Scalable Filtering of Large Graph-Coupled Hidden Markov Models

Ravi N. Haksar[1], Joseph Lorenzetti[2], and Mac Schwager[2]

***Abstract*— We consider the online filtering problem for a graph-coupled hidden Markov model (GHMM) with the Anonymous Influence property. Large-scale spatial processes such as forest fires, social networks, disease epidemics, and robot swarms are often modeled by GHMMs with this property. We derive a scalable online recursive algorithm to produce a belief over states for each HMM node in the GHMM at each time step, given a history of noisy observations. In contrast to prior work, our algorithm is tractable for the high-dimensional discrete state spaces of GHMMs with arbitrary graph structure, and our method scales linearly with the total number of HMMs, i.e., nodes in the graph. We demonstrate the accuracy and scaling of our method using simulation experiments of a wildfire model containing 10^{298} total states and a disease epidemic model containing 10^{18} states.**

I. Introduction

In this work, we consider the problem of producing sequential state estimates online for a class of discrete space and discrete time graph-coupled hidden Markov models (GHMMs). Online sequential inference for large GHMMs requires approximate methods due to the size of the state and observation spaces. For this reason, we leverage variational inference in our approach. Many large-scale, dynamic spatial processes of recent interest are described by this class of models: coupled HMMs have been used to model user interactions in a social network [1] and GHMMs were formulated to model the spread of a contagion in a population [2]. By creating a tractable online inference algorithm, we aim to generate additional interest in using large expressive models in real time for spreading spatial phenomena.

In previous work [3], we developed a scalable framework for generating control policies for graph-based Markov decision processes (GMDPs) which relied on perfect state observations. For this work, we consider realistic sensing models and build a scalable framework for producing state estimates to enable the use of our control framework with natural phenomena which inherently contain state uncertainty.

For a GHMM, each vertex in the graph corresponds to a standard HMM and edges between vertices describe the coupling interactions between HMMs. Many processes, such as forest fires (each tree is an HMM affected by neighboring trees), social networks (each person is an HMM affected by a few other people), disease epidemics (each community is an HMM affected by neighboring communities), and others are thus described by the GHMM framework [4].

This research was supported in part by NSF grant IIS-1646921. We are grateful for this support.

[1]Department of Mechanical Engineering, Stanford University, Stanford, USA rhaksar@stanford.edu

[2]Department of Aeronautics & Astronautics, Stanford University, Stanford, USA {jlorenze, schwager}@stanford.edu

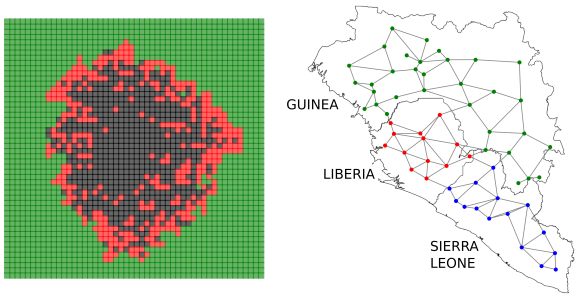

Fig. 1: Graph-coupled HMMs model spatial spreading phenomena such as forest fires (left) and the 2014 Ebola outbreak (right). We derive an online sequential estimation algorithm appropriate for GHMMs and demonstrate its benefits on simulations of forest fires and disease epidemics.

The online aspect creates further challenges, as any candidate method must provide a belief at each time step in a reasonable amount of computation time. Our method is most appropriate for GHMMs with a property common in large-scale spatial processes called "Anonymous Influence." Simply stated, a GHMM has Anonymous Influence if the transition distribution of HMMs relies on the *number* of neighbor HMMs in particular states and not the *identity* of these neighbors.

We develop a message-passing algorithm using variational inference (VI), similar in spirit to belief propagation (BP) methods, that is tractable for GHMMs with considerably large state and observation spaces. Prior work has presented many variations and improvements of standard VI and BP methods. However, these methods incur additional computational complexity or require additional structure and as a result are not appropriate for the models we consider. We review relevant literature in Section II.

The main contributions of this work are: (1) we approximate the evidence lower bound (ELBO) and prove the approximation is a lower bound to the ELBO; (2) we produce a tractable online state estimation algorithm that addresses the challenges in performing inference for GHMMs with arbitrary graph structure and size; and (3) we show our approach requires significantly less computation time to achieve comparable or better accuracy than benchmark methods on a forest fire model with 10^{298} states, and on an Ebola outbreak model derived from data with 10^{18} states.

The remainder of this paper is organized as follows. Section II reviews prior work. Section III describes the GHMM framework for time-invariant graphs and the Anonymous Influence property. Section IV formulates the online

sequential estimation problem and Section V outlines mean-field variational inference. In Section VI, we describe our novel online estimation algorithm for GHMMs. We present numerical results validating our approach in Section VII and provide concluding remarks in Section VIII.

II. Related Work

Exact inference for arbitrary Bayesian networks is an NP-hard problem [5] and we review methods that perform approximate inference. We consider GHMM models in which the equivalent graphical model representation contains many cycles due to the graph edge set. Therefore, methods that rely on a tree structure (e.g., belief propagation) or assume few cycles cannot be directly applied. Furthermore, we are interested in producing a distribution over states as opposed to a maximum-likelihood estimate (e.g., Viterbi algorithm).

Sampling Methods. Particle filters address some issues of exact online inference [6]. However, the number of particles required for a given accuracy increases with the state dimension [7] which makes these filters generally intractable for GHMMs. Particle filters [8], [9] and Kalman filters [10] for high-dimensional state spaces are appropriate for continuous dynamical system models and not the discrete state space models we consider. Other methods [2], [11]–[13] have been applied to relatively large or intractable models but do not scale to the model sizes we consider.

Variational Methods. Variational optimization methods have been applied to large discrete models for inference [14]–[19]. Notably, semi-implicit VI [19] optimizes bounds of the evidence lower bound (ELBO), but these bounds are not suitable for our approach and thus we develop our own approximation. While some methods [20] perform approximate inference for large datasets, it is unclear how to best adapt them for online inference as online applications have been limited to relatively small models [21]. Stochastic gradient methods typically require a differentiable distribution whereas we estimate arbitrary discrete distributions. Variational methods for high-dimensional models require a dynamical system model [22]–[24]. Other methods [25], [26] are based on exploiting distribution structure which we do not require for our approach.

Belief Propagation. Belief propagation (BP) methods can be derived using variational inference with energy approximations (e.g., Bethe or Kikuchi). Loopy belief propagation (LBP) has been shown to be effective in some discrete loopy graphical models [27] and we use LBP as a benchmark method. Generalized belief propagation (GBP) improves upon LBP [28] but incurs additional (worst-case exponential) complexity and is non-trivial to apply generally. We emphasize that our approach does not use energy approximations and therefore we do not follow the standard arguments or derivations of BP methods.

Our approach is algorithmically relatively simple compared to some prior work. Many other methods incur a level of complexity that is not suitable for online inference in GHMMs with large state spaces.

III. Graph-coupled HMMs

We recap the graph-coupled HMM (GHMM) framework for time-invariant graphs. Let $G = (V, E)$ be a (directed or undirected) graph with vertex set $V = \{1, \dots, n\}$ containing n vertices and edge set $E \subseteq V \times V$ [2]. Each vertex $i \in V$ corresponds to a standard HMM with latent state $x_i^t \in \mathcal{X}_i$ and observation $y_i^t \in \mathcal{Y}_i$ at time t. In a GHMM, the transition probabilities of HMM i are influenced by its neighbors. The *neighbor set* compactly describes the set of HMMs that influence a given HMM,

$$N(i) = \{j \mid (j, i) \in E\}.$$

Subscripts are used to describe the latent states or observations of a subset of HMMs, for example x_i^t for HMM i and $x_{N(i)}^t = \{x_j^t \mid j \in N(i)\}$ for the neighbors of HMM i. The latent state transition distribution for each HMM is,

$$p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}). \tag{1}$$

Observations for each HMM are based on the HMM only,

$$p_i(y_i^t \mid x_i^t). \tag{2}$$

Arbitrary measurement models, with $p_i(y_i^t \mid x_i^t, x_{M(i)}^t)$ and $M(i) \subseteq V$, can be used in our framework. However, the method derivation (Section VI) is specific to the transition (1) and measurement models (2) and there is no straightforward description for general models. Due to space constraints, we consider a single case to illustrate our approach and we plan to provide details of other cases in future work.

We omit the subscript for the combination of all HMM states or observations, $x^t = \{x_1^t, \dots, x_n^t\}$ and $y^t = \{y_1^t, \dots, y_n^t\}$. Summing (marginalizing) out all variables from a distribution is denoted by $\sum_{x^t} = \sum_{x_1^t} \cdots \sum_{x_n^t}$ to represent the n consecutive summations. We specify the subset for the marginalization of a limited number of variables, e.g., marginalizing out a neighbor set, $\sum_{x_{N(i)}^t}$. An example GHMM consisting of three HMMs is shown in Fig. 2. For HMM 1 in this example, the neighbor set contains HMM 2, $N(1) = \{2\}$. For HMM 2, $N(2) = \{3\}$ and for HMM 3, $N(3) = \{2\}$.

A. Anonymous Influence

We add structure by considering (1) as based on the number of neighbors in particular states rather than the identity of these neighbors. This property is called "Anonymous Influence" and we summarize the relevant ideas [1], [4].

We assume binary variables and the results extend to arbitrary discrete variables. For a set of binary variables $\mathcal{Z}$, the *count aggregator* (CA) is $\#(z) = \sum_{i=1}^{|\mathcal{Z}|} z_i$ and maps an instantiation of the variables $z \in \mathcal{Z}$ to the set $\{0, \dots, |\mathcal{Z}|\}$. A count aggregator can be thought of as a count of the number of variables that are "enabled," i.e., equal to one. A *count aggregator function* (CAF) is a function that uses a count aggregator to map a set of binary variables to the real numbers, $f : \mathcal{Z} \mapsto \mathbb{R}$. The notation $f(\#(z))$ is used to emphasize that f uses a CA.

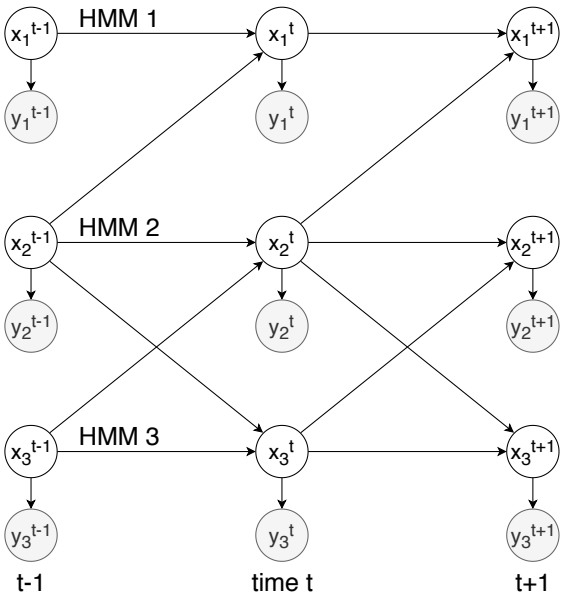

Fig. 2: Example GHMM graphical model for a process containing three HMMs. At each time step t, the objective is to produce a belief over states for each HMM i.

For a GHMM containing only binary variables, the transition distribution (1) requires specifying $2^{|N(i)|+1}$ values. If a CA is used to represent the influence of other HMMs, then the distribution (1) can be represented by a CAF,

$$p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) = p_i(x_i^t \mid x_i^{t-1}, \#(z_i^{t-1})). \tag{3}$$

This representation requires specifying $2(|N(i)|+1)$ values for the case of binary variables, a potentially significant reduction. For example, for $|N(i)| = 10$, the number of required values is reduced from 2048 to 22 which is nearly two orders of magnitude. In the next section, we describe the filtering problem in the context of GHMMs.

IV. Bayesian Sequential Estimation

The objective of sequential estimation for GHMMs is to produce the posterior distribution $p(x^t \mid y^{1:t})$ at each time step where $y^{1:t}$ denotes the history of observations up to time t, $y^{1:t} = (y^1, \ldots, y^t)$. The observation likelihood for the combination of all HMM states in the GHMM is described by the distribution,

$$p(y^t \mid x^t) = \prod_{i=1}^{n} p_i(y_i^t \mid x_i^t), \tag{4}$$

which assumes the observations y_i^t are conditionally independent given the HMM i latent state. The exact filter for producing $p(x^t \mid y^{1:t})$ for a discrete model is derived using Bayes' rule and is a recursive relationship,

$$p(x^t \mid y^{1:t}) \propto p(y^t \mid x^t) \sum_{x^{t-1}} p(x^t \mid x^{t-1}) p(x^{t-1} \mid y^{1:t-1}),$$

which is initialized by a prior at the initial time step, $p(x^0)$. The recursive Bayesian filter (RBF) [29] for the GHMMs considered in this work is thus,

$$p(x^t \mid y^{1:t}) \propto \Big(\prod_{i=1}^{n} p_i(y_i^t \mid x_i^t) \Big) \times \Big(\sum_{x^{t-1}} p(x^{t-1} \mid y^{1:t-1}) \prod_{i=1}^{n} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \Big). \tag{5}$$

We note here that the above expression does not simplify to a tractable form as we allow for arbitrary graph structure. In particular, we allow for models where a path exists between any two vertices. In the context of probabilistic graphical models, the model is considered to have many cycles (or loops). Variational inference provides a framework to approximate an intractable posterior distribution and we present the main ideas in the following section.

V. Mean-Field Variational Inference

Variational inference (VI) methods formulate an optimization problem to approximate an intractable posterior distribution [30]. The RBF for GHMMs (5) requires $(\prod_{i=1}^{n} |\mathcal{X}_i|) - 1$ values to specify $p(x^{t-1} \mid y^{t-1})$ and is intractable to compute for even a single time step despite the graph structure. Therefore, VI introduces a family of distributions $q(x^t) \in \mathcal{Q}$ and approximates the posterior $p(x^t \mid y^{1:t})$ by maximizing the evidence lower bound (ELBO),

$$\text{ELBO} = \mathbb{E}_{q(x^t)} \left[\log p(x^t, y^t \mid y^{1:t-1}) - \log q(x^t) \right], \tag{6}$$

where $\mathbb{E}_{q(x^t)}$ indicates the expectation is taken with respect to (w.r.t.) the distribution $q(x^t)$. We leverage the mean-field approximation [30] where the approximating distribution is factored, $q(x^t) = \prod_{i=1}^{n} q_i(x_i^t)$, and a discrete distribution (or variational factor) is associated with each HMM in the GHMM. This approximation reduces the representation size of the posterior and leads to,

$$\text{ELBO} = \sum_{x^t} \Big(\prod_{i=1}^{n} q_i(x_i^t) \Big) \log p(x^t, y^t \mid y^{1:t-1}) - \sum_{i=1}^{n} \sum_{x_i^t} q_i(x_i^t) \log q_i(x_i^t),$$

after substitution of $q(x^t)$ and algebraic simplification. A common approach to the above optimization problem finds a local optimum by iteratively optimizing each factor $q_i(x_i^t)$ while holding others fixed. First, isolating terms involving index i yields,

$$\text{ELBO} = \sum_{x_i^t} q_i(x_i^t) \mathbb{E}_{-i} \left[\log p(x^t, y^t \mid y^{1:t-1}) \right] - \sum_{x_i^t} q_i(x_i^t) \log q_i(x_i^t) + \text{other terms}, \tag{7}$$

where $\mathbb{E}_{-i}$ refers to the expectation taken w.r.t. the distribution $q(x^t)$ excluding factor $q_i(x_i^t)$, i.e., $\prod_{j=1, j\neq i}^{n} q_j(x_j^t)$.

Maximizing the ELBO w.r.t a single factor $q_i(x_i^t)$ yields the factor objective function $\mathcal{L}_i$,

$$\mathcal{L}_i = -D_{KL}(q_i(x_i^t) \,||\, \exp \mathbb{E}_{-i} \left[\log p(x^t, y^t \mid y^{1:t-1})\right]), \tag{8}$$

where D_{KL} is the Kullback-Leibler (KL) divergence and the "other terms" in (7) are constant w.r.t. $q_i(x_i^t)$ and thus do not influence the solution. Since the KL divergence equals zero when the argument distributions are identical, maximizing $\mathcal{L}_i$ leads to the update equation,

$$q_i(x_i^t) \propto \exp \mathbb{E}_{-i} \left[\log p(x^t, y^t \mid y^{1:t-1})\right]. \tag{9}$$

This approach is known as coordinate ascent VI and is closely related to other inference approaches, such as Gibbs sampling and message-passing algorithms [30]. We emphasize here that allowing arbitrary graph structure in a GHMM typically prevents simplification of the previous expression through structure in $\log p(x^t, y^t \mid y^{1:t-1})$. In the next section, we present a modification of this approach for GHMMs with arbitrary graph structure.

VI. Mean-Field Variational Inference for GHMMs with Anonymous Influence

A. Approximating the ELBO

For GHMMs with the mean-field assumption, the coordinate ascent update (9) for a single time step requires computing the joint probability,

$$\begin{aligned} &p(x^t, y^t \mid y^{1:t-1}) \\ &\propto p(y^t \mid x^t) \sum_{x^{t-1}} p(x^t \mid x^{t-1}) p(x^{t-1} \mid y^{1:t-1}) \\ &\propto \big(\prod_{i=1}^n p_i(y_i^t \mid x_i^t)\big) \sum_{x^{t-1}} \prod_{i=1}^n p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) r_i(x_i^{t-1}), \end{aligned} \tag{10}$$

where $r(x^{t-1}) = \prod_{i=1}^n r_i(x_i^{t-1}) \approx p(x^{t-1} \mid y^{t-1})$ is the approximate factored prior distribution. However, computing the above joint probability is typically intractable due to the required marginalization of all n HMMs.

Instead, a tractable computation of $\mathbb{E}_{-i}\left[p(y^t, x^t \mid y^{1:t-1})\right]$ is possible using a message-passing scheme. We discuss necessary approximations to the ELBO (6) and describe the scheme in the next section. We assume the joint probability satisfies a lower bound, $p(y^t, x^t \mid y^{1:t-1}) \geq \epsilon$ for $0 < \epsilon < 1$. Given a bound ϵ, an under-approximation to the logarithm function over the interval $[\epsilon, 1]$ is the line,

$$g(\theta) = \frac{\log \epsilon}{1 - \epsilon}(1 - \theta), \tag{11}$$

and $g(\theta) \leq \log \theta$ for $\theta \in [\epsilon, 1]$. Using (11) to approximate $\log p(x^t, y^t \mid y^{1:t-1})$ in (6) results in a surrogate ELBO,

$$\overline{\text{ELBO}} = \mathbb{E}_{q(x^t)}\left[g\left(p(x^t, y^t \mid y^{1:t-1})\right) - \log q(x^t)\right]. \tag{12}$$

Theorem 1. *The surrogate ELBO (12) is a lower bound to the original ELBO (6) if the joint probability is lower bounded,* $\epsilon \leq p(x^t, y^t \mid y^{1:t-1}) \leq 1$ *for any* $0 < \epsilon < 1$.

Proof. The difference between the surrogate ELBO (12) and the ELBO (6) is,

$$\begin{aligned} &\text{ELBO} - \overline{\text{ELBO}} \\ &= \mathbb{E}_{q(x^t)}\left[\log p(x^t, y^t \mid y^{1:t-1}) - \log q(x^t)\right] - \\ &\qquad \mathbb{E}_{q(x^t)}\left[g\left(p(x^t, y^t \mid y^{1:t-1})\right) - \log q(x^t)\right] \\ &= \mathbb{E}_{q(x^t)}\left[\log p(x^t, y^t \mid y^{1:t-1})\right] - \\ &\qquad \mathbb{E}_{q(x^t)}\left[g\left(p(x^t, y^t \mid y^{1:t-1})\right)\right] \geq 0 \\ &\Rightarrow \text{ELBO} \geq \overline{\text{ELBO}}\ \forall x^t, y^t. \end{aligned}$$

The expectation operator is linear and thus preserves the lower bound relationship of the approximation (11) to the logarithm function. The lower bound is valid for any combination of GHMM states x^t and observations y^t as the joint probability is bounded below by ϵ. □

Maximizing the surrogate ELBO (12) over the factors $q_i(x_i^t)$ indirectly maximizes the ELBO (7) via the lower bound relationship. Following the same derivation in Section V with the surrogate ELBO, the factor objective (8) is,

$$\hat{\mathcal{L}}_i = -D_{KL}(q_i(x_i^t) \,||\, \exp \mathbb{E}_{-i}\left[g\left(p(x^t, y^t \mid y^{1:t-1})\right)\right]).$$

The coordinate update (9) changes to,

$$\begin{aligned} q_i(x_i^t) &\propto \exp \mathbb{E}_{-i}\left[g\left(p(x^t, y^t \mid y^{1:t-1})\right)\right] \\ &\propto \exp g\left(\mathbb{E}_{-i}\left[p(x^t, y^t \mid y^{1:t-1})\right]\right), \end{aligned} \tag{13}$$

and the factors are now a function of the expectation of the joint probability, as desired, due to the linear approximation.

By inspection of (10), imposing a lower bound on the joint probability precludes combinations of all n HMM states and observations that have zero probability of occurring. It is rare for (10) to be exactly zero except for cases where some (or all) of the HMM states are known exactly without uncertainty or observations have no uncertainty, which is impractical in real applications. In practice, we round estimates of (10) that are lower than ϵ up to ϵ; this has the effect of introducing a small amount of noise to the posterior factors. For large GHMM models, probabilities naturally tend to zero, e.g., the observation distribution (4), since the product of probabilities less than one will approach zero. This approximation can therefore be seen as preventing the filter from rounding state estimates to zero and our numerical results suggest that this approach is effective; see Section VII. In practice, ϵ is a tuning parameter and is typically set to a small value to avoid adding excessive noise to the estimated distributions.

B. Message-passing Scheme

We now build a tractable message-passing scheme to estimate the quantity $\mathbb{E}_{-i}\left[p(x^t, y^t \mid y^{1:t-1})\right]$ required in (13) for each coordinate update. Substituting (10) leads to,

$$\begin{aligned} &\mathbb{E}_{-i}\left[p(x^t, y^t \mid y^{1:t-1})\right] \propto \\ &\sum_{\{x_j^t \mid j \in V, j \neq i\}} \left(\prod_{\substack{j=1 \\ j \neq i}}^n q_j(x_j^t)\right) \left(\prod_{i=1}^n p_i(y_i^t \mid x_i^t)\right) \times \\ &\qquad \left(\sum_{x^{t-1}} \prod_{i=1}^n p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) r_i(x_i^{t-1})\right). \end{aligned} \tag{14}$$

Algorithm 1 Message-passing Scheme

Given observations $\{y_1^t, \ldots, y_n^t\}$
for each HMM i **do**
  initialize message and posterior factor $q_k^0(x_i^t)$
for iteration $k = 1, \ldots$ **do**
  for each HMM i **do**
    Receive messages from neighbors $j \in N(i)$
    Calculate k^{th} estimate $E_i^k(x_i^t)$ using messages
    Update factor $q_i^k(x_i^t)$ with $E_i^k(x_i^t)$
    Create message for next iteration $k+1$

The message-passing scheme works as follows. Each HMM in the GHMM maintains an estimate of its posterior factor, $q_i^k(x_i^t)$, and an estimate of (14), $E_i^k(x_i^t)$; the superscript k on these quantities, and other quantities below, refers to the k^{th} estimate. For each iteration k of the scheme, each HMM i receives messages from the neighbor HMMs $j \in N(i)$ and generates estimate $E_i^k(x_i^t)$. This estimate then updates the posterior factor using (13). Lastly, an updated message is calculated for the next iteration $k+1$. Algorithm 1 summarizes this process. For clarity, we present the necessary relationships for the scheme below and provide the derivation in the Appendix.

First, the initial *message* $m_i^0(x_i^{t-1})$ for each HMM is its prior, $m_i^0(x_i^{t-1}) = r_i(x_i^{t-1})$. To compute joint probability estimates and updated messages, a *candidate message* is computed by each HMM,

$$c_i^k(x_i^{t-1}, x_i^t) = p_i(y_i^t \mid x_i^t) \sum_{x_{N(i)}^{t-1}} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \prod_{j \in N(i)} m_j^{k-1}(x_j^{t-1}). \quad (15)$$

Using this candidate message, the estimate of the joint probability computed by HMM i is,

$$\begin{aligned} E_i^k(x_i^t) &\propto \sum_{x_i^{t-1}} r_i(x_i^{t-1}) c_i^k(x_i^{t-1}, x_i^t) \\ &\approx \mathbb{E}_{-i}\left[p(x^t, y^t \mid y^{1:t-1})\right], \end{aligned} \quad (16)$$

and $q_i^k(x_i^t)$ is updated by using the above estimate with (13). Lastly, the updated message that is shared by each HMM at the next iteration is,

$$m_i^k(x_i^{t-1}) \propto r_i(x_i^{t-1}) \sum_{x_i^t} q_i^k(x_i^t) c_i^k(x_i^{t-1}, x_i^t). \quad (17)$$

Using (11) moves the expectation into the argument of (13) which allows the posterior factors to be used for marginalization of (14). This property is key for creating a tractable message-passing method. Otherwise, the quantity $\mathbb{E}_{-i}\left[\log p(x^t, y^t \mid y^{1:t-1})\right]$ is not tractable to compute as we do not allow for simplification based on the graph structure of the GHMM, on the properties of distributions for the posterior factors, or other properties. In addition, the estimates (16) are normalized to estimate the normalization constant of the joint probability (10), since the linear approximation does not allow this constant to be factored out.

Algorithm 2 Relaxed Anonymous Variational Inference (RAVI) for time step t

1: **Input:** prior factors $r_i(x_i^{t-1})$, graph G, state transition model $p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1})$, observations y_i^t and sensor model $p_i(y_i^t \mid x_i^t)$
2: **Output:** posterior factors $q_i(x_i^t)$
3: **Parameters:** iteration limit $K_{\max}$, lower bound ϵ, convergence criteria
4: **for** each HMM i **do**
5: initialize message $m_i^0(x_i^{t-1}) = r_i(x_i^{t-1})$
6: initialize factor $q_i^0(x_i^t)$
7: **for** iteration $k = 1, \ldots, K_{\max}$ **do**
8: **for** each HMM i **do**
9: Receive messages $\{m_j^{k-1}(x_j^{t-1}) \mid j \in N(i)\}$
10: Compute $c_i^k(x_i^{t-1}, x_i^t)$ with (15) or (18)
11: Estimate $E_i^k(x_i^t)$ using (16)
12: Update $q_i^k(x_i^t)$ by (13)
13: Compute $m_i^k(x_i^{t-1})$ with (17)
14: **if** factors $q_i^k(x_i^t)$ converge **then** terminate early
15: **return** posterior factors $q_i(x_i^t) = q_i^k(x_i^t)$

C. Simplifying with Anonymous Influence

Computing the candidate message (15) may be intractable as marginalizing out $x_{N(i)}^{t-1}$ requires considering $\prod_{j \in N(i)} |\mathcal{X}_j|$ values. If a HMM in the GHMM has many neighbors (large $|N(i)|$) or if the neighbors have large state spaces $|\mathcal{X}_j|$ then the computational cost may be significant. Therefore, we now exploit Anonymous Influence to address this potential issue. If the transition distribution (1) of a HMM relies on a count aggregator (CA) (as shown in (3)), then it is useful to create a count-aggregator function (CAF) $\tilde{m}_i^{k-1}(\#(z_i^{t-1}))$ to represent the received neighbor messages. Using this CAF leads to the modified candidate message,

$$c_i^k(x_i^{t-1}, x_i^t) = p_i(y_i^t \mid x_i^t) \sum_{z_i^{t-1}} p_i(x_i^t \mid x_i^{t-1}, \#(z_i^{t-1})) \tilde{m}_i^{k-1}(\#(z_i^{t-1})). \quad (18)$$

The marginalization for (18) is now with respect to $z_i^{t-1} \in [0, \ldots, |N(i)|]$ which has lower computational cost.

D. Algorithm Description

Algorithm 2, Relaxed Anonymous Variational Inference (RAVI), summarizes our filtering algorithm for producing posterior factors for a single time step. The factors $q_i(x_i^t)$ produced are then used as the priors $r_i(x_i^t)$ in the algorithm for the next time step. The main component is the message-passing scheme which is relatively straightforward to implement and can be used independently of the Anonymous Influence property (line 10). The posterior factors are initialized to any valid discrete distribution (line 6). The algorithm runs for a fixed number of iterations $K_{\max}$ unless the posterior factors converge (line 14). We provide a brief analysis of the computational complexity. Without Anonymous Influence, the complexity for a single time step

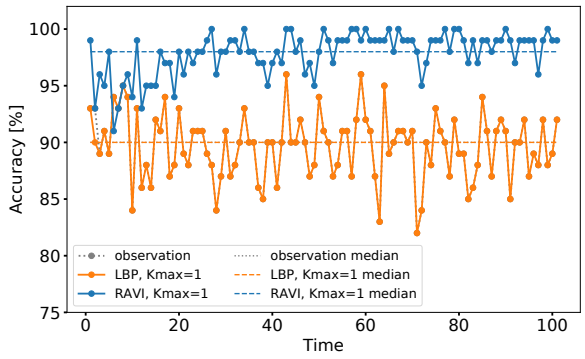

Fig. 3: Example filter results for a single model simulation. For each simulation, the simulation accuracy for a filter is the median accuracy over the entire time series. Here, LBP is the same as taking the observation as the estimate whereas RAVI improves the median accuracy by 8%.

is $\mathcal{O}(K_{\max} \max_{i \in V} \ |\mathcal{X}_i|^2 \prod_{j \in N(i)} |\mathcal{X}_j|)$ for an iteration limit $K_{\max}$. Exploiting the Anonymous Influence property reduces the complexity to,

$$\mathcal{O}(K_{\max} \max_{i \in V} \ \{|N(i)|2^{|N(i)|}, |\mathcal{X}_i|^2(|N(i)|+1) \mid i \in V\}).$$

This analysis shows the complexity is determined by HMMs with many neighbors (large $|N(i)|$), neighbors with large state spaces (large $|\mathcal{X}_j|$), and a large state space $|\mathcal{X}_i|$. Since typically $|N(i)|+1 \ll \prod_{j \in N(i)} |\mathcal{X}_j|$, Anonymous Influence reduces the computational cost of RAVI. In contrast, the RBF (5) requires $\mathcal{O}(\prod_{i \in V} |\mathcal{X}_i|^2)$ operations which is intractable for all but trivial model sizes.

VII. Experiments

Models. We model a forest as a square lattice of trees initialized with fire at the center [3]. Each tree state x_i^t is one of three values, $\mathcal{X}_i = \{H, F, B\} = \{\text{healthy}, \text{on fire}, \text{burnt}\}$, with dynamics given in Table III and $f_i^t = \sum_{j \in N(i)} \mathbf{1}_F(x_i^t)$ is the number of neighboring trees on fire. A west-to-east wind is also modeled by linearly varying the parameter α_i from 0.1 (left edge) to 0.4 (right edge) across the lattice. A constant value of $\beta = 0.9$ was used. We use forest sizes of 3×3, 10×10, and 25×25 that have 10^4, 10^{47}, and 10^{298} total states at each time step, respectively. For each tree, the true state is observed with 90% probability and the remaining states are observed with 5% probability.

We use a simplified model based on [3] for the 2014 Ebola outbreak in West Africa. Each vertex corresponds to a community and edges indicate transportation routes. The community state is one of two values, $\mathcal{X}_i = \{S, E\} = \{\text{healthy}, \text{infected}\}$, with dynamics given in Table II where $e_i^t = \sum_{j \in N(i)} \mathbf{1}_E(x_i^t)$ and $\eta = 0.08$. The model contains 62 communities, with 10^{18} total states at each time step, and the neighbor set size $|N(i)|$ ranges from one to seven. For each community, the true state is observed with 85% probability.

Comparison Methods. We implement the recursive Bayesian filter (5) (RBF) only for a small forest fire model as the method quickly becomes intractable. We also adopt loopy belief propagation (LBP) for the online sequential estimation problem by limiting the graphical model size to a maximum of $H = 3$ hidden and observed layers. If adding a new layer will exceed the model size limit, the model is reinitialized as only the new layer with a prior belief equal to the first layer of the previous model. Two message-passing iteration limits were tested, $K_{\max} = 1$ and $K_{\max} = 3$, and the approximation (11) used $\epsilon = 10^{-10}$. Both RAVI and LBP terminated early if less than 1% of HMMs stop changing their maximum likelihood belief. Lastly, we compare against taking each observation as the estimate for each time step. This may produce inconsistent estimates, e.g., a tree on fire changing to healthy within a time step in the forest fire model. All filters were initialized with the ground truth.

At each time step, the belief produced by the filter is converted to a maximum likelihood estimate for each HMM and compared to the true state. Each filter thus produces a time history of accuracies for a single simulation run; see Fig. 3. We compute the median accuracy over the time series and denote this quantity the *simulation accuracy* for a filter. We run 100 total simulations and report the minimum, median, and maximum simulation accuracy.

Results. Tables I and II present the results. The RBF and

TABLE I: Results for three forest sizes. Data are the median simulation accuracy for 100 simulations, and the superscript and subscript indicate the maximum and minimum simulation accuracy, respectively. A dash indicates where a method is computationally intractable. LBP performs well with enough iterations but does not scale. RAVI is comparably accurate and scales to larger model sizes.

Method		Forest Size		
		3×3	10×10	25×25
Observation		$88.9^{+11.1}_{-00.0}\%$	$90.0^{+1.0}_{-1.0}\%$	$90.0^{+0.4}_{-0.2}\%$
RBF		$100^{+00.0}_{-11.1}\%$	—	—
LBP	$K_{\max} = 1$	$88.9^{+11.1}_{-00.0}\%$	$90.0^{+1.0}_{-1.0}\%$	—
	$K_{\max} = 10$	$100^{+00.0}_{-11.1}\%$	$99.0^{+1.0}_{-1.0}\%$	—
RAVI	$K_{\max} = 1$	$100^{+00.0}_{-11.1}\%$	$98.0^{+1.0}_{-1.0}\%$	$98.1^{+0.5}_{-0.6}\%$
	$K_{\max} = 10$	$100^{+00.0}_{-11.1}\%$	$98.0^{+0.0}_{-2.0}\%$	$97.8^{+0.8}_{-1.0}\%$

TABLE II: Results for the disease epidemic model. Accuracy values are the median simulation accuracy over 100 simulations, with the superscript and subscript indicating the maximum and minimum simulation accuracy, respectively. Time values are the average time required to produce a belief for a single time step of a simulation. RAVI is comparably accurate to LBP as well as two orders of magnitude faster.

Method		Accuracy (%)	Time for Estimate (seconds)
Observation		$85.5^{+1.6}_{-1.6}$	
LBP	$K_{\max} = 1$	$85.5^{+1.6}_{-1.6}$	3.57
	$K_{\max} = 10$	$98.4^{+1.6}_{-0.0}$	12.82
RAVI	$K_{\max} = 1$	$96.8^{+0.0}_{-1.6}$	0.03
	$K_{\max} = 10$	$95.2^{+1.6}_{-1.6}$	0.10

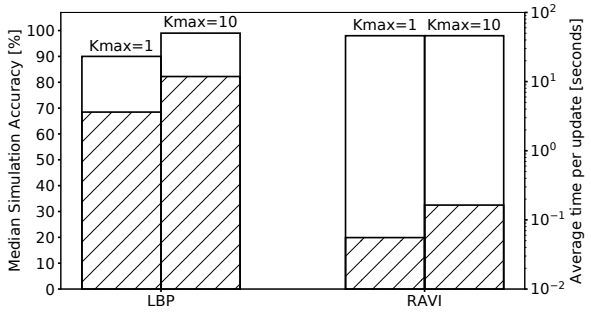

Fig. 4: Comparison of time required to update belief (hatched bar) and the median accuracy (unshaded bar) for a single time step for LBP and RAVI on the 10×10 forest size. Both methods have comparable accuracy but RAVI is approximately two orders of magnitude faster.

TABLE III: Tree dynamics for wildfire model. Blank entries are zero.

		x_i^{t+1}		
		H	F	B
x_i^t	H	$(1-\alpha_i)^{f_i^t}$	$1-(1-\alpha_i)^{f_i^t}$	
	F		β	$1-\beta$
	B			1

LBP are not suitable methods as they are only tractable for relatively small model sizes. In contrast, RAVI performs well and scales to the largest forest fire model size. RAVI is also more efficient than LBP as Fig. 4 and Table II show that LBP takes roughly 100 times longer produce a belief per time step. In addition, RAVI has low variance in the accuracy which is desirable for a filtering method. Lastly, for the two different iteration limits, the results of RAVI did not significantly change. Coordinate ascent methods are known to reach local optima and RAVI quickly finds a solution which does not change significantly with more iterations.

VIII. Conclusions

We derived a novel inference algorithm appropriate for online sequential state estimation of large-scale GHMMs with the property of Anonymous Influence. Compared to benchmark methods, our approach is accurate and can be used in time critical applications. Future work will focus on improving the approximations required for our method, such as using implicit distributions [19] and adding structure back to the approximate factored distribution [16], [17], [25]. The effectiveness of LBP is likely due to using information across time steps when performing inference. A key improvement of our method will be incorporating this aspect in our approach. We also plan to further characterize the theoretical properties of our approach, such as the convergence rate and quality of the local optima, as well as provide details regarding other forms of the state transition and observation models.

Appendix

The message-passing scheme produces iterative estimates of (14) for each HMM i in the GHMM. We derive the scheme

TABLE IV: Community dynamics for epidemic model.

		x_i^{t+1}	
		S	E
x_i^t	S	$(1-\eta)^{e_i^t}$	$1-(1-\eta)^{e_i^t}$
	E	0	1

and the use of messages to simplify the computation by describing the first two approximations for a given HMM i. The first approximation E_i^1 considers information from the neighbor HMMs $j \in N(i)$,

$$E_i^1(x_i^t) \propto \left[p_i(y_i^t \mid x_i^t)\right] \sum_{x_i^{t-1}} r_i(x_i^{t-1}) \left[\sum_{x_{N(i)}^{t-1}} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \prod_{j \in N(i)} r_j(x_j^{t-1})\right]. \tag{19}$$

The second approximation E_i^2 considers information from the neighbors $N(i)$ as well as the neighbors of neighbors $\bigcup_{j \in N(i)} N(j)$,

$$E_i^2(x_i^t) \propto p_i(y_i^t \mid x_i^t) \sum_{x_i^{t-1}} r_i(x_i^{t-1}) \sum_{x_{N(i)}^{t-1}} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \prod_{j \in N(i)} r_j(x_j^{t-1}) \sum_{x_j^t} q_j^1(x_j^t) \left[p_j(y_j^t \mid x_j^t) \sum_{x_{N(j)}^{t-1}} p_j(x_j^t \mid x_j^{t-1}, x_{N(j)}^{t-1}) \prod_{l \in N(j)} r_l(x_l^{t-1})\right]. \tag{20}$$

There is a common structure in the first (19) and second (20) approximations, as indicated by the large brackets in both expressions. Define the following as a *candidate message*,

$$c_i^1(x_i^{t-1}, x_i^t) = p_i(y_i^t \mid x_i^t) \sum_{x_{N(i)}^{t-1}} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \prod_{j \in N(i)} r_j(x_j^{t-1}). \tag{21}$$

The first approximation (19) then simplifies to,

$$E_i^1(x_i^t) \propto \sum_{x_i^{t-1}} r_i(x_i^{t-1}) c_i^1(x_i^{t-1}, x_i^t). \tag{22}$$

In addition, the second approximation (20) simplifies to,

$$E_i^2(x_i^t) \propto p_i(y_i^t \mid x_i^t) \sum_{x_i^{t-1}} r_i(x_i^{t-1}) \sum_{x_{N(i)}^{t-1}} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \prod_{j \in N(i)} \left[r_j(x_j^{t-1}) \sum_{x_j^t} q_j^1(x_j^t) c_j^1(x_j^{t-1}, x_j^t)\right], \tag{23}$$

by using the candidate messages of the neighbor HMMs $j \in N(i)$ computed during the first iteration of the scheme. Notably, the simplified second approximation now only requires information from the neighbors $N(i)$, as indicated by

brackets. If the neighbors produce a *message* to share,

$$m_j^1(x_j^{t-1}) \propto r_j(x_j^{t-1}) \sum_{x_j^t} q_j^1(x_j^t) c_j^1(x_j^{t-1}, x_j^t),$$

then the second approximation (23) further simplifies,

$$E_i^2(x_i^t) \propto \left[p_i(y_i^t \mid x_i^t) \right] \sum_{x_i^{t-1}} r_i(x_i^{t-1}) \left[\sum_{x_{N(i)}^{t-1}} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \prod_{j \in N(i)} m_j^1(x_j^{t-1}) \right]. \tag{24}$$

Finally, (24) shares a common structure with the first approximation (19), as indicated by brackets. If the candidate message (21) is modified for the second approximation to use the neighbors' messages instead of their priors,

$$c_i^2(x_i^{t-1}, x_i^t) = p_i(y_i^t \mid x_i^t) \sum_{x_{N(i)}^{t-1}} p_i(x_i^t \mid x_i^{t-1}, x_{N(i)}^{t-1}) \prod_{j \in N(i)} m_j^1(x_j^{t-1}), \tag{25}$$

then the second approximation E_i^2 simplifies again to,

$$E_i^2(x_i^t) \propto \sum_{x_i^{t-1}} r_i(x_i^{t-1}) c_i^2(x_i^{t-1}, x_i^t),$$

which mirrors the form of the simplified first approximation (22). The slightly different forms of the candidate message, (21) and (25), are reconciled by defining the initial message for all HMMs as the prior, $m_i^0(x_i^{t-1}) = r_i(x_i^{t-1})$. Subsequent approximations E_i^k, $k \geq 3$ continue to incrementally add influence from additional HMMs in the GHMM to improve the estimate of (14).

Finally, the general form of the candidate message is (15), the k^{th} approximation of (14) is (16), and the message for the next iteration is (17). The messages use an estimate of the posterior factor and the k^{th} estimate is determined by (13).

REFERENCES

[1] V. Raghavan, G. ver Steeg, A. Galstyan, and A. G. Tartakovsky, "Coupled hidden Markov models for user activity in social networks," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2013, pp. 1–6.

[2] W. Dong, A. S. Pentland, and K. A. Heller, "Graph-coupled HMMs for modeling the spread of infection," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2012, pp. 227–236.

[3] R. N. Haksar and M. Schwager, "Controlling large, graph-based MDPs with global control capacity constraints: An approximate LP solution," in *57th IEEE Conference on Decision and Control (CDC)*, Dec 2018, pp. 35–42.

[4] P. Robbel, F. A. Oliehoek, and M. J. Kochenderfer, "Exploiting anonymity in approximate linear programming: scaling to large multiagent MDPs," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 2537–2573.

[5] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artificial Intelligence*, vol. 42, no. 2-3, pp. 393–405, Mar. 1990.

[6] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, July 2000.

[7] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proceedings F - Radar and Signal Processing*, vol. 140, no. 2, pp. 107–113, 1993.

[8] N. Vaswani, A. Yezzi, Y. Rathi, and A. Tannenbaum, "Particle filters for infinite (or large) dimensional state spaces- part 1," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3, May 2006, pp. III–III.

[9] A. Beskos, D. Crisan, A. Jasra, K. Kamatani, and Y. Zhou, "A stable particle filter for a class of high-dimensional state-space models," *Advances in Applied Probability*, vol. 49, no. 1, p. 24–48, 2017.

[10] T. Seo, T. T. Tchrakian, S. Zhuk, and A. M. Bayen, "Filter comparison for estimation on discretized PDEs modeling traffic: Ensemble Kalman filter and minimax filter," in *55th IEEE Conference on Decision and Control (CDC)*, Dec 2016, pp. 3979–3984.

[11] A. Jasra, S. S. Singh, J. S. Martin, and E. McCoy, "Filtering via approximate Bayesian computation," *Statistics and Computing*, vol. 22, no. 6, pp. 1223–1237, Nov 2012.

[12] K. Fan, C. Li, and K. Heller, "A unifying variational inference framework for hierarchical graph-coupled HMM with an application to influenza infection," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 3828–3834.

[13] L. Bronstein and H. Koeppl, "Scalable inference using PMCMC and parallel tempering for high-throughput measurements of biomolecular reaction networks," in *55th IEEE Conference on Decision and Control (CDC)*, Dec 2016, pp. 770–775.

[14] A. Saeedi, T. D. Kulkarni, V. K. Mansinghka, and S. J. Gershman, "Variational particle approximations," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2328–2356, Jan. 2017.

[15] D. M. Blei, M. I. Jordan, and J. W. Paisley, "Variational Bayesian inference with stochastic search," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research. PMLR, 2012, pp. 1367–1374.

[16] M. D. Hoffman and D. M. Blei, "Structured stochastic variational inference," in *Artificial Intelligence and Statistics*, 2015.

[17] T. Salimans, D. Kingma, and M. Welling, "Markov chain Monte Carlo and variational inference: Bridging the gap," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 2015, pp. 1218–1226.

[18] A. Buchholz, F. Wenzel, and S. Mandt, "Quasi-Monte Carlo variational inference," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 668–677.

[19] M. Yin and M. Zhou, "Semi-implicit variational inference," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 5660–5669.

[20] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[21] V. Smidl and A. Quinn, "Variational Bayesian filtering," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5020–5030, 2008.

[22] Y. Lu, S. Khatibisepehr, and B. Huang, "A variational Bayesian approach to identification of switched ARX models," in *53rd IEEE Conference on Decision and Control (CDC)*, Dec 2014, pp. 2542–2547.

[23] K. Fujimoto and Y. Takaki, "On system identification for ARMAX models based on the variational Bayesian method," in *55th IEEE Conference on Decision and Control (CDC)*, Dec 2016, pp. 1217–1222.

[24] B. Ait-El-Fquih and I. Hoteit, "A variational Bayesian multiple particle filtering scheme for large-dimensional systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5409–5422, Oct 2016.

[25] L. K. Saul and M. I. Jordan, "Exploiting tractable substructures in intractable networks," in *Advances in Neural Information Processing Systems 8*. MIT Press, 1996, pp. 486–492.

[26] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, Dec. 2005.

[27] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 467–475.

[28] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, pp. 689–695.

[29] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[30] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.