Schedae Informaticae Vol. 23 (2014): 45–56
doi: 10.4467/20838476SI.14.004.3021

Markov State Space Aggregation via the Information Bottleneck Method

Bernhard C. Geiger
Institute for Communications Engineering
TU Munich
Arcisstraße 21, D-80333 Munich
e-mail: *geiger@ieee.org*

Abstract. Consider the problem of approximating a Markov chain by another Markov chain with a smaller state space that is obtained by partitioning the original state space. An information-theoretic cost function is proposed that is based on the relative entropy rate between the original Markov chain and a Markov chain defined by the partition. The state space aggregation problem can be sub-optimally solved by using the information bottleneck method.

Keywords: Markov chains, state space aggregation, coarse-graining, information bottleneck, relative entropy, lumpability.

1. Introduction

Markov models appear in many scientific disciplines such as systems biology, natural language processing, information theory, and automated control. These models are popular because the Markov property greatly simplifies analysis and simulation. However, sometimes the state space of a Markov model is too large to admit simulation or inference of model parameters from real-world data sets. For example, dealing with the state space explosion is a major challenge in n-gram word models [13].

One way to reduce the state space of a Markov chain is aggregation: A non-injective function induces a partition of the original state space, effectively grouping or aggregating states of the original chain together. The aggregated process, is a Markov chain whose state space consists of the groups of states rather than of the individual states. State space aggregation has attracted much attention during

the last years, e.g., in chemical reaction networks [14], control theory [24], or in [21], which used total variational distance for aggregation. Most relevant to our work are information-theoretic cost functions [2,7,11,23] and information-theoretic graph clustering [4,16,20].

Partitioning the state space does not suffice for aggregation. To obtain a Markov chain on the smaller state space, one has to choose transition probabilities. This paper draws heavily on our work [5] and chooses the transition probabilities by minimizing an information-theoretic cost function. We show that the same cost function can be relaxed in order to employ a standard method from the machine learning literature – the information bottleneck method [19] – to find the optimal partition of the original state space. Finally, we show that our proposed cost function is related to the concept of *lumpability*, i.e., a function of a Markov chain has the Markov property.

This paper accompanies an invited lecture at the Theoretical Foundations of Machine Learning Conference. The paper is written in tutorial-style and tries to give an overview of our state space aggregation method. For the mathematical details and rigorous proofs the reader is referred to [5].

2. Notation, Random Variables, and Markov Chains

In this work, upper case letters denote random variables (RVs), calligraphic letters their state space, and lower case letters their realizations; for example, the RV X may assume the value x, which is an element of the state space $\mathcal{X}$. Since we are dealing with discrete RVs on finite alphabets only, the distribution of X is determined by its probability mass function (PMF) p_X, where

$$\forall x \in \mathcal{X}: \quad p_X(x) := \Pr(X = x). \tag{1}$$

The joint PMF of several RVs and the conditional PMF of a set of RVs given another set of RVs are defined similarly.

We denote discrete-time stochastic processes by bold-faced letters, e.g., $\mathbf{X}$; their samples are RVs indexed by natural numbers, i.e., $X_1, X_2, \ldots$. Each RV X_n takes values from the same, finite, state space $\mathcal{X}$. The random processes considered in this work are *stationary*. In particular, the PMF of X_n is equal for all n and shall be denoted as p_X.

This work deals with Markov chains, i.e., stochastic processes that satisfy the Markov property so that the distribution of future samples depends exclusively on the present sample, but not on past samples. In other words, if $\mathbf{X}$ is a Markov chain with finite state space $\mathcal{X} = \{1, 2, \ldots, N\}$, then

$$\Pr(X_{n+1} = j | X_n = i, X_{n-1} = h, \ldots) = \Pr(X_{n+1} = j | X_n = i) \tag{2}$$

for all n and all realizations $j, i, h, \cdots \in \mathcal{X}$ [12, Def. 2.1.1]. We consider irreducible, aperiodic, time-homogeneous (i.e., the probability on the r.h.s. of (2) does not

depend on n), and stationary Markov chains; see [12] for terminology and basic results. Under these assumptions the behavior of $\mathbf{X}$ is uniquely determined by its transition probability matrix $\mathbf{P} = \{P_{i,j}\}$, where $P_{i,j} := \Pr(X_n = j | X_{n-1} = i)$. Its unique invariant distribution vector $\boldsymbol{\mu}$ with its i-th component given by

$$\mu_i := p_X(i) > 0 \tag{3}$$

satisfies $\boldsymbol{\mu}^T = \boldsymbol{\mu}^T \mathbf{P}$ [12, Thm. 4.1.6]. To guarantee that $\mathbf{X}$ is stationary, its initial distribution must coincide with the invariant distribution. For such a Markov chain we use the shorthand notation $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$.

Suppose we partition the state space $\mathcal{X}$ of the Markov chain $\mathbf{X}$ by a non-injective function $g\colon \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, M\}$ for $M < N$. In other words, g induces a partition of $\mathcal{X}$ by the preimages of the elements of $\mathcal{Y}$: For $l \in \mathcal{Y}$, $g^{-1}(l) := \{i \in \mathcal{X} \mid g(i) = l\}$ is an element of the partition of $\mathcal{X}$. The sequence of samples $Y_n := g(X_n)$ obtained by projecting the Markov chain through the function g defines another stationary stochastic process, which we will henceforth call the *projection* of $\mathbf{X}$. From $\boldsymbol{\mu}$, $\mathbf{P}$, and g the joint PMF of two consecutive samples of $\mathbf{X}$ and/or $\mathbf{Y}$ can be computed as follows:

$$p_{X_1,X_2}(i,j) = \mu_i P_{i,j} \tag{4}$$

$$p_{X_1,Y_2}(i,l) = \mu_i \sum_{j \in g^{-1}(l)} P_{i,j} \tag{5}$$

$$p_{Y_1,Y_2}(k,l) = \sum_{i \in g^{-1}(k)} \sum_{j \in g^{-1}(l)} \mu_i P_{i,j} \tag{6}$$

The conditional PMFs can be computed similarly.

It is well known that the projection $\mathbf{Y}$ in general loses the Markov property. The phenomenon where $\mathbf{Y}$ is an irreducible, aperiodic, and time-homogeneous Markov chain is called *lumpability* and has been treated in [12, §6.3], as well as in [9] and the references therein.

3. Relative Entropy

Our goal is to find a "good" aggregation of a Markov chain, and we need a measure for the dissimilarity between the distributions of two RVs or two stochastic processes. Possible measures are the total variational distance or Pearson's χ^2-divergence. Since we will be dealing with information-theoretic methods, relative entropy is the most immediate choice:

Definition 1 (Relative Entropy [1, Ch. 2.3]) *The relative entropy between two PMFs p_{X_1} and p_{X_2} with the same support $\mathcal{X}$ (corresponding to two RVs X_1 and X_2 with the same state space) is*

$$D(p_{X_1} || p_{X_2}) := \sum_{i:\ p_{X_1}(i) > 0} p_{X_1}(i) \log \frac{p_{X_1}(i)}{p_{X_2}(i)}. \tag{7}$$

Clearly, $D(p_{X_1}||p_{X_2})$ is finite only if $p_{X_2}(i) = 0$ implies $p_{X_1}(i) = 0$ for all $i \in \mathcal{X}$, or in short, if $p_{X_1} \ll p_{X_2}$. Moreover, $D(p_{X_1}||p_{X_2}) = 0$ if and only if the two PMFs are equal, i.e., $p_{X_1}(i) = p_{X_2}(i)$ for all $i \in \mathcal{X}$.

With this definition we can state a sufficient[1] condition for the Markov chain $\mathbf{X}$ and the function g such that the projection $\mathbf{Y}$ is a Markov chain:

Lemma 1 (Lumpability, adopted from [6, Thm. 9]) *For $\mathbf{X}$ being an irreducible, aperiodic, and stationary Markov chain and for a given function g, the projection $\mathbf{Y}$ is an irreducible, aperiodic, and stationary Markov chain if*

$$\mathbb{E}\left(D(p_{Y_2|X_1}(\cdot|X_1)||p_{Y_2|Y_1}(\cdot|g(X_1)))\right) = 0 \tag{8}$$

where the expectation is taken w.r.t. the distribution $\boldsymbol{\mu}$ of X_1.

Relative entropy, as a measure of dissimilarity between the distribution of RVs, can be extended to stochastic processes:

Definition 2 (Relative Entropy Rate [8, Ch. 10]) *The relative entropy rate between two stationary stochastic processes $\mathbf{X}_1$ and $\mathbf{X}_2$ with the same state space $\mathcal{X}$ is*

$$\bar{D}(\mathbf{X}_1||\mathbf{X}_2) := \lim_{n\to\infty} \frac{1}{n} D(p_{X_{1,1},X_{1,2},\ldots,X_{1,n}}||p_{X_{2,1},X_{2,2},\ldots,X_{2,n}}) \tag{9}$$

whenever the limit exists.

For this quantity to be finite, $p_{X_{1,1},X_{1,2},\ldots,X_{1,n}} \ll p_{X_{2,1},X_{2,2},\ldots,X_{2,n}}$ has to hold for all n. The limit exists, e.g., between a stationary stochastic process and a time-homogeneous Markov chain [8, Ch. 10] as well as between Markov chains (not necessarily stationary or irreducible) [15]. For example, for two Markov chains $\mathbf{X}_1 \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$ and $\mathbf{X}_2 \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}', \boldsymbol{\mu}')$, the relative entropy rate reads [15]

$$\bar{D}(\mathbf{X}_1||\mathbf{X}_2) = \sum_{i,j:\ P_{i,j}>0} \mu_i P_{i,j} \log \frac{P_{i,j}}{P'_{i,j}} \tag{10}$$

if $P'_{i,j} = 0 \Rightarrow P_{i,j} = 0$ (in short: $\mathbf{P} \ll \mathbf{P}'$).

In general, (10) can not be used to compute the relative entropy rate between a projection $\mathbf{Y}_1$ of a Markov chain $\mathbf{X}_1$ and Markov chain $\mathbf{Y}_2$ with state space $\mathcal{Y}$, because $\mathbf{Y}_1$ might not be Markov. However, in some cases this relative entropy rate can be bounded from above by (10), provided $\mathbf{X}_1$ and $\mathbf{X}_2$ are chosen appropriately.

Lemma 2 (Bound on the Relative Entropy Rate [5, Lem. 2]) *Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be irreducible, aperiodic, and stationary Markov chains with the same state space $\mathcal{X}$ having transition probability matrices $\mathbf{P}$ and $\mathbf{P}'$ such that $\mathbf{P} \ll \mathbf{P}'$. Let g: $\mathcal{X} \to \mathcal{Y}$, and let $\mathbf{Y}_1$ and $\mathbf{Y}_2$ be the projections of $\mathbf{X}_1$ and $\mathbf{X}_2$, respecively. Let additionally $\mathbf{X}_2$ be lumpable w.r.t. g, i.e., let $\mathbf{Y}_2$ be Markov. Then we have*

$$\bar{D}(\mathbf{Y}_1||\mathbf{Y}_2) \leq \bar{D}(\mathbf{X}_1||\mathbf{X}_2). \tag{11}$$

[1] In fact, the condition in [6, Thm. 9] is sufficient *and* necessary for *strong* lumpability in the sense of [12, § 6.3]. We omit the distinction between strong and weak lumpability in this paper.

4. Clustering and The Information Bottleneck Method

Relative entropy not only measures the dissimilarity between PMFs, it also allows us to measure the information two RVs X_1 and X_2 share: if two RVs are independent and hence do not share any information, their joint PMF equals the product of their marginal PMFs. Hence, two RVs share the more information the more their joint PMF differs from the product of their marginals:

Definition 3 (Mutual Information [1, Ch. 2.3]) *The mutual information of two RVs X and Y with state spaces $\mathcal{X}$ and $\mathcal{Y}$ is the relative entropy between their joint PMF and the product of their marginal PMFs:*

$$I(X;Y) := D(p_{X,Y}||p_X p_Y) = \sum_{i,k:\ p_{X,Y}(i,k)>0} p_{X,Y}(i,k) \log \frac{p_{X,Y}(i,k)}{p_X(i)p_Y(k)} \tag{12}$$

One is often interested in a compressed description Y of an observation X. Rate-distortion theory (e.g., [8, Ch. 9] or [1, Ch. 10]) deals with this problem from an information-theoretic point-of-view, and achieves compression by minimizing the mutual information between X and its compressed representation Y. At the same time, however, Y should be capable of delivering sufficient information about the observation X, which is achieved by simultaneously minimization a distortion criterion d (e.g., the mean-squared reconstruction error). This can be formulated as a variational problem

$$\arg\min_{p_{Y|X}} \ I(X;Y) + \beta d(X,Y) \tag{13}$$

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

Compressing observations becomes particularly interesting when the distortion is not measured w.r.t. the observation X itself, but w.r.t. some related RV C. For example, if X are observed features (e.g, directions of pencil strokes) and C is a class variable (e.g., the numbers from 0 to 9), then the compression Y should reveal as much information as possible about the class C, rather than about the observation X. The information bottleneck (IB) method [19] formulates exactly this problem by using relative entropy as distortion criterion, resulting in the following variational problem:

$$\arg\min_{p_{Y|X}} \ I(X;Y) - \beta I(C;Y) \tag{14}$$

In other words, Y should be a highly compressed version of X, but it should contain as much information about C as possible; β trades between these two objectives.

For $\beta \to \infty$ and with the restriction to deterministic compressions $p_{Y|X}$ determined by functions $g\colon \mathcal{X} \to \mathcal{Y}$, in [18] an iterative procedure, called *agglomerative IB* was introduced. It 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 merging step, but does not guarantee that the global optimum is achieved.

Sometimes the class variable C is not available or known, as it often happens with unlabelled data sets. In these cases, the cluster structure should reveal itself from the data. For the scenario in which the data is given only via a matrix of pairwise distances or dissimilarities, Friedman and Goldberger [4] suggested to convert the pairwise distance matrix to the transition probability matrix of a stationary Markov chain $\mathbf{X}$. Then, they looked for the function g maximizing the mutual information between consecutive samples of the projection, i.e., they maximized $I(Y_1; Y_2)$; in that sense, their work is strongly related to Dhillon's approach to information-theoretic co-clustering [3]. For the same setting of data given only via pairwise distances, also Tishby and Slonim [20] used the Markov chain approach. They, however, first let the Markov chain relax to some quasi-stable time point n, which is characterized by a slowly changing $I(X_1; X_k)$ for k close to n. Assuming that this X_n reveals a lot about the cluster structure of the data, they applied the agglomerative IB method to maximize $I(X_n; Y_1)$: In other words, they were looking for a function g, such that the projection Y_1 of the initial state X_1 reveals as much information about the relaxed state X_n as possible.

5. From Information-Theoretic Clustering to Markov State Space Aggregation

We now consider the problem of state space aggregation for Markov chains, namely the problem of defining a Markov chain on a smaller state space. In other words, given a Markov chain $\mathbf{X}$ with state space $\mathcal{X}$, we are interested in finding a Markov chain $\mathbf{Y}_{\mathcal{M}}$ with a smaller state space $\mathcal{Y}$ which is similar to $\mathbf{X}$ in a well-defined sense.

While Tishby and Slonim [20] employed a Markov chain approach for clustering of data given by pairwise distances, they did not recognize that their IB approach can be also employed for state space aggregation. Contrary to that, Friedman and Goldberger applied their method for pairwise clustering [4] also to state space aggregation of a Markovian movement model [7]. In addition to that, their cost function $I(Y_1; Y_2)$ is identical to the one employed by Deng, Mehta, and Meyn [2, Lem. 3], who focused their work solely on the state space aggregation of (nearly completely decomposable) Markov chains and introduced a connection between information-theoretic and spectral clustering. In the author's opinion, the coincidence of the cost functions of [4] and [2] is accidental: Deng, Mehta, and Meyn did not intend to maximize $I(Y_1; Y_2)$ as such, but proposed to minimize the relative entropy rate between the original Markov chain $\mathbf{X}$ and a Markov approximation of its projection $\mathbf{Y}$ to find the most suitable function g. Since these two Markov chains have state spaces of different cardinalities, they had to *lift* the Markov approximation of $\mathbf{Y}$ to the state space $\mathcal{X}$. In doing so, they employed the same lifting method as Vidyasagar [23] and arrived at the same cost function as [4].

Our approach to Markov state space aggregation introduced in [5] draws from these references: Employing relative entropy rate as a cost function, we have to resort to lifting; however, our lifting differs from the one proposed in [2, 23]. Then,

by relaxing the optimization problem, we show that the IB method can be used to obtain a state space aggregation; the relevance variable we use, however, is different from the one of [20]. This way, we try to solve the following problem:

Definition 4 (M-partition problem) *Given a Markov chain* $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$ *with* $\mathcal{X} = \{1, \dots, N\}$ *and a state space* $\mathcal{Y} = \{1, \dots, M\}$, $M < N$, *the* M-partition problem *searches for the function* $g\colon \mathcal{X} \to \mathcal{Y}$ *such that the relative entropy rate between the projection* $\mathbf{Y}$ *of* $\mathbf{X}$ *and its best Markov approximation is minimal, i.e., it solves*

$$\arg\min_{g \in [\mathcal{X} \to \mathcal{Y}]} \min_{\mathbf{Y}_{\mathcal{M}}} \{\bar{D}(\mathbf{Y} || \mathbf{Y}_{\mathcal{M}}) \mid \mathbf{Y}_{\mathcal{M}} \textit{ is Markov}\}. \tag{15}$$

The rationale for stating the problem this way is as follows: When looking for a Markov state space aggregation, it is desirable to choose the transition probabilities for $\mathbf{Y}_{\mathcal{M}}$ such that the resulting chain closely resembles the projection $\mathbf{Y}$ of the original chain, as this allows to train the Markov model $\mathbf{Y}_{\mathcal{M}}$ based on a realization of the projection. Choosing the function g in order to minimize $\bar{D}(\mathbf{Y}||\mathbf{Y}_{\mathcal{M}})$ consequently minimizes the training error.

The M-partition problem can be greatly simplified by noting that the best Markov approximation $\mathbf{Y}_{\mathcal{M}}$ (in the sense of relative entropy rate) to a given stationary stochastic process $\mathbf{Y}$ is given by the joint PMF of two consecutive samples. Mathematically, if for all $k, l \in \mathcal{Y}$

$$\nu_k = p_{Y_1}(k) \tag{16}$$

$$Q_{k,l} = p_{Y_2|Y_1}(l, k) = \frac{p_{Y_1,Y_2}(k, l)}{p_{Y_1}(k)} \tag{17}$$

then $\mathbf{Y}_{\mathcal{M}} \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \boldsymbol{\nu})$ minimizes $\bar{D}(\mathbf{Y}||\mathbf{Y}_{\mathcal{M}})$ [8, Cor. 10.4]. As a consequence, the double minimization problem in (15) reduces to a single minimization of $\bar{D}(\mathbf{Y}||\mathbf{Y}_{\mathcal{M}})$ over all functions $g \in [\mathcal{X} \to \mathcal{Y}]$.

Since for this relative entropy rate between a stationary stochastic process and a Markov chain no closed-form expression exists (except in the case of $\mathbf{Y}$ being a Markov chain), we *lift* $\mathbf{Y}_{\mathcal{M}}$ to a Markov chain on the original state space $\mathcal{X}$:

Definition 5 ($\mathbf{P}$-lifting [5, Def. 7]) *Given a Markov chain* $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$, *a function* g, *and the best Markov approximation* $\mathbf{Y}_{\mathcal{M}} \sim \mathrm{Mar}(\mathcal{Y}, \mathbf{Q}, \boldsymbol{\nu})$ *of the projection* $\mathbf{Y}$ *in the sense of relative entropy rate, the* $\mathbf{P}$-lifting *of* $\mathbf{Y}_{\mathcal{M}}$ *is a Markov chain* $\hat{\mathbf{X}} \sim \mathrm{Mar}(\mathcal{X}, \hat{\mathbf{P}}, \hat{\boldsymbol{\mu}})$, *where*

$$\hat{P}_{i,j} := P_{i,j} \frac{p_{Y_2|Y_1}(g(j), g(i))}{p_{Y_2|X_1}(g(j), i)} \overset{(a)}{=} \frac{P_{i,j}}{\sum_{h \in g^{-1}(g(j))} P_{i,h}} Q_{g(i),g(j)} \tag{18}$$

and where $\hat{\boldsymbol{\mu}}$ *is the unique invariant distribution vector of* $\hat{\mathbf{X}}$.

In (18), equality in (a) follows from (5) and (17), where for simplicity, in Definition 5 we assumed that $\sum_{h \in g^{-1}(g(j))} P_{ih} > 0$ for all $i, j \in \mathcal{X}$. This restriction is not present in [5].

The following theorem summarizes the main properties of our lifting method.

Theorem 1 (Properties of P-lifting [5, Thm. 1]) *Given a Markov chain* $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbf{P}, \boldsymbol{\mu})$*, a function* g*, the best Markov approximation* $\mathbf{Y}_{\mathcal{M}}$ *of the projection* $\mathbf{Y}$ *in the sense of relative entropy rate, and the* $\mathbf{P}$*-lifting* $\hat{\mathbf{X}}$ *of* $\mathbf{Y}_{\mathcal{M}}$*, we have*

1. $\mathbf{P} \ll \hat{\mathbf{P}}$
2. $\hat{\mathbf{X}} = \underset{\text{Markov } \tilde{\mathbf{X}}:\mathbf{Y}_{\mathcal{M}} \text{ is } g\text{-projection of } \tilde{\mathbf{X}}}{\arg\min} \bar{D}(\mathbf{X}||\tilde{\mathbf{X}})$
3. $\bar{D}(\mathbf{Y}||\mathbf{Y}_{\mathcal{M}}) \leq \bar{D}(\mathbf{X}||\hat{\mathbf{X}}) = \sum_{i,j:\ P_{i,j}>0} \mu_i P_{i,j} \log \frac{P_{i,j}}{\hat{P}_{i,j}}.$

Property 2 implies that $\hat{\mathbf{X}}$ is lumpable w.r.t. g and that $\mathbf{Y}_{\mathcal{M}}$ is the projection of $\hat{\mathbf{X}}$ w.r.t. g. Property 3 is a consequence of Properties 1 and 2 and Lemma 2.

If we thus relax the M-partition problem of Definition 4 by using $\bar{D}(\mathbf{X}||\hat{\mathbf{X}})$ as a cost function, we obtain a closed-form expression for minimization. Interestingly, if we substitute $\hat{\mathbf{P}}$ from Definition 5 into Property 3, we get

$$\bar{D}(\mathbf{X}||\hat{\mathbf{X}}) = \sum_{i,j:\ P_{i,j}>0} \mu_i P_{i,j} \log \frac{p_{Y_2|X_1}(g(j), i)}{p_{Y_2|Y_1}(g(j), g(i))} \tag{19}$$

$$= \mathbb{E}\left(D(p_{Y_2|X_1}(\cdot|X_1)||p_{Y_2|Y_1}(\cdot|g(X_1)))\right) \tag{20}$$

i.e., the relative entropy involved in the condition for lumpability in Lemma 1. In other words, relaxing the problem by applying Theorem 1 converts our original M-partition problem to the problem of finding the M-partition w.r.t. which the original chain $\mathbf{X}$ is "most lumpable".

We can also expand $\bar{D}(\mathbf{X}||\hat{\mathbf{X}})$ in the following way:

$$\begin{aligned}
\bar{D}(\mathbf{X}||\hat{\mathbf{X}}) &= \sum_{i,j:\ P_{i,j}>0} \mu_i P_{i,j} \log \left(\frac{p_{Y_2,X_1}(g(j), i)}{p_{Y_2,Y_1}(g(j), g(i))} \frac{p_{Y_1}(g(i))}{p_{X_1}(i)} \frac{p_{Y_2}(g(j))}{p_{Y_2}(g(j))} \right) \\
&= \sum_{i,l:\ p_{Y_2,X_1}(l,i)>0} p_{Y_2,X_1}(l, i) \left(\log \frac{p_{Y_2,X_1}(l, i)}{p_{X_1}(i) p_{Y_2}(l)} - \log \frac{p_{Y_2,Y_1}(l, g(i))}{p_{Y_1}(g(i)) p_{Y_2}(l)} \right) \\
&= I(X_1; Y_2) - I(Y_1; Y_2)
\end{aligned}$$

which is the formulation of the IB method in (14) for $\beta = 1$, a relevance RV Y_2, and with the goal of compressing X_1 to Y_1. Unfortunately, the IB method cannot be applied directly, because the relevance RV Y_2 depends on the function g, which is the object of the optimization. Hence, we apply the chain rule of mutual information [1, Thm. 2.5.2] in (a) and the data processing inequality [1, p. 35] in (b) to further relax the problem:

$$\bar{D}(\mathbf{X}||\hat{\mathbf{X}}) = I(X_1; Y_2) - I(Y_1; Y_2) \tag{21}$$

$$\overset{(c)}{=} I(X_1, Y_1; Y_2) - I(Y_1; Y_2) \tag{22}$$

$$\overset{(a)}{=} I(X_1; Y_2|Y_1) \tag{23}$$

$$\overset{(b)}{\leq} I(X_1; X_2|Y_1) \tag{24}$$

$$\overset{(a)}{=} I(X_1, Y_1; X_2) - I(Y_1; X_2) \tag{25}$$

$$\overset{(c)}{=} I(X_1; X_2) - I(Y_1; X_2) \tag{26}$$

where (c) is because Y_1 is a function of X_1. Since the first term does not depend on g, the relaxed problem essentially tries to maximize $I(Y_1; X_2)$, the information the current sample Y_1 of the projection contains about the future sample X_2 of the original Markov chain. As a side note, by applying the data processing inequality again in the last line, the problem is further relaxed to minimizing $I(X_1; X_2) - I(Y_1; Y_2)$, which is considered in [2, 4, 7].

Since the first term in (26) does not depend on the function g, we want to solve

$$\arg\max_{g \in [\mathcal{X} \to \mathcal{Y}]} I(Y_1; X_2). \tag{27}$$

Comparing this with the cost function Tishby and Slonim were applying in [20] (cf. also Section 4.), one can see that this can be accomplished by the agglomerative IB method. While we had to relax the original M-partition problem from Definition 4, we are now at a point where it is tractable by employing a standard method from the machine learning literature.

6. Examples

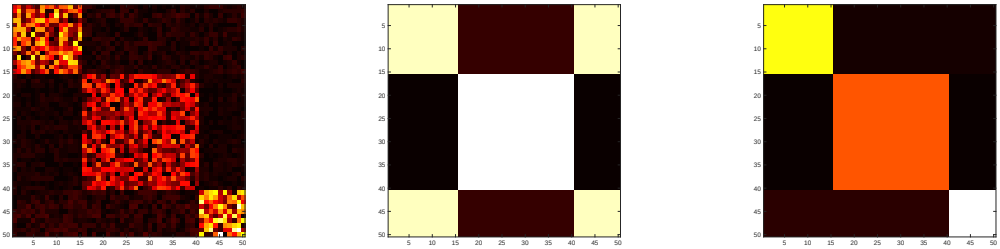
We illustrate our results by using a nearly completely decomposable Markov chain and a toy example from natural language processing. In both examples, the VLFeat Matlab implementation [22] of the agglomerative IB method was used.

6.1. A Nearly Completely Decomposable Markov Chain

In this first example we generated a Markov chain with 50 states, of which each 15, 25, and 10 are strongly interacting. In order to obtain an appropriate transition probability matrix $\mathbf{P}$, we generated a weighting matrix

$$\mathbf{W} = \begin{bmatrix} 0.9 \cdot \mathbf{1}_{15\times 15} & 0.1 \cdot \mathbf{1}_{15\times 25} & 0.1 \cdot \mathbf{1}_{15\times 10} \\ 0.1 \cdot \mathbf{1}_{25\times 15} & 0.9 \cdot \mathbf{1}_{25\times 25} & 0.1 \cdot \mathbf{1}_{25\times 10} \\ 0.1 \cdot \mathbf{1}_{10\times 15} & 0.1 \cdot \mathbf{1}_{10\times 25} & 0.9 \cdot \mathbf{1}_{10\times 10} \end{bmatrix} \tag{28}$$

where $\mathbf{1}_{n\times m}$ is a matrix full of ones with n rows and m columns. We multiplied this matrix element-wise with a matrix with random entries uniformly distributed on $[0, 1]$ and normalized the row sums of the resulting matrix to unity. Figure 1 shows both the original transition probability matrix and the obtained aggregations for $M = 2$ and $M = 3$. As it can be seen, the groups of strongly interacting states are identified correctly by our method.



(a) (b) (c)

Fig. 1. The nearly completely decomposable transition probability matrix (a) and the partitions obtained by using the agglomerative IB method. Blocks of the same color indicate that the corresponding states are mapped to the same output. Hot colors indicate high transition probabilities. For $M = 2$ in (b) it can be seen that the first and the third group of states (corresponding to states 1 to 15 and 41 to 50) are grouped together, while for $M = 3$ in (c) all three strongly interacting groups of states are identified correctly.

6.2. A Toy Example from Natural Language Processing

In this example we trained a letter bigram model[2] of an English translation of "Quo Vadis" by Henryk Sienkiewicz[3], i.e., we trained a Markov model of the co-occurrence of letters by determining the relative frequency of transitions between the letters of the text. We simplified the text beforehand by converting upper case to lower case letters, removing punctuation, and replacing all non-Latin characters by appropriate Latin ones (e.g., we replaced `'ç'` by `'c'` in "façade"). The state space of the Markov model thus consists of only the 26 letters of the Latin alphabet and the blank space.

We applied our state space aggregation method based on the agglomerative IB method for various cardinalities M of $\mathcal{Y}$. Most notably, for a bi-partition of the state space (i.e., for $M = 2$), all consonants are lumped together, leaving the five vowels and the blank space for the second group of states. For $M = 3$, the three groups are vowels, consonants, and the blank space, illustrating that the proposed method identifies clusters in accordance with human intuition based on the knowledge of English language. For aggregations to larger state spaces, individual letters crystallize as clusters on their own, e.g., the letters `'e'` and `'t'`. Analyzing the meaning of these aggregations in the light of language models is interesting, but has to be deferred to future work. Aggregations for various choices of M can be seen in Table 1.

[2]In one of his first introductions to Markov chains, Markov used letter bigram models: He trained – by hand – a model based on the first 20.000 letters of Pushkin's *Eugen Onegin*, see [10]. Roughly 35 years later, Shannon used bigram models of the English language in "A Mathematical Theory of Communication" [17].

[3]A copy of the text can be obtained from Project Gutenberg: http://www.gutenberg.org/ebooks/2853.

Tab. 1. Partitions of the letter state space obtained by agglomerative IB, shown for various values of M. Letters within brackets belong to the same element of the partition.

M	Partition
2	[,a,e,i,o,u],[b,c,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z]
3	[ ],[a,e,i,o,u],[b,c,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z]
4	[ ],[a,e,i,o,u],[b,c,h,j,k,l,m,n,p,q,r,v,w,z],[d,f,g,s,t,x,y]
7	[ ],[a,i,o,u],[e],[b,c,h,j,p,q,v,w,z],[d,f,g,s,x,y],[k,l,m,n,r],[t]
12	[ ],[a],[e],[i,u],[o],[b,j,p,q],[c,w],[d,f,g,s,x,y],[h,v,z],[k,l,m,r],[n],[t]

7. References

[1] Cover T.M., Thomas J.A., *Elements of Information Theory*, Wiley Interscience, Hoboken, NJ, 2nd edition, 2006.

[2] Deng K., Mehta P.G., Meyn S.P., *Optimal Kullback-Leibler aggregation via spectral theory of Markov chains*, IEEE Trans. Autom. Control 56 (12), Dec. 2011, pp. 2793–2808.

[3] Dhillon I., Mallela S., Modha D., *Information-theoretic co-clusterings,* Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Washington, D.C., Aug. 2003, pp. 89–98.

[4] Friedman A., Goldberger J., *Information theoretic pairwise clustering,* E. Hancock and M. Pelillo (Eds.), Proc. Similarity-Based Pattern Recognition, Springer Berlin LNCS 7953, 2013, pp. 106–119.

[5] Geiger B.C., Petrov T., Kubin G., Koeppl H., *Optimal Kullback-Leibler aggregation via information bottleneck,* Apr. 2013, Accepted for publication in IEEE Trans. Autom. Control; preprint available: arXiv:1304.6603 [cs.SY].

[6] Geiger B.C., Temmel C., *Lumpings of Markov chains, entropy rate preservation, and higher-order lumpability*, Dec. 2012. Accepted for publication in *J. Appl. Prob.*, preprint available: arXiv:1212.4375 [cs.IT].

[7] Goldberger J., Erez K., Abeles M., *A Markov clustering method for analyzing movement trajectories,* Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP), Thessaloniki, Aug. 2007, pp. 211–216.

[8] Gray R.M., *Entropy and Information Theory*, Springer, New York, NY, 1990.

[9] Gurvits L., Ledoux J., *Markov property for a function of a Markov chain: A linear algebra approach,* Linear Algebra Appl. 404, 2005, pp. 85–117.

[10] Hayes B., *First links in the Markov chain,* American Scientist 101, 2013, pp. 92–97.

[11] Katsoulakis M.A., Trashorras J., *Information loss in coarse-graining of stochastic particle dynamics,* J. Stat. Phys., 122 (1), 2006, pp. 115–135.

[12] Kemeny J.G., Snell J.L., *Finite Markov Chains*, Springer, 2nd edition, 1976.

[13] Manning C.D., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, 2nd edition, 2000.

[14] Petrov T., *Formal reductions of stochastic rule-based models of biochemical systems*, PhD thesis, ETH Zürich, 2013.

[15] Rached Z., Alajaji F., Campbell L.L., *The Kullback-Leibler divergence rate between Markov sources,* IEEE Trans. Inf. Theory 50 (5), May 2004, pp. 917–921.

[16] Raj A., Wiggins C.H., *An information-theoretic derivation of min-cut-based clustering,* IEEE Trans. Pattern Anal. Mach. Intell. 32 (6), June 2010, pp. 988–995.

[17] Shannon C.E., *A mathematical theory of communication,* Bell Systems Technical Journal 27, Oct. 1948, pp. 379–423, 623–656.

[18] Slonim N., Tishby N., *Agglomerative information bottleneck,* Advances in Neural Information Processing Systems (NIPS), Denver, CO, Nov. 1999, pp. 617–623.

[19] Tishby N., Pereira F.C., Bialek W., *The information bottleneck method,* Proc. Allerton Conf. on Communication, Control, and Computing, Monticello, IL, Sept. 1999, pp. 368–377.

[20] Tishby N., Slonim N., *Data clustering by Markovian relaxation and the information bottleneck method,* Advances in Neural Information Processing Systems (NIPS), Denver, CO, Nov. 2000.

[21] Tzortzis I., Charalambous C.D., Charalambous T., Hadjicostis C.N., Johansson M., *Approximation of Markov processes by lower dimensional processes via total variation metrics*, Oct. 2014, arXiv:1410.3976 [math.OC].

[22] Vedaldi A., Fulkerson B. *VLFeat: An open and portable library of computer vision algorithms,* 2008, http://www.vlfeat.org/.

[23] Vidyasagar M., *Reduced-order modeling of Markov and hidden Markov processes via aggregation,* Proc. IEEE Conf. on Decision and Control (CDC), Atlanta, GA, Dec. 2010, pp. 1810–1815.

[24] White L.B., Mahony R., Brushe G.D., *Lumpable hidden Markov models–model reduction and reduced complexity filtering,* IEEE Trans. Autom. Control 45 (12), Dec. 2000, pp. 2297–2306.