Optimal Kullback–Leibler Aggregation via Information Bottleneck

Bernhard C. Geiger, *Member, IEEE*, Tatjana Petrov, Gernot Kubin, *Member, IEEE*, and Heinz Koeppl

***Abstract*—In this paper, we present a method for reducing a regular, discrete-time Markov chain (DTMC) to another DTMC with a given, typically much smaller number of states. The cost of reduction is defined as the Kullback–Leibler divergence rate between a projection of the original process through a partition function and a DTMC on the correspondingly partitioned state space. Finding the reduced model with minimal cost is computationally expensive, as it requires an exhaustive search among all state space partitions, and an exact evaluation of the reduction cost for each candidate partition. Our approach deals with the latter problem by minimizing an upper bound on the reduction cost instead of minimizing the exact cost. The proposed upper bound is easy to compute and it is tight if the original chain is *lumpable* with respect to the partition. Then, we express the problem in the form of information bottleneck optimization, and propose using the agglomerative information bottleneck algorithm for searching a suboptimal partition greedily, rather than exhaustively. The theory is illustrated with examples and one application scenario in the context of modeling bio-molecular interactions.**

***Index Terms*—Information bottleneck method, lumpability, Markov chain, model reduction.**

I. Introduction

MARKOV models are ubiquitously used in scientific and engineering disciplines, for example, to understand chemical reaction systems, to model speech recognition and data sources, or in Markov decision processes in automated control. The popularity of these models arises because the Markov property often renders model analysis tractable and their simulation efficient. However, sometimes the state space of a Markov model (i.e., its alphabet) is too large to permit simulation, even when harnessing today's computing power. Indeed, in stochastic modeling in computational biology [1], or in n-gram word models in speech recognition [2], dealing with the state space explosion is a major challenge. Also in control theory, particularly for nearly completely decomposable Markov chains, state space reduction is an important topic [3], [4].

A direct way of reducing the state space of a Markov chain is aggregation: With the help of a partition function, groups of nodes in the original transition graph are aggregated, resulting in a graph with a smaller number of nodes. The aggregated process, or *aggregation*, can be any Markov chain over this smaller transition graph, depending on how the transition probabilities are chosen. Another way of reducing the state space of a Markov chain is to project realizations of the original chain through the partition function. The process thus obtained is called the projected process, or *projection*. Ideally, for a given partition function, aggregated and projected process should coincide. However, as the projected process is generally not Markov, the aggregation "closest" to the projection is sought instead, where closeness has to be defined appropriately. In this paper, we quantify the distance between the projection and the aggregation by the Kullback–Leibler divergence rate (KLDR).

We focus on finding the optimal aggregation for a given alphabet size, i.e., on finding the partition function for which the KLDR between the projection and aggregation is minimized. Although, for a given partition function, the aggregation closest to the projection is easy to obtain (cf. [3] or Lemma 3 in this work), finding the optimal partition function remains computationally expensive because: (i) it requires an exhaustive search among all partitions of a given alphabet size and (ii) it requires evaluating the KLDR for each candidate partition. In our approach, we relax the latter problem to evaluating an upper bound on the KLDR. More precisely, the aggregated Markov chain is *lifted* to the original alphabet; the KLDR between the lifted and the original Markov chain provides an upper bound which can be evaluated analytically [3], [5]. Further relaxing the problem allows its expression in terms of the information bottleneck method [6], casting the problem of state space reduction in terms of a widely used machine learning technique. As a result, we propose using the information bottleneck method for finding a suboptimal partition function in a greedy manner, thus obviating the complexity of an exhaustive search for the cost of optimality.

A. Contributions and Related Work

In control theory, model reduction and, in particular, state space aggregation of Markov models is an important topic. For example, White *et al.* analyzed aggregation of Markov and hidden Markov models in [7]. In particular, they presented a

Manuscript received July 23, 2013; revised November 5, 2013, May 21, 2014, and October 7, 2014; accepted October 21, 2014. Date of publication October 24, 2014; date of current version March 20, 2015. This work was supported by the Austrian Research Association under Project 06/12684, by the Swiss National Science Foundation (SNSF) under Grant PP00P2 128503/1, by the SystemsX.ch (the Swiss Inititative for Systems Biology), and by a SNSF Early Postdoc.Mobility Fellowship grant P2EZP2_148797. Recommended by Associate Editor H. S. Chang.

B. C. Geiger is with the Institute for Communications Engineering, TU Munich, D-80333 Munich, Germany, and the Signal Processing and Speech Communication Laboratory, Graz University of Technology, 8010 Graz, Austria (e-mail: geiger@ieee.org).

T. Petrov is with the Automatic Control Lab, ETH Zürich, Zürich, Switzerland, and the IST Austria, 3400 Klosterneuburg, Austria.

G. Kubin is with the Signal Processing and Speech Communication Laboratory, Graz University of Technology, 8010 Graz, Austria (e-mail: g.kubin@ieee.org).

H. Koeppl is with the Department of Electrical Engineering and Information Technology, TU Darmstadt, 64289 Darmstadt, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2014.2364971

linear algebraic condition for lumpable chains (see Definition 3 on page 4) and determined, for a given partition function, the best aggregation in terms of the Frobenius norm. Given the transition matrix of a Markov chain, they obtained a bi-partition of its state space via alternating projection. Aldhaheri and Khalil considered optimal control of nearly completely decomposable Markov chains and adapted Howard's algorithm to work on an aggregated model [4]. The work of Jia considers state aggregation of Markov decision processes optimal w.r.t. the value function and provides algorithms which perform this aggregation [8]. Aggregation of Markov chains with information-theoretic cost functions was considered by Deng *et al.* [3] and Vidyasagar [9], the first reference being the main inspiration of our work. Deng and Huang used the KLDR as a cost function to obtain a low-rank approximation of the original transition matrix via nuclear-norm regularization, thus preserving the cardinality of the state space [10].

The idea of lifting the aggregated chain to the original state space was used in, e.g., Deng *et al.* [3] and Katsoulakis *et al.* [11]. In [11], the authors realized that the Kullback–Leibler divergence between the resulting Markov chains provides an upper bound on the reduction cost; however, their work is focused on continuous-time Markov chains, which makes a detailed comparison with our work difficult. Compared to [3], our approach differs in the definition of the lifting and its consequences. More precisely, the lifting we use incorporates the one-step transition probabilities of the original chain, while the authors of [3] define lifting based only on the stationary distribution of the original chain. Consequently, while Deng *et al.* maximize the redundancy of the aggregated Markov chain, the lifting proposed here minimizes *information loss* in a well-defined sense. Moreover, the upper bound we obtain is better than the upper bound obtained in [3], and it is tight if the original chain is lumpable.

The connection to spectral graph theory observed in [3] does not apply in our case, to the best of our knowledge. More precisely, for nearly completely decomposable Markov chains, the optimal bi-partition of the alphabet is determined by the sign structure of the Fiedler vector, the eigenvector associated with the second largest eigenvalue. Despite spectral graph theory being employed for model reduction and Markov chain aggregation for some time (e.g., [12], [13]), the authors of [3] first showed a connection between this eigenvector-based aggregation method and an information-theoretic cost function.

In summary, by introducing a different lifting, we lose the connection to eigenvector-based aggregation, but instead gain the following:

1) Our lifting minimizes an upper bound on the KLDR between the projection and the aggregation, subject to the requirement that the lifted chain is lumpable.
2) The upper bound we obtain is tight if the original chain is lumpable.
3) Minimizing the upper bound proposed by our lifting minimizes information loss in a well-defined sense; this minimization, loosely speaking, yields the partition w.r.t. which the original chain is "most lumpable."
4) Relaxating the cost function allows applying the information bottleneck method for state space aggregation.

The connection to the information bottleneck method is most interesting: Recently, Vidyasagar investigated a metric between distributions on sets of different cardinalities [14], a problem very similar to the one considered in this work. He proposed an information-theoretic metric called the variation of information, and showed that the optimal reduced-order distribution on a set of given cardinality is obtained by *projecting* the original distribution. Specifically, the reduced-order distribution should have maximal entropy, which is equivalent to requiring that the partition function induces the minimum information loss; a suboptimal solution to this problem is given by the information bottleneck method, cf. [15].

We furthermore provide new insight in some aspects of the lifting proposed in [3]:

1) The KLDR between the original Markov chain and its aggregation, as defined in [3], is an upper bound on the KLDR between the projection and the aggregation.
2) Following [16], we introduce a compact matrix notation for the lifting introduced in [3], allowing us to provide new proofs for some of the results shown there.

In works related to (graph) clustering, information-theoretic cost functions are often used for error quantification. In particular, in [17], the authors use the information bottleneck method for partitioning a graph via assuming continuous-time graph diffusion. Moreover, in [18] and [19] pairwise distance measures between data points were used to define a stationary Markov chain, whose statistics are then used for clustering the data points. While [18] applies the information bottleneck method and obtains a result very similar to ours, the authors do not describe its importance for Markov chain aggregation. In [19], the authors employ the same cost function as [3] and present an iterative algorithm similar to the agglomerative information bottleneck method [20]. While their work focuses on pairwise clustering, they conclude by stating that their results can be employed for Markov chain aggregation. Most recently, the authors of [21] proposed graph clustering by defining a dissimilarity function between the original and the aggregated graph and subsequently applying deterministic annealing to find the best clustering. They define a composite graph to cope with the problem of comparing graphs with different sizes and apply their results to Markov chain aggregation using KLDR as a dissimilarity measure.

Although the present work focuses on stationary Markov chains, we conjecture that our results can be generalized to time-homogeneous Markov chains with a starting distribution different from the invariant distribution, since they are still *stationary in the asymptotic mean*, or AMS [22]. A more detailed discussion of AMS can be found in [23].

The extension to *stochastic aggregations*, i.e., an aggregation where the state space reduction is not performed by a deterministic partition function, but rather by a stochastic mapping, is not immediately possible, at least to the best of our knowledge. While a result similar to our Lemma 2 should hold also for stochastic mappings (cf. [24, Ch. 4.4]), it is not clear how the stochastic aggregation should be lifted to a Markov chain

on the original state space. Since deterministic mappings are preferable for their simplicity, we leave the topic of stochastic aggregations for future investigation.

B. Outline of the Paper

We start by introducing notation and information-theoretic quantities (Section II-A and B) and their application to Markov chains (Section II-C) and functions of Markov chains (Section II-D; introducing also the notion of lumpability). Turning to the problem of state space aggregation in Section III, we restate results linked to the lifting method proposed by [3] (Section IV) and present an alternative and its properties (Section V). Section VI connects the proposed lifting method to the notion of relevant information loss recently introduced in [15]; we exploit this connection in Section VII to show how the information bottleneck method can be employed for state space reduction. A few small examples are contained in Section VIII. The final section, Section IX, is devoted to a biologically inspired example.

II. Notation, Preliminaries, and Setup

A. Random Variables and Stochastic Processes

Let $(\Omega, \mathfrak{B}, \Pr)$ denote the probability space on which all random variables (RVs) and stochastic processes are defined. We denote RVs by upper case letters, e.g., Z, their (finite) alphabet by calligraphic letters, e.g., $\mathcal{Z}$, and realizations by lower case letters, e.g., z, where $z \in \mathcal{Z}$. For an index set $\mathbb{I} \subset \mathbb{N}$ with finite cardinality $\mathrm{card}(\mathbb{I})$, let $Z_{\mathbb{I}} := \{Z_i\}_{i\in\mathbb{I}}$; in particular, we abbreviate $Z_m^n := \{Z_m, Z_{m+1}, \ldots, Z_n\}$. The probability mass function (PMF) of Z is denoted by p_Z, where

$$\forall z \in \mathcal{Z}: \quad p_Z(z) := \Pr(Z = z). \tag{1}$$

The joint PMF $p_{Z_{\mathbb{I}}}$ of $Z_{\mathbb{I}}$ and the conditional PMF $p_{Z_{\mathbb{I}}|Z_{\mathbb{J}}}$ of $Z_{\mathbb{I}}$ given $Z_{\mathbb{J}}$ are defined similarly.

In this work, discrete-time, one-sided random process are denoted by bold-faced letters, e.g., $\mathbf{Z}$, and their (random) samples are indexed by the set of natural numbers, i.e., $\{Z_1, Z_2, \ldots\}$. We assume each RV Z_i takes values from the same, finite, alphabet $\mathcal{Z}$. The random processes considered in this work are *stationary*. In particular, the marginal distribution of Z_k is equal for all k and shall be denoted as p_Z.

B. Information-Theoretic Quantities

In the remainder of this work we will need

Definition 1 (Information-Theoretic Quantities [24, Ch. 2 and 4]): The *(joint) entropy* of a collection of RVs $Z_{\mathbb{I}}$, the *conditional entropy* of $Z_{\mathbb{I}}$ given $Z_{\mathbb{J}}$, and the *mutual information* between $Z_{\mathbb{I}}$ and $Z_{\mathbb{J}}$ are

$$H(Z_{\mathbb{I}}) := -\sum_{z_{\mathbb{I}} \in \mathcal{Z}^{\mathrm{card}(\mathbb{I})}} p_{Z_{\mathbb{I}}}(z_{\mathbb{I}}) \log p_{Z_{\mathbb{I}}}(z_{\mathbb{I}}) \tag{2a}$$

$$H(Z_{\mathbb{I}}|Z_{\mathbb{J}}) := H(Z_{\mathbb{I}\cup\mathbb{J}}) - H(Z_{\mathbb{J}}) = H(Z_{\mathbb{I}}, Z_{\mathbb{J}}) - H(Z_{\mathbb{J}}) \tag{2b}$$

$$I(Z_{\mathbb{I}}; Z_{\mathbb{J}}) := H(Z_{\mathbb{J}}) + H(Z_{\mathbb{I}}) - H(Z_{\mathbb{I}}, Z_{\mathbb{J}}). \tag{2c}$$

The *entropy rate* and the *redundancy rate* of a stationary stochastic process $\mathbf{Z}$ are

$$\bar{H}(\mathbf{Z}) := \lim_{n\to\infty} \frac{1}{n} H\left(Z_1^n\right) = \lim_{n\to\infty} H\left(Z_n | Z_1^{n-1}\right) \tag{2d}$$

$$\bar{R}(\mathbf{Z}) := H(Z) - \bar{H}(\mathbf{Z}) \overset{(a)}{\geq} 0 \tag{2e}$$

where $H(Z)$ is the entropy of the marginal distribution of $\mathbf{Z}$ and where (a) is due to the fact that conditioning reduces entropy [24, Theorem 2.6.5].

The redundancy rate is a measure of statistical dependence between the current sample and its past: For a process of independent, identically distributed RVs, $\bar{H}(\mathbf{Z}) = H(Z)$ and $\bar{R}(\mathbf{Z}) = 0$. Conversely, for a completely predictable process, $\bar{H}(\mathbf{Z}) = 0$ and $\bar{R}(\mathbf{Z}) = H(Z)$. In other words, the higher the redundancy rate, the lower the entropy rate and, thus, the less information is conveyed by the process in each time step.

We need another definition for the development of our results:

Definition 2 (Kullback–Leibler Divergence Rate): The Kullback–Leibler divergence rate (KLDR) between two stationary stochastic processes $\mathbf{Z}$ and $\mathbf{Z}'$ on the same finite alphabet $\mathcal{Z}$ is [25, Ch. 10]

$$\begin{aligned}\bar{D}(\mathbf{Z}\|\mathbf{Z}') &:= \lim_{n\to\infty} \frac{1}{n} D(Z_1^n \| Z_1'^n) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{z_1^n \in \mathcal{Z}^n} p_{Z_1^n}(z_1^n) \log \frac{p_{Z_1^n}(z_1^n)}{p_{Z_1'^n}(z_1^n)}\end{aligned} \tag{3}$$

whenever the limit exists and if $p_{Z_1^n} \ll p_{Z_1'^n}$ for all n, i.e., if for all n and all z_1^n

$$p_{Z_1'^n}(z_1^n) = 0 \Rightarrow p_{Z_1^n}(z_1^n) = 0. \tag{4}$$

The limit exists, e.g., between a stationary stochastic process and a time-homogeneous Markov chain [25] as well as between Markov chains (not necessarily stationary or irreducible) [5]. Roughly speaking, the KLDR between a process $\mathbf{Z}$ and its model $\mathbf{Z}'$ quantifies the number of bits necessary per time step to correct the model distribution to arrive at the true process distribution.

C. Markov Chains

Let $\mathbf{X}$ be a regular, i.e., irreducible and aperiodic, time-homogeneous Markov chain on the finite alphabet $\mathcal{X} = \{1, \ldots, N\}$ (see [16] for terminology). Its behavior is uniquely determined by its transition matrix $\mathbf{P} = \{P_{ij}\}$, where $P_{ij} := \Pr(X_n = j | X_{n-1} = i)$. The unique invariant distribution vector $\boldsymbol{\mu}$ with its i-th component given by

$$\mu_i := p_X(i) = \Pr(X_k = i) > 0 \tag{5}$$

satisfies $\boldsymbol{\mu}^T = \boldsymbol{\mu}^T \mathbf{P}$ [16, Th. 4.1.6]. For such a Markov chain we use the shorthand notation $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$. We assume furthermore that $\mathbf{X}$ is stationary, i.e., its initial distribution coincides with the invariant distribution.

With Definition 1, the entropy and the entropy rate of $\mathbf{X}$, as well as the KLDR between two Markov chains $\mathbf{X}$ and $\mathbf{X}'$

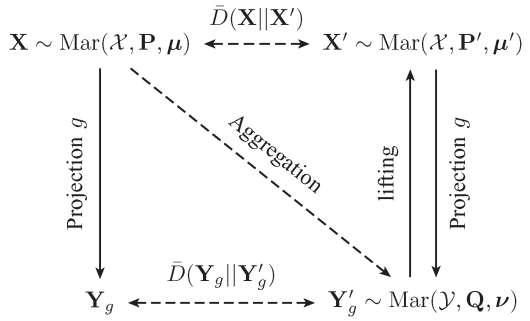

Fig. 1. Illustration of the problem: Assume a Markov chain $\mathbf{X}$ is given. We are interested in finding an aggregation of $\mathbf{X}$, i.e., a Markov chain $\mathbf{Y}'_g$ on a partition of the alphabet of $\mathbf{X}$. This partition defines a function g (and vice-versa), which allows us to define a process $\mathbf{Y}_g$ (via $Y_{g,n} := g(X_n)$), the projection of $\mathbf{X}$. Note that $\mathbf{Y}_g$ need not be Markov. Lifting $\mathbf{Y}'_g$ yields a Markov chain on the original alphabet, which can be projected to $\mathbf{Y}'_g$ using the function g.

on the same alphabet $\mathcal{X}$ with transition matrices $\mathbf{P}$ and $\mathbf{P}'$ are [24, p. 77], [5]

$$H(X) = -\sum_{i\in\mathcal{X}} \mu_i \log \mu_i \tag{6}$$

$$\bar{H}(X) = H(X_1|X_0) = -\sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log P_{ij} \tag{7}$$

$$\bar{D}(\mathbf{X}\|\mathbf{X}') = \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{P_{ij}}{P'_{ij}} \tag{8}$$

respectively, provided that $P'_{ij} = 0$ implies $P_{ij} = 0$ ($\mathbf{P} \ll \mathbf{P}'$).

D. Functions of Markov Chains

We partition the alphabet $\mathcal{X}$ of the Markov chain $\mathbf{X}$ by a surjective function $g : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, M\}$. In other words, g induces a partition of $\mathcal{X}$ by the preimages[1] of the elements of $\mathcal{Y}$. Projecting $\mathbf{X}$ through the function, i.e., $Y_n := g(X_n)$, defines another stochastic process $\mathbf{Y}$; in what follows we call this process the *projected process*, or simply the *projection* of $\mathbf{X}$ (see Fig. 1).

If $\mathbf{X}$ is stationary, then so is $\mathbf{Y}$. The following inequalities

$$H(Y) \le H(X) \tag{9}$$

$$\bar{H}(\mathbf{Y}) \le \bar{H}(\mathbf{X}) \tag{10}$$

$$\bar{R}(\mathbf{Y}) \le \bar{R}(\mathbf{X}) \tag{11}$$

hold by the data processing inequality [24], [26] and by [27].

It is well known that $\mathbf{Y}$ is not necessarily Markov. The case where $\mathbf{Y}$ is a regular, time-homogeneous Markov chain, gives rise to the notion of lumpability:

Definition 3 (Lumpability [16]): A Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$ is *lumpable* w.r.t. a function g, iff the process $\mathbf{Y}$ is a regular, time-homogeneous Markov chain with alphabet $\mathcal{Y}$, transition matrix $\mathbf{Q}$, and invariant distribution $\boldsymbol{\nu}$, i.e., iff $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \boldsymbol{\nu})$, for every initial distribution of $\mathbf{X}$.

[1]Given a state $j \in \mathcal{Y}$, with slight abuse of notation we write $g^{-1}(j)$ for the preimage of j under g, that is, $g^{-1}(j) := g^{-1}(\{j\}) = \{i \in \mathcal{X} | g(i) = j\}$.

In order to present conditions under which a Markov chain is lumpable, we need the following matrices: Let $\mathbf{V}$ be an $N \times M$ matrix with $V_{ij} := 1$ if $i \in g^{-1}(j)$ and zero otherwise (thus, every row contains exactly one 1). Furthermore, $\mathbf{U}^{\boldsymbol{\pi}}$ is an $M \times N$ matrix with zeros in the same positions as $\mathbf{V}^T$, but with otherwise positive row entries which sum to one. In other words, with $\boldsymbol{\pi}$ being a positive probability vector

$$U^{\boldsymbol{\pi}}_{ij} := \begin{cases} \frac{\pi_j}{\sum_{k\in g^{-1}(i)} \pi_k}, & \text{if } j \in g^{-1}(i) \\ 0, & \text{else.} \end{cases} \tag{12}$$

Lemma 1 (Conditions for Lumpability): A stationary Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$ is lumpable w.r.t. g iff for every positive probability vector $\boldsymbol{\zeta}$

$$\mathbf{V}\mathbf{U}^{\boldsymbol{\zeta}}\mathbf{P}\mathbf{V} = \mathbf{P}\mathbf{V}. \tag{13}$$

Then, $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \boldsymbol{\nu})$ with $\boldsymbol{\nu}^T = \boldsymbol{\mu}^T\mathbf{V}$ and

$$\mathbf{Q} = \mathbf{U}^{\boldsymbol{\mu}}\mathbf{P}\mathbf{V}. \tag{14}$$

Proof: See [16, Th. 6.3.5 and Ex. 6.3.3] and note that

$$\boldsymbol{\nu}^T = \boldsymbol{\nu}^T\mathbf{Q} = \boldsymbol{\nu}^T\mathbf{U}^{\boldsymbol{\mu}}\mathbf{P}\mathbf{V} = \boldsymbol{\mu}^T\mathbf{P}\mathbf{V} = \boldsymbol{\mu}^T\mathbf{V}. \tag{15}$$

■

The corresponding result for continuous-time Markov chains on a countable alphabet has been proven in [28, Th. 2].

While the KLDR between two Markov chains is easy to compute, for the KLDR between a (non-Markov) function of a Markov chain and a Markov chain no closed-form solution is available. In special cases, however, the former can act as an upper bound on the latter, provided the Markov chains are chosen appropriately. We make this precise in

Lemma 2: Let $\mathbf{X}$ and $\mathbf{X}'$ be stationary, time-homogeneous, regular Markov chains on the same alphabet $\mathcal{X}$ with transition matrices $\mathbf{P}$ and $\mathbf{P}'$. Let $\mathbf{P}' \gg \mathbf{P}$. We define two processes $\mathbf{Y}$ and $\mathbf{Y}'$ by $Y_n := g(X_n)$ and $Y'_n := g(X'_n)$, $g : \mathcal{X} \to \mathcal{Y}$. Let additionally $\mathbf{X}'$ be lumpable w.r.t. g. We have

$$\bar{D}(\mathbf{X}\|\mathbf{X}') \ge \bar{D}(\mathbf{Y}\|\mathbf{Y}'). \tag{16}$$

Proof: The inequality follows from the fact that the Kullback–Leibler divergence reduces under measurements (e.g., [25, Cor. 3.3] or [26, Ch. 2.4]), i.e., that for all n

$$D(X_1^n\|X_1'^n) \ge D(Y_1^n\|Y_1'^n)\,. \tag{17}$$

It thus remains to show that the limits exist.

Since $\mathbf{P}' \gg \mathbf{P}$, $\bar{D}(\mathbf{X}\|\mathbf{X}')$ exists and equals (8), cf. [5]. Since $\mathbf{X}'$ is lumpable, $\mathbf{Y}'$ is a regular, time-homogeneous Markov chain. Moreover, from $\mathbf{P}' \gg \mathbf{P}$ it follows that the process distribution of $\mathbf{Y}$ is absolutely continuous w.r.t. the process distribution of $\mathbf{Y}'$. This ensures the existence of $\bar{D}(\mathbf{Y}\|\mathbf{Y}')$ [25, Lemma 10.1] and completes the proof. ■

III. Problem Statement

Throughout the remainder of this work, we will stick to the following

Assumption 1: The discrete-time Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$ is stationary, i.e., the initial distribution equals its invariant distribution $\boldsymbol{\mu}$. The alphabet of $\mathbf{X}$ is $\mathcal{X} = \{1, \ldots, N\}$ and the partition function is $g : \mathcal{X} \to \mathcal{Y} = \{1, \ldots, M\}$ with $1 < M < N$. The *g-projection* of $\mathbf{X}$ is the stationary process $\mathbf{Y}_g$ over the alphabet $\mathcal{Y}$, whose samples are defined by

$$Y_{g,n} := g(X_n). \tag{18}$$

We are interested in performing model reduction by employing information-theoretic cost functions. In particular, we specify the M-partition problem, equivalently defined in [3]:

Definition 4 (M-Partition Problem): Given $\mathbf{X}$ and g as in Assumption 1, the *M-partition problem* searches for the partition function g such that the KLDR between the g-projection of $\mathbf{X}$ and its best Markov approximation is minimal, i.e., it solves

$$\arg\min_{g \in [\mathcal{X} \to \mathcal{Y}]} \min_{\mathbf{Y}'} \left\{ \bar{D}(\mathbf{Y}_g \| \mathbf{Y}') | \mathbf{Y}' \text{ is Markov} \right\}. \tag{19}$$

For a fixed partition function g, the best Markov approximation (in the sense of the KLDR) of the g-projection $\mathbf{Y}_g$ can be found analytically. With the matrix notation introduced in Section II-D we present

Lemma 3: Given $\mathbf{X}$, g and $\mathbf{Y}_g$ as in Assumption 1, let $\mathbf{Y}'_g$ denote the best Markov approximation of the g-projection in the sense of the KLDR, i.e.,

$$\mathbf{Y}'_g := \arg\min_{\mathbf{Y}'} \left\{ \bar{D}(\mathbf{Y}_g \| \mathbf{Y}') | \mathbf{Y}' \text{ is Markov} \right\}. \tag{20}$$

Then, $\mathbf{Y}'_g \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \boldsymbol{\nu})$ with $\boldsymbol{\nu}^T = \boldsymbol{\mu}^T \mathbf{V}$ and

$$\mathbf{Q} = \mathbf{U}^{\boldsymbol{\mu}} \mathbf{P} \mathbf{V} \tag{21}$$

which is a matrix notation for

$$Q_{kl} = \frac{\sum_{i \in g^{-1}(k)} \sum_{j \in g^{-1}(l)} \mu_i P_{ij}}{\sum_{i \in g^{-1}(k)} \mu_i}, \qquad k, l \in \mathcal{Y}. \tag{22}$$

Proof: See [25, Cor. 10.4]. ∎

From now on, we keep the notation $\mathbf{Y}'_g$ for the optimal aggregation (see Fig. 1) of $\mathbf{X}$, given a partition function g.

Remark 1: The same aggregation was declared being optimal in [29], although by using a different cost function. Also [3, Th. 1] declares this aggregation as being optimal, although the cost function there is the KLDR between the original chain $\mathbf{X}$ and the *lifting* of $\mathbf{Y}'_g$ (see Section IV below).

One thus obtains the transition matrix $\mathbf{Q}$ of the optimal Markov model $\mathbf{Y}'_g$ from the joint distribution of two consecutive samples of $\mathbf{Y}_g$. If this joint distribution completely specifies the process $\mathbf{Y}_g$, then $\mathbf{Y}_g \equiv \mathbf{Y}'_g$, i.e., $\mathbf{X}$ is lumpable (cf. Lemma 1). Note further that since $\mathbf{P}$ is the transition matrix of a regular Markov chain, so is $\mathbf{Q}$ [16, p. 140].

We can now define the aggregation error of $\mathbf{X}$ w.r.t. g:

Definition 5 (Aggregation Error): Given $\mathbf{X}$, g and $\mathbf{Y}_g$ as in Assumption 1 and $\mathbf{Y}'_g$ as in Lemma 3. Then

$$\bar{D}\big(\mathbf{Y}_g \| \mathbf{Y}'_g\big) \tag{23}$$

is the *aggregation error* of $\mathbf{X}$ w.r.t. g.

It immediately follows that the aggregation error is zero if $\mathbf{X}$ is lumpable.

Following [3], we split the M-partition problem into two subproblems: finding the best Markov approximation of the projected process $\mathbf{Y}_g$ (to which Lemma 3 provides the solution), and minimizing the aggregation error over all partition functions g with a range of cardinality M. Thus, the optimization problem stated in (19) translates to finding

$$\arg\min_{g \in [\mathcal{X} \to \mathcal{Y}]} \bar{D}\big(\mathbf{Y}_g \| \mathbf{Y}'_g\big). \tag{24}$$

IV. $\boldsymbol{\pi}$-Lifting: Bounding the Aggregation Error

Often, a direct evaluation of the aggregation error in Definition 5 is mathematically cumbersome, since $\mathbf{Y}_g$ is not necessarily Markov. The authors of [3] therefore suggested to *lift* the aggregation $\mathbf{Y}'_g$ to a Markov chain $\mathbf{X}'$ over the alphabet $\mathcal{X}$, which subsequently allows a computation of the KLDR.

The questions is now, whether there is a relation between the KLDR between $\mathbf{X}$ and the lifted chain $\mathbf{X}'$, and the aggregation error, $\bar{D}(\mathbf{Y}_g \| \mathbf{Y}'_g)$. Relying on Lemma 2, we will answer this question affirmatively.

Definition 6 ($\boldsymbol{\pi}$-Lifting [3, Def. 2]): Given $\mathbf{X}$, g and $\mathbf{Y}_g$ as in Assumption 1, $\mathbf{Y}'_g \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \boldsymbol{\nu})$ as in Lemma 3, and $\boldsymbol{\pi}$ a positive probability distribution over the alphabet $\mathcal{X}$. The *$\boldsymbol{\pi}$-lifting* of $\mathbf{Y}'_g$ w.r.t. g, denoted by $\mathbf{X}'^{\boldsymbol{\pi}}_g$, is a Markov chain over the alphabet $\mathcal{X}$ with transition matrix

$$\mathbf{P}' := \mathbf{V} \mathbf{Q} \mathbf{U}^{\boldsymbol{\pi}} \tag{25}$$

which is a matrix notation for

$$P'_{ij} = \frac{\pi_j}{\sum_{k \in g^{-1}(g(j))} \pi_k} Q_{g(i)g(j)}, \quad i, j \in \mathcal{X}. \tag{26}$$

Remark 2: An equivalent lifting method is suggested in [29] and [9], [30].

We conclude this section by presenting the elementary properties of $\boldsymbol{\pi}$-lifting. Properties 1), 2), and 3) appear also in [3]; the proofs can be found in Appendix A and, in contrast to the proofs in [3], appear in short matrix notation. To the best of the authors' knowledge, properties 4) and 5) are proved for the first time here.

Proposition 1 (Properties of $\boldsymbol{\pi}$-Lifting): Given $\mathbf{X}$, g and $\mathbf{Y}_g$ as in Assumption 1, $\mathbf{Y}'_g$ as in Lemma 3, and $\boldsymbol{\pi}$ some distribution over $\mathcal{X}$. Then, the $\boldsymbol{\pi}$-lifting $\mathbf{X}'^{\boldsymbol{\pi}}_g$ satisfies:

1) $\mathbf{X}'^{\boldsymbol{\pi}}_g$ is lumpable w.r.t. g (and $\mathbf{Y}'_g$ is the resulting g-projection);
2) The invariant distribution of $\mathbf{X}'^{\boldsymbol{\mu}}_g$ is $\boldsymbol{\mu}$;
3) $\boldsymbol{\mu} = \arg\min_{\boldsymbol{\pi}} \bar{D}(\mathbf{X} \| \mathbf{X}'^{\boldsymbol{\pi}}_g)$;
4) $\mathbf{P}' \gg \mathbf{P}$;
5) $\bar{D}(\mathbf{Y}_g \| \mathbf{Y}'_g) \leq \bar{D}(\mathbf{X} \| \mathbf{X}'^{\boldsymbol{\mu}}_g)$.

V. A Better Bound via $\mathbf{P}$-Lifting

In Section IV, we showed that the KLDR between $\mathbf{X}$ and the $\boldsymbol{\pi}$-lifting with $\boldsymbol{\pi} = \boldsymbol{\mu}$, $\mathbf{X}'^{\boldsymbol{\mu}}_g$, provides an upper bound on the aggregation error for a given partition function g. Unfortunately,

the bound is loose in the sense that for $\bar{D}(\mathbf{Y}_g\|\mathbf{Y}'_g)=0$, we may have $\bar{D}(\mathbf{X}\|\mathbf{X}'^{\boldsymbol{\mu}}_g)>0$; see also [9]. One of the reasons for this disadvantage of $\boldsymbol{\pi}$-lifting is that, by construction, the lifted process $\mathbf{X}'^{\boldsymbol{\pi}}_g$ does not contain information about the transition probabilities between states of $\mathbf{X}$. We therefore propose a lifting which takes into account the transition matrix $\mathbf{P}$ of the original process.

Definition 7 ($\mathbf{P}$-Lifting): Given $\mathbf{X}$, g and $\mathbf{Y}_g$ as in Assumption 1 and $\mathbf{Y}'_g\sim\mathrm{Mar}(\mathcal{Y},\mathbf{Q},\boldsymbol{\nu})$ as in Lemma 3. The $\mathbf{P}$-*lifting* of $\mathbf{Y}'_g$ w.r.t. g, denoted by $\mathbf{X}'^{\mathbf{P}}_g$, is a Markov chain over the alphabet $\mathcal{X}$ with a transition matrix $\hat{\mathbf{P}}$ given by

$$\hat{P}_{ij}:=\begin{cases}\frac{P_{ij}}{\sum_{k\in g^{-1}(g(j))}P_{ik}}Q_{g(i)g(j)}, & \text{if } \sum_{k\in g^{-1}(g(j))}P_{ik}>0\\ \frac{1}{\mathrm{card}(g^{-1}(g(j)))}Q_{g(i)g(j)}, & \text{if } \sum_{k\in g^{-1}(g(j))}P_{ik}=0.\end{cases}\tag{27}$$

One of the main contributions of this paper is to show that the KLDR between $\mathbf{X}$ and the $\mathbf{P}$-lifting $\mathbf{X}'^{\mathbf{P}}_g$ yields a better bound than the one obtained using $\boldsymbol{\pi}$-lifting. In Appendix B we prove

Theorem 1 (Properties of $\mathbf{P}$-Lifting): Given $\mathbf{X}$, g and $\mathbf{Y}_g$ as in Assumption 1 and $\mathbf{Y}'_g$ as in Lemma 3. Then, the $\mathbf{P}$-lifting $\mathbf{X}'^{\mathbf{P}}_g$ satisfies

1) $\mathbf{X}'^{\mathbf{P}}_g$ is lumpable w.r.t. g (and $\mathbf{Y}'_g$ is the resulting g-projection);
2) $\hat{\mathbf{P}}\gg\mathbf{P}$;
3) (minimizer)

$$\mathbf{X}'^{\mathbf{P}}_g=\underset{\hat{\mathbf{X}}:\mathbf{Y}'_g\text{ is }g-\text{projection of }\hat{\mathbf{X}}}{\arg\min}\bar{D}(\mathbf{X}\|\hat{\mathbf{X}})\tag{28}$$

4) (better bounds than π-lifting)

$$\bar{D}\big(\mathbf{Y}_g\|\mathbf{Y}'_g\big)\le\bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\mathbf{P}}_g\big)\le\bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\boldsymbol{\mu}}_g\big)\tag{29}$$

5) (tight bounds) If $\mathbf{X}$ is lumpable w.r.t. g

$$\bar{D}\big(\mathbf{Y}_g\|\mathbf{Y}'_g\big)=0\Leftrightarrow\bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\mathbf{P}}_g\big)=0.\tag{30}$$

Tightness follows from the fact that for a lumpable $\mathbf{X}$, the $\mathbf{P}$-lifting yields $\hat{\mathbf{P}}=\mathbf{P}$; the invariant distribution of $\hat{\mathbf{P}}$ trivially coincides with $\boldsymbol{\mu}$, the invariant distribution of $\mathbf{P}$. In general, however, the invariant distribution of $\hat{\mathbf{P}}$ differs from $\boldsymbol{\mu}$, contrasting the corresponding result for the $\boldsymbol{\pi}$-lifing (cf. Proposition 1, property 2).

Interestingly, the restriction to lumpable chains for the tightness result cannot be dropped: There are Markov chains $\mathbf{X}$ which are lumpable in a weaker sense (i.e., not for all initial distributions but, e.g., only for the invariant distribution) for which consequently the aggregation error vanishes, but for which the $\mathbf{P}$-lifting does not yield $\hat{\mathbf{P}}=\mathbf{P}$. A simple example of such a chain is given in [16, p. 139] (cf. Section VIII-D).

As this theorem shows, $\mathbf{P}$-lifting yields the best upper bound on the aggregation error achievable for Markov chains over the alphabet $\mathcal{X}$. This can also be explained intuitively, by expanding the KLDR as

$$\bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\mathbf{P}}_g\big)=\sum_{i,j\in\mathcal{X}}\mu_iP_{ij}\log\frac{P_{ij}}{\hat{P}_{ij}}\tag{31}$$

$$=\sum_{i,j\in\mathcal{X}}\mu_iP_{ij}\log\frac{\sum_{k\in\mathcal{S}_j}P_{ik}}{Q_{g(i)g(j)}}\tag{32}$$

$$\overset{(a)}{=}\sum_{i,j\in\mathcal{X}}\mu_iP_{ij}\log\frac{\left(\sum_{k\in\mathcal{S}_i}\mu_k\right)\left(\sum_{l\in\mathcal{S}_j}P_{il}\right)}{\sum_{k\in\mathcal{S}_i}\mu_k\sum_{l\in\mathcal{S}_j}P_{kl}}\tag{33}$$

$$=H(Y_{g,n}|Y_{g,n-1})-H(Y_{g,n}|X_{n-1})\tag{34}$$

where (a) is due to Lemma 3 and where $\mathcal{S}_i=g^{-1}(g(i))$. The last line corresponds to the difference between the upper and lower bounds on the entropy rate of a function of a Markov chain [24, Th. 4.5.1]; equality of these bounds implies Markovity of $\mathbf{Y}_g$, i.e., lumpability of $\mathbf{X}$ w.r.t. g [31, Th. 9]. In other words, minimizing this cost function yields the function g for which the projected process $\mathbf{Y}_g$ is "as Markov as possible."

VI. $\mathbf{P}$-Lifting and Information Loss

We now analyze how the KLDR between the original process and a $\mathbf{P}$-lifted process connects with the information loss induced by the projection function g. This parallels the analysis in [3], claiming that $\boldsymbol{\pi}$-lifting maximizes the redundancy rate[2] of the aggregated process. Interestingly, the cost function induced by $\mathbf{P}$-lifting *minimizes* a special notion of information loss introduced recently. Moreover, the latter analysis paves the way for solving the state space reduction problem using information-theoretic algorithms, such as the information bottleneck method (cf. Section VII).

Definition 8 (Relevant Information Loss [15]): Let X be an RV with finite alphabet $\mathcal{X}$, and let $Y:=g(X)$. Let S be another RV with alphabet $\mathcal{S}$ representing *relevant information*. The information loss *relevant w.r.t.* S is

$$L_S(X\to Y)=I(S;X)-I(S;Y)=I(X;S|Y).\tag{35}$$

A simple example to illustrate this notion of information loss is the following: Let S be a binary signal, and let X be this signal superimposed by noise, e.g., the output of a noisy communications channel. By passing X through a function g, e.g., a quantizer in a digital receiver, some information about S is lost. $L_S(X\to Y)$ does not quantify the information (about S) lost over the channel, but the *additional* information lost by quantizing the channel's output.

We now make a connection between Definition 8 and the KLDR between $\mathbf{X}$ and the $\mathbf{P}$-lifted chain, $\mathbf{X}'^{\mathbf{P}}_g$. To this end, let

$$g^\bullet:=\arg\min_g\bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\mathbf{P}}_g\big).\tag{36}$$

Recall from (34) that

$$\bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\mathbf{P}}_g\big)=H(Y_{g,n}|Y_{g,n-1})-H(Y_{g,n}|X_{n-1})\tag{37}$$

[2]The reader should not be misled by the fact that the redundancy rate satisfies $\bar{R}(\mathbf{Y}'_g)=I(Y_{g,0};Y_{g,1})$, i.e., that it is formulated as a mutual information. The fact that the current sample $Y_{g,0}$ shares much information with the future sample $Y_{g,1}$ only emphasizes that the process is redundant, i.e., that it conveys *little new information in each time step*.

which, by adding and subtracting $H(Y_{g,n})$ can be rewritten as

$$\begin{aligned}\bar{D}\big(\mathbf{X}\|\mathbf{X}_g'^{\mathbf{P}}\big) &= I(Y_{g,n};X_{n-1}) - I(Y_{g,n};Y_{g,n-1}) \\ &= L_{Y_{g,n}}(X_{n-1}\to Y_{g,n-1}) \end{aligned} \tag{38}$$

where $L_{Y_{g,n}}(X_{n-1}\to Y_{g,n-1})$ is the *information loss relevant w.r.t.* $Y_{g,n}$ induced by projecting X_{n-1} through the function g. Finding the optimal function $g^\bullet$ thus amounts to *minimizing* information loss.

To the present date, we could not verify if this cost function has an interpretation in spectral theory. However, as mentioned above, it minimizes the difference between first-order upper and lower bounds on the entropy rate of the projected process $\mathbf{Y}_g$ and, with [31, Th. 9], makes the projected process "as Markov as possible."

VII. The Information Bottleneck Method: A Possible Way to Model Reduction

In this section, we show that the state space reduction problem can be solved by a well-known information-theoretic algorithm: the information bottleneck method [6]. Since in [15] the information bottleneck (IB) method was reformulated in terms of relevant information loss, the results of Section VI are essential for the development of the following paragraphs.

Let X be a discrete RV representing an observation (e.g., the output of a noisy communications channel) or a data set. We are interested in a compressed representation Y of this RV. In rate-distortion theory (e.g., [25]) one pursues the goal to minimize the mutual information between X and its compression Y, $I(X;Y)$, subject to satisfying a certain distortion criterion d (e.g., the mean-squared reconstruction error). This can be cast as a variational problem

$$\arg\min_{p_{Y|X}} I(X;Y) + \beta d(X,Y) \tag{39}$$

where β is a Lagrange multiplier, and where stochastic compressions $p_{Y|X}(x,y) = \Pr(Y=y|X=x)$ are permitted.

The IB method takes up this approach by replacing the distortion measure by the negative mutual information between the compressed RV Y and a *relevant* RV S, representing the information one considers as meaningful and one wants to preserve (e.g., the binary input to the communications channel). The IB method therefore tries to solve

$$\arg\min_{p_{Y|X}} I(X;Y) - \beta I(S;Y) \tag{40}$$

where the minimization runs over all stochastic relationships and where β trades compression and preservation of information: A large value of β places emphasis on preservation of relevant information, while a small value leads to high compression. Typical applications of the IB method include word and document clustering [20], [32] or speech processing [33], [34].

With $\beta\to\infty$, we focus on the second term of (40) which can be rewritten with Definition 8 as

$$I(S;Y) = I(S;X) - L_S(X\to Y). \tag{41}$$

With the restriction to deterministic compressions $p_{Y|X}$ determined by functions $g\colon \mathcal{X}\to\mathcal{Y}$, one obtains a formulation of the IB method which minimizes the relevant information loss, i.e., which solves

$$\arg\min_{g\in[\mathcal{X}\to\mathcal{Y}]} L_S(X\to Y). \tag{42}$$

For this problem, in [20] an iterative procedure, called *agglomerative IB* (AIB) was introduced, which successively merges two elements of a partition of $\mathcal{X}$ until the desired cardinality M is reached. The method is greedy, i.e., it minimizes the information lost in each step [20], but does not guarantee that the global optimum (42), i.e., the least possible relevant information loss, is achieved.

Comparing (42) with (38), one can see that the relevant information $Y_{g,n}$ depends on g, i.e., on the object to be optimized. Since in such a case the IB method is not applicable directly, we relax the problem by applying [15, Cor. 1]:

$$L_{Y_{g,n}}(X_{n-1}\to Y_{g,n-1}) \le L_{X_n}(X_{n-1}\to Y_{g,n-1}). \tag{43}$$

Instead of minimizing $\bar{D}(\mathbf{X}\|\mathbf{X}_g'^{\mathbf{P}})$, we only minimize its upper bound given by (43). We thus look for

$$g^{IB} := \arg\min_g L_{X_n}(X_{n-1}\to Y_{g,n-1}). \tag{44}$$

The possibility to apply the IB method and its algorithms (e.g., AIB) for state space reduction comes at the cost of optimality. As the next subsection shows, this cost is not as high as one would expect.

A. Sub-Optimality of IB

By relaxing the M-partition problem to (44), one loses the property that the cost function minimizes the best upper bound on the aggregation error $\bar{D}(\mathbf{Y}_g\|\mathbf{Y}_g')$. However, the obtained upper bound is still better than $\bar{D}(\mathbf{X}\|\mathbf{X}_g'^{\boldsymbol{\mu}})$:

$$\begin{aligned} &L_{X_n}(X_{n-1}\to Y_{g,n-1}) \\ &\quad= H(X_n|Y_{g,n-1}) - H(X_n|X_{n-1}) && (45)\\ &\quad= H(X_n,Y_{g,n}|Y_{g,n-1}) - \bar{H}(\mathbf{X}) && (46)\\ &\quad= H(X_n|Y_{g,n},Y_{g,n-1}) + \underbrace{H(Y_{g,n}|Y_{g,n-1})}_{=\bar{H}(\mathbf{Y}_g')} - \bar{H}(\mathbf{X}) && (47)\\ &\quad\le H(X_n|Y_{g,n}) + \bar{H}\left(\mathbf{Y}_g'\right) - \bar{H}(\mathbf{X}) && (48)\\ &\quad= H(X) - H\left(Y_g'\right) + H\left(\mathbf{Y}_g'\right) - \bar{H}(\mathbf{X}) && (49)\\ &\quad= \bar{R}(\mathbf{X}) - \bar{R}\left(\mathbf{Y}_g'\right) && (50)\\ &\quad= \bar{D}\big(\mathbf{X}\|\mathbf{X}_g'^{\boldsymbol{\mu}}\big) && (51)\end{aligned}$$

where the last line is due to [3, Lemma 3].

The solution of the relaxed problem (44) might not coincide with the solution of (38). To be specific: Even if a Markov chain $\mathbf{X}$ is lumpable, neither the AIB nor the IB method implementing the relaxed optimization problem necessarily find

TABLE I
RESULTS OF EXAMPLE 1

Partition	KLDR ($\boldsymbol{\mu}$) bit/sample	KLDR ($\mathbf{P}$) bit/sample	$\hat{\boldsymbol{\mu}}$ $[0.347, 0.388, 0.265]^T$
$\{\{1,2\},\{3\}\}$	0.823	0.185	$[0.077, 0.658, 0.265]^T$
$\{\{1,3\},\{2\}\}$	0.808	0.317	$[0.065, 0.388, 0.546]^T$
$\{\{1\},\{2,3\}\}$	0.037	0.001	$[0.347, 0.388, 0.265]^T$

the optimal M-partition. We will elaborate on this topic in the example in Section VIII-C.

VIII. EXAMPLES

In this section, we illustrate our theoretical results at the hand of a few examples. In particular, we show the applicability of the information bottleneck method for Markov chain aggregation in Section VIII-B.

A. Example 1

We take the matrix given in [3, Sec. V.A]

$$\mathbf{P} = \begin{bmatrix} 0.97 & 0.01 & 0.02 \\ 0.02 & 0.48 & 0.50 \\ 0.01 & 0.75 & 0.24 \end{bmatrix} \tag{52}$$

and use three different functions g inducing the following partitions of $\mathcal{X}$: {{1,2},{3}}, {{1,3},{2}}, and {{1},{2,3}}.

For all the resulting aggregations, we compute upper bounds on the aggregation error using both the $\boldsymbol{\pi}$-lifting with $\boldsymbol{\pi} = \boldsymbol{\mu}$ and the $\mathbf{P}$-lifting. In addition to that, the invariant distributions of the $\mathbf{P}$-lifted Markov chains $\mathbf{X}_g'^{\mathbf{P}}$ are computed and compared to $\boldsymbol{\mu}$, the invariant distribution of the original chain $\mathbf{X}$. The results are shown in Table I.

As it can be seen, the partition {{1},{2,3}} yields the best results in terms of KLDR. Moreover, it can be seen that the KLDR using $\mathbf{P}$-lifting is smaller than the KLDR using $\boldsymbol{\pi}$-lifting in all three cases, as suggested by Theorem 1. However, unlike for $\boldsymbol{\pi}$-lifting with $\boldsymbol{\pi} = \boldsymbol{\mu}$, the invariant distribution obtained with our method depends on g and in general differs from $\boldsymbol{\mu}$. An exception is the optimal partition, where $\mathbf{Y}_g$ and $\mathbf{Y}'_g$ are very close in terms of the KLDR, i.e., where $\mathbf{X}$ is "nearly" lumpable w.r.t. g.

B. Example 2

In this example we took the transition matrix $\mathbf{P}$ from [3, Fig. 7] and applied the agglomerative information bottleneck method [20] to aggregate the chain,[3] as described in Section VII. As it can be seen in Fig. 2, the partitions of the alphabet appear to be reasonable and, for $M = 5$, coincide with the solution obtained in [3]. In essence, the aggregation reduces the alphabet to groups of strongly interacting states.

An interesting fact can be observed by looking at Fig. 3, which compares the KLDR curves for both lifting methods (the aggregation was obtained using the agglomerative IB method in both cases). While for $\boldsymbol{\pi}$-lifting the KLDR seems to be a function decreasing with increasing M, the same does not hold for $\mathbf{P}$-lifting: If a certain partition is "nearly" lumpable, the KLDR curve exhibits a local minimum (cf. Theorem 1). Trivially, global minima with value zero are obtained for $M = 1$ and $M = N$; thus, the curve depicted in Fig. 3 will decrease eventually if M is further increased.

These results are relevant for properly choosing the cardinality of the reduced state space: For $\boldsymbol{\pi}$-lifting, it was suggested that a change in slope of the KLDR indicates that a meaningful partition was obtained [3, Sec. V.D]. Utilizing the tighter bound from $\mathbf{P}$-lifting allows to choose the cardinality by detecting local minima.

C. Example 3

In this example we show that the relaxed optimization problem does not necessarily find the optimal partition. We start with a Markov chain $\mathbf{X}$ with state space $\mathcal{X} = \{1, 2, 3\}$ and investigate the bi-partition problem (i.e., $M = 2$). Let the transition matrix be given as

$$\mathbf{P} = \begin{bmatrix} 0.0475 & 0.9025 & 0.05 \\ 0.9025 & 0.0475 & 0.05 \\ 0.95 & 0.05 & 0 \end{bmatrix}. \tag{53}$$

Since this chain is lumpable for the partition {{1,2},{3}} (induced by the optimal function $g^\bullet$), one obtains

$$L_{Y_{g^\bullet,n}}(X_{n-1} \to Y_{g^\bullet,n-1}) = 0. \tag{54}$$

Computer simulations show, however, that this partition leads to a larger value of $H(X_n|Y_{g^\bullet,n-1})$ than the other two options (namely, 1.19 bit compared to 0.55 and 0.69 bit, respectively). Since here the AIB and IB methods coincide,[4] this example shows that the relaxation of the optimization problem does not necessarily lead to the optimal partition.

It is interesting to observe, however, that the information bottleneck method provides the same partition function as the method introduced in [3], namely {{1},{2,3}}. The eigenvalues of the additive reversibilization of $\mathbf{P}$ are $\lambda_1 = 1$, $\lambda_2 = -0.038$, and $\lambda_3 = -0.867$, the latter two inducing the partitions {{1,2},{3}} and {{1},{2,3}}, respectively. Hence, IB and the method in [3] respond with the solution related to the eigenvalue with the second-largest modulus, while the optimal solution remains to be related to the second-largest eigenvalue. This suggests a closer investigation of the interplay between the proposed cost function, its relaxation, and spectral theory, especially when the relevant eigenvalues are negative, cf. [3].

D. Example 4

We finally take an example from [16, p. 139], which shows that our upper bound on the aggregation error is not tight in general, but only for lumpable $\mathbf{X}$. To this end, let

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{6} & \frac{5}{6} \\ \frac{7}{8} & \frac{1}{8} & 0 \end{bmatrix}. \tag{55}$$

[3] We used the VLFeat Matlab implementation [35] of the agglomerative IB method.

[4] The bi-partition is obtained by merging two states.

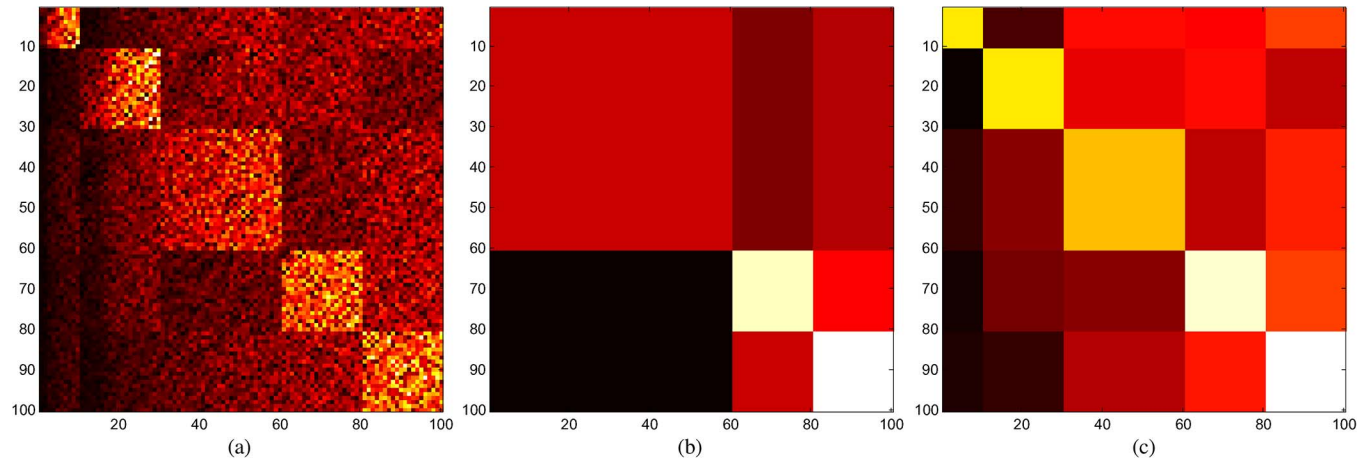

Fig. 2. An illustration of Example 2: The original transition matrix (a) and the partitions obtained by using the agglomerative information bottleneck method. Blocks of the same color indicate that the corresponding states are mapped to the same output. (b) $M = 3$ (c) and $M = 5$.

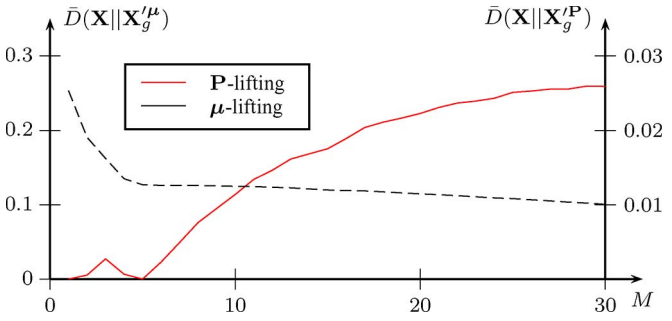

Fig. 3. KLDR for the $\mathbf{P}$- and the $\boldsymbol{\pi}$-lifting with $\boldsymbol{\pi} = \boldsymbol{\mu}$ ($\boldsymbol{\mu}$-lifting in the figure) for different cardinalities M of the aggregated chain's alphabet. Both curves were obtained using the agglomerative IB method. Note that the KLDRs according to the different liftings are displayed with different scales, and note that the graph shows only $M \leq 30 < 100 = N$.

As it can be verified easily, this chain is lumpable in the weak sense w.r.t. the partition $\{\{1\},\{2,3\}\}$, but not lumpable; i.e., $\mathbf{Y}_g$ is a Markov chain if $\mathbf{X}$ is initialized with the invariant distribution, but (13) is not fulfilled. To show that the bound is not tight, we observe that $\bar{D}(\mathbf{Y}_g \| \mathbf{Y}'_g) = 0$ but that, with

$$\hat{\mathbf{P}} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{7}{12} & \frac{5}{72} & \frac{25}{72} \\ \frac{7}{12} & \frac{5}{12} & 0 \end{bmatrix} \neq \mathbf{P} \tag{56}$$

we get $\bar{D}(\mathbf{X}\|\mathbf{X}'^{\mathbf{P}}_g) = 0.347 > 0$.

IX. Application to Models of Bio-Molecular Systems

Recent advances in measurement techniques brought the need for quantitative modeling in biology [1]. Markov models are a major tool used for modeling the stochastic nature of bio-molecular interactions in cells. However, even the simplest networks with only a few interacting species can result in very large Markov chains, in which case their analysis becomes computationally inefficient or prohibitive. In these cases, reducing the state space of the model, with minimal information loss, is an important challenge. We illustrate on an example how our reduction method can be used in such a scenario. The model defined by stochastic chemical kinetics evolves in continuous-time, following a continuous-time Markov chain (CTMC). We will show how our aggregation method can be applied to reduce this CTMC by aggregating a subordinated DTMC. The existing theory confirms that the resulting partition will also be suitable for the original CTMC.

For a well-mixed reaction system with molecular species $S_1, \ldots, S_n$, the state of a system is typically modeled by a multiset of species' abundances: $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X} \subseteq \mathbb{N}_0^n$. The dynamics of such a system are determined by a set of r reactions. The k-th reaction reads

$$\nu_{1k} S_1, \ldots, \nu_{nk} S_n \xrightarrow{c_k} \nu'_{1k} S_1, \ldots, \nu'_{nk} S_n \tag{57}$$

where $\nu_{ik} \in \mathbb{N}_0$ and $\nu'_{ik} \in \mathbb{N}_0$ denote the substrate and product stoichiometric coefficients of species i, respectively, and where c_k is the rate with which the reaction occurs. If the k-th reaction occurs, after being in the state $\mathbf{x}$, the next state will be $\mathbf{x} + (\boldsymbol{\nu}'_k - \boldsymbol{\nu}_k) = \mathbf{x} + \boldsymbol{\mu}_k$, where $\boldsymbol{\mu}_k$ is referred to as the stoichiometric change vector. The species multiplicities follow a continuous-time Markov chain and we denote the state of the system as the t-indexed random vector $\mathbf{X}(t) = (X_1(t), \ldots, X_n(t))$. The probability of moving to the state $\mathbf{x} + \boldsymbol{\mu}_k$ from $\mathbf{x}$ after time Δ is $\Pr(\mathbf{X}(t+\Delta) = \mathbf{x} + \boldsymbol{\mu}_k | \mathbf{X}(t) = \mathbf{x}) = \lambda_k(\mathbf{x})\Delta + o(\Delta)$, with λ_k the propensity of reaction k, the functional form of which is assumed to follow the principle of mass-action $\lambda_k(\mathbf{x}) = c_k \prod_{i=1}^n \binom{x_i}{\nu_{ik}}$ [36]. The generator matrix $\mathbf{R}$: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of the CTMC is determined by $\mathbf{R}(\mathbf{x}, \mathbf{x} + \boldsymbol{\mu}_k) = \lambda_k(\mathbf{x})$, $\mathbf{R}(\mathbf{x}, \mathbf{x}) = -\sum_{k=1}^n \mathbf{R}(\mathbf{x}, \mathbf{x} + \boldsymbol{\mu}_k)$, and zero otherwise.

To illustrate, assume that a gene G spontaneously turns on and off at rates c_1 and c_2 respectively, and that it regulates the expression of protein P. More precisely, whenever a gene is turned on, the protein is synthesized at a rate c_3, such that $c_1, c_2 \ll c_3$, that is, the gene activation is slow relative to the rate of protein synthesis. Such a system requires a stochastic model and it can be specified with the following set of reactions:

$$G_0 \underset{c_2}{\overset{c_1}{\rightleftharpoons}} G_1, \quad G_1 + P_0 \xrightarrow{c_3} G_1 + P, \quad P \xrightarrow{c_4} P_0$$

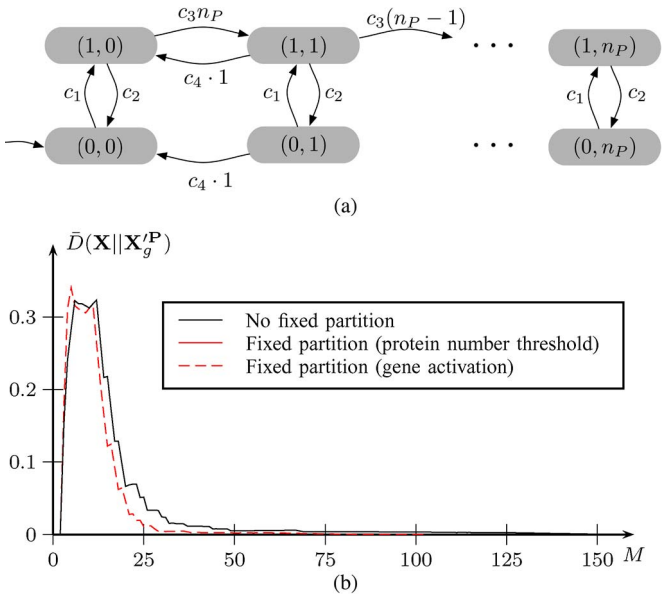

Fig. 4. Application to modeling bio-molecular interactions: (a) Continuous-time Markov chain assigned to the gene expression example. (b) Upper bound to the KLDR obtained by agglomerative information bottleneck method, for all partition sizes: (black) no partition class is fixed, (red) a partition class where the number of proteins is bigger than a threshold $T = 0.9n = 180$ is fixed, (red, dashed) all states where the gene is turned on (a cluster with 101 states) are fixed. The red line is not visible because it equals the black line. The aggregation error is displayed only up to a partition size of $M = 150$.

Here, P_0 is introduced to simplify the system so that the total number of proteins is n_P. This is arguably more realistic than the unlimited birth-death process, as P_0 represents the (limited) pool of amino-acid building blocks for the proteins. Finally, the protein can spontaneously degrade at rate c_4.

Since the Markov process assigned to the model of a biochemical reaction network evolves in continuous-time, we cannot directly apply our aggregation method to the CTMC model of a biochemical network. Instead, we aggregate the subordinated DTMC:

Definition (Subordinated DTMC): Let $\mathbf{X}$ be a CTMC over the state space $\mathcal{X}$ with generator matrix $\mathbf{R}$ and transient marginal distribution $\boldsymbol{\pi}$, such that $\pi_i(t) = \Pr(\mathbf{X}(t) = i)$. For $\lambda \geq \sup_{i\in\mathcal{X}} |\mathbf{R}_{ii}|$, let $\mathbf{P} := \mathbf{R}/\lambda + \mathbf{I}_N$ ($\mathbf{I}_N$ is an identity matrix of dimension N). The DTMC defined by $\mathbf{P}$ is the subordinated process of $\mathbf{X}$ with uniformization constant λ, denoted by $\mathbf{X}_\lambda$.

The subordinated[5] DTMC agrees with the original process in its transient distribution [37]. Moreover, the KLDR between the subordinated DTMCs equals the KLDR between the original CTMCs in the limit of a large uniformization constant, as can be shown by discretizing the time domain (see [38] for detailed presentation). The definition of the KLDR for CTMCs and its existence criterion can be found in [39, Ch. 6]. In the algorithm, we choose the uniformization constant $\lambda = \sup_{i\in\mathcal{X}} |\mathbf{R}_{ii}| + 1$.

For the initial vector $\mathbf{X}(0) = (1, n_P)$ (where the components denote copy numbers of G_1 and P respectively), the CTMC has $N = 2(n_P + 1)$ reachable states (Fig. 4(a)). After the chain exhibits stationary behavior, the algorithm is applied for $n_P = 100$ and for $M = 1, 2, \ldots, 202$. Moreover, the algorithm was adapted to search for the optimal partition after one partition class is fixed. This is desirable in scenarios where the modeler *a priori* wants to track the joint probability of all the states that satisfy a certain property. For example, one may be interested in *a priori* clustering those states for which the number of proteins is bigger than a given threshold $T = 0.9n_P$, or all the states where the gene is turned on (depicted in the top row in Fig. 4(a)). In Fig. 4(b), we compare the upper bound on the aggregation error for the optimal partition, and optimal partition upon fixing each of the two mentioned partition classes. The results confirm that our algorithm provides only suboptimal solutions, because, for example, lumping *a priori* all states with an activated gene yields a better bound than when no partition class is fixed. In particular, notice that for $M = 2$, lumping all states where the gene is turned on satisfies the criterion of lumpability, rendering the upper bound on the error to be within numerical precision.

[5]The DTMC $\mathbf{X}_\lambda$ is also called a *uniformized* or *randomized* chain.

X. Conclusion

In this work, we presented a new method for Markov chain state space reduction based on information-theoretic criteria. Specifically, the Kullback–Leibler divergence rate between the process obtained by simply partitioning the alphabet of the original chain and its best Markov approximation is employed as a cost function. The Kullback–Leibler divergence rate between the original chain and the *lifting* of the optimal Markov approximation was shown to yield an upper bound on the cost function.

By properly defining the lifting, we not only obtain the best upper bound under certain restrictions, but also a cost function which links the reduced-alphabet model to the notion of lumpability. In addition to that, it is shown that the information bottleneck method can be used for model reduction by relaxing the optimization problem. Future work shall investigate possible connections between the proposed cost function and the spectral theory of Markov chains, the extension to non-stationary Markov chains, and the generalization to stochastic aggregations.

Appendix A
Proof of Proposition 1

For the first property (which is also mentioned in [3]) we show that the condition

$$\mathbf{V}\mathbf{U}^{\boldsymbol{\zeta}}\mathbf{P}'\mathbf{V} = \mathbf{P}'\mathbf{V} \tag{58}$$

from Lemma 1 holds for all possible $\boldsymbol{\pi}$-liftings and for all positive probability vectors $\boldsymbol{\zeta}$. Letting $\mathbf{P}' = \mathbf{V}\mathbf{Q}\mathbf{U}^{\boldsymbol{\pi}}$

$$\mathbf{V}\mathbf{U}^{\boldsymbol{\zeta}}\mathbf{V}\mathbf{Q}\mathbf{U}^{\boldsymbol{\pi}}\mathbf{V} = \mathbf{V}\mathbf{Q}\mathbf{U}^{\boldsymbol{\pi}}\mathbf{V}. \tag{59}$$

Since $\mathbf{U}^{\boldsymbol{\pi}}\mathbf{V} = \mathbf{I}$ for all positive probability vectors $\boldsymbol{\pi}$, equality is achieved and the first result is proved.

For the second property (cf. [3, Tm. 2, Prop. 3]) note that with $\mathbf{P}' = \mathbf{V}\mathbf{Q}\mathbf{U}^{\boldsymbol{\mu}}$,

$$\boldsymbol{\mu}^T\mathbf{P}' = \boldsymbol{\mu}^T\mathbf{V}\mathbf{Q}\mathbf{U}^{\boldsymbol{\mu}} = \boldsymbol{\nu}^T\mathbf{Q}\mathbf{U}^{\boldsymbol{\mu}} = \boldsymbol{\nu}^T\mathbf{U}^{\boldsymbol{\mu}} = \boldsymbol{\mu}^T \tag{60}$$

where the second and third equality are due to Lemma 1 and the last follows from the definition of $\mathbf{U}^{\boldsymbol{\mu}}$ in (12).

For the third property we refer the reader to [3, Th. 1].

Next, observe that $Q_{g(i)g(j)} = 0$ implies $P_{ij} = 0$ (see (22), combined with the fact that $\boldsymbol{\mu}$ is positive [16, Th. 4.1.4]). For the entries of the lifted matrix we can now write

$$P'_{ij} = \frac{\pi_j}{\sum_{k\in g^{-1}(g(j))} \pi_k} Q_{g(i)g(j)}. \tag{61}$$

Since $\boldsymbol{\pi}$ is positive, it follows that $P'_{ij} = 0$ implies $P_{ij} = 0$, or equivalently, $\mathbf{P}' \gg \mathbf{P}$.

The last property immediately follows from Lemma 2; the lemma may be applied because $\mathbf{X}'^{\boldsymbol{\mu}}_g$ is lumpable to $\mathbf{Y}'_g$ (property 1) and $\mathbf{P}' \gg \mathbf{P}$ (property 4). This completes the proof.

Appendix B
Proof of Theorem 1

For the proof we note that another condition for lumpability is given by the entries of the matrix $\mathbf{R} = \mathbf{PV}$. In particular, iff for all $h, l \in \mathcal{Y}$ the elements

$$R_{il} := \sum_{j\in g^{-1}(l)} P_{ij} \tag{62}$$

are the same for all $i \in g^{-1}(h)$, the chain is lumpable w.r.t. g [16, Th. 6.3.2]. Using this with (27) one gets

$$\hat{R}_{il} = \sum_{j\in g^{-1}(l)} \hat{P}_{ij} = Q_{hl}. \tag{63}$$

Clearly, $\hat{R}_{il}$ assumes the same values for all $i \in g^{-1}(h)$, as required. This completes the proof of the first statement.[6]

The second statement is obvious from the definition of $\mathbf{P}$-lifting and from the proof of property 4 of Proposition 1.

For the third statement, we introduce an arbitrary lifting

$$\tilde{P}_{ij} = b_{ij} Q_{g(i)g(j)} \tag{64}$$

subject to $\sum_{j\in g^{-1}(l)} b_{ij} = 1$ for all $l \in \mathcal{Y}$ and all $i \in \mathcal{X}$. With (62), this condition is necessary and sufficient for lumpability of the lifted chain $\tilde{\mathbf{X}}$ with transition matrix $\tilde{\mathbf{P}}$. We write for the KLDR

$$\bar{D}(\mathbf{X}\|\tilde{\mathbf{X}}) = \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{P_{ij}}{\tilde{P}_{ij}} \tag{65}$$

$$= \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{P_{ij}}{b_{ij} Q_{g(i)g(j)}} \tag{66}$$

$$= \bar{H}\left(\mathbf{Y}'_{\mathbf{g}}\right) - \bar{H}(\mathbf{X}) + \sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{1}{b_{ij}}. \tag{67}$$

[6]Note that this statement holds for all stochastic matrices used for lifting, i.e., the lifting matrix does not have to be equal to the transition matrix of the original chain.

The last term can be written as

$$\sum_{i,j\in\mathcal{X}} \mu_i P_{ij} \log \frac{1}{b_{ij}} = \sum_{i\in\mathcal{X}} \mu_i \sum_{l\in\mathcal{Y}} R_{il} \sum_{j\in g^{-1}(l)} \frac{P_{ij}}{R_{il}} \log \frac{1}{b_{ij}}.$$

Here, the last term on the right is a cross-entropy, since both b_{ij} and P_{ij}/R_{il} are probability vectors on $g^{-1}(l)$. The cross-entropy is minimized[7] iff for all $j \in g^{-1}(l)$

$$b_{ij} = \frac{P_{ij}}{R_{il}} = \frac{P_{ij}}{\sum_{k\in g^{-1}(l)} P_{ik}}. \tag{68}$$

Since the sums over i and l are expecations, the minimum is achieved iff above condition holds also for all $i \in \mathcal{X}$ and all $l \in \mathcal{Y}$ for which $R_{il} > 0$. If $R_{il} = 0$, the assignment for b_{ij} is immaterial for $j \in g^{-1}(l)$.

To show that the $\mathbf{P}$-lifting indeed yields a better bound observe that with $H(Y_{g,n}|Y_{g,n-1}) = \bar{H}(\mathbf{Y}'_{\mathbf{g}})$

$$\bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\boldsymbol{\mu}}_g\big) - \bar{D}\big(\mathbf{X}\|\mathbf{X}'^{\mathbf{P}}_g\big)$$

$$= H(X) - \bar{H}(\mathbf{X}) - H\left(Y'_g\right) + H(Y_{g,n}|X_{n-1}) \tag{69}$$

$$= H(X_n) - H(X_n|X_{n-1}) - H(Y_{g,n}) + H(Y_{g,n}|X_{n-1}) \tag{70}$$

$$= I(X_n; X_{n-1}) - I(Y_{g,n}; X_{n-1}) \tag{71}$$

$$\geq 0 \tag{72}$$

by the data processing inequality. $\bar{D}(\mathbf{X}|\mathbf{X}'^{\mathbf{P}}_g) \geq \bar{D}(\mathbf{Y}_g\|\mathbf{Y}'_g)$ is obtained by Lemma 2, see Proposition 1, property 5).

For the fifth property, note that the sufficient and necessary condition for lumpability (13), namely that

$$R_{il} = \sum_{j\in g^{-1}(l)} P_{ij} = Q_{hl} \tag{73}$$

is the same for all $i \in g^{-1}(h)$, can be used in the definition of $\hat{\mathbf{P}}$:

$$\hat{P}_{ij} = \frac{P_{ij}}{\sum_{k\in\mathcal{S}_j} P_{ik}} Q_{g(i)g(j)} = \frac{P_{ij}}{\sum_{k\in\mathcal{S}_j} P_{ik}} R_{ig(j)} = P_{ij} \tag{74}$$

This proves the "$\Rightarrow$" part. The "$\Leftarrow$" part follows from Lemma 2. This completes the proof.

Acknowledgment

The authors wish to thank Kun Deng and Prashant Mehta for providing the data set for the example in Section VIII-B. The authors are particularly indebted to the reviewers providing helpful comments to improve the quality of the paper.

References

[1] D. Wilkinson, *Stochastic Modelling for Systems Biology*. Boca Raton, FL, USA: Taylor and Francis, 2006, ser. Chapman and Hall/CRC Mathematical and Computational Biology.

[7]This is a direct consequence of the fact that the Kullback–Leibler divergence vanishes if and only if the considered probability mass functions are equal [24, p. 31].

[2] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.
[3] K. Deng, P. G. Mehta, and S. P. Meyn, "Optimal Kullback–Leibler aggregation via spectral theory of Markov chains," *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2793–2808, Dec. 2011.
[4] R. W. Aldhaheri and H. K. Khalil, "Aggregation of the policy iteration method for nearly completely decomposable Markov chains," *IEEE Trans. Autom. Control*, vol. 36, no. 2, pp. 178–187, Feb. 1991.
[5] Z. Rached, F. Alajaji, and L. L. Campbell, "The Kullback–Leibler divergence rate between Markov sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 917–921, May 2004.
[6] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Allerton Conf. Commun., Control, Comput.*, Sep. 1999, pp. 368–377.
[7] L. B. White, R. Mahony, and G. D. Brushe, "Lumpable hidden Markov models-model reduction and reduced complexity filtering," *IEEE Trans. Autom. Control*, vol. 45, no. 12, pp. 2297–2306, Dec. 2000.
[8] Q.-S. Jia, "On state aggregation to approximate complex value functions in large-scale Markov decision processes," *IEEE Trans. Autom. Control*, vol. 56, no. 2, pp. 333–344, Feb. 2011.
[9] M. Vidyasagar, "Reduced-order modeling of Markov and hidden Markov processes via aggregation," in *Proc. IEEE Conf. Decision Control (CDC)*, Atlanta, Dec. 2010, pp. 1810–1815.
[10] QCK. Deng and D. Huang, "Model reduction of Markov chains via low-rank approximation," in *Proc. Amer. Control Conf. (ACC)*, Montreal, Canada, Jun. 2012, pp. 2651–2656.
[11] M. A. Katsoulakis and J. Trashorras, "Information loss in coarse-graining of stochastic particle dynamics," *J. Statist. Phys.*, vol. 122, no. 1, pp. 115–135, 2006.
[12] M. Meilă and J. Shi, "Learning segmentation by random walks," in *Advances in Neural Information Processing Systems (NIPS)*, Denver, CO, USA, Nov. 2000, pp. 1–7.
[13] T. Runolfsson and Y. Ma, "Model reduction of nonreversible Markov chains," in *Proc. IEEE Conf. Decision Control (CDC)*, New Orleans, LA, USA, Dec. 2007, pp. 3739–3744.
[14] M. Vidyasagar, "A metric between probability distributions on finite sets of different cardinalities and applications to order reduction," *IEEE Trans. Autom. Control*, vol. 57, no. 10, pp. 2464–2477, Oct. 2012.
[15] B. C. Geiger and G. Kubin, "Signal enhancement as minimization of relevant information loss," in *Proc. ITG Conf. on Systems, Communication and Coding*, Munich, Germany, Jan. 2013, pp. 1–6, extended version available: *arXiv:1205.6935 [cs.IT]*.
[16] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, 2nd ed. Berlin, Germany: Springer, 1976.
[17] A. Raj and C. H. Wiggins, "An information-theoretic derivation of min-cut-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 988–995, Jun. 2010.
[18] N. Tishby and N. Slonim, "Data clustering by Markovian relaxation and the information bottleneck method," *Advances in Neural Information Processing Systems (NIPS)*, 2001.
[19] A. Friedman and J. Goldberger, "Information theoretic pairwise clustering," in *SIMBAD*, vol. 7953, E. Hancock and M. Pelillo, Eds. Berlin, Germany: Springer, 2013, ser. LNCS, pp. 106–119.
[20] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems (NIPS)*. Cambridge, MA, USA: MIT Press, 1999, pp. 617–623.
[21] Y. Xu, S. M. Salapaka, and C. L. Beck, "Aggregation of graph models and Markov chains by deterministic annealing," *IEEE Trans. Autom. Control*, vol. 59, no. 10, pp. 2807–2812, Oct. 2014.
[22] J. C. Kieffer and M. Rahe, "Markov channels are asymptotically mean stationary," *Siam J. Mathemat. Anal.*, vol. 12, pp. 293–305, 1980.
[23] R. M. Gray, *Probability, Random Processes, Ergodic Properties*, 2nd ed. New York, NY, USA: Springer, 2009.
[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley Interscience, 2006.
[25] R. M. Gray, *Entropy and Information Theory*. New York, NY, USA: Springer, 1990.
[26] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA, USA: Holden Day, 1964.
[27] S. Watanabe and C. T. Abraham, "Loss and recovery of information by coarse observation of stochastic chain," *Information and Control*, vol. 3, no. 3, pp. 248–278, Sep. 1960.
[28] J. Feret, T. Henzinger, H. Koeppl, and T. Petrov, "Lumpability abstractions of rule-based systems," *Theoretical Comput. Sci.*, vol. 431, pp. 137–164, 2012.
[29] E. Weinan, L. Tiejun, and E. Vanden-Eijnden, "Optimal partition and effective dynamics of complex networks," *PNAS*, vol. 105, no. 23, pp. 7907–7912, Jun. 2008.
[30] M. Vidyasagar, "Kullback–Leibler divergence rate between probability distributions on sets of different cardinalities," in *Proc. IEEE Conf. Decision and Control*, Atlanta, GA, Dec. 2010, pp. 948–953.
[31] B. C. Geiger and C. Temmel, Lumpings of Markov Chains, Entropy Rate Preservation, Higher-Order Lumpability, Dec. 2012, accepted in J. Appl. Prob.; preprint available: *arXiv:1212.4375 [cs.IT]*.
[32] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2000, pp. 208–215.
[33] M. Wohlmayr, M. Markaki, and Y. Stylianou, "Speech-nonspeech discrimination based on speech-relevant spectrogram modulations," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Poznan, Poland, Sep. 2007, pp. 1551–1555.
[34] R. M. Hecht, E. Noor, and N. Tishby, "Speaker recognition by Gaussian information bottleneck," in *Proc. InterSpeech*, Brighton, U.K., Sep. 2009, pp. 1567–1570.
[35] A. Vedaldi and B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms, 2008. [Online]. Available: http://www.vlfeat.org/
[36] D. T. Gillespie, "Stochastic Simulation of Chemical Kinetics," *Ann. Rev. Phys. Chem.*, vol. 58, no. 1, pp. 35–55, 2007.
[37] J. R. Norris, *Markov Chains*. Cambridge, U.K.: Cambridge University Press, 1998, no. 2008.
[38] I. Cohn, T. El-Hay, N. Friedman, and R. Kupferman, "Mean field variational approximation for continuous-time Bayesian networks," *J. Mach. Learn. Res.*, vol. 9999, pp. 2745–2783, 2010.
[39] T. Petrov, "Formal reductions of stochastic rule-based models of biochemical systems," Ph.D. dissertation, ETHZ, 2013.

Bernhard C. Geiger (S'07–M'14) was born in Graz, Austria, in 1984. He received the Dipl.-Ing. degree in electrical engineering (with distinction) and the Dr. Techn. degree in electrical and information engineering (with distinction) from Graz University of Technology in 2009 and 2014, respectively.

In 2009, he joined the Signal Processing and Speech Communication Laboratory, Graz University of Technology, as a Project Assistant and took a position as a Research and Teaching Associate at the same lab in 2010. He is currently a postdoctoral researcher at the Institute for Communications Engineering, TU Munich. His research interests cover the intersection of information theory with system theory and signal processing, certain topics in the theory of Markov chains, and the analysis of GNSS acquisition.

Tatjana Petrov received the M.Sc. degree in theoretical computer science from the University of Novi Sad, Serbia, and the Ph.D. degree from the Swiss Federal School of Technology Zurich (ETH Zurich). In her Ph.D. work, in the scope of SystemsX (the Swiss Initiative for Systems Biology), towards understanding complex signaling pathway dynamics, she developed a formal framework for exact and approximate reductions of stochastic rule-based models of complex biochemical networks.

She is a postdoctoral Fellow at IST Austria, where her current work focuses on formal verification approaches to modeling and analysis of complex systems, with a major interest in stochastic and hybrid systems, interface theories, and applications to molecular biology.

Gernot Kubin (M'84) was born in Vienna, Austria, on June 24, 1960. He received the Dipl.-Ing. degree in 1982 and the Dr. Techn. degree (*sub auspiciis praesidentis*) in 1990 in electrical engineering from the Technical University of Vienna, Vienna, Austria.

He is a Professor of Nonlinear Signal Processing and has been Head of the Signal Processing and Speech Communication Laboratory (SPSC) at the Technical University (TU) of Graz, Graz, Austria, since 2000. At TU Graz, he has been Dean of Studies in EE-Audio Engineering 2004–2007, Chair of the Senate 2007–2010 and 2013–present, and he has coordinated the Doctoral School in Information and Communications Engineering since 2007. Earlier international appointments include: CERN Geneva/CH 1980, TU Vienna 1983–2000, Erwin Schroedinger Fellow at Philips Natuurkundig Laboratorium Eindhoven/NL 1985, AT&T Bell Labs Murray Hill/USA 1992–1993 and 1995, KTH Stockholm/S 1998, and Global IP Sound Sweden and USA 2000–2001 and 2006, University of California (UC) San Diego and UC Berkeley/USA 2006, and UT Danang, Vietnam, 2009. In 2011, he co-founded Synvo GmbH, a start-up in the area of speech synthesis for mobile devices, and he holds leading positions in several national research centres for academia-industry collaboration such as the Vienna Telecommunications Research Centre FTW 1999–now (Key Researcher and Board of Governors), the Christian Doppler Laboratory for Nonlinear Signal Processing 2002–2010 (Founding Director), the Competence Network for Advanced Speech Technologies COAST 2006–2010 (Scientific Director), the COMET Excellence Projects Advanced Audio Processing AAP 2008–2013 and Acoustic Sensing and Design 2013–now (Key Researcher), and in the National Research Network on Signal and Information Processing in Science and Engineering SISE 2008–2011 (Principal Investigator) funded by the Austrian Science Fund. Since 2011, His research interests are in nonlinear signals and systems, computational intelligence, as well as speech and audio communication. He has authored or co-authored over 150 peer-reviewed publications and ten patents.

Dr. Kubin is an elected member of the Speech and Language Processing Technical Committee of the IEEE.

Heinz Koeppl received the M.Sc. degree in physics from Graz Karl–Franzens University, Graz, Austria, in 2001 and the Ph.D. degree in electrical engineering from Graz University of Technology in 2004.

He was a postdoctoral Fellow at the University of California at Berkeley and at EPF Lausanne Switzerland until 2009, and an Assistant Professor at ETH Zurich, Switzerland, until 2013. He is currently a full Professor of Electrical Engineering at the Technische Universitaet Darmstadt, Darmstadt, Germany. His research interests include the analysis and reconstruction of stochastic biomolecular networks and self-organizing phenomena such as flocking and self-assembly.

Dr. Koeppl received the Erwin Schrödinger fellowship in 2005, the IFAC Fred Margulies Ph.D. thesis award in 2006, the CNRS/Max-Planck postdoctoral fellowship in 2008, the SNSF Professorship in 2010, and the IBM Faculty Award in 2014.