matrix and U stands for a uniform distribution. A normalization step is used to ensure each row is a valid conditional distribution. The Gaussian emission probability for each hidden state is parametrized by a random mean and variance. The mean is sampled from $U[-d, d]$ and the variance is sampled from $U[0.1d, 0.5d]$.

We first demonstrate that our algorithm gives excellent estimation of the distributions of the hidden states of aggregate HMMs. We randomly generate M trajectories from the underlying HMM and produce the observations $\mathbf{o}_t,\ t = 1, \ldots, T$ based on these trajectories. We then use CO-CFB to estimate the hidden state distribution $\mathbf{n}_t$. This is compared to the ground truth $\mathbf{n}_t^*$ with respect to 1-norm $Error = \sum_{t=1}^{T} \|\mathbf{n}_t - \mathbf{n}_t^*\|_1$. Figure 2(a) depicts the estimation error for two different population sizes. We observe that the estimation errors are small in both cases and get smaller as the iteration step increases. Further Figure 2(b) shows the complexity of our algorithm with respect to population size M. To further evaluate the efficiency of our algorithm, we test it over different state dimension d and HMM length T. The results are displayed in Figure 2(c)-(d), from which we observe that the CO-CFB algorithm has great scalability.

Next, we consider a realistic experiment of estimating aggregate robot location estimation with noisy continuous observations. Specifically, we consider a total of $M = 1000$ robots deployed in a geographical area represented by a 10×10 grid such that the (hidden) state dimension is $d = 100$. We let the robots move in the area with a certain dynamics. The observations are recorded as noisy continuous locations in the area and the individual robot's association is anonymized (unknown). The position transition probability follows the log-linear distribution described in [7] and a Gaussian observation with covariance $5 \times \mathcal{I}$. We simulate the dynamics for $T = 10$ time-steps and then utilize our CO-CFB algorithm to estimate the aggregate robot locations from the noisy continuous (noisy) observation trajectories. The leftmost plot in Figure 3 depicts the ground truth of the aggregate robot locations and the continuous anonymized observations are depicted in the middle. The estimated aggregate robot locations are shown in the right of Figure 3 and it can be qualitatively observed that it is close to the ground truth (aggregate) locations. Figure 4 depicts the comparison (in terms of L1-error) of our method against the observation-only estimation, which assigns every observation point to closest choice among the finite robot locations.

V. Conclusion

In this paper, we proposed an efficient algorithm to solve aggregate inference problems for HMMs with continuous observations. This algorithm is based on the CFB algorithm [7] which was designed for aggregate inference for HMMs with discrete observations. The latter was extended to the setting with continuous observations in this work through a reformulation of a key step in the algorithm which enables updating using samples from the observations. One limitation of CO-CFB is that its computational complexity scales linearly with M. This is in contrast to the CFB algorithm whose complexity is independent of M. Moreover, our algorithm is not applicable to continuous state space.



(a) (b) (c) (d)

Fig. 2: (a) Empirical convergence behavior in terms of L_1-error with respect to the underlying state distribution for $M = 50$ and $M = 200$ with $T,\ d = 20$. (b) - (d) Illustration of number of iterations required for convergence for different $M,\ T$ and d, respectively. We use stopping criterion as relative change per node less than 1×10^{-3}. The default setting follows $T = 20, d = 20, M = 200$. Here each "iteration" comprises of updating all the messages involved in the underlying HMM. Each experiment is run with 10 different random seeds and the plots show both means and variances.

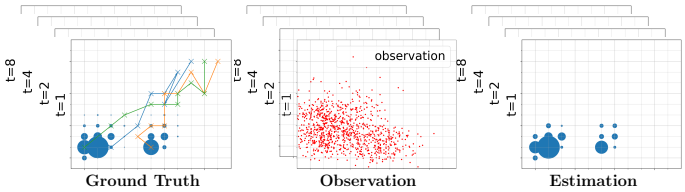


Fig. 3: Performance of our algorithm on aggregate robots location estimation. Left figure shows the simulated underlying robots location distribution, where we plot three trajectories. Middle figure illustrates noisy continuous observations with unknown individual associations. Right figure demonstrates the estimation results from the continuous observations. In these plots, the size of the blue circle at each grid point is proportional to the number of robots at that grid, based on ground truth or inference.

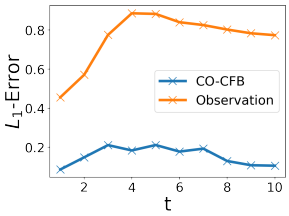


Fig. 4: CO-CFB results in a lower location estimation error compared with observation-only estimation. The latter estimates robots distribution by observation-based nearest neighbors.

REFERENCES

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.

[3] G. Welch, G. Bishop, *et al.*, "An introduction to the Kalman filter," 1995.

[4] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, vol. 2, no. 3, pp. 347–369, 1964.

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

[6] I. Haasler, A. Ringh, Y. Chen, and J. Karlsson, "Estimating ensemble flows on a hidden Markov chain," in *58th IEEE Conference on Decision and Control*, 2019.

[7] 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.

[8] 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.

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

[10] L. Vilnis, D. Belanger, D. Sheldon, and A. McCallum, "Bethe projections for non-local inference," *arXiv preprint arXiv:1503.01397*, 2015.

[11] Y. Chen and J. Karlsson, "State tracking of linear ensembles via optimal mass transport," *IEEE Control Systems Letters*, vol. 2, no. 2, pp. 260–265, 2018.

[12] S. Zeng, "Sample-based population observers," *Automatica*, vol. 101, pp. 166–174, 2019.

[13] J. W. Kim and P. G. Mehta, "Feedback particle filter for collective inference," *arXiv preprint arXiv:2010.06655*, 2020.

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

[15] I. Haasler, A. Ringh, Y. Chen, and J. Karlsson, "Multimarginal optimal transport with a tree-structured cost and the Schrödinger bridge problem," *SIAM Journal on Control and Optimization*, vol. 59, no. 4, pp. 2428–2453, 2021.

[16] 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.

[17] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[18] M. Basseville, "Detecting changes in signals and systems—a survey," *Automatica*, vol. 24, no. 3, pp. 309–326, 1988.

[19] B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4091–4104, 2006.

[20] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[21] L. Nenna, "Numerical methods for multi-marginal optimal transportation," Ph.D. dissertation, 2016.

[22] B. Pass, "On the local structure of optimal measures in the multi-marginal optimal transportation problem," *Calculus of Variations and Partial Differential Equations*, vol. 43, no. 3-4, pp. 529–536, 2012.

[23] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative Bregman projections for regularized transportation problems," *SIAM Journal on Scientific Computing*, vol. 37, no. 2, pp. A1111–A1138, 2015.