Reduced Complexity HMM Filtering With Stochastic Dominance Bounds: A Convex Optimization Approach

Vikram Krishnamurthy, *Fellow, IEEE*, and 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. We consider an HMM consisting of an X-state Markov chain with transition matrix P. By using convex optimization methods for nuclear norm minimization with copositive constraints, we construct low rank stochastic matrices $\underline{P}$ and $\bar{P}$ so that the optimal filters using $\underline{P}, \bar{P}$ provably lower and upper bound (with respect to a partially ordered set) the true filtered distribution at each time instant. Since $\underline{P}$ and $\bar{P}$ are low rank (say R), the computational cost of evaluating the filtering bounds is $O(XR)$ 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, explicit bounds are given on the variational norm between the true posterior and the upper and lower bounds in terms of the Dobrushin coefficient.**

***Index Terms*—Hidden Markov model filter, stochastic dominance, copositive matrix, nuclear norm minimization, importance sampling filter, Dobrushin coefficient.**

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 $B_{xy} = \mathbb{P}(y_k = y | x_k = x)$ denote the observation likelihood probabilities. With $y_{1:k}$ denoting the sequence of observations from time 1 to k, define the posterior state probability mass function (pmf)

$$\begin{aligned}\pi_k(i) &= \mathbb{P}(x_k = i | y_{1:k}), \quad i \in \{1, 2, \ldots, X\},\\ \pi_k &= [\,\pi_k(1), \ldots, \pi_k(X)\,]'. \end{aligned} \tag{1}$$

Manuscript received June 22, 2014; revised September 16, 2014; accepted October 02, 2014. Date of publication October 13, 2014; date of current version November 06, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Petr Tichavsky. This work was supported in part by NSERC and SSHRC, Canada. Part of this work was completed when V. Krishnamurthy was visiting KTH in 2013.

V. Krishnamurthy is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4 Canada (e-mail: vikramk@ece.ubc.ca).

C. R. Rojas is with the ACCESS Linnaeus Centre, Department of Automatic Control, KTH Royal Institute of Technology, SE 100 44 Stockholm, Sweden (e-mail: crro@kth.se).

Digital Object Identifier 10.1109/TSP.2014.2362886

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

$$\pi_{k+1} = T(\pi_k, y_{k+1}; P) \triangleq \frac{B_{y_{k+1}} P' \pi_k}{\mathbf{1}' B_{y_{k+1}} P' \pi_k}. \tag{2}$$

Here, $B_y = \operatorname{diag}(B_{1y}, \ldots, B_{Xy})$ is a diagonal X-dimensional matrix of observation likelihoods and $\mathbf{1}$ 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 π_{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.

A. 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 $\underline{P}$ and $\bar{P}$, such that $\underline{P} \preceq P \preceq \bar{P}$. Here $\preceq$ denotes a *copositive ordering* defined in Section II. The Bayesian filters using $\underline{P}$ and $\bar{P}$, are guaranteed to sandwich the true posterior distribution π_k at any time k as

$$T(\pi_{k-1}, y_k; \underline{P}) \leq_r T(\pi_{k-1}, y_k; P) \leq_r T(\pi_{k-1}, y_k; \bar{P}) \tag{3}$$

where $T(\cdot)$ denotes the filtering recursion (2) and $\leq_r$ denotes monotone likelihood ratio (MLR) stochastic dominance defined in Section II. What (3) says is that at any time k, the true posterior $\pi_k = T(\pi_{k-1}, y_k, P)$ can be sandwiched in the partially ordered set specified by the above stochastic dominance constraints. Moreover, if P is a TP2 matrix[1], then (3) can be globalized to say that if $\underline{\pi}_0 \leq_r \pi_0 \leq_r \bar{\pi}_0$, then

$$\underline{\pi}_k \leq_r \pi_k \leq_r \bar{\pi}_k, \quad \text{for all time } k \tag{4}$$

where $\underline{\pi}_k$ and $\bar{\pi}_k$ denote posteriors computed using $\underline{P}$ and $\bar{P}$.

The MLR stochastic order $\leq_r$ used in (3) and (4) is a partial order on the set of distributions. A crucial property of the MLR order is that it is closed under conditional expectations. This

[1]TP2 matrices are defined in Definition 3.

makes it very useful in Bayesian estimation [3]–[6]. An important consequence of the sandwich result (4) is that the conditional mean state estimates are sandwiched as $\underline{x}_k \leq \hat{x}_k \leq \bar{x}_k$ for all time k. Indeed, the second and all higher moments also are sandwiched.

Finally, in Section II-D we generalize the above result to multivariate HMMs by using the multivariate TP2 stochastic order. Such multivariate HMMs provide a useful example of large scale HMMs. The TP2 order was pioneered by Karlin [7], see also Whitt's classic paper [8].

2) Construction of Low Rank Transition Matrices via Nuclear Norm Minimization: Section III uses state-of-the-art convex optimization methods to construct low rank transition matrices $\underline{P}$ and $\bar{P}$. A low rank R ensures that the lower and upper bounds to the posterior can be computed with $O(RX)$ rather than $O(X^2)$ computational cost. The transition matrices $\underline{P}$ 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 and serves as a convex surrogate of the rank of a matrix [9]. The construction of low rank transition matrices $\underline{P}$ and $\bar{P}$ is formulated as a convex optimization problem on the *cone* of copositive matrices.[2] These computations are performed offline without affecting the computational cost of the real time filtering algorithm.

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 (3) or (4) 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 $\hat{\pi}_k$ lies in the partially ordered set $\underline{\pi}_k \leq_r \hat{\pi}_k \leq_r \bar{\pi}_k$ 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 $\underline{\pi}_k$ and $\bar{\pi}_k$.

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

[2]A symmetric matrix M is *copositive* if $\pi' M \pi \geq 0$ for all positive vectors π. (Thus the set of positive definite matrices is a subset of the set of copositive matrices. In this paper, π are probability mass function vectors.)

B. 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 [10], [11] 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 contributions 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. Such results have been developed in the context of stochastic control and Bayesian games in [5], [12], [3], [13], [4] 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 [9]. Inspired by the seminal work of Candès and Tao [14], 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 [15], [16].

There has been extensive work in signal processing on posterior Cramér-Rao bounds for nonlinear filtering [17]; see also [18] 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 $\underline{P}$ and $\bar{P}$ are constructed such that $\underline{P} \preceq P \preceq \bar{P}$ (in terms of a copositive ordering), the filtered estimates computed using $\underline{P}$ 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 $\underline{P}$ 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 π_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 $x_{1:k-1}, y_{1:k-1}$. For the case $y_k \in \{1, 2, \ldots, Y\}$, denote the observation likelihood probabilities as $B_{xy} = \mathbb{P}(y_k = y | x_k = x)$. For 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 π_k lives in an $X-1$ dimensional unit simplex Π comprising of X-dimensional probability vectors π. That is,

$$\Pi \triangleq \{\pi \in \mathbb{R}^X : \mathbf{1}'\pi = 1, \ \pi(i) \geq 0\}. \tag{5}$$

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)

$$\hat{x}_k \triangleq \mathbb{E}\{x_k | y_{0:k}; P\} = g'\pi_k, \ \text{where } g = [1, 2, \ldots, X]'. \tag{6}$$

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

$$\hat{x}_k^{\mathrm{MAP}} \triangleq \mathrm{argmax}_{i \in \{1,2,\ldots,X\}} \pi_k(i). \tag{7}$$

We refer to the posterior π_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 [19]. Recall that Π is the unit simplex defined in (5).

Definition 1 (Monotone Likelihood Ratio (MLR) Dominance): Let $\pi_1, \pi_2 \in \Pi$ be any two probability vectors. Then π_1 is greater than π_2 with respect to the MLR ordering—denoted as $\pi_1 \geq_r \pi_2$—if

$$\pi_1(i)\pi_2(j) \leq \pi_2(i)\pi_1(j), \quad i < j, \quad i, j \in \{1, \ldots, X\}. \tag{8}$$

Similarly $\pi_1 \leq_r \pi_2$ if $\leq$ in (8) is replaced by a $\geq$.
The MLR stochastic order is useful since it is closed under conditional expectations. That is, $X \geq_r Y$ implies $\mathbb{E}\{X|\mathcal{F}\} \geq_r \mathbb{E}\{Y|\mathcal{F}\}$ for any two random variables X, Y and sigma-algebra $\mathcal{F}$ [12], [7], [8], [19].

Definition 2 (First Order Stochastic Dominance, [19]): Let $\pi_1, \pi_2 \in \Pi$. Then π_1 first order stochastically dominates π_2—denoted as $\pi_1 \geq_s \pi_2$—if $\sum_{i=j}^X \pi_1(i) \geq \sum_{i=j}^X \pi_2(i)$ for $j = 1, \ldots, X$.

The following result is well known [19]. It says that MLR dominance implies first order stochastic dominance, and it gives a necessary and sufficient condition for stochastic dominance.

Result 1 ([19]): (i) Let $\pi_1, \pi_2 \in \Pi$. Then $\pi_1 \geq_r \pi_2$ implies $\pi_1 \geq_s \pi_2$.
(ii) Let $\mathcal{V}$ 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 $\pi_1 \geq_s \pi_2$ iff for all $v \in \mathcal{V}, v'\pi_1 \geq v'\pi_2$.

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 every subsequent row with respect to the MLR order.

2) Copositivity: The following definitions of copositive ordering of stochastic matrices will be used extensively in the paper. Let $\mathbb{R}_+^X$ denote the subset of $\mathbb{R}^X$ comprised of non-negative vectors—this is termed as the positive orthant.

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 that 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:

$$M^{(m)}(Q, P) \triangleq Q_m P'_{m+1} + P_{m+1} Q'_m - P_m Q'_{m+1} - Q_{m+1} P'_m. \tag{9}$$

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

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

Intuition: It will be proved in Theorem 1 below that

$$Q \preceq P \text{ iff } Q'\pi \leq_r P'\pi \text{ for all } \pi \in \Pi. \tag{10}$$

That is, the ordering of transition matrices $Q \preceq P$ is equivalent to the MLR ordering $(\leq_r)$ of the optimal predictor updates for all $\pi \in \Pi$. Moreover, (10) is equivalent to the optimal filter updates satisfying $T(\pi, y; Q) \leq_r T(\pi, y; P)$ for any observation y and posterior π. In other words the $\preceq$ ordering of transition matrices preserves the MLR ordering $\leq_r$ of posterior distributions computed via the optimal filter. 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 $\preceq$ is a partial order over the set of stochastic matrices, i.e., $\preceq$ 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 (2) 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 $\underline{P}$ and $\bar{P}$ such that the filtering recursion using these matrices form lower and upper bounds to π_k in the MLR stochastic dominance sense. Due to the low rank of $\underline{P}$ and $\bar{P}$, the cost involved in computing these lower and upper bounds to π_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 $\underline{P}$ and $\bar{P}$ instead of the original transition matrix P, we need further notation to distinguish between the posteriors and estimates computed using $P, \underline{P}$ and $\bar{P}$. Let

$$\underbrace{\pi_{k+1} = T(\pi_k, y_{k+1}; P)}_{\text{optimal}}, \quad \underbrace{\bar{\pi}_{k+1} = T(\bar{\pi}_k, y_{k+1}; \bar{P})}_{\text{upper}}, \quad \underbrace{\underline{\pi}_{k+1} = T(\underline{\pi}_k, y_{k+1}; \underline{P})}_{\text{lower}}$$

denote the posterior updated using the optimal filter (2) with transition matrices $P, \bar{P}$ and $\underline{P}$, respectively. Also, as in (6), with $g = (1, 2, \ldots, X)'$, the conditional mean estimates of the underlying state computed using $\underline{P}$ and $\bar{P}$, respectively, will be denoted as

$$\begin{aligned} \underline{x}_k &\triangleq \mathbb{E}\{x_k | y_{0:k}; \underline{P}\} = g'\underline{\pi}_k, \\ \bar{x}_k &\triangleq \mathbb{E}\{x_k \,|\, y_{0:k}; \bar{P}\} = g'\bar{\pi}_k. \end{aligned} \tag{11}$$

In analogy to (7), denote the "hard" MAP state estimates computed using $\underline{P}$ and $\bar{P}$ as

$$\underline{x}^{\text{MAP}}{}_k \triangleq \text{argmax}_i \underline{\pi}_k(i), \quad \bar{x}_k^{\text{MAP}} \triangleq \text{argmax}_i \bar{\pi}_k(i). \tag{12}$$

The following is the main result of this section. Recall the definitions 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 $T(\pi, y; \underline{P})$ where $T(\cdot)$ 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 $\underline{P}$ and $\bar{P}$ such that $\underline{P} \preceq P \preceq \bar{P}$ (recall $\preceq$ is defined in Definition 5).
2) Suppose transition matrices $\underline{P}$ and $\bar{P}$ are constructed such that $\underline{P} \preceq P \preceq \bar{P}$. Then for any y and $\pi \in \Pi$, the filtering updates satisfy the sandwich result
$$T(\pi, y; \underline{P}) \leq_r T(\pi, y; P) \leq_r T(\pi, y; \bar{P}).$$
3) Suppose P is TP2 (Definition 3). Assume the filters $T(\pi, y; P), T(\pi, y; \bar{P})$ and $T(\pi, y; \underline{P})$ are initialized with common prior π_0. Then the posteriors satisfy
$$\underline{\pi}_k \leq_r \pi_k \leq_r \bar{\pi}_k, \quad \text{for all time } k = 1, 2, \ldots$$
As a consequence for all time $k = 1, 2, \ldots,$
   a) The "soft" conditional mean state estimates defined in (6), (11) satisfy $\underline{x}_k \leq \hat{x}_k \leq \bar{x}_k$.
   b) The "hard" MAP state estimates defined in (7), (12) satisfy $\underline{x}^{\text{MAP}}{}_k \leq \hat{x}_k^{\text{MAP}} \leq \bar{x}_k^{\text{MAP}}$.

■

Statement 1 says that for any transition matrix P, there always exist transition matrices $\underline{P}$ and $\bar{P}$ such that $\underline{P} \preceq P \preceq \bar{P}$ (copositivity dominance). Actually if P is TP2, then one can trivially construct the tightest rank 1 bounds $\underline{P}$ and $\bar{P}$ as shown in Section III-A.

Given existence of $\underline{P}$ and $\bar{P}$, the next step is to optimize the choice of $\underline{P}$ and $\bar{P}$—that is the subject of Section III where nuclear norm minimization is used to construct sparse eigenvalue matrices $\underline{P}$ and $\bar{P}$.

Statement 2 says that for any prior π and observation y, the one step filtering updates using $\underline{P}$ and $\bar{P}$ constitute lower and upper bounds to 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 (see Result 1), the conditional mean estimates satisfy $\underline{x}_k \leq \hat{x}_k \leq \bar{x}_k$.

Why MLR Dominance?: The proof of Theorem 1 in the Appendix uses the result that $\pi \leq_r \bar{\pi}$ implies that the filtered update $T(\pi, y; P) \leq_r T(\bar{\pi}, y; P)$. Such a result does not hold with first order stochastic dominance $\leq_s$—that is, $\pi \leq_s \bar{\pi}$ does not imply that $T(\pi, y; P) \leq_s T(\bar{\pi}, y; P)$. In other words, the MLR order is closed with respect to conditional expectations. This the reason why we use the MLR order (and its multivariate generalization called the TP2 order defined below) in this paper.

Examples of TP2 Matrices: Several classes of transition matrices satisfy the TP2 property, see [20], [21], [7], ([22], pp. 99–100) and Karlin's classic book [23]. They are widely used in structural results in stochastic control [5], [12]. The left-to-right Bakis HMM used in speech recognition [24] has an upper triangular transition matrix which has a TP2 structure under mild conditions, e.g., if the upper triangular elements in row i are $(1 - P_{ii})/(X - i)$ then P is TP2 if $P_{ii} \leq 1/(X - i)$.

In the numerical examples section, we use the fact that the matrix exponential of any tridiagonal generator matrix is TP2 ([25], pp. 154). TP2 matrices can be constructed systematically starting with the first row and then generating each subsequent row to MLR dominate the previous row. Each new row can be constructed via a LP feasibility test since given the previous row, the elements of the next row lie in a convex polytope.

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 [19], [8], [7]. We consider a highly stylized example which will serve as a reproducible way of constructing large scale HMMs in numerical studies of Section VI.

Consider L independent Markov chains, $x_k^{(l)}, l = 1, 2 \ldots, L$ with transition matrices $A^{(l)}$. Define the joint process

$x_k = (x_k^{(1)}, \ldots, x_k^{(L)})$. Let $\mathbf{i} = (i_1, \ldots, i_L)$ and $\mathbf{j} = (j_1, \ldots, j_L)$ denote the vector indices where each index $i_l, j_l \in \{1, \ldots, X\}, \quad l = 1, \ldots, L$.

Suppose the observation process recorded at a sensor has the conditional probabilities $B_{\mathbf{i},y} = \mathbb{P}(y_k = y | x_k = \mathbf{i})$. 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 $P(x_k | y_{1:k})$. This is equivalent to HMM filtering of the process x_k with transition matrix $P = A^{(1)} \otimes \cdots \otimes A^{(L)}$ where $\otimes$ denotes Kronecker product. 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 (in general, mapping the vector index into a scalar index does not yield univariate distributions that are MLR orderable). We use the totally positive (TP2) stochastic order, which is a natural multivariate generalization of the MLR order [19], [8], [7]. Denote the element-wise minimum and maximum vectors

$$\begin{aligned} \mathbf{i} \wedge \mathbf{j} &= [\min(i_1, j_1), \ldots, \min(i_L, j_L)]', \\ \mathbf{i} \vee \mathbf{j} &= [\max(i_1, j_1), \ldots, \max(i_L, j_L)]'. \end{aligned} \tag{13}$$

Denote the L-variate posterior at time k as

$$\pi_k(\mathbf{i}) = \mathbb{P}\left(x_k^{(1)} = i_1, x_k^{(2)} = i_2, \ldots, x_k^{(L)} = i_L | y_{1:k}\right)$$

and let Π denote the space of all such L-variate posteriors.

Definition 6 (TP2 Ordering[3]): Let π_1 and π_2 denote any two L-variate probability mass functions. Then $\pi_1 \underset{\text{TP2}}{\geq} \pi_2$ if $\pi_1(\mathbf{i})\pi_2(\mathbf{j}) \leq \pi_1(\mathbf{i} \vee \mathbf{j})\pi_2(\mathbf{i} \wedge \mathbf{j})$.
If π_1 and π_2 are univariate, then this definition is equivalent to the MLR ordering $\pi_1 \geq_r \pi_2$. Indeed, just like the MLR order, the TP2 order is closed under conditional expectations [7]. Next define $P'\pi$ such that its $\mathbf{j}$-th component is $\sum_{\mathbf{i}} P_{\mathbf{ij}}\pi(\mathbf{i})$. In analogy to Definition 5 and (10), given two transition matrices $\underline{P}$ and P, we say that

$$P \succeq_M \underline{P}, \text{ if } P'\pi \underset{\text{TP2}}{\geq} \underline{P}'\pi \quad \text{for all } \pi \in \Pi. \tag{14}$$

The main result regarding filtering of multivariate HMMs is as follows:

Theorem 2: Consider an L-variate HMM where each transition matrix satisfies $A^{(l)} \succeq \underline{A}^{(l)}$ for $l = 1, \ldots, L$ (where $\succeq$ is interpreted as in Definition 5). Then
(i) $A^{(1)} \otimes \cdots \otimes A^{(L)} \succeq_M \underline{A}^{(1)} \otimes \cdots \otimes \underline{A}^{(L)}$.
(ii) Theorem 1 holds for the posterior and state estimates with $\geq_r$ replaced by $\underset{\text{TP2}}{\geq}$. ■

[3]In general the TP2 order is not reflexive. A multivariate distribution P is said to be multivariate TP2 (MTP2) if $P \underset{\text{TP2}}{\geq} P$ holds, i.e., $P(\mathbf{i})P(\mathbf{j}) \leq \mathbf{P}(\mathbf{i} \vee \mathbf{j})\mathbf{P}(\mathbf{i} \wedge \mathbf{j})$. Actually this definition of reflexivity also applies to stochastic matrices. That is, if $\mathbf{i}, \mathbf{j} \in \{1, \ldots, X\}$ are scalar indices then MTP2 and TP2 (Definition 3) are identical for a stochastic matrix, see [7].

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

III. Convex Optimization to Compute Low Rank Transition Matrices $\underline{P}, \bar{P}$

It only remains to give algorithms for constructing low rank transition matrices $\underline{P}$ and $\bar{P}$ that yield the lower and upper bounds $\underline{\pi}_k$ and $\bar{\pi}_k$. These involve convex optimization [26], [27] for minimizing the nuclear norm. *The computation of $\underline{P}$ and $\bar{P}$ 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 $\underline{P}$ and $\bar{P}$ 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 $\underline{P}, \bar{P}$ Without Rank Constraint

Given a TP2 matrix P, the transition matrices $\underline{P}$ and $\bar{P}$ such that $\underline{P} \preceq P \preceq \bar{P}$ can be constructed straightforwardly via an LP solver. With $\underline{P}_1, \underline{P}_2, \ldots, \underline{P}_X$ denoting the rows of $\underline{P}$, a sufficient condition for $\underline{P} \preceq P$ is that $\underline{P}_i \leq_r P_1$ for any row i. Hence, the rows $\underline{P}_i$ 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 $\bar{P}$, where it is sufficient to construct $\bar{P}_i \geq_r P_X$.

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

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

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

$$\text{Minimize rank of } X \times X \text{ matrix } \underline{P} \tag{15}$$

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

$$\mathbf{Cons}(\Pi, \underline{P}, m) \equiv \begin{cases} M^{(m)}(\underline{P}, P) \text{ is copositive on } \Pi & \text{(16a)} \\ \|P'\pi - \underline{P}'\pi\|_1 \leq \epsilon \text{ for all } \pi \in \Pi & \text{(16b)} \\ \underline{P} \geq 0, \quad \underline{P}\mathbf{1} = \mathbf{1}. & \text{(16c)} \end{cases}$$

Recall M is defined in (9) and (16a) is equivalent to $\underline{P} \preceq P$. The constraints $\mathbf{Cons}(\Pi, \underline{P}, m)$ are convex in matrix $\underline{P}$, since (16a) and (16c) are linear in the elements of $\underline{P}$, and (16b) is convex (because norms are convex). The constraints (16a), (16c) are exactly the conditions of Theorem 1, namely that $\underline{P}$ is a stochastic matrix satisfying $\underline{P} \preceq P$.

The convex constraint (16b) is equivalent to $\|\underline{P} - P\|_1 \leq \epsilon$, where $\|\cdot\|_1$ denotes the induced 1-norm for matrices.[4]

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

1) The objective (15) is replaced with the reweighted nuclear norm (see Section III-B1 below).
2) Optimization over the copositive cone (16a) is achieved via a sequence of simplicial decompositions (Section III-B2 below).

1) Reweighted Nuclear Norm: Since the rank is a non-convex function of a matrix, direct minimization of the rank (15) is computationally intractable. Instead, we follow the approach developed by Boyd and coworkers [26], [27] to minimize the iteratively reweighted nuclear norm. As mentioned earlier, inspired by Candès and Tao [14], there has been much recent interest in minimizing nuclear norms for constructing matrices with sparse eigenvalue sets or equivalently low rank. Here we compute $\underline{P}, \bar{P}$ by minimizing their nuclear norms subject to copositivity conditions that ensure $\underline{P} \preceq P \preceq \bar{P}$.

Let $\|\cdot\|_*$ denote the nuclear norm, which corresponds to the sum of the singular values of a matrix, The re-weighted nuclear norm minimization proceeds as a *sequence* of convex optimization problems indexed by $n = 0, 1, \ldots$. Initialize $\underline{P}^{(0)} = I$. For $n = 0, 1, \ldots$, compute $X \times X$ matrix

$$\begin{aligned} \underline{P}^{(n+1)} &= \mathrm{argmin}_{\underline{P}} \|\underline{W}_1^{(n)} \underline{P}\, \underline{W}_2^{(n)}\|_* \\ \text{subject to}: \quad & \text{constraints } \mathbf{Cons}(\Pi, \underline{P}, m),\ m = 1, \ldots, X-1 \\ & \text{namely, } (16\text{a}), (16\text{b}), (16\text{c}). \end{aligned} \tag{17}$$

Notice that at iteration $n+1$, the previous estimate, $\underline{P}^{(n)}$ appears in the cost function of (17) in terms of weighting matrices $\underline{W}_1^{(n)}, \underline{W}_2^{(n)}$. These weighting matrices are evaluated iteratively as

$$\begin{aligned} \underline{W}_1^{(n+1)} &= \left(\left[\underline{W}_1^{(n)}\right]^{-1} U \Sigma U^T \left[\underline{W}_1^{(n)}\right]^{-1} + \delta I \right)^{-1/2}, \\ \underline{W}_2^{(n+1)} &= \left(\left[\underline{W}_2^{(n)}\right]^{-1} V \Sigma V^T \left[\underline{W}_2^{(n)}\right]^{-1} + \delta I \right)^{-1/2} \end{aligned} \tag{18}$$

Here $\underline{W}_1^{(n)} \underline{P}^{(n)} \underline{W}_2^{(n)} = U \Sigma V^T$ is a reduced singular value decomposition, starting with $\underline{W}_1^{(0)} = \underline{W}_2^{(0)} = I$ and $\underline{P}^0 = P$. Also δ is a small positive constant in the regularization term δI. In numerical examples of Section VI, we used `YALMIP` with `MOSEK` and `CVX` to solve the above convex optimization problem.

The intuition behind the reweighting iterations is that as the estimates $\underline{P}^{(n)}$ converge to the limit $\underline{P}^{(\infty)}$, the cost function becomes approximately equal to the rank of $\underline{P}^{(\infty)}$.

2) Simplicial Decomposition for Copositive Programming: Problem (17) is a convex optimization problem in $\underline{P}$. However, one additional issue needs to be resolved: the constraints (16a) involve a copositive cone and cannot be solved directly by standard interior point methods. To deal with the copositive constraints (16a), we use the state-of-the-art simplicial decomposition method detailed in [16]. The nice key idea used in [16] is summarized in the following proposition.

Proposition 1 ([16]): Let Λ denote any sub-simplex of the belief space Π, with vertices $v_1, \ldots, v_X$. Then a sufficient condition for copositive condition (16a) to hold on Λ is that $v_i' M^{(m)}(\underline{P}, P) v_j \geq 0$ for all pairs of vertices v_i, v_j. ■

Proposition 1 allows us to replace the constraints (16a)–(16c) with constraints of the type

$$\overline{\mathbf{Cons}}(\Pi, \underline{P}, m) \equiv \begin{cases} v_i' M^{(m)}(\underline{P}, P) v_j \geq 0 \text{ for all vertices } v_i, v_j \text{ of } \Pi \\ \|P'\pi - \underline{P}'\pi\|_1 \leq \epsilon \text{ for all } \pi \in \Pi \\ \underline{P} \geq 0, \quad \underline{P}\mathbf{1} = \mathbf{1}. \end{cases}$$

Let $\Lambda_j \triangleq \{\Lambda_1^J, \ldots, \Lambda_L^J\}$ denote the set of subsimplices at iteration J that constitute a partition of Π, i.e., for each $i = 1, \ldots, L, \Lambda_i^J = \mathrm{conv}\{v_1^{i,J}, \ldots, v_X^{i,J}\}$ is the convex hull of the vertices $v_1^{i,J}, \ldots, v_X^{i,J}$. 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 (17), $n = 1, 2, \ldots$ with constraints $\overline{\mathbf{Cons}}(\Lambda_1^J, \underline{P}, m), \overline{\mathbf{Cons}}(\Lambda_2^J, \underline{P}, m), \ldots, \overline{\mathbf{Cons}}(\Lambda_L^J, \underline{P}, m), m = 1, \ldots, X-1$.
2) **if** nuclear norm $\|\underline{W}_1^{(n)} \underline{P}\, \underline{W}_2^{(n)}\|_*$ decreases compared to that in iteration $J-1$ by more than a pre-defined tolerance, partition Λ_J into Λ_{J+1} as detailed in [16]: Among all subsimplices and pairs of vertices, choose the pair $(v_r^{i,J}, v_s^{i,J})$ satisfying $(v_r^{i,J})' M^{(m)}(\underline{P}, P) v_s^{i,J} = 0$ with the longest distance between them and pick the two subsimplices that share the edge formed by these vertices, $(\Lambda_i^J, \Lambda_j^J)$. Then, subdivide these subsimplices along the midpoint $w = (v_r^{i,J} + v_s^{i,J})/2$ of this edge, i.e., if $\Lambda_i^J = \mathrm{conv}\{v_1^{i,J}, \ldots, v_r^{i,J}, \ldots, v_s^{i,J}, \ldots, v_X^{i,J}\}$, replace it in Λ_{J+1} by the two new subsimplices $\mathrm{conv}\{v_1^{i,J}, \ldots, w, \ldots, v_s^{i,J}, \ldots, v_X^{i,J}\}$ and $\mathrm{conv}\{v_1^{i,J}, \ldots, v_r^{i,J}, \ldots, w, \ldots, v_X^{i,J}\}$, and similarly for Λ_j^J.
   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 $\|\underline{W}_1^{(n)} \underline{P}\, \underline{W}_2^{(n)}\|_*$, 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 [16] for details on simplex partitioning. We emphasize again that the algorithms in this section for computing $\underline{P}$ and $\bar{P}$ 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

[4]The three statements $\|P'\pi - \underline{P}'\pi\|_1 \leq \epsilon$, $\|\underline{P} - P\|_1 \leq \epsilon$ and $\sum_{i=1}^X \|(P' - \underline{P}')_{:,i}\|_1 \pi(i) \leq \epsilon$ are all equivalent since $\|\pi\|_1 = 1$ because π is a probability vector (pmf)

sample path of the optimal filter to the partially ordered set $\underline{\pi}_k \leq_r \pi_k \leq_r \bar{\pi}_k$ at each time k. The next question is: Given these estimates $\underline{\pi}_k$ and $\bar{\pi}_k$, how to construct an algorithm to estimate π_k? That is, how can the reduced complexity bounds $\underline{\pi}_k$ and $\bar{\pi}_k$ be exploited to estimate the posterior π_k? We present a filtering algorithm that is inspired by recent results in stochastic linear solvers [28], [29]. The algorithm uses importance sampling for matrix-vector multiplication together with Gibbs sampling to ensure that the estimated posterior $\hat{\pi}_k$ lies in the partially ordered set $\underline{\pi}_k \leq_r \hat{\pi}_k \leq_r \bar{\pi}_k$. The resulting estimate that exploits the upper and lower bounds has a lower variance than the unconditional estimator since $\operatorname{var}(\hat{\pi}_k | \underline{\pi}_k \leq_r \hat{\pi}_k \leq_r \bar{\pi}_k) \leq \operatorname{var}(\hat{\pi}_k)$.

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 [30] and body area networks [31], [32]. 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. 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.

A. Stochastic Dominance Constrained Importance Sampling Filtering Algorithm

Suppose at time $k-1$, we have an estimate $\hat{\pi}_{k-1}$ of the optimal filtered estimate π_{k-1} together with the reduced complexity bounds $\underline{\pi}_{k-1}$ and $\bar{\pi}_{k-1}$ such that $\underline{\pi}_{k-1} \leq_r \hat{\pi}_{k-1} \leq_r \bar{\pi}_{k-1}$. Algorithm 1 computes an estimate $\hat{\pi}_k$ of the optimal filtered estimate π_k. Steps 0 and 1 were detailed in Sections II and III. Step 2 computes an estimate $\hat{\pi}_{k|k-1}$ of the optimal predictor $\pi_{k|k-1}$ using Monte-Carlo importance sampling for matrix-vector multiplication $P'\hat{\pi}_{k-1}$ so that the constraint $\underline{\pi}_{k|k-1} \leq_r \hat{\pi}_{k|k-1} \leq_r \pi_{k|k-1}$ holds. Step 3 then computes the filtered posterior at time k with $O(X)$ computations as $\hat{\pi}_k \propto B_{y_k} \hat{\pi}_{k|k-1}$ (Bayes rule). By Theorem 1, this updated posterior is guaranteed to satisfy $\underline{\pi}_k \leq_r \hat{\pi}_k \leq_r \bar{\pi}_k$.

Discussion of Step 2 in Algorithm 1: Step 2 computes the components $j = 1, \ldots, X$ of predicted estimate $\hat{\pi}_{k|k-1}$. At the j-th iteration of Step 2, the aim is to simulate samples $\{\hat{\pi}^{(n)}_{k|k-1}(j); n = 1 \ldots, N\}$ from the j-th row of $P'\hat{\pi}_{k-1}$ so that these samples satisfy $\alpha_j \leq \hat{\pi}^{(n)}_{k|k-1}(j) \leq \beta_j$. It is easily seen that this is equivalent to the MLR constraint $\underline{\pi}_{k|k-1} \leq_r \hat{\pi}_{k|k-1} \leq_r \pi_{k|k-1}$ for the j-th component in terms of the $(j-1)$-th component. Building up these samples sequentially for $j = 1, 2, \ldots, X$ constitutes a Gibbs sampling approach to ensure that the constraints hold and is a special case of adaptive importance sampling.[5] Equation (20) together with the if-then statement preceding (21) achieves this objective. Next, the update in (21), namely $\hat{\pi}^{(n)}_{k|k-1}(j) = \frac{P_{i_n, j}\, \hat{\pi}_{k-1}(i_n)}{q_j(i_n)}$, is simply an importance sampling Monte-Carlo estimator for the j-th component of matrix vector product $P'\hat{\pi}_{k-1}$. For example, if the importance pmf q_j is chosen as $\hat{\pi}_{k-1}$, then i_n is simulated from $\hat{\pi}_{k-1}$ and $\hat{\pi}^{(n)}_{k|k-1}(j) = P_{i_n j}$.

The set F_j consists of the indices n of accepted samples $\{\hat{\pi}^{(n)}_{k|k-1}(j); n = 1 \ldots, N\}$ that satisfy the constraints. Finally, (22) is the average of these accepted samples.

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 $\hat{\pi}_k$.

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

Step 1: Evaluate predicted & filtered upper/lower bounds

$$\underline{\pi}_{k|k-1} = \underline{P}' \underline{\pi}_{k-1}, \quad \underline{\pi}_k = \frac{B_{y_k} \underline{\pi}_{k|k-1}}{\mathbf{1}' B_{y_k} \underline{\pi}_{k|k-1}},$$
$$\bar{\pi}_{k|k-1} = \bar{P}' \bar{\pi}_{k-1}, \quad \bar{\pi}_k = \frac{B_{y_k} \bar{\pi}_{k|k-1}}{\mathbf{1}' B_{y_k} \bar{\pi}_{k|k-1}}. \tag{19}$$

Step 2: Compute predicted estimate $\hat{\pi}_{k|k-1}$ using $\underline{\pi}_k, \bar{\pi}_k$.

for $j = 1$ **to** X **do**

Evaluate stochastic dominance path bounds α_j and β_j: $\alpha_1 = 0, \quad \beta_1 = 1$ and for $j > 1$,

$$\alpha_j = \frac{\underline{\pi}_{k|k-1}(j)\, \hat{\pi}_{k|k-1}(j-1)}{\underline{\pi}_{k|k-1}(j-1)},$$
$$\beta_j = \min\left\{ \frac{\bar{\pi}_{k|k-1}(j)\, \hat{\pi}_{k|k-1}(j-1)}{\bar{\pi}_{k|k-1}(j-1)}, 1 - \sum_{i=1}^{j-1} \bar{\pi}_{k|k-1}(i) \right\}. \tag{20}$$

for iterations $n = 1$ **to** N **do**

Sample $i_n \in \{1, \ldots, X\}$ from importance pmf q_j.

if $\frac{P_{i_n, j}\, \hat{\pi}_{k-1}(i_n)}{q(i_n)} \in [\alpha_j, \beta_j]$ **then**

$$\text{Set } F_j = F_j \cup \{n\} \text{ and } \hat{\pi}^{(n)}_{k|k-1}(j) = \frac{P_{i_n, j}\, \hat{\pi}_{k-1}(i_n)}{q_j(i_n)}. \tag{21}$$

end if

end for

Compute estimated predictor as

$$\hat{\pi}_{k|k-1}(j) = \frac{1}{|F_j|} \sum_{n \in F_j} \hat{\pi}^{(n)}_{k|k-1}(j). \tag{22}$$

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

end for

Step 3: Compute filtered estimate $\hat{\pi}_k = \frac{B_{y_k} \hat{\pi}_{k|k-1}}{\mathbf{1}' B_{y_k} \hat{\pi}_{k|k-1}}$.

The key point in Algorithm 1 is the reduced variance compared to the unconstrained estimator since $\operatorname{var}(\hat{\pi}_k | \underline{\pi}_k \leq_r \hat{\pi}_k \leq_r \bar{\pi}_k) \leq$

[5]We thank Eric Moulines and Olivier Cappe of ENST for mentioning this.

$\mathrm{var}(\hat{\pi}_k)$. If the stochastic dominance constraints are not exploited, then instead of (20),

$$\alpha_j = 0 \text{ and } \beta_j = 1 - \sum_{n=1}^{j-1} \bar{\pi}_{k+1|k}(n). \tag{23}$$

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 $q_j(i) = \hat{\pi}_k(i)$, in which case (21) becomes: If $P_{i_n,j} \in [\alpha_j, \beta_j]$ then $\hat{\pi}_{k|k-1}(j) \leftarrow P_{i_{n+1},j}$.
2) The optimal importance function, which minimizes the variance of $\hat{\pi}_{k|k-1}$, is $q_j(i) \propto P_{ij}\,\hat{\pi}_{k-1}(i)$. 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 $q_j(i) \propto \underline{P}_{ij}\hat{\pi}_{k-1}(i)$ or $q_j(i) \propto \bar{P}_{ij}\hat{\pi}_{k-1}(i)$. These have already been computed and therefore no extra computations are required. These are particularly useful when $\underline{P}$ and $\bar{P}$ are constructed to minimize the distance between the bounds and the actual posterior (as discussed in Section III-B).

One can add an optional step below (21) to increase the sampling efficiency—if a particular index i_n does not satisfy the constraint, then there is no need to simulate it again; simulation of this index can be eliminated by setting the corresponding probability $q_j(i_n) = 0$.

Note that Algorithm 1 is not a particle filter. In particular, (21) is simply a Monte-Carlo evaluation of the matrix multiplication $P'\hat{\pi}_{k-1}$ and is motivated by techniques in [29], [28]. Without the sample path constraints, (21) reduces to Algorithm 1 of [28]. Degeneracy issues that plague particle filtering do not arise. For N iterations at each time instant k, Algorithm 1 has $O(X(N+R))$ computational cost where R is the rank of $\underline{P}$. In comparison a particle filter with N particles involves $O(N)$ computational cost.

B. Importance Sampling Filter for Computing Lower Bound

Given the lower bound matrix $\underline{P}$ of rank R, $\underline{\pi}_k$ can be computed exactly using (19) with $O(RX)$ computations. An alternative method is to exploit the rank R and estimate $\underline{\pi}_k$ by using Monte-Carlo importance sampling methods similar to Algorithm 1. Consider the singular value decomposition of $\underline{P}$:

$$\underline{P}'\pi = \sum_{r=1}^{R} \sigma_r v_r u_r' \pi \tag{24}$$

where we have minimized rank R via the nuclear norm minimization algorithm of Section III-B. Algorithm 2 presents the importance sampling filter for $\underline{\pi}_k$ (computing the upper bound is similar).

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 $\underline{\pi}_k \leq_r \hat{\pi}_k \leq_r \bar{\pi}_k$ in the computation of $\hat{\pi}_k$. (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 $\underline{\pi}_k \leq_r \hat{\pi}_k \leq_r \bar{\pi}_k$ involves $O(NX)$ computations. Finally, in the particle filtering folklore, the so called 'optimal' choice for the importance density is $q(x_k|x_{0:k-1}, y_{1:k}) = \mathbb{P}(x_k|x_{k-1}, y_k)$ with particle weight update $w_k^{(n)} = w_{k-1}^{(n)} \sum_{j=1}^{X} P_{x_{k-1}^{(n)} j} B_{j y_k}$. For each particle, this requires $O(X)$ computations and hence $O(NX)$ for N particles.

V. Analysis of Bounds

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

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

for $r = 1$ to R **do**

for iterations $n = 1$ **to** N **do**

Sample $i_n \in \{1, \ldots, X\}$ from importance pmf q_r.

Set $\hat{u}_r(n) = \frac{u_r(n)\,\hat{\underline{\pi}}_{k-1}(i_n)}{q_r(i_n)}$.

end for

Set $\hat{u}_r = \frac{1}{N}\sum_{n=1}^{N} \hat{u}_r(n)$.

end for

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

Compute filtered posterior $\hat{\underline{\pi}}_k = \frac{B_{y_k}\hat{\underline{\pi}}_{k|k-1}}{\mathbf{1}' B_{y_k}\hat{\underline{\pi}}_{k|k-1}}$.

The main result, namely, Theorem 1 above, was an *ordinal* bound: It said that we can compute reduced complexity filters $\underline{\pi}_k$ and $\bar{\pi}_k$ such that the posterior π_k of the original filtering problem is lower and upper bounded on the partially ordered set $\underline{\pi}_k \leq_r \pi_k \leq_r \bar{\pi}_k$ for all time k. Moreover, by minimizing (17), we computed transition matrices $\underline{P}$ and $\bar{P}$ so that $\|P'\pi - \underline{P}'\pi\|_1 \leq \epsilon$ 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 $\|\underline{\pi}_k - \pi_k\|_1$ in terms of ϵ. These bounds together with Theorem 1 give a complete characterization of the reduced complexity filters.

In order to present the main result, we first define the Dobrushin coefficient:

Definition 7 (Dobrushin Coefficient): For a transition matrix P, the Dobrushin coefficient of ergodicity is

$$\rho(P) = \frac{1}{2} \max_{i,j} \sum_{l \in \{1,2,\ldots,X\}} |P_{il} - P_{jl}|. \tag{25}$$

Note that $\rho(P)$ lies in the interval $[0, 1]$. Also $\rho(P) = 0$ implies that the process $\{x_k\}$ is independent and identically distributed (iid). In words: the Dobrushin coefficient of ergodicity $\rho(P)$ is the maximum variational norm[6] between two rows of the transition matrix P.

The following is the main result of this section:

Theorem 3: Consider an HMM with transition matrix P and state levels g. Let $\epsilon > 0$ denote the user defined parameter in constraint (16b) of convex optimization problem (17) and let $\underline{P}$ denote the solution. Then

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

$$\mathbb{E}_y|g'(T(\pi,y;P)-T(\pi,y;\underline{P}))| \leq \epsilon \sum_y \max_{i,j} g'(I-T(\pi,y;\underline{P})\mathbf{1}')B_y(e_i-e_j). \quad (26)$$

Here $\mathbb{E}_y$ is with respect to the measure $\sigma(\pi,y;P) = \mathbf{1}'B_yP\pi = \mathbb{P}(y_k=y|\pi_{k-1}=\pi)$ which is the denominator of the Bayesian filtering update $T(\cdot)$.

2) The sample paths of the filtered posteriors have the following explicit bounds at each time k:

$$\|\pi_k - \underline{\pi}_k\|_1 \leq \frac{\epsilon}{\max\{F(\underline{\pi}_{k-1},y_k)-\epsilon,\, \mu(y_k)\}} + \frac{\rho(\underline{P})\,\|\pi_{k-1}-\underline{\pi}_{k-1}\|_1}{F(\underline{\pi}_{k-1},y_k)} \quad (27)$$

Here $\rho(\underline{P})$ denotes the Dobrushin coefficient of the transition matrix $\underline{P}$ and $\underline{\pi}_k$ is the posterior computed using the HMM filter with $\underline{P}$, and

$$F(\underline{\pi},y) = \frac{\mathbf{1}'B_y\underline{P}'\underline{\pi}}{\max_i B_{i,y}}, \quad \mu(y) = \frac{\min_i B_{iy}}{\max_i B_{iy}}. \quad (28)$$

∎

Theorem 3 gives explicit upper bounds between the filtered distributions using transition matrices $\underline{P}$ and P. Similar bounds hold for $\bar{P}$ and are omitted. The approach used here in terms of the Dobrushin coefficient is similar to that in [1]. We also refer to [33] for stability results on filters involving mismatch between the true system and the approximating filter model. In [34] stability results are presented for filters involving mismatch between the true system and the averaged system for two-time scale systems.

The bounds are useful since their computation involves the reduced complexity filter with transition matrices $\underline{P}$—the original transition matrix P is not used. In numerical examples below, we illustrate (26).

[6]It is conventional to use the variational norm to measure the distance between two probability distributions. Recall that given probability mass functions α and β on $\{1,2,\ldots,X\}$, the variational norm is $\|\alpha-\beta\|_{\mathrm{TV}} = \frac{1}{2}\|\alpha-\beta\|_1 = \frac{1}{2}\sum_{i\in\{1,2,\ldots,X\}}|\alpha(i)-\beta(i)|$. So the variational norm is just half the l_1 norm between two probability mass functions.

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 Section 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 σ_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 $P = A\underbrace{\otimes\cdots\otimes}_{5\text{times}}A$ 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 [7], the Kronecker products of A preserve the TP2 property implying that P is TP2.

Using the above procedure, we constructed a 3125×3125 TP2 transition matrix P as shown in (29) at the bottom of the page.

2) Off-Line Optimization of Lower Bound via Convex Optimization: We used the semidefinite optimization `solvesdp` solver from `MOSEK` with `YALMIP` and `CVX` to solve[7] the convex optimization problem (17) for computing the upper and lower bound transition matrices $\underline{P}$ and $\bar{P}$. To estimate the rank of the resulting transition matrices, we consider the costs (17), which correspond approximately to the number of singular values larger than δ (defined in (18)). The reweighted nuclear norm algorithm is run for 5 iterations, and the simplicial algorithm is stopped as soon as the cost decreased by less than 0.01.

[7]See http://web.cvxr.com/cvx/doc/ for a complete documentation of CVX.

$$Q = \begin{bmatrix} -0.8147 & 0.8147 & 0 & 0 & 0 \\ 0.4529 & -0.5164 & 0.06350 & 0 & \\ 0 & 0.4567 & -0.7729 & 0.3162 & 0 \\ 0 & 0 & 0.0488 & -0.1880 & 0.1392 \\ 0 & 0 & 0 & 0.5469 & -0.5469 \end{bmatrix},$$
$$A = \exp(2Q), \quad P = A\underbrace{\otimes\cdots\otimes}_{5\text{ times}}A. \quad (29)$$

TABLE I

RANKS OF LOWER BOUND TRANSITION MATRICES $\underline{P}$ EACH OF DIMENSION 3125×3125 OBTAINED AS SOLUTIONS OF THE NUCLEAR NORM MINIMIZATION PROBLEM (17) FOR SIX DIFFERENT CHOICES OF ϵ APPEARING IN CONSTRAINT (16B). NOTE $\epsilon = 0$ CORRESPONDS TO $\underline{P} = P$ AND $\epsilon = 2$ CORRESPONDS TO THE IID CASE

ϵ	0	0.4	0.8	1.2	1.6	2
R (rank of $\underline{P}$)	3125 ($\underline{P} = P$)	800	232	165	40	1 (iid)

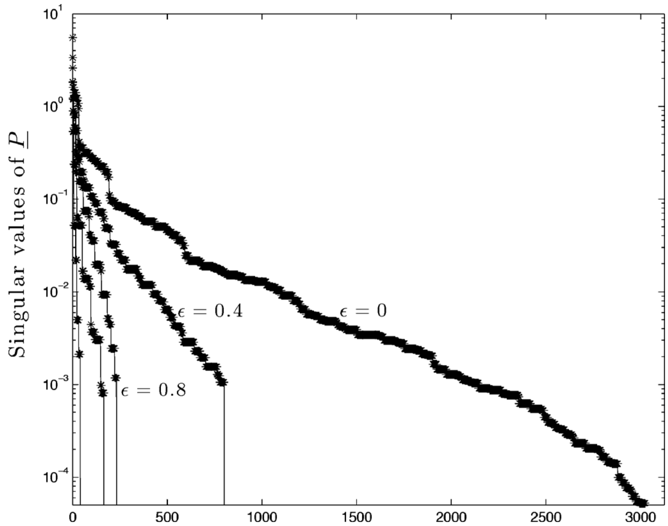


Fig. 1. Plot of 3125 singular values of P and singular values of five different transition matrices $\underline{P}$ parametrized by ϵ in Table I. The transition matrix P (corresponding to $\epsilon = 0$) of dimension 3125×3125 is specified in (29).

To save space we present results only for the lower bounds. We computed[8] 5 different lower bound transition matrices $\underline{P}$ by solving the nuclear norm minimization problem (17) for 5 different choices of $\epsilon \in \{0.4, 0.8, 1.2, 1.6, 2\}$ defined in constraint (16b).

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

3) Performance of Lower Complexity Filters: At each time k, the reduced complexity filter $\underline{\pi}_k = T(\underline{\pi}_{k-1}, y_k; \underline{P})$ incurs computational cost of $O(XR)$ where $X = 3125$ and R is specified in Table I. For each matrix $\underline{P}$ and noise variances σ_v^2 in the range $(0, 2.25]$ we ran the reduced complexity HMM filter $T(\pi, y; \underline{P})$ 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 $\underline{\pi}_k$ provably lower bounds the true posterior π_k as $\underline{\pi}_k \leq_r \pi_k$ for all time k. Therefore the conditional mean estimates satisfy $\underline{x}_k \leq \hat{x}_k$ for all k.

[8]In practice the following preprocessing is required: $\underline{P}$ computed via the semidefinite optimization solver has several singular values close to zero—this is the consequence of nuclear norm minimization. These small singular values need to be truncated exactly to zero thereby resulting in the computational savings associated with the low rank. How to choose the truncation threshold? We truncated those singular values of $\underline{P}$ to zero so that the resulting matrix $\hat{\underline{P}}$ (after rescaling to a stochastic matrix) satisfies the normalized error bound $\frac{\|\underline{P}'\pi - \hat{\underline{P}}\pi\|_2}{\|\underline{P}\|_2 \|\pi\|_2} \leq \frac{\|\hat{\underline{P}} - \underline{P}\|_2}{\|\underline{P}\|_2} \leq 0.01$, i.e., negligible error. The rescaling to a stochastic matrix involved subtracting the minimum element of the matrix (so every element is non-negative) and then normalizing the rows. This transformation does not affect the rank of the matrix thereby maintaining the low rank. For notational convenience, we continue to use $\underline{P}$ instead of $\hat{\underline{P}}$.

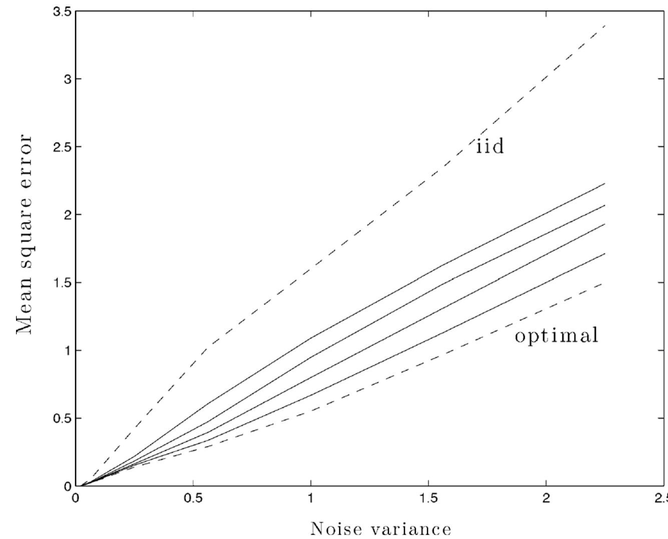


Fig. 2. Mean Square Error of lower bound reduced complexity filters computed using five different transition matrices $\underline{P}$ summarized in Table I. The transition matrix P of dimension 3125×3125 is specified in (29). The four solid lines (lowest to highest curve) are for $\epsilon = 0.4, 0.8, 1.2, 1.6$. The optimal filter corresponds to $\epsilon = 0$, while the iid approximation corresponds to $\epsilon = 2$.

4) Stochastic Dominance Constrained Importance Sampling Algorithm 1: Here we illustrate the performance of three different suboptimal state predictors compared to the optimal predictor, namely

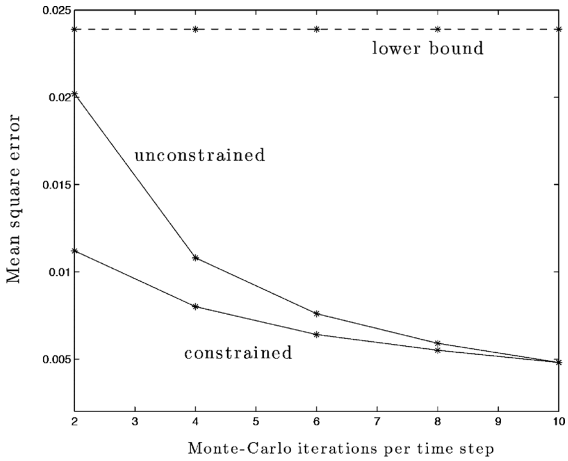
1) Stochastic dominance predictor based on the upper/lower bounds using (22) of Algorithm 1. We ran this for 5 different values of N, namely, 2,4,6,8,10 iterations at each time.
2) Stochastic dominance predictor without exploiting bounds, namely (22), (23), again for the 5 values of N.
3) Reduced complexity lower bound predictor $\underline{\pi}_{k|k-1}$ computed using the lower bound transition matrix $\underline{P}$.

With $\hat{\pi}_{k|k-1}$ denoting any of the above three suboptimal predictors, we computed the mean square error averaged over a million simulations as follows:

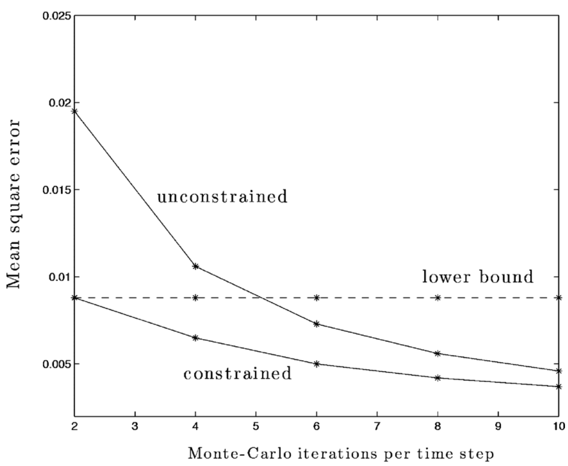
$$\text{MSE} = \frac{1}{10^6} \sum_{m=1}^{10^6} \|\hat{\pi}_{k|k-1} - \pi_{k|k-1}\|^2 \text{ with } \pi_{k-1} = \pi^{(m)}. \tag{30}$$

Here $\pi_{k|k-1}$ denotes the optimal predictor and $\pi^{(m)}$ was sampled uniformly from the $5^5 - 1$ dimensional unit simplex.

Fig. 3(a) and (b) display MSE of all 3 predictors listed above. Fig. 3(a) corresponds to $\epsilon = 2$, resulting in $\underline{P}$ of rank 1. Fig. 3(b) corresponds to $\epsilon = 1.6$, resulting in $\underline{P}$ of rank 40. For the constrained and unconstrained algorithms, naturally, more iterations n per time step yield more accurate estimates and a smaller



(a)



(b)

Fig. 3. Average mean square error (MSE) defined in (30) between optimal predictor and three suboptimal predictors described in Section VI-4, namely the constrained importance sampling predictor (denoted as "constrained"), the unconstrained importance sampling predictor (denoted as "unconstrained") and the reduced complexity lower bound predictor (denoted as "lower bound"). The transition matrix P of dimension 3125×3125 is specified in (29). Recall from Table I that $\epsilon = 2$ corresponds to the iid transition matrix $\underline{P}$ of rank 1, while $\epsilon = 1.6$ corresponds to $\underline{P}$ of rank 40. (a) $\epsilon = 2$, (b) $\epsilon = 1.6$.

MSE. The MSE of the reduced complexity lower bound predictor is displayed with a dashed line. (Recall the performance of the lower bound estimates with these transition matrices were reported in Section VI-3.) The figures show that substantial reductions in the mean square error occur by exploiting the stochastic dominance constraints, even for the iid lower bound case ($\epsilon = 2$).

5) Explicit Bounds: We now illustrate the explicit bound (26). We chose the same 3125 state Markov chain with $\underline{P}, P$ as above and a tridiagonal observation matrix

$$B_{xy} = \begin{cases} b & \text{if } y = x \\ \frac{1}{2}(1-b) & \text{if } y = x-1 \text{ or } y = x+1. \end{cases} \tag{31}$$

We evaluated the right hand side of the bound (26) normalized by $\|g\|_1$ for 5 different choices of $\epsilon \in \{0.4, 0.8, 1.2, 1.6, 2\}$ defined in constraint (16b). (Recall from Table I that these correspond to 5 different choices of $\underline{P}$.) Fig. 4 displays these bounds for three different observation matrices, namely $b = 0.9, b = 0.8$ and $b = 0.5$. The figure shows that the bounds have two properties that are intuitive: First as ϵ get smaller, the approximation $(\underline{P} - P)'\pi$ gets tighter and so one would expect that $\mathbb{E}_y|g'(T(\pi, y; P) - T(\pi, y; \underline{P}))|$ is smaller. This is reflected in the upper bound displayed in the figure. Second, for larger values of b, the "smaller" the noise and so the higher the estimation accuracy. Again the bounds reflect this.

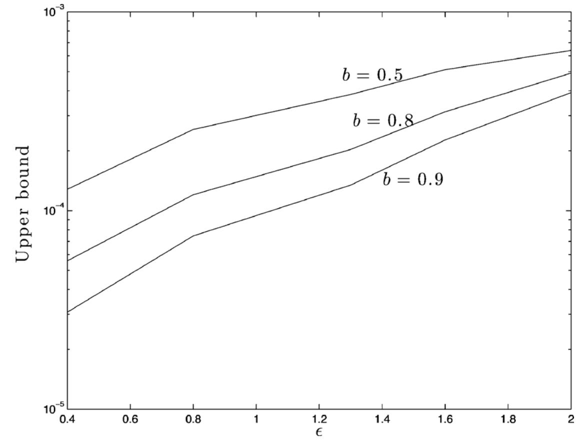


Fig. 4. The "upper bound" in the figure denotes the right hand side of (26) normalized by $\|g\|_1$. The values displayed are for five different values of ϵ corresponding to five different transition matrices $\underline{P}$ whose ranks are given in Table I. The observation matrix parametrized by b is specified in (31).

VII. Discussion

Reduced Complexity Predictors: If one were interested in constructing reduced complexity HMM predictors (instead of filters), the results in this paper are straightforwardly relaxed using first order dominance $\leq_s$ instead of MLR dominance $\leq_r$ as follows: Construct $\underline{P}$ by nuclear norm minimization as in (17), where (16a) is replaced by the linear constraints $\underline{P}_i \leq_s P_i$, on the rows $i = 1, \ldots, X$, and (16b), (16c) hold. Thus the construction of $\underline{P}$ is a standard convex optimization problem and the bound $\underline{P}'\pi \leq_s P'\pi$ holds for the optimal predictor for all $\pi \in \Pi$.

Further, if $\underline{P}$ is chosen so that its rows satisfy the linear constraints $\underline{P}_i \leq_s \underline{P}_{i+1}, i = 1, \ldots, X-1$, then the following global bound holds for the optimal predictor: $(\underline{P}')^k\pi \leq_s (P')^k\pi$ for all time k and $\pi \in \Pi$. A similar result holds for the upper bounds in terms of $\bar{P}$.

It is instructive to compare this with the filtering case, where we imposed a TP2 condition on P for the global bounds (4) to hold wrt $\leq_r$. We could have equivalently imposed a TP2 constraint on $\underline{P}$ and allow P to be arbitrary for the global filtering bounds (4) to hold, however the TP2 constraint is non-convex.

Finally, keep in mind that the predictor bounds in terms of $\leq_s$ do not hold if a filtering update is performed since $\leq_s$ is not closed wrt conditional expectations (unlike $\leq_r$).

Summary: The main idea of the paper was to develop reduced complexity HMM filtering algorithms with provable sample path bounds. At each iteration, the optimal HMM filter has $O(X^2)$ computations and our aim was to derive reduced complexity upper and lower bounds with complexity $O(XR)$ where $R \ll X$. The paper consisted of 4 main results. Theorem 1 showed that one can construct transition matrices $\underline{P}$ and $\bar{P}$ and lower and upper bound beliefs $\underline{\pi}_k$ and $\bar{\pi}_k$ that sandwich the true posterior π_k as $\underline{\pi}_k \leq_r \pi_k \leq_r \bar{\pi}_k$, for all time $k = 1, 2, \ldots$. Theorem 2 generalized this to multivariate TP2 orders. Section III used copositive programming methods to construct low rank transition matrices $\underline{P}$ and $\bar{P}$ of rank R by minimizing the nuclear norm to guarantee $\|\underline{P}'\pi - P'\pi\|_1 \leq \epsilon$ and $\|\bar{P}'\pi - P'\pi\|_1 \leq \epsilon$ over the space of all posteriors Π. Finally, Theorem 3 derived explicit bounds between the optimal estimates and the reduced complexity estimates.

It is interesting that the derivation of MLR stochastic dominance bounds in this paper involves copositivity conditions. There is a rich literature in copositivity including computational aspects [15], [16]. 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.

Appendix

A. Proof of Theorem 1

1. Choose $\underline{P} = [e_1, \ldots, e_1]'$ and $\bar{P} = [e_X, \ldots, e_X]'$ where e_i is the unit X-dimensional vector with 1 in the ith position. Then clearly, $\underline{P} \succeq P \succeq \bar{P}$. These correspond to extreme points on the space of matrices with respect to copositive dominance.

2. We show this in 2 steps. In the first step, we show that for the predictor: $\underline{P} \preceq P \equiv \underline{P}'\pi \leq_r P'\pi$. By definition $\underline{P}'\pi \leq_r P'\pi$ is equivalent to

$$\sum_i \sum_j P_{im} \underline{P}_{j,\bar{m}} \pi_i \pi_j - \sum_{i'} \sum_{j'} \underline{P}_{i'm} P_{j',\bar{m}} \pi_{i'} \pi_{j'} \leq 0$$

for each $\bar{m} > m$. This is equivalent to the above condition holding for $m = 1, \ldots, X-1$ and $\bar{m} = m+1$. This is equivalent to the copositivity ordering $\underline{P} \preceq P$ of Definition 5.

In the second step, we show that the filtered updates satisfy $T(\pi, y; \underline{P}) \leq_r T(\pi, y; P)$. Denote $\underline{\pi} = \underline{P}'\pi$ and $\pi = P'\pi$ and from the first step $\underline{\pi} \leq_r \pi$. It is straightforwardly verified that $\pi \geq_r \underline{\pi}$ implies $\frac{B_y \pi}{\mathbf{1}' B_y \pi} \geq_r \frac{B_y \underline{\pi}}{\mathbf{1}' B_y \underline{\pi}}$ for any observation likelihood B_y. (This crucial property of closure under Bayes' rule makes the MLR stochastic order ideal for this paper.)

3. Suppose $\underline{\pi}_k \leq_r \pi_k$. Then by Statement 2, $T(\underline{\pi}_k, y_{k+1}; \underline{P}) \leq_r T(\underline{\pi}_k, y_{k+1}; P)$. Next since P is TP2, it follows that $\underline{\pi}_k \leq_r \pi_k$ implies $T(\underline{\pi}_k, y_{k+1}; P) \leq_r T(\pi_k, y_{k+1}; P)$. Combining the two inequalities yields $T(\underline{\pi}_k, y_{k+1}; \underline{P}) \leq_r T(\pi_k, y_{k+1}; P)$, or equivalently $\underline{\pi}_{k+1} \leq_r \pi_{k+1}$. Finally, MLR dominance implies first order dominance which by Result 1 implies dominance of means thereby proving 3(a).

To prove 3(b) we need to show that $\underline{\pi} \leq_r \pi$ implies $\arg\max_i \underline{\pi}(i) \leq \arg\max_i \pi(i)$. This is shown by contradiction: Let $i^* = \text{argmax}_i \pi_i$ and $j^* = \text{argmax}_j \underline{\pi}_j$. Suppose $i^* \leq j^*$. Then $\pi \geq_r \underline{\pi}$ implies $\pi(i^*) \leq \frac{\underline{\pi}(i^*)}{\underline{\pi}(j^*)} \pi(j^*)$. Since $\frac{\underline{\pi}(i^*)}{\underline{\pi}(j^*)} \leq 1$, we have $\pi(i^*) \leq \pi(j^*)$ which is a contradiction since i^* is the argmax for $\pi(i)$.

B. Proof of Theorem 2

If suffices to show that $\underline{A} \preceq A \Rightarrow \underline{A} \otimes \underline{A} \preceq_M A \otimes A$. (The proof for repeated Kronecker products then follows by induction.) Consider the TP2 ordering in Definition 6. The indices $\mathbf{i} = (j, n)$ and $\mathbf{j} = (f, g)$ 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 $\mathbf{i} = (j, g-1)$ and $\mathbf{j} = (j-1, g)$, it follows that $\underline{A} \otimes \underline{A} \preceq A \otimes A$ is equivalent to

$$\sum_m \sum_l A_{m,g-1} \underline{A}_{l,g} \times \sum_i \sum_k (A_{ij} \underline{A}_{k,j+1} - A_{i,j+1} \underline{A}_{k,j}) \pi_{im} \pi_{kl} \leq 0$$

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

C. Proof of Theorem 3 and Auxiliary Results

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 using prior π and observation y is

$$\mathcal{B}(\pi, y) = \frac{B_y \pi}{\mathbf{1}' B_y \pi}.$$

Theorem 4: Consider any two posterior probability mass functions $\pi, \tilde{\pi} \in \Pi$. Then:

1) The variational distance in the Bayesian update satisfies

$$\|\mathcal{B}(\pi, y) - \mathcal{B}(\tilde{\pi}, y)\|_{\text{TV}} \leq \frac{\max_i B_{i,y}}{\mathbf{1}' B_y \pi} \|\pi - \tilde{\pi}\|_{\text{TV}}.$$

(Recall that the variational distance is half the l_1 norm).

2) The normalization term in Bayes' rule satisfies

$$\mathbf{1}' B_y \pi \geq \max\{\mathbf{1}' B_y \tilde{\pi} - \|\pi - \tilde{\pi}\|_1 \max_i B_{iy}, \ \min_i B_{iy}\}.$$

Proof: We refer to [1] for a textbook treatment of similar proofs on more general spaces.

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

$$\begin{aligned} g'\mathcal{B}(\pi, y) - \mathcal{B}(\tilde{\pi}, y) &= g'\mathcal{B}(\pi, y) - \frac{B_y \tilde{\pi}}{\mathbf{1}' B_y \pi} + \frac{B_y \tilde{\pi}}{\mathbf{1}' B_y \pi} - \mathcal{B}(\tilde{\pi}, y) \\ &= \frac{1}{\mathbf{1}' B_y \pi} g' \left[I - \mathcal{B}(\tilde{\pi}, y)\mathbf{1}'\right] B_y (\pi - \tilde{\pi}). \end{aligned} \quad (32)$$

Applying the result[9] (see Theorem 5 below for proof) that for any vector $f \in \mathbb{R}^X$,

$$|f'(\pi - \tilde{\pi})| \leq \max_{i,j} |f_i - f_j| \|\pi - \tilde{\pi}\|_{\text{TV}} \quad (33)$$

[9]This inequality is tighter than Holder's inequality which is $|f'(\pi - \tilde{\pi})| \leq 2 \max_i |f_i| \, \|\pi - \tilde{\pi}\|_{\text{TV}}$.

to the right hand side of the above equation yields,

$$|g'(\mathcal{B}(\pi,y)-\mathcal{B}(\tilde{\pi},y))| \leq \frac{1}{\mathbf{1}'B_y\pi}\max_{i,j}|f_i-f_j|\|\pi-\tilde{\pi}\|_{\text{TV}}$$

where $f_i = g'[I-\mathcal{B}(\tilde{\pi},y)\mathbf{1}']B_y e_i$ and $f_j = g'[I-\mathcal{B}(\tilde{\pi},y)\mathbf{1}']B_y e_j$.

So

$$|f_i-f_j| = |g_iB_{i,y}-g'\mathcal{B}(\tilde{\pi},y)B_{i,y} \\ -(g_jB_{j,y}-g'\mathcal{B}(\tilde{\pi},y)B_{j,y})|.$$

Since $\mathcal{B}(\tilde{\pi},y)$ is a probability vector, clearly $|g'\mathcal{B}(\tilde{\pi},y)| \leq \max_i|g_i|$. This together with the fact that $B_{i,y}$ are non-negative implies

$$\max_{i,j}|f_i-f_j| \leq 2\max_i|g_i|\max_i B_{i,y}.$$

So denoting $\|g\|_\infty = \max_i|g_i|$, we have

$$|g'\mathcal{B}(\pi;B_y)-\mathcal{B}(\tilde{\pi};B_y)| \leq 2\frac{\|g\|_\infty\max_i B_{i,y}}{\mathbf{1}'B_y\pi}\|\pi-\tilde{\pi}\|_{\text{TV}}.$$

Finally applying the result that $\|f\|_1 = \max_{\|g\|_\infty=1}|g'f|$ for $g\in\mathbb{R}^X$ (see ([35], pp. 267)), yields

$$\begin{aligned}\|\mathcal{B}(\pi,y)&-\mathcal{B}(\tilde{\pi},y)\|_1\\ &= \max_{\|g\|_\infty=1}|g'\mathcal{B}(\pi,y)-\mathcal{B}(\tilde{\pi},y)|\\ &\leq \max_{\|g\|_\infty=1}2\frac{\|g\|_\infty\max_i B_{i,y}}{\mathbf{1}'B_y\pi}\|\pi-\tilde{\pi}\|_{\text{TV}}.\end{aligned}$$

2) Statement 2: Applying Holder's inequality yields

$$|\mathbf{1}'B_y(\pi-\tilde{\pi})| \leq \|\mathbf{1}'B_y\|_\infty\|\pi-\tilde{\pi}\|_1 = \max_i B_{iy}\,\|\pi-\tilde{\pi}\|_1$$

implying that

$$\mathbf{1}'B_y\pi \geq \mathbf{1}'B_y\tilde{\pi}-\|\pi-\tilde{\pi}\|_1\max_i B_{iy}. \tag{34}$$

Also clearly $\mathbf{1}'B_y\pi \geq \min_i B_{iy}\mathbf{1}'\pi = \min_i B_{iy}$. Combining this with (34) proves the result.

3) Proof of Theorem 3: With the above results we are now ready to prove the theorem.

Part 1: Since $T(\pi_k,y;P)\geq_r T(\pi_k,y;\underline{P})$ and g is a vector with increasing elements, therefore $g'(T(\pi_k,y;P) \geq g'T(\pi_k,y;\underline{P})$. Applying (32) with $\pi = P'\pi_k$ and $\tilde{\pi} = \underline{P}'\pi_k$ yields

$$\begin{aligned}g'(T(\pi_k,y;P)&-T(\pi_k,y;\underline{P}))\\ &= \frac{1}{\sigma(\pi,y;P)}g'\left[I-T(\pi,y,\underline{P})\mathbf{1}'\right]B_y(P-\underline{P})'\pi\end{aligned}$$

where $\sigma(\pi,y;P) = \mathbf{1}'B_yP'\pi$. Then (33) yields

$$\begin{aligned}&g'(T(\pi_k,y;P)-T(\pi_k,y;\underline{P}))\\ &\leq \max_{i,j}\frac{1}{\sigma(\pi,y;P)}g'[I-T(\pi,y,\underline{P})\mathbf{1}']B_y(e_i-e_j)\|P'\pi-\underline{P}'\pi\|_{\text{TV}}\end{aligned}$$

Since $\|P'\pi-\underline{P}'\pi\|_{\text{TV}} \leq \epsilon$, taking expectations with respect to the measure $\sigma(\pi,y;P)$, completes the proof.

Part 2: The triangle inequality for norms yields

$$\begin{aligned}\|\pi_{k+1}&-\underline{\pi}_{k+1}\|_{\text{TV}}\\ &= \|T(\pi_k,y_{k+1};P)-T(\underline{\pi}_k,y_{k+1};\underline{P})\|_{\text{TV}}\\ &\leq \|T(\pi_k,y_{k+1};P)-T(\pi_k,y_{k+1};\underline{P})\|_{\text{TV}}\\ &\quad+\|T(\pi_k,y_{k+1};\underline{P})-T(\underline{\pi}_k,y_{k+1};\underline{P})\|_{\text{TV}}.\end{aligned}\tag{35}$$

Consider the first normed term in the right hand side of (35). Applying Theorem 4(1) with the notation $\pi = P'\pi_k$ and $\tilde{\pi} = \underline{P}'\pi_k$ yields

$$\begin{aligned}\|T(\pi_k,y;P)&-T(\pi_k,y;\underline{P})\|_{\text{TV}}\\ &\leq \frac{\max_i B_{i,y}\|P'\pi_k-\underline{P}'\pi_k\|_{\text{TV}}}{\mathbf{1}'B_y\underline{P}'\pi_k}\\ &\leq \frac{\epsilon}{2}\frac{\max_i B_{i,y}}{\mathbf{1}'B_y\underline{P}'\pi_k}\\ &\leq \frac{\max_i B_{i,y}\,\epsilon/2}{\max\{\mathbf{1}'B_y\underline{P}'\underline{\pi}_k-\epsilon\max_i B_{iy},\min_i B_{iy}\}}.\end{aligned}\tag{36}$$

The second last inequality follows from the construction of $\underline{P}$ satisfying (16b) (recall the variational norm is half the l_1 norm). The last inequality follows from Theorem 4(2).

Consider the second normed term in the right hand side of (35). Applying Theorem 4(i) with notation $\pi = \underline{P}'\pi_k$ and $\tilde{\pi} = \underline{P}'\underline{\pi}_k$ yields

$$\begin{aligned}\|T(\pi_k,y;\underline{P})&-T(\underline{\pi}_k,y;\underline{P})\|_{\text{TV}}\\ &\leq \frac{\max_i B_{i,y}\|\underline{P}'\pi_k-\underline{P}'\underline{\pi}_k\|_{\text{TV}}}{\mathbf{1}'B_y\underline{P}'\underline{\pi}_k}\\ &\leq \frac{\max_i B_{i,y}\,\rho(\underline{P})\,\|\pi_k-\underline{\pi}_k\|_{\text{TV}}}{\mathbf{1}'B_y\underline{P}'\underline{\pi}_k}\end{aligned}\tag{37}$$

where the last inequality follows from the submultiplicative property of the Dobrushin coefficient. Substituting (36) and (37) into the right hand side of the triangle inequality (35) proves the result.

Proof of (33): This is given in the following theorem. See ([1], pp. 93) for a more general setting.

Theorem 5: $|f'(\pi-\tilde{\pi})| \leq \max_{i,j}|f_i-f_j|\|\pi-\tilde{\pi}\|_{\text{TV}}$.

Let $\pi_i^+, i\in\{1,2,\ldots,X\}$ denote those elements of $\pi-\tilde{\pi}$ that are non-negative. Let $\pi_j^-, j\in\{1,2,\ldots,X\}$ denote the magnitude of elements of $\pi-\tilde{\pi}$ that are negative. Then

$$\begin{aligned}|f'(\pi-\tilde{\pi})| &= |f'\pi^+-f'\pi^-|\\ &= \left|\sum_i\sum_j\frac{f_i\pi_i^+\pi_j^-}{\sum_n\pi_n^-}-\sum_i\sum_j\frac{f_j\pi_i^+\pi_j^-}{\sum_m\pi_m^+}\right|\\ &\leq \sum_i\sum_j\left|\frac{f_i}{\sum_n\pi_n^-}-\frac{f_j}{\sum_m\pi_m^+}\right|\pi_i^+\pi_j^-\\ &\leq \max_{i,j}\left|\frac{f_i}{\sum_n\pi_n^-}-\frac{f_j}{\sum_m\pi_m^+}\right|\sum_m\pi_m^+\sum_n\pi_n^-\\ &= \max_{i,j}\left|f_i\sum_m\pi_m^+-f_j\sum_n\pi_n^-\right|.\end{aligned}$$

But $\sum_m\pi_m^+ = \sum_n\pi_n^- = \|\pi-\tilde{\pi}\|_{\text{TV}}$.

ACKNOWLEDGMENT

The authors would like to thank B. Wahlberg at KTH for facilitating the visit of the first author.

REFERENCES

[1] O. Cappe, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models*. Berlin, Germany: Springer-Verlag, 2005.
[2] R. Elliott, L. Aggoun, and J. Moore, *Hidden Markov Models—Estimation and Control*. Berlin, Germany: Springer-Verlag, 1995.
[3] V. Krishnamurthy, "Bayesian sequential detection with phase-distributed change time and nonlinear penalty—A Lattice programming POMDP approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 7096–7124, Oct. 2011.
[4] V. Krishnamurthy, "How to schedule measurements of a noisy Markov chain in decision making?," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 4440–4461, Sep. 2013.
[5] W. Lovejoy, "Some monotonicity results for partially observed Markov decision processes," *Oper. Res.*, vol. 35, no. 5, pp. 736–743, Sep.–Oct. 1987.
[6] W. Whitt, "A note on the influence of the sample on the posterior distribution," *J. Amer. Statist. Assoc.*, vol. 74, pp. 424–426, 1979.
[7] S. Karlin and Y. Rinott, "Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions," *J. Multivar. Anal.*, vol. 10, no. 4, pp. 467–498, Dec. 1980.
[8] W. Whitt, "Multivariate monotone likelihood ratio and uniform conditional stochastic order," *J. Appl. Probab.*, vol. 19, pp. 695–701, 1982.
[9] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1235–1256, 2009.
[10] Q. Zhang, G. Yin, and J. Moore, "Two-time-scale approximation for Wonham filters," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1706–1715, May 2007.
[11] G. Yin, Q. Zhang, J. Moore, and Y. Liu, "Continuous-time tracking algorithms involving two-time-scale Markov chains," *IEEE Trans. Signal Process.*, vol. 53, no. 12, pp. 4442–4452, 2005.
[12] U. Rieder, "Structural results for partially observed control models," *Methods Models Oper. Res.*, vol. 35, no. 6, pp. 473–490, 1991.
[13] V. Krishnamurthy, "Quickest detection POMDPs with social learning: Interaction of local and global decision makers," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5563–5587, Aug. 2012.
[14] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2009.
[15] S. Bundfuss and M. Dür, "Algorithmic copositivity detection by simplicial partition," *Linear Algebra Appl.*, vol. 428, no. 7, pp. 1511–1523, 2008.
[16] S. Bundfuss and M. Dür, "An adaptive linear approximation algorithm for copositive programs," *SIAM J. Optimiz.*, vol. 20, no. 1, pp. 30–53, 2009.
[17] P. Tichavsky, C. Muravchik, and A. Nehorai, "Posterior Cramer-Rao bounds for discrete-time nonlinear filtering," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1386–1396, May 1998.
[18] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Cedar Knolls, NJ, USA: Artech, 2004.
[19] A. Muller and D. Stoyan, *Comparison Methods for Stochastic Models and Risk*. New York, NY, USA: Wiley, 2002.
[20] M. Kijima, *Markov Processes for Stochastic Modelling*. Boca Raton, FL, USA: Chapman and Hall, 1997.
[21] J. Keilson and A. Kester, "Monotone matrices and monotone Markov processes," *Stochastic Processes Appl.*, vol. 5, no. 3, pp. 231–241, 1977.
[22] F. Gantmacher, *Matrix Theory*. New York, NY, USA: Chelsea Pub. Co., 1960, vol. 2.
[23] S. Karlin, *Total Positivity*. Stanford, CA, USA: Stanford Univ. Press, 1968.
[24] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.
[25] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes*. New York, NY, USA: Academic, 1981.
[26] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. Amer. Control Conf. (ACC'01)*, 2001, vol. 6, pp. 4734–4739.
[27] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices," presented at the 2003 Amer. Contr. Conf., 2003.
[28] S. Eriksson-Bique, M. Solbrig, M. Stefanelli, S. Warkentin, R. Abbey, and I. Ipsen, "Importance sampling for a Monte Carlo matrix multiplication algorithm, with application to information retrieval," *SIAM J. Sci. Comput.*, vol. 33, no. 4, pp. 1689–1706, 2011.
[29] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication," *SIAM J. Comput.*, vol. 36, no. 1, pp. 132–157, 2006.
[30] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Cedar Knolls, NJ, USA: Artech House, 1999.
[31] J. Boger, P. Poupart, and J. Hoey, "A decision-theoretic approach to task assistance for persons with dementia," in *Proc. Int. Joint Conf. Artif. Intell.*, 2005, pp. 1293–1299.
[32] M. Pollack, L. Brown, and D. Colbry, "Autominder: An intelligent cognitive orthotic system for people with memory impairment," *Robot. Autonom. Syst.*, vol. 44, pp. 273–282, 2003.
[33] O. Techakesari, J. J. Ford, and D. Nešić, "Practical stability of approximating discrete-time filters with respect to model mismatch," *Automatica*, vol. 48, no. 11, pp. 2965–2970, 2012.
[34] H. Kushner, *Weak Convergence and Singularly Perturbed Stochastic Control and Filtering Problems*. Boston, MA, USA: Birkhauser, 1990.
[35] R. Horn and C. Johnson, *Matrix Analysis*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2012.

Vikram Krishnamurthy (S'90–M'91–SM-99–F'05) received his Ph.D. from the Australian National University in 1992. Currently he is a professor and Canada Research Chair at the University of British Columbia, Vancouver, Canada.

His research interests include statistical signal processing, biomolecular simulation and the dynamics and control of social networks. He served as distinguished lecturer for the IEEE Signal Processing Society and Editor in Chief of IEEE JOURNAL SELECTED TOPICS IN SIGNAL PROCESSING. He received an honorary doctorate from KTH (Royal Institute of Technology), Sweden in 2013.

Cristian R. Rojas (M'13) was born in 1980. He received the M.S. degree in electronics engineering from the Universidad Técnica Federico Santa María, Valparaíso, Chile, in 2004, and the Ph.D. degree in electrical engineering at The University of Newcastle, NSW, Australia, in 2008.

Since October 2008, he has been with the Royal Institute of Technology, Stockholm, Sweden, where he is currently an Assistant Professor of the Automatic Control Lab, School of Electrical Engineering. His research interests lie in system identification and signal processing.