Reduced Complexity Filtering with Stochastic Dominance Bounds: A Convex Optimization Approach

Vikram Krishnamurthy, Fellow, IEEE Cristian R. Rojas, Member, IEEE

Abstract—This paper uses stochastic dominance principles to construct upper and lower sample path bounds for Hidden Markov Model (HMM) filters. Given a HMM, by using convex optimization methods for nuclear norm minimization with copositive constraints, we construct low rank stochastic matrices \( P \) and \( \bar{P} \) so that the optimal filters using \( P \) and \( \bar{P} \) provably lower and upper bound (with respect to a partially ordered set) the true filtered distribution at each time instant. Since \( P \) and \( \bar{P} \) are low rank (say \( R \)), the computational cost of evaluating the filtering bounds is \( O(RX) \) instead of \( O(X^2) \). A Monte-Carlo importance sampling filter is presented that exploits these upper and lower bounds to estimate the optimal posterior. Finally, using the Dobrushin coefficient, explicit bounds are given on the variational norm between the true posterior and the upper and lower bounds.

I. INTRODUCTION

This paper is motivated by the filtering problem involving estimating a large dimensional finite state Markov chain given noisy observations. With \( k \) denoting discrete time, consider an \( X \)-state discrete time Markov chain \( \{x_k\} \) observed via a noisy process \( \{y_k\} \). Here \( x_k \in \{1, 2, \ldots, X\} \) where \( X \) denotes the dimension of the state space. Let \( P \) denote the \( X \times X \) transition matrix and \( B_{xy} = P(y_k = y|x_k = x) \) denote the observation likelihood probabilities. With \( y_k|x_k \) denoting the sequence of observations from time 1 to \( k \), define the posterior state probability mass function

\[
\pi_k(i) = P(x_k = i|y_1:k), \quad i \in \{1, 2, \ldots, X\}, \quad \pi_k = (\pi_k(1), \ldots, \pi_k(X))^T.
\]

It is well known [1], [2] that the optimal Bayesian filter (Hidden Markov Model filter) for computing the \( X \)-dimensional posterior vector \( \pi_k \) at each time \( k \) is of the form

\[
\pi_{k+1} = T(\pi_k, y_{k+1}; P) \triangleq \frac{B_{y_{k+1}} P \pi_k}{1'B_{y_{k+1}} P \pi_k}.
\]

Here, \( B_y = \text{diag}(B_{y1}, \ldots, B_{yX}) \) is a diagonal \( X \times X \)-dimensional matrix of observation likelihoods and \( 1 \) denotes the \( X \)-dimensional column vector of ones.

Due to the matrix-vector multiplication \( P' \pi_k \) in (2), the computational cost for evaluating the posterior \( \pi_{k+1} \) at each time \( k \) is \( O(X^2) \). This quadratic computational cost \( O(X^2) \) can be excessive for large state space dimension \( X \).

Motivation and Main Results

This paper addresses the question: Can the optimal filter be approximated with reduced complexity filters with provable sample path bounds? We derive reduced-complexity filters with computational cost \( O(RX) \) where \( R \ll X \). There are four main results in this paper.

1. Stochastic Dominance Bounds: Theorem 1 presented in Section II asserts that for any transition matrix \( P \), one can construct two new transition matrices \( \bar{P} \) and \( \bar{P} \), such that \( P \preceq \bar{P} \preceq \bar{P} \). Here \( \preceq \) denotes a copositive ordering defined in Section II. The Bayesian filters using \( P \) and \( \bar{P} \), are guaranteed to sandwich the true posterior distribution \( \pi_k \) at any time \( k \) as

\[
T(\pi_{k-1}, y_k; P) \preceq T(\pi_{k-1}, y_k; \bar{P}) \preceq T(\pi_{k-1}, y_k; \bar{P}) \tag{3}
\]

where \( T(\cdot) \) denotes the filtering recursion (2) and \( \preceq \) denotes monotone likelihood ratio (MLR) stochastic dominance.

2. Construction of low rank transition matrices via nuclear norm minimization: Sec. III uses state-of-the-art convex optimization methods for nuclear norm minimization with copositive constraints, we construct low rank stochastic matrices \( P \) and \( \bar{P} \) so that the optimal filters using \( P \) and \( \bar{P} \) provably lower and upper bound (with respect to a partially ordered set) the true filtered distribution at each time instant. Since \( P \) and \( \bar{P} \) are low rank (say \( R \)), the computational cost of evaluating the filtering bounds is \( O(RX) \) instead of \( O(X^2) \). A Monte-Carlo importance sampling filter is presented that exploits these upper and lower bounds to estimate the optimal posterior. Finally, using the Dobrushin coefficient, explicit bounds are given on the variational norm between the true posterior and the upper and lower bounds.

\[1\] TP2 matrices are defined in Definition 3.
than $O(X^2)$ computational cost. The transition matrices $P$ and $\bar{P}$ are constructed as low rank matrices by minimizing their nuclear norms. Matrices with small nuclear norm exhibit sparseness in the set of eigenvalues or equivalently low rank. The nuclear norm is the sum of the singular values of a matrix. The construction of low rank transition matrices $P$ and $\bar{P}$ is formulated as a convex optimization problem on the cone of copositive matrices. These computations are performed offline without affecting the computational cost of the real time filter.

3 Stochastic Dominance Constrained Monte-Carlo Importance Sampling Filter: In monitoring systems, it is of interest to detect when the underlying Markov chain is close to a target state. Using the reduced complexity filtering bounds outlined above, a monitoring system would want to switch to the full complexity filter when the filtering bounds approach the target state. A natural question is: How can the reduced complexity filtering bounds be exploited to estimate the true posterior? Section IV presents an importance sampling Monte-Carlo method for matrix vector multiplication that is inspired by recent results in stochastic linear solvers. The algorithm uses Gibbs sampling to ensure that the estimated posterior $\bar{\pi}_k$ lies in the partially ordered set $\bar{\pi}_k \leq_r \bar{\pi}_k \leq \bar{\pi}_k$ at each time $k$. Numerical experiments show that this stochastic dominance constrained algorithm yields estimates with substantially reduced mean square errors compared to the unconstrained algorithm – in addition, by construction the estimates are provably sandwiched between $\bar{\pi}_k$ and $\bar{\pi}_k$.

4. Analytical Bounds on Variational Distance: Given the low complexity bounds $\bar{\pi}_k$ and $\bar{\pi}_k$ such that $\bar{\pi}_k \leq \bar{\pi}_k \leq r \bar{\pi}_k$, a natural question is: How tight are the bounds? Theorem 3 presents explicit analytical bounds on the deviation of the true posterior $\pi_k$ (which is expensive to compute) from the lower and upper bounds $\bar{\pi}_k$ and $\bar{\pi}_k$ in terms of the Dobrushin coefficient of the transition matrix. It yields useful analytical bounds that (can be computed without evaluating the posterior $\pi_k$) for quantifying how the stochastic dominance constraints sandwich the true posterior as time evolves.

Related Work

The area of constructing approximate filters for estimating the state of large scale Markov chains has been well studied both in discrete and continuous time. Most works assume that the Markov chain has two-time scale dynamics (e.g. the Markov chain is nearly completely decomposable). This two-time scale feature is then exploited to construct suitable filtering approximations on the slower time scale. In comparison, the framework in the current paper does not assume a two-time scale Markov chain. Indeed, our results are finite sample results that do not rely on asymptotics.

The main tools used in this paper are based on monotone likelihood ratio (MLR) stochastic dominance and associated monotone structural results of the Bayesian filtering update.

A symmetric matrix $M$ is copositive if $\pi' M \pi \geq 0$ for all positive vectors $\pi$. (Thus the set of positive definite matrices is a subset of the set of copositive matrices. In this paper, $\pi$ are probability mass function vectors.) Such results have been developed in the context of stochastic control and Bayesian games in [4], [11], [3] but have so far not been exploited to devise efficient filtering approximations. To the best of our knowledge, constructing upper and lower sample path bounds to the optimal filter in terms of stochastic orders is new – and the copositivity constraints presented in this paper yield a constructive realization of these bounds. Recently, [3] use similar copositive characterizations to derive structural results in stochastic control.

Optimizing the nuclear norm as a surrogate for rank has been studied as a convex optimization problem in several papers, see for example [8]. Inspired by the seminal work of Candès and Tao [12], there has been much recent interest in minimizing nuclear norms in the context of sparse matrix completion problems. Algorithms for testing for copositive matrices and copositive programming have been studied recently in [13], [14], [15].

There has been extensive work in signal processing on posterior Cramér-Rao bounds for nonlinear filtering [15]; see also [16] for a textbook treatment. These yield lower bounds to the achievable variance of the conditional mean estimate of the optimal filter. However, unlike the current paper, such posterior Cramér-Rao bounds do not give constructive algorithms for computing upper and lower bounds for the sample path of the filtered distribution. The sample path bounds proposed in this paper have the attractive feature that they are guaranteed to yield lower and upper bounds to both hard and soft estimates of the optimal filter.

II. STOCHASTIC DOMINANCE OF FILTERS AND COPOSITIVITY CONDITIONS

Theorem 1 below is the main result of this section – it shows that if stochastic matrices $P$ and $\bar{P}$ are constructed such that $P \succeq P \succeq \bar{P}$ (in terms of a copositive ordering), the filtered estimates computed using $P$ and $\bar{P}$ are guaranteed to sandwich the optimal filtered estimate in terms of the monotone likelihood ratio order. This section sets the stage for Section III where the construction of low rank matrices $P$ and $\bar{P}$ is formulated as a convex optimization problem on a copositive cone; and also Section IV where algorithms that exploit this result are presented.

A. Signal Model and Optimal Filter

Consider an $X$-state discrete time Markov chain $\{x_k\}$ on the state space $\{1, 2, \ldots, X\}$. Suppose $x_0 \in \{1, 2, \ldots, X\}$ has a prior distribution $\pi_0$. The $X \times X$-dimensional transition probability matrix $P$ comprises of elements $P_{ij} = \mathbb{P}(x_{k+1} = j | x_k = i)$.

The Markov process $\{x_k\}$ is observed via a noisy process $\{y_k\}$ where at each time $k$, $y_k \in \{1, 2, \ldots, Y\}$ or $y_k \in \mathbb{R}^m$. As is widely assumed in optimal filtering, we make the conditional independence assumption that $y_k$ given $x_k$ is statistically independent of $x_{k-1} : x_{k-1}$. For the case $y_k \in \{1, 2, \ldots, Y\}$, denote the observation likelihood probabilities as $B_{xy} = \mathbb{P}(y_k = y | x_k = x)$. The case $y_k \in \mathbb{R}^m$, $B_{xy}$ is the conditional probability density. (For readability to an
engineering audience, unified notation with respect to the Lebesgue and counting measures is avoided.)

With \( \pi_k(i) \) denoting the posterior defined in (1), the optimal filter is given by (2). Note that the posterior \( \pi_k \) lives in an \( X \) -dimensional unit simplex \( \Pi \) comprising of \( X \)-dimensional probability vectors \( \pi \). That is,
\[
\Pi = \{ \pi \in \mathbb{R}^X : 1^T \pi = 1, \ \pi(i) \geq 0 \}.
\] (5)

Finally, since the state space is \( \{1, 2, \ldots, X\} \), the conditional mean estimate of the state computed using the observations \( y_{0:k} \) is (we avoid using the notation of sigma algebras)
\[
\hat{x}_k \triangleq \mathbb{E}\{x_k|y_{0:k}; P\} = g' \pi_x, \quad \text{where } g = [1, 2, \ldots, X]'.
\] (6)

In some applications, rather than the “soft” state estimate provided by the conditional mean, one is interested in the “hard” valued maximum aposteriori estimate defined as
\[
x_k^{\text{MAP}} \triangleq \arg\max_{i \in \{1, 2, \ldots, X\}} \pi_k(i).
\] (7)

We refer to posterior \( \pi_k \) in (2) and state estimates (6), (7) computed using transition matrix \( P \) as the “optimal filtered estimates” to distinguish them from the lower and upper bound filters.

B. Some Preliminary Definitions

We introduce here some key definitions that will be used in the rest of the paper.

1) Stochastic Dominance: We start with the following standard definitions involving stochastic dominance [17]. Recall that \( \Pi \) is the unit simplex defined in (5).

Definition 1 (Monotone Likelihood Ratio (MLR) Dominance): Let \( \pi_1, \pi_2 \subset \Pi \) be any two probability vectors. Then \( \pi_1 \) is greater than \( \pi_2 \) with respect to the MLR ordering – denoted as \( \pi_1 \succeq \pi_2 \) if
\[
\pi_1(i) \pi_2(j) \leq \pi_2(i) \pi_1(j), \quad i < j, \quad i, j \in \{1, 2, \ldots, X\}.
\] (8)

Similarly \( \pi_1 \succeq \pi_2 \) if \( \pi_1 \) is replaced by a \( \geq \). The MLR stochastic order is useful since it is closed under conditional expectations. That is, \( X \succeq Y \) implies \( \mathbb{E}\{X|F\} \succeq \mathbb{E}\{Y|F\} \) for any two random variables \( X, Y \) and sigma-algebra \( F \) [11], [6], [7], [17].

Definition 2 (First order stochastic dominance, [17]): Let \( \pi_1, \pi_2 \subset \Pi \). Then \( \pi_1 \) first order stochastically dominates \( \pi_2 \) – denoted as \( \pi_1 \succeq_1 \pi_2 \) if \( \sum_{i=1}^X \pi_1(i) \geq \sum_{i=1}^X \pi_2(i) \) for \( j = 1, \ldots, X \).

The following result is well known [17]. It says that MLR dominance implies first order stochastic dominance, and it gives a necessary and sufficient condition for stochastic dominance.

Result 1 ([17]): (i) Let \( \pi_1, \pi_2 \subset \Pi \). Then \( \pi_1 \succeq_1 \pi_2 \) implies \( \pi_1 \succeq_1 \pi_2 \).
(ii) Let \( V \) denote the set of all \( X \) dimensional vectors \( v \) with nondecreasing components, i.e., \( v_1 \leq v_2 \leq \cdots \leq v_X \). Then \( \pi_1 \succeq_1 \pi_2 \) if and only if all \( v \in V, \nu' \pi_1 \geq \nu' \pi_2 \).

Definition 3 (Total Positivity of order 2): A transition matrix \( P \) is totally positive of order 2 (TP2) if every second order minor of \( P \) is non-negative. Equivalently, every row is dominated by a subsequent row with respect to the MLR order.

2) Copositivity: The following definitions of copositive matrices and a copositive ordering of stochastic matrices will be used extensively.

Definition 4 (Copositivity on simplex): An arbitrary \( X \times X \) matrix \( M \) is copositive if \( \pi' M \pi \geq 0 \) for all \( \pi \in \Pi \), or equivalently, if \( \pi' M \pi \geq 0 \) for all \( \pi \in \mathbb{R}_+^X \).

The definition says copositivity on the unit simplex and positive orthant are equivalent. Clearly positive semidefinite matrices and non-negative matrices are copositive.

Given two \( X \times X \) dimensional transition matrices \( P \) and \( Q \), we now define a sequence of matrices \( M^{(m)}(Q, P) \), indexed by \( m = 1, 2, \ldots, X - 1 \), as follows: Each \( M^{(m)}(Q, P) \) is a symmetric \( X \times X \) matrix of the form:
\[
M^{(m)}(Q, P) = Q_m P_{m+1} + P_{m+1} Q'_m - P_m Q'_{m+1} - Q_{m+1} P'_m
\] (9)

Here \( P_m \) and \( Q_m \), respectively, denote the \( m \)-th column of matrix \( P \) and \( Q \).

Definition 5 (Copositive Ordering \( \succeq \) of Stochastic Matrices): Given two \( X \times X \) transition matrices \( P \) and \( Q \), we say \( Q \succeq P \) (equivalently, \( P \succeq Q \)) if all the matrices \( M^{(m)}(Q, P) \), \( m = 1, 2, \ldots, X - 1 \), defined in (9), are copositive.

Intuition: The ordering of transition matrices \( Q \succeq P \) implies that the optimal filtering updates satisfy \( T(\pi, y; Q) \leq_r T(\pi, y; P) \) for any observation \( y \) and posterior \( \pi \). (Recall the Bayesian update \( T(\pi, y, \cdot) \) is defined in (2) and \( \leq_r \) denotes the MLR order.) In other words the \( \succeq \) ordering of transition matrices preserves the MLR ordering \( \leq_r \) of posterior distributions computed via the optimal filter. This property will be proved in Theorem 1 below. This is a crucial property that will be used subsequently in deriving lower and upper bounds to the optimal filtered posterior. It is easily verified that \( \succeq \) is a partial order over the set of stochastic matrices, i.e., \( \succeq \) satisfies reflexivity, antisymmetry and transitivity.

C. Upper and Lower Sample Path Stochastic Dominance Bounds to Posterior

With the above definitions, we are now ready to state the main result of this section. Recall that the original filtering problem seeks to compute \( \pi_{k+1} = T(\pi_k, y_{k+1}; P) \) using the filtering update (3) with transition matrix \( P \) and involves \( O(X^2) \) multiplications. This can be excessive for large \( X \). Our goal is to construct low rank transition matrices \( \bar{P} \) and \( P \) such that the filtering recursion using these matrices form lower and upper bounds to \( \pi_k \) in the MLR stochastic dominance sense. Due to the low rank of \( P \) and \( \bar{P} \), the cost involved in computing these lower and upper bounds to \( \pi_k \) at each time \( k \) will be \( O(XR) \) where \( R \ll X \) (for example, \( R = O(\log X) \)).

Since we plan to compute filtered estimates using \( P \) and \( \bar{P} \) instead of the original transition matrix \( P \), we need further notation to distinguish between the posteriors and estimates computed using \( P \) and \( \bar{P} \). Let
\[
\pi_{k+1} = T(\pi_k, y_{k+1}; P), \quad \bar{\pi}_{k+1} = T(\bar{\pi}_k, y_{k+1}; \bar{P}),
\]
\[
\tilde{\pi}_{k+1} = T(\tilde{\pi}_k, y_{k+1}; \tilde{P})
\]

lower
denote the posterior updated using optimal filter \(\mathcal{E}\) with transition matrices \(P\), \(\bar{P}\) and \(\bar{P}\), respectively. Also, similar to \(\mathcal{E}\), with \(g = (1, 2, \ldots, X)\), the conditional mean estimates of the underlying state computed using \(\bar{P}\) and \(\bar{P}\), respectively, will be denoted as
\[
\bar{\pi}_k \triangleq \mathbb{E}\{x_k|y_{0:k}; \bar{P}\} = g'\bar{\pi}_k, \quad \bar{x}_k \triangleq \mathbb{E}\{x_k|y_{0:k}; \bar{P}\} = g'\bar{\pi}_k.
\]
In analogy to \((7)\), denote the “hard” MAP state estimates computed using \(\bar{P}\) and \(\bar{P}\) as
\[
\bar{\pi}_k^{\text{MAP}} \triangleq \arg\max_i \bar{\pi}_k(i), \quad \bar{x}_k^{\text{MAP}} \triangleq \arg\max_i \bar{\pi}_k(i).
\] (11)

The following is the main result of this section. Recall the definition of copositivity ordering \(\preceq\), MLR dominance and TP2 in Section II-B.

**Theorem 1 (Stochastic Dominance Sample-Path Bounds):** Consider the filtering updates \(T(\pi, y; P), T(\pi, y; \bar{P})\) and \(T(\pi, y; \bar{P})\) where \(T(\cdot, \cdot)\) is defined in \((2)\) and \(P\) denotes the transition matrix of the original filtering problem.

1) For any transition matrix \(P\), there exist transition matrices \(\hat{P}\) and \(\check{P}\) such that \(P \preceq \hat{P} \preceq \check{P}\) (recall \(\preceq\) is defined in Definition \(\text{[3]}\)).

2) Suppose transition matrices \(P\) and \(\hat{P}\) are constructed such that \(P \preceq \hat{P} \preceq P\). Then for all \(y\) and \(\pi \in \Pi\), the filtering updates satisfy the sandwich result
\[
T(\pi, y; \hat{P}) \leq T(\pi, y; \check{P}) \leq T(\pi, y; P).
\]

3) Suppose \(P\) is TP2 (Definition \(\text{[3]}\)). Assume the filters \(T(\pi, y; P), T(\pi, y; \hat{P})\) and \(T(\pi, y; \check{P})\) are initialized with common prior \(\pi_0\) at time 0. Then the posteriors satisfy
\[
\pi_k \preceq \pi_k \preceq \pi_k, \quad \text{for all time } k = 1, 2, \ldots
\]
As a consequence for all time \(k = 1, 2, \ldots\),

a) The “soft” conditional mean state estimates defined in \((6), (10)\) satisfy \(\bar{\pi}_k \leq \bar{x}_k \leq \bar{x}_k\).

b) The “hard” MAP state estimates defined in \((7), (11)\) satisfy \(\bar{\pi}_k^{\text{MAP}} \leq \bar{x}_k^{\text{MAP}} \leq \bar{x}_k^{\text{MAP}}\).

Statement 1 says that for any transition matrix \(P\), there always exist transition matrices \(\hat{P}\) and \(\check{P}\) such that \(P \preceq \hat{P} \preceq \check{P}\) (copositivity dominance). An obvious but useless construction is \(\hat{P} = [e_1, \ldots, e_1]')\) and \(\check{P} = [e_X, \ldots, e_X]')\) where \(e_i\) is the unit \(X\)-dimensional vector with 1 in the \(i\)th position. These correspond to extreme points on the space of matrices with respect to copositive dominance. Given existence of \(\hat{P}\) and \(\check{P}\), the next step is to optimize the choice of \(\hat{P}\) and \(\check{P}\) - that is the subject of Sec III where nuclear norm minimization is used to construct sparse eigenvalue matrices \(\hat{P}\) and \(\check{P}\).

Statement 2 says that for any prior \(\pi\) and observation \(y\), the one step update of the filter lower and upper bounds the original filtering problem.

Statement 3 globalizes Statement 2 and asserts that with the additional assumption that the transition matrix \(P\) of the original filtering problem is TP2, then the upper and lower bounds hold for all time. Since MLR dominance implies first order stochastic dominance (Result \(\text{[1]}\)), the conditional mean estimates satisfy \(\bar{\pi}_k \leq \bar{x}_k \leq \bar{x}_k\).

**Why MLR Dominance?:** The proof of Theorem \(\text{[1]}\) in the appendix uses the result that \(\pi \preceq_\pi \pi\) implies that the filtered update \(T(\pi, y; P) \leq T(\pi, y; P)\). Such a result does not hold with first order stochastic dominance \(\leq_s -\) that is, \(\pi \preceq_\pi \pi\) does not imply that \(T(\pi, y; P) \leq T(\pi, y; P)\). In other words, the MLR order is closed with respect to conditional expectations. This the reason why we use the MLR order in this paper.

**D. Stochastic Dominance Bounds for Multivariate HMMs**

We conclude this section by showing how the above bounds can be generalized to multivariate HMMs – the main idea is that MLR dominance is replaced by the multivariate TP2 (totally positive of order 2) stochastic dominance \([17], [7], [6]\). We consider a highly stylized example which will serve as a reproducible way of constructing large scale HMMs in numerical studies of Sec VI.

Consider \(L\) independent Markov chains, \(x_{(l)}\), \(l = 1, 2, \ldots, L\) with transition matrices \(A^{(l)}\). Define the joint process \(x_k = (x_{(1)}(k), \ldots, x_{(L)}(k))\). Suppose the observation process recorded at a sensor has the conditional probabilities \(B_{x,y} = \mathcal{P}(y_k = y|x_k = x)\). Even though the individual Markov chains are independent of each other, since the observation process involves all \(L\) Markov chains, computing the filtered estimate of \(x_k\), requires computing and propagating the joint posterior \(P(x_{(l)}|y_{1:k})\). This is equivalent to HMM filtering the process \(x_k\) with transition matrix \(P = A^{(1)} \otimes \cdots \otimes A^{(L)}\) where \(\otimes\) denotes Kronecker product. For example, if each process \(x^{(l)}\) has \(S\) states, then \(P\) is an \(S^L \times S^L\) matrix and the computational cost of the HMM filter at each time is \(O(S^{2L})\) which is excessive for large \(L\).

A naive application of the results of the previous sections will not work, since the MLR ordering does not apply to the multivariate case. Instead, we use the totally positive (TP2) stochastic order, which is a multivariate generalization of the MLR order. Let \(i = (i_1, \ldots, i_L)\) and \(j = (j_1, \ldots, j_L)\) denote the indices of two \(L\)-variate probability mass functions Denote
\[
i \wedge j = [\min(i_1, j_1), \ldots, \min(i_L, j_L)'],
\]
\[
i \vee j = [\max(i_1, j_1), \ldots, \max(i_L, j_L)']'.
\] (12)

**Definition 6 (TP2 ordering and Reflexive TP2 distributions):** Let \(P\) and \(Q\) denote any two \(L\)-variate probability mass functions. Then:

(i) \(P \succeq_{TP2} Q\) if \(P(i)Q(j) \leq P(i \wedge j)Q(i \wedge j)\). If \(P\) and \(Q\) are univariate, then this definition is equivalent to the MLR ordering \(P \succeq_{ MLR} Q\) defined above.

(ii) A multivariate distribution \(P\) is said to be multivariate TP2 (MTP2) if \(P \succeq_{TP2} P\) holds, i.e., \(P(i)P(j) \leq P(i \vee j)P(i \wedge j)\). If \(i, j \in \{1, \ldots, X\}\) are scalar indices, this is equivalent to saying that an \(X \times X\) matrix \(P\) is MTP2 if all second order minors are non-negative.

With suitable notational abuse, in analogy to Definition \(\text{[5]}\) given two transition matrices \(\hat{P}\) and \(\check{P}\) and a multivariate belief \(\pi\), we say
\[
P \succeq_{TP2} \hat{P} \text{ if } P'\pi \succeq_{TP2} \check{P}'\pi.
\] (13)
The main result regarding filtering of multivariate HMMs is as follows:

**Theorem 2:** Consider an L-variate HMM where each transition matrix satisfies $A^{(i)} \preceq A^{(i)}$ for $l = 1, \ldots, L$ (where $\prec$ is interpreted as in Definition 5). Then (i) Then $A^{(1)} \otimes \cdots \otimes A^{(L)} \preceq A^{(1)} \otimes \cdots \otimes A^{(L)}$ where $\preceq$ is interpreted as $\leq$. (ii) Theorem 1 holds for the posterior and state estimates with $\geq_r$ replaced by $\geq$. 

We need to qualify statement (ii) of Theorem 2 since for multivariate HMMs, the conditional mean $\hat{x}_k$ and MAP estimate $x_k^{MAP}$ are L-dimensional vectors. The inequality $\hat{x}_k \preceq \hat{x}_k$ of statement (ii) is interpreted as the component wise partial order on $\mathbb{R}^L$, namely, $\hat{x}_k(l) \preceq \hat{x}_k(l)$ for all $l = 1, \ldots, L$. (A similar result applies for the upper bounds.)

III. CONVEX OPTIMIZATION TO COMPUTE LOW RANK TRANSITION MATRICES $\tilde{P}$, $\tilde{\tilde{P}}$

It only remains to give algorithms for constructing low rank transition matrices $\tilde{P}$ and $\tilde{\tilde{P}}$ that yield the lower and upper bounds $\pi_k$ and $\pi_k$. These involve convex optimization [18], [19] for minimizing the nuclear norm. The computation of $\tilde{P}$ and $\tilde{\tilde{P}}$ is independent of the observation sample path and so the associated computational cost is irrelevant to the real time filtering. Recall that the motivation is as follows: If $P$ and $\tilde{P}$ have rank $R$, then the computational cost of the filtering recursion is $O(RX)$ instead of $O(X^2)$ at each time $k$.

A. Construction of $\tilde{P}$, $\tilde{\tilde{P}}$ without rank constraint

Given a TP2 matrix $P$, the transition matrices $\bar{P}$ and $\tilde{\tilde{P}}$ such that $\bar{P} \preceq P \preceq \tilde{\tilde{P}}$ can be constructed straightforwardly via an LP solver. With $\bar{P}_1, \bar{P}_2, \ldots, \bar{P}_X$ denoting the rows of $\bar{P}$, a sufficient condition for $\bar{P} \preceq P$ is that $\bar{P}_i \preceq P_1$ for any row $i$. So the rows $\bar{P}_i$ satisfy linear constraints with respect to $P_1$ and can be straightforwardly constructed via an LP solver. A similar construction holds for the upper bound $\tilde{\tilde{P}}$, where it is sufficient to construct $\tilde{\tilde{P}}_i \preceq P_X$.

**Rank 1 bounds:** If $P$ is TP2, an obvious construction is to construct $\bar{P}$ and $\tilde{\tilde{P}}$ as follows: Choose rows $\bar{P}_1 = P_1$ and $\tilde{\tilde{P}}_1 = P_X$ for $i = 1, 2, \ldots, X$. These yield rank 1 matrices $\bar{P}$ and $\tilde{\tilde{P}}$. It is clear from Theorem 2 that $\bar{P}$ and $\tilde{\tilde{P}}$ constructed in this manner are the tightest rank 1 lower and upper bounds.

B. Nuclear Norm Minimization Algorithms to Compute Low Rank Transition Matrices $\bar{P}$, $\tilde{\tilde{P}}$

In this subsection we construct $\bar{P}$ and $\tilde{\tilde{P}}$ as low rank transition matrices subject to the condition $\bar{P} \preceq P \preceq \tilde{\tilde{P}}$. To save space we consider the lower bound transition matrix $\bar{P}$; construction of $\tilde{\tilde{P}}$ is similar. Consider the following optimization problem for $\bar{P}$:

\[
\text{Minimize rank of } X \times X \text{ matrix } \bar{P} \quad (14)
\]

subject to the constraints $\text{Cons}(\Pi, m)$ for $m = 1, 2, \ldots, X-1$, where for $\epsilon > 0$,

\[
\begin{align*}
\text{Cons}(\Pi, m) & \equiv \begin{cases} 
M^{(m)}(\bar{P}, P) \text{ is copositive on } \Pi 
\end{cases} & (15a) \\
\|P' \pi - P \pi\|_1 & \leq \epsilon \text{ for all } \pi \in \Pi & (15b) \\
P & \geq 0, P' \bar{P} = 1. & (15c)
\end{align*}
\]

Recall $M$ is defined in (9). The constraints $\text{Cons}(\Pi, m)$ are convex in matrix $\bar{P}$, since (15a) is linear in the elements of $\bar{P}$ and (15b) is convex because norms are convex. The constraints (15a), (15c) are exactly the conditions of Theorem 1. Recall that (15a) is equivalent to $\bar{P} \preceq P$. The convex constraint (15b) is equivalent to $\| P' - P\|_1 \leq \epsilon$, where $\| \cdot \|_1$ denotes the induced 1-norm for matrices.

To solve the above problem, we proceed in two steps:

1) The objective (14) is replaced with the reweighted nuclear norm (Sec. III-B1 below).

2) Optimization over the copositive cone (15a) is achieved via a sequence of simplicial decompositions (Sec. III-B2 below).

1) Reweighted Nuclear Norm: Since the rank is a non-convex function of a matrix, direct minimization of the rank (14) is computationally intractable. Instead, we follow the approach developed by Boyd and coworkers [18], [19] to minimize the iteratively reweighted nuclear norm. As mentioned earlier, inspired by Cand`es and Tao [12], there has been much recent interest in minimizing nuclear norms for constructing matrices with sparse eigenvalue sets or equivalently low rank. Here we compute $\bar{P}$, $\tilde{\tilde{P}}$ by minimizing their nuclear norms subject to copositivity conditions that ensure $\bar{P} \preceq P \preceq \tilde{\tilde{P}}$.

The re-weighted nuclear norm minimization proceeds as a sequence of convex optimization problems indexed by $n = 0, 1, \ldots$. Initialize $P^{(0)} = I$. For $n = 0, 1, \ldots$, compute $X \times X$ matrix

\[
P^{(n+1)} = \arg\min_{P_2} \|W_{1}^{(n)} P W_{2}^{(n)}\|_*,
\]

subject to: constraints $\text{Cons}(\Pi, m)$, $m = 1, \ldots, X-1$ namely, (15a), (15b), (15c).

Here $\| \cdot \|_*$ denotes the nuclear norm, which corresponds to the sum of the singular values of a matrix, and the weighting matrices $W_{1}^{(n)}$, $W_{2}^{(n)}$ are evaluated iteratively as

\[
\begin{align*}
W_{1}^{(n+1)} &= (W_{1}^{(n)})^{-1} U S V T (W_{1}^{(n)})^{-1} + \delta I)^{-1/2}, \\
W_{2}^{(n+1)} &= (W_{2}^{(n)})^{-1} U S V T (W_{2}^{(n)})^{-1} + \delta I)^{-1/2}. 
\end{align*}
\]

Here $W_{1}^{(n)} P^{(n)} W_{2}^{(n)} = U S V T$ is a reduced singular value decomposition, starting with $W_{1}^{(0)} = W_{2}^{(0)} = I$ and $P^{(0)} = P$. Also $\delta$ is a small positive constant in the regularization term $\delta I$. In numerical examples of Sec. VI we used YALMIP with MOSEK and CVX to solve the above convex optimization problem.

Let us explain the above sequence of convex optimization problems. Notice that at iteration $n + 1$, the previous estimate, $P^{(n)}$ appears in the cost function of (16) in terms of weighing matrices $W_{1}^{(n)}$, $W_{2}^{(n)}$. The intuition behind the reweighting

\[3\] The three statements $\|P' \pi - P \pi\|_1 \leq \epsilon$, $\|P - \bar{P}\|_1 \leq \epsilon$ and $\sum_{i=1}^{X} \|P' - P\|_1 \pi(i) \leq \epsilon$ are all equivalent since $\|\pi\|_1 = 1$. 

subject to the constraints $\text{Cons}(\Pi, m)$ for $m = 1, 2, \ldots, X-1$, where for $\epsilon > 0$,
iterations is that as the estimates $P^{(n)}$ converge to the limit $P^{(\infty)}$, the cost function becomes approximately equal to the rank of $P^{(\infty)}$.

2) Simplicial Decomposition for copositive programming: Problem [16] is a convex optimization problem in $P$. However, one additional issue needs to be resolved: the constraints (15a) involve a copositive cone and cannot be solved directly by standard interior point methods. To deal with the copositive constraints (15a), we use the state-of-the-art simplicial decomposition method proposed in [14]. The nice key idea used in [14] is summarized in the following proposition.

**Proposition 1 (14):** Let $\Lambda$ denote any sub-simplex of the belief space $\Pi$. Then a sufficient condition for copositive condition (15a) to hold is that as the estimates $\pi$ converge to the limit $\pi^{(\infty)}$, the cost function becomes approximately equal to the rank of $\pi^{(\infty)}$.

Problem (16) is a convex optimization problem in $P$. However, one additional issue needs to be resolved: the constraints (15a) involve a copositive cone and cannot be solved directly by standard interior point methods. To deal with the copositive constraints (15a), we use the state-of-the-art simplicial decomposition method proposed in [14]. The nice key idea used in [14] is summarized in the following proposition.

**Proposition 1 (14):** Let $\Lambda$ denote any sub-simplex of the belief space $\Pi$. Then a sufficient condition for copositive condition (15a) to hold is that as the estimates $\pi$ converge to the limit $\pi^{(\infty)}$, the cost function becomes approximately equal to the rank of $\pi^{(\infty)}$.

Problem (16) is a convex optimization problem in $P$. However, one additional issue needs to be resolved: the constraints (15a) involve a copositive cone and cannot be solved directly by standard interior point methods. To deal with the copositive constraints (15a), we use the state-of-the-art simplicial decomposition method proposed in [14]. The nice key idea used in [14] is summarized in the following proposition.

**Proposition 1 (14):** Let $\Lambda$ denote any sub-simplex of the belief space $\Pi$. Then a sufficient condition for copositive condition (15a) to hold is that as the estimates $\pi$ converge to the limit $\pi^{(\infty)}$, the cost function becomes approximately equal to the rank of $\pi^{(\infty)}$.

Let $\Lambda_j \triangleq \{\Lambda_1^J, \ldots, \Lambda_J^J\}$ denote the set of subsimplices at iteration $J$ that constitute a partition of $\Pi$. Proposition 1 along with the nuclear norm minimization leads to a finite dimensional convex optimization problem that can be solved via the following 2 step algorithm:

**for** iterations $J = 1, 2, \ldots$,

1) Solve the sequence of convex optimization problems [16], $n = 1, 2, \ldots$ with constraints $\text{Cons}(\Lambda_j^J, m), \text{Cons}(\Lambda_j^J, m), \ldots, \text{Cons}(\Lambda_j^J, m)$, $m = 1, \ldots, X - 1$.

2) If nuclear norm $\|W_1(n), P W_2(n)\|_{\ast}$ decreases compared to that in iteration $J - 1$ by more than a pre-defined tolerance, systematically partition $\Lambda_j$ as described in [14] into $\Lambda_{j+1}$.

Set $J = J + 1$ and go to Step 1.

else Stop.

The iterations of the above simplicial algorithm lead to a sequence of decreasing costs $\|W_1(n), P W_2(n)\|_{\ast}$, hence the algorithm can be terminated as soon as the decrease in the cost becomes smaller than a pre-defined value (set by the user); please see [14] for details on simplex partitioning. We emphasize again that the algorithms in this section for computing $P$ and $P$ are off-line and do not affect the real time filtering computations.

### IV. STOCHASTIC DOMINANCE CONSTRAINED IMPORTANCE SAMPLING FILTER

So far we have constructed reduced complexity lower and upper stochastic dominance bounds that confine the posterior sample path of the optimal filter to the partially ordered set $\pi_k \preceq \pi_k \preceq \pi_k$ at each time $k$. The next question is: Given the estimates $\pi_k$ and $\pi_k$, how to construct an algorithm to estimate $\pi_k$? That is, how can the bounds $\pi_k$ and $\pi_k$ be exploited to estimate the posterior $\pi_k$? We present a filtering algorithm that is inspired by recent results in stochastic linear solvers [20], [21]. The algorithm uses importance sampling for matrix-vector multiplication together with Gibbs sampling to ensure that the estimated posterior $\pi_k$ lies in the partially ordered set $\pi_k \preceq \pi_k \preceq \pi_k$.

Why? Running a reduced complexity estimator and then switching to a high resolution estimator when an event of interest occurs, arises in monitoring systems, cued sensing in adaptive target tracking systems [22] and body area networks [23], [24]. In these examples, it is of interest to detect when the underlying Markov chain is close to a target state. A sensor monitoring the state of a noisy Markov chain can compute the reduced complexity filtering bounds cheaply. Since the reduced complexity bounds provably sandwich the true posterior, as soon as these bounds get close to a target state, the sensor switches to a higher resolution (complexity) estimator. For example, in cued target tracking, when a target’s state approaches a high threat level, the reduced complexity tracker can cue (deploy) a higher resolution (complexity) tracker.

**Remark:** We emphasize at the outset that obviously, $P(x_k|y_1:k) = P(x_k|y_1:k, \pi_k, \pi_k)$ since $\pi_k$ and $\pi_k$ are $y_1:k$ measurable. That is, the posterior (and therefore, the conditional mean estimate) is exactly the same whether or not the upper and lower bounds are used. (In other words, since the upper and lower bounds were computed using the same observations as the conditional mean estimate, they cannot be used to obtain a better conditional mean estimate.) This section deals with estimating the posterior - the posterior estimate conditioned on the upper and lower bounds has a lower variance than the unconditional estimator.

#### A. Stochastic Dominance Constrained Importance Sampling Filtering Algorithm

Suppose we have an estimate $\hat{\pi}_{k-1}$ of the posterior such that $\pi_{k-1} \preceq \hat{\pi}_{k-1} \preceq \bar{\pi}_{k-1}$. Algorithm 1 constructs an estimate $\hat{\pi}_{k-1}$ using Monte-Carlo important sampling methods for matrix-vector multiplication so that the predicted distributions satisfy $\pi_{k-1} \preceq \hat{\pi}_{k-1} \preceq \bar{\pi}_{k-1}$. Once $\pi_{k-1}$ is constructed, the filtered posterior at time $k$ is straightforwardly computed with $O(X)$ computations as $\hat{\pi}_k \propto B_{y_k} \pi_{k-1}$ (Bayes rule). Moreover, by Theorem 1 [14] this updated posterior is guaranteed to satisfy $\pi_{k-1} \preceq \hat{\pi}_k \preceq \bar{\pi}_k$.

**Algorithm 1**: Algorithm 1 is equivalent to $\alpha_j \leq \hat{\pi}_{k-1}(j) \leq \hat{\pi}_{k-1}(j-1)$, $\pi_{k-1}(j-1)$. This in turn is equivalent to the sample path bound $\pi_{k-1}(j) \preceq \hat{\pi}_{k-1}(j) \preceq \bar{\pi}_{k-1}(j-1)$, $\pi_{k-1}$, $\pi_{k-1}$. The key point in Algorithm 1 is that the reduced variance compared to the unconstrained estimator since $\text{var}(\hat{\pi}|\pi \preceq \hat{\pi} \preceq \bar{\pi}) = \text{var}(\hat{\pi})$.

If the stochastic dominance constraints are not exploited, then $\alpha_j = 0$ and $\beta_j = 1 - \sum_{l=1}^{j-1} \bar{\pi}_{k+1}(k) \in [19]$. The condition [20] uses Gibbs sampling to ensure that the constraints hold - this is simply a special case of adaptive importance sampling.

**Choice of Importance Distribution:** In Algorithm 1 the importance distribution $q_j$ is an X-dimension probability vector. There are several choices for the importance distribution $q_j$.

1) An obvious choice is $q_j(i) = \pi_k(i)$, in which case $P_{ij}$ becomes: If $P_{ij} \in [\alpha_j, \beta_j]$ then $\pi_{k-1}(j) \leftarrow P_{ij}$.

2) The optimal importance function, which minimizes the variance of $\pi_{k-1}(j-1)$, is $q_j(i) \propto P_{ij} \pi_{k-1}$. This is not useful since evaluating it requires $O(X)$ multiplications for each $j$ and therefore $O(X^2)$ multiplications in total.

3) A near optimal choice is to choose $q_j(i) \propto P_{ij} \pi_{k-1}$ or $q_j(i) \propto P_{ij} \pi_{k-1}(i)$. These have already been computed.

---

4We thank Eric Moulines of ENST for mentioning this.
Algorithm 1 Stochastic Dominance Constrained Importance Sampling Filter at time $k$

**Aim:** Given posterior estimate $\hat{\pi}_{k-1}$, lower bound $\underline{\pi}_{k-1}$ and upper bound $\bar{\pi}_{k-1}$, evaluate $\bar{\pi}_k$.

**Step 0 (offline):** Given TP2 transition matrix $P$, compute low rank $P$ and $\bar{P}$ with $P \preceq \bar{P} \preceq \bar{P}$ by minimizing nuclear norm (Sec.[III-B]).

**Step 1:** Evaluate predicted & filtered upper/lower bounds

$$
\underline{\pi}_{k|k-1} = \underline{P}' \underline{\pi}_{k-1}, \quad \bar{\pi}_k = \frac{B_{y_k} \underline{\pi}_{k|k-1}}{1/B_{y_k} \bar{\pi}_{k|k-1}},
$$

$$
\bar{\pi}_{k|k-1} = \bar{P}' \bar{\pi}_{k-1}, \quad \bar{\pi}_k = \frac{B_{y_k} \bar{\pi}_{k|k-1}}{1/B_{y_k} \bar{\pi}_{k|k-1}}.
$$

(18)

**Step 2:** Compute estimate $\bar{\pi}_k$ using $\underline{\pi}_k$ and $\bar{\pi}_k$.

for $j = 1 \to X$

Evaluate stochastic dominance path bounds $\alpha_j$ and $\beta_j$:

$$
\alpha_j = 0, \quad \beta_j = 1 
$$

and for $j > 1$,

$$
\alpha_j = \frac{\bar{\pi}_{k|k-1}(j-1) \underline{\pi}_{k|k-1}(j-1)}{\bar{\pi}_{k|k-1}(j)},
$$

$$
\beta_j = \min \left\{ \frac{\bar{\pi}_{k|k-1}(j) \underline{\pi}_{k|k-1}(j-1)}{\bar{\pi}_{k|k-1}(j-1)}, 1 - \sum_{l=1}^{j-1} \underline{\pi}_{k|k-1}(l) \right\}. \tag{19}
$$

for iterations $l = 1 \to L$

Sample $i_l \in \{1, \ldots, X\}$ from importance probability mass function $q_{jl}$,

if $P_{i_l,j} \bar{\pi}_{k|k-1}(i_l) \in [\alpha_j, \beta_j]$ then

Set $F_j = F_j \cup \{i\}$ and $\bar{\pi}_{k|k-1}(j) = \frac{P_{i_l,j} \bar{\pi}_{k|k-1}(i_l)}{q_{jl}(i_l)} \tag{20}$

end if

end for

Set $\bar{\pi}_{k|k-1}(j) = \frac{1}{|F_j|} \sum_{l \in F_j} \bar{\pi}_{k|k-1}(j)$.

(IF $F_j$ is empty, set $\bar{\pi}_{k|k-1}(j) = \underline{\pi}_{k|k-1}(j)$)

end for

Compute filtered posterior estimate $\bar{\pi}_k = \frac{B_{y_k} \bar{\pi}_{k|k-1}}{1/B_{y_k} \bar{\pi}_{k|k-1}}$

where $R$ is the rank of $P$. In comparison a particle filter with $L$ particles involves $O(L)$ computational cost.

B. Importance Sampling Filter for Computing Lower Bound

Given the lower bound matrix $P$ of rank $R$, $\bar{\pi}_k$ can be computed exactly using (18) with $\hat{O}(RX)$ computations. An alternative method is to exploit the rank $R$ and estimate $\bar{\pi}_k$ by using Monte-Carlo importance sampling methods similar to Algorithm 1. Consider the singular value decomposition of $P$:

$$
P' \pi = \sum_{r=1}^{R} \sigma_r v_r u_r' \pi \tag{21}
$$

where we have minimized rank $R$ via the nuclear norm minimization algorithm of Sec.[III-B]. Algorithm 2 presents the importance sampling filter for $\pi$ (the upper bound is similar).

Algorithm 2 Importance Sampling Filter for estimating lower bound $\underline{\pi}_k$ at time $k$

**Aim:** Given lower bound estimate $\hat{\pi}_{k-1}$, evaluate lower bound $\underline{\pi}_k$.

for $r = 1 \to R$

for iterations $l = 1 \to L$

Sample $i_l \in \{1, \ldots, X\}$ from importance probability mass function $q_{jr}$.

Set $\hat{u}_r(l) = u_{r}(i_l) \frac{p_{r}(i_l)}{q_{jr}(i_l)}$

end for

Set $\hat{u}_r = \frac{1}{L} \sum_{l=1}^{L} \hat{u}_r(l)$

end for

Set $\underline{\pi}_{k|k-1} = \sum_{r=1}^{R} \sigma_r v_r \hat{u}_r$ (where the vectors $\sigma_r, v_r, r = 1, \ldots, R$ are precomputed).

Compute filtered posterior $\underline{\pi}_k = \frac{B_{y_k} \underline{\pi}_{k|k-1}}{1/B_{y_k} \underline{\pi}_{k|k-1}}$

The choice of importance sampling distributions $q$ is similar to that for Algorithm 1.

C. Stochastic Dominance Constrained Particle Filter – A Non-result

Given the abundance of publications in particle filtering, it is of interest to obtain a particle filtering algorithm that exploits the upper and lower bound constraints to estimate the posterior. Unfortunately, since particle filters propagate trajectories and not marginals, we were unable to find a computationally efficient way of enforcing the MLR constraints $\underline{\pi}_k \leq_r \bar{\pi}_k \leq_r \bar{\pi}_k$ in the computation of $\bar{\pi}_k$. (If we propagated the marginals, then the algorithm becomes identical to Algorithm 1). Also, since MLR comparison of two $X$-dimensional posteriors involves $O(X)$ multiplications, projecting $L$ particles to the polytope $\underline{\pi}_k \leq_r \bar{\pi}_k \leq_r \bar{\pi}_k$ involves $O(LX)$ computations. Finally, in the particle filtering folklore, the so called ‘optimal’ choice for the importance density is $q(x_k|x_{0:k-1}, y_{1:k}) = P(x_k|x_{k-1}, y_k)$ with particle weight update $w_k(l) = w_k(l-1) \sum_{j=1}^{X} P_{x_{k-1:p}} B_{y_k}$.

For each particle, this requires $O(X)$ computations and hence $O(LX)$ for $L$ particles.
V. Analysis of Bounds

Our main result, namely, Theorem 1 above, is an ordinal bound: It said that we can compute reduced complexity filters \( \tilde{\pi}_k \) and \( \bar{\pi}_k \) such that the posterior \( \pi_k \) of the original filtering problem is lower and upper bounded on the partially ordered set: \( \tilde{\pi}_k \leq \pi_k \leq \bar{\pi}_k \) for all time \( k \). Moreover, by minimizing (16), we computed transition matrices \( P \) and \( \bar{P} \) so that \( \|P\pi - \bar{P}\pi\|_1 \leq \epsilon \) and \( \|P'\pi - \bar{P}'\pi\|_1 \leq \epsilon \).

In this section we construct cardinal bounds – that is, an explicit analytical bound is developed for \( \|\pi_k - \bar{\pi}_k\| \) and therefore \( |x_k - \bar{x}_k| \) in terms of \( \epsilon \). These bounds together with Theorem 1 give a complete characterization of the reduced complexity filters.

To estimate the rank \( \rho(P) \) of the original transition matrix \( P \) we consider the case where \( \rho(P) \) is the maximum variational norm \( \|F(P)\|_1 \) between two rows of the transition matrix \( P \).

The following is the main result of this section:

**Theorem 3:** Consider a HMM with transition matrix \( P \) and state levels \( g \). Let \( \epsilon > 0 \) denote the user defined parameter in constraint (15) of convex optimization problem (16) and let \( \bar{P} \) denote the solution. Then

1) The expected absolute deviation between one step of filtering using \( P \) versus \( \bar{P} \) is upper bounded as:

\[
\mathbb{E}_y |g'(T(\pi, y; P) - T(\pi, y; \bar{P}))| \leq \epsilon \sum_{i,j} \max_{l \in \{1,2,\ldots,X\}} |P_{il} - \bar{P}_{il}|. \tag{20}
\]

2) The sample paths of the filtered posteriors and conditional means have the following explicit bounds at each time \( k \):

\[
|\pi_k - \bar{\pi}_k|_1 \leq \epsilon \max \{F(\bar{\pi}_{k-1}, y) - \epsilon, \mu(y_k)\} + \frac{\rho(\bar{P})}{F'(\bar{\pi}_{k-1}, y)} \|\pi_{k-1} - \bar{\pi}_{k-1}\|_1. \tag{21}
\]

Theorem 3 gives explicit upper bounds between the filtered distributions using transition matrices \( P \) and \( \bar{P} \). The \( \mathbb{E}_y \) in (20) is with respect to the measure \( \sigma(\pi, y; P) = 1' B_y P \pi \) which corresponds to \( \mathbb{P}(y_k = y|\pi_{k-1} = \pi) \). Similar bounds hold for \( \bar{P} \) and are omitted.

The bounds are useful since their computation involves the reduced complexity filter with transition matrices \( \bar{P} = \bar{P} \) – the original transition matrix \( P \) is not used. In numerical examples below, we illustrate (21).

VI. Numerical Examples

In this section we present numerical examples to illustrate the behavior of the reduced complexity filtering algorithms proposed in this paper. To give the reader an easily reproducible numerical example of large dimension, we construct a 3125 state Markov chain according to the multivariate HMM construction detailed in Sec.II-D. Consider \( L = 5 \) independent Markov chains \( x_k^{(l)}, l = 1,\ldots,5 \), each with 5 states. The observation process is

\[
y_k = \sum_{l=1}^{5} x_k^{(l)} + v_k
\]

where the observation noise \( v_k \) is zero mean iid Gaussian with variance \( \sigma_v^2 \). Since the observation process involves all 5 Markov chains, computing the filtered estimate requires propagating the joint posterior. This is equivalent to defining a \( 5^5 = 3125 \) state Markov chain with transition matrix \( P = A \otimes \cdots \otimes A \) where \( \otimes \) denotes Kronecker product. The optimal HMM filter incurs \( 5^{10} \approx 10 \) million computations at each time step \( k \).

1) Generating TP2 Transition Matrix: To illustrate the reduced complexity global sample path bounds developed in Theorem 1 we consider the case where \( P \) is TP2. We used the following approach to generate \( P \): First construct \( A = \exp(Qt) \), where \( Q \) is a tridiagonal generator matrix (nonnegative off-diagonal entries and each row adds to 0) and \( t > 0 \). Karlin’s classic book [25, pp.154] shows that \( A \) is then TP2. Second, as shown in [61], the Kronecker products of \( A \) preserve the TP2 property implying that \( P \) is TP2.

Using the above procedure, we constructed a 3125 \( \times \) 3125 TP2 transition matrix \( P \) as follows:

\[
Q = \begin{bmatrix}
0.8147 & 0.8147 & 0 & 0 & 0 \\
0.4529 & -0.5164 & 0.0650 & 0 \\
0 & 0.4567 & -0.7729 & 0.3162 & 0 \\
0 & 0 & 0.0488 & -0.1880 & 0.1392 \\
0 & 0 & 0 & 0.5469 & -0.5469
\end{bmatrix}, \quad A = \exp(2Q), \quad P = A \otimes \cdots \otimes A. \tag{25}
\]

2) Off-line Optimization of Lower Bound via Convex Optimization: We used the semidefinite optimization solver from MOSEK with YALMIP and CVX to solve the convex optimization problem (16) for computing the upper and lower bound transition matrices \( \bar{P} \) and \( P \). To estimate the rank of the resulting transition matrices, we consider the costs (16), which correspond approximately to the number of singular
satisfies the normalized error bound 

\[ \| P - P^* \|_2 \leq \frac{\| \hat{P} - \pi \|_2}{2} \| \hat{P} \|_2 \leq 0.01, \]

thereby implying that approximating \( \hat{P} \) by \( P \) results in negligible error. For notational convenience, we continue to use \( \hat{P} \) instead of \( P \).

In each case, after computing the low rank matrix \( \hat{P} \), small singular values of \( \hat{P} \) were truncated to zero. The resulting matrix was then made stochastic by subtracting the minimum element of the matrix (thereby every element is non-negative) and then normalizing the rows. Both transformations do not affect the rank of the matrix. It was ensured that the resulting matrix \( \hat{P} \) satisfies the normalized error bound 

\[ \| P - P^* \|_2 \leq \frac{\| \hat{P} - \pi \|_2}{2} \leq 0.01, \]

thereby implying that approximating \( \hat{P} \) by \( P \) results in negligible error. For notational convenience, we continue to use \( \hat{P} \) instead of \( P \).

To save space we present results only for the lower bounds. We computed 5 different lower bound transition matrices \( \hat{P} \) by solving the nuclear norm minimization problem [16] for 5 different choices of \( \epsilon \in \{0.4, 0.8, 1.2, 1.6, 2\} \) defined in constraint [15b].

Table I displays the ranks of these 5 transition matrices \( \hat{P} \), and also the rank of \( \hat{P} \) which corresponds to the case \( \epsilon = 0 \). The low rank property of \( \hat{P} \) can be visualized by displaying the singular values. Fig. 1 displays the singular values of \( \hat{P} \) and \( \hat{P} \). When \( \epsilon = 2 \), the rank of \( \hat{P} \) is 1 and models an iid chain; \( \hat{P} \) then simply comprises of repetitions of the first row of \( P \). As \( \epsilon \) is made smaller the number of singular values increases. For \( \epsilon = 0 \), \( \hat{P} \) coincides with \( P \).

3) Performance of Lower Complexity Filters: At each time \( k \), the reduced complexity filter \( \hat{\pi}_k = T(\hat{\pi}_{k-1}, y_k; \hat{P}) \) incurs computational cost of \( O(XR) \) where \( X = 3125 \) and \( R \) is specified in Table 1. For each matrix \( \hat{P} \) and noise variances \( \sigma_y^2 \) in the range \((0, 2.25]\) we ran the reduced complexity HMM filter \( T(\pi, y; \hat{P}) \) for a million iterations and computed the average mean square error of the state estimate. These average mean square error values are displayed in Fig. 2. As might be intuitively expected, Fig. 2 shows that the reduced complexity filters yield a mean square error that lies between the iid approximation \((\epsilon = 2)\) and the optimal filter \((\epsilon = 0)\). In all cases, as mentioned in Theorem 1, the estimate \( \hat{\pi}_k \) provably lower bounds the true posterior \( \pi_k \) as \( \hat{\pi}_k \leq r \pi_k \) for all time \( k \). Therefore the conditional mean estimates satisfy \( \hat{\pi}_k \leq \hat{x}_k \) for all \( k \).

4) Stochastic Dominance Constrained Importance Sampling Algorithm 7. Recall Algorithm 1 computes the predicted posterior \( \bar{\pi}_{k|k-1} \) by exploiting the lower and upper bound stochastic dominance constraints. To illustrate the performance of Algorithm 1 we computed the mean square error between the estimated posterior using Algorithm 1 and optimal predictor, that is, \((\bar{\pi}_{k|k-1} - \pi_{k|k-1})^2\) averaged over a million belief states \( \pi_{k-1} \) sampled uniformly from the \( 5^5 - 1 \) dimensional

| \( \epsilon \) | 0 | 0.4 | 0.8 | 1.2 | 1.6 | 2 |
|---|---|---|---|---|---|---|
| \( R \) (rank of \( \hat{P} \)) | 3125 | 3125 | 800 | 232 | 165 | 40 |
| Posterior \( \hat{\pi}_k \) | 1 (iid) |

TABLE I: Ranks of lower bound transition matrices \( \hat{P} \) each of dimension \( 3125 \times 3125 \) obtained as solutions of the nuclear norm minimization problem [16] for six different choices of \( \epsilon \) appearing in constraint [15b]. Note \( \epsilon = 0 \) corresponds to \( \hat{P} = P \) and \( \epsilon = 2 \) corresponds to the iid case.

![Fig. 1: Plot of 3125 singular values of \( P \) and singular values of five different transition matrices \( \hat{P} \) parametrized by \( \epsilon \) in Table I](image1)

![Fig. 2: Mean Square Error of lower bound reduced complexity filters computed using five different transition matrices \( P \) summarized in Table I](image2)

![Fig. 3: Mean Square Error between optimal predictor and stochastic dominance constrained importance sampling predictor of Algorithm 1](image3)
whose ranks are given in Table I. The
and
≤
and
−
⪯
of rank 1. The
observation matrix parametrized by
b
is specified in (27).

We ran Algorithm [1] for 5 different values of
L
, namely,
2, 4, 6, 8, 10 iterations at each time step. Naturally, the more iterations
L
per time step, the more accurate the estimate. Fig.3(a) and 3(b) display these mean square errors for the constrained importance
sampling filter for 5 values of
L
. Fig.3(a) corresponds to
ε
= 2, resulting in
P
of rank 1. Fig.3(b) corresponds to
ε
= 1.6, resulting in
P
of rank 40. Recall the performance of the lower bound estimates
with these transition matrices were reported in Sec.VI-3. Fig.3(a) and 3(b) also display these mean square errors for the constrained importance
sampling filter for 5 values of
b
. The figures show that the bounds
get smaller,
ε
gets tighter and so one would
expect that
P
for larger values of
b
is reflected in the upper bound displayed in the figure. Second,
A. Proof of Theorem 1
π
k
implies
P
k
≤
P
k+1
, or equivalently
P
k
≤
P
k+1
. Combining the two inequalities yields
T(πk, yk+1; Pk) ≤
T(πk, yk+1; P)
. Next since
P
is TP2, it follows that
πk
≤
πk
implies
T(πk, yk+1; Pk) ≤
T(πk, yk+1; P). Thus
T(πk, yk+1; P
+1
) ≤
T(πk, yk+1; P)
. Finally, MLR dominance implies first order dominance of means thereby proving 3(a).

To prove 3(b) we need to show that
π
≤
π
implies
arg max
i
π
(i)
= arg max
i
π
(i)
. This is shown by contradiction: Let
i∗ = arg max
i
π
(i)
and
j∗ = arg max
j
π
(j)
. Suppose
i∗ ≤ j∗ . Then
π
≥
π
implies
π(i∗) ≤
π(j∗) . Since
π(i∗) ≤ 1, we have
π(i∗) ≤
π(j∗) which is a contradiction since
i∗
is the argmax for
π
(i).
VII. DISCUSSION
The main idea of the paper is to develop reduced complexity HMM
filtering algorithms with provable sample path bounds. At each iteration, the optimal HMM filter has
O(X2)
computations and our aim was to derive reduced complexity upper and lower bounds with complexity
O(XR)
including computational aspects [13], [14]. In future work it is worthwhile extending the bounds in this paper to copositive kernels for
continuous state filtering problems. Such results could yield guaranteed sample path bounds for general nonlinear filtering problems.
B. Proof of Theorem 2

If suffices to show that \( A \preceq A \implies A \otimes A \preceq A \otimes A \). (The proof for repeated Kronecker products then follows straightforwardly by induction.) Consider the TP2 ordering in Definition 6. The indices \( i = (j, n) \) and \( j = (f, g) \) are each two dimensional. There are four cases: \((j < f, n < g), (j < f, n > g), (j > f, n < g), (j > f, n > g)\). TP2 dominance for the first and last cases are trivial to establish. We now show TP2 dominance for the third case (the second case follows similarly): Choosing the indices \( i = (j, g - 1) \) and \( j = (j - 1, g) \), it follows that \( A \otimes A \preceq A \otimes A \) is equivalent to

\[
\sum_m \sum_l A_{m,g-1} A_{l,g} \sum_i (A_{ij} A_{k,j-1} - A_{i,j+1} A_{k,j}) \pi_{im} \pi_{kl} \leq 0
\]

So a sufficient condition is that for any non-negative numbers \( \pi_{im} \) and \( \pi_{kl} \), \( \sum_i \sum_k (A_{ij} A_{k,j-1} - A_{i,j+1} A_{k,j}) \pi_{im} \pi_{kl} \leq 0 \) which is equivalent to \( A \preceq A \) by Definition 5.

C. Proof of Theorem 3

We start with the following theorem that characterizes the \( l_1 \) (equivalently, variational distance) in the classical Bayes' rule. Recall that the Bayes' rule update prior \( \pi \) and observation \( y \)

\[
\mathcal{B}(\pi, y) = \frac{B_y \pi}{\mathcal{V} B_y \pi}.
\]

(Of course this is the same as the optimal transition operator being identity).

**Theorem 4:** Consider any two posterior probability mass functions \( \pi, \tilde{\pi} \in \mathcal{P} \). Then:

1. The variational distance in the Bayesian update satisfies

\[
||\mathcal{B}(\pi, y) - \mathcal{B}(\tilde{\pi}, y)||_{TV} \leq \max_i B_i y ||\pi - \tilde{\pi}||_{TV}.
\]

(Remark the variational distance is half the TV norm).  

2. The normalization term in Bayes' rule satisfies

\[
1' B_y \pi \geq \max \{1' B_y \tilde{\pi} - \varepsilon \max_i B_i y, \min_i B_i y\}.
\]

**Proof:** We refer to [11] for a textbook treatment of similar proofs on more general spaces.

1) **Statement 1:** For any \( g \in \mathbb{R}^X \),

\[
g' \left( \mathcal{B}(\pi, y) - \mathcal{B}(\tilde{\pi}, y) \right) = g' \left( \mathcal{B}(\pi, y) - \frac{B_y \tilde{\pi}}{1' B_y \pi} - \frac{B_y \pi}{1' B_y \pi} - \mathcal{B}(\tilde{\pi}, y) \right) = \frac{1}{1' B_y \pi} g' \left[ I - \mathcal{B}(\pi, y) 1' \right] B_y (\pi - \tilde{\pi}).
\]  

Applying the result [11] that for any vector \( f \in \mathbb{R}^X \),

\[
|g'(\pi - \tilde{\pi})| \leq \max_{i,j} |f_i - f_j| ||\pi - \tilde{\pi}||_{TV}
\]

to the right hand side of the above equation yields,

\[
|g' \left( \mathcal{B}(\pi, y) - \mathcal{B}(\tilde{\pi}, y) \right)| \leq \frac{1}{1' B_y \pi} \max_{i,j} |f_i - f_j| ||\pi - \tilde{\pi}||_{TV}
\]

3) **Proof of Theorem 3** With the above results we are now ready to prove the theorem. The triangle inequality for norms yields

\[
||\pi_{k+1} - \pi_k||_{TV} = ||T(\pi_k, y_{k+1}; P) - T(\pi_k, y_{k+1}; \tilde{P})||_{TV}
\]

\[
\leq ||T(\pi_k, y_{k+1}; P) - T(\pi_k, y_{k+1}; \tilde{P})||_{TV}
\]

\[
+ ||T(\pi_k, y_{k+1}; \tilde{P}) - T(\pi_k, y_{k+1}; \tilde{P})||_{TV}.
\]

Part 1: Consider the first normed term in the right hand side of (31). Applying (28) with \( \pi = P' \pi_k \) and \( \tilde{\pi} = P' \tilde{\pi}_k \) yields

\[
g'(T(\pi_k, y; P) - T(\pi_k, y; P')) = \frac{1}{\sigma(\pi, y; P)} g' \left[ I - T(\pi_k, y; P) \right] B_y (P - P') \pi
\]

where \( \sigma(\pi, y; P) = 1' B_y P' \). Then (29) yields

\[
g'(T(\pi_k, y; P) - T(\pi_k, y; P')) \leq \max_{i,j} \frac{1}{\sigma(\pi, y; P)} g' \left[ I - T(\pi_k, y; P) \right] B_y (e_i - e_j) ||P' \pi - P' \tilde{\pi}||_{TV}
\]

Since \( ||P' \pi - P' \tilde{\pi}||_{TV} \leq \epsilon \), taking expectations with respect to the measure \( \sigma(\pi, y; P) \), completes the proof of the first assertion.
Part 2: Applying Theorem 4(i) with the notation \( \pi = P' \pi_k \) and \( \tilde{\pi} = P' \tilde{\pi}_k \) yields

\[
\|T(\pi_k; y; P) - T(\tilde{\pi}_k, y; P)|\|_\infty \leq \frac{\max_i B_{i,y} \|P' \pi_k - \tilde{\pi}' \pi_k\|_{1V}}{1' B_y P' \pi_k} 
\leq \frac{\epsilon}{2} \max_i B_{i,y} \frac{\epsilon/2}{\max_i B_{i,y}, \min_i B_{i,y}}.
\]

The second last inequality follows from the construction of \( P \) satisfying (15b) (recall the variational norm is half the \( l_1 \) norm). The last inequality follows from Theorem 4(ii).

Consider the second normed term in the right hand side of (31). Applying Theorem 4(i) with notation \( \pi = P' \pi_k \) and \( \tilde{\pi} = P' \tilde{\pi}_k \) yields

\[
\|T(\pi_k; y; P) - T(\tilde{\pi}_k, y; P)|\|_\infty \leq \frac{\max_i B_{i,y} \|P' \pi_k - \tilde{\pi}' \pi_k\|_{1V}}{1' B_y P' \pi_k} 
\leq \frac{\max_i B_{i,y} \rho(P) \|\pi_k - \tilde{\pi}_k\|_{1V}}{1' B_y P' \pi_k} \tag{33}
\]

where the last inequality follows from the submultiplicative property of the Dobrushin coefficient. Substituting (32) and (33) into the right hand side of the triangle inequality (31) proves the result.

REFERENCES

[1] O. Cappe, E. Moulines, and T. Ryden, Inference in Hidden Markov Models., Springer-Verlag, 2005.
[2] R. Elliott, L. Aggoun, and J. Moore, Hidden Markov Models – Estimation and Control. New York: Springer-Verlag, 1995.
[3] V. Krishnamurthy, “Bayesian sequential detection with phase-distributed change time and nonlinear penalty – a lattice programming POMDP approach,” IEEE Transactions on Information Theory, vol. 57, no. 3, pp. 7096–7124, Oct. 2011.
[4] W. Lovejoy, “Some monotonicity results for partially observed Markov decision processes,” Operations Research, vol. 35, no. 5, pp. 736–743, Sept.-Oct. 1987.
[5] W. Whitt, “A note on the influence of the sample on the posterior distribution,” Journal American Statistical Association, vol. 74, pp. 424–426, 1979.
[6] S. Karlin and Y. Rinott, “Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions,” Journal of Multivariate Analysis, vol. 10, no. 4, pp. 467–498, December 1980.
[7] W. Whitt, “Multivariate monotone likelihood ratio and uniform conditional stochastic order,” Journal Applied Probability, vol. 19, pp. 695–701, 1982.
[8] Z. Lin and L. Vandenberghe, “Interior-point method for nuclear norm approximation with application to system identification,” SIAM Journal on Matrix Analysis and Applications, vol. 31, no. 3, pp. 1235–1256, 2009.
[9] Q. Zhang, G. Yin, and J. Moore, “Two-time-scale approximation for Wonham filters,” Information Theory, IEEE Transactions on, vol. 53, no. 5, pp. 1706–1715, 2007.
[10] G. Yin, Q. Zhang, J. Moore, and Y. Liu, “Continuous-time tracking algorithms involving two-time-scale Markov chains,” Signal Processing, IEEE Transactions on, vol. 53, no. 12, pp. 4442–4452, 2005.
[11] U. Rieder, “Structural results for partially observed control models,” Methods and Models of Operations Research, vol. 35, no. 6, pp. 473–490, 1991.
[12] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2053–2080, May 2009.
[13] S. Bundfuss and M. Dür, “Algorithmic copositivity detection by simplicial partition,” Linear Algebra and its Applications, vol. 426, no. 7, pp. 1511–1523, 2008.
[14] ——–, “An adaptive linear approximation algorithm for copositive programs,” SIAM Journal on Optimization, vol. 20, no. 1, pp. 30–53, 2009.