This is a review submitted to Mathematical Reviews/MathSciNet.

Reviewer Name: Konstantopoulos, Takis

Mathematical Reviews/MathSciNet Reviewer Number: 68397

Address:
Department of Mathematics
Uppsala University
PO Box 480
SE-75106 Uppsala
SWEDEN
takiskonst@gmail.com

Author: Geiger, Bernhard C.; Temmel, Christoph

Title: Lumpings of Markov chains, entropy rate preservation, and higher-order lumpability.

MR Number: MR3301292

Primary classification:

Secondary classification(s):

Review text:

The paper studies two related issues regarding functions of an irreducible and aperiodic Markov chain with values in a finite state space $\mathcal{X}$. The first is criteria for preservation or loss of information by a function of the chain. The second issue concerns the preservation of Markov property (of, possibly, higher order) by functions.

Regarding the first issue, let $g : \mathcal{X} \to \mathcal{Y} = g(\mathcal{X})$ be a deterministic function (called "lumping" or "coloring"). Say that a sequence $(x_0, \dots, x_n)$ of states is *realizable* if it is possible for the chain to assume these states consecutively. Call such a sequence *identifiable* if, given the endpoints x_0 and x_n and the coloring $g(x_0), g(x_1), \dots, g(x_n)$, one can uniquely identify the sequence $(x_0, \dots, x_n)$. The *split-merge index* $\mathcal{K}$ is the largest n for which all such sequences are identifiable, with $\mathcal{K} = \infty$ if sequences of all lengths are identifiable. Let $X = (X_n, n \in \mathbb{Z})$ be the (necessarily unique) stationary version of the chain and set $Y_n := g(X_n)$ for the stationary sequence of its coloring. Let $\bar{H}(X)$, $\bar{H}(Y)$ be the entropy rates of the two stationary (and, necessarily, ergodic) sequences. Then $\bar{H}(X|Y)$ represents the information (rate) lost in trying to identify X by observing Y; necessarily, $\bar{H}(X|Y) = \bar{H}(X) - \bar{H}(Y) \geq 0$. For a sequence $(y_1, \dots, y_n)$ of colors let $R(y_1, \dots, y_n)$ be the set of realizable sequences $(x_1, \dots, x_n)$ of states with colors $g(x_i) = y_i$, $1 \leq i \leq n$. Define T_n to be a random variable with cardinality equal to the cardinality of $R(Y_1, \dots, Y_n)$, where $Y_1, \dots, Y_n$ are n consecutive colors from the stationary sequence Y. The first result states that

$$\bar{H}(X|Y) > 0 \iff \mathcal{K} < \infty \iff \limsup_{n\to\infty} T_n^{1/n} \geq C, \text{ a.s., for some constant } C > 1.$$

In the complementary case, $\bar{H}(X|Y) = 0$ (no information loss) we have that $\limsup_{n\to\infty} T_n$ is a.s. bounded below a deterministic constant. A sufficient condition for zero information loss is that the colored chain be *single entry* (SE) meaning that, given a state $x \in \mathcal{X}$ and a color $y \in \mathcal{Y}$ we can uniquely identify (if any) the state $x' \in g^{-1}(y)$ with $p(x, x') > 0$.

For the second issue, say that $(Y_n = g(X_n), n \geq 0)$ is weakly k-lumpable (k a positive integer) if it is a k-th order Markov chain when (X_n) is assumed to be stationary. Say that it is strongly k-lumpable if there exists a k-th order transition probability function $q(y_0, \dots, y_{k-1}; y)$ on $\mathcal{Y}$ such that, for any probability distribution ν on $\mathcal{X}$, if the law of X_0 is ν then $\mathbb{P}(Y_{m+k} = y|Y_m = y_0, \dots, Y_{m+k-1} = y_{k-1}) = q(y_0, \dots, y_{k-1}; y)$, for all $m \geq 0$ and all $y_0, \dots, y_{k-1}$, $y \in \mathcal{Y}$. Strong k-lumpability is equivalent to

$$H(Y_k \mid X_0, Y_1, \dots Y_{k-1}) = H(Y_k \mid Y_1, \dots Y_{k-1}),$$

A sufficient condition for strong k-lumpability is that the chain be weakly k-lumpable and single entry, in the aforementioned sense. Another sufficient condition for strong k-lumpability is that the chain have the *single forward k-sequence* property meaning that there is at most one realizable sequence $x_0, \dots, x_{k-1}$ with colors $y, y_1, \dots, y_{k-1}$, respectively, for any sequence of such colors.

The paper also presents several simple examples which exemplify the relationships between the introduced notions.