J. Appl. Prob. **51,** 1114–1132 (2014)
Printed in England

LUMPINGS OF MARKOV CHAINS, ENTROPY RATE PRESERVATION, AND HIGHER-ORDER LUMPABILITY

BERNHARD C. GEIGER,* *Graz University of Technology*
CHRISTOPH TEMMEL,** *Graz University of Technology and VU University Amsterdam*

Abstract

A lumping of a Markov chain is a coordinatewise projection of the chain. We characterise the entropy rate preservation of a lumping of an aperiodic and irreducible Markov chain on a finite state space by the random growth rate of the cardinality of the realisable preimage of a finite-length trajectory of the lumped chain and by the information needed to reconstruct original trajectories from their lumped images. Both are purely combinatorial criteria, depending only on the transition graph of the Markov chain and the lumping function. A lumping is strongly k-lumpable, if and only if the lumped process is a kth-order Markov chain for each starting distribution of the original Markov chain. We characterise strong k-lumpability via tightness of stationary entropic bounds. In the sparse setting, we give sufficient conditions on the lumping to both preserve the entropy rate and be strongly k-lumpable.

Keywords: Lumping; entropy rate loss; functional hidden Markov model; strong lumpability; higher-order Markov chain

2010 Mathematics Subject Classification: Primary 60J10
Secondary 60G17; 94A17; 60G10; 65C40

1. Introduction

The *entropy rate* of a stationary stochastic process is the average number of bits per time step needed to encode the process. A *lumping* of a (stationary) Markov chain is a coordinatewise projection of the chain by a *lumping function*. The resulting (stationary) *lumped stochastic process* is also called a *functional hidden Markov model* [8]. One can transform every hidden Markov model on finite state and observation spaces into this setting [8, Section IV.E]. In general, the lumped process loses the Markov property [13] and has a lower entropy rate than the original Markov chain due to the aggregation of states [21], [24].

Our first result characterises the structure of entropy rate preserving lumpings of stationary Markov chains over a finite state space. The *realisable preimage* is the set of finite paths in the transition graph associated with the Markov chain having the same image. The key property is the behaviour of the growth of this random set. It is also described by the ability of two such paths, once split, to join again. We document a strong dichotomy between the preservation and loss case: a uniform finite bound on the lost entropy and an almost surely finite growth

Received 17 January 2013; revision received 14 November 2013.
* Postal address: Institute for Communications Engineering, Technische Universiät München, Theresienstraße 90, 80333 Munich. Email address: geiger@ieee.org
** Postal address: Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. Email address: math@temmel.me

in the former, and a linearly growing entropy loss and an almost surely exponential growth in the latter.

In particular, a positive transition matrix always implies an entropy rate loss for a nonidentity lumping. We state a sufficient condition on a lumping of a Markov chain with a nonpositive transition matrix to preserve the entropy rate. Carlyle's representation [6] of a finite-state stationary stochastic process as a lumping of a Markov chain on an at most countable state space fulfills this condition.

Lumpings resulting in higher-order Markov chains are highly desirable from a simulation point of view. Our second result characterises such lumpings by the equality of natural entropic bounds with the entropy rate of the lumped process in the stationary setting. A first equality that holds only for entropies depending on the lumped process is equivalent to *weak lumpability*, i.e. the lumped process is a higher-order Markov chain in the stationary setting. A second equality involving entropies also using the underlying Markov chain in the stationary case is equivalent to *strong lumpability*, i.e. the lumped process is a higher-order Markov chain, *for every* initial distribution. Our characterisation is an information-theoretic complement to Gurvits and Ledoux's [13] linear algebraic approach to characterise lumpability.

We state a sufficient condition on the transition graph and the lumping function to preserve the entropy rate and be strongly k-lumpable. The condition is fulfilled on nontrivial lower-dimensional subspaces of the space of transition matrices. This complements Gurvits and Ledoux's [13] result that lumpings having higher-order Markov behaviour are nowhere dense.

2. Main results

2.1. Preliminaries

We let $\mathbb{N} := \{1, 2, \dots\}$ and $\mathbb{N}_0 := \{0, 1, 2, \dots\}$. We write $[n, m] := \{k \in \mathbb{N}_0 \colon n \le k \le m\}$ and abbreviate $[n] := [1, n]$. A vector $\boldsymbol{x}$ subscripted by a set A is the subvector of elements indexed by this set: $\boldsymbol{x}_A := (x_n)_{n \in A}$.

We recall information-theoretic basics from [7, Chapters 2 and 4]. Let ld denote the binary logarithm. By continuous extension, we assume that $0 \operatorname{ld} 0 = 0$. The *Shannon entropy* of a random variable Z taking values in a finite set $\mathcal{Z}$ is

$$H(Z) := -\sum_{z \in \mathcal{Z}} \mathbb{P}(Z = z) \operatorname{ld} \mathbb{P}(Z = z).$$

The *conditional entropy* of Z given W is defined by

$$H(Z \mid W) := \sum_{w \in \mathcal{W}} \mathbb{P}(W = w) H(Z \mid W = w).$$

Successive conditioning reduces entropy:

$$H(Z) \ge H(Z \mid W_1) \ge H(Z \mid W_1, W_2). \tag{1}$$

For a stationary stochastic process $Z := (Z_n)_{n \in \mathbb{N}_0}$ on a finite state space $\mathcal{Z}$, the *entropy rate* is

$$\bar{H}(Z) := \lim_{n \to \infty} \frac{1}{n} H(Z_{[n]}) = \lim_{n \to \infty} H(Z_n \mid Z_{[n-1]}). \tag{2}$$

The left-hand limit in (2) is the limit of the normalised *block entropy* $H(Z_{[n]})$. By stationarity and (1), the $H(Z_n \mid Z_{[n-1]})$ in the right-hand limit of (2) are monotonically decreasing.

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

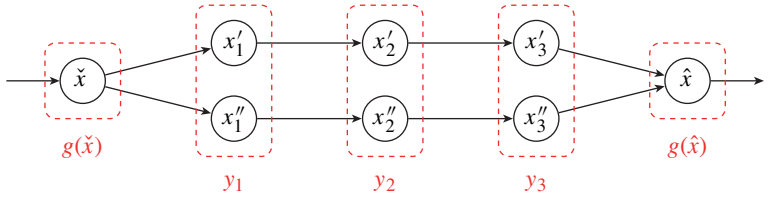

FIGURE 1: A section of trajectory space, with time running left to right. The two realisable length-5 trajectories $(\check{x}, x'_1, x'_2, x'_3, \hat{x})$ and $(\check{x}, x''_1, x''_2, x''_3, \hat{x})$ have the same lumped image $(g(\check{x}), y_1, y_2, y_3, g(\hat{x}))$. Thus, $\mathcal{K} \leq 3$. The lumped states $\{g(\check{x}), y_1, y_2, y_3, g(\hat{x})\}$ need not be distinct, e.g. it might be that $y_1 = y_2 = g(\hat{x})$. If $\mathcal{K} = 3$, then the minimality of $\mathcal{K}$ implies that $x'_i \neq x''_i$, for $i \in [3]$.

2.2. Setting

Let $X := (X_n)_{n\in\mathbb{N}_0}$ be an irreducible, aperiodic, time-homogeneous Markov chain on the finite state space $\mathcal{X}$. It has transition matrix P with invariant probability measure μ. We assume that X is stationary, that is $X_0 \sim \mu$. The *lumping function* g is $\mathcal{X} \to \mathcal{Y}$ and surjective. We assume g to be nontrivial, that is $2 \leq |\mathcal{Y}| < |\mathcal{X}|$. Without loss of generality, we extend g to $\mathcal{X}^n \to \mathcal{Y}^n$ coordinatewise for arbitrary $n \in \mathbb{N}$. The *lumped process* of X under g is the stationary stochastic process $Y := (Y_n)_{n\in\mathbb{N}_0}$ defined by $Y_n := g(X_n)$. We refer to this setup as the *lumping* (P, g).

The lumping induces a *conditional entropy rate* [9], [24], which characterises the average information loss per time step:

$$\bar{H}(X \mid Y) := \lim_{n\to\infty} \frac{1}{n} H(X_{[n]} \mid Y_{[n]}) = \bar{H}(X) - \bar{H}(Y).$$

Our main question is whether $\bar{H}(X \mid Y)$ is positive or zero, speaking of *entropy rate loss* or *entropy rate preservation* respectively. Entropy rate preservation does not imply that we can reconstruct the original process from the lumped process without entropy loss. See Figure 4 in Section 2.4 for an example.

The *transition graph* G of the Markov chain X is the directed graph with vertex set $\mathcal{X}$ and an edge (x, x'), if and only if $\mathbb{P}(X_1 = x' \mid X_0 = x) > 0$. A length-$n$ trajectory $\boldsymbol{x} \in \mathcal{X}^n$ is *realisable*, if and only if $\mathbb{P}(X_{[n]} = \boldsymbol{x}) > 0$, equivalent to being a directed path in G. A key structural property of G is its *split-merge index* with respect to g:

$$\begin{aligned}\mathcal{K} := \inf\{ n \in \mathbb{N} \mid \text{there exist } \check{x}, \hat{x} \in \mathcal{X},\ \boldsymbol{y} \in \mathcal{Y}^n,\ \text{and } \boldsymbol{x}',\ \boldsymbol{x}'' \in g^{-1}(\boldsymbol{y}),\ \boldsymbol{x}'' \neq \boldsymbol{x}',\\ \text{such that both } \mathbb{P}(X_0 = \check{x},\ X_{[n]} = \boldsymbol{x}',\ X_{n+1} = \hat{x}) > 0 \text{ and}\\ \mathbb{P}(X_0 = \check{x},\ X_{[n]} = \boldsymbol{x}'',\ X_{n+1} = \hat{x}) > 0\}. \end{aligned} \tag{3}$$

The split-merge index is the shortest length of the differing part of a pair of finite, different, and realisable trajectories with common start and end points, and the same lumped image if such a pair exists. Otherwise, let $\mathcal{K} = \infty$. If $\mathcal{K} < \infty$, every pair of sequences $\boldsymbol{x}', \boldsymbol{x}'' \in \mathcal{X}^{\mathcal{K}}$ fulfilling (3) is not only different, but differs in every coordinate by virtue of the infimum in (3). See Figure 1 for an example with $\mathcal{K} \leq 3$.

2.3. Characterisation of the entropy rate loss

In this section we present the characterisation of the entropy rate loss of a lumping in terms of $\mathcal{K}$ and the growth rate of the cardinality of the realisable preimage. The *realisable preimage*

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

of a lumped trajectory $\boldsymbol{y} \in \mathcal{Y}^n$ are the realisable trajectories in its preimage:

$$R(\boldsymbol{y}) := \{\boldsymbol{x} \in g^{-1}(\boldsymbol{y}) \colon \boldsymbol{x} \text{ is realisable}\}.$$

The *preimage count of length n* of the lumping (P, g) is the cardinality of the *realisable preimage* of a random lumped trajectory of length n:

$$T_n := |R(Y_{[n]})| = \sum_{\boldsymbol{x} \in g^{-1}(Y_{[n]})} [\mathbb{P}(X_{[n]} = \boldsymbol{x}) > 0].$$

Here, the right-hand side sums over Iverson brackets. Our first main result characterises entropy rate preservation.

Theorem 1. *It holds that*

$$\begin{aligned} \bar{H}(X \mid Y) > 0 \quad &\iff \quad \mathcal{K} < \infty \\ &\iff \quad \textit{there exists } C > 1 \textit{ such that } \mathbb{P}\Big(\liminf_{n\to\infty} \sqrt[n]{T_n} \geq C\Big) = 1, \qquad (4a) \\ \bar{H}(X \mid Y) = 0 \quad &\iff \quad \mathcal{K} = \infty \\ &\iff \quad \textit{there exists } C < \infty \textit{ such that } \mathbb{P}\Big(\sup_{n\to\infty} T_n \leq C\Big) = 1. \qquad (4b) \end{aligned}$$

The proofs of all statements in this section are given in Section 3. The constant C in Theorem 1 is an explicit function of (P, g); see (31) for (4a) and (16) for (4b). Likewise, an explicit lower bound for the entropy rate loss in case (4a) is stated in (28), implying that the entropy loss grows at least linearly in the sequence length.

Theorem 1 reveals a dichotomy in behaviour of the entropy of the lumping. If $\mathcal{K}$ is infinite, then no split-merge situations as in Figure 1 occur. Thus, all finite trajectories of X can be reconstructed from its lumped image and knowledge of its endpoints. Therefore, the only entropy loss occurs at those endpoints and is finite. This yields uniform finite bounds on the conditional block entropies and the preimage count. If $\mathcal{K}$ is finite, then at least two different realisable length-$(\mathcal{K}+2)$ trajectories of X with the same lumped image split and merge (see Figure 1). Such a split-merge leads to a finite entropy loss. The ergodic theorem ensures that this situation occurs linearly often in the block length, thus leading to a linear growth of the conditional block entropy. This implies an entropy rate loss. In particular, the conditional block entropy of a lumping never exhibits sublinear and unbounded growth.

If no split-merge situation occurs, then realisable trajectories with the same lumped image must be parallel. This constraint bounds their number. First, this yields a uniform bound on the conditional block entropies for lengths smaller than $\mathcal{K}$.

Proposition 1. *We have, for all n,*

$$n - 2 < \mathcal{K} \quad \Longrightarrow \quad H(X_{[n]} \mid Y_{[n]}) \leq 2 \operatorname{ld}(|\mathcal{X}| - |\mathcal{Y}| + 1). \qquad (5)$$

Second, the finiteness of $\mathcal{X}$ implies that either a split-merge situation of low trajectory length exists or no split-merge situation exists at all.

Proposition 2. *In case (4a), we have*

$$\mathcal{K} \leq \sum_{y \in \mathcal{Y}} |g^{-1}(y)|(|g^{-1}(y)| - 1).$$

If P is positive, i.e. all its entries are positive, then G is the complete directed graph and $\mathcal{K} = 1$. The following corollary is an immediate consequence.

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

Corollary 1. *If P is positive, then $\bar{H}(X \mid Y) > 0$.*

Thus, entropy rate preserving lumpings must have sufficiently sparse transition matrices P. The examples depicted in Figures 2 and 3 preserve the entropy rate without satisfying the sufficient conditions from Section 2.5.

2.4. Characterisation of strong k-lumpability

The case of the lumped process retaining the Markov property is desirable from a computational and modelling point of view. However, in general, the lumped process Y does not possess the Markov property [13], [17]. Nevertheless, one may hope that the lumped process belongs to the larger and still desirable class of higher-order Markov chains.

Definition 1. A stochastic process $Z := (Z_n)_{n\in\mathbb{N}_0}$ is a kth-order homogeneous Markov chain (denoted by Z is HMC(k)), if and only if, for all n $\in \mathbb{N}$, $m \in [k, n]$, $z_n \in \mathcal{Z}$, and $z \in \mathcal{Z}^m$,

$$\begin{aligned} &\mathbb{P}(Z_{[n-m,n-1]} = z) > 0 \\ \implies\quad &\mathbb{P}(Z_n = z_n \mid Z_{[n-m,n-1]} = z) = \mathbb{P}(Z_n = z_n \mid Z_{[n-k,n-1]} = z_{[n-k,n-1]}). \end{aligned} \tag{6}$$

The entropy rate of a HMC(k) is as straightforward as we would expect.

Proposition 3. *Let $Z := (Z_n)_{n\in\mathbb{N}_0}$ be a stationary stochastic process on $\mathcal{Z}$. Then*

$$Z \text{ is HMC}(k) \quad\iff\quad \bar{H}(Z) = H(Z_k \mid Z_{[0,k-1]}). \tag{7}$$

The proof of this proposition is given in Section 4. We investigate lumpings, where the lumped process is HMC(k).

Definition 2. (*Extension of [17, Definition 6.3.1].*) A lumping (P, g) of a stationary Markov chain is *weakly k-lumpable*, if and only if Y is HMC(k). It is *strongly k-lumpable*, if and only if this holds for each distribution of X_0 and the transition probabilities of Y are independent of this distribution.

A direct expression of the entropy rate of the lumped process Y is intrinsically complicated [3]. However, there are asymptotically tight, monotone decreasing, upper and lower bounds.

Lemma 1. ([7, Theorem 4.5.1, p. 86].) *In our setup, we have*

$$H(Y_n \mid Y_{[n-1]}, X_0) \le \bar{H}(Y) \le H(Y_n \mid Y_{[0,n-1]}) \quad \text{for all } n \in \mathbb{N}. \tag{8}$$

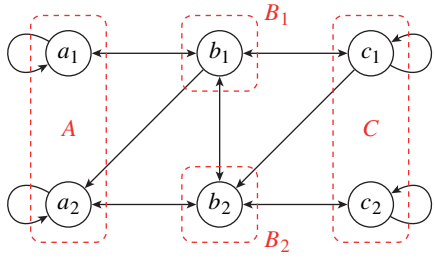

FIGURE 2: The transition graph of a Markov chain with the lumping represented by dashed boxes. The lumping preserves the entropy rate without satisfying the single-entry property of Section 2.5. The loops at a_1 and a_2 on the left-hand side, and at c_1 and c_2 on the right-hand side, prevent the lumped process from being a kth-order homogeneous Markov chain, for every k, given that the loop probabilities are different.

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

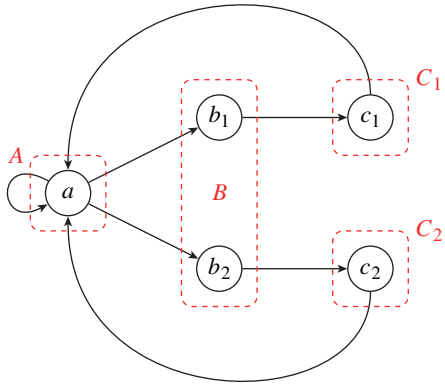

FIGURE 3: The transition graph of a Markov chain with the lumping represented by dashed boxes. The lumping is not single entry (violated by transitions from a into B). On the other hand, the existence of the uniquely represented states C_1 and C_2 allows us to distinguish between the trajectories (a, b_1, c_1, a) and (a, b_2, c_2, a). Therefore, the lumping preserves the entropy rate. Furthermore, this lumping is weakly 1-lumpable and strongly 2-lumpable, but not strongly 1-lumpable. Hence, it shows that the single-entry property is neither necessary for entropy rate preservation nor for weak k-lumpability. This also applies to the single forward k-sequence, a subclass of single entry.

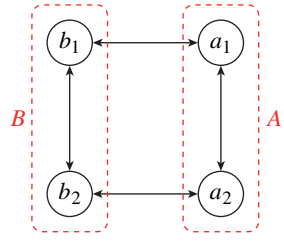

FIGURE 4: The transition graph of a Markov chain with the lumping represented by dashed boxes. The lumping is single entry and, thus, preserves the entropy rate. Furthermore, if all transitions have probability $\frac{1}{2}$, it is strongly 1-lumpable and, thus, $H(Y_1 \mid X_0) = H(Y_1 \mid Y_0)$ (see Theorem 2). However, observing an arbitrarily long trajectory of the lumped process does not determine the current preimage state. Whence, (P, g) does not satisfy the single forward k-sequence property for every k. Therefore, the single forward k-sequence property is neither necessary for entropy rate preservation nor for strong lumpability.

In the stationary setting, the equality on the right-hand side of (8) for $n = k$, together with Proposition 3, implies that Y is HMC(k), i.e. (P, g) is weakly k-lumpable. If there is also equality on the left-hand side of (8) for $n = k$, then knowledge of the distribution of X_0 delivers no additional information about Y_k. In other words, Y is HMC(k) for every starting distribution. Our second main result characterises higher-order lumpability.

Theorem 2. *The following statements are equivalent:*

$$H(Y_k \mid Y_{[k-1]}, X_0) = H(Y_k \mid Y_{[0,k-1]}), \tag{9a}$$

$$X \text{ is strongly } k\text{-lumpable.} \tag{9b}$$

The proof of Theorem 2 is given in Section 4. We stress the fact that (9a) is a condition only on the stationary setting, whereas (9b) deals with all starting distributions. Theorem 2 is an information-theoretic equivalent to Gurvits and Ledoux's characterisation [13, Theorems 2

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

and 6] of k-lumpability via a linear algebraic description of invariant subspaces. A classic example [17, p. 139] shows that weak k-lumpability alone is not sufficient for (9). Moreover, the examples presented in Figures 3 and 4, and Example 1 are strongly lumpable for some k without satisfying the sufficient condition from Section 2.5.

2.5. Sufficient conditions

We present easy to check sufficient conditions for the preservation of the entropy rate and strong k-lumpability. Their proofs are in Section 5. The conditions depend only on the transition graph G and the lumping function g.

Our first sufficient condition preserves the entropy rate.

Definition 3. A lumping (P, g) is *single entry* (SE), if and only if, for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, there exists $x' \in g^{-1}(y)$ such that

$$\mathbb{P}(X_1 = x'' \mid X_0 = x) = 0 \quad \text{for all } x'' \in g^{-1}(y) \setminus \{x'\}, \tag{10}$$

i.e. there is *at most one* edge from a given state x into the preimage $g^{-1}(y)$.

The SE lumpings are entropy rate preserving.

Proposition 4. *If (P, g) is SE, then $\bar{H}(X \mid Y) = 0$.*

Figures 2 and 3 show that SE is not necessary for entropy rate preservation.

Corollary 2. *If (P, g) is SE and weakly k-lumpable then it is strongly k-lumpable.*

Proof. The proof of Proposition 4 shows that SE implies equality on the left-hand side of (8) for all n. Weak k-lumpability implies equality on the right-hand side of (8) for $n = k$. Therefore, Theorem 2 applies.

An example of a lumping satisfying the conditions of the corollary is given in Figure 4. That a lumping can be SE without being strongly lumpable, or strongly lumpable without being SE is shown in Figure 5 and in Example 1, respectively.

Our second sufficient condition preserves the entropy rate and guarantees higher-order lumpability.

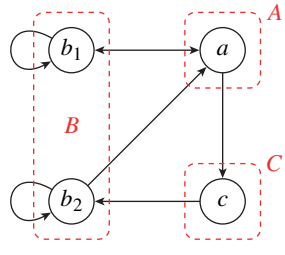

FIGURE 5: The transition graph of a Markov chain with the lumping represented by dashed boxes. The lumping is SE. The loops at b_1 and b_2 imply that the lumped process is not HMC(k) for every k, regardless of the distribution of X_0. This is easily seen by the inability to differentiate between n consecutive b_1 and n consecutive b_2. When starting in B and as long as $\mathbb{P}(X_1 = a \mid X_0 = b_1) \neq \mathbb{P}(X_1 = a \mid X_0 = b_2)$ and $\mathbb{P}(X_1 = b_1 \mid X_0 = b_1) \neq \mathbb{P}(X_1 = b_2 \mid X_0 = b_2)$, this long sequence of Bs prevents determining the probability of entering A. Thus, it is neither SFS(k) nor strongly k-lumpable for each k.

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

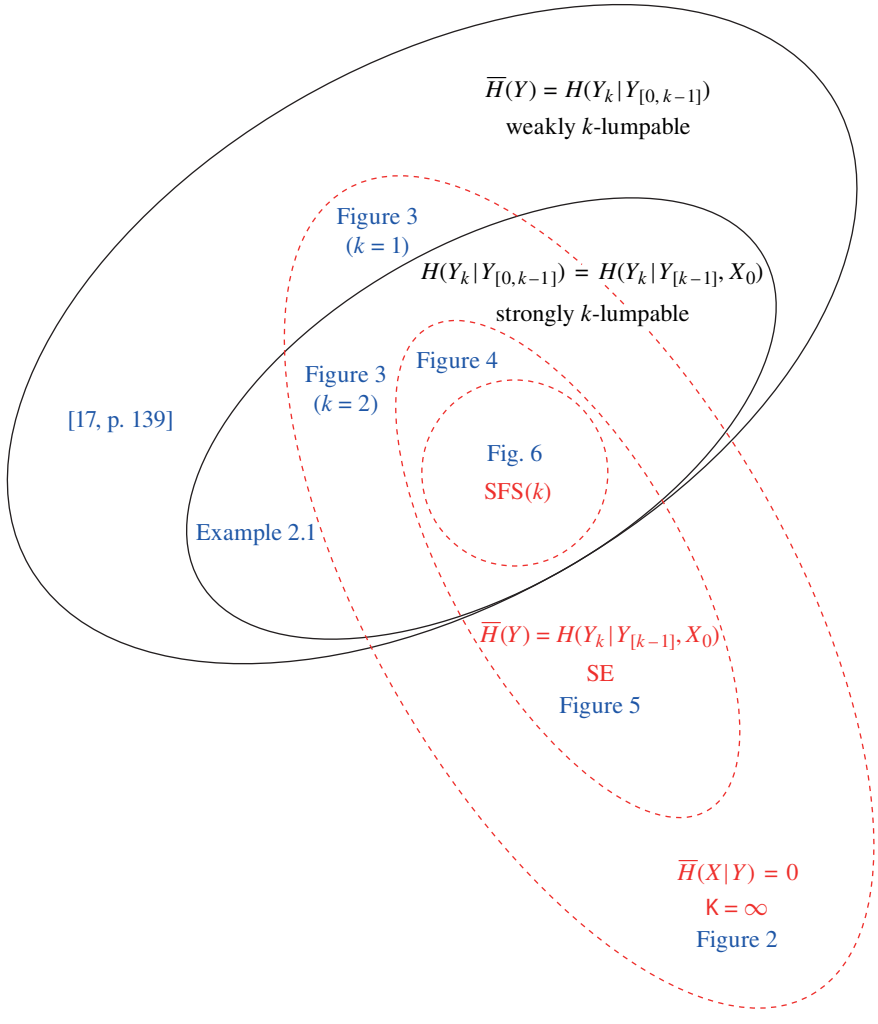

FIGURE 6: Venn diagram of the relation between different classes and location of counterexamples in this paper.

Definition 4. For $k \geq 2$, a lumping (P, g) has the *single forward k-sequence property* (denoted by SFS(k)), if and only if, for all $\boldsymbol{y} \in \mathcal{Y}^{k-1}$ and $y \in \mathcal{Y}$, there exists $\boldsymbol{x}' \in g^{-1}(\boldsymbol{y})$ such that

$$\mathbb{P}(X_{[k-1]} = \boldsymbol{x} \mid Y_{[k-1]} = \boldsymbol{y}, X_0 = x) = 0, \quad \text{for all } x \in g^{-1}(y),\ \boldsymbol{x} \in g^{-1}(\boldsymbol{y}) \setminus \{\boldsymbol{x}'\}, \tag{11}$$

i.e. there is *at most one* realisable sequence in the preimage $g^{-1}(\boldsymbol{y})$ starting in y.

The SFS(k) property implies entropy rate preservation and strong k-lumpability.

Proposition 5. *If (P, g) is SFS(k), then it is strongly k-lumpable and SE.*

Figure 6 gives an overview of the various classes and examples, in particular showing that the sufficient conditions are not necessary.

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

Example 1. Consider the following transition matrix, where the lines divide lumped states:

$$P := \left[\begin{array}{cc|cc} 0. & 0.4 & 0 & 0 \\ 0.3 & 0.2 & 0.1 & 0.4 \\ \hline 0.2 & 0.05 & 0.375 & 0.375 \\ 0.2 & 0.05 & 0.375 & 0.375 \end{array}\right].$$

This lumping is strongly 2-lumpable and satisfies (9a) with

$$\bar{H}(Y) = H(Y_2 \mid Y_{[0,1]}) = H(Y_2 \mid Y_1, X_0) = 0.733$$

(with an accuracy of 0.001). However, it does not preserve entropy: $1.480 = \bar{H}(X) > \bar{H}(Y)$, whence it is neither SE nor SFS(2).

2.6. Further discussion

The study of functions of Markov chains has a long tradition, in particular, whether a function of a Markov chain possesses the Markov property or not [5], [22]. Kemeny and Snell [17] coined the term *lumpability* for retaining the Markov property. Gurvits and Ledoux [13] analysed higher-order lumpability, as we use in this work. They showed that the class of Markov chains being lumpable is nowhere dense.

A related problem is the *identification problem*, initially posed by Blackwell and Koopmanns [4]: given a stationary process on a finite state space, is it representable by a lumping of a Markov chain? The question of existence of a finite state space representation has a long tradition [1], [11], [15], without a definite algorithmic solution. Two results from research into this topic have a connection to the present work. Firstly, Carlyle [6] showed that every stationary stochastic process on a finite state space is representable as a lumping of a Markov chain on an at most countable state space. The representation is SE. If it involves a Markov chain on a finite state space then Proposition 4 guarantees entropy rate preservation of the representation.

Secondly, Gilbert [11] showed that the distribution of a lumping of a finite-state Markov chain is uniquely determined by the distribution of m consecutive samples, where m depends on the cardinalities of the input and output alphabet. This does not contradict the nowhere

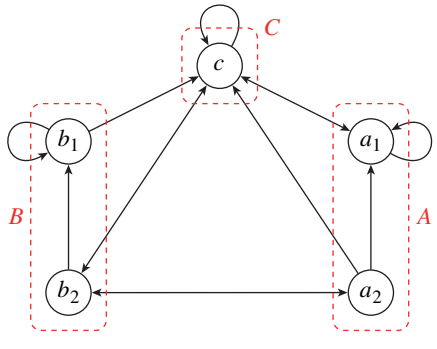

FIGURE 7: The transition graph of a Markov chain with the lumping represented by dashed boxes. After at most two steps one either enters a new lumped state at a unique original state or is circling in either b_1 or a_1. Hence, this lumping is SFS(2) and not strongly 1-lumpable. The space of Markov chains with this transition graph contains at least the interior of a multi-simplex in $\mathbb{R}^{13}$, parametrised by eigth parameters (13 directed edges minus five nodes).

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

dense result of Gurvits and Ledoux [13], since the construction of the process distribution is different from a product of conditional distributions (as it is in the case of lumpability).

Moreover, the nowhere dense property does not prevent our results from being practically relevant. In particular, our sufficient condition holds for nontrivial lower-dimensional subspaces of the space of Markov transition matrices; see Figure 7. In other words, if the transition matrix is sufficiently sparse, one can hope that the lumping satisfies some of our sufficient conditions. More generally, one can hope that for a given Markov model there exists a lumping function with a desired output alphabet size such that the resulting lumping satisfies our sufficient conditions. Sparse transition matrices appear, e.g. in n-gram models in automatic speech recognition [2, Table 1], chemical reaction networks [14], [16], [25], and link prediction and path analysis [23]. That the sufficient conditions for entropy preservation and weak k-lumpability are not overly restrictive was recently shown for a letter bi-gram model [10]: the bi-gram model exhibited the SFS(2)-property and, thus, permitted lossless compression.

In the nonstationary case, i.e. with X_0 having a different distribution than the invariant one, we are still *stationary in the asymptotic mean* [12], [18]. In particular, we have entropy rates and an ergodic theorem. Hence, all statements of this paper should generalise to this setting. Whether we can drop the restriction to aperiodic and irreducible chains is a more difficult question.

We give crude upper bounds on the algorithmic complexity of checking the properties introduced in the present paper. By Proposition 2, determining the finiteness and value of $\mathcal{K}$ takes at most $\mathcal{O}(\exp((1+|\mathcal{X}|^2)\log|\mathcal{Y}|))$ steps. We can check the SE property in $\mathcal{O}(|\mathcal{X}|^2)$ steps and the SFS(k) property in $\mathcal{O}(|\mathcal{X}|^k)$ steps. Finally, the verification of strong k-lumpability via (9a) requires $\mathcal{O}(|\mathcal{X}|^{k+1})$ steps. The last bound is of a similar order as Gurvits and Ledoux's algorithm for weak k-lumpability [13, Section 2.2.2].

There is another notion of information loss through lumping: Lindqvist [19] discussed sufficient statistics for estimating X_0 from Y_n. Gurvits and Ledoux introduced g-observability [13, Section 3] for determining X_0 from $Y_{[0,n]}$. Simple examples show that entropy rate preservation is independent of g-observability.

3. Proof of entropy rate preservation

Proof of Theorem 1. Statement (4) follows from the mutually exhaustive implications

$$\mathcal{K} < \infty \quad \Longrightarrow \quad \bar{H}(X \mid Y) > 0, \tag{12a}$$

$$\mathcal{K} = \infty \quad \Longrightarrow \quad \bar{H}(X \mid Y) = 0, \tag{12b}$$

and

$$\mathcal{K} < \infty \quad \Longrightarrow \quad \text{there exists } C > 1 \text{ such that } \mathbb{P}\Big(\liminf_{n\to\infty} \sqrt[n]{T_n} \geq C\Big) = 1, \tag{13a}$$

$$\mathcal{K} = \infty \quad \Longrightarrow \quad \text{there exists } C < \infty \text{ such that } \mathbb{P}\Big(\sup_{n\to\infty} T_n \leq C\Big) = 1. \tag{13b}$$

The proofs of implications (12b), (13b), and Proposition 1 are given in Section 3.1 and the proofs of implications (12a), (13a), and Proposition 2 are given in Section 3.4. Sections 3.2 and 3.3 contain technical results about Markov chains needed in the proof of the loss case in Section 3.4.

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

3.1. The preservation case

The definition of $\mathcal{K}$ in (3) implies that lumped trajectories of length less than $\mathcal{K}$ have a unique preimage contingent on the endpoints, i.e. if $n < \mathcal{K}$, then, for all $\check{x}, \hat{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}^n$,

$$\begin{aligned} & \mathbb{P}(X_0 = \check{x},\ Y_{[n]} = \boldsymbol{y},\ X_{n+1} = \hat{x}) > 0 \\ \Longrightarrow\quad & \text{there exists exactly one } \boldsymbol{x} \in \mathcal{X}^n \text{ such that} \\ & \mathbb{P}(X_{[n]} = \boldsymbol{x} \mid X_0 = \check{x},\ Y_{[n]} = \boldsymbol{y},\ X_{n+1} = \hat{x}) = 1. \end{aligned} \tag{14}$$

Proof of Proposition 1. We assume that $n - 2 < \mathcal{K}$. The unique preimage (14) implies that the conditional entropy of the interior of a block, given its lumped image and the states at its ends, is 0:

$$\begin{aligned} & H(X_{[2,n-1]} \mid X_1, X_n, Y_{[n]}) \\ & \quad = \sum_{\substack{\boldsymbol{y} \in \mathcal{Y}^n \\ \check{x}, \hat{x} \in \mathcal{X}}} \mathbb{P}(X_1 = \check{x}, X_n = \hat{x}, Y_{[n]} = \boldsymbol{y}) \underbrace{H(X_{[2,n-1]} \mid X_1 = \check{x}, X_n = \hat{x}, Y_{[n]} = \boldsymbol{y})}_{=0 \text{ by (14)}} \\ & \quad = 0. \end{aligned} \tag{15}$$

We apply the chain rule of entropy (compare [7, p. 22]) to decompose the conditional block entropy into its interior and its boundary. The interior vanishes by (15) and the entropy at the endpoints is maximal for the uniform distribution:

$$\begin{aligned} H(X_{[n]} \mid Y_{[n]}) &= H(X_{[2,n-1]} \mid X_1, X_n, Y_{[n]}) + H(X_1, X_n \mid Y_{[n]}) \\ &\le 0 + H(X_1, X_n \mid Y_1, Y_n) \\ &\le H(X_1 \mid Y_1) + H(X_n \mid Y_n) \\ &\le 2 \max\{\operatorname{ld} |g^{-1}(y)| \colon y \in \mathcal{Y}\} \\ &\le 2 \operatorname{ld}(|\mathcal{X}| - |\mathcal{Y}| + 1). \end{aligned}$$

Proof of (12b). As $\mathcal{K} = \infty$, the bound from (5) holds uniformly. Thus,

$$\bar{H}(X \mid Y) = \lim_{n \to \infty} \frac{1}{n} H(X_{[n]} \mid Y_{[n]}) \le \lim_{n \to \infty} \frac{2 \operatorname{ld}(|\mathcal{X}| - |\mathcal{Y}| + 1)}{n} = 0.$$

Proof of (13b). Recall, we assume that $\mathcal{K} = \infty$. We show, for all $\boldsymbol{y} \in \mathcal{Y}^n$ with $\mathbb{P}(Y_{[n]} = \boldsymbol{y}) > 0$, we have

$$\mathbb{P}(T_n \le (|\mathcal{X}| - |\mathcal{Y}| + 1)^2 \mid Y_{[n]} = \boldsymbol{y}) = 1. \tag{16}$$

This implies (13b). To show (16), we use (14) to bound

$$\begin{aligned} & \sum_{\boldsymbol{x} \in g^{-1}(\boldsymbol{y})} [\mathbb{P}(X_{[n]} = \boldsymbol{x}) > 0] \\ & \quad = \sum_{x_1, x_n \in g^{-1}(\boldsymbol{y}_{\{1,n\}})} [\mathbb{P}(X_1 = x_1, X_n = x_n \mid Y_{[n]} = \boldsymbol{y}) > 0] \\ & \qquad \times \sum_{\boldsymbol{x} \in g^{-1}(\boldsymbol{y}_{[2,n-1]})} [\mathbb{P}(X_{[2,n-1]} = \boldsymbol{x} \mid X_1 = x_1, X_n = x_n, Y_{[2,n-1]} = \boldsymbol{y}_{[2,n-1]}) > 0] \end{aligned}$$

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

$$\begin{aligned}
&\le \sum_{x_1, x_n \in g^{-1}(\boldsymbol{y}_{\{1,n\}})} [\mathbb{P}(X_1 = x_1, X_n = x_n \mid Y_{[n]} = \boldsymbol{y}) > 0] \\
&\le |g^{-1}(\boldsymbol{y}_{\{1,n\}})| \\
&\le (|\mathcal{X}| - |\mathcal{Y}| + 1)^2.
\end{aligned}$$

3.2. Nonoverlapping traversal instants

The main result of this section is an almost-sure *linear lower growth bound* for nonoverlapping occurrences of a fixed, finite pattern in a realisation in Proposition 6.

Let $Z := (Z_n)_{n\in\mathbb{N}}$ be a stationary stochastic process taking values in $\mathcal{Z}$. The *occupation instants* of a state z is the set of indices

$$\mathcal{O}_Z^z(n) := \{i \in [n] \colon Z_i = z\}.$$

The classic *occupation time* [20, Section 6.4] is the cardinality of the occupation instants. The *traversal instants* of a sequence $\boldsymbol{z} \in \mathcal{Z}^k$ is the set of indices

$$\mathcal{T}_Z^{\boldsymbol{z}}(n) := \{i \in [n-k+1] \colon Z_{[i,i+k-1]} = \boldsymbol{z}\}.$$

The *nonoverlapping traversal instants* of a sequence $\boldsymbol{z} \in \mathcal{Z}^k$ is the set of indices

$$\mathcal{N}_Z^{\boldsymbol{z}}(n) := \{i \in [n-k+1] \colon Z_{[i,i+k-1]} = \boldsymbol{z} \text{ for all } j \in [i+1, i+k-1] \colon Z_{[j,j+k-1]} \neq \boldsymbol{z}\}$$

where we select lower indices greedily.

For $k \in \mathbb{N}$, the *k-transition process* $Z^{(k)}$ of Z is the stochastic process on $\mathcal{Z}^k$ with marginals

$$\mathbb{P}(Z^{(k)}_{[n]} = (\boldsymbol{z}^i)_{i=1}^n) = \mathbb{P}(Z_{[n-1]} = (\boldsymbol{z}^i_{\{1\}})_{i=1}^{n-1}, Z_{[n,n+k-1]} = \boldsymbol{z}^n)$$

if, for all $i \in [n-1]$, $\boldsymbol{z}^i_{[2,k]} = \boldsymbol{z}^{i+1}_{[k-1]}$, and 0 otherwise. Obvious relations are

$$\mathcal{T}_Z^{\boldsymbol{z}}(n) = \mathcal{O}^{\boldsymbol{z}}_{Z^{(k)}}(n-k) \tag{17a}$$

and

$$\mathcal{N}_Z^{\boldsymbol{z}}(n) \subseteq \mathcal{T}_Z^{\boldsymbol{z}}(n) \quad \text{with} \quad |\mathcal{N}_Z^{\boldsymbol{z}}(n)| \ge \frac{1}{k}|\mathcal{T}_Z^{\boldsymbol{z}}(n)|. \tag{17b}$$

Proposition 6. *Let $\boldsymbol{s} \in \mathcal{X}^k$ be realisable with $p := \mathbb{P}(X_{[k]} = \boldsymbol{s} \mid X_1 = \boldsymbol{s}_{\{1\}}) > 0$. Then*

$$\mathbb{P}\left(\liminf_{n\to\infty} \frac{1}{n}|\mathcal{N}_X^{\boldsymbol{s}}(n)| \ge \frac{p\mu(\boldsymbol{s}_{\{1\}})}{k}\right) = 1 \tag{18a}$$

and

$$\lim_{n\to\infty} \mathbb{P}\left(|\mathcal{N}_X^{\boldsymbol{s}}(n)| \ge \left(\frac{p\mu(\boldsymbol{s}_{\{1\}})}{k} - \varepsilon\right)n\right) = 1 \quad \textit{for all } \varepsilon > 0. \tag{18b}$$

Lemma 2. (Ergodic theorem [26, Theorem 3.55, p. 69].) *For every homogeneous, irreducible, and aperiodic Markov chain $Z := (Z_n)_{n\in\mathbb{N}}$ on a finite state space $\mathcal{Z}$ with invariant measure ν, all $f : \mathcal{Z} \to \mathbb{R}$, and each starting distribution $\alpha \in \mathcal{M}_1(\mathcal{Z})$ of Z_1, we have*

$$\mathbb{P}_\alpha\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n f(Z_i) = \int_{\mathcal{Z}} f(z)\,\mathrm{d}\nu(z) =: \nu(f)\right) = 1.$$

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

Proof of Proposition 6. Statement (18b) is a direct consequence of (18a).

The k-transition process $X^{(k)}$ of X is a homogeneous Markov chain with transition probabilities

$$\mathbb{P}(X_2^{(k)} = \boldsymbol{x}' \mid X_1^{(k)} = \boldsymbol{x}) = \begin{cases} \mathbb{P}(X_{k+1} = \boldsymbol{x}'_{\{k\}} \mid X_k = \boldsymbol{x}_{\{k\}}) & \text{if } \boldsymbol{x}_{[2,k]} = \boldsymbol{x}'_{[k-1]}, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, as X is irreducible and aperiodic then so is $X^{(k)}$. Its invariant measure $\mu^{(k)}$ fulfils $\mu^{(k)}(\boldsymbol{x}) = \mu(\boldsymbol{x}_{\{1\}}) \prod_{i=1}^{k-1} \mathbb{P}(X_2 = \boldsymbol{x}_{\{i+1\}} \mid X_1 = \boldsymbol{x}_{\{i\}})$.

Let f be the indicator function of $\boldsymbol{s}$. We use (17a), (17b), and Lemma 2 to obtain

$$\begin{aligned} \mathbb{P}\left(\liminf_{n\to\infty} \frac{1}{n} |\mathcal{N}_X^{\boldsymbol{s}}(n)| \geq \frac{\mu^{(k)}(f)}{k} \right) &\geq \mathbb{P}\left(\liminf_{n\to\infty} \frac{1}{n} |\mathcal{T}_X^{\boldsymbol{s}}(n)| \geq \mu^{(k)}(f) \right) \\ &= \mathbb{P}\left(\liminf_{n\to\infty} \frac{1}{n} |\mathcal{O}_{X^{(k)}}^{\boldsymbol{s}}(n-k)| \geq \mu^{(k)}(f) \right) \\ &= \mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n} |\mathcal{O}_{X^{(k)}}^{\boldsymbol{s}}(n)| \geq \mu^{(k)}(f) \right) \\ &= 1. \end{aligned}$$

Finally, $\mu^{(k)}(f) = \mu^{(k)}(\boldsymbol{s}) = p\mu(\boldsymbol{s}_{\{1\}})$.

3.3. Conditional Markov property

In this section we present two technical statements about discrete Markov processes. Let $X := (X_n)_{n\in\mathbb{N}_0}$ be a stochastic process on the cartesian product $\mathcal{A} := \prod_{n\in\mathbb{N}_0} \mathcal{A}_n$ of the finite sets $(\mathcal{A}_n)_{n\in\mathbb{N}_0}$. For $A \subseteq \mathbb{N}_0$, let $\mathcal{A}_A := \prod_{n\in A} \mathcal{A}_n$. In the remainder of this section, we assume that all conditional probabilities are well defined. The process X is Markov, if and only if, for all $n, m \in \mathbb{N}$, $m \leq n$, $s_n \in \mathcal{A}_n$, and $\boldsymbol{s}_{[n-m,n-1]} \in \mathcal{A}_{[n-m,n-1]}$:

$$\mathbb{P}(X_n = s_n \mid X_{[n-m,n-1]} = \boldsymbol{s}_{[n-m,n-1]}) = \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1}). \tag{19}$$

We denote by $A \Subset \mathbb{N}_0$ the fact that A is a *finite subset* of $\mathbb{N}_0$. The first statement is a factorisation of conditional probabilities over disjoint index blocks: for all $\varnothing \neq B_0, A_1, B_1, \ldots, B_{m-1}, A_m, B_m \Subset \mathbb{N}_0$, $A \cap B = \varnothing$, where $A := \biguplus_{i=1}^m A_i$ and $B := \biguplus_{i=0}^m B_i$, $\boldsymbol{x}_A \in \mathcal{A}_A$, and $\boldsymbol{x}_B \in \mathcal{A}_B$ (for all $i \in [m]$, $b_i^- := \min(B_{i-1}) < \min(A_i)$ and $\max(A_i) < \min(B_i) =: b_i^+$):

$$\mathbb{P}(X_A = \boldsymbol{x}_A \mid X_B = \boldsymbol{x}_B) = \prod_{i=1}^m \mathbb{P}(X_{A_i} = \boldsymbol{x}_{A_i} \mid X_{b_i^-} = x_{b_i^-}, X_{b_i^+} = x_{b_i^+}). \tag{20}$$

Secondly, a Markov process retains the Markov property under a *cartesian conditioning*: for all $\varnothing \neq C \Subset \mathbb{N}_0$, and where $S_C := \prod_{n\in C} S_n$ with $S_n \subseteq \mathcal{A}_n$:

$$(X \mid X_C \in S_C) \text{ is Markov.} \tag{21}$$

To prove this, we need the following intermediate statements. For all $n \in \mathbb{N}$, $\varnothing \neq B \subseteq [0, n-1]$, $x_n \in \mathcal{A}_n$, and $\boldsymbol{x}_B \in \mathcal{A}_B$:

$$\mathbb{P}(X_n = x_n \mid X_B = \boldsymbol{x}_B) = \mathbb{P}(X_n = x_n \mid X_{\max(B)} = x_{\max(B)}). \tag{22}$$

For all $\varnothing \neq A, B \Subset \mathbb{N}_0$, $b := \max(B) < \min(A)$, $\boldsymbol{x}_A \in \mathcal{A}_A$, and $S_B \subseteq \{x_b\} \times \mathcal{A}_{B\setminus\{b\}}$:

$$\mathbb{P}(X_A = \boldsymbol{x}_A \mid X_B \in S_B) = \mathbb{P}(X_A = \boldsymbol{x}_A \mid X_b = x_b). \tag{23}$$

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

Proof of (22). Let $b := \max(B)$, $C := [b+1, n-1]$, and $D := [\min(B), b] \setminus B$. We use (19) to obtain

$$\begin{aligned}
&\mathbb{P}(X_n = x_n \mid X_B = \boldsymbol{x}_B) \\
&\quad = \frac{\sum_{\boldsymbol{x}_C, \boldsymbol{x}_D} \mathbb{P}(X_n = x_n, X_C = \boldsymbol{x}_C, X_B = \boldsymbol{x}_B, X_D = \boldsymbol{x}_D)}{\mathbb{P}(X_B = \boldsymbol{x}_B)} \\
&\quad = \frac{\sum_{\boldsymbol{x}_C, \boldsymbol{x}_D} \mathbb{P}(X_n = x_n, X_C = \boldsymbol{x}_C \mid X_B = \boldsymbol{x}_B, X_D = \boldsymbol{x}_D)\mathbb{P}(X_B = \boldsymbol{x}_B, X_D = \boldsymbol{x}_D)}{\mathbb{P}(X_B = \boldsymbol{x}_B)} \\
&\quad = \frac{\sum_{\boldsymbol{x}_C} \mathbb{P}(X_n = x_n, X_C = \boldsymbol{x}_C \mid X_b = x_b) \sum_{\boldsymbol{x}_D} \mathbb{P}(X_B = \boldsymbol{x}_B, X_D = \boldsymbol{x}_D)}{\mathbb{P}(X_B = \boldsymbol{x}_B)} \\
&\quad = \mathbb{P}(X_n = x_n \mid X_b = x_b).
\end{aligned}$$

Proof of (20). For $A \Subset \mathbb{N}_0$, we abbreviate the event $E_A := [X_A = \boldsymbol{x}_A]$. Apply (22) in order to obtain

$$\begin{aligned}
&\mathbb{P}(X_A = \boldsymbol{x}_A \mid X_B = \boldsymbol{x}_B) \\
&\quad = \frac{\mathbb{P}(X_A = \boldsymbol{x}_A, X_B = \boldsymbol{x}_B)}{\mathbb{P}(X_B = \boldsymbol{x}_B)} \\
&\quad = \frac{\mathbb{P}(E_{B_0}) \prod_{i=1}^m \mathbb{P}(E_{B_i \setminus \{b_i^+\}} \mid (E_{A_j})_{j \le i}, (E_{B_j})_{j<i}, E_{b_i^+}) \mathbb{P}(E_{b_i^+}, E_{A_i} \mid (E_{A_j})_{j<i}, (E_{B_j})_{j<i})}{\mathbb{P}(E_{B_0}) \prod_{i=1}^m \mathbb{P}(E_{B_i \setminus \{b_i^+\}} \mid (E_{B_j})_{j<i}, E_{b_i^+}) \mathbb{P}(E_{b_i^+} \mid (E_{B_j})_{j<i})} \\
&\quad = \prod_{i=1}^m \frac{\mathbb{P}(E_{B_i \setminus \{b_i^+\}} \mid E_{b_i^+}) \mathbb{P}(E_{b_i^+}, E_{A_i} \mid E_{b_i^-})}{\mathbb{P}(E_{B_i \setminus \{b_i^+\}} \mid E_{b_i^+}) \mathbb{P}(E_{b_i^+} \mid E_{b_i^-})} \\
&\quad = \prod_{i=1}^m \mathbb{P}(E_{A_i} \mid E_{b_i^-}, E_{b_i^+}).
\end{aligned}$$

Proof of (23). Let $b := \max(B)$. We apply (22) to obtain

$$\begin{aligned}
\mathbb{P}(X_A = \boldsymbol{x}_A \mid X_B \in S_B) &= \frac{\sum_{\boldsymbol{x}_B \in S_B} \mathbb{P}(X_A = \boldsymbol{x}_A, X_B = \boldsymbol{x}_B)}{\mathbb{P}(X_B \in S_B)} \\
&= \frac{\sum_{\boldsymbol{x}_B \in S_B} \mathbb{P}(X_A = \boldsymbol{x}_A \mid X_b = x_b)\mathbb{P}(X_B = \boldsymbol{x}_B)}{\mathbb{P}(X_B \in S_B)} \\
&= \mathbb{P}(X_A = \boldsymbol{x}_A \mid X_b = x_b).
\end{aligned}$$

Proof of (21). Let $n, m \in \mathbb{N}$ with $m \le n$. Let $B := [n-m, n-1]$, $x_n \in \mathcal{S}_n$, and $\boldsymbol{x}_B \in \mathcal{S}_B$. Let $C_+ := C \setminus [0, n-1]$ and $C_- := C \cap [0, n-1]$. Thus, $S_C = S_{C_-} \times S_{C_+}$. We apply (23) twice to show that $(X \mid X_C \in S_C)$ fulfils (19) and is thus Markov:

$$\begin{aligned}
&\mathbb{P}(X_n = x_n \mid X_B = \boldsymbol{x}_B, X_C \in S_C) \\
&\quad = \frac{\mathbb{P}(X_n = x_n, X_{C_+} \in S_{C_+}, X_B = \boldsymbol{x}_B, X_{C_-} \in S_{C_-})}{\mathbb{P}(X_B = \boldsymbol{x}_B, X_C \in S_C)} \\
&\quad = \frac{\mathbb{P}(X_n = x_n, X_{C_+} \in S_{C_+} \mid X_B = \boldsymbol{x}_B, X_{C_-} \in S_{C_-})\mathbb{P}(X_B = \boldsymbol{x}_B, X_{C_-} \in S_{C_-})}{\mathbb{P}(X_{C_+} \in S_{C_+} \mid X_B = \boldsymbol{x}_B, X_{C_-} \in S_{C_-})\mathbb{P}(X_B = \boldsymbol{x}_B, X_{C_-} \in S_{C_-})} \\
&\quad = \frac{\mathbb{P}(X_n = x_n, X_{C_+} \in S_{C_+} \mid X_{n-1} = x_{n-1})}{\mathbb{P}(X_{C_+} \in S_{C_+} \mid X_{n-1} = x_{n-1})} \\
&\quad = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{C_+} \in S_{C_+})
\end{aligned}$$

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

$$= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{C_+} \in S_{C_+}, X_{C_-} \in S_{C_-})$$
$$= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_C \in S_C).$$

3.4. The loss case

We start with derivations common to the proofs of (12a) and (13a). We assume that $\mathcal{K} < \infty$. Equation (3) is equivalent to the existence of $\check{x}$, $\hat{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}^{\mathcal{K}}$, and $\boldsymbol{x} \in g^{-1}(\boldsymbol{y})$ with

$$0 < \mathbb{P}(X_0 = \check{x}, X_{[\mathcal{K}]} = \boldsymbol{x}, X_{\mathcal{K}+1} = \hat{x}) < \mathbb{P}(X_0 = \check{x}, Y_{[\mathcal{K}]} = \boldsymbol{y}, X_{\mathcal{K}+1} = \hat{x}). \tag{24}$$

Let $\boldsymbol{s} := (\check{x}, \boldsymbol{x}, \hat{x})$. The *unreconstructable set of trajectories* $\mathcal{H}$ is

$$\mathcal{H} := \{\check{x}\} \times g^{-1}(\boldsymbol{y}) \times \{\hat{x}\}.$$

Equation (3) implies that $\mathcal{H}$ contains at least two elements with positive probability. If we pass through $\mathcal{H}$, then we incur an *entropy loss* L:

$$L := H(X_{[\mathcal{K}]} \mid X_{[0,\mathcal{K}+1]} \in \mathcal{H}) > 0. \tag{25}$$

Let $\mathcal{I}$ be the random set of indices marking the start of nonoverlapping runs of $X_{[n]}$ through $\mathcal{H}$, that is

$$\mathcal{I} := \{i \in [n - \mathcal{K} - 1] \colon X_{[i,i+\mathcal{K}+1]} \in \mathcal{H} \text{ and, for all } j \in [i+1, i+\mathcal{K}+1], X_{[j,j+\mathcal{K}+1]} \notin \mathcal{H}\},$$

where we select lower indices greedily. Taking $\boldsymbol{s}$ from just after (24), we lower bound the tail probability of the cardinality of $\mathcal{I}$ by that of $\mathcal{N}_X^{\boldsymbol{s}}(n)$:

$$\mathbb{P}(|\mathcal{I}| \geq m) \geq \mathbb{P}(|\mathcal{N}_X^{\boldsymbol{s}}(n)| \geq m) \quad \text{for all } m \in \mathbb{N}. \tag{26}$$

Finally, let

$$\alpha := \frac{\mathbb{P}(X_{[\mathcal{K}+2]} = \boldsymbol{s})}{2(\mathcal{K}+2)} > 0.$$

Proof of (12a). We claim that, for every $m \in \mathbb{N}$,

$$H(X_{[n]} \mid Y_{[n]}) \geq \mathbb{P}(|\mathcal{I}| \geq m) H(X_{[n]} \mid Y_{[n]}, |\mathcal{I}| \geq m) \geq \mathbb{P}(|\mathcal{I}| \geq m) m L. \tag{27}$$

Combining (26) and (27) for $m = \alpha n$ with (18b), we arrive at (12a)

$$\begin{aligned} \bar{H}(X \mid Y) &= \lim_{n\to\infty} \frac{1}{n} H(X_{[n]} \mid Y_{[n]}) \\ &\geq \lim_{n\to\infty} \frac{1}{n} \mathbb{P}(|\mathcal{I}| \geq \alpha n) \alpha n L \\ &\geq \alpha L \lim_{n\to\infty} \mathbb{P}(|\mathcal{N}_X^{\boldsymbol{s}}(n)| \geq \alpha n) \\ &= \alpha L \\ &> 0. \end{aligned} \tag{28}$$

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

It remains to prove (27). We fix $m, n \in \mathbb{N}$. For $I \subseteq [n]$ with $\mathbb{P}(\mathcal{I} = I) > 0$ and each $i \in I$, we derive the indices of the block $B_i := [i, i + \mathcal{K} + 1]$ and its interior $\widehat{B}_i := [i + 1, i + \mathcal{K}]$. Their unions are $B := \biguplus_{i \in I} B_i$ and $\widehat{B} := \biguplus_{i \in I} \widehat{B}_i$, respectively. Hence,

$$\begin{aligned} H(X_{[n]} \mid Y_{[n]}, \mathcal{I} = I) &\geq H(X_{\widehat{B}} \mid X_{[n]\setminus B}, \text{for all } i \in I\colon X_{B_i} \in \mathcal{H}) && \text{(29a)} \\ &= H(X_{\widehat{B}} \mid \text{for all } i \in I\colon X_{B_i} \in \mathcal{H}) && \text{(29b)} \\ &= \sum_{i \in I} H(X_{\widehat{B}_i} \mid X_{B_i} \in \mathcal{H}) && \text{(29c)} \\ &= |I| \times L, && \text{(29d)} \end{aligned}$$

where in (29a) we discard all information outside $\widehat{B}$ and condition on it, in (29b) we apply the conditional factorisation (20) to remove every condition except the block ends, in (29c) we apply the conditional factorisation (20) to the Markov process $(X \mid X_B \in \mathcal{H}^{|I|})$ (as $\mathcal{H}$ is a cartesian product), and in (29d) we conclude by stationarity and the minimum loss (25). Hence,

$$\begin{aligned} H(X_{[n]} \mid Y_{[n]}, |\mathcal{I}| \geq m) &= \sum_{I \subseteq [n],\, |I| \geq m} \mathbb{P}(\mathcal{I} = I \mid |\mathcal{I}| \geq m) H(X_{[n]} \mid Y_{[n]}, \mathcal{I} = I) \\ &\geq \sum_{I \subseteq [n],\, |I| \geq m} \mathbb{P}(\mathcal{I} = I \mid |\mathcal{I}| \geq m) \times |I| \times L \\ &\geq mL. \end{aligned}$$

Proof of (13a). Taking $\boldsymbol{s}$ from just after (24), we have

$$T_n \geq 2^{|\mathcal{N}_X^{\boldsymbol{s}}(n)|}. \tag{30}$$

Thus, (18a) and (30) imply that

$$\begin{aligned} \liminf_{n \to \infty} \sqrt[n]{T_n} &\geq \liminf_{n \to \infty} \exp\left((\log 2) \frac{1}{n} |\mathcal{N}_X^{\boldsymbol{s}}(n)| \right) \\ &= \exp\left((\log 2) \liminf_{n \to \infty} \frac{1}{n} |\mathcal{N}_X^{\boldsymbol{s}}(n)| \right) \\ &\overset{\mathbb{P}-\text{a.s.}}{\geq} \exp((\log 2)\alpha) \qquad (31) \\ &= 2^{\alpha} > 1. \end{aligned}$$

Proof of Proposition 2. Let $x_0, x_{\mathcal{K}+1}, \boldsymbol{y}, \boldsymbol{x}', \boldsymbol{x}''$ be as in (3). Suppose that $\mathcal{K} > K := \sum_{y \in \mathcal{Y}} |g^{-1}(y)|(|g^{-1}(y)| - 1)$ and $\mathcal{K} > 1$. We apply the *pigeonhole principle*, first to every $x \in g^{-1}(y)$ and then to each $g^{-1}(y)$ for every $y \in \operatorname{supp} \boldsymbol{y}$. This ensures that the two trajectories intersect:

$$\text{there exists } m \in [\mathcal{K}] \text{ such that } \boldsymbol{x}'_{\{m\}} = \boldsymbol{x}''_{\{m\}}. \tag{32}$$

Choose m to satisfy (32). If $m = 1$, then $\boldsymbol{x}'_{\{1\}}, x_{\mathcal{K}+1}, \boldsymbol{y}_{[2,\mathcal{K}]}, \boldsymbol{x}'_{[2,\mathcal{K}]}$, and $\boldsymbol{x}''_{[2,\mathcal{K}]}$ fulfill the conditions in (3). If $m > 1$, then $x_0, \boldsymbol{x}'_{\{m\}}, \boldsymbol{y}_{[m-1]}, \boldsymbol{x}'_{[m-1]}$, and $\boldsymbol{x}''_{[m-1]}$ fulfill the conditions in (3). Both cases lead to $\mathcal{K} < \mathcal{K}$, which is a contradiction.

4. Proof of strong k-lumpability

For (conditional) probabilities, we use the shorthand notation

$$\mathbb{P}(Z = z) = p_Z(z) \quad \text{and} \quad \mathbb{P}(Z_1 = z_1 \mid Z_2 = z_2) = p_{Z_1 \mid Z_2}(z_1 \mid z_2),$$

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

where we always assume that the latter is well defined, i.e. that $p_{Z_2}(z_2) > 0$. Recall that the *conditional mutual information* of Z_1 and Z_2 given Z_3 is

$$I(Z_1; Z_2 \mid Z_3) := H(Z_1 \mid Z_3) - H(Z_1 \mid Z_2, Z_3).$$

The conditional mutual information vanishes, if and only if Z_1 and Z_2 are conditionally independent given Z_3 [7, Theorem 2.6.3].

Proof of Proposition 3. The right-hand side of (7) is equivalent to

$$\begin{aligned} 0 &= H(Z_k \mid Z_{[0,k-1]}) - \bar{H}(Z) \\ &= H(Z_k \mid Z_{[0,k-1]}) - \lim_{n\to\infty} H(Z_n \mid Z_{[0,n-1]}) \\ &= \lim_{n\to\infty} (H(Z_n \mid Z_{[n-k,n-1]}) - H(Z_n \mid Z_{[0,n-1]})) \\ &= \lim_{n\to\infty} I(Z_n; Z_{[0,n-k-1]} \mid Z_{[n-k,n-1]}). \end{aligned}$$

By stationarity, the sequence in the last limit increases monotonically in n. A limit value of zero is equivalent to

$$\begin{aligned} & p_{Z_n \mid Z_{[n-k,n-1]}}(\cdot \mid z) p_{Z_{[0,n-k-1]} \mid Z_{[n-k,n-1]}}(\cdot \mid z) \\ &\quad = p_{Z_n, Z_{[0,n-k-1]} \mid Z_{[n-k,n-1]}}(\cdot \mid z) \\ &\quad = p_{Z_n \mid Z_{[0,n-1]}}(\cdot \mid \cdot, z) p_{Z_{[0,n-k-1]} \mid Z_{[n-k,n-1]}}(\cdot \mid z) \quad \text{for all } n \in \mathbb{N}, \end{aligned}$$

where the first equality holds $p_{Z_{[n-k,n-1]}}$-a.s. The equality between the first and last line is equivalent to the higher-order Markov property (6).

Proof of Theorem 2. The equivalence in (9) follows from the equivalence of its two statements to the following technical property: for all $y', y \in \mathcal{Y}$, $\boldsymbol{y} \in \mathcal{Y}^{k-1}$, and $x \in g^{-1}(y)$:

$$\begin{aligned} & p_{Y_k, Y_{[k-1]}, X_0}(y', \boldsymbol{y}, x) > 0 \\ & \quad \Longrightarrow \quad 0 < p_{Y_k \mid Y_{[k-1]}, X_0}(y' \mid \boldsymbol{y}, x) = p_{Y_k \mid Y_{[k-1]}, Y_0}(y' \mid \boldsymbol{y}, y). \end{aligned} \tag{33}$$

The equivalence between (9a) and (33) is in Proposition 7, and the equivalence between (9b) and (33) is in Proposition 8.

Proposition 7. *For a lumping (P, g), property (9a) is equivalent to (33).*

Proof. We rewrite (9a) as

$$\begin{aligned} 0 &= H(Y_k \mid Y_{[0,k-1]}) - H(Y_k \mid Y_{[k-1]}, X_0) \\ &= H(Y_k \mid Y_{[0,k-1]}) - H(Y_k \mid Y_{[0,k-1]}, X_0) \\ &= I(Y_k; X_0 \mid Y_{[0,k-1]}). \end{aligned}$$

For all $y' \in \mathcal{Y}$, $\boldsymbol{y} \in \mathcal{Y}^k$, and $x \in \mathcal{X}$ with $p_{Y_k, Y_{[0,k-1]}, X_0}(y', \boldsymbol{y}, x) > 0$, this is equivalent to

$$0 < p_{Y_k, X_0 \mid Y_{[0,k-1]}}(\cdot \mid \boldsymbol{y}) = p_{Y_k \mid Y_{[0,k-1]}}(\cdot \mid \boldsymbol{y}) p_{X_0 \mid Y_{[0,k-1]}}(\cdot \mid \boldsymbol{y}).$$

Division in the previous line equals (33).

Proposition 8. *A lumping (P, g) is strongly k-lumpable, if and only if (33) holds.*

Proof. This is a straightforward generalization of the proof for the case $k = 1$ in [17].

https://doi.org/10.1239/jap/1421763331 Published online by Cambridge University Press

5. Proofs of the sufficient conditions

We use the shorthand notation introduced at the beginning of Section 4.

Proof of Proposition 4. We have

$$H(Y_k \mid X_{k-1}) \le H(Y_k \mid Y_{[k-1]}, X_0) \le \bar{H}(Y) \le \bar{H}(X) = H(X_k \mid X_{k-1}),$$

where the first and the second inequality are due to [7, Theorem 4.5.1, p. 86] (compare Lemma 1) and the third inequality is due to data processing [9], [24]. The SE property implies that $p_{X_k, X_{k-1}}$-a.s.

$$p_{Y_k \mid X_{k-1}}(y \mid x) = p_{X_k \mid X_{k-1}}(x'(x, y) \mid x),$$

where $x'(x, y)$ is a unique endpoint of the edge existing by (10). Thus, the outer terms in the above chain of inequalities coincide, yielding $\bar{H}(Y) = \bar{H}(X)$. This completes the proof.

Proof of Proposition 5. First, we show that SFS(k) is a subclass of SE, implying preservation of entropy. If SE does not hold, then there exist states $y^\star \in \mathcal{Y}$ and $x^\star \in \mathcal{X}$ such that at least two states $x', x'' \in g^{-1}(y^\star)$ have positive transition probabilities from $x^\star$. Choose a realisable path $\boldsymbol{x}_{[0,k-3]}$, with positive transition probability from x_{k-3} to $x^\star$. Let $\boldsymbol{y} = (g(\boldsymbol{x}_{[k-3]}), g(x^\star), y^\star) \in \mathcal{Y}^{k-1}$. We have

$$p_{X_{[k-1]} \mid Y_{[k-1]}, X_0}(\boldsymbol{x}_{[k-3]}, x^\star, x' \mid \boldsymbol{y}, \boldsymbol{x}_{\{0\}}) > 0$$

and

$$p_{X_{[k-1]} \mid Y_{[k-1]}, X_0}(\boldsymbol{x}_{[k-3]}, x^\star, x'' \mid \boldsymbol{y}, \boldsymbol{x}_{\{0\}}) > 0.$$

This contradicts the definition of SFS(k) (11).

Second, we show that SFS(k) implies strong k-lumpability of (P, g). We check (33) and then conclude via Proposition 8. We have $p_{Y_{[k-1]}, X_0}$-a.s. a unique $\boldsymbol{x}'(g(X_0), \boldsymbol{Y}_{[k-1]}) \in \mathcal{X}^{k-1}$ fulfilling (11). Hence,

$$\begin{aligned}
&p_{Y_k \mid Y_{[k-1]}, X_0}(y \mid \boldsymbol{y}, x) \\
&\quad = p_{Y_k \mid Y_{[k-1]}, X_{[k-1]}, X_0}(y \mid \boldsymbol{y}, \boldsymbol{x}'(g(x), \boldsymbol{y}), x) \underbrace{p_{X_{[k-1]} \mid Y_{[k-1]}, X_0}(\boldsymbol{x}'(g(x), \boldsymbol{y}) \mid \boldsymbol{y}, x)}_{=1 \text{ by (11)}} \\
&\quad = p_{Y_k \mid X_{[k-1]}, X_0}(y \mid \boldsymbol{x}'(g(x), \boldsymbol{y}), x) \\
&\quad = p_{Y_k \mid X_{k-1}}(y \mid \boldsymbol{x}'(g(x), \boldsymbol{y})_{\{k-1\}}) \quad \text{(by the Markov property of } X\text{)}
\end{aligned}$$

is independent of x and (33) holds.

Acknowledgements

We thank Gernot Kubin and Wolfgang Woess for establishing contact between us, leading to our joint investigation of this topic. We are particularly indebted to our anonymous reviewer for his/her thoughtful comments and encouragement to elaborate the section concerning k-lumpability.

References

[1] Anderson, B. D. O. (1999). The realization problem for hidden Markov models. *Math. Control Signals Systems* **12,** 80–120.

[2] Brown, P. F. *et al.* (1992). Class-based n-gram models of natural language. *Comput. Linguist.* **18,** 467–479.

[3] Blackwell, D. (1957). The entropy of functions of finite-state Markov chains. In *Trans 1st Prague Conf. Inf. Theory, Statist. Decision Functions, (Liblice, 1956)*. Publishing House of the Czechoslovak Academy of Sciences, Prague, pp. 13–20.
[4] Blackwell, D. and Koopmans, L. (1957). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **28,** 1011–1015.
[5] Burke, C. J. and Rosenblatt, M. (1958). A Markovian function of a Markov chain. *Ann. Math. Statist.* **29,** 1112–1122.
[6] Carlyle, J. W. (1967). Identification of state-calculable functions of finite Markov chains. *Ann. Math. Statist.* **38,** 201–205.
[7] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd edn. John Wiley, Hoboken, NJ.
[8] Ephraim, Y. and Merhav, N. (2002). Hidden Markov processes. *IEEE Trans. Inf. Theory* **48,** 1518–1569.
[9] Geiger, B. C. and Kubin, G. (2011). Some results on the information loss in dynamical systems. In *Proc. IEEE Internat. Symp. Wireless Commun. Systems (ISWSC)*, IEEE, New York, pp. 794–798, 2011. Extended version available at http://uk.arxiv.org/abs/1106.2404.
[10] Geiger, B. C. and Temmel, C. (2013). Information-preserving Markov aggregation. In *Proc. IEEE Information Theory Workshop (ITW)*, IEEE, New York, pp. 258–262. Extended version available at http://uk.arxiv.org/abs/1304.0920.
[11] Gilbert, E. J. (1959). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **30,** 688–697.
[12] Gray, R. M. (1990). *Entropy and Information Theory*. Springer, New York.
[13] Gurvits, L. and Ledoux, J. (2005). Markov property for a function of a Markov chain: a linear algebra approach. *Linear Algebra Appl.* **404,** 85–117.
[14] Heiner, M., Rohr, C., Schwarick, M. and Streif, S. (2010). A comparative study of stochastic analysis techniques. In *Proc. 8th Internat. Conf. Comput. Meth. Systems Biol.*, ACM, New York, pp. 96–106.
[15] Heller, A. (1965). On stochastic processes derived from Markov chains. *Ann. Math. Statist.* **36,** 1286–1291.
[16] Henzinger, T. A., Mikeev, L., Mateescu, M. and Wolf, V. (2010). Hybrid numerical solution of the chemical master equation. In *Proc. 8th Internat. Conf. Comput. Meth. Systems Biol.*, ACM, New York, pp. 55–65.
[17] Kemeny, J. G. and Snell, J. L. (1976). *Finite Markov Chains*. Springer, New York.
[18] Kieffer, J. C. and Rahe, M. (1981). Markov channels are asymptotically mean stationary. *SIAM J. Math. Anal.* **12,** 293–305.
[19] Lindqvist, B. (1978). On the loss of information incurred by lumping states of a Markov chain. *Scand. J. Statist.* **5,** 92–98.
[20] Parzen, E. (1999). *Stochastic Processes* (Classics Appl. Math. **24**). Society for Industrial and Applied Mathematics, Philadelphia, PA.
[21] Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, CA.
[22] Rogers, L. C. G. and Pitman, J. W. (1981). Markov functions. *Ann. Prob.* **9,** 573–582.
[23] Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains. *Comput. Networks* **33,** 377–386.
[24] Watanabe, S. and Abraham, C. T. (1960). Loss and recovery of information by coarse observation of stochastic chain. *Inf. Control* **3,** 248–278.
[25] Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC, Boca Raton, FL.
[26] Woess, W. (2009). *Denumerable Markov chains. Generating Functions, Boundary Theory, Random Walks on Trees*. European Mathematical Society, Zürich.