2012 American Control Conference
Fairmont Queen Elizabeth, Montréal, Canada
June 27-June 29, 2012

Model Reduction of Markov Chains via Low-Rank Approximation

Kun Deng and Dayu Huang

Abstract— **This paper is concerned with the model reduction for Markov chain models. The goal is to obtain a *low-rank approximation* to the original Markov transition matrix. A nuclear-norm regularized optimization problem is proposed for this purpose, in which the Kullback-Leibler divergence rate is used to measure the similarity between two Markov chains, and the nuclear norm is used to approximate the rank function. An efficient iterative optimization algorithm is developed to compute the solution to the regularized problem. The effectiveness of this approach is demonstrated via numerical examples.**

I. Introduction

Markov chain models are important modeling tools to a number of applications in electrical engineering, system biology, computer science, and economics [1]. A fundamental problem arising from practice is the large dimension of the state space, which causes computational difficulties in *decision making* (e.g. Markov Decision Process), *performance analysis* (e.g. queuing networks or multi-agent systems), and *simulation* (e.g. complex dynamical systems), etc. Methods and algorithms are needed for reducing the complexity of Markov chains without incurring much loss of information or performance.

The most straightforward approach to the problem of model reduction for Markov chains is based on the aggregation of *states* [2]. After aggregation, a reduced Markov chain with a smaller number of *super-states* is obtained. In our recent work [3], [4], the aggregation-based method is studied and various algorithms are developed accordingly. It is shown that the reduced Markov model can be easily obtained once the optimal aggregation is found. But finding the exact optimal aggregation requires computing the solution to a combinatorial optimization problem whose complexity grows exponentially fast in the size of the state space. In the prior work [4], a heuristic based on non-convex relaxations is justified for *nearly completely decomposable* Markov chains.

In this paper, we propose another approach to mitigate the complexity of large-scale Markov chains. The idea is to find a smaller number of principal "features" to represent or approximate the Markov chain. Motivated by the Singular Value Decomposition method, we choose the features as certain singular-values and vectors, which may be not exactly the same as those of original chain. The number of the features is specified by the *rank* of transition matrix.

The goal of this paper is to find reduced Markov chains that closely approximate the original chain but have much lower rank. The closeness of two Markov chains is quantified by the Kullback-Leibler divergence rate (K-L rate) [5], which is widely used by many researchers for comparing Markov chains, particularly for model reduction problems [3], [6], [7]. To obtain low-rank approximations, we need to work with the rank function, a non-smooth non-convex function. No efficient algorithm can find the exact optimal solution. Using nuclear norm as a convex surrogate for the rank function, we propose a convex optimization problem: We find the transition matrices that minimizes the K-L rate plus the nuclear-norm regularization term. An efficient iterative algorithm based on proximity operators is developed to compute the optimal solution.

The rest of the paper is organized as follows: In Section II, we give the exact definition of low rank approximations and the formulation of the nuclear-norm regularized optimization problem. In Section III, we describe an efficient proximal algorithm used to compute the solution to the optimization problem. The effectiveness of the proposed approach is demonstrated via numerical examples in Section IV.

A. Notations

Let $\mathbb{M}$ denote the Hilbert space of all $n \times n$ real-valued matrices, equipped with the inner product $\langle Y, Z\rangle := \mathrm{Tr}(Y^\top Z) = \sum_i \sum_j Y_{ij} Z_{ij}$. The norm induced by this inner product is the *Frobenius norm* $\|\cdot\|_F$.

$$\|Y\|_F := \langle Y, Y\rangle^{\frac{1}{2}}.$$

The *operator norm* of a matrix is equal to its largest singular value:

$$\|Y\| := \sigma_1(Y).$$

The *nuclear norm* of a matrix is equal to the sum of all its singular values, i.e.,

$$\|Y\|_* := \sum_{i=1}^{r} \sigma_i(Y).$$

Let C be a nonempty subset of $\mathbb{M}$. The indicator function of C is

$$\iota_C(x) = \begin{cases} 0, & \text{if } x \in C \\ +\infty, & \text{if } x \notin C. \end{cases}$$

If C is closed and convex, then the projection of $Y \in \mathbb{M}$ onto C is unique and denoted by $\mathcal{P}_C(Y)$.

II. Low-rank Approximation of Markov Chains

A. Markov chains and its spectral decomposition

We consider a first-order discrete-time homogeneous Markov chain $\{x_k\}_{k\geq 0}$ evolving on a finite state space $\mathsf{N} =$

The authors are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 West Main Street, Urbana, IL 61801. Emails: {kundeng2,dhuang8}@illinois.edu.

$\{1, \ldots, n\}$ with the initial distribution ν and the transition matrix P. For any integer time $k \geq 0$,

$$\text{Prob}(x_{k+1} = j \mid x_k = i) = P_{ij}, \quad i, j \in \mathsf{N}.$$

Assumption 1 *All Markov chains considered in this paper are assumed to have* positive *transition matrix, i.e., $P_{ij} > 0$ for any $i, j \in \mathsf{N}$, or $P > 0$ in short.*

A Markov chain satisfying Assumption 1 is irreducible and aperiodic, thus has a unique *invariant distribution* π satisfying $\pi_i > 0$ for any $i \in \mathsf{N}$ and

$$\pi^\top P = \pi^\top.$$

Assumption 2 *All Markov chains considered in this paper are assumed to be* reversible *with respect to its invariant distribution π, i.e., $\Pi P = P^\top \Pi$, where $\Pi := \text{diag}(\pi)$.*

As shown in [1], all eigenvalues of a reversible Markov chain are real-valued. Denote them by $\{\lambda_1, \ldots, \lambda_n\}$, sorted by their magnitude in a decreasing order. It follows from Assumption 1 and the Perron-Frobenius theorem [1] that $\lambda_1 = 1$ and $u^1 = \mathsf{1}$ where $\mathsf{1}$ is a vector with all one entries, and $|\lambda_i| < 1$ for $i \geq 2$. Consider the n-dimensional real space $\mathbb{R}^n$ with the *π-weighted inner product*: For any $u, v \in \mathbb{R}^n$,

$$\langle u, v \rangle_\pi := u^\top \Pi v.$$

The right-eigenvectors $\{u^1, \ldots, u^n\}$ is chosen to form an orthonormal basis of the space $\mathbb{R}^n$ under the π-weighted inner product, i.e., $\langle u^i, u^j \rangle_\pi = \delta_{ij}$, for any $i, j \in \mathsf{N}$, where the function $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise. If the rank of P is r, then there are r nonzero eigenvalues. The spectral decomposition of P is given by

$$P = \sum_{i=1}^{r} \lambda_i u^i (u^i)^\top \Pi. \tag{1}$$

For a given Π, we introduce the mapping $\mathcal{S} : \mathbb{M} \to \mathbb{M}$, where $\mathcal{S}(Y) := \Pi^{\frac{1}{2}} Y \Pi^{-\frac{1}{2}}$ for $Y \in \mathbb{M}$. Since $\Pi^{\frac{1}{2}}$ has full rank, $\mathcal{S}$ is bijective. We denote its inverse operator by $\mathcal{S}^{-1}(Y) := \Pi^{-\frac{1}{2}} Y \Pi^{\frac{1}{2}}$. For a reversible transition matrix P, the spectral decomposition of $\mathcal{S}(P)$ is obtained from (1):

$$\mathcal{S}(P) = \sum_{i=1}^{r} \lambda_i \tilde{u}^i (\tilde{u}^i)^\top \tag{2}$$

where $\tilde{u}^i = \Pi^{\frac{1}{2}} u^i$. Clearly, $\mathcal{S}(P)$ is a symmetric matrix and its singular values are equal to absolute values of the nonzero eigenvalues of P [8]. If the rank of P is r, then the rank of $\mathcal{S}(P)$ is also r; and the singular value decomposition of $\mathcal{S}(P)$ can be taken as $\mathcal{S}(P) = U \Sigma V^\top$ with $U = [\tilde{u}^1 | \ldots | \tilde{u}^r]$, $\Sigma = \text{diag}(\{|\lambda_i|\}_{1 \leq i \leq r})$, $V = [\text{sign}(\lambda_1)\tilde{u}^1 | \ldots | \text{sign}(\lambda_r)\tilde{u}^r]$. The operator norm of $\mathcal{S}(P)$ satisfies

$$\|\mathcal{S}(P)\| = 1. \tag{3}$$

B. Low-rank approximation

Recall that the objective of this paper is to obtain a *low-rank approximation* of the given Markov chain. This approximation needs to preserve some important properties of the original Markov chain:

Definition 1 *Consider a reversible Markov chain defined on the state space* N *with the transition matrix P. Let π denote its invariant distribution and $r = \mathsf{rank}\,(P)$. Given an integer $1 \leq m \leq r$, a* low-rank approximation *or* rank-m approximation *of the original chain is a Markov chain defined on* N *whose the transition matrix Q satisfies the following requirement:*

- Rank*: $\mathsf{rank}\,(Q) = m$.*
- Invariant distribution*: $\pi^\top Q = \pi^\top$.*
- Probability transition matrix*: $Q \geq 0$ and $Q\mathsf{1} = \mathsf{1}$.*

For a given Markov chain, there are infinitely many transition matrices that satisfy the above three requirements. Among them, we want to find those that "best" approximate the original Markov chain with respect to some metrics.

When the Frobenius norm is used as the metric, it is well-known from the Eckart-Young Theorem [9] that the following matrix minimizes the metric among all matrices satisfying the rank and invariant distribution requirements:

$$Q^\star = \sum_{i=1}^{m} \lambda_i u^i (u^i)^\top \Pi.$$

However, in general, $Q^\star$ is not a *non-negative matrix* [10]; hence it does not satisfy the third requirement in Definition 1.

In this paper, we use the Kullback-Leibler divergence rate function as the metric. The problem of finding the optimal low-rank approximation can be posed as the following optimization problem:

$$\begin{array}{ll} \min\limits_{Q \in \mathbb{M}} & R(P\|Q) \\ \text{s.t.} & Q\mathsf{1} = \mathsf{1}, \quad \pi^\top Q = \pi^\top \\ & Q \geq 0, \quad \mathsf{rank}\,(Q) \leq m \end{array} \tag{4}$$

where the K-L rate between two Markov chains is given by [5]

$$R(P\|Q) := \sum_{i,j \in \mathsf{N}} \pi_i P_{ij} \log\left(P_{ij}/Q_{ij}\right).$$

C. Nuclear norm regularization problem

The optimization problem (4) is not a convex programming problem due to the non-convex constraint $\mathsf{rank}\,(Q) \leq m$. To our best knowledge, there does not exist an efficient algorithm to find the global optimum of (4).

Our approach to approximate the solution to (4) is to employ a heuristic method based on the nuclear norm, a *convex approximation* of the rank function. As shown in [11], the nuclear norm is a convex lower bound of the rank function on the unit ball defined with respect to the operator norm. In fact, it is the tightest convex lower bound:

Lemma 1 (Theorem 2.2 of [11]) *The nuclear norm* $\|Y\|_*$ *is the* convex envelope *of the function* $\mathsf{rank}\,(Y)$ *on the set* $\{Y \in \mathbb{M} : \|Y\| \leq 1\}$.

For a transition matrix Q, we have $\mathsf{rank}\,(\mathcal{S}(Q)) = \mathsf{rank}\,(Q)$ and $\|\mathcal{S}(Q)\| = 1$ (given in (3)). Thus $\|\mathcal{S}(Q)\|_*$ is the convex envelop of the function $\mathsf{rank}\,(\mathcal{S}(Q))$. As an approximation to the original rank-constrained optimization problem, we consider the following *nuclear norm regularization problem*:

$$\begin{array}{ll} \min\limits_{Q\in\mathbb{M}} & R(P\|Q) + t\|\mathcal{S}(Q)\|_* \\ \text{s.t.} & Q\mathsf{1} = \mathsf{1}, \quad \pi^\top Q = \pi^\top. \end{array} \tag{5}$$

The regularization parameter $t \geq 0$ provides a tradeoff between the rank of approximation and the quality of approximation for any feasible solution Q. The optimization problem (5) is clearly convex. Note that the non-negativity constraint on Q (i.e., $Q \geq 0$) is automatically satisfied by the optimal solution because the domain of logarithm function is the positive real line.

D. Characterization of optimal solution

In the following theorem, the *uniqueness*, *positiveness*, and *reversibility* of the optimal solution to (5) is established for any fixed $t \geq 0$. The proof can be found in our technical report [12].

Theorem 2 *For any fixed* $t \geq 0$*, the optimization problem* (5) *admits a* unique *optimal solution* $Q^*(t)$*, which is a positive and reversible transition matrix.*

The following proposition characterizes the rank-1 solution of the problem (5). The proof can be found in our technical report [12].

Proposition 3 *The problem* (5) *attains a unique rank-one solution* $Q^*(t) = \mathsf{1}\pi^\top$ *if and only if* $t \geq |\lambda_2|$.

Unfortunately, we don't have rigorous results for the existence and the characterization of rank-m solutions with $m \geq 2$.

III. Proximal Splitting Algorithm

The regularized problem (5) is a convex optimization problem. For relatively small-scale problems, the interior point algorithms based on semi-definite programming can be directly applied to solve (5) with high efficiency and accuracy. But these algorithms require the *second-order information* to solve a suitable Newton system for descent directions [8]. The memory requirements for computing descent directions quickly explode for practical large-scale Markov chains.

In this section, we propose an iterative algorithm that does not require the second-order information. It is based on *splitting proximity operators of convex functions*. Due to a favorable structure of the problem (5), the associated proximal operators have closed-form expressions (see Section III-D below for more details), making the proposed algorithm very efficient.

A. Unconstrained representation of optimization problem

Let $X = \mathcal{S}(Q)$. The set of feasible solutions of (5) is mapped to the following set via the operator $\mathcal{S}$:

$$\Omega := \{X \in \mathbb{M} : X\pi^{\frac{1}{2}} = \pi^{\frac{1}{2}},\ (\pi^{\frac{1}{2}})^\top X = (\pi^{\frac{1}{2}})^\top\}.$$

We then equivalently represent the nuclear norm regularization problem (5) in an unconstrained form:

$$\min_{X\in\mathbb{M}} \quad f_1(X) + f_2(X) + f_3(X) \tag{6}$$

where $f_1(X) := R(P\|\mathcal{S}^{-1}(X))$, $f_2(X) := t\|X\|_*$, and $f_3(X) := \iota_\Omega(X)$. The unique optimal solution to (6) is denoted by X^*. The optimal solution to (5) is then given by $Q^* = \mathcal{S}^{-1}(X^*)$.

A major difficulty in solving (6) stems from the fact that the functions f_2 and f_3 are *not differentiable*, which rules out conventional smooth optimization techniques [8]. The *proximal splitting method* [13]–[15], on the other hand, does not require the computation of (sub)gradients of f_1, f_2 or f_3. Instead each iteration involves the evaluation of *proximity operators*.

B. Proximity operator

The *proximity operator* of a convex function is a natural extension of the notion of a *projection operator* [16]: The projection $\mathcal{P}_C(Y)$ of $Y \in \mathbb{M}$ onto the nonempty closed convex set $C \subset \mathbb{M}$ is the unique solution to the following problem

$$\min_{Z\in\mathbb{M}} \iota_C(Z) + \tfrac{1}{2}\|Y - Z\|_F^2. \tag{7}$$

The proximity operator is defined by replacing ι_C in (7) with an arbitrary lower semicontinuous convex function f:

Definition 2 *Let* f *be a lower semicontinuous convex function. The* proximity operator of f $\mathrm{prox}_f : \mathbb{M} \to \mathbb{M}$ *is defined as:*

$$\mathrm{prox}_f(Y) := \arg\min_{Z\in\mathbb{M}} \left\{ f(Z) + \tfrac{1}{2}\|Y - Z\|_F^2 \right\}.$$

The value of the proximity operator of f satisfies the relation:

$$Z = \mathrm{prox}_f(Y) \quad \Leftrightarrow \quad Y - Z \in \partial f(Z) \tag{8}$$

where ∂f denotes the subdifferential of the function f [8].

C. Proximal splitting algorithm

Based on ideas proposed in [15], we develop a proximal splitting algorithm for solving the problem (6) as follows: First note that all three functions f_1, f_2, and f_3 in the objective function of (6) are proper, convex, and lower semicontinuous. A matrix $X \in \mathbb{M}$ is an optimal solution to (6) if and only if

$$\mathsf{0} \in \partial f_1(X) + \partial f_2(X) + \partial f_3(X). \tag{9}$$

Using operator splitting techniques [13], we obtain the following equivalent conditions to (9): For any $\gamma > 0$, there

exist $B_i, W_i \in \mathbb{M}$ for $i = 1, 2, 3$ such that

$$X = \tfrac{1}{3}(W_1 - B_1 + W_2 - B_2 + W_3 - B_3), \quad (10)$$
$$X = W_1, \quad X = W_2, \quad X = W_3, \quad (11)$$
$$(B_1 + X - W_1) \in \gamma \partial f_1(W_1), \quad (12)$$
$$(B_2 + X - W_2) \in \gamma \partial f_2(W_2), \quad (13)$$
$$(B_3 + X - W_3) \in \gamma \partial f_3(W_3). \quad (14)$$

Since f_1, f_2, and f_2 are lower semicontinuous convex functions, we can further represent (12)–(14) using proximity operators (see (8)):

$$W_1 = \mathrm{prox}_{\gamma f_1}(B_1 + X), \quad (15)$$
$$W_2 = \mathrm{prox}_{\gamma f_2}(B_2 + X), \quad (16)$$
$$W_3 = \mathrm{prox}_{\gamma f_3}(B_3 + X). \quad (17)$$

Equations (10), (11), and (15)–(17) form a set of fixed point equations involving proximity operators. This leads to the following fixed point iterative scheme in [15], which can be used to solve (6) with a fixed $t \geq 0$:

Proximal Splitting Algorithm
Initialization: Start with matrices $B_1^{(0)} = B_2^{(0)} = B_3^{(0)} = \mathsf{0}$, and $W_1^{(0)} = W_2^{(0)} = W_3^{(0)} = P$, scalars $\gamma > 0$ and $\epsilon_{\mathrm{tol}} > 0$.
Loop: For $k = 0, 1, \ldots$, repeat until a stopping criterion is reached

$$X^{(k+1)} = \tfrac{1}{3}(W_1^{(k)} - B_1^{(k)} + W_2^{(k)} - B_2^{(k)} + W_3^{(k)} - B_3^{(k)}),$$
$$W_1^{(k+1)} = \mathrm{prox}_{\gamma f_1}(B_1^{(k)} + X^{(k+1)}),$$
$$W_2^{(k+1)} = \mathrm{prox}_{\gamma f_2}(B_2^{(k)} + X^{(k+1)}),$$
$$W_3^{(k+1)} = \mathrm{prox}_{\gamma f_3}(B_3^{(k)} + X^{(k+1)}),$$
$$B_1^{(k+1)} = B_1^{(k)} + X^{(k+1)} - W_1^{(k+1)},$$
$$B_2^{(k+1)} = B_2^{(k)} + X^{(k+1)} - W_2^{(k+1)},$$
$$B_3^{(k+1)} = B_3^{(k)} + X^{(k+1)} - W_3^{(k+1)}.$$

Stopping criterion: Stop the algorithm if $\|X^{k+1} - X^k\|_F / \|X_k\|_F < \epsilon_{\mathrm{tol}}$.
Output: $\widehat{X}^{\mathrm{opt}} := X^{(k+1)}$ is taken as the approximate optimal solution to (6).

Due to the *nonexpansive property* of the proximity operators [13], the sequence $\{X_k\}_{k \in \mathbb{N}}$ generated by the proximal splitting algorithm converges to the unique solution of (6) as $\epsilon_{\mathrm{tol}} \to 0$, for any $\gamma > 0$ (see [15] for rigorous proofs).

D. Closed-form expressions of proximity operators

In this section, we show that proximity operators used in the proximal splitting algorithm all have closed-form expressions; hence this algorithm can be efficiently implemented in practice.

1) The proximity operator associated with the K-L rate: Note that the function $f_1(Y)$ is differentiable with respect to Y. The following lemma follows from the first-order optimality condition for the optimization problem in the definition of $\mathrm{prox}_{\gamma f_1}$. The proof can be found in our technical report [12].

Lemma 4 *For any $\gamma > 0$ and $Y \in \mathbb{M}$,*

$$(\mathrm{prox}_{\gamma f_1}(Y))_{ij} = \tfrac{1}{2}\left(Y_{ij} + \sqrt{Y_{ij}^2 + 4\gamma \pi_i P_{ij}}\right), \quad i, j \in \mathsf{N}.$$

2) The proximity operator associated with the nuclear norm: Consider the singular value decomposition of a matrix $Y \in \mathbb{M}$ of rank r

$$Y = U \Sigma V^\top, \quad \Sigma = \mathrm{diag}(\sigma_i)$$

where U and V are both $n \times r$ matrices with orthonormal columns, and the singular values $\{\sigma_i\}_{1 \leq i \leq r}$ are all positive. For each $\tau \geq 0$, the *singular value shrinkage operator* defined in [14] is given by

$$\mathcal{D}_\tau(Y) := U \Sigma_\tau V^\top, \quad \Sigma_\tau = \mathrm{diag}((\sigma_i - \tau)_+).$$

This operator applies a soft-thresholding to all singular values of Y, effectively shrinking them towards zero. The following lemma shows that the proximity operator associated with the nuclear norm is equal to a singular value thresholding operator. The proof can be found in various recent references (see e.g. Theorem 2.1 of [14]).

Lemma 5 *For any $\gamma > 0$ and $Y \in \mathbb{M}$,*

$$\mathrm{prox}_{\gamma f_2}(Y) = \mathcal{D}_{t\gamma}(Y).$$

In particular, note that if many of the singular values of Y are below the threshold τ, the rank of $\mathcal{D}_\tau(Y)$ may be considerably lower than that of Y. This is essential for the proposed algorithm to generate low-rank solutions.

3) The proximity operator associated with the indicator function: Since Ω is a closed affine subspace of $\mathbb{M}$, the projection of any $Y \in \mathbb{M}$ onto Ω is a unique point $\mathcal{P}_\Omega(Y)$ in Ω. The proximity operator associated with the indicator function $f_3 = \iota_\Omega$ is equal to a projection operator:

Lemma 6 *For any $\gamma > 0$ and $Y \in \mathbb{M}$,*

$$\mathrm{prox}_{\gamma f_3}(Y) = \mathcal{P}_\Omega(Y).$$

Note that

$$Y \in \Omega \quad \Longleftrightarrow \quad A\ \mathrm{vec}(Y) = b$$

where

$$A = \begin{bmatrix} I \otimes (\pi^{\frac{1}{2}})^\top \\ (\pi^{\frac{1}{2}})^\top \otimes I \end{bmatrix}, \quad b = \begin{bmatrix} \pi^{\frac{1}{2}} \\ \pi^{\frac{1}{2}} \end{bmatrix},$$

with $\otimes$ denoting the Kronecker product of matrices, and $\mathrm{vec}(Y)$ denotes the $n^2 \times 1$ vector obtained by stacking the columns of Y together, and $\mathrm{vec}^{-1}(y)$ converts the $n^2 \times 1$ vector y back into an $n \times n$ matrix column by column. We have an explicit formula:

$$\mathcal{P}_\Omega(Y) := \mathrm{vec}^{-1}(\mathrm{vec}(Y) + A^\top (AA^\top)^{-1}(b - A\ \mathrm{vec}(Y))).$$

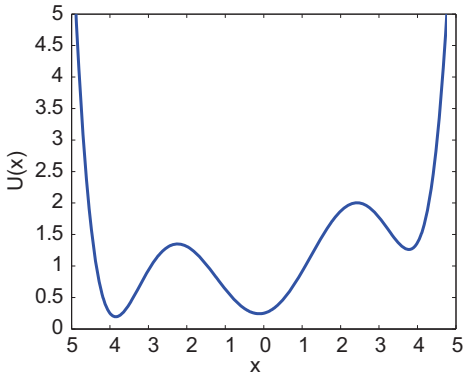


Fig. 1. Three-well potential function U.

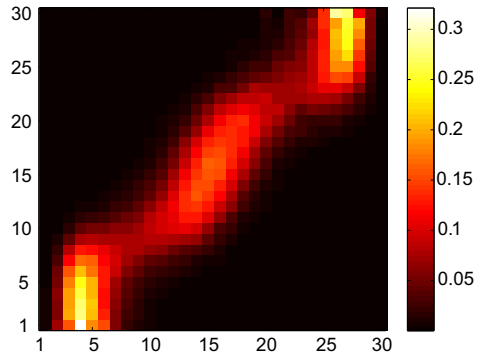


Fig. 2. A typical color plot of the 30×30 transition matrix P for the diffusion-induced Markov chain.

IV. Numerical Results

In this section, we use a numerical example to illustrate the theoretical results and demonstrate the algorithms described above.

We consider a continuous time continuous state space diffusion process $\{x_t\}_{t\geq 0}$ governed by the following stochastic differential equation:

$$dx_t = -\frac{1}{\alpha}\nabla U(x_t)dt + \frac{\beta}{\alpha}dw_t \tag{18}$$

where α and β are two parameters, $\{w_t\}_{t\geq 0}$ denotes a standard Brownian motion, and the potential $U : \mathbb{R} \to \mathbb{R}_+$ is a smooth function. The following three-well potential [17] is considered in this paper:

$$U(x) = \frac{1}{200}(0.5x^6 - 15x^4 + 119x^2 + 28x + 50)$$

which is depicted in Fig. 1.

For $\alpha, \beta > 0$, the diffusion process $\{x_t\}$ is an ergodic reversible Markov process. For numerical analysis, we choose $\alpha = 1.5$ and $\beta = 1.65$ in which the diffusion process exhibits clear occasional transition behaviors among three wells of U.

The SDE (18) is used to construct a finite-time finite-state space Markov chain as follows:

- First we simulate a discrete-time realization $\{x_k\}$ of the process $\{x_t\}$ in time interval $[0, 10000]$ by using Euler-Maruyama scheme with step size $\Delta t = 0.02$ and $K = 5 \times 10^5$ time steps.
- In the specific realization that we studied, the time series $\{x_k\}$ stay in the interval $[-5, 5]$. Then we uniformly discretize the interval $[-5, 5]$ into 30 equal-sized intervals $B_1, \ldots, B_{30}$. We take these 30 intervals as 30 states of the constructed Markov chain.
- From the time series, we compute the transition probabilities of the Markov chain between spatial discretization intervals with lag time $\tau = 1$; i.e., we compute the probability for going from B_i to B_j as

$$P_{ij} = \frac{\#\{t_k : x_k \in B_i, x_{k+\tau/\Delta t} \in B_j, k = 1, \ldots, K\}}{\#\{t_k : x_k \in B_i, k = 1, \ldots, K\}}$$

The color plot of the resulting 30×30 transition matrix P is shown in Fig. 2.

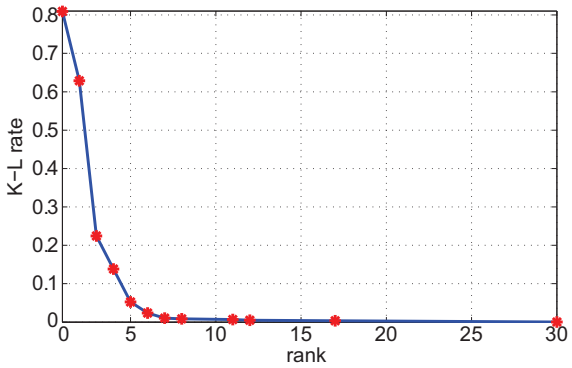
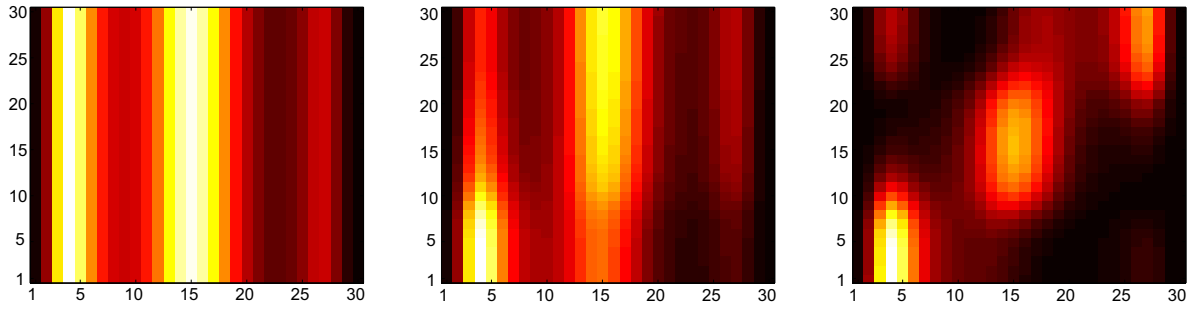


Fig. 3. The K-L rate of best rank-m approximations for $m = 1, \ldots, 30$.

In Fig. 2, we clearly observe three clusters centered around the intervals belonging to the vicinity of the minima of the potential function U, and a transition region between these clusters belonging to the vicinity of the saddle point in the potential function.

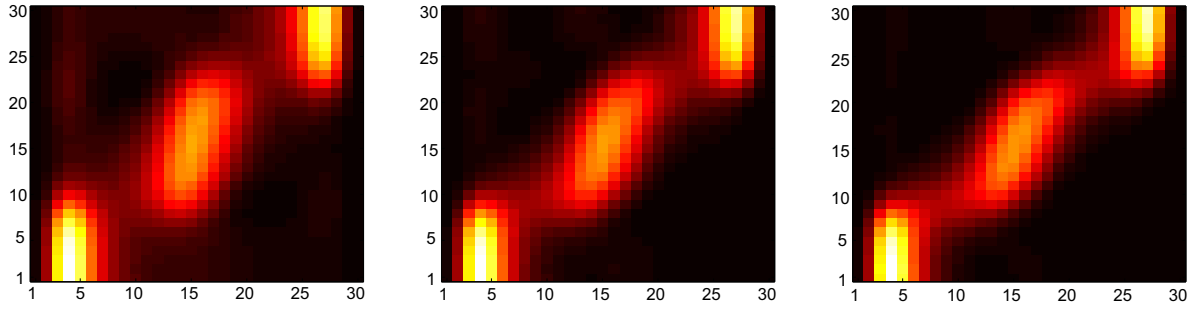
Applying the proximal splitting algorithm described in Section III-C to the 30×30 transition matrix P, we obtain low-rank approximations for different t. The optimal solutions $Q^*(t)$ have low-rank structures, and the rank of $Q^*(t)$ decreases monotonically with respect to the regulation parameter t. For $m = 1, \ldots, 30$, the best rank-m approximation is obtained at the minimal possible t_i such that the rank of the optimal solution $Q^*(t)$ is m with the smallest possible KL rate. In Fig. 4, we depicts the best rank-1, rank-2, and rank-3 approximations. In Fig. 5, we depict the best rank-4, rank-5, and rank-6 approximations.

In Fig. 3, we depict the K-L rate of the best rank-m approximations for $m = 1, \ldots, 30$. For low-rank approximations with $m \geq 7$, the approximation error is less than 0.0104, and the error decreases very slowly with respect to m. Thus one can take rank-6 approximation $Q^*(t_6)$ as a good candidate low-rank approximation to the constructed Markov chain.



(a) (b) (c)

Fig. 4. The color plot of transition matrices of (a) the rank-1 approximation, (b) the rank-2 approximation, (b) the rank-3 approximation of P.



(a) (b) (c)

Fig. 5. The color plot of transition matrices of (a) the rank-4 approximation, (b) the rank-5 approximation, and (c) the rank-6 approximation of P.

V. Conclusions

In this paper, we presented a model reduction method for Markov chains based on low-rank approximations. The Kullback-Leibler divergence rate was used to measure the similarity between the original Markov chain and the reduced one. The nuclear norm was used as a convex surrogate for the rank function. A nuclear-norm regularized optimization problem was proposed to approximately find the optimal low-rank approximations. Based on proximal splitting techniques, an efficient iterative optimization algorithm was developed to compute the optimal solutions to the regularized problem. We demonstrated the effectiveness of the proposed approach using a numerical examples.

Acknowledgment

The authors gratefully acknowledge Prof. Sean Meyn and Prof. Prashant Mehta in initially proposing the ideas of low-rank approximation of Markov chains.

References

[1] P. Bremaud, *Markov Chains – Gibbs Fields, Monte Carlo Simulation, and Queues*. New York, NY: Springer, 1999.

[2] H. A. Simon and A. Ando, "Aggregation of variables in dynamical systems," *Econometrica*, vol. 29, no. 2, pp. 111–138, 1961.

[3] K. Deng, Y. Sun, P. G. Mehta, and S. P. Meyn, "An information-theoretic framework to aggregate a Markov chain," in *Proceedings of American Control Conference*, St. Louis, MO, 2009, pp. 731–736.

[4] K. Deng, P. G. Mehta, and S. P. Meyn, "Optimal Kullback-Leibler aggregation via the spectral theory of Markov chains," *IEEE Trans. Automat. Contr.*, vol. 56, no. 12, pp. 2793–2808, 2011.

[5] 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.

[6] L. Xie, V. Ugrinovskii, and I. R. Petersen, "Probabilistic distances between finite-state finite-alphabet Hidden Markov Models," *IEEE Trans. Automat. Contr.*, vol. 50, no. 4, pp. 505–511, 2005.

[7] M. Vidyasagar, "Reduced-order modeling of Markov and Hidden Markov Processes via aggregation," in *Proceedings of IEEE Conference of Decision and Control*, Atlanta, GA, 2010, pp. 1810–1815.

[8] J. B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. New York, NY: Springer-Verlag, 2001.

[9] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[10] W. E, T. Li, and E. Vanden-Eijnden, "Optimal partition and effective dynamics of complex networks," *PNAS*, vol. 105, no. 23, pp. 7907–7912, 2008.

[11] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, 2002.

[12] K. Deng and D. Huang, "Optimal Kullback-Leibler aggregation via the spectral theory of Markov chains," 2012. [Online]. Available: https://netfiles.uiuc.edu/kundeng2/www/paper/dhACC12J.pdf

[13] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, pp. 1168–1200, 2005.

[14] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. OPTIM.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[15] S. Setzer, G. Steidl, and T. Teuber, "Deblurring poissonian images by split bregman techniques," *J. Vis. Commun. Image R.*, vol. 21, pp. 193–199, 2010.

[16] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *C. R. Acad. Sci. Paris Sér. A Math.*, vol. 255, pp. 2897–2899, 1962.

[17] 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.