2009 American Control Conference
Hyatt Regency Riverfront, St. Louis, MO, USA
June 10-12, 2009

An Information-Theoretic Framework to Aggregate a Markov Chain

Kun Deng, Yu Sun, Prashant G. Mehta, and Sean P. Meyn

Abstract— **This paper is concerned with an information-theoretic framework to aggregate a large-scale Markov chain to obtain a reduced order Markov model. The Kullback-Leibler (K-L) divergence rate is employed as a metric to measure the distance between two stationary Markov chains. Model reduction is obtained by considering an optimization problem with respect to this metric. The solution is just the optimal aggregated Markov model. We show that the solution of the bi-partition problem is given by an eigenvalue problem. To construct a reduced order model with m super-states, a recursive algorithm is proposed and illustrated with examples.**

I. Introduction

An important new tool for understanding multi-scale phenomenon is based on the spectral theory of Markov models. For a stationary Markov chain on a finite dimensional state space, the second eigenvalue is precisely the rate of convergence to stationarity. A more recent contribution to the spectral theory of Markov models is the use of the second eigenvector (or eigenfunction) to obtain the intuition regarding dynamics, as well as methods for aggregation in complex models. In dynamical systems settings, this technique was introduced as a heuristic in [1], [2] to obtain a state-space decomposition based on an analysis of the *Perron cluster of eigenvalues* for the generator of a Markov process. The technique has been applied in diverse settings: [1] considers analysis of the nonlinear chaotic dynamics of Chua's circuit model, [2] concerns molecular models, and [3] treats transport phenomena in building systems. In each of these papers, it is shown through numerical examples that the associated eigenvectors carry significant information regarding dynamics. In particular, its sign-structure can be used to obtain the partition information for defining super-states. Theory to support this aggregation technique is contained in [4], [5], based on a change of measure similar to what is used to establish large deviations asymptotically, following [6]. These results may be regarded as an extension to the classical Wentzell–Freidlin theory [7].

The spectral method has close connections to both the classical notion of *nearly completely decomposable* Markov chain (NCDMC) [8] and the notion of *cut* in spectral graph theory [9]. For an NCDMC, the state space can be naturally divided into *groups* with strong interactions within each group and weak interactions among different groups. Such a decomposition is consistent with sign-structure of the second eigenvector and an assumption on a spectral gap between the second and the rest of the eigenvalues. Aggregations of states within each group can then be justified using a singular perturbation framework (see [10]): over the long time-period that weak interactions become significant, the strongly interacting states within each group can be treated as an aggregated super-state.

A related notion is that of a *cut* that is used for partitioning a graph. A symmetric Markov chain may be represented as an undirected graph where vertices of the graph denote states of the Markov chain and (weights on) edges represent the transition probabilities between states. A cut is defined to be a certain normalized sum of weights that are removed to obtain a bi-partition (into two groups) of the graph. In terms of a minimal cut, the optimal solution for the bi-partition problem is described by the sign-structure of the second eigenvector of the Markov transition matrix [9]. The resulting decomposition algorithms have been extensively used in applications including image segmentation, clustering and graph partitioning.

The objective of this paper is to examine decomposition, aggregation and model reduction issues for Markov chains in information-theoretic terms. The goal is to construct an information-theoretic basis for both interpreting classical and more recent spectral methods, and deriving new error bounds and algorithms for model reduction of Markov chains. In particular, we seek to explain the significance of the second eigenvector in these terms.

The consideration of this paper rests on the use of Kullback-Leibler divergence rate metric for *stationary* Markov chains [11]. With K-L metric, the model reduction problem is expressed as an optimization problem. Taking bi-partition problem as an example, the solution is shown to be given by the sign-structure of the second eigenvector consistent with the spectral theory of Markov models. To construct a reduced order model with m super-states, a recursive algorithm is proposed and illustrated with examples.

The remainder of this paper is organized as follows: In Section II, we define the metric used to compare Markov chains. In Section III, we pose an optimization problem with respect to this metric and describe the results for bi-partition case. In Section IV, we describe several examples.

II. Metric

A. Preliminaries and notations

We consider a first-order homogeneous Markov chain $\mathcal{X}_t$ defined on a finite dimensional state space $\mathcal{N} = \{1, 2, \ldots, n\}$ (see [12] for terminology). The following notations are adopted throughout the paper: The state value at time t is

The authors are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 West Main Street, Urbana, IL 61801. Email: kundeng2@uiuc.edu, yusun2@uiuc.edu, mehtapg@uiuc.edu, and meyn@control.csl.uiuc.edu

978-1-4244-4524-0/09/$25.00 ©2009 AACC

denoted as $X(t)$, the initial condition $X(0)$ is denoted as $x_0 \in \mathcal{N}$, $\nu_0 \in \mathcal{P}(\mathcal{N})$ is used to denote the probability distribution of the initial condition x_0. The transition probability between states is described by a $n \times n$ stochastic matrix P whose ij^{th} entry is given by

$$P_{ij} = \text{Prob}(X(t+1) = j \mid X(t) = i), \quad i, j \in \mathcal{N}. \tag{1}$$

A Markov chain is said to be *stationary* if it has a unique *stationary distribution* π such that

$$\pi = \pi P, \tag{2}$$

where $\pi_i > 0$ for all $i \in \mathcal{N}$. We use the tuple (π, P) to denote a stationary Markov chain P with stationary distribution π.

B. K-L divergence rate for Markov chains on $\mathcal{N}$

For two stationary Markov chains (π, P) and (θ, Q) defined on the same state space $\mathcal{N}$, the K-L divergence rate is given by the following formula (see [11]):

$$R(P \parallel Q) = \sum_{i,j \in \mathcal{N}} \pi_i P_{ij} \log \left(\frac{P_{ij}}{Q_{ij}} \right). \tag{3}$$

To ensure $R(P \parallel Q)$ is finite, we require P to be *absolutely continuous* w.r.t. Q, i.e. $Q_{ij} = 0 \Rightarrow P_{ij} = 0$.

C. K-L divergence rate for Markov chains on different state spaces

In this paper, we are interested in comparing two Markov chains P and Q defined on $\mathcal{N}$ and $\mathcal{M}$ respectively. The relationship between $\mathcal{N}$ and $\mathcal{M}$ is described by a partition function ϕ.

Definition 1 (partition function) *Let $\mathcal{N} = \{1, 2, \ldots, n\}$ and $\mathcal{M} = \{1, 2, \ldots, m\}$ be two finite dimensional state spaces with $m \leq n$. A* partition function $\phi : \mathcal{N} \mapsto \mathcal{M}$ *is a surjective function from $\mathcal{N}$ onto $\mathcal{M}$. For $k \in \mathcal{M}$, $\phi^{-1}(k)$ denotes the k^{th} group in $\mathcal{N}$.*

Since we already have a formula for comparing two Markov chains on the same state space (see (3)), the strategy is to use the partition function ϕ to *lift* the Markov chain Q to the original state space $\mathcal{N}$. The lifted Markov chain is denoted as $\hat{Q}$.

Definition 2 (μ-lifting of Q) *Let ϕ be a partition function on $\mathcal{N}$ and μ be a probability measure on $\mathcal{P}(\mathcal{N})$. Let $\mathcal{M}$ denote the range of ϕ and Q be a Markov transition matrix on $\mathcal{M}$. Then* μ-lifting of Q *with the partition function ϕ is a Markov matrix on $\mathcal{N}$ defined as*

$$\hat{Q}^{(\mu)}_{ij}(\phi) = \frac{\mu_j}{\sum_{k \in \psi(j)} \mu_k} Q_{\phi(i)\phi(j)}, \quad i, j \in \mathcal{N} \tag{4}$$

where $\psi(j) = \phi^{-1} \circ \phi(j) \subset \mathcal{N}$ denotes the set of states belonging to the same group as the j^{th} state.

The definition of the K-L divergence rate is extended to two chains on *different state spaces* using the lifted chain:

Definition 3 *Let (π, P) denote a stationary Markov chain on $\mathcal{N}$ and (θ, Q) a Markov chain on $\mathcal{M}$. Then*

$$R^{(\phi)}(P \parallel Q) \stackrel{\Delta}{=} \min_{\mu \in \mathcal{P}(\mathcal{N})} R(P \parallel \hat{Q}^{(\mu)}(\phi)), \tag{5}$$

where $\hat{Q}^{(\mu)}(\phi)$ denotes the μ-lifting of Q with the partition function ϕ.

Theorem 1 *Suppose that (π, P) is a stationary Markov chain on $\mathcal{N}$, ϕ is a partition function with range $\mathcal{M}$ with $m \leq n$, and (θ, Q) is a stationary Markov chain on $\mathcal{M}$. Then, there is a unique matrix $\hat{Q}^{(\mu^*)}(\phi)$ that achieves the minimum in* (5)*. The optimizer μ^* can be taken to be the stationary distribution of P:*

$$\pi \in \arg\min_{\mu \in \mathcal{P}(\mathcal{N})} R(P \parallel \hat{Q}^{(\mu)}(\phi)). \tag{6}$$

Note that the theorem does not say that μ^* is unique. A probability measure μ minimizes (5) if and only if there exists constants $\{K_l,\ l \in \mathcal{M}\}$ satisfying

$$\frac{\pi_j}{\mu_j} = K_l, \quad \forall j \in \phi^{-1}(l),\ l \in \mathcal{M}. \tag{7}$$

The corresponding matrix $\hat{Q}^{(\mu)}(\phi)$ then coincides with $\hat{Q}^{(\pi)}(\phi)$.

Proof of Theorem 1: On denoting,

$$R_\phi(P \parallel Q) = \sum_{i,j \in \mathcal{N}} \pi_i P_{ij} \log \left(\frac{P_{ij}}{Q_{\phi(i)\phi(j)}} \right) \tag{8}$$

the K-L divergence rate (3) between (π, P) and the lifted Markov chain $(\hat{\theta}(\phi), \hat{Q}^{(\mu)}(\phi))$ is expressed as

$$\begin{aligned} R(P \parallel \hat{Q}^{(\mu)}(\phi)) &= \sum_{i,j \in \mathcal{N}} \pi_i P_{ij} \log \left(\frac{P_{ij}}{\hat{Q}^{(\mu)}_{ij}(\phi)} \right), \\ &= R_\phi(P \parallel Q) - \sum_{i,j \in \mathcal{N}} \pi_i P_{ij} \log \frac{\mu_j}{\sum_{k \in \psi(j)} \mu_k}, \\ &= \underbrace{R_\phi(P \parallel Q)}_{\text{term (i)}} - \underbrace{\sum_{j \in \mathcal{N}} \pi_j \log \frac{\mu_j}{\sum_{k \in \psi(j)} \mu_k}}_{\text{term (ii)}}, \end{aligned} \tag{9}$$

where we used the fact that $\pi_j = \sum_{i \in \mathcal{N}} \pi_i P_{ij}$ (see (2)).

In (9), term (i) is independent of the probability measure μ and term (ii) is independent of the Markov transition matrices P and Q. Thus, we only need to consider the term (ii) of $R(P \parallel \hat{Q}^{(\mu)}(\phi))$ to find the optimal $\mu \in \mathcal{P}(\mathcal{N})$. Using (9) and setting $l = \phi(j)$, we have

$$\begin{aligned} &R(P \parallel \hat{Q}^{(\pi)}(\phi)) - R(P \parallel \hat{Q}^{(\mu)}(\phi)) \\ &= \sum_{l \in \mathcal{M}} \Big(\sum_{k \in \phi^{-1}(l)} \pi_k \Big) \log \frac{\sum_{k \in \phi^{-1}(l)} \pi_k}{\sum_{k \in \phi^{-1}(l)} \mu_k} - \sum_{j \in \mathcal{N}} \pi_j \log \frac{\pi_j}{\mu_j}, \\ &\leq \sum_{j \in \mathcal{N}} \pi_j \log \frac{\pi_j}{\mu_j} - \sum_{j \in \mathcal{N}} \pi_j \log \frac{\pi_j}{\mu_j} = 0, \end{aligned}$$

where the *log sum inequality* (see [13]) is used and the equality holds if and only if (7) is satisfied.

Thus $R(P \parallel \hat{Q}^{(\pi)}(\phi)) \leq R(P \parallel \hat{Q}^{(\mu)}(\phi))$ for all $\mu \in \mathcal{P}(\mathcal{N})$. Note that the optimal choice of probability measure is not unique and $\mu = \pi$ is one of the optimal choice which minimizes $R(P \parallel \hat{Q}^{(\mu)}(\phi))$. ∎

The formula (9) in the proof of Theorem 1 will also be useful in obtaining further results. For this purpose, we summarize the formula for the case $\mu = \pi$ in the following:

Lemma 2 *Under the assumptions of Theorem 1,*

$$R(P \parallel \hat{Q}^{(\pi)}(\phi)) = R_\phi(P \parallel Q) - S(\pi, \phi), \qquad (10)$$

where R_ϕ is defined in (8)*, and*

$$S(\pi, \phi) \quad = \quad \sum_{j \in \mathcal{N}} \pi_j \log \frac{\pi_j}{\sum_{k \in \psi(j)} \pi_k}. \qquad (11)$$

III. Optimization Problem

A. Problem statement

Let (π, P) be a given stationary Markov chain on $\mathcal{N}$. The *m-partition problem*, is to find the partition function $\phi : \mathcal{N} \mapsto \mathcal{M}$ and the optimal aggregated Markov chain (θ, Q) such that $R^{(\phi)}(P \parallel Q)$ is minimized:

$$\begin{array}{ll} \min\limits_{\phi, Q} & R^{(\phi)}(P \parallel Q) \\ \text{s.t.} & \sum_{l \in \mathcal{M}} Q_{kl} = 1, \quad k \in \mathcal{M} \\ & Q_{kl} \geq 0, \quad k, l \in \mathcal{M} \end{array} \qquad (12)$$

where $R^{(\phi)}(P \parallel Q) = R(P \parallel \hat{Q}^{(\pi)}(\phi))$ and constraints arise due to the stochastic property of Markov transition matrix.

The optimization problem (12) is a *mixed-integer nonlinear program*. In general, it is intractable for Markov chains with large state spaces.

B. Optimal solution of Q

It turns out that the main difficulty in solving (12) is in finding the optimal partition function. The following theorem shows that for a fixed (say an optimal) partition function ϕ, the solution of Q that solves (12) can be easily obtained.

Theorem 3 *Let (π, P) be a given Markov chain on $\mathcal{N}$ and ϕ be a partition function defined on $\mathcal{N}$ with the range $\mathcal{M}$. For problem (12), if ϕ is fixed, the optimal solution of Q is given by*

$$Q_{kl}(\phi) = \frac{v^{(k)} \Pi P v^{(l)'}}{v^{(k)} \Pi v^{(k)'}}, \quad k, l \in \mathcal{M} \qquad (13)$$

where $\Pi = diag(\pi)$, $v^{(k)'}$ is the transpose of $v^{(k)}$, and $v^{(k)}$ is a 1-by-n row vector whose i^{th} entry is given by

$$v_i^{(k)} = \begin{cases} 1 & \text{if } \phi(i) = k, \\ 0 & \text{otherwise.} \end{cases} \qquad (14)$$

The stationary distribution of Q is given by

$$\theta_k(\phi) = v^{(k)} \Pi v^{(k)'}, \quad k \in \mathcal{M}. \qquad (15)$$

Proof: Noting that $R^{(\phi)}(P \parallel Q)$ is a convex function with respect to Q, we introduce the Lagrange function for the optimization problem (12)

$$L = R^{(\phi)}(P \parallel Q) + \sum_{k \in \mathcal{M}} \lambda_k (\sum_{l \in \mathcal{M}} Q_{kl} - 1), \qquad (16)$$

where $\{\lambda_k, \ k \in \mathcal{N}\}$ are Lagrange multipliers. Applying Lemma 2, we have

$$L = R_\phi(P \parallel Q) - S(\pi, \phi) + \sum_{k \in \mathcal{M}} \lambda_k (\sum_{l \in \mathcal{M}} Q_{kl} - 1). \qquad (17)$$

On taking the derivative with respect to Q_{kl},

$$\begin{aligned} \frac{\partial L}{\partial Q_{kl}} =& \frac{\partial}{\partial Q_{kl}} \left(R_\phi(P \parallel Q) + \sum_{k \in \mathcal{M}} \lambda_k (\sum_{l \in \mathcal{M}} Q_{kl} - 1) \right) \\ =& - \frac{\sum_{i \in \phi^{-1}(k)} \sum_{j \in \phi^{-1}(l)} \pi_i P_{ij}}{Q_{kl}} + \lambda_k. \end{aligned} \qquad (18)$$

Setting the right hand side of (18) equal to zero, we have

$$Q_{kl} = \frac{\sum_{i \in \phi^{-1}(k)} \sum_{j \in \phi^{-1}(l)} \pi_i P_{ij}}{\lambda_k}, \quad k, l \in \mathcal{M}. \qquad (19)$$

The Lagrange multipliers $\{\lambda_k, \ k \in \mathcal{N}\}$ are obtained by using the constraints

$$1 = \sum_{l \in \mathcal{M}} Q_{kl} = \frac{\sum_{i \in \phi^{-1}(k)} \pi_i}{\lambda_k} \sum_{j \in \mathcal{N}} P_{ij},$$

where we used the fact that $\phi : \mathcal{N} \mapsto \mathcal{M}$ is a surjective function, so

$$\sum_{l \in \mathcal{M}} \sum_{j \in \phi^{-1}(l)} P_{ij} = \sum_{j \in \mathcal{N}} P_{ij}.$$

Now, noting that P is a stochastic matrix, we have

$$1 = \frac{\sum_{i \in \phi^{-1}(k)} \pi_i}{\lambda_k} \quad \Rightarrow \quad \lambda_k = \sum_{i \in \phi^{-1}(k)} \pi_i. \qquad (20)$$

Finally, substituting (20) into (19), we get

$$Q_{kl}(\phi) = \frac{\sum_{i \in \phi^{-1}(k)} \sum_{j \in \phi^{-1}(l)} \pi_i P_{ij}}{\sum_{i \in \phi^{-1}(k)} \pi_i}, \quad k, l \in \mathcal{M}$$

which is just the formula shown in (13). Stationarity of (15) with respect to $Q(\phi)$ follows from a trivial calculation. ∎

For a given partition, (13) gives the entries of optimal $Q(\phi)$. We denote it as $Q(v^{(1)}, v^{(2)}, \ldots, v^{(m)})$, where indicator functions $\{v^{(k)}, \ k \in \mathcal{M}\}$ are defined by partition function ϕ (see (14)). Using this notation, the m-partition problem becomes finding only the partition function ϕ such that $R^{(\phi)}(P \parallel Q(v^{(1)}, v^{(2)}, \ldots, v^{(m)}))$ is minimized:

$$\min_{\phi : \mathcal{N} \mapsto \mathcal{M}} R^{(\phi)}(P \parallel Q(v^{(1)}, v^{(2)}, \ldots, v^{(m)})). \qquad (21)$$

C. Optimal partition function for the bi-partition problem

In this section, we consider the special case of (21), where $m = |\mathcal{M}| = 2$. This is referred to as the *bi-partition problem*. In this case, $v^{(2)} = \mathbf{1} - v^{(1)}$. We denote $v = v^{(1)}$ and use the notation $Q(v)$ to denote $Q(v^{(1)}, v^{(2)})$. We refer to v as a *bi-partition function*. For a given v, the optimal solution $Q(v)$ is a 2×2 matrix whose entries are obtained using Theorem 3.

Lemma 4 *Let (π, P) be a given Markov chain on $\mathcal{N}$ and ϕ be a given bi-partition function defined on $\mathcal{N}$ with range $\mathcal{M} = \{1, 2\}$. The optimal solution of Q is given by*

$$Q(v) = \begin{bmatrix} \frac{\alpha(v)}{\beta(v)} & \frac{\beta(v)-\alpha(v)}{\beta(v)} \\ \frac{\beta(v)-\alpha(v)}{1-\beta(v)} & \frac{1-2\beta(v)+\alpha(v)}{1-\beta(v)} \end{bmatrix}, \tag{22}$$

where $v_i = \mathbb{1}_{\phi^{-1}(1)}(i)$, $\Pi = diag(\pi)$, $\alpha(v) = v\Pi P v'$ and $\beta(v) = v\Pi v'$. The stationary distribution of Q is given by

$$\theta(v) = [\beta(v), 1 - \beta(v)]. \tag{23}$$

Proof: In the notation of Theorem 3, $v^{(1)} = v$ and $v^{(2)} = \mathbf{1} - v$ where $\mathbf{1}$ denotes a 1-by-n row vector with all ones. Substituting $v^{(1)}$ and $v^{(2)}$ into formula (13), we have

$$Q(v) = \begin{bmatrix} \frac{v\Pi P v'}{v\Pi v'} & \frac{v\Pi P(\mathbf{1}-v)'}{v\Pi v'} \\ \frac{(\mathbf{1}-v)\Pi P v'}{(\mathbf{1}-v)\Pi(\mathbf{1}-v)'} & \frac{(\mathbf{1}-v)\Pi P(\mathbf{1}-v)'}{(\mathbf{1}-v)\Pi(\mathbf{1}-v)'} \end{bmatrix}. \tag{24}$$

Noting that $v\Pi P(\mathbf{1} - v)' = (\mathbf{1} - v)\Pi P v' = \beta(v) - \alpha(v)$ and $(\mathbf{1} - v)\Pi(\mathbf{1} - v)' = 1 - \beta(v)$, we get entries of Q in (22). The stationary distribution (23) of Q directly follows from (15) in Theorem 3. ∎

Using Lemma 4, we can represent Q in the optimal form (22) in terms of bi-partition function v. Then using Lemma 2, we can represent $R^{(\phi)}(P \parallel Q(v))$ as

$$R^{(\phi)}(P \parallel Q(v)) = \underbrace{(H_1(Q) - H_0(\theta))}_{\text{term (i)}} - \underbrace{(H_1(P) - H_0(\pi))}_{\text{term (ii)}}, \tag{25}$$

where H_0 denotes the standard entropy and H_1 is its Markovian analog:

$$H_1(Q) = -\sum_{k,l \in \mathcal{M}} \theta_k(v) Q_{kl}(v) \log Q_{kl}(v),$$

$$H_0(\theta) = -\sum_{k \in \mathcal{M}} \theta_k(v) \log \theta_k(v),$$

$$H_1(P) = -\sum_{i,j \in \mathcal{N}} \pi_i P_{ij} \log P_{ij}, \; H_0(\pi) = -\sum_{i \in \mathcal{N}} \pi_i \log \pi_i.$$

These formulae are obtained by substituting (22) and (23) into (10). Note that term (ii) in (25) is independent of the bi-partition function v. Thus the optimization problem (21) is equivalent to the following problem,

$$\min_{v} \quad [H_1(Q(v)) - H_0(\theta(v))]. \tag{26}$$

Since both $Q(v)$ and $\theta(v)$ can be represented in terms of $\alpha(v)$ and $\beta(v)$, then we define

$$\begin{aligned} F(\alpha, \beta) &\triangleq H_1(Q) - H_0(\theta) \\ &= -2(\beta - \alpha)\log(\beta - \alpha) - (1 - 2\beta + \alpha)\log(1 - 2\beta + \alpha) \\ &\quad - \alpha \log \alpha + 2\beta \log \beta + 2(1 - \beta)\log(1 - \beta) \end{aligned} \tag{27}$$

We are interested in choosing v that minimizes $F(\alpha(v), \beta(v))$:

$$\min_{v_i \in \{0,1\}} \quad F(\alpha(v), \beta(v)). \tag{28}$$

An exact solution of (28) may be obtained by searching over 2^n possibilities for vector v. To obtain an approximate solution, one may consider relaxing integer constraints on v_i. One particular relaxation is to let $v_i \in \mathbb{R}$ and consider an optimization problem:

$$\min_{v_i \in \mathbb{R}} \quad F(\alpha(v), \beta(v)). \tag{29}$$

To obtain the solution, we take the derivative of the function $F(\alpha(v), \beta(v))$ with respect to v. Setting the derivative equal to zero, after some algebraic manipulations, we obtain the following necessary condition for a minimizer v^*:

$$\frac{d\alpha}{dv}(v^*) = \lambda^*(\alpha(v^*), \beta(v^*)) \frac{d\beta}{dv}(v^*), \tag{30}$$

where

$$\lambda^*(\alpha, \beta) = \frac{\log \frac{(1-\beta)^2(\beta-\alpha)^2}{\beta^2(1-2\beta+\alpha)^2}}{\log \frac{(\beta-\alpha)^2}{\alpha(1-2\beta+\alpha)}}.$$

Substituting the formulae for $\alpha(v)$ and $\beta(v)$ (see Lemma 4) into (30), we see that an optimal solution v^* of the relaxed problem (29) solves the following eigenvalue problem:

$$\hat{P} v^{*\prime} = \lambda^* \Pi v^{*\prime}, \tag{31}$$

where $\lambda^* \triangleq \lambda^*(\alpha(v^*), \beta(v^*))$, and $\hat{P} = \frac{\Pi P + P'\Pi}{2}$ is a symmetric matrix.

As a result, a solution to (29) may be obtained by considering a generalized eigenvalue problem (31). We denote its eigenvalues as $\{1, \lambda_2, \ldots, \lambda_n\}$ sorted in a decreasing order. Although we do not give details here, we propose that the optimal solution to (29) is obtained by setting

$$\lambda^* = \max\{|\lambda_2|, |\lambda_n|\}. \tag{32}$$

Let $u^{(2)}$ denote the associated eigenvector corresponding to the second largest eigenvalue in magnitude. Using $u^{(2)}$, a sub-optimal partition function to the original problem (28) may be obtained as,

$$v_i = \begin{cases} 1 & \text{if } u_i^{(2)} \geq 0 \\ 0 & \text{otherwise} \end{cases}, \quad i \in \mathcal{N}. \tag{33}$$

Based on these considerations, we conclude that: For a nearly completely decomposable Markov chain (NCDMC) (π, P), when $\lambda_2(P) \approx 1$ and $|\lambda_n(P)| < \lambda_2(P)$, a solution to the bi-partition problem may be obtained by considering the sign-structure of the second eigenvector. We note that the solution proposed here is consistent with the results of [4], [5], [9], [14].

D. A recursive algorithm to obtain m partitions

Since the bi-partition problem can be solved by considering the second eigenvector, we propose a recursive bi-partition algorithm to obtain a sub-optimal solution for the m-partition problem (12). The K-L divergence rate serves as the error bound for model reduction and the recursive algorithm is described as follows:

During the m^{th}-iteration of the algorithm, we assume that a partition with m groups (or super-states) is given. The objective of the m^{th} iteration is to obtain a refinement that has $m + 1$ groups. For $i = 1, \ldots, m$, we denote $P^{(i)}$ to be the sub-Markov transition matrix that describes the transition

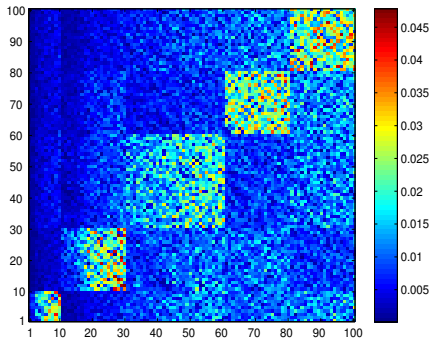


Fig. 1. The graph of the 100-state Markov chain.

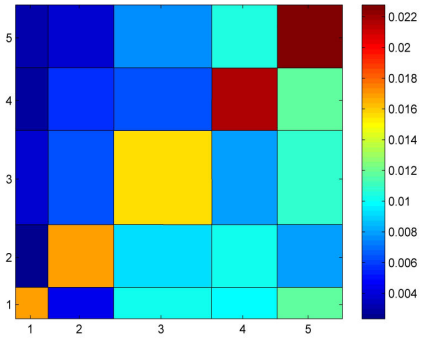


Fig. 2. The graph of the aggregated 5-state Markov chain obtained using the recursive algorithm.

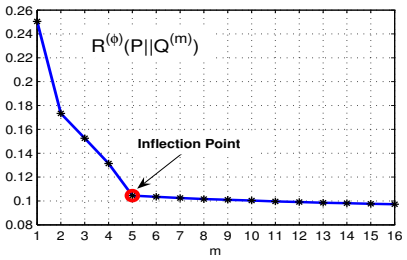


Fig. 3. The K-L divergence rate as a function of the number of aggregated states.

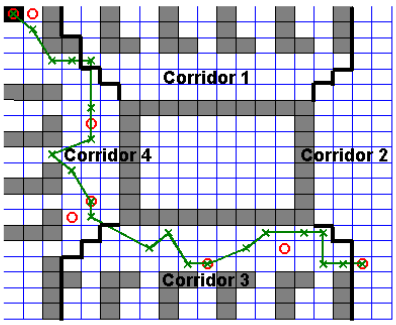


Fig. 4. Layout of the Building Model: $-\times-\times-$ corresponds to a single sample path of the agent, $\circ$ are the sensors, gray grid spaces are walls, and the black grid space is the exit. The agent starts in the bottom-right corner and proceeds toward the exit by a random walk.

probabilities within the i^{th} group. The i^{th} group is split into two sub-groups according to the sign-structure of the second eigenvector for $(\Pi^{(i)}, \hat{P}^{(i)})$ (see (31) and (33)). The spectral split of the i^{th} group alone provides a partition of the states into $m+1$ groups. We denote this partition as $\phi^{(i)}$, and use it to evaluate the optimal reduced order model $Q^{(i)}$ according to (13). From the resulting m possible choices of $m+1$ partitions, we select the one that minimizes $R^{(\phi)}(P \parallel Q^{(i)})$, i.e.,

$$i_{\min} = \underset{i\in\{1,\ldots,m\}}{\arg\min}\ R^{(\phi)}(P \parallel Q^{(i)}).$$

The $m+1$ super-states correspond to the original $m-1$ super-states from the previous iteration together with two super-states obtained from the spectral split of the $i^{\text{th}}_{\min}$ super-state. The associated reduced order model is given by $Q^{(i_{\min})}$.

IV. Examples and Discussions

In this section, we present some examples to illustrate the theoretical results described in preceding sections.

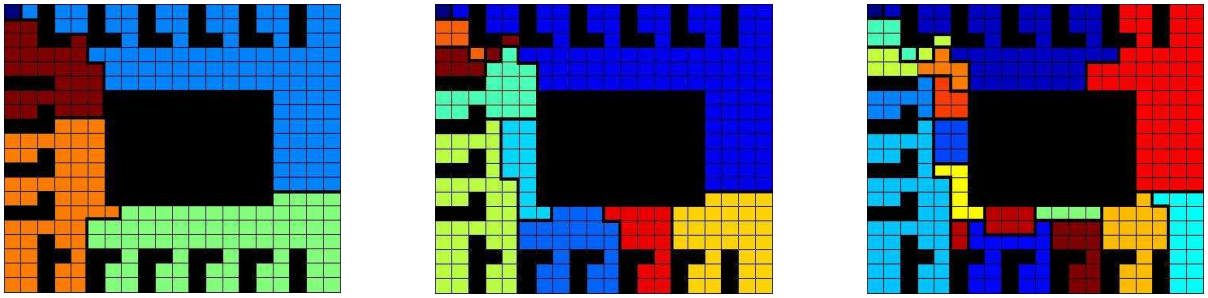
A. Block partitioning example

The 100-state stationary Markov chain for this example is taken from [15]. Fig. 1 depicts the transition probabilities for this chain. The cold colors indicate weak interactions (small transition probabilities), and warm colors indicate strong interactions (large transition probabilities) between states. The color plot suggests that the Markov chain is nearly completely decomposable with five groups.

In the following, we employ the recursive algorithm to obtain a reduced order model. With $m = 1$, all states are aggregated into a single group and $R^{(\phi)}(P \parallel Q^{(1)}) = 0.247$. The bi-partition problem ($m = 2$) is solved by considering the sign-structure of the second eigenvector for $(\Pi, \hat{P})$ (see (31) and (33)). The resulting 2-state Markov model has error $R^{(\phi)}(P \parallel Q^{(2)}) = 0.176$. The recursive algorithm correctly identifies five groups in the fifth recursion. Fig. 2 depicts the transition probabilities for the resulted reduced order model $Q^{(5)}$. Fig. 3 depicts the error bound (K-L divergence rate) as a function of the number of aggregated states $m = |\mathcal{M}|$. The error bound plot shows that $R^{(\phi)}(P \parallel Q^{(m)})$ decreases rapidly from $m = 1$ to $m = 5$. With $m = 5$, $R^{(\phi)}(P \parallel Q^{(5)}) = 0.105$. After five strongly interacting groups have been identified, additional super-states in the reduced order model ($m > 5$) do not significantly decrease the error bound.

B. Building example

We consider a grid-based model [16] of an agent movement in a large building as shown in Fig. 4. We denote the successive locations of the agent by $\{X(0), X(1), \ldots, X(t)\}$. $\mathcal{N} = \{1, 2, \ldots, n\}$ denote the cells in the building that can be occupied by the agent ($n = 255$ for the building considered here). We use a sub-stochastic matrix P to denote its Markovian transition probabilities: For a special node e, called the *exit node*, we have $P_{ej} = 0$



(a) (b) (c)

Fig. 6. Depicts (a) 4-aggregation, (b) 9-aggregation, and (c) 16-aggregation of the building plane by using the initial distribution ν_0.

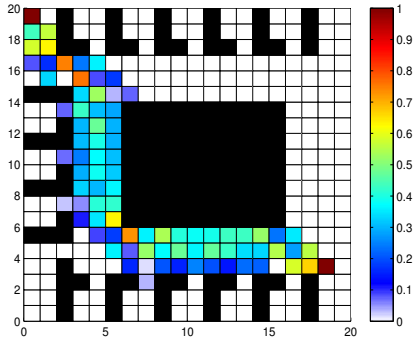


Fig. 5. Depicts the pseudo-stationary distribution $\frac{1}{\|\hat{\pi}\|}\hat{\pi}$, where $\hat{\pi} = \nu_0(I-P)^{-1}$ and $\nu_0 \in \mathcal{P}(\mathcal{N})$ denotes the initial distribution.

for all $j \in \mathcal{N}$. The agent leaves the building via the exit node. For all other nodes i, it is assumed that $P_i.$ is an honest probability measure: An agent at node i will move according to this distribution. If $\sum_j P_{ij} < 1$, then the agent leaves the building from node i with probability $(1-\sum_j P_{ij})$. Finally, it is assumed that each node is transient in the sense that the agent eventually exits the building. This assumption is expressed by requiring the matrix P to be transient.

Even though, we only present results for stationary Markov chains, similar generalization also applies to the transient case where K-L divergence rate is replaced by K-L divergence, and stationary distribution is replaced by the *pseudo-stationary distribution* $\frac{1}{\|\hat{\pi}\|}\hat{\pi}$, where $\hat{\pi} = \nu_0(I-P)^{-1}$ and $\nu_0 \in \mathcal{P}(\mathcal{N})$ denotes the initial distribution. Intuitively, $\frac{1}{\|\hat{\pi}\|}\hat{\pi}_i$ denotes the fraction of the expected time spent in the i^{th} node before the agent exits the building.

For a transient Markov chain, one requires the knowledge of the initial distribution ν_0 regarding the agent starting location x_0. It is used to obtain the pseudo-stationary distribution $\frac{1}{\|\hat{\pi}\|}\hat{\pi}$ as depicted in Fig. 5. Fig. 6 summarizes the aggregations obtained using the recursive algorithm (for $m = 4, 9, 16$). We make following two observations:

1) The $\frac{1}{\|\hat{\pi}\|}\hat{\pi}$ is supported primarily in left and bottom corridors (see Fig. 5). As a result, one obtains finer aggregations of states in these corridors with increasing values of m (see Fig. 6).
2) For large values of m, the groups show non-uniform aggregation of states even in the same corridor. This is due to the nature of the assumed agent movement dynamics. These transition probabilities are obtained by perturbing the baseline best route of an agent to the exit. As a result, states inside offices have small probability of visit and these states are grouped together into groups with larger number of states.

REFERENCES

[1] O. Junge and M. Dellnitz, "Almost invariant sets in Chua's circuit," *Int. J. Bif. and Chaos*, vol. 7, pp. 2475–85, 1997.

[2] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid Monte Carlo," *J. Comput. Phys., Special Issue on Computational Biophysics*, vol. 151, pp. 146–168, 1999.

[3] P. G. Mehta, M. Dorobantu, and A. Banaszuk, "Graph-based multi-scale analysis of building system transport models," in *Procs. of American Controls Conference*, Minneapolis, 2006, pp. 1110–1115.

[4] W. Huisinga, S. P. Meyn, and C. Schütte, "Phase transitions and metastability in Markovian and molecular systems," *Ann. Appl. Probab.*, vol. 14, no. 1, pp. 419–458, 2004.

[5] S. P. Meyn, G. Hagen, G. Mathew, and A. Banasuk, "On complex spectra and metastability of Markov models," in *Proc. of the IEEE Conf. on Decision & Control*, December 2008.

[6] I. Kontoyiannis and S. P. Meyn, "Spectral theory and limit theorems for geometrically ergodic Markov processes," *Ann. Appl. Probab.*, vol. 13, pp. 304–362, 2003, presented at the INFORMS Applied Probability Conference, NYC, July, 2001.

[7] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, "Metastability in stochastic dynamics of disordered mean-field models," *Probab. Theory Related Fields*, vol. 119, no. 1, pp. 99–161, 2001.

[8] H. A. Simon and A. Ando, "Aggregation of variables in dynamical systems," *Econometrica*, vol. 28, pp. 111–138, 1961.

[9] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[10] R. G. Phillips and P. V. Kokotovic, "A singular perturbation approach to modeling and control of Markov chains," *IEEE Trans. Automat. Contr.*, vol. 26, no. 5, pp. 1087–1094, 1981.

[11] Z. Rached, F. Alalaji, and L. L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Trans. Info. Thy.*, vol. 50, no. 5, pp. 917–921, 2004.

[12] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. London: Springer-Verlag, 1993, online: `http://black.csl.uiuc.edu/~meyn/pages/book.html`.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. New York, NY: John Wiley & Sons, Inc., 1991.

[14] W. Huisinga, "Metastability of markovian systems: A transfer operator approach in application to molecular dynamics," Ph.D. dissertation, Free University Berlin, 2001.

[15] M. Meila and L. Xu, "Multiway cuts and spectral clustering," May 2003, Technical Report 442.

[16] J. Niedbalski, K. Deng, P. G. Mehta, and S. Meyn, "Model reduction for reduced order estimation in traffic models," in *Proceeding of American Control Conference*, 2008, pp. 914–919.