Abstract—This paper develops a generalization of the balanced truncation algorithm applicable to a class of discrete-time stochastic jump linear systems. The approximation error, which is captured by means of the stochastic $L_2$ gain, is bounded from above by twice the sum of singular numbers associated to the truncated states, similar to the case of linear time-invariant systems. A two step model reduction algorithm for hidden Markov models is also developed. The first step relies on the aforementioned balanced truncation algorithm due to a topological equivalence established between hidden Markov models and a subclass of stochastic jump linear systems. In a second step the positivity constraints, which reflect the hidden Markov model structure, are enforced by solving a low dimensional optimization problem.

Index Terms—Balanced truncation, error bound, finite state machines, hidden Markov models, jump systems, model reduction, reduced order systems, stochastic automata, stochastic hybrid systems, stochastic systems.

I. INTRODUCTION

A. Previous Work on Model Reduction

The concept of model reduction is pervasive in all areas where system theoretic ideas have been applied. The objective is to find an adequate simplified model for a given complex system. A particular reduction algorithm is judged upon the certificates it provides concerning the algorithmic cost involved, the level of complexity reduction achieved and the accuracy of the reduced order model.

In the context of linear time-invariant (LTI) systems a distance measure relevant to robustness analysis is the $H_{\infty}$ system norm of the error system between the original and the reduced order model. Balanced truncation and optimal Hankel model reduction are two related reduction algorithms, which are suboptimal, but are accompanied by provable a priori bounds to the aforementioned metric.

Balanced realizations were originally proposed in the controls literature in [48]. Optimal Hankel model reduction was developed in [30] and is inspired by the results in [1]. The derivation of the associated error bounds in continuous and discrete-time settings can be found in [2], [20], [30] and [36]. Both methods are based on the computation of the singular values of the Hankel operator of an LTI system and are frequently termed as singular value decomposition (SVD) based approximation methods [3]. Their strength lies in the guarantees they provide, however they are rather expensive in terms of algorithmic cost. They require computations of the order $O(n^3)$ where $n$ is the number of states of the original system and this prevents their applicability in truly large scale applications.

In the case of balanced truncation one possibility to alleviate the computational burden is to relax the exact balancing requirement and seek approximate solutions to the large scale Lyapunov equations involved, see for instance [33] and [38]. Another way to perform approximate balanced truncation is by using empirical gramians obtained from simulation of the system along prespecified trajectories, [66] and [46]. The use of empirical gramians is inspired by the proper orthogonal decomposition method, [9], which is a popular reduction technique in the computational fluid mechanics community.

Balanced truncation has been investigated also outside the realm of LTI systems. In [7] a generalization to multidimensional and uncertain systems in the linear-fractional framework is presented. Linear parameter-varying systems are investigated in [68] and linear time-varying systems are the subject of [45] and [54].

Apart from the SVD based approximation methods, a widely used class of reduction algorithms for LTI systems was generated by establishing a connection between the Krylov subspace projection methods used in numerical linear algebra and approximation by matching moments of the transfer function, [3], [22], [27], [31], [32], [39] and the references therein. The moment matching based methods can make full use of the sparsity of the original model, and offer the possibility of being implemented utilizing the benefit of matrix vector multiplications, thus they are computationally attractive, suitable for large scale applications. On the other hand they do not provide a priori quality certificates, in terms of guaranteeing stability of the reduced order model or being accompanied by bounds on the $H_{\infty}$ norm of the error system.

B. Previous Work on Stochastic Jump Linear Systems

Jump linear systems (JLS’s) are abstractions of hybrid systems, which combine continuous and discrete dynamics. They form an extension of LTI systems, in the sense that they use state update laws which are linear with respect to the analog state, with matrix coefficients depending on a quantized auxiliary input.
They are often referred to as multi-modal systems, with the transition between the different modes of operation being controlled by a switching signal, which can be interpreted as an exogenous parametric input. In this work it is assumed that the switching signal takes values in a finite set and that it follows an unconstrained stochastic evolution. In particular the parametric input is a sequence of independently identically distributed random variables. Accordingly, the resulting system will be referred to as an i.i.d. jump linear system (iid-JLS).

There is a large body of literature in the fields of statistics, econometrics and control theory pertaining to the class of JLS’s with randomly varying parameters. Fundamental contributions on the stochastic jump linear control problem in a continuous-time setting can be found in [43], [61] and [67]. Subsequently various analysis and synthesis results applicable to linear time-invariant systems have been extended to JLS’s with randomly varying parameters, especially to the class of Markov jump linear systems. A comprehensive review of this material can be found in [50] and the references therein. Stochastic JLS’s arise naturally in connection with networked control systems when the effects of variable sampling rates, link failures, delays and other communication constraints are modeled in a probabilistic framework. A review article on networked control systems is [35].

C. Previous Work on Hidden Markov Models

Hidden Markov Models (HMM’s) are one of the most basic and widespread modeling tools for discrete-time stochastic processes, which take values on a finite alphabet. A comprehensive review paper is [21]. The following account of the literature is heavily biased by the scope of this work.

One of the first references that makes use of the concept of a HMM is [57], where HMM’s with discrete inputs were considered as models for noisy, finite state communication channels. This class of models is commonly referred to as Probabilistic Automata [51]. Another early encounter with HMM’s can be found in [13]. That work provided motivation for the subsequent investigation of the stochastic realization problem. Given a finite valued, stationary process, find necessary and sufficient conditions for it to be equivalent, in the distributional sense, with a function of a Markov chain. Those conditions were derived in [34]. Finite state stochastic processes were associated with algebraic modules and it was shown that the stochastic realization problem is essentially a question of polyhedral convexity. The work in [34] was translated in more transparent system theoretic terms in [52], however both approaches are non constructive. A remedy to that was provided in [4], where the employment of convergence results on infinite products of nonnegative matrices led to a realization algorithm, which is based on asymptotic arguments and is semi-constructive in nature. In that work it was assumed at the outset that the given process has an HMM realization of unknown order. Conditions that are essentially necessary and sufficient and are stated in terms of the given process alone, without making the a priori assumption that it has an HMM realization, were derived recently in [64]. It is worth mentioning that the minimal stochastic realization problem is open up to this date, in fact the simpler problem of minimal realization of linear time-invariant positive systems lacks a complete solution too. A review paper on the positive realization problem of LTI systems is [8].

Applications of HMM’s are met across the spectrum of engineering and science in fields as diverse as speech processing, computational biology and financial econometrics, see for example [42], [53] and [12] respectively. Part of the wide applicability of HMM’s is owed to the fact that a stationary, ergodic, discrete-time, finite valued stochastic process can be expressed as the limit, in a weak convergence sense, of a sequence of HMM’s, see for instance [40]. A multidimensional version of this result in the context of hidden Markov random fields is proven in [44]. Very often though the cardinality of the state space of the underlying Markov chain renders the use of a given HMM for statistical inference or decision making purposes as infeasible, motivating the investigation of possible algorithms that compress the state space without incurring much loss of information. In [65] it was suggested that the concept of approximate lumpability can be used in the context of model reduction of HMM’s. Approximate realization by HMM’s has been studied in conjunction with the nonnegative matrix factorization problem in [24] and further results can be found in [25].

D. Contributions of the Paper

The contributions of this paper are twofold and consist of a balanced truncation algorithm for discrete-time i.i.d. jump linear systems and a model reduction algorithm for discrete-time, finite state, finite alphabet hidden Markov models.

In the case of jump linear systems, the main point of the reduction algorithm is the formulation of two dissipation inequalities, which in conjunction with a suitably defined storage function enable the derivation of a reduced order model with certifiable guarantees of quality. In particular each reduced order model is accompanied by a provable a priori upper bound on the approximation error, which is measured in terms of the stochastic $L_2$ gain. This result can be considered as generalization of the corresponding balanced truncation algorithm for linear time-invariant systems.

In the case of hidden Markov models a two stage reduction algorithm is developed. Initially hidden Markov models are mapped to a certain class of i.i.d. jump linear systems. The image of the high dimensional hidden Markov model in this class is simplified by means of the aforementioned balanced truncation algorithm. In the second stage, the reduced order jump linear system is modified, so that it meets the constraints of an image of a hidden Markov model. This is achieved by solving a low dimensional non convex optimization problem. Numerical simulation results provide evidence that the proposed algorithm computes accurate reduced order hidden Markov models, while achieving a compression of the state space by orders of magnitude.

In addition, to provide algorithmic support for handling the model reduction, a polynomial time algorithm for computing the $L_2$ distance of two HMM’s is provided.

E. Notation and Mathematical Preliminaries

The set of nonnegative integers is denoted by $\mathbb{N}$, the set of positive integers by $\mathbb{Z}_+$ and the set of real numbers by $\mathbb{R}$. For $n \in \mathbb{Z}_+$, let $\mathbb{R}^n$ denote the Euclidean $n$-space. The transpose of a column vector $x \in \mathbb{R}^n$ is $x'$. For $x \in \mathbb{R}^n$ let $|x|^2 = x'x$ denote the square of the Euclidean norm. For $P \in \mathbb{R}^{n \times n} \quad P > 0$ indicate that it is a positive definite matrix and the notation $[x|x]_P$ stands for $x'Px$, the square of the weighted norm of $x \in \mathbb{R}^n$. The set of positive definite matrices is $\mathbb{S}_+$. For $P \in \mathbb{R}^{n \times n}$, $P > 0$ indicates that $P$ is a positive definite matrix.

The contributions of this paper are twofold and consist of a balanced truncation algorithm for discrete-time i.i.d. jump linear systems and a model reduction algorithm for discrete-time, finite state, finite alphabet hidden Markov models.

In the case of jump linear systems, the main point of the reduction algorithm is the formulation of two dissipation inequalities, which in conjunction with a suitably defined storage function enable the derivation of a reduced order model with certifiable guarantees of quality. In particular each reduced order model is accompanied by a provable a priori upper bound on the approximation error, which is measured in terms of the stochastic $L_2$ gain. This result can be considered as generalization of the corresponding balanced truncation algorithm for linear time-invariant systems.

In the case of hidden Markov models a two stage reduction algorithm is developed. Initially hidden Markov models are mapped to a certain class of i.i.d. jump linear systems. The image of the high dimensional hidden Markov model in this class is simplified by means of the aforementioned balanced truncation algorithm. In the second stage, the reduced order jump linear system is modified, so that it meets the constraints of an image of a hidden Markov model. This is achieved by solving a low dimensional non convex optimization problem. Numerical simulation results provide evidence that the proposed algorithm computes accurate reduced order hidden Markov models, while achieving a compression of the state space by orders of magnitude.

In addition, to provide algorithmic support for handling the model reduction, a polynomial time algorithm for computing the $L_2$ distance of two HMM’s is provided.

E. Notation and Mathematical Preliminaries

The set of nonnegative integers is denoted by $\mathbb{N}$, the set of positive integers by $\mathbb{Z}_+$ and the set of real numbers by $\mathbb{R}$. For $n \in \mathbb{Z}_+$, let $\mathbb{R}^n$ denote the Euclidean $n$-space. The transpose of a column vector $x \in \mathbb{R}^n$ is $x'$. For $x \in \mathbb{R}^n$ let $|x|^2 = x'x$ denote the square of the Euclidean norm. For $P \in \mathbb{R}^{n \times n} \quad P > 0$ indicate that it is a positive definite matrix and the notation $[x|x]_P$ stands for $x'Px$, the square of the weighted norm of $x \in \mathbb{R}^n$. The set of positive definite matrices is $\mathbb{S}_+$. For $P \in \mathbb{R}^{n \times n}$, $P > 0$ indicates that $P$ is a positive definite matrix.
**II. Preliminaries on i.i.d. Jump Linear Systems**

### A. System Model

Let \( n, m, q, N \in \mathbb{Z}_+ \), \( \Theta = \{1, \ldots, N\} \). Define an i.i.d. jump linear system \( \mathbf{L} \) as a dynamical system with input \( f(k) \in \mathbb{R}^m \), system mode \( \theta(k) \in \Theta \), state variable \( x(k) \in \mathbb{R}^n \), and output \( y(k) \in \mathbb{R}^q \), related by the state space equations

\[
\begin{align*}
\dot{x}(k+1) &= A[\theta(k)]x(k) + B[\theta(k)]f(k) \\
y(k) &= C[\theta(k)]x(k), \quad k \in \mathbb{N}
\end{align*}
\]

where the state space matrices involved have compatible dimensions. The system mode \( \theta \), which can also be interpreted as an exogenous parametric input, is assumed to be a sequence of independent identically distributed random variables. The statistical description of \( \theta \) is encoded by nonnegative numbers \( p_i = \text{Pr}[\theta(k) = i], i \in \Theta \). The input \( f \) is assumed to be deterministic, i.e., each \( f(k) \) is a random variable with zero variance. A reduced order model candidate is denoted by \( \hat{\mathbf{L}} \) and it is required that it lies in the same class of iid-JLS’s. Its state space representation will be accordingly

\[
\begin{align*}
\dot{x}(k+1) &= \hat{A}[\theta(k)]x(k) + \hat{B}[\theta(k)]f(k) \\
y(k) &= \hat{C}[\theta(k)]x(k), \quad k \in \mathbb{N}
\end{align*}
\]

where \( \dot{x}(k) \in \mathbb{R}^n \) and \( \hat{p} < n \). In order to quantify the fidelity of \( \hat{\mathbf{L}} \), an error system \( E_{\mathbf{L}}\hat{\mathbf{L}} \) is introduced, whose inputs are the common inputs \( f(k), \theta(k) \) of \( \mathbf{L} \) and \( \hat{\mathbf{L}} \) and whose output is the difference of their outputs, namely \( e(k) = y(k) - \hat{y}(k), k \in \mathbb{N} \) (see Fig. 1).

### B. Stability and Stochastic \( L_2 \) Gain

The relevant stability concept to this work is that of mean square stability.

**Definition 2.1.** [17], [18] The iid-JLS \( \mathbf{L} \) with \( f(k) = 0, \forall k \in \mathbb{N} \) is mean square stable, if for every initial condition \( x(0) \in \mathbb{R}^n \)

\[
\mathbb{E}[|x(k)|^2] \rightarrow 0 \text{ as } k \rightarrow \infty.
\]

Introduce the linear operator \( \mathcal{L} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m} \), where

\[
\mathcal{L}[X] = \sum_{i=1}^{N} p_i A[i]X A[i].
\]

Note that \( \mathcal{L} \) is monotonic in its argument in the sense that \( \mathcal{L}[X_1] \geq \mathcal{L}[X_2] \) whenever \( X_1 \geq X_2 \).

**Theorem 2.1:** [50] The following statements are equivalent:

- The iid-JLS \( \mathbf{L} \) is mean square stable.
- For every \( Q \in \mathbb{R}^{n \times n}, Q > 0 \) the linear matrix equation

\[
\mathcal{L}[U] - U = -Q,
\]

has a unique solution \( U = \sum_{k=0}^{\infty} \mathcal{L}^k[Q] \) and \( U > 0 \).
- The spectral radius of \( \mathcal{L} \) is smaller than one, \( r_s[\mathcal{L}] < 1 \).

**Definition 2.2.** [50] The stochastic \( L_2 \) gain of the iid-JLS \( \mathbf{L} \) (denoted by \( \gamma_\mathbf{L} \)) is the non-negative real number defined by

\[
\gamma_\mathbf{L}^2 = \sup_{f \in \mathcal{S}_2} \sum_{k=0}^{\infty} \mathbb{E}[|x(k)|^2],
\]

where the supremum is taken under the assumption \( x(0) = 0 \).

**Lemma 2.1.** [56] Let a positive definite matrix \( G = G^T > 0 \) and a real number \( \gamma > 0 \) be such that

\[
\gamma^2 |f|^2 + V[x] \geq \sum_{i=1}^{N} p_i \left( |C[i]x|^2 + V[A[i]x + B[i]f] \right), \quad \forall x \in \mathbb{R}^n, \forall f \in \mathbb{R}^m
\]

where \( V : \mathbb{R}^n \rightarrow [0, \infty), V[x] = x'Gx \). Then the stochastic \( L_2 \) gain of the iid-JLS system (1) does not exceed \( \gamma \).

**Theorem 2.2.** If the iid-JLS \( \mathbf{L} \) is mean square stable, then its stochastic \( L_2 \) gain is finite.

**Proof:** Lemma 2.1 will be employed in this proof. Note that (4) can be written equivalently in matrix form as

\[
W = \sum_{i=1}^{N} p_i \begin{bmatrix} A[i] & B[i] \\ 0 & I_q \end{bmatrix}^T \begin{bmatrix} G & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A[i] & B[i] \\ 0 & I_q \end{bmatrix} \\
\leq \begin{bmatrix} W_{11} & W_{12} \\ W_{12}' & W_{22} \end{bmatrix} \begin{bmatrix} G & 0 \\ 0 & \gamma^2 I_m \end{bmatrix}
\]

where \( W : \mathbb{R}^n \rightarrow [0, \infty), W[x] = x'Gx \). By taking the Schur complement one obtains the following set of sufficient conditions for (5) to hold:

\[
W_{11} < G \quad \text{and} \quad W_{22} - W_{12}'(W_{11} - G)^{-1}W_{12} < \gamma^2 I_m.
\]

Mean square stability implies existence of \( \hat{P} \in \mathbb{R}^{n \times n}, \hat{P} > 0 \), such that \( \sum_{i=1}^{N} p_i A[i] \hat{P} A[i] - \hat{P} \leq \hat{P} < 0 \). Set \( G = \alpha \hat{P} \) with \( \alpha \in \mathbb{R}, \alpha \geq 1 \). Note that (6) can be satisfied by taking \( \alpha \) large enough. Subsequently note that for a fixed value of \( \alpha \), (7) can always be satisfied by taking \( \gamma \) large enough.
C. Reduction by State Truncation

This is a review of the concept of model reduction by means of state truncation. One starts out with the state space representation of an i.i.d.-JLS $\mathbf{L}$ as in (1) and applies an invertible coordinate transformation $\mathbf{x}(k) = T\hat{x}(k)$. The new state vector $\hat{x}(k)$ is then partitioned as $\hat{x}(k) = [\hat{x}_1(k), \hat{x}_2(k)]^T$, where $\hat{x}_1(k) \in \mathbb{R}^n$ corresponds to the states that are to be retained, $\hat{x}_2(k) \in \mathbb{R}^n$ to the states that are to be removed and $\hat{n} = n - r$. The transformed state space matrices are partitioned accordingly as

$$
\hat{A} [\theta(k)] = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} [\theta(k)], \quad \hat{B} [\theta(k)] = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \end{bmatrix} [\theta(k)]
$$

and the state space representation of $\hat{\mathbf{L}}$ of order $\hat{n}$ is given by (2), where

$$
\hat{A} [\theta(k)] = \hat{A}_{11} [\theta(k)], \quad \hat{B} [\theta(k)] = \hat{B}_1 [\theta(k)]
$$

In order to shorten subsequent notation, it will be convenient to think of the state variable of the reduced system submerged in the original state space, namely $\mathbb{R}^n$. Let $\hat{x}_1(k) = [\hat{x}_1(k), \hat{x}_2(k)]^T \in \mathbb{R}^n$, then in a slight abuse of notation $\hat{\mathbf{L}}$ will also be used to denote the system with state space representation

$$
\hat{x}_1(k+1) = (I_n - E_r) \left( \hat{A} [\theta(k)] \hat{x}_1(k) + \hat{B} [\theta(k)] f(k) \right)
$$

$$
\hat{y}(k) = \hat{C} [\theta(k)] \hat{x}_1(k), \quad k \in \mathbb{N}
$$

(8)

where

$$
E_r = \begin{bmatrix} 0 & 0 \\ 0 & I_r \end{bmatrix} \in \mathbb{R}^{n \times n}.
$$

III. REDUCTION OF I.I.D. JUMP LINEAR SYSTEMS

A. Dissipation Inequalities

The point of departure is a mean square stable i.i.d.-JLS $\mathbf{L}$. The model reduction procedure developed for this class of systems relies on the computation of $U \in \mathbb{R}^{n \times n}$, $U > 0$ and $R \in \mathbb{R}^{n \times n}$, $R > 0$, such that two dissipation inequalities, referred to as the output and input dissipation inequality, are fulfilled. It will be shown in the Appendix that solutions to these inequalities are guaranteed given the mean square stability assumption.

1) Output Dissipation Inequality:

$$
|x(k)|_U^2 \geq \sum_{i=1}^N p_i \left( |A[i]x|_U^2 + |C[i]x|^2 \right), \quad \forall x \in \mathbb{R}^n.
$$

(9)

Using the above relation one can bound the average energy of the output of the system when it is excited by initial conditions only.

Lemma 3.1: Let $T \in \mathbb{Z}_+$ and consider the unforced sequence $\{f(0), \ldots, f(T)\}$ response of $\mathbf{L}$ to the initial condition $x(0) \in \mathbb{R}^n$. For an arbitrary $T_0 \in \mathbb{N}$, such that $T_0 < T$ one has

$$
\sum_{k=0}^T \mathbf{E} \left[ |y(k)|^2 \right] \leq \mathbf{E} \left[ |x(T_0)|_U^2 \right].
$$

Proof: The dissipation inequality (9) implies in the unforced case for arbitrary $x(0) \in \mathbb{R}^n$

$$
\mathbf{E} \left[ |x(k+1)|_U^2 \right] + \mathbf{E} \left[ |y(k)|^2 \right] \leq |x(k)|_U^2.
$$

Taking the expectation on both sides and employing the law of iterated expectations leads to

$$
\mathbf{E} \left[ |x(k+1)|_U^2 \right] + \sum_{k=0}^T \mathbf{E} \left[ |y(k)|^2 \right] \leq \mathbf{E} \left[ |x(T_0)|_U^2 \right].
$$

Sum the above relation from $k = T_0$ to $k = T$ to obtain

$$
\mathbf{E} \left[ |x(T+1)|_U^2 \right] + \sum_{k=0}^T \mathbf{E} \left[ |y(k)|^2 \right] \leq \mathbf{E} \left[ |x(T_0)|_U^2 \right].
$$

Noticing that $\mathbf{E}[|x(T+1)|_U^2] \geq 0$ leads to the desired result.

In the case where $N = 1$, so that $\mathbf{L}$ reduces to an LTI system, relation (9) is satisfied with equality using $U = W_c$, the observability gramian of the system.

2) Input Dissipation Inequality:

$$
|x(k)|_R^2 + |f(k)|^2 \geq \sum_{i=1}^N p_i \left( |A[i]x|_R^2 + |B[i]f|^2 \right), \quad \forall x \in \mathbb{R}^n, \forall f \in \mathbb{R}^m.
$$

(10)

The above relation can be used to bound the average weighted energy of the reachable set of the system when it starts at rest and the input energy is fixed.

Lemma 3.2: Let $T \in \mathbb{Z}_+$ and consider the evolution of $\mathbf{L}$ under the assumption $x(0) = 0$. Then, for an arbitrary input sequence $\{f(0), \ldots, f(T)\}$ one has

$$
\sum_{k=0}^T |f(k)|^2 \geq \mathbf{E} \left[ |x(T+1)|_R^2 \right], \quad \forall f \in \mathbb{R}^m, \quad k \in \{0 \ldots T\}
$$

Proof: Let $x(k) \in \mathbb{R}^n, f(k) \in \mathbb{R}^m$, then (10) implies

$$
\mathbf{E} \left[ |x(k+1)|_R^2 |x(k)|^2 \right] \leq |x(k)|_R^2 + |f(k)|^2.
$$

By applying the law of iterated expectations one gets

$$
\mathbf{E} \left[ |x(k+1)|_R^2 \right] \leq \mathbf{E} \left[ |x(k)|_R^2 + |f(k)|^2 \right], \quad \forall f \in \mathbb{R}^m.
$$

Sum the above inequalities over the interval from $k = 0$ to $k = T$ and use that $x(0) = 0$ to obtain the desired result.

When $N = 1$ and assuming that the resulting LTI system is controllable, relation (10) is satisfied with equality using $R = W_c^{-1}$, where $W_c$ is the controllability gramian of the system.
B. Upper Bound on the Approximation Error

In this section it will be shown how to reduce the order of a given mean square stable system $L$ by means of state truncation and obtain an upper bound on the stochastic $L_2$ gain of the resulting error system $E_{L \leftrightarrow L'}$.

**Theorem 3.1.** Consider a mean square stable system $L$ of order $n$ and a positive definite matrix $W$, such that

$$W = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \beta_1 I_{r_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \beta_n I_{r_n} \end{bmatrix}$$

(11)

where $r_1 + \ldots + r_n = r$. Suppose that the matrix $U = W$ satisfies (9) and $R = W^{-1}$ satisfies (10). Let $\tilde{L}$ be the reduced order model obtained by truncating the last $r$ states of $L$. Then, the stochastic $L_2$ gain of the error system $E_{L \leftrightarrow \tilde{L}}$ is bounded from above by twice the sum of the distinct entries on the diagonal of $\Sigma_2$:

$$\gamma_{E_{L \leftrightarrow \tilde{L}}} \leq 2(\beta_1 + \ldots + \beta_n)$$

(12)

**Proof.** The proof will proceed by successive truncation of the last $r$, $r_{n-1}$, $\ldots$, $r_1$ states. Let $\tilde{L}_k$ denote the reduced system obtained by truncating the last $r_k$ states and $E_{L_k \leftrightarrow \tilde{L}_k}$ the corresponding error system between $\tilde{L}_k$ and $L$. The state space representation of $\tilde{L}_k$ is

$$\hat{x}_k(k+1) = (I_n - E_{r_k}) (A[k] \hat{x}_k(k) + B[k] f(k))$$

$$\hat{y}_k(k) = C[k] \hat{x}_k(k), \quad k \in \mathbb{N},$$

where $\hat{x}_k(k) \in V_{\tilde{L}_k}$, $\{ x \in \mathbb{R}^n \mid x = [x', 0]' \}, \quad z \in \mathbb{R}^{n-r_k}$.

The following signals will shorten subsequent notation:

$$z_k(k) = x(k) + \hat{x}_k(k)$$

$$\delta_k(k) = x(k) - \hat{x}_k(k)$$

$$h_k[k] = A[k] \hat{x}_k(k) + B[k] f(k)$$

$$e_k(k) = y(k) - \hat{y}_k(k), \quad k \in \mathbb{N}.$$

One obtains accordingly

$$z_k(k+1) = A[k] z_k(k) + 2B[k] f(k) - E_{r_k} h_k[k]$$

$$\delta_k(k+1) = A[k] \delta_k(k) + E_{r_k} h_k[k]$$

$$e_k(k) = C[k] \delta_k(k), \quad k \in \mathbb{N}.$$

In a first step it will be shown that

$$\gamma_{E_{L_k \leftrightarrow \tilde{L}_k}} \leq 2\beta_k$$

(13)

According to Lemma 2.1 it is sufficient to find a quadratic storage function $V: \mathbb{R}^n \times V_{\tilde{L}_k} \to \mathbb{R}_+$, such that $V(x, \tilde{x}) \geq 0$, $\forall x \in \mathbb{R}^n$, $\forall \tilde{x} \in V_{\tilde{L}_k}$, $V[0, 0] = 0$ and

$$\sum_{i=1}^N p_i \left| C[i] \delta_i \right|^2 + \Delta V \leq 4\beta^2 i f^2,$$

$$\forall x \in \mathbb{R}^n, \forall \tilde{x} \in V_{\tilde{L}_k}, \forall f \in \mathbb{R}^n,$$

where

$$\Delta V = \sum_{i=1}^N p_i V[x(+), \hat{x}_i(+)] - V[x, \hat{x}_i]$$

$$x(+) = A[i] x + B[i] f$$

$$\hat{x}_i(+) = (I_n - E_{r_i}) (A[i] \hat{x}_i + B[i] f).$$

A quadratic storage function candidate is given by

$$V[x, \hat{x}_i] = \beta^2_i |x + \hat{x}_i|^2_{W^{-1}} + |x - \hat{x}_i|^2_{W} = \beta^2_i |z_i|^2_{W^{-1}} + |\delta_i|^2_W.$$

Let $x \in \mathbb{R}^n$, $\tilde{x} \in V_{\tilde{L}_k}$, $f \in \mathbb{R}^m$, one has

$$\Delta V = \sum_{i=1}^N p_i \left| A[i] \delta_i + E_{r_i} h_k[i] \right|^2$$

$$+ \beta^2_i \sum_{i=1}^N p_i \left| A[i] z_i + 2B[i] f - E_{r_i} h_k[i] \right|^2_{W^{-1}} +$$

$$- \beta^2_i |z_i|^2_{W^{-1}} - |\delta_i|^2_W.$$

Expanding the individual terms in the above expressions, one obtains

$$\Delta V = \sum_{i=1}^N \sum_{i=1}^N p_i \left| A[i] \delta_i \right|^2_{W^{-1}} - |\delta_i|^2_W$$

$$+ \beta^2_i \sum_{i=1}^N p_i \left| A[i] z_i + 2B[i] f - E_{r_i} h_k[i] \right|^2_{W^{-1}}$$

$$+ 2\beta_i \sum_{i=1}^N p_i \left| E_{r_i} h_k[i] \right|^2$$

$$+ 2\beta_i \sum_{i=1}^N p_i \left( E_{r_i} h_k[i] \right)^2$$

$$2 \beta_i \sum_{i=1}^N p_i \left( E_{r_i} h_k[i] \right)^2$$


(15)

Applying the dissipation inequality (9) to the first two terms of (15) gives

$$\sum_{i=1}^N p_i \left| A[i] \delta_i \right|^2_{W^{-1}} - |\delta_i|^2_W \leq \sum_{i=1}^N p_i \left| C[i] \delta_i \right|^2.$$

Applying the dissipation inequality (10) to the second line in (15) gives

$$\beta^2_i \sum_{i=1}^N p_i \left| A[i] z_i + 2B[i] f - A[i] \delta_i \right|^2_{W^{-1}} \leq 4\beta^2_i |f|^2.$$

For the last term of (15) note that $A[i] z_i + 2B[i] f - A[i] \delta_i = 2E_{r_i}[f]$, and that $E_{r_i}^2 = E_{r_i}$. Using the above relations one obtains

$$\Delta V + \sum_{i=1}^N p_i \left| C[i] \delta_i \right|^2 \leq 2\beta_i \sum_{i=1}^N p_i \left| E_{r_i} h_k[i] \right|^2.$$
Since $2 \beta_k \sum_{i=1}^N p_i |E_i| \beta_i [E_i]_R^{2} \geq 0$ relation (14) is satisfied, completing the first part of the proof.

Let $W_{s}$ be the submatrix of $W$ corresponding to the retained states.

$$W_{s} = \begin{bmatrix} \Sigma_{1} & 0 \\ 0 & \Sigma_{2}s \end{bmatrix}, \Sigma_{2}s = \begin{bmatrix} \beta_{1}I_{r_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \beta_{s-1} I_{r_{s-1}} \end{bmatrix}$$

where $r_1, \ldots, r_{s-1} = r - r_s$. Consider the system $\hat{L}_0$ and note that $W_{s}$ and $W_{s}^{-1}$ satisfy the corresponding output and input dissipation inequality respectively. Thus, if the last $r_{s-1}$ states from $\hat{L}_s$ are truncated and if one denotes the resulting system by $\hat{L}_{s-1}$ and the error system between $\hat{L}_s$, $\hat{L}_{s-1}$ by $E_{t_1} \hat{L}_{s-1}$, then by repeating the above argument

$$\gamma_{E_{t_1} \hat{L}_{s-1}} \leq 2 \beta_{s-1}.$$ 

Similarly,

$$\gamma_{E_{t_1-1} \hat{L}_{s-1}} \leq 2 \beta_{s-1} - j, \quad j \in \{1, \ldots, s-2\}.$$ 

In the above notation $\hat{L}_1 = \hat{L}$. The error signal $e(k)$ between the original system $L$ and the final reduced system $\hat{L}$ can be decomposed as $e(k) = e_1(k) + \ldots + e_s(k)$, thus by applying the triangle inequality on the stochastic $L_2$ gains of the corresponding systems, one has

$$\gamma_{E_{t_1} \hat{L}_{s-1}} \leq \gamma_{E_{t_1} \hat{L}_{s-1}} + \ldots + \gamma_{E_{t_1} \hat{L}_{s-1}}$$

and the desired result (12) follows.

C. Computational Considerations

1) Computation of $W$ Given $U$, $R$: Feasibility of (9) and (10) for a mean square stable system $L$ is shown in the Appendix. Subsequent computation of $W$ conformal to the assumptions of theorem 3.1 follows from the standard argument of simultaneous diagonalization of two positive definite matrices, [3]. The next lemma can be verified by a straightforward calculation.

**Lemma 3.3:** Let $U \in \mathbb{R}^{n \times n}$, $R \in \mathbb{R}^{n \times n}$ be positive definite matrices, which satisfy (9) and (10) respectively. There exists an invertible state transformation $x(k) = T \hat{x}(k), T \in \mathbb{R}^{n \times n}$, such that

$$|\hat{x}|_F^2 \geq \sum_{i=1}^{N} p_i \left( |A[i]|_F^2 + |C[i]|_F^2 \right), \forall \hat{x} \in \mathbb{R}^n$$

$$|\hat{x}|_F^2 + |f|_R^2 \geq \sum_{i=1}^{N} p_i \left( |A[i]|_R^2 + |B[i]|_R^2 \right), \forall \hat{x} \in \mathbb{R}^n, \forall f \in \mathbb{R}^m.$$ 

where $\hat{T} = T^T U T = W$, $R = T^T R T = W^{-1}$ and $W$ is diagonal.

**Proof:** Compute a Cholesky factorization $H^{-1} = FF'$ and subsequently an eigenvalue decomposition $F^T U F = HW^2 H'$, where $H H' = I_n$ and $W$ is a positive definite diagonal matrix. The required transformation matrix is given by

$$T = FHW^{-\frac{1}{2}}$$

2) Selecting $U$ and $R$: Assume as before that $U \in \mathbb{R}^{n \times n}$, $R \in \mathbb{R}^{n \times n}$ are positive definite matrices, which satisfy (9) and (10) respectively. Note that $U R^{-1} = T^T W^2 T'$, thus

$$\text{trace}[U^{-1} R] = \sum_{i=1}^{n} \frac{1}{\lambda_i^2}$$

where $\lambda_i$, $i \in \{1, \ldots, n\}$ are the eigenvalues of $W$. Given that the error bound (12) is controlled by the sum of the nonrepeated eigenvalues corresponding to the truncated states, a reasonable objective would be to maximize (17) over all positive definite matrices $U$ and $R$ that satisfy (9) and (10) respectively. Note that in the objective function the smaller eigenvalues corresponding to the states to be truncated are more heavily penalized, which is desirable from an error bound point of view. However this is a nonconvex problem and thus has to be relaxed for sake of computational tractability.

a) Computation of a minimal solution to the output dissipation inequality:

**Lemma 3.4:** The output dissipation inequality possesses a minimal solution.

**Proof:** Let $Q = \sum_{i=1}^{N} p_i C[i] C[i]' \geq 0$. Note that (9) can be written equivalently as

$$L[U] - U \leq -Q.$$ 

Consider also (3) and let $U_\perp$ denote its solution. Subtracting (3) from (18) and by letting $\Delta = U - U_\perp$, one gets

$$L[\Delta] - \Delta = -Q \leq 0.$$ 

Mean square stability implies $r_{\sigma[L]} < 1$ and $\Delta = \sum_{i=0}^{\infty} C[i] Q_{\Delta}$ solves the above Lyapunov like equation. By construction $\Delta \geq 0$ proving the minimality of $U \perp$ among all solutions of (9).

The matrix $U_\perp$ is guaranteed to be positive semidefinite. If in fact $U_{\perp}$ ends up having nontrivial nullspace one can truncate the associated states with zero reduction error. Note also that if the given system is mean square observable [41], which is equivalent with $\exists k \in \mathbb{Z}_+: \sum_{i=1}^{k-1} C[i] > 0$, then $U_\perp$ is strictly positive definite. For a fixed $R$ the smaller $U$ is the larger the objective (17) becomes, thus obtaining a minimal solution $U = U_\perp$ to (18) is desirable from an error bound point of view. The matrix $U_\perp$ can be computed as the limit of the nondecreasing sequence $\{U(k)\}$, where

$$U(k+1) = Q + L[U(k)], U(0) = Q, k \in \mathbb{N}.$$ 

The Convergence to the Fixed Point $U_{\perp}$ is exponential.

b) Computation of a solution to the input dissipation inequality: When it comes to relation (10) or its equivalent form

$$-R + \sum_{i=1}^{N} p_i A[i] R A[i]' + \sum_{i=1}^{N} p_i A[i] R B[i] R B[i]' \leq 0$$

the situation is different. Assuming that $N = 1$ and that the resulting LTI system is controllable, the inverse of the controllability gramian corresponds to a maximal solution of (20). The
gramian itself is computed as the solution to a Lyapunov equation. In the case of an iid-JLS system, where \( N > 1 \), (20) does not necessarily possess a maximal solution. Let \( N = 2 \) and define

\[
R_i = \arg \max_{R \succ 0} \text{trace}(U_i R), \quad i \in \{1, 2\} \tag{21}
\]

subject to (20), where \( U_i > 0, i \in \{1, 2\} \). If there was a maximal solution to (20), then one should have \( R_1 = R_2 \). This not the case though, for instance consider

\[
U_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0.2 & 0.0 \\ 0.3 & 0.5 \end{bmatrix},
\]

\[
U_2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{bmatrix},
\]

\[
B_1 = B_2 = \begin{bmatrix} 0.4 & 0.2 \\ 0.2 & 0.4 \end{bmatrix}, \quad p_1 = p_2 = \frac{1}{2}.
\]

Solving the optimization problem (21) subject to (20) gives

\[
R_1 \approx \begin{bmatrix} 17.4 & -16.9 \\ -16.9 & 22.0 \end{bmatrix}, \quad R_2 \approx \begin{bmatrix} 15.2 & -16.2 \\ -16.2 & 24.9 \end{bmatrix}.
\]

Given the lack of a maximal solution to (20), a reasonable formulation for reduction purposes is to obtain a positive definite matrix \( R \) that maximizes (17) subject to (20).

This optimization problem is a semidefinite program, which is convex and can be solved efficiently using interior point methods [16]. However this step of the reduction algorithm is the limiting factor since the computational cost for obtaining \( R \) is higher than the the matrix product iterations (19) required for computing \( U \).

3) Reducing the Computational Complexity in the Case of a Stronger Stability Condition: In this section a stronger condition than mean square stability is assumed. Given that condition it is shown how the solution to a certain linear matrix equation, which is dual to (3) can be used to obtain a matrix \( R \), which satisfies (10). Solving this linear matrix equation instead of the previously proposed semidefinite program is advantageous from a computational standpoint. Consider a mean square stable iid-JLS \( L \), with state space representation as in (1) and suppose that

\[
\exists \exists P > 0, P \in \mathbb{R}^{n \times n} : \sum_{i=1}^{N} A[i]'P A[i] = P < 0, \tag{22}
\]

The above relation can be interpreted as a stability condition. Note that (22) implies mean square stability, but the opposite is not true. Introduce the linear monotonotic operator \( T : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n} \),

\[
T[X] = \sum_{i=1}^{N} A[i]'X A[i].
\]

Following similar arguments as in theorem 2.1 one can show that (22) is equivalent to \( r_{r_e}[T] < 1 \). Let \( \mathcal{S} : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n} \),

\[
\mathcal{S}[X] = \sum_{i=1}^{N} A[i]'X A[i] '
\]

and note that \( \mathcal{S} \) is the adjoint of \( T \), so \( T^* = \mathcal{S} \) and thus \( r_{r_e}[T] = r_{r_e}[\mathcal{S}] \). Define

\[
\tilde{B} = \sum_{i=1}^{N} B[i]B[i]' + \epsilon I_n > 0
\]

where \( \epsilon \) is a small nonnegative real number to ensure strict positive definiteness of \( \tilde{B} \) and consider the linear matrix equation

\[
\tilde{S}[\tilde{R}] - \tilde{R} = -\tilde{B}.
\]

Given that \( r_{r_e}[\tilde{S}] < 1 \) the unique solution \( \tilde{R} > 0 \) to (23) can be obtained as the limit of the nondecreasing sequence \( \{ \tilde{R}_n(k) \} \), where

\[
\tilde{R}_n(k+1) = \tilde{B} + \tilde{S} [\tilde{R}_n(k) ], \quad \tilde{R}_n(0) = \tilde{B}, k \in \mathbb{N}.
\]

Lemma 3.5: Let \( \tilde{R} \) be the positive definite square root of \( \tilde{R} \), then

\[
R_n = \tilde{R}^{-1} \text{ satisfies (10).}
\]

Proof: From (23) it follows:

\[
A[i]' \tilde{R} A[i]' - \tilde{R} \leq -B[i]' B[i], \quad i \in \Theta.
\]

Let \( R^{1/2} > 0 \) denote the positive definite square root of \( R \), then one has

\[
R^{1/2} A[i]' \tilde{R} R^{1/2} + \tilde{R}^{1/2} B[i]' B[i]' \tilde{R}^{1/2} \leq I_n, i \in \Theta
\]

\[
\left\| \begin{bmatrix} R^{1/2} A[i]' \tilde{R} R^{1/2} & \tilde{R}^{1/2} B[i]' \end{bmatrix} \right\|_2 \leq 1, \quad i \in \Theta.
\]

The adjoint of the operator that appears in the left hand side in the above relation is also a contraction, thus

\[
\left\| \begin{bmatrix} R^{1/2} A[i]' & \tilde{R}^{1/2} B[i] \end{bmatrix} \right\|_2 \leq 1, \quad i \in \Theta
\]

and consequently

\[
\left[ A[i]' \tilde{R}^{1/2} B[i]' \right] \tilde{R}^{-1} \left[ A[i]' \right] B[i] \leq \left[ \begin{bmatrix} \tilde{R}^{-1} & 0 \\ 0 & I_m \end{bmatrix}, i \in \Theta.
\]

Let \( R = \tilde{R}^{-1} \), by multiplying each of the above inequalities by \( p_i, i \in \Theta \), and subsequently summing them up, one obtains (20), which is equivalent to (10).

IV. A NUMERICAL EXAMPLE

To illustrate the model reduction algorithm developed in this paper, consider a network control example based on [55, 56].

A one dimensional platoon consists of \( m + 1 \) vehicles. Let \( x_0 \) denote the position of the lead car and \( x_i, i \in \{1, \ldots, m\} \) denote the position of the \( i \)th follower in the platoon. The spacing error is given by \( e_i(t) = x_{i-1}(t) - x_i(t) - \delta, i \in \{1, \ldots, m\} \), where \( \delta \) is the desired vehicle spacing, which is constant. It is assumed that \( x_0(0) = 0 \) and that there is no initial spacing error, \( e_i(0) = 0, i \in \{1, \ldots, m\} \).

Two control schemes have been designed, whose goal is to achieve disturbance attenuation between the leader motion, which is considered as a reference signal and the spacing error among any two successive followers in the platoon. The assumptions are that every vehicle has the same simple
model of a double integrator with first order actuator dynamics,
\[ X_i(s) = H(s)U_i(s) + i(\delta/s), \quad i \in \{1, \ldots, m\}, \]
where
\[ H(s) = \frac{1}{M s^2(\tau s + 1)} \]
and \( M = 1 \), \( \tau = 0.1 \). The control loop runs at a sampling time of \( T_s = 20\) ms and a zero-order hold is used on the control input of each vehicle.

The first scheme is decentralized, based on local measurements from on-board sensors. Its performance cannot be satisfactory due to fundamental limitations, which have been elaborated in [55]. The second control scheme utilizes information about the lead car and exhibits better performance. However it requires communication between the lead car and the followers \( \{2, \ldots, m\} \), which occurs through a wireless network idealized as a Bernoulli channel. As a simplifying assumption corruption occurs in all communication channels simultaneously, if at all, thus the network can be modeled with two states. The probability of a communication error is denoted by \( p \) and it is further assumed that errors are always detected. If the leader motion is transmitted with an error to the followers \( \{2, \ldots, m\} \), then the first control scheme based on the on-board measurements is implemented for that particular sample. If the communication is error free, then the second control scheme is utilized. The first control scheme, is called predecessor following and the control law is of the form:
\[ U_i(s) = K(s)E_i(s), \quad i \in \{1, \ldots, m\}. \]
The second control scheme, which requires communication, is called predecessor and leader following and is of the form:
\[ U_i(s) = K_p(s)E_i(s) + K_l(s) \left( X_0(s) - X_i(s) - \frac{i\delta}{s} \right), \quad i \in \{1, \ldots, m\}. \]
The control parameters are
\[ K(s) = \frac{2s + 1}{0.05s + 1} \]
and \( K_p(s) = K_l(s) = 0.5K(s) \), note that with this choice of parameters the first follower uses the same control law regardless of the state of the network. For the exposition of this paper, what is important is not the actual control design, but the fact that the closed loop system is an iid-JLS, which can serve the purposes of demonstrating the reduction algorithm.

An example where \( m = 12 \) is considered. The stochastic \( L_2 \) gain from the reference signal \( x_0(t) \) to the spacing error between the last two followers, \( e_{12}(t) \) is calculated for various communication error rates \( p \) based on the original model \( \mathbf{L} \), which has 48 states and the reduced model \( \mathbf{\hat{L}} \), obtained by means of the balanced truncation algorithm, which has 6 states. The semidefinite programs involved in the calculation of the stochastic \( L_2 \) gain and the reduction process are solved using standard tools in MATLAB such as SeDuMi [60] together with YALMIP [47]. The results are depicted in Fig. 2.

Let \( \beta_{E_{LL}} \) denote the upper bound to the approximation error in terms of the stochastic \( L_2 \) gain, namely the right hand side of (12). Fig. 3 depicts the upper bound \( \beta_{E_{LL}} \), as well as the actual realized approximation error \( \gamma_{E_{LL}} \) for various communication error rates \( p \).

V. PRELIMINARIES ON HIDDEN MARKOV MODELS

A. Hidden Markov Models and Their Statistical Description

Hidden Markov Models can be defined in many equivalent ways. The basic definitions and notation introduced in the context of realization theory of HMM’s will be used. One can find them for instance in slightly varying language in [4], [23], [52], [63], [64]. Same goes for the proof of the single lemma in this section.

Let \( \{y(t)\} \) be a discrete-time, stationary stochastic process over some fixed probability space \( \{\Omega, \mathcal{F}, \mathbf{P}\} \), with values on a finite set \( \mathbf{Y} \), which has \( N \in \mathbb{Z}_+ \), \( N \geq 2 \) elements. The set \( \mathbf{Y} \) is called the alphabet and its elements are referred to as letters.
For a given \( Y \), define \( Y^* \) as the set of all finite sequences of elements of \( Y \), including the empty sequence, denoted by \( \emptyset \). The finite sequences of letters are called words or strings, if needed they are surrounded by quotation marks to avoid confusion. The set \( Y^* \), referred to as the language, is equipped with a non-commutative “multiplication” operation defined as concatenation of strings and the identity element is the empty sequence \( \emptyset \), i.e. \( Y^* \) is a monoid.

Let \( v \) be a word, its length is denoted by \(|v|\), and by convention \(|\emptyset| = 0\). The set of all strings of length \( k \in \mathbb{N} \) is denoted by \( Y^k \). The concatenation of \( u \) and \( v \) is written as \( uv \), and \( |uv| = |v| + |u| \). Strings are read from right to left, in the sense that in the expression \( vuv \), \( u \) is followed by \( v \). The strict future of the process after time \( t \) is denoted by \( Y^+ = \{ y(t), y(t+1), \ldots \} \) and \( Y^{-} = \{ y(t-1), y(t-2), \ldots \} \) denotes its past and present. Let \( v = "u_k \ldots u_1" \in Y^* \) the notation \( \{ Y^+ \equiv v \} \) stands for the event \( \{ \omega \in \Omega | y(t+k) = u_k, \ldots, y(t-1) = u_1 \} \), by convention \( \{ Y^+ \equiv \emptyset \} = \Omega \).

**Definition 5.1:** The probability function of the process \( \{ y(t) \} \) is a map \( p : Y^* \rightarrow \mathbb{R}_+ \), where

\[
p[v] = \Pr \left[ Y^+ \equiv v \right], \forall v \in Y^*, \forall t \in \mathbb{Z}.
\]

Note that since the process is stationary, the values of \( p[v] \) in the above definition does not depend on \( t \). It can be readily verified, that the probability function satisfies the properties:

\[
p[\emptyset] = 1
\]

(24)

\[
p[v] \in [0, 1], \forall v \in Y^*
\]

(25)

\[
P[v] = \sum_{u \in Y^k} P[vu], \forall v \in Y^*, k \in \mathbb{N}.
\]

(26)

**Definition 5.2:** Let \( \{ y(t) \} \), \( \{ \tilde{y}(t) \} \) be discrete-time, stationary stochastic processes over the same alphabet \( Y \). The two stochastic processes are equivalent if \( \forall t \in \mathbb{Z}, \forall v \in Y^* \)

\[
\Pr \left[ Y^+ \equiv v \right] = \Pr \left[ \tilde{Y}^+ \equiv v \right].
\]

(27)

According to the definition above the two stochastic processes must only coincide in their probability laws in order to be equivalent. They don’t have to be defined on the same underlying probability space \( \{ \Omega, \mathcal{F}, P \} \). In the context of this work when referring to a stationary stochastic process over the alphabet \( Y \), one is thinking of an equivalence class of processes in the sense of (27). No explicit distinction between the members of the equivalence class is made, the concept of strong realization is not used, its only the statistical description that matters.

**Definition 5.3:** A discrete-time, stationary process \( \{ y(t) \} \) over the alphabet \( Y \) has a realization as a stationary HMM of size \( n \in \mathbb{Z}_+ \), \( n \geq 2 \) if there exists a pair of discrete-time, stationary stochastic processes \( \{ x(t) \}, \{ \tilde{x}(t) \} \) taking values on the finite sets \( X = \{ 1, \ldots, n \} \) and \( Y \) respectively, such that \( \{ y(t) \} \) and \( \{ \tilde{y}(t) \} \) are equivalent, the joint process \( \{ x(t), \tilde{x}(t) \} \) is a Markov process and \( \forall \sigma \in X^*, \forall v \in Y^*, \) the following “splitting property” holds:

\[
\Pr \left[ x^+ \equiv \sigma, \tilde{x}^+ \equiv v | X^- , \tilde{Y}^- \equiv \tilde{v} \right] = \Pr \left[ x^+ \equiv \sigma, Y^+ \equiv v | x(t) \right].
\]

The above definition insures that \( \{ x(t) \} \) is by itself a Markov chain of order \( n \), meaning

\[
\Pr \left[ x^+ \equiv \sigma | X^- \right] = \Pr \left[ x^+ \equiv \sigma | x(t) \right].
\]

It also insures that \( \{ \tilde{y}(t) \} \) is a probabilistic function of the Markov chain \( \{ x(t-1) \} \) in the sense that

\[
\Pr \left[ \tilde{Y}^+ \equiv v | X^- \right] = \Pr \left[ \tilde{Y}^+ \equiv v | x(t) \right].
\]

Consider the map \( M : Y^* \rightarrow \mathbb{R}^{n \times n} \) where

\[
M[v]_{ij} = \Pr \left[ x(t+|v|) = i, \tilde{x}^+ \equiv v | x(t) = j \right], i, j \in X, \forall v \in Y^*, \forall t \in \mathbb{N}.
\]

Note that the state transition matrix of the underlying Markov process \( \{ x(t) \} \) is given by

\[
\Pi = \sum_{v \in Y} M[v],
\]

Let \( 1_n \in \mathbb{R}^n \) be a column vector of one’s. Consider \( \pi \in \mathbb{R}^n_+ \), such that \( \Pi \pi = \pi, \Pi_n \pi = 1 \). The vector \( \pi \) corresponds to an invariant distribution of \( \{ x(t) \} \), which is unique if the Markov process has a single ergodic class. Since the processes \( \{ y(t) \} \) and \( \{ \tilde{y}(t) \} \) are equivalent, one has

\[
\Pr \left[ Y^+ \equiv v \right] = \Pr \left[ \tilde{Y}^+ \equiv v \right], \forall t \in \mathbb{Z}, \forall v \in Y^*.
\]

**Lemma 5.1:** Consider \( k \in \mathbb{N}, v = "v_k v_{k-1} \ldots v_1" \in Y^k \), one has

\[
M[v] = M[v_k] \cdots M[v_1], \quad M[\emptyset] = I_n
\]

(28)

\[
p[v] = \Pi_n M[v] \pi.
\]

The preceding lemma shows that if a given stationary process \( \{ y(t) \} \) over the alphabet \( Y \) has a realization as a stationary HMM of size \( n \), then its probability function is completely encoded by the ordered triple \( H = \{ 1_n, \{ M[v], v \in Y \}, \pi \} \). Accordingly in the following discussion referring to a HMM of size \( n \) will be in terms of the ordered triple \( H = \{ 1_n, \{ M[v], v \in \}

Fig. 4. Eigenvalues of \( W \), logarithmic scale on y axis.
to, thus is denoted by $\mathcal{H}_{n,Y}$.

B. Generalized Automata

The concept of a generalized automaton was formally introduced in [62]. Prior to that, with a slightly different terminology the same objects appeared in connection with quasi-realizations of discrete-time, finite valued stochastic processes of finite rank, see [63] for more information and the references therein. Generalized automata (GA) are equivalent to recognizable Formal Power Series in several noncommuting indeterminates with real coefficients, which have been frequently used in the study of formal languages in theoretical computer science, see for instance [10], [19], [58] and in connection with realization problems of multi-linear, state-affine, uncertain and stochastic systems, see [6], [26], [37], [59] respectively. Consider a finite set $Y$ of $N$ elements, where $N \geq 2$.

Definition 5.4: A generalized automaton of size $n$ over the alphabet $Y$ is defined as an ordered triple $G = (c, \{A[v], v \in Y\}, b)$, where $c \in \mathbb{R}^n$, $A : Y \to \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$.

Some of the references in the literature including [62] incorporate a finite state space $X$ of cardinality $n$ in the definition, thus making $G$ an ordered quadruple. This is not pursued in this work since no explicit use of the state space $X$ is being made. Let $v = v_1 \ldots v_n \in Y^*$, where $k \in \mathbb{N}$, the domain of $A$ is extended from $Y$ to $Y^*$ by means of the homomorphism


Definition 5.5: The word function of $G$ is a map $q_G : Y^* \to \mathbb{R}$, where

$$q_G[v] = c'A[v]b, \quad \forall v \in Y^*.$$ 

The set of all GA of size $n$ over the alphabet $Y$ is denoted by $G_{n,Y}$. A HMM can be interpreted as a generalized automaton. Let $H = (1, \{M[v], v \in Y\}, \pi) \in G_{n,Y}$ then clearly $H \in G_{n,Y}$ with $q_H[v] = p_H[v], \forall v \in Y^*$, thus $H \in G_{n,Y}$.

Definition 5.6: Adapted from [4], [63]. A quasi-realization of a discrete-time, finite rank, finite valued, stationary stochastic process $\{g(t)\}$ over the alphabet $Y$ is a generalized automaton $G = (c, \{A[v], v \in Y\}, b)$ over the same alphabet, whose word function satisfies

$$q[v] = \Pr \left[ Y_t^+ = v \right] = c'A[v]b, \quad \forall v \in Y^*.$$ 

and additionally

$$c' = c' \left( \sum_{v \in Y} A[v] \right) \quad (28)$$

$$b = \left( \sum_{v \in Y} A[v] \right)b. \quad (29)$$

The quasi-realization is minimal if the size of the automaton equals the rank of the given process. Minimal quasi-realizations are also termed as regular. Due to (28), (29), the word function of a quasi-realization is a valid probability function in the sense that it satisfies the axioms (24)–(26). The underlying parameters of the automaton, being arbitrary real numbers, do not necessarily have a probabilistic interpretation. The connection between discrete-time, finite valued, stationary stochastic processes of finite rank and GA, has been long recognized in the literature, essentially in the work of [28]. Quasi-realizations offer several advantages over realizations by HMM’s. There exist linear algebra based constructive algorithms for obtaining a minimal quasi-realization. The rank of the process is a lower bound to the size of a minimal realization by a HMM, thus minimal quasi-realizations are more economical. There are no constructive algorithms for obtaining a minimal realization of a finite rank process by a HMM. Even more so there are finite rank processes, where a realization by a HMM does not exist.

The use of quasi-realizations in problems of classification, prediction and inference can be found in [29], [49]. A polynomial algorithm for checking equivalence between two quasi-realizations is given in [5].

The main drawback of quasi-realizations is related to answering the inverse question: given a generalized automaton does it correspond to a quasi-realization of some discrete-time, finite valued, stationary stochastic processes of finite rank. The constraints (28), (29) can be readily verified, the difficulty lies with checking $q[v] \geq 0, \forall v \in Y^*$. Even in the case where the parameters of a generalized automaton have integer entries, checking nonnegativity of its word function for all strings in the language is an undecidable problem [14]. For the case of a single letter alphabet NP-hardness complexity results can be found in [15]. Of course nonnegativity of the word function can always be verified for finite subsets of the language.

Definition 5.7: Let $T \in \mathbb{Z}_+$, a quasi-realization of horizon $T$ of a discrete-time, finite rank, finite valued, stationary stochastic process $\{g(t)\}$ over the alphabet $Y$ is a generalized automaton $G = (c, \{A[v], v \in Y\}, b)$ over the same alphabet, whose word function satisfies

$$q[v] = \Pr \left[ Y_T^+ = v \right] = c'A[v]b, \quad |v| \leq T,$$

and additionally (28), (29) hold.

VI. MODEL REDUCTION FOR HIDDEN MARKOV MODELS

Consider a stationary HMM $H \in \mathcal{H}_{n,Y}$, where $n \geq 3$. Suppose also that $H$ satisfies the technical assumption (31), which will be further elaborated in the next section. The objective of model reduction is to find a stationary HMM $\tilde{H} = (1, \{M[v], v \in Y\}, \pi) \in \mathcal{H}_{n,Y}$, where $\tilde{n} < n$, such that

$$p_{\tilde{H}}[v] \approx p_H[v], \quad \forall v \in Y^* \quad (30)$$

thus obtaining a more economical but approximate description of the statistics of the original model.

A. A Stability Condition

The language $Y^*$ is countable and one can impose a lexical order on it, for instance the first-lexicial order, $f_l : \mathbb{Z}_+ \to Y^*$. Let accordingly $p : \mathbb{Z}_+ \to [0,1]$, where $p(i) = p_{f_l(i)}, i \in \mathbb{Z}_+$. The assumption in the next lemma can be interpreted as a stability condition, which ensures that $p_{\tilde{H}} \in L_2$. 

Authorized licensed use limited to: MIT Libraries. Downloaded on March 02,2010 at 18:04:21 EST from IEEE Xplore. Restrictions apply.
Lemma 6.1: If $\mathbf{H} \in \mathcal{H}_{n,Y}$ and
\[
\exists P \in \mathbb{R}^{n \times n}, P > 0 : \sum_{y \in Y} M'[y]PM[y] - P < 0 \quad (31)
\]
then $\mathbf{P}_H \in l_2$.

Proof: Let $Q = 1, \mathbf{I}_n$. By virtue of (31) the linear matrix equation
\[
\sum_{y \in Y} M'[y]WM[y] - W = -Q
\]
has the unique solution,
\[
W = \sum_{v \in Y^*} M[v]Q M[v]
\]
and
\[
\pi'W \pi = \sum_{v \in Y^*} \pi[v] P[v] = \sum_{k=1}^{\infty} (\mathbf{P}_H(k))^2 = ||\mathbf{P}_H||^2.
\]

The assumption in lemma 6.1 is not too restrictive. A sufficient condition for (31) to hold is that none of the entries of $M[y]$ take the value of 1 (the maximal possible value), which implies
\[
\gamma \left[ \sum_{y \in Y} M'[y] \otimes M'[y] \right] < 1
\]
and consequently (31). The stationary HMM’s, which satisfy the assumption of the lemma above will be termed as stable and the corresponding subset of $\mathcal{H}_{n,Y}$ will be denoted by $\mathcal{H}^*_{n,Y}$.\]

B. Topological Equivalence Between GA and iid-JLS’s

In relation (30) the distance metric between the two HMM’s was not stated explicitly. A relevant distance metric will be specified in this section. Let $\mathcal{G}^*_{n,Y}$ denote the subset stable GA in $\mathcal{G}_{n,Y}$, in the sense that for any $\mathbf{G} = (c, \{A[v], v \in Y\}, b) \in \mathcal{G}_{n,Y}^*$
\[
\exists P \in \mathbb{R}^{n \times n}, P > 0 : \sum_{y \in Y} A'[y]PA[y] - P < 0. \quad (32)
\]
Let $\mathcal{L}^*_{n,Y}$ stand for the space of SISO mean square stable iid-JLS’s of order $n$, whose discrete mode takes values in $Y = \{1, 2, 3\}$, and
\[
\Pr[\theta(t) = y] = \frac{1}{N}, \forall y \in Y, \forall t \in \mathbb{N}.
\]
A system in $\mathcal{L}^*_{n,Y}$ is encoded by the ordered triple $\mathbf{L} = (c, \{A[v], v \in Y\}, b)$. This notation implies that $\mathbf{L}$ has the state space realization
\[
x(t+1) = A[\theta(t)]x(t) + b f(t) \\
y(t) = c'x(t), \quad t \in \mathbb{N}.
\]

Definition 6.1: The map $\phi_{n,Y} : \mathcal{G}^*_{n,Y} \rightarrow \mathcal{L}^*_{n,Y}$ is defined by
\[
\mathbf{L} = \phi_{n,Y} [\mathbf{G} = (c, \{A[v], v \in Y\}, b)] = (c, \sqrt{N}A[v], v \in Y, b).
\]
The map above is a well defined bijective map since mean square stability of $\phi_{n,Y} [\mathbf{G}] = \mathbf{L}$ is equivalent to (32). Consider $\mathbf{L} = \phi_{n,Y} [\mathbf{G}]$ and $\tilde{\mathbf{L}} = \phi_{n,Y} [\tilde{\mathbf{G}}]$ and suppose that the state space representations of the two jump linear systems are related by a nonsingular state transformation, then it is straightforward to verify that $q_{\mathbf{G}}[v] = q_{\tilde{\mathbf{G}}}[v], \forall v \in Y^*$. Let $\mathcal{L}^*_{Y} = \cup_{n \in \mathbb{Z}_+} \mathcal{L}^*_{n,Y}$ and similarly $\mathcal{G}^*_{n,Y} = \cup_{n \in \mathbb{Z}_+} \mathcal{G}^*_{n,Y}$ and note that both unions are disjoint.

Definition 6.2: The distance between $\mathbf{G} \in \mathcal{G}^*_{n,Y}$ and $\tilde{\mathbf{G}} \in \mathcal{G}^*_{\tilde{n},\tilde{Y}}$ is defined by
\[
d_{\mathcal{G}^*_{n,Y}} [\mathbf{G}, \tilde{\mathbf{G}}] := d_{\mathcal{L}^*_{n,Y}} [\mathbf{L}, \tilde{\mathbf{L}}] = \gamma_{\mathcal{L}^*_{n,Y}}\]
where $\mathbf{L} = \phi_{n,Y} [\mathbf{G}]$ and $\tilde{\mathbf{L}} = \phi_{\tilde{n},\tilde{Y}} [\tilde{\mathbf{G}}]$. The metric defined above can be interpreted as an $L2$ induced metric on the space of stable GA over the alphabet $Y$. Since $\mathcal{H}^*_{n,Y} \subset \mathcal{G}^*_{n,Y} \forall n \in \mathbb{Z}_+$, the metric is automatically induced on the space of stable stationary HMM’s over the alphabet $Y$.

Lemma 6.2: Let $\mathbf{G} \in \mathcal{G}^*_{n,Y}$ and $\tilde{\mathbf{G}} \in \mathcal{G}^*_{\tilde{n},\tilde{Y}}$. One has
\[
\sum_{v \in Y^*} (q_{\mathbf{G}}[v] - q_{\tilde{\mathbf{G}}}[v])^2 \leq \left( d_{\mathcal{G}^*_{n,Y}} [\mathbf{G}, \tilde{\mathbf{G}}] \right)^2
\]
Proof: Let $\mathbf{L} = \phi_{n,Y} [\mathbf{G}]$ and $\tilde{\mathbf{L}} = \phi_{\tilde{n},\tilde{Y}} [\tilde{\mathbf{G}}]$. Set $x(0) = 0$ and apply the input $f = (1,0,0,\ldots)$ to the error system $\mathbf{E}_{\mathcal{L}^*_{n,Y}}$. One obtains then
\[
\sum_{v \in Y^*} (q_{\mathbf{G}}[v] - q_{\tilde{\mathbf{G}}}[v])^2 = \sum_{t=0}^{\infty} \mathbf{E} [\epsilon(t)]^2
\]
and also note that by definition of the stochastic $L2$ gain one has $\sum_{t=0}^{\infty} \mathbf{E} [\epsilon(t)]^2 \leq \gamma_{\mathcal{L}^*_{n,Y}}$ for any unit norm input.\]

C. Problem Statement

Now that a distance metric has been introduced the model reduction problem can be restated as follows. Given a stationary HMM $\mathbf{H} \in \mathcal{H}^*_{n,Y}$, where $n \geq 3$, find a stationary HMM $\tilde{\mathbf{H}} = (1, \{M[v], v \in Y\}, \tilde{\pi}) \in \mathcal{H}^*_{\tilde{n},\tilde{Y}}$, where $\tilde{n} < n$, such that $d_{\mathcal{G}^*_{n,Y}} [\mathbf{H}, \tilde{\mathbf{H}}]$ is minimized. Consider the subset of $\mathcal{L}^*_{n,Y}$ defined as $\mathcal{M}^*_{n,Y} = \phi_{n,Y} [\mathcal{H}^*_{n,Y}]$. Every $\tilde{\mathbf{L}} \in \mathcal{M}^*_{n,Y}$ has state space representation of the form
\[
\hat{x}(t+1) = \sqrt{N}M[\theta(t)] \hat{x}(t) + \hat{\pi} f(t) \\
\hat{y}(t) = 1_{\tilde{n}}' \hat{x}(t), \quad t \in \mathbb{N},
\]
and its state space matrices satisfy the constraints
\[
M[v]_{ij} \geq 0 \quad \forall i, j \in \{1, \ldots, \tilde{N}\}, \forall v \in Y \quad (33)
\]
\[
1_{\tilde{n}}' = 1_{\tilde{n}}' \left( \sum_{v \in Y} M[v] \right) \quad (34)
\]
\[
1 = 1_{\tilde{n}}' \hat{\pi} \quad (35)
\]
\[ \hat{\gamma}_i \geq 0 \quad \forall i \in \{1, \ldots, \hat{N}\} \]  
(36)

\[ \hat{\pi} = \left( \sum_{v \in \mathcal{Y}} \hat{M}[v] \right) \hat{\pi}. \]  
(37)

Let \( \mathbf{L} = \phi_{n, \mathcal{Y}}[\mathbf{H}] \), the model reduction problem is equivalently stated as

\[ \min_{\mathbf{L} \in \mathcal{M}_{n, \mathcal{Y}}} \gamma_{E_{\mathbf{L}}, f} \]

This is a nonlinear program (NLP) and the error system, which appears in the objective function has \( n + \hat{n} \) states, thus making this formulation potentially unwieldy. As a remedy to that the following two step suboptimal reduction algorithm is proposed.

D. The Two Step Reduction Algorithm

1) Step 1: Apply the balanced truncation algorithm developed for iid-JLS to \( \mathbf{L} = \phi_{n, \mathcal{Y}}[\mathbf{H}] \) leading to the reduced order model \( \hat{\mathbf{L}} \in \mathcal{L}_{\hat{n}, \mathcal{Y}} \), where \( \hat{n} < n \) and

\[ \gamma_{E_{\mathbf{L}}, f} \leq \epsilon_1 \]

with \( \epsilon_1 \) being an a priori computable error bound as in (12). It is important to note that the stability assumption (31) corresponds to (22) and therefore one can avoid the SDP formulation for the computation of \( \hat{R} \) and instead obtain \( \hat{R} = R^{-1} \) by solving a linear matrix equation as in (23).

2) Step 2: The constraints, which reflect the HMM structure are enforced in a second step by solving the NLP

\[ \min_{\mathbf{E} \in \mathcal{E}_{\mathbf{H}, \mathcal{Y}}} \sum_{v \in \mathcal{Y}^*} \left( q_{\mathcal{E}}[v] - p_{\mathbf{H}}[v] \right)^2 \]  
(38)

where \( \mathcal{G} = \phi_{\hat{n}, \mathcal{Y}}[\hat{\mathbf{L}}] \). An algorithm on how to solve (38) is provided below for reasons of completeness. The algorithm is motivated by the feasible directions method, see [11]. Introduce the error automaton \( \mathbf{E}_{\mathbf{H}, \mathcal{G}} = \{ c_e, \{ A_e[v], v \in \mathcal{Y} \}, \{ b_e \} \} \) of size \( \hat{n} + \hat{n} \) over the alphabet \( \mathcal{Y} \), where

\[ c_e = \begin{bmatrix} \hat{c} \\ -1_{\hat{n}} \end{bmatrix}, \quad A_e[v] = \begin{bmatrix} A[v] & 0 \\ 0 & \hat{M}[v] \end{bmatrix}, \quad v \in \mathcal{Y}, \quad b_e = \begin{bmatrix} \hat{b} \\ \hat{\pi} \end{bmatrix} \]

and note that \( q_{E_{\mathbf{H}, \mathcal{G}}}[v] = q_{\mathcal{G}}[v] - p_{\mathbf{H}}[v], \forall v \in \mathcal{Y}^* \). The decision parameter set is denoted by \( \mathcal{M} \) and is the set of \( \hat{\mathbf{M}} : \mathcal{Y} \rightarrow \mathbb{R}^{\hat{n} \times \hat{n}} \) such that (33) with strict inequality and (34) hold. Using a strict inequality in (33) ensures that the underlying Markov chain has a single ergodic class and \( \pi \) is uniquely determined by (35), (37). The decision parameter set is then open and convex and every element in that set gives a finite cost. The NLP can be then stated as

\[ \min_{\hat{\mathbf{M}} \in \mathcal{M}} h'_e[Wb_e] \]  
(39)

subject to the constraints

\[ -c_e A_e[v] W A_e[v] - W \]  
(40)

and (35), (37).

At \( t = 0 \) initialization occurs by selecting \( \hat{M}(0) \in \mathcal{M} \) randomly. This can be done by picking all the matrix entries out of a uniform distribution in \([0, 1]\) and then scaling such that (34) is met. Given \( \hat{M}(0), \hat{\pi}(0) \) is determined by the eigenvalue problem (35), (37), \( W(0) \) by solving the linear matrix equation (40) and the initial cost is \( f(0) = h'_e(W(0)b(0)) \).

Given \( \hat{M}(s) \) at instant \( t = s \in \mathbb{N} \), the direction of descent \( \delta \hat{M}(s) \) is determined from the linearized version of the problem. In particular one solves the linear program

\[ \min_{\hat{\mathbf{M}} : \mathcal{Y} \rightarrow \mathbb{R}^{n \times n}} 2h'_e(W(s)\delta b(s) + h'_e(s)\delta W(s)b(s)) \]  
(41)

subject to the constraints

\[ 0 = \sum_{v \in \mathcal{Y}} \delta A'_e[v] W(s) A_e[v] + \sum_{v \in \mathcal{Y}} A'_e[v] W(s)\delta A_e[v] + \sum_{v \in \mathcal{Y}} A'_e[v] \delta W(s) A_e[v] - \delta W(s) \]

\[ \delta \hat{\pi}(s) = \left( \sum_{v \in \mathcal{Y}} \delta \hat{M}(s)[v] \right) \hat{\pi}(s) + \left( \sum_{v \in \mathcal{Y}} \hat{M}(s)[v] \right) \delta \hat{\pi}(s), \]

where

\[ \delta A_e[v] = \begin{bmatrix} 0 \\ 0 \\ \delta \hat{M}(s)[v] \end{bmatrix}, \quad v \in \mathcal{Y}, \delta b_e = \begin{bmatrix} 0 \\ \delta \hat{\pi}(s) \end{bmatrix}. \]

The next feasible point \( \hat{M}(s+1) \) is determined according to

\[ \hat{M}(s+1) = \hat{M}(s) + \alpha(s) \delta \hat{M}(s), \]

where \( \alpha(s) \in [0, 1] \) is the step size. It is selected so that \( f(s+1) < f(s) \). Note that for every possible value of \( \alpha(s) \), \( \delta \hat{M}(s+1) \) is computed from (35), (37), \( W(s+1) \) from (40) and the cost is \( f(s+1) = h'_e(W(s+1)b(s+1)) \). The largest possible step size is determined using bisection.

The overall algorithm terminates once \( f(s) - f(s+1) \) falls below a prespecified threshold.

One can then a posteriori compute

\[ \sum_{v \in \mathcal{Y}^*} (q_{\mathcal{G}}[v] - p_{\mathbf{H}}[v])^2 = \epsilon_2^2. \]

After the first step of the reduction algorithm one has according to lemma 6.2

\[ \sum_{v \in \mathcal{Y}^*} (q_{\mathcal{E}}[v] - p_{\mathbf{H}}[v])^2 \leq \gamma_{E_{\mathbf{L}}, f} \leq \epsilon_1^2 \]

thus by using the triangle inequality one gets

\[ \sum_{v \in \mathcal{Y}^*} (p_{\mathbf{H}}[v] - p_{\mathbf{H}}[v])^2 \leq (\epsilon_1 + \epsilon_2)^2. \]

3) Comments: In typical applications the first step of the reduction algorithm delivers a very accurate approximation of the original model with \( \hat{n} \ll n \), thus making the NLP of the second
step much more manageable in comparison with the NLP formulation in the problem statement of Section VI-C, where the error system has size $n + \tilde{n}$.

One can consider solving in the second step the NLP

$$\min_{\mathbf{L} \in \mathcal{M}_{n,n}} \gamma_{\mathbf{L},T}.$$  \hfill (42)

In such a case one works with the same metric in the first and second step of the reduction algorithm. However this NLP is less tractable than the one suggested in the previous section. The reason is that computation of the cost in (42), which is in terms of the stochastic $L_2$ gain requires the solution of a semidefinite program at each iteration, which is more expensive than solving a single linear matrix equation as described in the previous section.

Instead of searching for a HMM in the second step, one can relax the constraints and search for a generalized automaton, which corresponds to a quasi-realization of finite horizon. This can be done either by solving a NLP or by ensuring that the projection matrices involved in the balanced truncation step lead to a reduced order automaton, which satisfies the constraints (28), (29) as well as well as $q[0] = 1$. The latter constraints can be satisfied by adding one more direction to the dominant eigenspace determined by the balanced truncation algorithm.

Although the whole discussion has been geared towards model reduction of stationary HMM’s, the non stationary case can be handled within the framework of this paper. One just needs to interpret the vector $\pi$ not as an invariant distribution, but as an initial distribution for the underlying Markov process and when it comes to model reduction the constraint (37) can be dropped, the rest remains the same.

VII. A NUMERICAL EXAMPLE

Denote by $S_n$ the set of all possible permutations of the set $\{1, \ldots, n\}$, $n \in \mathbb{Z}_+$. Let $\sigma \in S_n$ and denote by $\Gamma_\sigma$ the corresponding permutation matrix. Set $n = 100$ and pick a $\sigma \in S_n$ at random. Define the matrix $\Sigma = \Gamma_\sigma + \epsilon \Delta$, where $\epsilon = 0.02$ and $\Delta \in \mathbb{R}^{n \times n}$. Each entry of $\Delta$ is drawn from a uniform distribution on the unit interval, $[0, 1]$. Let $w = 1_{n, \Sigma}$ and form the diagonal matrix $D$ with $D_{ii} = w_i^{-1}$, $i \in \{1, \ldots, n\}$. One can verify that the matrix $\Pi = \Sigma D$ is column stochastic, thus it can be considered as the transition matrix of a Markov chain evolving on $X = \{s_1, \ldots, s_n\}$. Note also that $Pr(\Pi_{ij} > 0) = 1$, $\forall i, j \in \{1, \ldots, n\}$, due to the way that $\Pi$ was generated, thus the stationary distribution corresponding to $\Pi$ is unique. Consider the stationary Markov process $\{x(t)\}$ on $X$ with transition matrix $\Pi$ and obtain a stationary HMM $H$ over $Y = \{0, 1\}$ by defining the output process as a deterministic function of the state process. In particular $x(t) = 0$ if $x(t) \in \{s_1, \ldots, s_{50}\}$ and $y(t) = 1$ if $x(t) \in \{s_{51}, \ldots, s_{100}\}$. The HMM $H$ generated following the procedure described above is used to demonstrate the reduction algorithm. Fig. 4 depicts the eigenvalues of the matrix $W$, which according to (12) control the error bound between $H$ and $\tilde{G}$. Overall there is an evident decay. The HMM $H$ is truncated to a generalized automaton $\tilde{G}$ of size $\tilde{n} = 8$. In the second step by solving the NLP (38) a HMM of size $\tilde{n} = 4$ is obtained.

Let $T = 8$, it is determined a posteriori that $q_{G} > 0$, $\forall v \in \mathcal{Y}^T$ and

$$\sum_{v \in \mathcal{Y}^*} (p_H[v] - q_{G}[v])^2 = 1.08 \times 10^{-6},$$

$$\sum_{v \in \mathcal{Y}^*} (q_{G}[v] - p_H[v])^2 = 5.09 \times 10^{-4},$$

Introduce the function

$$p_H^T = \begin{cases} p_H & \text{if } p_H[v] > 0 \\ \text{undefined} & \text{otherwise} \end{cases}$$

and define similarly

$$q_{G}^T = \begin{cases} q_{G} & \text{if } q_{G}[v] > 0 \\ \text{undefined} & \text{otherwise} \end{cases}$$

The function $p_H^T$ has the obvious conditional probability interpretation. In this example $p_H^T$, $q_{G}^T$ are defined $\forall v \in \mathcal{Y}^T$. Let $G_{1} = G_{2} = [q_{G}[v] - q_{G}[v]]$ and $G_{1} = [q_{G}[v] - q_{G}[v]]$.

The algorithm achieved a compression by a factor of 25 with good quality, especially if one focusses on the approximation of $p_H$. The larger errors in the right column of the Table I are due to the fact that even small absolute errors in approximating strings of low probability of occurrence can reflect heavily due to the way that $p_H$. Table II shows the error measures over $\mathcal{Y}^T$. Let $\mathcal{Y}^T = \{v \in \mathcal{Y}^T | p_H[v] > \delta\}$ and consider for instance the case where $\delta = 0.03$, which is depicted in Table II.

In the future alternative formulations for the second step are worth pursuing, as well as ways of obtaining a certifiable quasi-realization of finite horizon.

APPENDIX

A. Feasibility of the Generalized Dissipation Inequalities

Given the mean square stability assumption it will be shown in this part of the appendix that there exists $\hat{U} > 0$, which satisfies (9) and $R > 0$, which satisfies (10).

Mean square stability is equivalent with the existence of $\hat{U} > 0$, such that

$$\sum_{i=1}^{N} p_i A[i]^{T} \hat{U} A[i] - \hat{U} < 0.$$  \hfill (43)
Relation (9) is equivalent to

\[ \sum_{i=1}^{N} p_i A[i]^T U A[i] - U \leq \sum_{i=1}^{N} p_i C[i]^T C[i]. \]

By virtue of the above two relations, if one sets \( U = \alpha \tilde{U} \) and takes \( \alpha > 0 \) large enough, the dissipation inequality (9) can always be satisfied by some positive definite matrix \( U \). Relation (10) is equivalent to (20). By taking the Schur complement, one can see that it is sufficient to compute \( R > 0 \) such that

\[ -R + \sum_{i=1}^{N} p_i A[i]^T R A[i] < 0 \quad (44) \]

and

\[ \sum_{i=1}^{N} p_i B[i]^T R B[i] - W_{11}^T W_{11}^{-1} W_{12} < I \quad (45) \]

where

\[ W_{11} = -R + \sum_{i=1}^{N} p_i A[i]^T R A[i] \]
\[ W_{12} = \sum_{i=1}^{N} p_i B[i]^T R B[i]. \]

Let \( \tilde{U} \) satisfy (43) and set \( R = \alpha \tilde{U} \) with \( \alpha > 0 \). Condition (44) is equivalent to mean square stability and thus feasible for all positive values of \( \alpha \). Concerning condition (45) note that both terms on the left hand side of the relation scale linearly with \( \alpha \). Thus, by taking the positive parameter \( \alpha \) to be small enough one can also satisfy (10) with the choice of \( R = \alpha \tilde{U} \).

ACKNOWLEDGMENT

The authors would like to thank the reviewers for valuable comments and suggestions and V. Blondel for [14].

REFERENCES


Georgios Kotsalis received the Ph.D. degree from the Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, in 2006, the Dipl.-Ing. degree from the Department of Mechanical Engineering, Swiss Federal Institute of Technology (ETH), Zurich, and the Financial Technology Certificate from the Sloan School of Management, Massachusetts Institute of Technology, Cambridge.

While at MIT, he was a member of the Laboratory of Information and Decision Sciences. He is currently a Post Doctoral Researcher at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. During 2007, he was a Post Doctoral Researcher at the Department of Automatic Control, Lund University, Lund, Sweden. His research interests are in the broad areas of system theory and control with current emphasis on complexity reduction of dynamical systems, robust decision making in uncertain competitive environments and mechanism design in dynamic auction settings.

Dr. Kotsalis received the ETH Medal.

Alexandre Megretski received the Ph.D. degree in control theory from Leningrad University, Leningrad, Russia, in 1988.

He was a Researcher at the Royal Institute of Technology, Sweden, University of Newcastle, Australia, and a faculty member at Iowa State University, Ames. He is now a Professor of electrical engineering at the Massachusetts Institute of Technology (MIT), Cambridge, working on nonlinear systems (analysis, identification, model reduction, and design) and optimization.

Munther A. Dahleh was born in 1962. He received the B.S. degree from Texas A & M University, College Station, in 1983 and the Ph.D. degree from Rice University, Houston, TX, in 1987, all in electrical engineering.

Since then, he has been with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, where he is now a Full Professor. He has been a Visiting Professor with the Department of Electrical Engineering, California Institute of Technology, Pasadena, for the Spring of 1993. He has held consulting positions with several companies in the U.S. and abroad. He is co-author of Control of Uncertain Systems: A Linear Programming Approach (Englewood Cliffs, NJ: Prentice-Hall) and Computational Methods for Controller Design (Berlin, Germany: Springer). He is interested in problems at the interface of robust control, filtering, information theory, and computation which include control problems with communication constraints and distributed mobile agents with local decision capabilities. He is also interested in model reduction problems for discrete-alphabet hidden Markov models and universal learning approaches for systems with both continuous and discrete alphabets. He is also interested in the interface between systems theory and neurobiology, and in particular, in providing an anatomically consistent model of the motor control system.