Model reduction, optimal prediction, and the Mori-Zwanzig representation of Markov chains

C. L. Beck,* S. Lall§, T. Liang† and M. West‡

***Abstract*— Model reduction methods from diverse fields—including control, statistical mechanics and economics—aimed at systems that can be represented by Markov chains, are discussed in terms of their general properties and common features. These methods include decomposability, optimal prediction techniques, and Mori-Zwanzig representations. Our objective in this paper is to present a survey of and highlight connections between the approaches pursued in different fields, and demonstrate application of the methods on a set of well-known examples.**

I. INTRODUCTION

There has been a long and ubiquitous need for model reduction methods, with a history of research in fields as varied as economics [23], speech and signal processing [15], Internet analysis [24], and statistical mechanics [2], [3], to name a few. The common goal in all of this research is to find a simple mathematical model that adequately represents the behavior of a given complex system. Any particular reduction algorithm is then judged upon the level of complexity reduction achieved, how closely the reduced model captures the given system behavior, and the computational complexity of implementing the algorithm. In this paper we focus on model reduction from a controls and dynamical systems perspective, with our specific goal being the reduction of large scale Markov chain representations for complex systems. We begin with a brief overview of research in this area.

One of the most standard approaches to model reduction of Markov chains has been to aggregate states into meta-states based on the concept of *completely decomposable* and *nearly completely decomposable* systems, first introduced in [23]. In this setting, the aggregated states captured by each of the meta-states have dynamics which evolve along a similar short-run time scale, whereas the interactions between the meta-states evolve on a long-run time scale. This approach provides the basis for singular perturbation methods such as those proposed in [20] and discussed more recently in [27]. Similar state-aggregation approaches for Markov chains have been established which are directly related to the property of *lumpability* [13], [26], [21], where lumpability of a Markov chain refers to the partitioning of the states of the chain into aggregated sets which exhibit similar dynamics and observation statistics. An alternate approach for analyzing multi-scale behavior is based on the use of spectral graph partitioning methods, where we view the Markov model as a graph with edge weights defined by the entries of the transition matrix. Tools from spectral graph partitioning can be applied to multi-partition the Markov-based graph into invariant subgraphs, or to determine the level of connectivity or interaction between aggregated states [10], [22], [18].

Model reduction has also been long considered from the standpoint of statistical mechanics. Here it is standard to describe a system in terms of *microstates* on a fine scale, with *macrostates* being observable quantities characterized by a probability distribution over an ensemble of microstates. Given dynamics on the microstates, it is natural to ask what dynamics are induced on the macrostates, with an answer provided by so-called Mori-Zwanzig theory [17], [28] (see also [8]). Recently much work has been done on the theory of optimal prediction [2], [3], which is concerned with the stochastic dynamics on coarse observables of a system. This latter work is particularly aimed at understanding numerical simulation, where in applications such as fluid dynamics the coarse observables are values of the state on grid points and *sub-grid* information has been lost. The question then is how best to update the coarse variables given a time history of coarse variables and statistical information about the unobserved fine-scale model.

There remain a number of open issues on this topic. In particular, error bounds on the reduction processes noted herein have yet to be determined. We note that for Hidden Markov Models (HMMs), a preliminary generalization of the classical balanced truncation methods established in the 1980s for linear time-invariant systems ([16], [11], [7], [12]) has been suggested [14], in which a two-stage process is proposed with balanced truncation error bounds established for the first stage.

The main contributions of the present paper are to highlight the equivalence between and unify existing reduction results from various fields, ranging from the study of HMMs in control theory to the techniques of system reduction using statistical methods, specifically the optimal prediction framework [2], [3], and to cast these in a form useful for applications. We discuss these ideas in the context of several well-known sample problems.

II. MARKOV CHAINS WITH OBSERVATIONS

Consider a finite state space $X = \mathbb{Z}_n = \{1, \ldots, n\}$ and a finite observation space $Y = \mathbb{Z}_m$. Letters $i, j, k, \ldots$ are used for variables in X and $a, b, c, \ldots$ for those in Y. The spaces of scalar-valued functions on these spaces are denoted

*Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, beck3@illinois.edu

§ Department of Electrical Engineering, and Department of Aeronautics and Astronautics, Stanford University, lall@stanford.edu

† Department of Aeronautics and Astronautics, Stanford University, tzuchen@stanford.edu

‡ Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, mwest@illinois.edu

$\mathcal{C}_X \cong \mathbb{R}^n$ and $\mathcal{C}_Y \cong \mathbb{R}^m$, and we also consider the space $\mathcal{C}_{X\times Y} \cong \mathbb{R}^{nm}$ of scalar-valued functions on $X \times Y$. The spaces of probability distributions on X, Y, and $X \times Y$ are then naturally identified with subsets of the dual spaces $\mathbb{R}^{n*}$, $\mathbb{R}^{m*}$, and $\mathbb{R}^{nm*}$, respectively. Function evaluation is written either $f(i)$ or f_i for $f \in \mathcal{C}_X$ and $i \in X$, similarly for $\mathcal{C}_Y$, and $h(i,a) = h_{ia}$ for $h \in \mathcal{C}_{X\times Y}$ and $(i,a) \in X \times Y$. We identify both primal elements $f \in \mathcal{C}_X$ and dual elements $d \in \mathcal{C}_X^*$ with column vectors, so the natural pairing is given by $\langle d, f\rangle = d^T f$. We do not use an implicit summation notation.

Consider a Markov chain consisting of a sequence of random variables $\{x(t) \mid t = 0, 1, 2, \ldots\}$ on the state space X defined by the transition matrix P, together with conditional observation random variables $\{y(t) \mid t = 0, 1, 2, \ldots\}$ defined by B. The states and observations have the Markov property:

$$\begin{aligned}\operatorname{Prob}\Big(x(t+1) = j \mid x(0) = i_0, \ldots x(t) = i_t,& \\ y(0) = a_0, \ldots, y(t) = a_t\Big)& \\ = \operatorname{Prob}\Big(x(t+1) = j \mid x(t) = i_t\Big)& \quad (1)\\ \operatorname{Prob}\Big(y(t+1) = b \mid x(0) = i_0, \ldots x(t+1) = i_{t+1},& \\ y(0) = a_0, \ldots, y(t) = a_t\Big)& \\ = \operatorname{Prob}\Big(y(t+1) = b \mid x(t+1) = i_{t+1}\Big)& \quad (2)\end{aligned}$$

and the transition probabilities are defined by:

$$P_{ij} = \operatorname{Prob}\Big(x(t+1) = j \mid x(t) = i\Big) \quad (3)$$

$$B_{ia} = \operatorname{Prob}\Big(y(t) = a \mid x(t) = i\Big). \quad (4)$$

This model is therefore termed a Hidden Markov Model (HMM) and is frequently used in a variety of signal processing applications. With the addition of control inputs, it becomes a partially observable Markov decision process (POMDP).

Distributions on X then evolve according to $d(t+1) = P^T d(t)$ for $d(t) \in \mathcal{C}_X^*$, while functions evolve by the adjoint $g(t) = Pg(t+1)$, so that they satisfy the natural duality pairing $g(t+1)^T d(t+1) = g(t)^T d(t)$. For simplicity, we will assume that P is irreducible and positive recurrent, so that it has a unique invariant distribution π satisfying $P^T\pi = \pi$ and $\pi_i > 0$ for all i [19]. In addition, we assume that B has no zero columns, so $(B^T\pi)_a > 0$ for all a. If the invariant distribution is not unique or not positive then the theory below must be extended to choose an invariant distribution and to restrict consideration to the support of π, or to consider the recurrent classes of P individually.

III. Optimal prediction

The evolution of functions by the Markov chain can equivalently be defined by $(Pg)_i = E[g(x(t+1)) \mid x(t) = i]$, so Pg is the optimal predictor of $g(x(t+1))$ that depends only on $x(t)$. That is, $Pg = \operatorname{argmin}_{g'} E[(g'(x(t)) - g(x(t+1)))^2]$. To find a reduced chain $\bar{P}$ that is defined only on the observable y we take the optimal prediction given by

$$\bar{P}f = \lim_{t\to\infty} \operatorname*{argmin}_{f'} E\Big[\big(f'(y(t)) - f(y(t+1))\big)^2\Big]. \quad (5)$$

This means that $(\bar{P}f)_a = \lim_{t\to\infty} E[f(y(t+1)) \mid y(t) = a]$, and hence $\bar{P}_{ab} = \lim_{t\to\infty} \operatorname{Prob}\Big(y(t+1) = b \mid y(t) = a\Big)$. Computing the conditional expectation thus gives the explicit formula

$$\bar{P}_{ab} = \frac{\sum_{i,j} \pi_i P_{ij} B_{ia} B_{jb}}{\sum_i \pi_i B_{ia}}. \quad (6)$$

IV. Operator representation

An alternative method of deriving the expression for the optimal predictor Markov chain $\bar{P}$ is useful for deriving the associated Mori-Zwanzig representation. We begin by defining $\Psi_X g$ to be the best predictor of $g(x)$ that depends only on y, and $\Psi_Y f$ to be the best predictor of $f(y)$ that depends only on x. That is,

$$\Psi_X g = \lim_{t\to\infty} \operatorname*{argmin}_{f} E\Big[\big(f(y(t)) - g(x(t))\big)^2\Big] \quad (7)$$

$$\Psi_Y f = \lim_{t\to\infty} \operatorname*{argmin}_{g} E\Big[\big(g(x(t)) - f(y(t))\big)^2\Big]. \quad (8)$$

Note that these limits exist if the chain is irreducible and aperiodic. These maps thus have pointwise forms $(\Psi_X g)_a = \lim_{t\to\infty} E[g(x(t)) \mid y(t) = a]$ and $(\Psi_Y f)_i = \lim_{t\to\infty} E[f(y(t)) \mid x(t) = i]$. Using Ψ_X and Ψ_Y we can now see that

$$\begin{aligned}(\Psi_X P \Psi_Y f)_a &= \lim_{t\to\infty} E[P\Psi_Y f(x(t+1)) \mid y(t+1) = a] \\ &= \lim_{t\to\infty} E[\Psi_Y f(x(t)) \mid y(t+1) = a] \\ &= \lim_{t\to\infty} E[f(y(t)) \mid y(t+1) = a]\end{aligned}$$

and hence

$$\bar{P} = \Psi_X P \Psi_Y. \quad (9)$$

To compute explicit forms for Ψ_Y and Ψ_X and hence $\bar{P}$, we define the natural inclusion maps $i_X : \mathcal{C}_X \to \mathcal{C}_{X\times Y}$ and $i_Y : \mathcal{C}_Y \to \mathcal{C}_{X\times Y}$ by $(i_X g)_{ia} = g_i$ and $(i_Y f)_{ia} = f_a$. The adjoints of the inclusions are the summations $(i_X^* d)_i = \sum_a d_{ia}$. and $(i_Y^* d)_a = \sum_i d_{ia}$.

The invariant distribution on $X \times Y$ is given by $\pi_i B_{ij}$, which we can use to define a natural map $L : \mathcal{C}_{X\times Y} \to \mathcal{C}_{X\times Y}^*$ by $(Lf)_{ia} = f_{ia}\pi_i B_{ia}$. This then induces the natural inner products L on $\mathcal{C}_{X\times Y}$ and L^{-1} on $\mathcal{C}_{X\times Y}^*$.

Using this additional notation we can rewrite (7) and (8) as

$$\Psi_X g = \operatorname*{argmin}_f \big\| i_Y f - i_X g\big\|_L^2 \quad (10)$$

$$\Psi_Y f = \operatorname*{argmin}_g \big\| i_X g - i_Y f\big\|_L^2. \quad (11)$$

That is, $\Psi_X g$ is the i_Y-preimage of the L-orthogonal projection of $i_X g$ onto the range of i_Y, and similarly for Ψ_Y. To

evaluate these projections we draw the commutative diagram

$$\begin{array}{ccccc} (\mathcal{C}_X, M_X) & \xrightarrow{i_X} & (\mathcal{C}_{X\times Y}, L) & \xleftarrow{i_Y} & (\mathcal{C}_Y, M_Y) \\ \downarrow{\scriptstyle M_X} & & \downarrow{\scriptstyle L} & & \downarrow{\scriptstyle M_Y} \\ (\mathcal{C}_X^*, M_X^{-1}) & \xleftarrow[i_X^*]{} & (\mathcal{C}_{X\times Y}^*, L^{-1}) & \xrightarrow[i_Y^*]{} & (\mathcal{C}_Y^*, M_Y^{-1}) \end{array} \quad (12)$$

where M_X and M_Y are the induced inner products on $\mathcal{C}_X$ and $\mathcal{C}_Y$. These can be readily evaluated to give

$$M_X = i_X^* L i_X = \mathrm{diag}(\pi) \quad (13)$$

$$M_Y = i_Y^* L i_Y = \mathrm{diag}(B^T\pi) \quad (14)$$

The projections (10) and (11) are thus given by $\Psi_X = M_Y^{-1} i_Y^* L i_X$ and $\Psi_Y = M_X^{-1} i_X^* L i_Y$, which evaluate to

$$\Psi_X = M_Y^{-1} B^T M_X \quad (15)$$

$$\Psi_Y = B. \quad (16)$$

From (9) we thus have the reduced transition matrix

$$\bar{P} = M_Y^{-1} B^T M_X P B, \quad (17)$$

which is the same as the expression (6).

V. Properties of optimal predictor chains

Once an optimal predictor chain $\bar{P}$ has been formed from a Markov chain P with observations B, we can easily obtain a number of elementary properties for the reduced chain.

Theorem 5.1: The invariant distribution of $\bar{P}$ is $\bar{\pi} = \Psi_Y^T \pi = B^T\pi$, and $\pi = \Psi_X^T \bar{\pi}$. Hence $M_Y = \mathrm{diag}(\bar{\pi})$.

Proof: Check that $\pi = \Psi_X^T B^T \pi$ using (15) and the facts $\mathrm{diag}(B^T\pi)^{-1} B^T \pi = 1$ and $M_X B 1 = \pi$. Then compute $\bar{P}^T B^T \pi$ using (9) and $P^T\pi = \pi$. ∎

Theorem 5.2: If P is reversible and B is full rank then $\bar{P}$ is reversible.

Proof: Recall that the chain represented by P is reversible if and only if $M_X P$ is symmetric, and compute $M_Y \bar{P} = B^T M_X P B$. ∎

Theorem 5.3: If P and $\bar{P}$ are reversible then they represent random walks with weights $W = M_X P$ and $\bar{W} = M_Y \bar{P}$ related by $\bar{W} = B^T W B$.

Proof: Compute $\bar{W} = M_Y \bar{P} = B^T M_X P B$. ∎

VI. Mori-Zwanzig representation

The previous sections found the best Markov model $\bar{P}$ on the observation space Y for the Markov chain P on the state space X. The reduced model $\bar{P}$ can be regarded as the leading-order term in a representation of the evolution on Y as a non-Markovian system, known as the the Mori-Zwanzig representation. This is constructed as follows.

Theorem 6.1: Take linear operators $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times n}$ and $R \in \mathbb{R}^{n\times m}$. Define a linear system by $x(t+1) = Ax(t)$ with measurements $y(t) = Cx(t)$ and a reconstruction $\hat{x}(t) = Ry(t)$. Then

$$y(t+1) = CARy(t) + \sum_{k=0}^{t-1} C\Big(A(I-RC)\Big)^{t-k} ARy(k) + C\Big(A(I-RC)\Big)^{t+1} x(0). \quad (18)$$

Proof: Write $x(t) = RCx(t) + (I-RC)x(t) = Ry(t) + (I-RC)Ax(t-1)$ so that $y(t+1) = CAx(t) = CARy(t) + CA(I-RC)Ax(t-1)$, and then recursively substitute $x(k)$ for $k = (t-1), \ldots, 0$. ∎

In the case of optimal prediction the maps are $A = P^T$, $C = \Psi_Y^T$ and $R = \Psi_X^T$, so the evolution of a distribution $h(t)$ in $\mathcal{C}_Y^*$ is given by

$$h(t+1) = \bar{P}^T h(t) + \sum_{k=0}^{t-1} \left[\Psi_X P\Big((I-\Psi_Y\Psi_X)P\Big)^{t-k}\Psi_Y\right]^T h(k) + \left[\Big((I-\Psi_Y\Psi_X)P\Big)^{t+1}\Psi_Y\right]^T x(0). \quad (19)$$

The leading-order term here is the optimal predictor chain $\bar{P}$, while the second and third terms are often called the "history" and "noise" terms, as they respectively represent contributions from the past observations and the unobservable components of the initial state.

This approach to predicting the output y is closely tied to the construction of observers in control theory. A standard construction is to let a state estimate $\hat{x}$ evolve according to

$$\hat{x}(t+1) = A\hat{x}(t) + R(y(t+1) - CA\hat{x}(t)),$$

driven by the measurements y. Notice that if the initial condition is correct, and the measurements are noise free, then this estimates the state exactly, for any R, A, and C. To see the connection with the Mori-Zwanzig representation, write the above filtering update rule as

$$\hat{x}(t+1) = (I-RC)A\hat{x}(t) + Ry(t+1)$$

and notice that it defines a convolution map from $y(1), \ldots, y(t)$ to $\hat{x}(t)$. Now let $\hat{y}(t+1)$ be the prediction $\hat{y}(t+1) = CA\hat{x}(t)$. Replacing $\hat{x}(t)$ by the convolution expression gives (18). It is also worth noting that, when the measurements are noisy, and when the the evolution of x is itself stochastic, the Kalman filter gives an optimal choice for R, in the sense that it minimizes the mean square error between $x(t)$ and $\hat{x}(t)$ for all t.

VII. Deterministic observations

We say that B is a deterministic observation if $B_{ij} \in \{0,1\}$ for all i, j. As B is a stochastic matrix, this means that exactly one entry in each row is non-zero, and so there is a map $b : X \to Y$ from states to observations.

We then have that $(Bf)(x) = f(b(x))$, and for $h \in \mathcal{C}_{X\times Y}$ we have $E[h(x,y)] = E[h(x,b(x))]$. Thus

$$\Psi_X g = \lim_{t\to\infty} \operatorname*{argmin}_f E\Big[\big(f(b(x(t))) - g(x(t))\big)^2\Big] \quad (20)$$

$$= \operatorname*{argmin}_f \big\|Bf - g\big\|_{M_X}^2 \quad (21)$$

From this we see that $\Psi_X g$ is the B-preimage of the M_X-orthogonal projection of g onto the range of B. To evaluate

this it is convenient to draw the commutative diagram

$$\begin{array}{ccc} (\mathcal{C}_X, M_X) & \xleftarrow{B} & (\mathcal{C}_Y, M_Y) \\ {\scriptstyle M_X}\downarrow & & \downarrow{\scriptstyle M_Y} \\ (\mathcal{C}_X^*, M_X^{-1}) & \xrightarrow[B^T]{} & (\mathcal{C}_Y^*, M_Y^{-1}) \end{array} \tag{22}$$

Hence the best approximator $\Psi_X = (B^T M_X B)^{-1} B^T M_Y = B^\dagger$ is the pseudo-inverse of B with respect to the M_X inner-product.

We thus have that $\bar{P} = B^\dagger P B$. Furthermore, $\mathbb{P} : \mathcal{C}_X^* \to \mathcal{C}_X^*$ defined by $\mathbb{P} = B^{\dagger T} B^T$ is the M_X^{-1}-orthogonal projector onto the range of $M_X B$. Using this we can write the Mori-Zwanzig representation as

$$h(t+1) = \bar{P}^T h(t) + \sum_{k=0}^{t-1} B^T \Big(P^T (I - \mathbb{P})\Big)^{t-k} P^T B^{\dagger T} h(k) + B^T \Big(P^T (I - \mathbb{P})\Big)^{t+1} x(0). \tag{23}$$

VIII. Examples

In this section, we briefly illustrate connections of the methods presented in the paper to well-known examples.

A. Example: Fluid mixing in a channel

We consider a passive mixing process in a long, thin channel. Fluids with two different colors (with intensities 1 and 0) are injected in one end of the channel and flow through the channel, driven by body force only. The channel walls are not flat, but rather have structure that passively mix the fluid. These structures are periodic with period ℓ_x and the cross section of the channel has dimension ℓ_y by ℓ_z. The channel is assumed to be long enough so that the velocity field inside is fully developed and hence also periodic with period ℓ_x. Such passive mixing channels have been physically constructed for microfluidic systems by [25].

Due to the periodicity of the velocity field, we can evolve the fluid color distribution by considering periodically spaced cross-sections X in the channel and computing the map $S : X \to X$ that maps one-period's in-flow cross-section to the out-flow cross-section. That is, if a fluid particle enters at the point x_0 then $S(x_0)$ is the point at which it exits the periodic channel segment. Iterating this map thus corresponds to the particle motion through the cross-sections down the channel.

Note that the map S contains no molecular diffusion information, so it causes no mixing of the fluid colors. One obvious way to simulate the system is to evolve the boundary of the two colors by S. However, this boundary will grow in length exponentially, making this very computationally expensive. As an alternative, it is common to use low-order Markov chain approximations. Markov chain models of systems such as fluid mixing are very convenient both analytically and numerically. They are cheap because only matrix-vector products are needed to compute with them, and because they are linear much useful information can be obtained from their spectra.

The mixing channel is shown in Figure 1. Simulation results for the exact boundary evolution and a Markov model approximation are shown in Figure 2 for increasing times.

One standard way of constructing a Markov chain approximation to the fluid map S is by forming the optimal prediction chain. Here we will work formally, and will assume that functions are Lebesgue-integrable and that measures are absolutely continuous with respect to the Lebesgue measure where necessary. Let $S : X \to X$ be the fluid map and π satisfying $S\pi = \pi$ be a non-vanishing invariant distribution.

For real-valued functions c on X we define a map $d_h : c \mapsto \vec{c}$ to be a real-valued function $\vec{c}$ on a grid on X with $1/h^2$ cells, given by the conditional observation of c over each grid cell. For simplicity, we will use regular mesh grids on X with grid size h in each direction, giving a total of $1/h^2$ states indexed from 1 to $1/h^2$. The function $d_h c$ is thus defined as

$$d_h c = \vec{c} = [c_1 \ c_2 \ \dots \ c_{1/h^2}]^T, \tag{24}$$

where $\vec{c} \in \mathbb{R}^{1/h^2}$ and $\vec{c}_i = \int_{a_i} c(x)\,dx$ for $i = 1, 2, \dots, 1/h^2$, and a_i refers to the i_{th} grid cell.

In the following, we simply match the aformentioned with the results we have derived, taking observation space $Y = \mathbb{R}^{1/h^2}$. It is straightforward to verify that

$$\Psi_Y = d_h^T \tag{25}$$

$$\Psi_X = \text{diag}(d_h \pi)^{-1} d_h \text{diag}(\pi) \tag{26}$$

and hence, using (13) we have

$$\bar{P} = \text{diag}(d_h \pi)^{-1} d_h \text{diag}(\pi) P d_h^T \tag{27}$$

where P is the Frobenious-Perron operator of S. For a non-singular (invertible measure-preserving) map S, P satisfies

$$(Pc)(x) = c(S^{-1}(x)) \tag{28}$$

for any function c on X.

By noting that $P^T \mu = S^{-1}(\mu)$ for any probability measure μ on X, it is easier to obtain $\bar{P}^T$ first, i.e.,

$$\bar{P}^T = d_h P^T \text{diag}(\pi) d_h^T \text{diag}(d_h \pi)^{-1} \tag{29}$$

and replace the P^T above by S^{-1} in computation.

Moreover, one can determine the entries of $\bar{P}$ explicitly as

$$\bar{P}_{ij} = \frac{\int_{S^{-1}(a_i) \cap a_j} \pi(x)\,dx}{\int_{a_i} \pi(x)\,dx} \text{ for } i, j = 1, \dots, 1/h^2. \tag{30}$$

This Markov chain model is simply a finite dimensional linear advection model of the underlying system with first order accuracy, and a sample trajectory is plotted in Figure 2.

B. Example: Random walk on a hypercube and Ehrenfests' Urn

In this example and the next we consider cases in which the reduction is exact. That is,

$$(\Psi_X^T \Psi_Y^T - I)(P^T)^t x(0) = 0 \text{ for } t = 0, 1, \dots \tag{31}$$

so $x(t)$ evolves in the null space of $\Psi_X^T \Psi_Y^T - I$ and the state can be reconstructed exactly from the observations.

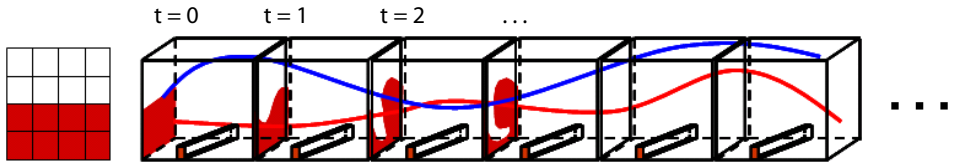

Fig. 1. A microfluidic mixing channel model. The channel geometry is periodic and the periodic flow solution is taken. The red and blue lines are representative fluid streamlines, the left half-red grid is the input fluid colors, and t denotes the different cross-sections of the channel.

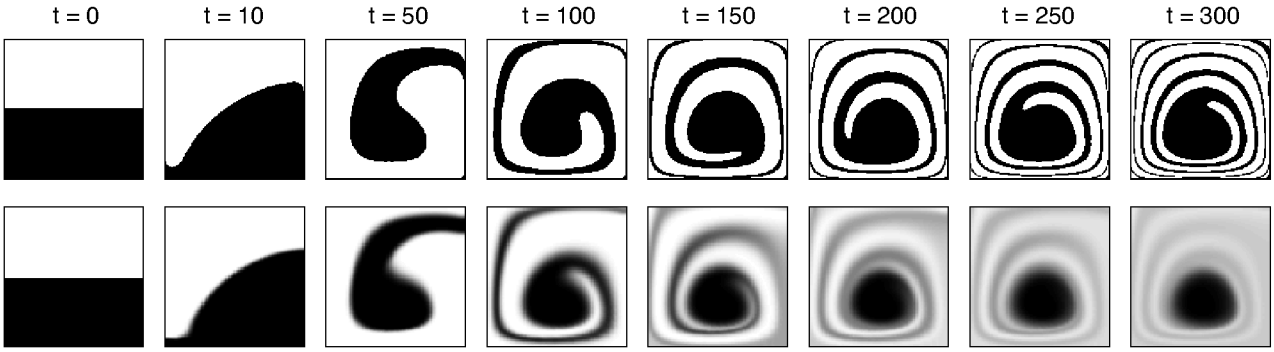

Fig. 2. Cross sections of the mixing channel at times $t = 0$, 10, 50, 100, 150, 200, and 250. The top row is the exact advection of the color boundary, computed by integrating the streamlines. The bottom row is the simulation of the reduced Markov chain model of the particle map, where diffusion occurs because of the inexactness of the reduced Markov chain $\bar{P}$.

The selection of the observation matrix B to achieve exact reduction is clearly related to the initial state $x(0)$. The appropriate choice of B with respect to some given $x(0)$ may not be straightforward, but can lead to the exact reduction of a large Markov chain for which the reduced system maintains certain desirable properties. Such techniques are widely used in the study of Markov chain problems; we see here two examples.

Consider the random walk on an n-dimensional hypercube: a particle starts at $\mathbf{0}$ and moves to one of its nearest neighbors, or stays fixed, with equal probability at each step. This process can be modeled simply as a Markov chain with 2^n states [5]. However, one can also observe 1-norm of each state. For example, on a 3-D cube, the corners with coordinates $(0,0,0)$, $(0,1,0)$ and $(1,0,1)$ would correspond to observations 0, 1 and 2, respectively. The observation matrix B is thus

$$B_{ij} = \begin{cases} 1 & \text{if the } i\text{th corner has 1-norm } j \\ 0 & \text{otherwise} \end{cases}$$

This observation maps 2^n states to $n+1$ observation values. The resulting optimal predictor chain on $\mathbb{Z}_{n+1}$ is called Ehrenfests' Urn and has been widely studied [9]. Simulating both the original and reduced system, starting at the initial state $\mathbf{0}$ and 0, respectively, gives

$$y(t) = B^T x(t) \text{ for } t = 0, 1, \ldots$$

That is, for this particular initial condition, the optimal prediction chain is exact.

C. Example: Card riffle-shuffling

Riffle shuffling is the most common method of shuffling cards. A deck of n cards is cut into two parts and the parts are riffled together. The basic model of this procedure was introduced by Gilbert and Shannon [6]. A review of related studies and an interesting phenomenon called *cutoff* can be found in [4]. In this model, the deck is cut into two piles according to the binomial distribution so the chance that pile one has j cards is $\binom{n}{j}/2^n$. One then sequentially drops cards from the bottom of each of the two piles according to the following rule: if at some stage pile one has A cards and pile two has B cards, drop the next card from pile one with probability $A/(A+B)$. This is continued until the two piles are exhausted, resulting in a shuffled pile of cards, and then repeated. This process defines a Markov chain with $n!$ states (i.e., permutations of card order). It is much harder to determine what observation matrix B one should choose in this problem. However, Bayer and Diaconis[1] show that it is enough to study only the number of rising sequences of the permutations, i.e, consider the following deterministic observation B,

$$B_{ij} = \begin{cases} 1 & \text{if the } i\text{th permutation has } j \text{ rising sequences} \\ 0 & \text{otherwise.} \end{cases}$$

The reduced system has n states only, and again, it can be shown that when the initial permutation has only one rising sequence, (31) is satisfied and so the resulting optimal predictor chain is exact.

IX. Conclusions

In this paper we have presented a review of several closed related approaches to model reduction of Markov chains, and discussed example applications. These methods view the system dynamics in terms of large and small scales, where only the large scale dynamics is explicitly retained and a conditional expectation is used to describe the affects of the small scale dynamics on the large scale dynamics. Alternatively, the well-known process of decomposing a Markov chain model is based on a time-scale separation where it is assumed one subset of the Markov chain is transient, or non-essential, and hence is dropped for long-run simulations. One comparative advantage to using the optimal prediction methods we describe is thus that statistical information from the small scale are available for use in the evaluating the large scale dynamics.

References

[1] Dave Bayer and Persi Diaconis. Trailing the dovetail shuffle to its lair. *The Annals of Applied Probability*, 2(2):294–313, 1992.

[2] A. J. Chorin and O. H. Hald. *Stochastic Tools in Mathematics and Science*. Springer Science and Business, 2006.

[3] A. J. Chorin, O. H. Hald, and R. Kupferman. Optimal prediction and the Mori-Zwanzig representation of irreversible processes. *Proceedings of the National Academy of Sciences*, 97:2968–2973, 2000.

[4] Persi Diaconis. *Groups, combinatorics and geometry*, chapter Mathematical developments from the analysis of riffle shuffling, pages 73–97. World Scientific Publishing, 2001.

[5] Persi Diaconis, R. L. Graham, and J. A. Morrison. Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random Structures and Algorithms*, 1(1):51–72, 1990.

[6] Gilbert E. Theory of shuffling. Technical report, Bell Laboratories, 1955.

[7] D. Enns. *Model Reduction for Control System Design*. PhD thesis, Stanford University, 1984.

[8] D. Evans and G. Morriss. *Statistical mechanics of nonequilibrium liquids*. Cambridge University Press, 2008.

[9] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley, New York, 2nd edition, 1971.

[10] M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovakian Math. J.*, 25:619–633, 1975.

[11] K. Glover. All optimal Hankel-norm approximations of linear multivariable systems and their l_∞ error bounds. *Int. Journal of Control*, 39:1115–1193, 1984.

[12] D. Hinrichsen and A. J. Pritchard. An improved error estimate for reduced order models of discrete-time systems. *IEEE Transactions on Automatic Control*, 35:317–320, 1991.

[13] J. G. Kemeny and J. L. Snell. *Finite Markov chains*. Springer-Verlag, 1983.

[14] G. Kotsalis, A. Megretski, and M. Dahleh. Balanced truncation for a class of stochastic jump linear systems and model reduction for Hidden Markov Models. *IEEE Transactions on Automatic Control*, 53:2543–2557, 2008.

[15] L. Mitiche, A.B.H. Adamou-Mitiche, and Daoud Berkani. Low-order model for speech signals. *Signal Processing*, 84:1805–1811, 2004.

[16] B. C. Moore. Principle component analysis of linear systems: Controllability, observability and model reduction. *IEEE Transactions on Automatic Control*, 26:17–32, 1981.

[17] H. Mori. Transport, collective motion, and Brownian motion. *Prog. Theor. Phys*, 33(3):423–455, 1965.

[18] J. S. Niedbalski, K. Deng, P. G. Mehta, and S. Meyn. Model reduction for reduced order estimation in traffic models. In *Proceedings, American Control Conference*, 2008.

[19] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[20] R. G. Phillips and P. V. Kokotovic. A singular perturbation approach to modeling and control of Markov chains. *IEEE Transactions on Automatic Control*, 26:1087–1094, 1981.

[21] Z. Ren and B. H. Krogh. State aggregation in Markov Decision Processes. In *Proceedings, IEEE Conference on Decision and Control*, 2002.

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[23] H. A. Simon and A. Ando. Aggregation of variables in dynamic systems. *Econometrica*, 29:111–138, 1961.

[24] R. Srikant. *The Mathematics of Internet Congestion Control*. Birkhauser, 2004.

[25] Abraham D. Stroock, Stephan K. W. Dertinger, Armand Ajdari, Igor Mezić, Howard A. Stone, and George M. Whitesides. Chaotic mixer for microchannels. *Science*, 295(5555):647–651, January 2002.

[26] L. B. White, R. Mahony, and G. D. Brusche. Lumpable Hidden Markov Models — Model reduction and reduced complexity filtering. *IEEE Transactions on Automatic Control*, 45:2297–2306, 2000.

[27] G. Yin, Q. Zhang, and G. Badowski. Decomposition and aggregation of large-dimensional Markov chains in discrete time. In *Proceedings, IEEE Conference on Decision and Control*, 2001.

[28] R. Zwanzig. Problems in nonlinear transport theory. In L. Garrido, editor, *Systems far from equilibrium*. Springer, Berlin, 1980.