Capacity of Finite State Channels Based on Lyapunov Exponents of Random Matrices

Tim Holliday, *Member, IEEE*, Andrea Goldsmith, *Fellow, IEEE*, and Peter Glynn

***Abstract*—The finite-state Markov channel (FSMC) is a time-varying channel having states that are characterized by a finite-state Markov chain. These channels have infinite memory, which complicates their capacity analysis. We develop a new method to characterize the capacity of these channels based on Lyapunov exponents. Specifically, we show that the input, output, and conditional entropies for this channel are equivalent to the largest Lyapunov exponents for a particular class of random matrix products. We then show that the Lyapunov exponents can be expressed as expectations with respect to the stationary distributions of a class of continuous-state space Markov chains. This class of Markov chains, which is closely related to the prediction filter in hidden Markov models, is shown to be nonirreducible. Hence, much of the standard theory for continuous state-space Markov chains cannot be applied to establish the existence and uniqueness of stationary distributions, nor do we have direct access to a central limit theorem (CLT). In order to address these shortcomings, we utilize several results from the theory of random matrix products and Lyapunov exponents. The stationary distributions for this class of Markov chains are shown to be unique and continuous functions of the input symbol probabilities, provided that the input sequence has finite memory. These properties allow us to express mutual information and channel capacity in terms of Lyapunov exponents. We then leverage this connection between entropy and Lyapunov exponents to develop a rigorous theory for computing or approximating entropy and mutual information for finite-state channels with dependent inputs. We develop a method for directly computing entropy of finite-state channels that does not rely on simulation and establish its convergence. We also obtain a new asymptotically tight lower bound for entropy based on norms of random matrix products. In addition, we prove a new functional CLT for sample entropy and apply this theorem to characterize the error in simulated estimates of entropy. Finally, we present numerical examples of mutual information computation for intersymbol interference (ISI) channels and observe the capacity benefits of adding memory to the input sequence for such channels.**

***Index Terms*—Finite-state channel, hidden Markov model, Lyapunov exponent, random matrices, Shannon capacity.**

Manuscript received July 18, 2003; revised October 19, 2005. This work was supported by the National Science Foundation under Grant CMS-0120912 and by the Office of Naval Research under Grant N00014-05-0168. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Yokohama, Japan, June/July 2003.

T. Holliday is with Princeton University, Princeton, NJ 08544 USA (e-mail: thollida@princeton.edu).

A. Goldsmith and P. Glynn are with Stanford University, Stanford, CA 94305 USA (e-mail: andrea@systems.stanford.edu; glynn@stanford.edu).

Communicated by A. Kavčić, Assoicate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2006.878230

I. Introduction

In this work, we develop the theory required to compute entropy and mutual information for Markov channels with dependent inputs using Lyapunov exponents. We model the channel as a finite state discrete-time Markov chain (DTMC). Each state in the DTMC corresponds to a memoryless channel with finite input and output alphabets. The capacity of the Markov channel is well known for the case of perfect state information at the transmitter and receiver. We consider the case where only the transition dynamics of the DTMC are known (i.e., no state information).

This problem was orginally considered by Mushkin and Bar-David [38] for the Gilbert–Elliot channel. Their results show that the mutual information for the Gilbert–Elliot channel can be computed as a continuous function of the stationary distribution for a continuous-state space Markov chain. The results of [38] were later extended by Goldsmith and Varaiya [24] to Markov channels with independent and identically distributed (i.i.d.) inputs and channel transition probabilities that are not dependent on the input symbol process. The key result of [24] is that the mutual information for this class of Markov channels can be computed as expectations with respect to the stationary distribution of the channel prediction filter. In both of these papers, it is noted that these results fail for non-i.i.d. input sequences or if the channel transitions depend on the input sequence. These restrictions rule out a number of interesting problems. In particular, intersymbol interference (ISI) channels do not fall into the above frameworks. In a more general setting, Blackwell originally considered the connection between entropy rates and the prediction filter in [10]. The assumptions in [10] are somewhat different than those used here, and we detail these issues in Sections III-D and V-B of this paper.

Recently, several authors ([1]–[3], [44], [31]) have proposed simulation-based methods for computing the sample mutual information of finite-state channels. A key advantage of the proposed simulation algorithms is that they can compute mutual information for Markov channels with non-i.i.d. input symbols as well as for ISI channels. In particular, simulation-based algorithms were used in [2] to quantify the capacity increase associated with input memory. All of the simulation-based results use similar methods to compute the sample mutual information of a simulated sequence of symbols that are generated by a finite-state channel. These works rely on the Shannon–McMillan–Breiman theorem to ensure convergence of the sample mutual information to the expected mutual information. However, the theory of simulation for these channels is still incomplete. We will show in Section VI that the Markov chain we must simulate is typically *not* irreducible. Hence, we

cannot apply the standard theory of simulation to construct rigorous confidence intervals or error bounds on the simulated estimate of mutual information. The lack of error bounds for these simulation estimates means that we cannot determine, *a priori*, the length of time needed to run a simulation nor can we determine the termination time by observing the simulated data [28]. Furthermore, simulations used to compute mutual information may require extremely long "startup" times to remove initial transient bias. Examples demonstrating this problem can be found in [20] and [13]. We will discuss this issue in more detail in Section VI and Appendix I of this paper.

Our goal in this paper is to present a detailed and rigorous treatment of the computational and statistical properties of entropy and mutual information for finite-state channels with dependent inputs. Our first result, which we will exploit throughout this paper, is that the entropy rate of a symbol sequence generated by a Markov channel is equivalent to the largest Lyapunov exponent for a product of random matrices. This connection between entropy and Lyapunov exponents provides us with a substantial body of theory that we may apply to the problem of computing entropy and mutual information for finite-state channels. In addition, this result provides many interesting connections between the theory of dynamic systems, hidden Markov models, and information theory, thereby offering a different perspective on traditional notions of information-theoretic concepts. A number of recent papers have explored these issues. The authors of [39] and [40] consider closed-form computation of entropy rates for finite-state channels in particular asymptotic regimes. In [27], conditions similar to those in Theorem 6 of this paper are provided that guarantee analyticity of entropy rates for finite-state channels. In [23], the authors provide new low-complexity bounds on the computation of Lyapunov exponents, which can be directly applied to entropy.

Our results fall into two categories: extensions of previous research and entirely new results—we summarize the new results first. We provide a new connection between entropy and Lyapunov exponents that allows us to prove several new theorems for entropy and mutual information of finite-state channels. In particular, we present new lower bounds for entropy in terms of matrix and vector norms for products of random matrices (Section VI and Appendix I). We also provide an explicit connection between computation of Lyapunov exponents, entropy, and the prediction filter problem in hidden Markov models (Section IV). In conjunction with the Lyapunov exponent results, we utilize ideas from continuous-state space Markov chains to prove the following new results:

- a method for directly computing entropy and mutual information for finite-state channels (Theorem 7);
- a functional central limit theorem (CLT) for sample entropy under easily verifiable conditions (Theorem 8);
- a functional CLT for a simulation based estimate of entropy (Theorem 8);
- a rigorous confidence interval methodology for simulation based computation of entropy (Section VI and Appendix I);
- a rigorous method for bounding the amount of initialization bias present in a simulation-based estimate (Appendix I).

A functional CLT is a stronger form of the standard CLT. It shows that the sample entropy, when viewed as a function of the amount of observed data, can be approximated by a Brownian motion. In addition, the proof techniques utilized in Section V, which exploit the contraction property of positive matrices, will likely be of interest to some readers. This powerful property allows us to prove existence, uniqueness, and continuity of a stationary distribution for the generalized hidden Markov prediction filter, even though the filter process is not irreducible.

In addition to the above new results, we provide several extensions of the work presented in [24]. In [24], the authors show that mutual information can be computed as a function of the conditional channel state probabilities (where the conditioning is on past input/output symbols). Moreover they show that for the case of i.i.d. inputs, the conditional channel probabilities converge weakly and their limit distributions are continuous functions of the input probabilities and channel parameters. In this paper, we will show that all of these properties hold for a much more general class of finite-state channels and inputs (Sections III–V). Furthermore, we also strengthen the results in [24] to show that the conditional channel probabilities converge exponentially fast. In addition, we apply results from Lyapunov exponent theory to show that there may be cases where entropy and mutual information for finite-state channels are not "well-behaved" quantities (Section IV). For example, we show that the conditional channel probability process does not have even the weakest form of irreducibility (i.e., Harris recurrence), and may have multiple stationary distributions. While entropy is always a continuous function of the input probabilities and channel parameters [15, Theorem 4.4.1], we show that the conditional channel probability process may have discontinuous stationary distributions. This lack of continuity could prove significant when computing functions of hidden Markov models other than entropy.

The rest of this paper is organized as follows. In Section II, we show that the conditional entropy of the output symbols given the input symbols can be represented as a Lyapunov exponent for a product of random matrices. We show that this property holds for *any* ergodic input sequence. In Section III, we show that under stronger conditions on the input sequence, the entropy of the outputs and the joint input/output entropy are also Lyapunov exponents. In Section IV, we show that entropies can be computed as expectations with respect to the stationary distributions of a class of continuous-state space Markov chains. We also provide an example in which such a Markov chain has multiple stationary distributions. In Section V, we provide conditions on the finite-state channel and input symbol process that guarantee uniqueness and continuity of the stationary distributions for the continuous-state space Markov chains. In Section VI, we discuss both simulation-based and non-simulation-based computation of entropy and mutual information for finite-state channels. In Section VII, we present numerical examples of computing mutual information for finite-state channels with general inputs. Finally, in Appendix I, we present a rigorous treatment of the theory required to construct simulation-based estimates of entropy.

II. Markov Channels With Ergodic Input Symbol Sequences

We consider here a communication channel with (channel) state sequence $C = (C_n\colon n \geq 0)$, input symbol sequence $X = (X_n\colon n \geq 0)$, and output symbol sequence $Y = (Y_n\colon n \geq 0)$. The channel states take values in $\mathcal{C}$, whereas the input and output symbols take values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. In this paper, we shall adopt the notational convention that if $s = (s_n\colon n \geq 0)$ is any generic sequence, then for $m, n \geq 0$

$$s_m^{m+n} = (s_m, \ldots, s_{m+n})$$

denotes the finite segment of s starting at index m and ending at index $m+n$.

In this section, we show that the conditional entropy of the output symbols given the input symbols can be represented as a Lyapunov exponent for a product of random matrices. In order to state this relation, we only require that the input symbol sequence be stationary and ergodic. Unfortunately, we cannot show an equivalence between unconditional entropies and Lyapunov exponents for such a general class of inputs. In Section III, we will discuss the equivalence of Lyapunov exponents and unconditional entropies for the case of Markov-dependent inputs.

A. Channel Model Assumptions

With the above notational convention in hand, we are now ready to describe this section's assumptions on the channel model. While some of these assumptions will be strengthend in the following sections, we will use the same notation for the channel throughout the paper.

A1: $C = (C_n\colon n \geq 0)$ is a stationary finite-state irreducible Markov chain, possessing transition matrix $R = (R(c_n, c_{n+1})\colon c_n, c_{n+1} \in \mathcal{C})$. In particular

$$P\left(C_0^n = c_0^n\right) = r(c_0) \prod_{j=0}^{n-1} R(c_j, c_{j+1})$$

for $c_0^n \in \mathcal{C}$, where $r = (r(c) : c \in \mathcal{C})$ is the unique stationary distribution of C.

A2: The input symbol sequence $X = (X_n\colon n \geq 0)$ is a stationary ergodic sequence independent of C.

A3: The output symbols $\{Y_n\colon n \geq 0\}$ are conditionally independent given X and C, so that

$$P\left(Y_0^n = y_0^n \,|\, C, X\right) = \prod_{j=0}^{n-1} P(Y_j = y_j \,|\, C, X)$$

for $y_0^n \in \mathcal{Y}^{n+1}$.

A4: For each triplet $(c_0, c_1, x) \in \mathcal{C}^2 \times \mathcal{X}$, there exists a probability mass function $q(\cdot \,|\, c_0, c_1, x)$ on $\mathcal{Y}$ such that

$$P(Y_j = y \,|\, C, X) = q(y \,|\, C_j, C_{j+1}, X_j).$$

The dependence of Y_j on C_{j+1} is introduced strictly for mathematical convenience that will become clear shortly. While this extension does allow us to address noncausal channel models, it is of little practical use.

B. The Conditional Entropy as a Lyapunov Exponent

Let the stationary distribution of the channel be represented as a row vector $r = (r(c)\colon c \in \mathcal{C})$, and let e be a column vector in which every entry is equal to one. Furthermore, for $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let $G_{(x,y)} = (G_{(x,y)}(c_0, c_1)\colon c_0, c_1 \in \mathcal{C})$ be the square matrix with entries

$$G_{(x,y)}(c_0, c_1) = R(c_0, c_1) q(y \,|\, c_0, c_1, x).$$

Observe that

$$\begin{aligned} &P\left(Y_0^n = y_0^n \,|\, X_0^n = x_0^n\right) \\ &\quad = \sum_{c_0, \ldots, c_{n+1}} r(c_0) \prod_{j=0}^{n} R(c_j, c_{j+1}) q(y_j \,|\, c_j, c_{j+1}, x_j) \\ &\quad = \sum_{c_0, \ldots, c_{n+1}} r(c_0) \prod_{j=0}^{n} G_{(x_j, y_j)}(c_j, c_{j+1}) \\ &\quad = r G_{(x_0, y_0)} G_{(x_1, y_1)} \cdots G_{(x_n, y_n)} e. \end{aligned}$$

Taking logarithms, dividing by n, and letting $n \to \infty$ we conclude that

$$H(Y \,|\, X) = -\lim_{n \to \infty} \frac{1}{n} \mathrm{E} \log P\left(Y_0^n \,|\, X_0^n\right) = -\lambda(Y \,|\, X) \quad (1)$$

where

$$\lambda(Y \,|\, X) = \lim_{n \to \infty} \frac{1}{n} \mathrm{E} \log \left(r G_{(X_0, Y_0)} G_{(X_1, Y_1)} \cdots G_{(X_n, Y_n)} e \right). \quad (2)$$

The quantity $\lambda(Y \,|\, X)$ is the largest Lyapunov exponent (or, simply, Lyapunov exponent) associated with the sequence of random matrix products $(G_{(X_0, Y_0)} G_{(X_1, Y_1)} \cdots G_{(X_n, Y_n)} : n \geq 0)$. Lyapunov exponents have been widely studied in many areas of applied mathematics, including discrete linear differential inclusions (Boyd *et al.* [11]), statistical physics (Ravishankar [45]), mathematical demography (Cohen [14]), percolation processes (Darling [16]), and Kalman filtering (Atar and Zeitouni [7]).

Let $\|\cdot\|$ be any matrix norm for which $\|A_1 A_2\| \leq \|A_1\| \cdot \|A_2\|$ for any two matrices A_1 and A_2. Within the Lyapunov exponent literature, the following result is of central importance.

Theorem 1: Let $(B_n\colon n \geq 0)$ be a stationary ergodic sequence of random matrices for which $\mathrm{E} \log(\max(\|B_0\|, 1)) < \infty$. Then, there exists a deterministic constant λ (known as the Lyapunov exponent) such that

$$\frac{1}{n} \log \|B_1 B_2 \cdots B_n\| \to \lambda \quad \text{a.s.}$$

as $n \to \infty$. Furthermore

$$\begin{aligned} \lambda &= \lim_{n \to \infty} \frac{1}{n} \mathrm{E} \log \|B_1 \cdots B_n\| \\ &= \inf_{n \geq 1} \frac{1}{n} \mathrm{E} \log \|B_1 \cdots B_n\|. \end{aligned}$$

The standard Proof of Theorem 1 is based on the subadditive ergodic theorem due to Kingman [33].

Note that for $\|A\|_\infty \triangleq \max\{\sum_{c_1} |A(c_0,c_1)|\colon c_0 \in \mathcal{C}\}$

$$\begin{aligned}\min_{c\in\mathcal{C}} r(c)\|G_{(X_0,Y_0)}G_{(X_1,Y_1)}\cdots G_{(X_n,Y_n)}\|_\infty \\ \le rG_{(X_0,Y_0)}G_{(X_1,Y_1)}\cdots G_{(X_n,Y_n)}e \\ \le \|G_{(X_0,Y_0)}G_{(X_1,Y_1)}\cdots G_{(X_n,Y_n)}\|_\infty.\end{aligned}$$

The positivity of r therefore guarantees that

$$\begin{aligned}\frac{1}{n}\mathrm{E}\log(rG_{(X_0,Y_0)}G_{(X_1,Y_1)}\cdots G_{(X_n,Y_n)}e) \\ -\frac{1}{n}\mathrm{E}\log\|G_{(X_0,Y_0)}G_{(X_1,Y_1)}\cdots G_{(X_n,Y_n)}\|_\infty \to 0 \quad (3)\end{aligned}$$

as $n \to \infty$, so that the existence of the limit in (2) may be deduced either from information theory (Shannon–McMillan–Breiman theorem) or from random matrix theory (Theorem 1).

For certain purposes, it may be useful to know that the conditional entropy $H(Y\,|\,X)$ (i.e., $\lambda(Y\,|\,X)$) is smooth in the problem data. Here, "smooth in the problem data" means the Lyapunov exponent is differentiable with respect to the channel transition probabilities R, as well as the probability mass function for the output symbols $q(\cdot\,|\,C_j, C_{j+1}, X_j)$. Arnold *et al.* [4, Corollary 4.11] provide sufficient conditions under which the Lyapunov exponent $\lambda(Y\,|\,X)$ is analytic in the entries of the random matrices. However, in our case, perturbations in R and q simultaneously affect both the entries of the $G_{(X,Y)}$'s and the probability distribution generating them. Therefore, the results of [4] are difficult to apply in our setting. This coupling between the matrix entries for R and the mass function q has also been studied in [39], which examines continuity properties of entropy in several asymptotic regimes associated with the binary-symmetric channel. A recent result [27] actually shows that the entropy (or Lyapunov exponent) in this formulation is an analytic function of the problem data. Their condition for analyticity of entropy is identical to our condition in Theorem 6 for continuity of the stationary distribution for the conditional channel probabilities.

It is in fact remarkable that this relationship holds for $H(Y\,|\,X)$ at this level of generality (i.e., for any stationary and ergodic input sequence). The main reason we can state this result is that the amount of memory in the inputs is irrelevant to the computation of *conditional* entropy. However, at the current level of generality, there is no obvious way to compute $H(Y)$ itself. As a consequence, the mutual information rate $I(X,Y)$ cannot be computed, and calculation of capacity for the channel is infeasible. In Section III, we strengthen our hypotheses on X so as to ensure computability of mutual information in terms of Lyapunov exponents.

III. Markov Channels With Markov-Dependent Input Symbols

In this section, we prove a number of important connections between entropy, Lyapunov exponents, Markov chains, and hidden Markov models (HMMs). First, we present examples of channels that can be modeled using Markov-dependent inputs. We then show that any entropy quantity (and hence mutual information) for these channels can be represented as a Lyapunov exponent. In addition to proving the Lyapunov exponent connection with entropy, we must also develop a framework for computing such quantities and evaluating their properties. To this end, we show that symbol entropies for finite-state channels with Markov-dependent inputs can be computed as expectations with respect to the stationary distributions of a class of Markov chains. Furthermore, we will also show that in some cases the Markov chain of interest is an augmented version of the well-known channel prediction filter from HMM theory.

A. Channel Assumptions and Examples

In this section (and throughout the rest of this paper), we will assume the following.

B1: $C = (C_n\colon n \ge 0)$ satisfies A1.

B2: The input/output symbol pairs $\{(X_i, Y_i) : i \ge 0\}$ are conditionally independent given C, so that

$$P\,(X_0^n = x_0^n, Y_0^n = y_0^n |\, C) = \prod_{i=0}^{n} P(X_i = x_i, Y_i = y_i \,|\, C)$$

for $x_0^n \in \mathcal{X}^{n+1}, y_0^n \in \mathcal{Y}^{n+1}$.

B3: For each pair $(c_0, c_1) \in \mathcal{C}^2$, there exists a probability mass function $q(\cdot\,|\,c_0, c_1)$ on $\mathcal{X}$ x $\mathcal{Y}$ such that

$$P(X_i = x, Y_i = y \,|\, C) = q(x, y \,|\, C_i, C_{i+1}).$$

Again, the noncausal dependence of the symbols is introduced strictly for mathematical convenience. It is clear that typical causal channel models fit into this framework.

A number of important channel models are subsumed by B1–B3, in particular channels with ISI and dependent inputs. We now outline a number of channel examples.

Example 1: The Gilbert–Elliot channel is a special case in which $\mathcal{X}, \mathcal{Y}$, and $\mathcal{C}$ are all binary sets.

Example 2: Goldsmith and Varaiya [24] consider the special class of finite-state Markov channels for which the input symbols are i.i.d. with q in B1–B3 taking the form

$$q(x, y \,|\, c_0, c_1) = q_1(x) q_2(y \,|\, c_0, x)$$

for some functions q_1, q_2.

Example 3: Suppose that we wish to model a channel in which the input symbol sequence is Markov. Specifically, suppose that X is an aperiodic irreducible finite-state Markov chain on $\mathcal{X}$, independent of C. Assume that the output symbols $\{Y_n\colon n \ge 0\}$ are conditionally independent given (X, C), with

$$P(Y_i = y \,|\, X, C) = q(y \,|\, X_i, X_{i+1}, C_i, C_{i+1}).$$

To incorporate this model into our framework, we augment the channel state (artificially), forming $\tilde{C}_i = (C_i, X_i)$. Note that $\tilde{C} = (\tilde{C}_n\colon n \ge 0)$ is a finite-state irreducible Markov chain on $\tilde{\mathcal{C}} = \mathcal{C}$ x $\mathcal{X}$ under our assumptions. The triplet $(\tilde{C}, X, Y)$ then satisfies the requirements demanded of (C, X, Y) in B1–B3.

Example 4: In this example, we extend Example 3 to include the possibility of ISI. As above, we assume that X is an aperiodic irreducible finite-state Markov chain, independent of $\mathcal{C}$. Suppose that the output symbol sequence Y satisfies

$$P\left(Y_{n+1}=y\,|C_0^{n+1},X_0^{n+1},Y_0^n\right)=q(y\,|\,C_{n+1},X_{n+1},X_n)$$

with $q(y\,|\,c_1,x_1,x_0)>0$ for all $(y,c_1,x_1,x_0)\in\mathcal{Y}x\mathcal{C}$ x $\mathcal{X}^2$. To incorporate this model, we augment the channel state to include previous input symbols. Specifically, we set $\hat{C}_n=(C_n,X_n,X_{n-1})$ and use $(\hat{C},X,Y)$ to validate the requirements of B1–B3.

B. Entropies as Lyapunov Exponents

With the channel model described by B1–B3, each of the entropies $H(X)$, $H(Y)$, and $H(X,Y)$ turn out to be Lyapunov exponents for products of random matrices (up to a change in sign).

Proposition 1: For $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, let

$$\begin{aligned}G_x^X&=\left(G_x^X(c_0,c_1):c_0,c_1\in\mathcal{C}\right)\\G_y^Y&=\left(G_y^Y(c_0,c_1):c_0,c_1\in\mathcal{C}\right)\\G_{(x,y)}^{(X,Y)}&=\left(G_{(x,y)}^{(X,Y)}(c_0,c_1):c_0,c_1\in\mathcal{C}\right)\end{aligned}$$

be $|\mathcal{C}|$x$|\mathcal{C}|$ matrices with entries given by

$$\begin{aligned}G_x^X(c_0,c_1)&=R(c_0,c_1)\sum_y q(x,y\,|\,c_0,c_1)\\G_y^Y(c_0,c_1)&=R(c_0,c_1)\sum_x q(x,y\,|\,c_o,c_1)\\G_{(x,y)}^{(X,Y)}(c_0,c_1)&=R(c_0,c_1)q(x,y\,|\,c_0,c_1).\end{aligned}$$

Assume B1–B3. Then $H(X)=-\lambda(X)$, $H(Y)=-\lambda(Y)$, and $H(X,Y)=-\lambda(X,Y)$, where $\lambda(X)$, $\lambda(Y)$, and $\lambda(X,Y)$ are the Lyapunov exponents defined as the following limits:

$$\begin{aligned}\lambda(X)&=\lim_{n\to\infty}\frac{1}{n}\log\left\|G_{X_1}^XG_{X_2}^X\cdots G_{X_n}^X\right\|\text{ a.s.}\\\lambda(Y)&=\lim_{n\to\infty}\frac{1}{n}\log\left\|G_{Y_1}^YG_{Y_2}^Y\cdots G_{Y_n}^Y\right\|\text{ a.s.}\\\lambda(X,Y)&=\lim_{n\to\infty}\frac{1}{n}\log\left\|G_{(X_1,Y_1)}^{(X,Y)}\cdots G_{(X_n,Y_n)}^{(X,Y)}\right\|\text{ a.s.}\end{aligned}$$

The proof of the preceding proposition is virtually identical to the argument of Theorem 1, and is therefore omitted.

At this point, it is useful to provide a bit of intuition regarding the connection between Lyapunov exponents and entropy. Following the development of Section II we can write

$$P(X_1,\ldots,X_n)=rG_{X_1}^XG_{X_2}^X\cdots G_{X_n}^Xe \tag{4}$$

where r is the stationary distribution for the channel C. Using Proposition 1 and (3), we can interpret the Lyapunov exponent λ_X as the average exponential rate of growth for the probability of the sequence X. Since $P(X_1,\ldots,X_n)\to 0$ as $n\to\infty$ for any nontrivial sequence, the rate of growth will be negative. (If the probability of the input sequence does not converge to zero then $H(X)=0$.)

This view of the Lyapunov exponent facilitates a straightforward information-theoretic interpretation based on the notion of typical sequences. From Cover and Thomas [15], the typical set A_ϵ^n is the set of sequences $x_1,\ldots,x_n$ satisfying

$$2^{-n(H(X)+\epsilon)}\le P(X_1=x_1,\ldots,X_n=x_n)\le 2^{-n(H(X)-\epsilon)}$$

and $P(A_\epsilon^n)>1-\epsilon$ for n sufficiently large. Hence, we can see that, asymptotically, any observed sequence must be a typical sequence with high probability. Furthermore, the asymptotic exponential rate of growth of the probability for any typical sequence must be $-H(X)$ or $\lambda(X)$. This intuition will be useful in understanding the results presented in the next subsection where we show that $\lambda(X)$ can also be viewed as an expectation rather than an asymptotic quantity.

C. A Markov Chain Representation for Lyapunov Exponents

Proposition 1 establishes that the mutual information $I(X,Y)=H(X)+H(Y)-H(X,Y)$ can be easily expressed in terms of Lyapunov exponents, and that the channel capacity involves an optimization of the Lyapunov exponents relative to the input symbol distribution. However, the above random matrix product representation is of little use when trying to prove certain properties for Lyapunov exponents, nor does it readily facilitate computation. In order to address these issues, we will now show that the Lyapunov exponents of interest in this paper can also be represented as expectations with respect to the stationary distributions for a class of Markov chains.

From this point onward, we will focus our attention on the Lyapunov exponent $\lambda(X)$, since the conclusions for $\lambda(Y)$ and $\lambda(X,Y)$ are analogous. In much of the literature on Lyapunov exponents for i.i.d. products of random matrices, the basic theoretical tool for analysis is a particular continuous state space Markov chain [22]. Since our matrices are not i.i.d. we will use a slightly modified version of this Markov chain, namely

$$\begin{aligned}Z_n&=\left(\frac{wG_{X_1}^X\cdots G_{X_n}^X}{\left\|wG_{X_1}^X\cdots G_{X_n}^X\right\|},C_n,C_{n+1}\right)\\&=(\tilde{p}_n,C_n,C_{n+1}).\end{aligned}$$

Here, w is a $|\mathcal{C}|$-dimensional stochastic (row) vector, and the norm appearing in the definition of Z_n is any norm on $\Re^{|\mathcal{C}|}$. If we view $wG_{X_1}^X\cdots G_{X_n}^X$ as a vector, then we can interpret the first component of Z as the direction of the vector at time n. The second and third components of Z determine the probability distribution of the random matrix that will be applied at time n. We choose the normalized direction vector

$$\tilde{p}_n=\frac{wG_{X_1}^X\cdots G_{X_n}^X}{\left\|wG_{X_1}^X\cdots G_{X_n}^X\right\|} \tag{5}$$

rather than the vector itself because $wG_{X_1}^X\cdots G_{X_n}^X\to 0$ as $n\to\infty$, but we expect some sort of nontrivial steady-state behavior for the normalized version. The structure of Z should make sense given the intuition discussion in the previous subsection. If we want to compute the average rate of growth (i.e., the average one-step growth) for $\left\|wG_{X_1}^X\cdots G_{X_n}^X\right\|$ then all we

should need is a stationary distribution on the space of directions combined with a distribution on the space of matrices.

The steady-state theory for Markov chains on continuous state space, while technically sophisticated, is a highly developed area of probability. The Markov chain Z allows one to potentially apply this set of tools to the analysis of the Lyapunov exponent $\lambda(X)$. Assuming for the moment that Z has a steady-state Z_∞, we can then expect to find that

$$Z_n = (\tilde{p}_n, C_n, C_{n+1}) \Rightarrow Z_\infty \triangleq (\tilde{p}_\infty, C_\infty, \tilde{C}_\infty) \tag{6}$$

as $n \to \infty$, where $C_\infty, \tilde{C}_\infty \in \mathcal{C}, \tilde{p}_0 = w$ and

$$\tilde{p}_n \triangleq \frac{wG^X_{X_1} \cdots G^X_{X_n}}{\left\| wG^X_{X_1} \cdots G^X_{X_n} \right\|} = \frac{\tilde{p}_{n-1} G^X_{X_n}}{\left\| \tilde{p}_{n-1} G^X_{X_n} \right\|} \tag{7}$$

for $n \geq 1$. If w is positive, the same argument as that leading to (3) shows that

$$\frac{1}{n} \log \left\| G^X_{X_1} \cdots G^X_{X_n} \right\| - \frac{1}{n} \log \left\| wG^X_{X_1} \cdots G^X_{X_n} \right\| \to 0 \text{ a.s.} \tag{8}$$

as $n \to \infty$, which implies

$$\lambda(X) = \lim_{n \to \infty} \frac{1}{n} \log \left\| wG^X_{X_1} \cdots G^X_{X_n} \right\|. \tag{9}$$

Furthermore, it is easily verified that

$$\log \left\| wG^X_{X_1} \cdots G^X_{X_n} \right\| = \sum_{j=1}^{n} \log \left(\left\| \tilde{p}_{j-1} G^X_{X_j} \right\| \right). \tag{10}$$

Relations (8) and (10) together guarantee that

$$\lambda(X) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \log \left(\left\| \tilde{p}_{j-1} G^X_{X_j} \right\| \right) \text{ a.s.} \tag{11}$$

In view of (6), this suggests that

$$\lambda(X) = \sum_{x \in \mathcal{X}} \mathrm{E} \log \left(\left\| \tilde{p}_\infty G^X_x \right\| \right) R(C_\infty, \tilde{C}_\infty) q(x | C_\infty, \tilde{C}_\infty) \tag{12}$$

where $q(x \,|\, c_0, c_1) \triangleq \sum_y q(x, y \,|\, c_0, c_1)$. Recall the above discussion regarding the intuitive interpretation of Lyapunov exponents and entropy and suppose we apply the 1-norm, given by $||w||_1 \triangleq \sum_c |w(c)|$, in (12). Then the representation (12) computes the expected exponential rate of growth for the probability $P(X_1, \ldots, X_n)$, where the expectation is with respect to the stationary distribution of the continuous state space Markov chain Z.[1] Thus, assuming the validity of (6), computing the Lyapunov exponent effectively amounts to computing the stationary distribution of the Markov chain Z. Because of the importance of this representation, we will return to providing rigorous conditions guaranteeing the validity of such representations in Sections IV and V.

D. The Connection to HMMs

As noted above, Z is a Markov chain regardless of the choice of norm on $\Re^{|\mathcal{C}|}$. If we specialize to the 1-norm, it turns out that the first component of Z can be viewed as the prediction filter for the channel given the input symbol sequence X.

Proposition 2: Assume B1–B3, and let $w = r$, the stationary distribution of the channel C. Then, for $n \geq 0$ and $c \in \mathcal{C}$,

$$\tilde{p}_n(c) = P(C_{n+1} = c \,|\, X_1^n).$$

Proof: The result follows by an induction on n. For $n = 0$, the result is trivial. For $n = 1$, note that

$$\begin{aligned} P(C_2 = c \,|\, X_1) &= \frac{\sum_{c_0} r(c_0) R(c_0, c) q(X_1 \,|\, c_0, c)}{\sum_{c_0, c_1} r(c_0) R(c_0, c_1) q(X_1 \,|\, c_0, c_1)} \\ &= \frac{(rG^X_{X_1})(c)}{\left\| rG^X_{X_1} \right\|_1}. \end{aligned}$$

The general induction step follows similarly. □

It turns out that the prediction filter $(\tilde{p}_n \colon n \geq 0)$ is itself Markov, without appending (C_n, C_{n+1}) to $\tilde{p}_n$ as a state variable.

Proposition 3: Assume B1–B3 and suppose $w = r$. Then, the sequence $\tilde{p} = (\tilde{p}_n \colon n \geq 0)$ is a Markov chain taking values in the continuous state space $\mathcal{P} = \{w : w \geq 0, \|w\|_1 = 1\}$. Furthermore

$$\left\| \tilde{p}_n G^X_x \right\|_1 = P(X_{n+1} = x \,|\, X_1^n).$$

Proof: See Appendix II-A.

In view of Proposition 3, the terms appearing in the sum (10) have interpretations as conditional entropies, namely

$$-\mathrm{E} \log \left(\left\| \tilde{p}_{j-1} G^X_{X_j} \right\|_1 \right) = H \left(X_{j+1} \,|\, X_1^j \right) \tag{13}$$

so that the formula (11) for $\lambda(X)$ can be interpreted as the well-known representation for $H(X)$ in terms of the averaged conditional entropies

$$\begin{aligned} H(X) &= \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n) \\ &= \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} H(X_{j+1} \,|\, X_1^j) \\ &= -\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathrm{E} \log \left(\left\| \tilde{p}_{j-1} G^X_{X_j} \right\|_1 \right) \\ &= -\lambda(X). \end{aligned}$$

It should be noted that Blackwell originally proposed an expected value formulation similar to (12) utilizing a prediction filter in his 1957 paper [10]. There he considered a filter that conditioned on an *infinite* history of observations, such as $P(C_n = c \,|\, X_{-\infty}^n)$. His filter also contained the channel state as an augmenting variable to ensure that it was always a stationary Markov chain. By conditioning on the infinite past, Blackwell was able to remove many of the convergence issues that we discuss in the next section. He also provided rather strong conditions for uniqueness of the stationary distribution for his prediction filter. We contrast those with our results

[1] Note that while (12) holds for any choice of norm, the 1-norm provides the most intuitive interpretation.

in Section V, which are based on more intuitive and general conditions.

For the single-sided filter that does not assume an infinite history, an expected value representation of $H(X)$, similar in spirit to (12), is a well-known result in the HMM literature [18]. Note, however, that the analysis of the hidden Markov prediction filter $(\tilde{p}_n : n \geq 0)$ with $w = r$ is only a special case of the problem we consider here. First, the above conditional entropy interpretation of $\log(\|p_{j-1}G^X_{X_j}\|_1)$ holds only when we choose to use the 1-norm. Moreover, the above interpretations also require that we initialize $\tilde{p}$ with $\tilde{p}_0 = r$, the stationary distribution of the channel C (i.e., Proposition 3 does not hold). Hence, if we want to use an arbitrary initial vector we must use the multivariate process Z, which is always a Markov chain. In the next section, we introduce a new process that is similar to the prediction filter and permits analysis with any choice of norm and any initial distribution.

We note, in passing, that in [24] it is shown that the prediction filter can be non-Markov in certain settings. However, we can include these non-Markov examples in our Markov framework by augmenting the channel states as in Examples 3 and 4. Thus, our process $\tilde{p}$ for these examples can be Markov without violating the conclusions in [24].

IV. Computing the Lyapunov Exponent as an Expectation

In the previous section, we showed that the Lyapunov exponent $\lambda(X)$ can be directly computed as an expectation with respect to the stationary distribution of the Markov chain Z. However, in order to make this statement rigorous, we must first prove that Z in fact has a stationary distribution. Furthermore, we should also determine if the stationary distribution for Z is unique and if this distribution is a continuous function of the input symbol and channel transition probabilities.

As it turns out, the Markov chain Z with $Z_n = (\tilde{p}_n, C_n, C_{n+1})$ is a very cumbersome theoretical tool for analyzing many properties of Lyapunov exponents. The main difficulty is that we must carry around the extra augmenting variables (C_n, C_{n+1}) in order to make Z a Markov chain. Unfortunately, we cannot utilize the channel prediction filter $\tilde{p}$ alone since it is only a Markov chain when $\tilde{p}_0 = r$. In order to prove properties such as existence and uniqueness of a stationary distribution for a Markov chain, we must be able to characterize the Markov chain's behavior for any initial point.

In this section, we introduce a new Markov chain p, which we will refer to as the "$\mathcal{P}$-chain." It is closely related to the prediction filter $\tilde{p}$ and, in some cases, will be identical to the prediction filter. However, the Markov chain p possess one important additional property—it is always a Markov chain regardless of its initial point. The reason for introducing this new Markov chain is that the asymptotic properties of p are the same as those of the prediction filter $\tilde{p}$ (we show this in Section V), and the analysis of p is substantially easier than that of Z. Therefore, the results we are about to prove for p can be applied to $\tilde{p}$ and hence the Lyapunov exponent $\lambda(X)$.

A. The Channel $\mathcal{P}$-Chain

We will define the random evolution of the $\mathcal{P}$-chain using the following algorithm.

Algorithm A:

1) Initialize $n = 0$ and $p_0 = w \in \mathcal{P}$, where
$$\mathcal{P} = \{w : w \geq 0, \|w\|_1 = 1\}.$$
2) Generate $\tilde{X} \in \mathcal{X}$ from the probability mass function $(\|p_n G^X_x\|_1 : x \in \mathcal{X})$.
3) Set
$$p_{n+1} = \frac{p_n G^X_{\tilde{X}}}{\|p_n G^X_{\tilde{X}}\|_1}.$$
4) Set $n = n + 1$ and return to 2.

The output produced by Algorithm A clearly exhibits the Markov property, for any initial vector $w \in \mathcal{P}$. Let $p^w = (p^w_n : n \geq 0)$ denote the output of Algorithm A when $p_0 = w$. Proposition 3 proves that for $w = r, p^r$ coincides with the sequence $\tilde{p}^r = (\tilde{p}^r_n : n \geq 0)$, where $\tilde{p}^w = (\tilde{p}^w_n : n \geq 0)$ for $w \in \mathcal{P}$ is defined by the recursion (also known as the forward Baum equation)

$$\tilde{p}^w_n = \frac{w G^X_{X_1} \cdots G^X_{X_n}}{\left\| w G^X_{X_1} \cdots G^X_{X_n} \right\|_1} \tag{14}$$

where $X = (X_n : n \geq 1)$ is a stationary version of the input symbol sequence. Note that in the above algorithm the symbol sequence $\tilde{X}$ is determined in an unconventional fashion. In a traditional filtering problem, the symbol sequence X follows an exogenous random process and the channel state predictor uses the observed symbols to update the prediction vector. However, in Algorithm A, the probability distribution of the symbol $\tilde{X}_n$ *depends on the random vector* p_n, hence the symbol sequence $\tilde{X}$ is not an exogenous process. Rather, the symbols are generated according to a probability distribution determined by the state of the $\mathcal{P}$-chain. Proposition 3 establishes a relationship between the prediction filter $\tilde{p}^w$ and the $\mathcal{P}$-chain p^w when $w = r$. As noted above, we shall need to study the relationship for arbitrary $w \in \mathcal{P}$. Proposition 4 provides the key link.

Proposition 4: Assume B1–B3. Then, if $w \in \mathcal{P}$

$$p^w_n = \frac{w G^X_{X_1(w)} \cdots G^X_{X_n(w)}}{\left\| w G^X_{X_1(w)} \cdots G^X_{X_n(w)} \right\|_1}$$

where $X(w) = (X_n(w) : n \geq 1)$ is the input symbol sequence when C_1 is sampled from the mass function w. In particular

$$P(X_1(w) = x_1, \ldots, X_n(w) = x_n) = w G^X_{x_1} \cdots G^X_{x_n} e.$$

Proof: See Appendix II-B.

Indeed, Proposition 4 is critical to the remaining analysis in this paper and therefore warrants careful examination. In Algorithm A, the probability distribution of the symbol

$\tilde{X}_n$ depends on the state of the Markov chain $\tilde{p}_n^w$. This dependence makes it difficult to explicitly determine the joint probability distribution for the symbol sequence $\tilde{X}_1, \ldots, \tilde{X}_n$. Proposition 4 shows that we can take an alternative view of the $\mathcal{P}$-chain. Rather than generating the $\mathcal{P}$-chain with an endogenous sequence of symbols $\tilde{X}_1, \ldots, \tilde{X}_n$, we can use the exogenous sequence $X_1(w), \ldots, X_n(w)$, where the sequence $X(w) = (X_n(w): n \geq 1)$ is the input sequence generated when the channel is initialized with the probability mass function w. In other words, we can view the chain $\tilde{p}_n^w$ as being generated by a stationary channel C, whereas the $\mathcal{P}$-chain p_n^w is generated by a *nonstationary version* of the channel $C(w)$, using w as the initial channel distribution. Hence, the input symbol sequences for the Markov chains $\tilde{p}^w$ and p^w can be generated by two different versions of the same Markov chain (i.e., the channel). In Section V, we will use this critical property (along with some results on products of random matrices) to show that the asymptotic behaviors of $\tilde{p}^w$ and p^w are identical.

The stochastic sequence $\tilde{p}^w$ is the prediction filter that arises in the study of "HMMs." As is natural in the filtering theory context, the filter $\tilde{p}^w$ is driven by the exogenously determined observations X. On the other hand, it appears that p^w has no obvious filtering interpretation, except when $w = r$. However, for reasons discussed above, p^w is the more appropriate object for us to study. As is common in the Markov chain literature, we shall frequently choose to suppress the dependence on w, choosing to denote the Markov chain as $p = (p_n: n \geq 0)$.

B. The Lyapunov Exponent as an Expectation

Our goal now is to analyze the steady-state behavior of the Markov chain p and show that the Lyapunov exponent can be computed as an expectation with respect to p's stationary distribution. In particular, if p has a stationary distribution we should expect

$$H(X) = -\sum_{x \in \mathcal{X}} \mathrm{E} \log \left(\left\| p_\infty G_x^X \right\|_1 \right) \left\| p_\infty G_x^X \right\|_1 \tag{15}$$

where p_∞ is a *random vector* distributed according to p's stationary distribution.

As mentioned earlier in this section, the "channel $\mathcal{P}$-chain" p that arises here is closely related to the prediction filter $\tilde{p} = (\tilde{p}_n: n \geq 0)$ that arises in the study of "HMMs." A sizeable literature exists on steady-state behavior of prediction filters for HMMs. An excellent recent survey of the HMM literature can be found in Ephraim and Merhav [18]. However, this literature involves significantly stronger hypotheses than we shall make in this section, potentially ruling out certain channel models associated with Examples 3 and 4. We shall return to this issue in Section V, in which we strengthen our hypotheses to ones comparable to those used in the HMM literature. We also note that the Markov chain p, while closely related to $\tilde{p}$, requires somewhat different methods of analysis.

Theorem 2: Assume B1–B3 and let

$$\mathcal{P}^+ = \{w \in \mathcal{P}: w(c) > 0, c \in \mathcal{C}\}.$$

Then we have the following.

i) For any stationary distribution π of $p = (p_n: n \geq 0)$

$$H(X) \leq -\sum_{x \in \mathcal{X}} \int_{\mathcal{P}} \log \left(\left\| w G_x^X \right\|_1 \right) \left\| w G_x^X \right\|_1 \pi(dw).$$

ii) For any stationary distribution π satisfying $\pi(\mathcal{P}^+) = 1$

$$H(X) = -\sum_{x \in \mathcal{X}} \int_{\mathcal{P}} \log \left(\left\| w G_x^X \right\|_1 \right) \left\| w G_x^X \right\|_1 \pi(dw).$$

Proof: See Appendix II-C.

Note that Theorem 2 suggests that $p = (p_n: n \geq 0)$ may have multiple stationary distributions. The following example shows that this may indeed occur, even in the presence of B1–B3.

Example 5: Suppose $\mathcal{C} = \{1, 2\}$, and $\mathcal{X} = \{1, 2\}$, with

$$R = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

and

$$G_1^X = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \quad G_2^X = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}.$$

Then, both π_1 and π_2 are stationary distributions for p, where

$$\pi_1 \left(\left(\frac{1}{2}, \frac{1}{2} \right) \right) = 1$$

and

$$\pi_2((0,1)) = \pi_2((1,0)) = \frac{1}{2}.$$

Theorem 2 leaves open the possibility that stationary distributions with support on the boundary of $\mathcal{P}$ will fail to satisfy (15). Furstenberg and Kifer [22] discuss the behavior of $p = (p_n: n \geq 0)$ when p has multiple stationary distributions, some of which violate (15) (under an invertibility hypotheses on the G_x^X's). Theorem 2 also fails to resolve the question of existence of a stationary distribution for p. To remedy this situation we impose additional hypotheses:

B4: $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$.

B5: For each (x, y) for which $P(X_0 = x, Y_0 = y) > 0$, the matrix $G_{(x,y)}^{(X,Y)}$ is row-allowable (i.e., it has no row in which every entry is zero).

Theorem 3: Assume B1–B5. Then, $p = (p_n: n \geq 0)$ possesses a stationary distribution π.

Proof: See Appendix II-D.

As we shall see in the next section, much more can be said about the channel $\mathcal{P}$-chain $p = (p_n: n \geq 0)$ in the presence of strong positivity hypotheses on the matrices $\{G_x^X: x \in \mathcal{X}\}$. The Markov chain $p = (p_n: n \geq 0)$, as studied in this section, is challenging largely because we permit a great deal of sparsity in the matrices $\{G_x^X: x \in \mathcal{X}\}$. The challenges we face here are largely driven by the inherent complexity of the behavior that Lyapunov exponents can exhibit in the presence of such sparsity. We will alleviate these problems in the next section through additional assumptions on the aperiodicty of the channel as well as the conditional probability distributions on the input/output symbols.

V. The Stationary Distribution of the Channel $\mathcal{P}$-Chain Under Positivity Conditions

In this section, we introduce extra conditions that guarantee the existence of a unique stationary distribution for the Markov chains p and $\tilde{p}^r$. By necessity, the discussion in this section and the resulting proofs in the Appendix are rather technical. Hence we will first summarize the results of this section and then prove further details.

The key assumption we will make in this section is that the probability of observing any symbol pair (x, y) is strictly positive for any valid channel transition (i.e., if $R(c_0, c_1)$ is positive)—recall that the probability mass function for the input/output symbols $q(x, y \mid c_0, c_1)$ depends on the channel transition rather than just the channel state. This assumption, together with aperiodicity of R, will guarantee that the random matrix product $G^X_{X_1(w)} \cdots G^X_{X_n(w)}$ can be split into a product of *strictly positive* random matrices. We then exploit the fact that strictly positive matrices are strict contractions on

$$\mathcal{P}^+ = \{w \in \mathcal{P}\colon w(c) > 0, c \in \mathcal{C}\}$$

for an appropriate distance metric. This contraction property allows us to show that both the prediction filter $\tilde{p}^r$ and the $\mathcal{P}$-chain p converge exponentially fast to the same limiting random variable. Hence, both p and $\tilde{p}^r$ have the same unique stationary distribution that we can use to compute the Lyapunov exponent $\lambda(X)$. This result is stated formally in Theorem 5. In Theorem 6, we show that the stationary distribution of the $\mathcal{P}$-chain (and hence the prediction filter) is a continuous function of both the transition matrix R and the symbol probabilities $q(x, y \mid c_0, c_1)$.

A. The Contraction Property of Positive Matrices

We assume here the following.

B6: The transition matrix R is aperiodic.

B7: For each $(c_0, c_1, x, y) \in \mathcal{C}^2 \times \mathcal{X} \times \mathcal{Y}$, $q(x, y \mid c_0, c_1) > 0$ whenever $R(c_0, c_1) > 0$.

Under B6–B7, all the matrices

$$\left\{G^X_x, G^Y_y, G^{(X,Y)}_{(x,y)}\colon x \in \mathcal{X}, y \in \mathcal{Y}\right\}$$

exhibit the same (aperiodic) sparsity pattern as R. That is, the matrices have the same pattern of zero and nonzero elements. Note that under B1 and B6, R^l is strictly positive for some finite value of l. So

$$H_j = G^X_{X_{(j-1)l+1}} \cdots G^X_{X_{jl}}$$

is strictly positive for $j \geq 0$. The key mathematical property that we shall now repeatedly exploit is the fact that positive matrices are contracting on $\mathcal{P}^+$ in a certain sense.

For $v, w \in \mathcal{P}^+$, let

$$d(v, w) = \log\left(\frac{\max_c (v(c)/w(c))}{\min_c (v(c)/w(c))}\right).$$

The distance $d(v, w)$ is called "Hilbert's projective distance" between v and w, and is a metric on $\mathcal{P}^+$; see page 90 of Seneta [46]. For any nonnegative matrix T, let

$$\tau(T) = \frac{1 - \theta(T)^{-1/2}}{1 + \theta(T)^{-1/2}},$$

where

$$\theta(T) = \max_{c_0, c_1, c_2, c_3} \left(\frac{T(c_0, c_3)T(c_1, c_4)}{T(c_0, c_4)T(c_1, c_3)}\right).$$

Note that $\tau(T) < 1$ if T is strictly positive (i.e., if all the elements of T are strictly positive).

Theorem 4: Suppose $v, w \in \mathcal{P}^+$ are row vectors. Then, if T is strictly positive

$$d(vT, wT) \leq \tau(T) d(v, w).$$

For a proof, see Seneta [46, pp. 100–110]. The quantity $\tau(T)$ is called "Birkhoff's contraction coefficient."

Our first application of this idea is to establish that the asymptotic behavior of the channel $\mathcal{P}$-chain p and the prediction filter $\tilde{p}$ coincide. Note that for $n \geq l, \tilde{p}^r_n$ and $\tilde{p_n}^w$ both lie in $\mathcal{P}^+$, so $d(\tilde{p}^r_n, \tilde{p}^w_n)$ is well defined for $n \geq l$. Proposition 5 will allow us to show that $\tilde{p}^w = (\tilde{p}^w_n : n \geq 0)$ has a unique stationary distribution. Proposition 6 will allow us to show that $p^w = (p^w_n\colon n \geq 0)$ must have the same stationary distribution as $\tilde{p}^w$.

Proposition 5: Assume B1–B4 and B6–B7. If $w \in \mathcal{P}$, then

$$d(\tilde{p}^r_n, \tilde{p}^w_n) = \mathrm{O}(e^{-\alpha n}) \text{ a.s.}$$

as $n \to \infty$, where

$$\alpha \triangleq -(\log \beta)/l$$
$$\beta \triangleq \max\left\{\tau\left(G^X_{x_1} \cdots G^X_{x_l}\right) : p(X_1 = x_1, \ldots, X_l = x_l) > 0\right\} < 1.$$

Proof: The proof follows a greatly simplified version of the proof for Proposition 6, and is therefore omitted.

Proposition 6: Assume B1–B4 and B6–B7. For $w \in \mathcal{P}$, there exists a probability space upon which

$$d(p^w_n, \tilde{p}^r_n) = \mathrm{O}(e^{-\alpha n}) \text{ a.s., as } n \to \infty.$$

Proof: See Appendix II-E.

The Proof of Proposition 6 relies on Proposition 4 and a coupling arguement that we will summarize here. Recall from Proposition 4 that we can view $\tilde{p}^r_n$ and p^w_n as being generated by a stationary and nonstationary versions of the channel C, respectively. The key idea is that the nonstationary version of the channel will eventually couple with the stationary version. Furthermore, the nonstationary version of the symbol sequence $X(w)$ will also couple with the stationary version X. Once this coupling occurs, say at time $T < \infty$, the symbol sequences $(X_n(w)\colon n > T)$ and $(X_n\colon n > T)$ will be identical. This means that for all $n > T$, the matrices applied to $\tilde{p}^r_n$ and p^w_n

will also be identical. This allows us to apply the contraction result from Theorem 4 and complete the proof.

B. A Unique Stationary Distribution for the Prediction Filter and the $\mathcal{P}$-Chain

We will now show that there exists a limiting random variable p^*_∞ such that $\tilde{p_n}^r \Rightarrow p^*_\infty$ as $n \to \infty$. In view of Propositions 5 and 6, this will ensure that for each $w \in \mathcal{P}, p^w_n \Rightarrow p^*_\infty$ as $n \to \infty$. To prove this result, we will use an idea borrowed from the theory of "random iterated functions"; see Diaconis and Freedman [17]. Let $X = (X_n\colon -\infty < n < \infty)$ be a doubly infinite stationary version of the input symbol sequence, and put $\chi_n = X_{-n}$ for $n \in \mathcal{Z}$. Then

$$rG^X_{X_1} \cdots G^X_{X_n} \stackrel{\mathcal{D}}{=} rG^X_{\chi_n} \cdots G^X_{\chi_1} \tag{16}$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. Put $p^*_0 = r$ and

$$p^*_n = \frac{rG^X_{\chi_n} \cdots G^X_{\chi_1}}{\left\| rG^X_{\chi_n} \cdots G^X_{\chi_1} \right\|}$$

for $n \geq 0$, and

$$H^*_n = G^X_{\chi_{nl}} \cdots G^X_{\chi_{(n-1)l+1}}.$$

Then

$$\begin{aligned} d(p^*_{nl}, p^*_{(n-1)l}) &= d(rH^*_n H^*_{n-1} \cdots H^*_1, rH^*_{n-1} \cdots H^*_1) \\ &\leq \tau(H^*_{n-1} \cdots H^*_1) d(rH^*_n, r) \\ &\leq \beta^n d(rH^*_n, r) \\ &\leq \beta^n d_* \end{aligned}$$

where

$$d_* = \max\{d(rG^X_{x_1} \cdots G^X_{x_l}, r) : P(X_1 = x_1, \ldots, X_l = x_l) > 0\}.$$

It easily follows that $(p^*_n\colon n \geq 0)$ is a.s. a Cauchy sequence, so there exists a random variable p^*_∞ such that $p^*_n \to p^*_\infty$ a.s. as $n \to \infty$. Furthermore

$$d(p^*_\infty, r) \leq (1-\beta)^{-1} d_*. \tag{17}$$

The constant d_* can be bounded in terms of easier-to-compute quantities. Note that

$$\begin{aligned} d\left(rG^X_{x_1} G^X_{x_2}, r\right) &\leq d\left(rG^X_{x_1} G^X_{x_2}, rG^X_{x_2}\right) + d\left(rG^X_{x_2}, r\right) \\ &\leq \tau\left(G^X_{x_2}\right) d\left(rG^X_{x_1}, r\right) + d\left(rG^X_{x_2}, r\right) \\ &\leq 2d^* \end{aligned}$$

where $d^* = \max\{d(rG^X_x, r) : x \in \mathcal{X}\}$. Repeating the argument $l-2$ additional times yields the bound $d_* \leq ld^*$. The above argument proves parts ii) and iii) of the following result.

Theorem 5: Assume B1–B4 and B6–B7. Let $K = \{w : d(w, r) \leq (1-\beta)^{-1} ld^*\} \subseteq \mathcal{P}^+$. Then we have the following.

i) $p = (p_n : n \geq 0)$ has a unique stationary distribution π.
ii) $\pi(K) = 1$ and $\pi(\cdot) = P(p^*_\infty \in \cdot)$.
iii) For each $w \in \mathcal{P}$, $p^w_n \Rightarrow p^*_\infty$ as $n \to \infty$.
iv) K is absorbing for $(p_{ln} : n \geq 0)$, in the sense that $P(p^w_{ln} \in K) = 1$ for $n \geq 0$ and $w \in K$.

Proof: See Appendix II-F for the proofs of parts i) and iv), parts ii) and iii) are proved above.

Applying Theorem 2, we may conclude that under B1–B4 and B6–B7, the channel $\mathcal{P}$-chain has a unique stationary distribution π on $\mathcal{P}^+$ satisfying

$$H(X) = -\sum_{x \in \mathcal{X}} \int_{\mathcal{P}} \log\left(\left\|wG^X_x\right\|_1\right) \left\|wG^X_x\right\|_1 \pi(dw). \tag{18}$$

It is interesting to note that Blackwell also provided uniqueness conditions for the stationary distribution of his infinite memory prediction filter in [10]. There he required that the rows of the channel transition matrix R be nearly identical with no element very near zero. The constraints we present in this section are much more general as we do not require either of Blackwell's properties. Perhaps even more interesting is that Blackwell conjectured that irreducibility of R should be sufficient to guarantee uniqueness, though he did not prove it. From our preceding results it appears this conjecture was close to the mark. Rather than an irreducibility constraint on R, we show that (assumptions B6 and B7) an irreducibility constraint on the family of random matrices $G_{(X,Y)}$ is required in order to guarantee uniqueness of the stationary distribution for our prediction filter.

We can also use our Markov chain machinery to establish continuity of the stationary distribution for the the $\mathcal{P}$-chain as a function of R and q. Such a continuity result is of theoretical importance in optimizing the mutual information between X and Y, or when computing functions of the channel estimator in HMMs. The following theorem generalizes the continuity result of Goldsmith and Varaiya [24] obtained in the setting of i.i.d. input symbol sequences.

Theorem 6: Assume B1–B4 and B6–B7. Suppose that $(R_n\colon n \geq 1)$ is a sequence of transition matrices on $\mathcal{C}$ for which $R_n \to R$ as $n \to \infty$. Also, suppose that for $n \geq 1, q_n(\cdot \mid c_0, c_1)$ is a probability mass function on $\mathcal{X} \times \mathcal{Y}$ for each $(c_0, c_1) \in \mathcal{C}^2$ and that $q_n \to q$ as $n \to \infty$. If π_n is the stationary distribution of the $\mathcal{P}$-chain associated with the channel model characterized by (R_n, q_n), then $\pi_n \Rightarrow \pi$ as $n \to \infty$.

Proof: See Appendix II-K.

We should note here that continuity of the entropies $H(X), H(Y)$, and $H(X,Y)$ are easily shown using standard bounds in Cover and Thomas [15, Theorem 4.4.1]. The result in Theorem 6 proves continuity of $E_{\pi_n}[h(\cdot)]$ for *any* continuous function $h : \mathcal{P} \to \Re$. These results essentially require that the stationary distribution for the $\mathcal{P}$-chain have support strictly on the interior of the simplex. It is interesting to note that entropy is a rather well-behaved function in this regard and is still continuous when that distribution has support on the simplex boundary. In [27], it is shown that our condition for continuity of a general function is also required to prove analyticity of entropy.

VI. Numerical Methods for Computing Entropy

In this section, we discuss numerical methods for computing the entropy $H(X)$. In the first subsection, we will discuss simulation-based methods for computing sample entropy. Recently, several authors [2], [31], [44], [1] have proposed similar simulation algorithms for entropy computation. However, a number of important theoretical and practical issues regarding simulation-based estimates for entropy remain to be addressed. In particular, there is currently no general method for computing confidence intervals for simulated entropy estimates. Furthermore, there is no method for determining how long a simulation must run in order to reach "steady state." We will summarize the key difficulties surrounding these issues below. In Appendix I, we present a new CLT for sample entropy. This new theorem allows us to compute rigorous confidence intervals for simulated estimates of entropy. We also present a method for computing the initialization bias in entropy simulations, which together with the confidence intervals, allows us to determine the appropriate run time of a simulation.

In the second subsection, we present two methods for directly computing the entropy $H(X)$. We first present asymptotically tight upper and lower bounds for the entropy in terms of matrix norms. These bounds are likely to be of more use in proof techniques rather than actual computation since their computational complexity is quite high. The second method develops a discrete approximation to the Markov chain p and its stationary distribution. We show that the discrete approximation for the stationary distribution can be used to approximate $H(X)$. We also show that the approximation for $H(X)$ converges to the true value of $H(X)$ as the discretization intervals for p become finer.

In general, none of these computational methods (simulation based or direct) should necessarily be considered superior. While the simulation methods may have issues with initialization bias (thereby requiring extremely long run times), the complexity of the direct methods increases exponentially with the number of states in the channel. In fact, it is well known that Lyapunov exponents can be extremely difficult to compute [23], [50], [4] and the computational performance of simulation-based methods versus direct computation can be highly model dependent [20]. Indeed, Tsitsiklis and Blondel [50] study computational complexity issues associated with calculating Lyapunov exponents (and hence entropies for our class of channel models). They prove that, in general, Lyapunov exponents are not algorithmically approximable. However, their class of problem instances contain nonirreducible matrices. Consequently, in the presence of the irreducibility assumptions made in this paper, the question of computability remains open.

A. Simulation-Based Computation of the Entropy

One consequence of Theorem 1 in Section III is that we can use simulation to calculate our entropy rates by applying Algorithm A and using the process $\tilde{p}$ to create the following estimator

$$H_n(X) = -\frac{1}{n}\sum_{j=0}^{n-1} \log\left(\left\|\tilde{p}_j^r G^X_{X_{j+1}}\right\|_1\right). \tag{19}$$

Although the authors of [2], [31], [44], [1] did not make the connection to Lyapunov exponents and products of random matrices, they propose a similar version of this simulation-based algorithm in their work. More generally, a version of this simulation algorithm is a common method for computing Lyapunov exponents in chaotic dynamic systems literature [20]. Indeed, as noted both in [1] and [20], simulation is often the only option available for this computation problem due to the high complexity of direct computation methods. When applying simulation to this problem we must consider two important theoretical questions:

1) "How long should we run the simulation?"
2) "How accurate is our simulated estimate?"

In general, there exists a well-developed theory for answering these questions when the simulated Markov chain is "well behaved." For continuous-state space Markov chains such as $\tilde{p}$ and p the term "well behaved" usually means that the Markov chain is Harris recurrent (see [37] for the theory of Harris chains). The key condition required to show that a Markov chain is Harris recurrent is the notion of ϕ-irreducibility. Consider the Markov chain $p = (p_n : n \geq 0)$ defined on the space $\mathcal{P}$ with Borel sets $\mathcal{B}(\mathcal{P})$. Define τ_A as the first return time to the set $A \in \mathcal{P}$. Then, the Markov chain $p = (p_n : n \geq 0)$ is ϕ-irreducible if there exists a nontrivial measure ϕ on $\mathcal{B}(\mathcal{P})$ such that for every state $w \in \mathcal{P}$

$$\phi(A) > 0 \Rightarrow P_w(\tau_A < \infty) > 0. \tag{20}$$

However,[2] the Markov chains p and $\tilde{p}$ are never irreducible, as illustrated by the following example.

Suppose we wish to use simulation to compute the entropy of an output symbol process from a finite-state channel. Further, suppose that the output symbols are binary, hence the random matrices $G^Y_{Y_n}$ can only take two values, say G^Y_0 and G^Y_1. corresponding to output symbols 0 and 1, respectively. Suppose we initialize $\tilde{p}_0 = r$ and examine the possible values for $\tilde{p}_n$. Notice that for any n, the random vector $\tilde{p}_n$ can take on only

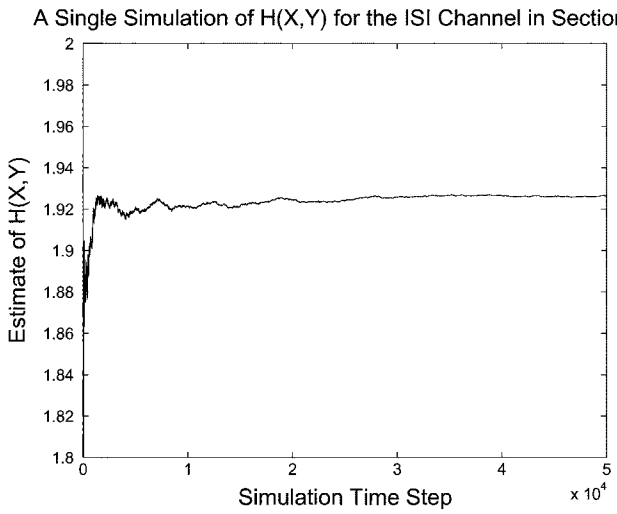

Fig. 1. A Single simulation trace from time 0 to 50 000. The estimate is $\lambda_{XY} = -H(X,Y)$ for the ISI channel we consider in Section VII.

[2]In [47], the authors assumed the augmented filter process Z was Harris recurrent. Though that publication and a following extended version gained traction in the Information Theory community, it should be noted that those results are incorrect.

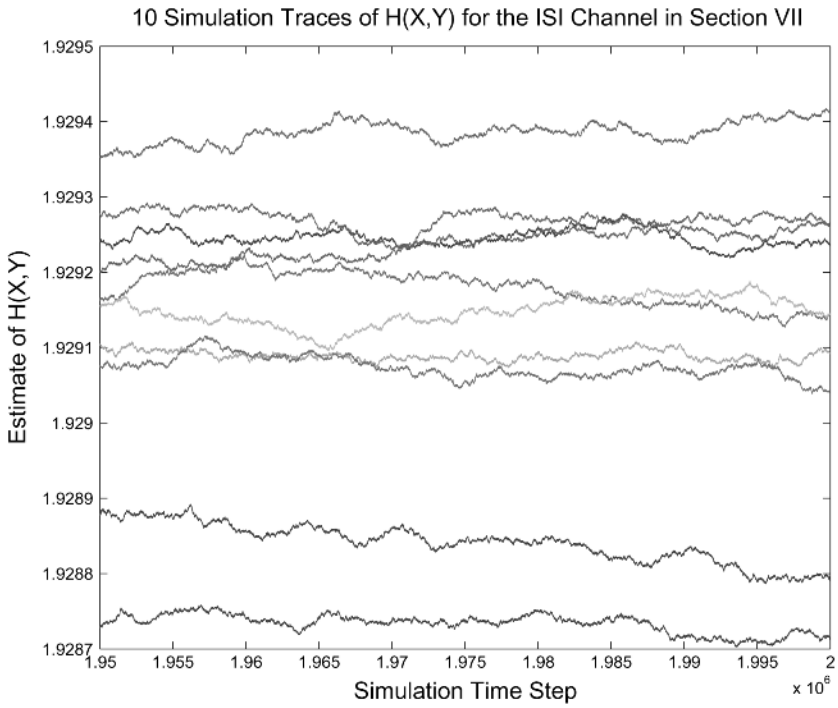

Fig. 2. Ten simulation traces from time 1,950,000 to 2,000,000. The estimate is $\lambda_{XY} = -H(X,Y)$ for the ISI channel we consider in Section VII.

a finite number of values, where each possible value is determined by one of the n-length permutations of the matrices G_0^Y and G_1^Y, and the initial condition $\tilde{p}_0$. One can easily find two initial vectors belonging to $\mathcal{P}$ for which the supports of their corresponding $\tilde{p}_n$'s are disjoint for all $n \geq 0$. This contradicts (20). Hence, the Markov chain $\tilde{p}$ has infinite memory and is not irreducible.

This technical difficulty means that we cannot apply the standard theory of simulation for continuous state-space Harris recurrent Markov chains to this problem. The authors of [44], [2], [34] note an important exception to this problem for the case of ISI channels with Gaussian noise. When Gaussian noise is added to the output symbols the random matrix $G_{Y_n}^Y$ is selected from a continuous population. In this case, the Markov chain $\tilde{p}$ is in fact irreducible and standard theory applies. However, since we wish to simulate any finite-state channel, including those with finite symbol sets, we cannot appeal to existing Harris chain theory to answer the two questions raised earlier regarding simulation-based methods.

Given the infinite memory problem discussed above we should pay special attention to the first question regarding simulation length. In particular, we need to be able to determine how long a simulation must run for the Markov chain $\tilde{p}$ to be "close enough" to steady state. The bias introduced by the initial condition of the simulation is known as the "initial transient," and for some problems its impact can be quite significant. For example, in the Numerical Results section of this paper, we will compute mutual information for two different channels. Using the above simulation algorithm, we estimated $\lambda_{XY} = -H(X,Y)$ for the ISI channel model in Section VII. Figs. 1 and 2 contain graphs of several traces taken from our simulations at different time intervals and resolutions.

The first figure shows a single sample path starting from time 0 to time 50 000. From this perspective it certainly appears that the sample path has converged (note that Fig. 2 in [1] shows similar behavior). However, closer examination suggests that this may not be the case. Fig. 2 shows ten sample paths taken from independent simulations after 2 000 000 iterations. Here we can see the simulation traces have minor fluctuations along each sample path. Furthermore, the variations in each trace are smaller than the distance between traces. This illustrates the potential numerical problems that arise when using simulation to compute Lyapunov exponents. Even though the traces in Figs. 1 and 2 are close in a relative sense, we have no way of determining which trace is "the correct" one or if any of the traces have even converged at all. Perhaps, even after 2 000 000 channel iterations, we are still stuck in an initial transient. Indeed, this type of behavior is frequently observed in Lyapunov exponent calculation, see Fig. 2 in [20] for an example.

In Appendix I, we develop a rigorous method for computing bounds on the initialization bias. This allows us to compute an explicit (although possibly very conservative) bound on the time required for the simulation to reach steady state. We also present a less conservative but more computationally intensive simulation-based bound in the same section.

Another means for addressing the uncertainty in simulated estimates is to develop confidence intervals. In order to produce confidence intervals we need access to a CLT for the sample entropy $H_n(X)$. Unfortunately, since $\tilde{p}^r$ is not irreducible we cannot apply the standard CLT for functions of Harris recurrent Markov chains. Therefore, in the first section of Appendix I we develop a new functional CLT for the sample entropy of finite-state channels. The "functional" form of the CLT implies the ordinary CLT. However, it also provides some stronger results which assist us in creating confidence intervals for simulated estimates of entropy. Using the techniques developed in Appendix A, we plot a 95% confidence interval for the simulation example in Fig. 3. If the simulation has reached steady state,

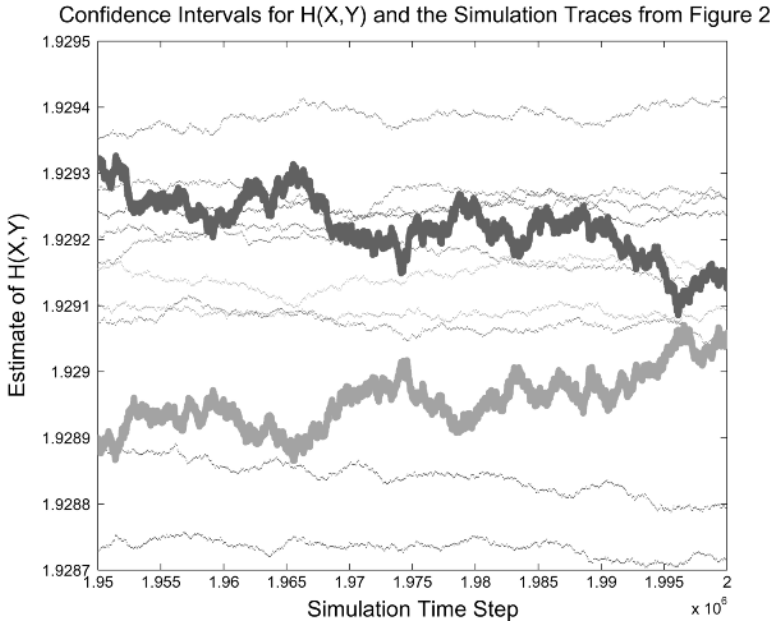

Fig. 3. Confidence interval for the middle trace in Fig. 2. The 95% confidence interval is plotted with bold lines. Notice that half of the traces fall outside the 95% interval.

and the functional CLT can be applied, then we should expect 95% of the traces to fall within the interval. In Fig. 3, the confidence interval is for the middle simulation trace and is shown in bold; the original traces from Fig. 2 are shown as dotted lines. Notice that five of the ten simulated traces fall outside the 95% confidence interval after 2 000 000 iterations. Moreover, if we chose to plot a confidence interval for the lower-most trace then eight of ten traces would fall outside the interval. Even after 2 000 000 iterations we still have not removed the initialization bias from the simulation and will have to run it for much longer to achieve a rigorous answer. In fact, the results in Appendix I-B suggest that we should run the above example for over 8 000 000 iterations to remove most of the initialization bias. Alternatively, if we believe the estimators in Fig. 3 are derived from a stationary sequence and a CLT applies; then we should expect the traces to converge on the order of $\sqrt{n}$ [28]. Clearly, this is not the case and the reason is that the sequences have not coupled with their stationary versions yet. Given the technical nature of the remaining discussion on simulation based estimates of entropy we direct the reader to Appendix I for the details on this topic.

We should note that in many practical applications the above convergence issue may not be a significant problem. It is likely that the traces in Figs. 1–3 are close to the correct answer. Specifically, it appears that we have achieved three significant figures of accuracy, *provided* that we do not converge to a new equilibrium when the initial transient is removed. It is unlikely that this pathology will occur, but we cannot rule it out. To make a rigorous probabilistic statement regarding the value of entropy we require the CLT and confidence interval methodology described herein. In addition, the convergence problem noted in Fig. 3 can cause significant accuracy issues when estimating the mutual information of a "low-SNR" channel. Estimating the mutual information involves computing the difference between estimates of $H(Y)$ and $H(Y \mid X)$. In the low-SNR regime we should expect these entropies to be very close in value (i.e., the mutual information will be small). Hence, the small relative errors observed in Figs. 2 and 3 can cause significant absolute errors in the estimated mutual information. In these cases, a direct computation algorithm may prove useful. In the next subsection, we discuss an algorithm that allows us to directly compute the entropy rates for a Markov channel—thus avoiding many of the convergence problems arising from a simulation based algorithm (at the potential cost of computational complexity).

B. Bounds and Direct Computation of the Entropy

Recall that if $g(w,x) = \log(1/||wG_x^X||_1)$, then (15) shows that

$$\mathrm{E}g(\tilde{p}_{n-1}^r, X_n) = H\left(X_{n+1} \,|X_1^n\right). \tag{21}$$

But the stationarity of X implies that $H(X_{n+1} \mid X_1^n) \searrow H(X)$ as $n \to \infty$; see Cover and Thomas [15]. So

$$H(X) \leq \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}g\left(\tilde{p}_{j-1}^r, X_j\right) \tag{22}$$

for $n \geq 1$. Note that this is precisely the bound from [15, Theorem 4.4.1] expressed as a matrix norm. For small values of n, this upper bound can be numerically computed by summing over all possible paths for X_1^n. We note, in passing, that Theorem 5 shows that $H(X) \leq \sup\{g(w) : w \in K\}$. An additional upper bound on $H(X)$ can be obtained from the existing general theory on lower bounds for Lyapunov exponents; see, for example, Key [32].

We obtain a new lower bound on the entropy $H(X)$ through an upper bound on the Lyapunov exponent $\lambda(X)$. According to Theorem 1, we therefore have the matrix norm lower bound

$$H(X) \geq -\frac{1}{n}\mathrm{E}\log\left\|G_{X_1}^X \cdots G_{X_n}^X\right\| \tag{23}$$

for $n \geq 1$. As in the case of (22), this lower bound can be numerically computed for small values of n by summing over all possible paths for X_1^n. Observe that both our upper bound and lower bound converge to $H(X)$ as $n \to \infty$, so that our bounds on $H(X)$ are asymptotically tight. From [15], the upper bound is guaranteed to converge monotonically to $H(X)$, but no monotonicity guarantee exists for the lower bound. In both cases, the computational complexity required to compute these bounds for large values of n can be quite high. Hence, these bounds are likely to be more useful in proof techniques or in computing probability one bounds on entropy. An additional upper bound on the Lyapunov exponent (i.e., lower bound on entropy) can be found in [23].

We conclude this section with a discussion of a second direct approximation scheme for entropy. This numerical discretization scheme computes $H(X)$ by approximating the stationary distribution of p via that of an appropriately chosen finite-state Markov chain. Specifically, for $n \geq 1$, let $E_{1n}, \ldots, E_{nn}$ be a partition of $\mathcal{P}$ such that

$$\sup_{1\leq j\leq n} \sup\{\|w_1 - w_2\| : w_1, w_2 \in E_{jn}\} \to 0 \tag{24}$$

as $n \to \infty$. For each set E_{in}, choose a representative point $w_{\mathrm{in}} \in E_{\mathrm{in}}$. Approximate the channel $\mathcal{P}$-chain via the Markov chain $(p_{n,i} : i \geq 0)$, where

$$P(p_{n,1} = w_{jn} \,|\, p_{n,0} = w) = P(p_1 \in E_{jn} \,|\, p_0 = w_{\mathrm{in}})) \tag{25}$$

for $w \in E_{\mathrm{in}}, 1 \leq i, j \leq n$. Then, $p_{ni} \in \mathcal{P}_n \triangleq \{w_{1n}, \ldots, w_{nn}\}$ for $i \geq 1$. Furthermore, any stationary distribution π_n of $(p_{ni} : i > 0)$ concentrates all its mass on $\mathcal{P}_n$. Its mass function $(\pi_n(w_{ni}) : 1 \leq i \leq n)$ must satisfy the finite linear system

$$\pi_n(w_{nj}) = \sum_{i=1}^{n} \pi_n(w_{ni}) P(p_1 \in E_{jn} \,|\, p_0 = w_{\mathrm{in}}) \tag{26}$$

for $1 \leq j \leq n$. Once π_n has been computed, one can approximate $H(X)$ by

$$\underline{H}_n(X) = \sum_{i=1}^{n} g(w_{ni})\pi_n(w_{ni}). \tag{27}$$

The following theorem proves that $\underline{H}_n(X)$ can provide a good approximation to $H(X)$ when n is large enough.

Theorem 7: Assume B1–B4 and B6–B7. Then, $\underline{H}_n(X) \to H(X)$ as $n \to \infty$.

Proof: See Appendix II-J.

Note that Theorem 5 asserts that the set $K \in \mathcal{P}$ is absorbing for $(p_{n,l} : n \geq 0)$. Thus, we could also compute $H(X)$ numerically by forming a discretization for the l-step "skeleton chain" $(p_{n,l} : n \geq 0)$ over K only. This has the advantage of shrinking the set over which the discretization is made, but at the cost of having to discretize the l-step transition structure of the channel $\mathcal{P}$-chain.

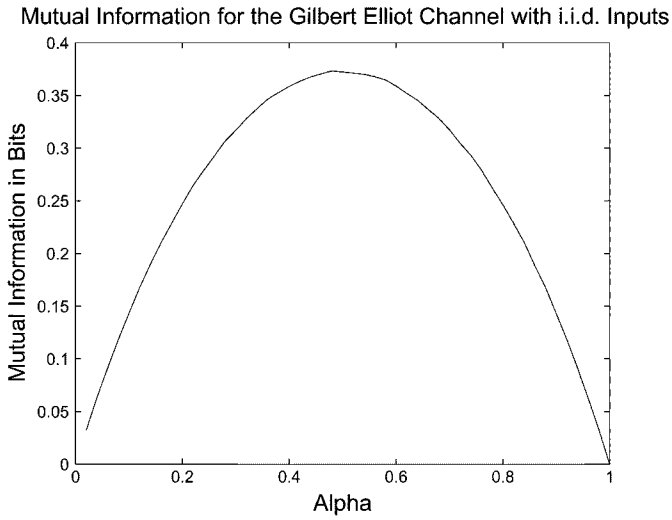

Fig. 4. Mutual Information of the Gilbert–Elliot channel with i.i.d. inputs.

VII. Numerical Examples

In this section, we present two numerical examples. The first example examines the mutual information for a Gilbert–Elliot channel with i.i.d. and Markov-modulated inputs. We know that i.i.d. inputs are optimal for this channel, so we should see no difference between the maximum mutual information for i.i.d. inputs and Markov-modulated inputs. Hence, we can view the first example as a check to ensure that our theory and algorithm are working properly. The second example considers i.i.d. and Markov modulated inputs for an ISI channel. In this case, we do see a difference in the maximum mutual information achieved by the different inputs.

A. Gilbert–Elliot Channel

The Gilbert–Elliot channel [48] is modeled by a simple two-state Markov chain with one "good" state and one "bad" state. In the good (resp., bad) state the probability of successfully transmitting a bit is p_G (resp., p_B). We use the good/bad naming convention for the states since $p_G > p_B$. The transition matrix for our example channel Z is

$$P = \begin{vmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{3}{4} \end{vmatrix}.$$

We consider two different types of inputs for this channel. The first case is that of i.i.d. inputs. In every time slot we set the input symbol to 0 with probability α and 1 with probability $1 - \alpha$. The graph in Fig. 4 plots the mutual information as α ranges from 0 to 1.

Next we examine the mutual information for the Gilbert–Elliot channel using Markov-modulated inputs. We define a second two-state Markov chain with transition matrix

$$Q = \begin{vmatrix} \frac{7}{8} & \frac{1}{8} \\ \frac{1}{2} & \frac{1}{2} \end{vmatrix}$$

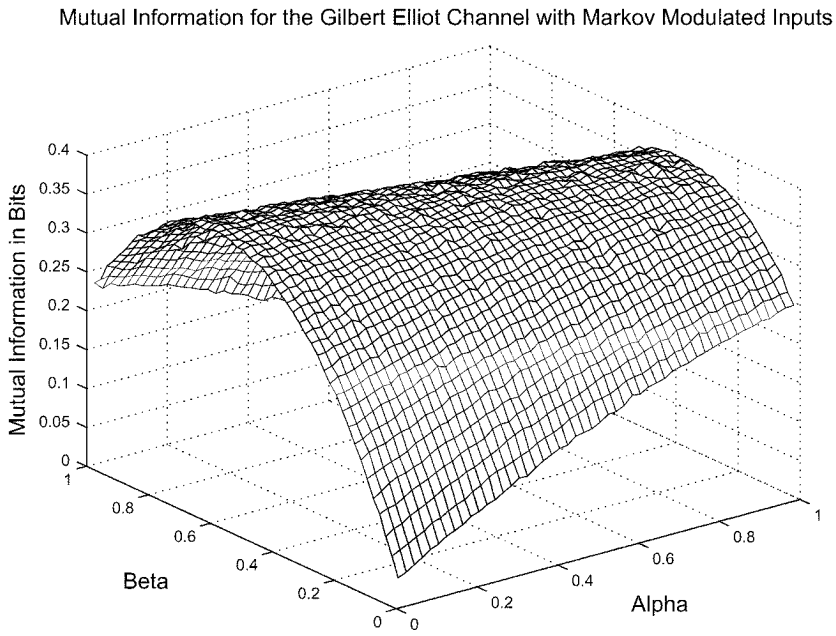

Fig. 5. Mutual information of the Gilbert–Elliot channel with Markov-modulated inputs.

TABLE I
CONDITIONAL OUTPUT SYMBOL PROBABILITIES

X_n	Y_{n-1}	Z_n	$P(Y_n = 0 \mid X_n, Y_{n-1}, Z_n)$
0	0	0	.95
0	0	1	.8
0	1	0	.4
0	1	1	.3
1	0	0	..6
1	0	1	..7
1	1	0	.05
1	1	1	.2

that assigns probability distributions to the inputs. If the Markov chain is in state 1 we set the input to 0 with probability α. If the Markov chain is in state 2 we set the input to 0 with probability β. Using the formulation from Section IV, we must combine the Markov chain for the channel and the Markov chain for the inputs into one channel model. Hence, we now have a four-state channel. The graph in Fig. 5 plots the mutual information for this channel as both α and β range from 0 to 1.

Notice that the maximum mutual information is identical to the i.i.d. case. In fact, there appears to be a curve consisting of linear combinations of α and β where the maximum is achieved. This provides a good check of the algorithm's validity as this result agrees with theory.

B. ISI Channel

The next numerical example examines the mutual information for an ISI channel. The model we will use here allows the output symbol at time $n+1$ to depend on the output symbol at time n (i.e., $p(y_{n+1} \mid z^{n+1}, x^{n+1}, y^n) = p(y_{n+1} \mid z_{n+1}, x_{n+1}, y_n)$). Again, Z is modeled as a simple two-state Markov chain with transition matrix

$$P = \begin{vmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{vmatrix}$$

The conditional probability distribution of Y_{n+1} for each combination of x_{n+1}, y_n, and z_{n+1} is listed in Table I.

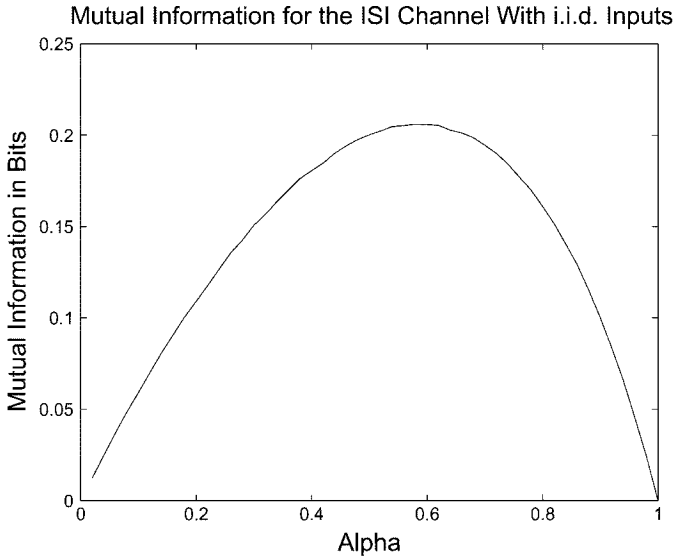

Fig. 6. Mutual information of an ISI channel with i.i.d. inputs.

We use the same input setup from the above Gilbert–Elliot example. Figs. 6 and 7 plot the mutual information for the i.i.d. inputs case and the Markov modulated inputs case. We can see that adding memory to the inputs for the ISI channel increases maximum mutual information by a small amount (approximately 8%).

VIII. CONCLUSION

We have formulated entropy and mutual information for finite-state channels in terms of Lyapunov exponents for products of random matrices, yielding

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= \lambda_X + \lambda_Y - \lambda_{XY}. \end{aligned}$$

We showed that the Lyapunov exponents can be computed as expectations with respect to the stationary distributions of a class of continuous state-space Markov chains. Furthermore, we

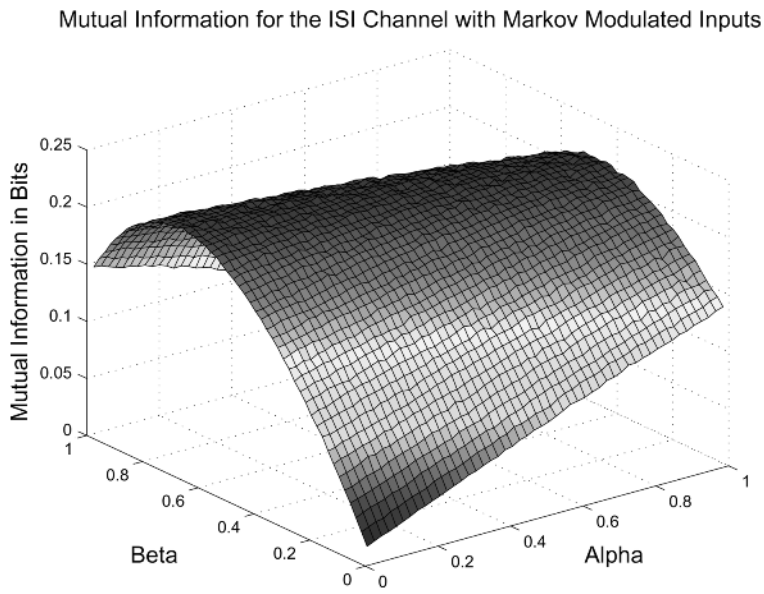

Fig. 7. Mutual information of an ISI channel with Markov-modulated inputs.

showed these stationary distributions are continuous functions of the input symbol distribution and the transition probabilities for the finite-state channel—thereby allowing us to write channel capacity in terms of Lyapunov exponents

$$\begin{aligned} C &= \max_{p(X)} I(X,Y) \\ &= \max_{p(X)} [H(X) + H(Y) + H(X,Y)] \\ &= \max_{p(X)} [\lambda_X + \lambda_Y - \lambda_{XY}]. \end{aligned}$$

These results extend work by previous authors to finite-state channel models that include ISI channels and channels with non-i.i.d. inputs.

In addition, we presented rigourous numerical methods for computing Lyapunov exponents through direct computation and by using simulations. The rigorous simulation formulation required us to develop a new functional CLT for sample entropy, as well as bounds on the initialization bias inherent in simulation. Our proposed direct computation is based on the representation of Lyapunov exponents as expectations and avoids many of the convergence problems that often arise in simulation-based computation of Lyapunov exponents.

Finally, we presented numerical results for the mutual information of a Gilbert–Elliot channel and an ISI channel. In both cases we computed the mutual information resulting from i.i.d. and Markov modulated inputs. In the Gilbert–Elliot case our numerical results agreed with known theory. For the ISI channel our results showed that adding memory to the inputs can indeed increase mutual information.

Appendix I
Rigorous Simulation of Lyapunov Exponents and Sample Entropy

In this appendix, we provide a rigorous treatment of the theory required to construct simulation-based estimates of Lyapunov exponents and entropy. This requires us to prove a new CLT for Lyapunov exponents and sample entropy as well as a means to analyze the initialization bias in these simulations. The proofs of the proposition and theorems in this appendix will be presented in Appendix II (along with the proofs of all the other propositions and theorems in the paper).

A. A Central Limit Theorem (CLT) for the Sample Entropy

The sample entropy for X based on observing X_0^n is given by

$$H_n(X) = -\frac{1}{n}\sum_{j=0}^{n-1} \log\left(\left\|\tilde{p}_j^r G_{X_{j+1}}^X\right\|_1\right) \tag{28}$$

where $\tilde{p}_0^r = r$, and $X = (X_n : n \geq 1)$ is a stationary version of the input symbol sequence. In this section, we provide a proof of the CLT for $H_n(X)$ under easily verifiable conditions on the channel model. At the same time, we also provide a CLT for the simulation estimator $\hat{H}_n(X)$.

The key to a CLT for $H_n(X)$ is to use methods associated with the CLT for Markov chains. (Note that we can not apply the CLTs for Harris chains directly, since our chains are not ϕ-irreducible.) To obtain our desired CLT, we shall represent $H_n(X)$ in terms of an appropriate martingale. The required CLT will then follow via an application of the martingale central limit theorem. This idea also appears in Heyde [26], but the conditions appearing there are difficult to verify in the current circumstances. Set

$$\tilde{g}(w,x) = \log\left(1/\left\|wG_x^X\right\|_1\right) - H(X) \tag{29}$$

and

$$\tilde{g}(w) = \sum_x \log\left(1/\left\|wG_x^X\right\|_1\right)\left\|wG_x^X\right\|_1 - H(X). \tag{30}$$

Then, according to Proposition 3

$$\begin{aligned} H_n(X) - H(X) &= \frac{1}{n}\sum_{j=0}^{n-1} \tilde{g}\left(\tilde{p}_j^r, X_{j+1}\right) \\ &= \frac{1}{n}\sum_{j=0}^{n-1} \tilde{g}\left(p_j^r, X_{j+1}\right) \\ &= \frac{1}{n}\sum_{j=0}^{n-1} \left[\tilde{g}\left(p_j^r, X_{j+1}\right) - \tilde{g}\left(p_j^r\right)\right] \\ &\quad + \frac{1}{n}\sum_{j=0}^{n-1} \tilde{g}\left(p_j^r\right). \end{aligned}$$

Note that $\tilde{g}(p_n^r, X_{n+1}) - \tilde{g}(p_n^r)$ is a martingale difference that is adapted to X_1^{n+1}. It remains only to provide a martingale representation for $\tilde{g}(p_n^r)$.

To obtain this representation, we suppose (temporarily) that there exists a solution $k(\cdot)$ to the equation

$$k(w) - \mathrm{E}[k(p_1) \,|\, p_0 = w] = \tilde{g}(w). \tag{31}$$

Equation (31) is known as Poisson's equation for the Markov chain p, and is a standard tool for proving Markov chain CLTs; see Maigret [36] and Glynn and Meyn [25]. In the presence of a solution k to Poisson's equation

$$\sum_{j=0}^{n-1} \tilde{g}(p_j) = \sum_{j=1}^{n} (k(p_j) - \mathrm{E}[k(p_j) \,|\, p_{j-1}] + k(p_0) - k(p_n))$$

each of the terms $k(p_{n+1}) - \mathrm{E}[k(p_{n+1}) \,|\, p_n]$ is then a martingale difference that is adapted to X_1^{n+1}, thereby completing our development of a martingale representation for $H_n(X)$.

To implement this approach, we need to establish existence of a solution k to (31). A related result appears in the HMM paper of Le Gland and Mevel [34]. However, their analysis studies not the channel $\mathcal{P}$-chain, but instead focuses on the Markov chain $((\tilde{p}_n, C_n, X_n) : n \geq 0)$. Furthermore, they assume that $\tilde{g}$ is Lipschitz (which is violated over $\mathcal{P}$ for our choice of $\tilde{g}$).

Let $A = \{w \in \mathcal{P} : d(w, r) \leq 2(1-\beta)^{-1} l d^*\}$. Set

$$\begin{aligned} \gamma(A) &= \sup\{|g(x) - g(y)|/d(x,y) : x, y \in A, x \neq y\} \\ m(A) &= \sup\{\tilde{g}(x) : x \in A\} \\ \lambda &= \max_{c_0 \in \mathcal{C}} \{\min_{c_1 \in \mathcal{C}} R^l(c_0, c_1)\}. \end{aligned}$$

Note that $\gamma(A)$ and $m(A)$ are finite, because A does not include the boundaries of $\mathcal{P}$.

Proposition 7: The function k defined by

$$k(w) = \sum_{n=0}^{\infty} \mathrm{E}\tilde{g}\,(p_n^w) \tag{32}$$

satisfies (31) for each $w \in K$. Furthermore, $\sup\{|k(w)| : w \in K\} \leq \nu$, where

$$\nu \triangleq 2\gamma(A)(1-\beta)^{-2} l^2 d^* + m(A) l/\lambda. \tag{33}$$

Proof: See Appendix II-G.

We are now ready to state the main result of this section. We shall show that the sample entropy, when viewed as a function of the amount of observed data, can be approximated by Brownian motion. This so-called "functional" form of the CLT, known in the probability literature as a functional central limit theorem (FCLT), implies the ordinary CLT. This stronger form will prove useful in developing confidence interval procedures for simulation-based entropy computation methods. A rigorous statement of the FCLT involves weak convergence on the function space $D[0,\infty)$, the space of right continuous functions with left limits on $[0,\infty)$. See Ethier and Kurtz [19] for a discussion of this notion of weak convergence.

Let $p^\infty = (p_n^\infty : n \geq 0)$ be the channel $\mathcal{P}$-chain when initiated under the stationary distribution π.

Theorem 8: Assume B1–B4 and B6–B7. Then

$$\epsilon^{\frac{1}{2}}(\hat{H}_{\lfloor \cdot/\epsilon \rfloor}(X) - H(X)) \Rightarrow \sigma B(\cdot) \tag{34}$$

and

$$\epsilon^{\frac{1}{2}}(H_{\lfloor \cdot/\epsilon \rfloor}(X) - H(X)) \Rightarrow \eta B(\cdot) \tag{35}$$

as $\epsilon \downarrow 0$, where $B = (B(t) : t \geq 0)$ is a standard Brownian motion and $\Rightarrow$ denotes weak convergence on $D[0,\infty)$. Furthermore

$$\begin{aligned} \sigma^2 &= 2\mathrm{E}k(p_\infty)\tilde{g}(p_\infty) - \mathrm{E}\tilde{g}^2(p_\infty) && (36) \\ &\leq 2\nu m(A) && (37) \end{aligned}$$

and

$$\begin{aligned} \eta^2 &= \mathrm{E}\left[(g\,(p_0^\infty, X_1) + k\,(p_1^\infty) - k(p_0^\infty))^2\right] && (38) \\ &\leq (\sup\{\tilde{g}(w,x) : w \in A, x \in \mathcal{X}\} + 2\nu)^2. && (39) \end{aligned}$$

Proof: See Appendix II-H.

Theorem 8 proves that the sample entropy satisfies a CLT, and provides computable upper bounds on the asymptotic variances of $H_n(X)$ and $\hat{H}_n(X)$.

B. Simulation Methods for Computing Entropy

Computing the entropy requires calculation of an expectation with respect to the channel $\mathcal{P}$-chain's stationary distribution. Note that because the channel $\mathcal{P}$-chain is not ϕ-irreducible, it effectively remembers its initial condition p_0 forever. In particular, the support of the random vectors p_n's mass function is effectively determined by the choice of p_0. Because of this memory issue, it seems appropriate to pay special attention in this setting to the issue of the "initial transient." Specifically, we shall focus first on the question of how long one must simulate $(p_n : n \geq 0)$ in order that $(g(p_n) : n \geq 0)$ is effectively in "steady state," where

$$g(w) = \sum_x \log\left(1 \,/\, \left\|wG_x^X\right\|_1\right) \|wG_x^X\|_1. \tag{40}$$

The Proof of Proposition 7 shows that for $n \geq 0$ and $w \in K$,

$$\begin{aligned} &|\mathrm{E}g\,(p_n^w) - H(X)| \\ &\quad \leq 2\gamma(A)\beta^{\lfloor n/l \rfloor}(1-\beta)^{-1} l d^* + m(A)(1-\lambda)^{\lfloor n/l \rfloor} \end{aligned} \tag{41}$$

so that we have an explicit computable bound on the rate at which the initialization bias decays to zero. One problem with the upper bound (41) is that it tends to be very conservative. For example, the key factors β and λ that determine the exponential rate at which the initialization bias decays to zero are clearly (very) loose upper bounds on the contraction rate and coupling rate for the chain, respectively.

We now offer a simulation-based numerical method for diagnosing the level of initialization bias. The method requires that we estimate $H(X)$ via the (slightly) modified estimator

$$\hat{H}'_n(X) = \frac{1}{n}\sum_{j=0}^{n-1} g\left(\tilde{p}_j^w\right). \tag{42}$$

The Proof of Theorem 7 shows that $\hat{H}'_n(X)$ satisfies precisely the same FCLT as does $\hat{H}_n(X)$. For $c \in \mathcal{C}$, let $\tilde{p}_n^c = \tilde{p}_n^{e_c}$, where e_c is the unit vector in which unit probability mass is assigned to c. The $\tilde{p}_n$'s are generated via a common sequence of random matrices, namely

$$\tilde{p}_n^c = \frac{e_c G_{X_1}^X \cdots G_{X_n}^X}{\left\| e_c G_{X_1}^X \cdots G_{X_n}^X \right\|_1} \tag{43}$$

where $(X_n : n \geq 1)$ is a stationary version of the input symbol sequence. Also, let

$$\Gamma_n = \left\{ \sum_c v(c)\tilde{p}_n^c : v = (v(c) : c \in \mathcal{C}) \in \mathcal{P} \right\}$$

be the convex hull of the random vectors $\{\tilde{p}_n^c : c \in \mathcal{C}\}$. Finally, for $w \in \Re^d$, let $||w||_2 = (\sum_i w(c)^2)^{1/2}$ be the Euclidean norm of w.

Proposition 8: Assume B1–B4 and B6–B7. Then, for $w \in \mathcal{P}$

$$\left|\mathrm{E}g\left(\tilde{p}_n^w\right) - H(X)\right| \leq \mathrm{E}\sup\{||\nabla g(x)||_2 : x \in \Gamma_n\} \max_{c_0,c_1} ||\tilde{p}_n^{c_0} - \tilde{p}_n^{c_1}||_2 .$$

Proof: See Appendix II-I.

With Proposition 8 in hand, we can now describe our simulation-based algorithm for numerically bounding the initialization bias. In particular, let

$$W = \sup\{||\nabla g(x)||_2 : x \in \Gamma_n\} \max_{c_0,c_1} ||\tilde{p}_n^{c_0} - \tilde{p}_n^{c_1}||_2. \tag{44}$$

Suppose that we independently simulate m copies of the random variable W, thereby providing $W_1, \ldots, W_m$ of W. Then $m^{-1}\sum_{i=1}^m W_i$ is, for large m, a simulation-based upper bound on the bias of $g(\tilde{p}_n^w)$. Such a simulation-based bound can be used to plan one's simulation. In particular, as a rule of thumb, the sample size n used to compute $\hat{H}'_n(X)$ should ideally be at least a couple of orders of magnitude greater than the value n_0 at which the initialization bias is practically eliminated. Exercising some care in the selection of the sample size n is important in this setting. There is significant empirical evidence in the literature suggesting that the type of steady-state simulation under consideration here can require surprisingly large sample sizes in order to achieve convergence; see [20].

In the remainder of this section, we take the point of view that initialization bias has been eliminated (either by choosing a sample size n for the simulation that is so large that the initial transient is irrelevant, or because we have applied the bounds above so as to reduce the impact of such bias).

To provide a confidence interval for $H(X)$, we appeal to the FCLT for $\hat{H}_n(X)$ derived in Section VI. For $n \geq 2$ and $1 \leq k \leq m$, set

$$\hat{H}_{nk}(X) = \frac{1}{n/m} \sum_{j=\lfloor (k-1)n/m \rfloor}^{\lfloor kn/m \rfloor - 1} g\left(p_j^r\right). \tag{45}$$

The FCLT of Theorem 8 guarantees that

$$\sqrt{n/m}(\hat{H}_{n1}(X) - H(X), \ldots, \hat{H}_{nm}(X) - H(X)) \Rightarrow \sigma(N_1(0,1), \ldots, N_m(0,1)) \tag{46}$$

as $n \to \infty$, where the random variables

$$N_1(0,1), \ldots, N_m(0,1) \tag{47}$$

are i.i.d. Gaussian random variables with mean zero and unit variance. It follows that

$$\frac{\sqrt{m}(\hat{H}_n(X) - H(X))}{\sqrt{\frac{1}{m-1}\sum_{i=1}^n (\hat{H}_{ni}(X) - \hat{H}_n(X))^2}} \Rightarrow t_{m-1} \tag{48}$$

as $n \to \infty$, where t_{m-1} is a Student-t random variable [28] with $m-1$ degrees of freedom. Hence, if we select z such that $P(-z \leq t_{m-1} \leq z) = 1 - \delta$, the random interval

$$\left[\hat{H}_n(X) - z\frac{s_n}{\sqrt{m}}, \hat{H}_n(X) + z\frac{s_n}{\sqrt{m}}\right] \tag{49}$$

is guaranteed to be an asymptotic $100(1-\delta)\%$ confidence interval for $H(X)$, where

$$s_n = \sqrt{\frac{1}{m-1}\sum_{i=1}^m (\hat{H}_{ni}(X) - \hat{H}_n(X))^2}. \tag{50}$$

We have proved the following theorem.

Theorem 9: Assume B1–B4 and B6–B7. Then, if $\sigma^2 > 0$,

$$P\left(H(X) \in \left[\hat{H}_n(X) - z\frac{s_n}{\sqrt{m}}, \hat{H}_n(X) + z\frac{s_n}{\sqrt{m}}\right]\right) \to 1 - \delta$$

as $n \to \infty$.

We conclude this section with a brief discussion of variance reduction techniques that can be applied in conjunction with the simulation-based estimators $\hat{H}_n(X)$ and $\hat{H}'_n(X)$. Recall that X_{n+1} is the realization of $\hat{X}$ that arises in step 2 of Algorithm A when generating p_{n+1}. Set

$$\hat{H}_n^c(\lambda, x) = \hat{H}_n(X) - \lambda\frac{1}{n}\sum_{j=1}^n f(X_j) \tag{51}$$

where

$$f(x) = \log\left(\mathrm{sp}(G_x^X)\right) - \mathrm{E}\log\left(\mathrm{sp}(G_{X_\infty}^X)\right) \tag{52}$$

where $\mathrm{sp}(G_x^X)$ is the spectral radius (or Perron–Frobenius eigenvalue) of G_x^X, and X_∞ has the input symbol sequence's steady-state distribution. Note that $\mathrm{E}\log(\mathrm{sp}(G_{X_\infty}^X))$ can be easily computed, so that the $f(X_j)$'s can be easily calculated during the

course of the simulation (by precomputing $\log(\mathrm{sp}(G_x^X))$ for each $x \in \mathcal{X}$ prior to running the simulation). Clearly

$$\frac{1}{n}\sum_{j=1}^{n} f(X_j) \to 0 \text{ a.s.} \tag{53}$$

(by the strong law for finite-state Markov chains) so

$$\hat{H}_n^c(\lambda, X) \to H(X) \text{ a.s.} \tag{54}$$

as $n \to \infty$. The idea is to select λ so as to minimize the asymptotic variance of $\hat{H}_n^c(\lambda, X)$.

The quantity $n^{-1}\sum_{j=1}^{n} f(X_j)$ is known in the simulation literature as a control variate; see Bratley, Fox, and Schrage [12] for a discussion of how to estimate the optimal value of λ from the simulated data. We choose $n^{-1}\sum_{j=1}^{n} f(X_j)$ as a control variate because we expect the $f(X_j)$'s to be strongly correlated with the $g(p_j)$'s. It is when the correlation is high that we can expect control variates to be most effective in reducing variance.

We can also try to improve the simulation's efficiency by taking advantage of the regenerative structure of the X_j's. This idea is easiest to implement in conjunction with the estimator $\hat{H}_n'(X)$. Suppose that $\hat{H}_n'(X)$ is obtained by simulation of a stationary version of the (C_j, X_j)'s. Set $T_0 = 1$, and put $T_{n+1} = \inf\{m > T_n \colon C_m = C_1\}$. Then, conditional on C_1, the sequence $(T_n \colon n \geq 0)$ is a sequence of regeneration times for the X_j's; see Asmussen [5] for a definition of regeneration. It follows that, conditional on $C_1, (\tilde{H}_n : n \geq 1)$ is a sequence of i.i.d. random matrices, where

$$\tilde{H}_n = G_{X_{T_{n-1}}}^X \cdots G_{X_{T_n-1}}^X. \tag{55}$$

Then

$$\tilde{p}_{T_n-1}^w = \frac{w\tilde{H}_1 \cdots \tilde{H}_n}{\left\|w\tilde{H}_1 \cdots \tilde{H}_n\right\|_1} \stackrel{\mathcal{D}}{=} \frac{w\tilde{H}_{\sigma(1)} \cdots \tilde{H}_{\sigma(n)}}{\left\|w\tilde{H}_{\sigma(1)} \cdots \tilde{H}_{\sigma(n)}\right\|} \tag{56}$$

for any permutation $\sigma = (\sigma(i) : 1 \leq i \leq n)$ of the integers 1 through n. Hence, given a simulation to time T_n, we may obtain a lower variance estimator by averaging

$$g\left(\frac{w\tilde{H}_{\sigma(1)} \cdots \tilde{H}_{\sigma(n)}}{\left\|w\tilde{H}_{\sigma(1)} \cdots \tilde{H}_{\sigma(n)}\right\|}\right) \tag{57}$$

over a certain number of permutations σ. One difficulty with this method is that it is expensive to compute such a "permutation estimator." It is also unclear whether the variance reduction achieved compensates adequately for the increased computational expenditure.

APPENDIX II
PROOFS OF THEOREMS AND PROPOSITIONS

A. Proof of Proposition 3

For any function $g \colon \mathcal{P} \to [0, \infty)$, observe that

$$\begin{aligned}
&\mathrm{E}[g(\tilde{p}_{n+1}) \,|\, \tilde{p}_0^n] \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[g(\tilde{p}_{n+1}) \,|X_1^n\right] | \, \tilde{p}_0^n\right] \\
&\quad = \mathrm{E}\left[\sum_x g\left(\frac{\tilde{p}_n G_x^X}{||\tilde{p}_n G_x^X||_1}\right) P\left(X_{n+1} = x \,|X_1^n\right) |\tilde{p}_0^n\right].
\end{aligned} \tag{58}$$

On the other hand, **B1**–**B3** imply that

$$\begin{aligned}
&P(X_{n+1} = x \,|X_1^n) \\
&\quad = \mathrm{E}\left[P\left(X_{n+1} = x \,\middle|X_1^n, C_1^{n+2}\right) \middle| \, X_1^n\right] \\
&\quad = \mathrm{E}\left[P(X_{n+1} = x \,|\, C_{n+1}, C_{n+2}) \,|X_1^n\right] \\
&\quad = \mathrm{E}\left[q(x \,|\, C_{n+1}, C_{n+2}) \,|X_1^n\right] \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[q(x \,|\, C_{n+1}, C_{n+2}) \,\middle|C_1^{n+1}\right] \middle| \, X_1^n\right] \\
&\quad = \mathrm{E}\left[\mathrm{E}[q(x \,|\, C_{n+1}, C_{n+2}) \,|\, C_{n+1}] \,|X_1^n\right] \\
&\quad = \mathrm{E}\left[\sum_{c_{n+2}} q(x \,|\, C_{n+1}, c_{n+2}) R(C_{n+1}, c_{n+2}) \middle| \, X_1^n\right] \\
&\quad = \left\|\tilde{p}_n G_x^X\right\|_1.
\end{aligned}$$

Hence,

$$\mathrm{E}\left[g(\tilde{p}_{n+1}) \,|\tilde{p}_0^n\right] = \sum_x g\left(\frac{\tilde{p}_n G_x^X}{||\tilde{p}_n G_x^X||_1}\right) \left\|\tilde{p}_n G_x^X\right\|_1 \tag{59}$$

proving the Markov property.

B. Proof of Proposition 4

Let $X = (X_n : n \geq 1)$ be the sequence of $\tilde{\mathcal{X}}$-values sampled at step 2 of Algorithm A. Clearly

$$p_n^w = \frac{wG_{X_1}^X \cdots G_{X_n}^X}{||wG_{X_1}^X \cdots G_{X_n}^X||_1}. \tag{60}$$

Note that

$$\begin{aligned}
&P(X_1 = x_1, \ldots, X_n = x_n) \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[I(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, \right.\right. \\
&\qquad \left.\left. X_n = x_n) \,\middle|X_1^{n-1}\right]\right] \\
&\quad = \mathrm{E}\left[I(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \left\|p_{n-1} G_{x_n}^X\right\|_1\right] \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[I(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \right.\right. \\
&\qquad \left.\left. \times \left\|p_{n-1} G_{x_n}^X\right\|_1\right] \left|X_1^{n-2}\right.\right] \\
&\quad = \mathrm{E}\left[I(X_1 = x_1, \ldots, X_{n-2} = x_{n-2}) \left\|p_{n-2} G_{X_{n-1}}^X\right\|_1 \right. \\
&\qquad \left. \cdot \frac{||p_{n-2} G_{X_{n-1}}^X G_{X_n}^X||_1}{\left\|p_{n-2} G_{X_{n-1}}^X\right\|_1}\right].
\end{aligned}$$

Repeating this process $n-2$ more times proves that

$$P(X_1 = x_1, \ldots, X_n = x_n) = \left\|wG_{x_1}^X G_{x_1}^X \cdots G_{x_n}^X\right\|_1 \tag{61}$$

as desired.

C. Proof of Theorem 2

For $(p, x) \in \mathcal{P}$ x $\mathcal{X}$ and $n \geq 1$, put

$$g(p, x) = -\log(||pG_x^X||_1) \tag{62}$$

and

$$g_n(p, x) = \min(g(p, x), n). \tag{63}$$

Note that the sequence $(X_n: n \geq 1)$ generated by Algorithm A (when initiated by p_0 having stationary distribution π) is such that $(p, X) = ((p_n, X_{n+1}) : n \geq 0)$ is stationary. Since

$$\left\|p_0 G^X_{X_1} \cdots G^X_{X_n}\right\| \leq \left\|G^X_{X_1} \cdots G^X_{X_n}\right\|_\infty \tag{64}$$

it follows that

$$\frac{1}{n}\sum_{j=0}^{n-1} g(p_j, X_{j+1}) \geq -\frac{1}{n}\log\left\|G^X_{X_1} \cdots G^X_{X_n}\right\|_\infty. \tag{65}$$

Hence,

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} g(p_j, X_{j+1}) \geq H(X) \text{ a.s.} \tag{66}$$

Fatou's lemma then yields

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \mathrm{E}g(p_j, X_{j+1}) \geq H(X). \tag{67}$$

But the stationarity of (p, X) shows that

$$\frac{1}{n}\sum_{j=0}^{n-1} \mathrm{E}g(p_j, X_{j+1}) = \mathrm{E}g(p_0, X_1). \tag{68}$$

Conditioning on p_0 then gives part i).

For part ii), we need to argue that when $p_0 \in \mathcal{P}^+$

$$\mathrm{E}g(p_0, X_1) \leq H(X).$$

In this case, (11) ensures that

$$\frac{1}{n+1}\sum_{j=0}^{n} g(p_j, X_{j+1}) \to H(X) \text{ a.s.} \tag{69}$$

as $n \to \infty$. According to Birkhoff's ergodic theorem

$$\frac{1}{n+1}\sum_{j=0}^{n} g_n(p_j, X_{j+1}) \to \mathrm{E}[g_n(p_0, X_1) \,|\, \mathcal{I}] \text{ a.s.} \tag{70}$$

as $n \to \infty$, where $\mathcal{I}$ is (p, X)'s invariant σ-algebra. Since $g_n \leq g$, (69) and (70) prove that

$$H(X) \geq \mathrm{E}[g_n(p_0, X_0) \,|\, \mathcal{I}] \text{ a.s.} \tag{71}$$

Taking expectations in (71) and applying the Monotone Convergence Theorem completes the proof that $H(X) \geq \mathrm{E}g(p_0, X_1)$. □

D. Proof of Theorem 3

Because $G^X_{X_1}$ is row-allowable a.s., $\|wG^X_{X_1}\|_1 > 0$ a.s. for every $w \in \mathcal{P}$. So $(p_n: n \geq 0)$ is a Markov chain that is well defined for every initial vector $w \in \mathcal{P}$. Also, for bounded and continuous $h : \mathcal{P} \to \Re$

$$\mathrm{E}[h(p_{n+1}) \,|\, p_n = w] = \sum_x h\left(\frac{wG^X_x}{\|wG^X_x\|_1}\right)\left\|wG^X_x\right\|_1 \tag{72}$$

is continuous in $w \in \mathcal{P}$. It follows that the Markov chain $(p_n : n \geq 0)$ is a Feller Markov chain (see Karr [30, p. 44] for a definition) living on a compact state space. The chain p therefore possesses a stationary distribution π; see [30]. □

E. Proof of Proposition 6

Recall Proposition 4, which stated that

$$p^w_n = \frac{wG^X_{X_1(w)} \cdots G^X_{X_n(w)}}{\left\|wG^X_{X_1(w)} \cdots G^X_{X_n(w)}\right\|_1} \tag{73}$$

where $X(w) = (X_n(w) : n \geq 1)$ is the input symbol sequence when C_1 is sampled from the mass function w. In particular

$$P(X_1(w) = x_1, \ldots, X_n(w) = x_n) = wG^X_{x_1} \cdots G^X_{x_n} e.$$

Hence we can generate the Markov chain p^w_n using nonstationary versions of the channel $C(w)$ and symbol sequence $X(w)$. Since the channel transition matrix R is aperiodic, we can find a probability space upon which $(X_n(w) : n \geq 1)$ can be coupled to a stationary version of X, call it $(X^*_n: n \geq 1)$; see, for example, Lindvall [35].

Consequently, there exists a finite-valued random time T (i.e., the coupling time) so that $X_n(w) = X^*_n$ for $n \geq T$. Since $G^X_{X(w)} = G_{X^*_n}$ for $n \geq T$, Theorem 4 shows that

$$d\left(p^w_{nl+T}, \tilde{p}^r_{nl+T}\right) \leq \beta^{n-1} d\left(p^w_{l+T}, \tilde{p}^r_{l+T}\right) \tag{74}$$

from which the result follows. □

F. Proof of Theorem 5

We need to prove parts i) and iv). Because $p^w_n \Rightarrow p^*_\infty$, we have

$$\frac{1}{n}\sum_{j=0}^{n-1} P(p^w_j \in \cdot) \Rightarrow \pi(\,\cdot\,) \tag{75}$$

as $n \to \infty$. Assumptions B6-B7 ensure that each of the matrices in $\{G^X_x : x \in \mathcal{X}\}$ is row-allowable, so the proof of Theorem 3 shows that $(p_n : n \geq 0)$ is Feller on $\mathcal{P}$. The limiting distribution π must therefore be a stationary distribution for $p = (p_n : n \geq 0)$; see [35]. For the uniqueness, suppose $\tilde{\pi}$ is a second stationary distribution. Then

$$\tilde{\pi}(\,\cdot\,) = \int_{\mathcal{P}} \tilde{\pi}(dw)\frac{1}{n}\sum_{j=0}^{n-1} P\left(p^w_j \in \cdot\right). \tag{76}$$

Taking limits as $n \to \infty$ and using (75) proves that $\tilde{\pi} = \pi$, establishing uniqueness.

To prove part iv), we first prove that $P(\tilde{p}^w_l \in K) = 1$ for $w \in K$. Recall that $d(\tilde{p}^r_l, r) \leq ld^*$. So

$$\begin{aligned} d(\tilde{p}^w_l, r) &\leq d\left(\tilde{p}^w_l, \tilde{p}^r_l\right) + d\left(\tilde{p}^r_l, r\right) \\ &\leq \tau(H_1) d(w, r) + ld^* \\ &\leq \beta(1-\beta)^{-1} ld^* + ld^* \\ &= (1-\beta)^{-1} ld^*. \end{aligned}$$

Hence, $P(\tilde{p}_l^w \in K) = 1$, so that $P(\tilde{p}_l^w \in K \,|\, C_1 = c) = 1$ for all $c \in \mathcal{C}$. However, according to Proposition 4

$$P\left(p_l^w \in K\right) = \sum_i P\left(\tilde{p}_l^w \in K \,|\, C_1 = c\right) w(c) \tag{77}$$

so we may conclude that $P(p_l^w \in K) = 1$ for $w \in K$. Since p^w is Markov, it follows that $P(p_{nl}^w \in K) = 1$ for $n \geq 0$ and $w \in K$. □

G. Proof of Proposition 7

We start by showing that if $w \in K$, then $P(p_n^w \in A) = 1$ for $n \geq 0$. (Part iv) of Theorem 5 proves this if n is a multiple of l. Here, we show this for all $n \geq 0$.) According to the Proof of Theorem 5, it suffices to establish that $P(\tilde{p}_n^w \in A) = 1$ for $n \geq 0$ and $w \in K$. Now, for $0 \leq j < l$

$$\begin{aligned} d(\tilde{p}_{lk+j}^w, r) &\leq d\left(\tilde{p}_{lk+j}^w, \tilde{p}_{lk+j}^r\right) + d\left(\tilde{p}_{lk+j}^r, r\right) \\ &\leq d(w, r) + d\left(\tilde{p}_{lk+j}^r, r\right) \\ &\leq (1-\beta)^{-1} l d^* + d\left(\tilde{p}_{lk+j}^r, r\right). \end{aligned}$$

The same argument as used to show that $p_n^* \to p_\infty^*$ a.s. as $n \to \infty$ shows that

$$d\left(\tilde{p}_{lk+j}^r, r\right) \leq \sum_{i=1}^{n} \beta^i l d^* \leq (1-\beta)^{-1} l d^*. \tag{78}$$

It follows that $P(\tilde{p}_{lk+j}^w \in A) = 1$.

The key to analyzing the expectations appearing in the definition of k is to observe that

$$\begin{aligned} \mathrm{E}\tilde{g}(p_n^w) &= \mathrm{E}\tilde{g}\left(p_n^w\right) - \mathrm{E}\tilde{g}(p_\infty) \\ &= \mathrm{E}\left(\tilde{g}(p_n^w) - \tilde{g}(\tilde{p}_n^w)\right) + \mathrm{E}\left(\tilde{g}(\tilde{p}_n^w) - \tilde{g}(\tilde{p}_n^r)\right) \\ &\quad + \mathrm{E}(\tilde{g}(p_n^*) - \tilde{g}(p_\infty^*)). \end{aligned}$$

Suppose that $n = kl + j$ for $0 \leq j \leq l$. Since $\tilde{p}_n^w, \tilde{p}_n^r \in A$ for $w \in K$, we have

$$\begin{aligned} \left|\tilde{g}\left(\tilde{p}_n^w\right) - \tilde{g}\left(\tilde{p}_n^r\right)\right| &\leq \gamma(A) d(\tilde{p}_n^w, \tilde{p}_n^r) \\ &\leq \gamma(A)\beta^k d(w, r) \\ &\leq \gamma(A)\beta^k (1-\beta)^{-1} l d^*. \end{aligned}$$

Also,

$$\begin{aligned} |\tilde{g}(p_n^*) - \tilde{g}(p_\infty^*)| &\leq \gamma(A) d(p_{kl+j}^*, p_\infty^*) \\ &\leq \gamma(A) \sum_{i=k}^{\infty} d\left(p_{il+j}^*, p_{(i+1)l+j}^*\right)\Big) \\ &\leq \gamma(A) \sum_{i=k}^{\infty} \beta^i l d^* \\ &= \gamma(A)\beta^k (1-\beta)^{-1} l d^*. \end{aligned}$$

To analyze $\tilde{g}(p_n^w) - \tilde{g}(p_n^r)$, we couple the Markov chain $(X_n(w) : n \geq 1)$ to its stationary version $X_n(r) : n \geq 1)$, as in the proof of Proposition 5. If T is the coupling time

$$\begin{aligned} \mathrm{E}\left|\tilde{g}(p_n^w) - \tilde{g}(p_n^r)\right| &= \mathrm{E}\left|\tilde{g}\left(p_n^w\right) - \tilde{g}\left(p_n^r\right)\right| I(T > n) \\ &\leq m(A) P(T > n) \\ &\leq m(A)(1-\lambda)^{\lfloor n/l \rfloor}; \end{aligned}$$

the bound $P(T > n) \leq (1-\lambda)^{\lfloor n/l \rfloor}$ can be found in Asmussen, Glynn, and Thorisson [5]. Summing our bounds over n, it follows that the sum defining k converges absolutely for each $w \in K$, and the sum is dominated by the bound appearing in the statement of the proposition. Given that the sum converges, it is then straightforward to show that k satisfies (31). □

H. Proof of Theorem 8

Note that

$$\epsilon^{\frac{1}{2}} \sum_{j=0}^{\lfloor t/\epsilon \rfloor} \tilde{g}(p_j^\infty) = \epsilon^{\frac{1}{2}} \sum_{j=1}^{\lfloor t/\epsilon \rfloor + 1} D_j + k\left(p_0^\infty\right) - k\left(p_{\lfloor t/\epsilon \rfloor}^\infty\right) \tag{79}$$

where

$$D_j = k(p_j^\infty) - \mathrm{E}[k(p_j^\infty) \,|\, k(p_{j-1}^\infty)]$$

is a martingale difference. Since π is the unique stationary distribution of p, it follows that p^∞ is an ergodic stationary sequence; see [5]. Because $\mathrm{E}\eta_j^2 \leq \nu < \infty$, Theorem 23.1 of Billingsley [9] applies, so that

$$\epsilon^{\frac{1}{2}} \sum_{j=1}^{\lfloor \cdot/\epsilon \rfloor} D_j \Rightarrow \sigma B(\,\cdot\,) \tag{80}$$

as $\epsilon \downarrow 0$ in $D[0, \infty)$, where $\sigma^2 = \mathrm{E}D_1^2$. Since $k(p_n^\infty)$ is a.s. a bounded sequence (by Proposition 7), it follows that

$$\epsilon^{\frac{1}{2}} \sum_{j=1}^{\lfloor \cdot/\epsilon \rfloor} \tilde{g}(p_j^\infty) \Rightarrow \sigma B(\,\cdot\,) \tag{81}$$

as $\epsilon \downarrow 0$ in $D[0, \infty)$.

We now represent p_n^∞ and p_n^r in the form

$$\begin{aligned} p_n^\infty &= \frac{p_0^\infty G_{X_1}^X \cdots G_{X_n}^X}{\left\| p_0^\infty G_{X_1}^X \cdots G_{X_n}^X \right\|} \\ p_n^r &= \frac{r G_{X_1}^X \cdots G_{X_n}^X}{\left\| r G_{X_1}^X \cdots G_{X_n}^X \right\|} \end{aligned}$$

where $(X_n : n \geq 1)$ is a common stationary version of the input symbol sequence (and is correlated with p_0^∞). But

$$d\left(p_n^\infty, p_n^r\right) \leq \beta^{\lfloor n/l \rfloor} d(p_0^\infty, r) \tag{82}$$

so

$$|\tilde{g}\left(p_n^{\infty}\right)-\tilde{g}\left(p_n^{r}\right)| \leq \gamma(A)\beta^{\lfloor n/l \rfloor}(1-\beta)^{-1} l d^{*}. \tag{83}$$

Hence,

$$\epsilon^{\frac{1}{2}} \sup_{t\geq 0}\left|\sum_{n=0}^{\lfloor t/\epsilon \rfloor}\left(\tilde{g}(p_n^{\infty})-\tilde{g}(p_n^{r})\right)\right| \leq \epsilon^{\frac{1}{2}}\gamma(A)(1-\beta)^{-2} l^2 d^{*}. \tag{84}$$

Thus, we may conclude that

$$\epsilon^{\frac{1}{2}}\left(\hat{H}_{\lfloor \cdot/\epsilon \rfloor}(X)-H(X)\right)=\epsilon^{\frac{1}{2}}\sum_{j=0}^{\lfloor \cdot/\epsilon \rfloor}\tilde{g}\left(p_j^{r}\right) \Rightarrow \sigma B(\cdot) \tag{85}$$

as $\epsilon \downarrow 0$ in $D[0,\infty)$.

We can now simplify the expression for σ^2. Note that

$$\begin{aligned}\mathrm{E}D_1^2 &= \mathrm{E}k^2\left(p_1^{\infty}\right)-\mathrm{E}\left(\mathrm{E}\left[k\left(p_1^{\infty}\right)\middle| p_0^{\infty}\right]\right)^2\\ &= \mathrm{E}k^2\left(p_0^{\infty}\right)-\mathrm{E}\left(k(p_0^{\infty})-\tilde{g}\left(p_0^{\infty}\right)\right)^2\\ &= 2\mathrm{E}k\left(p_0^{\infty}\right)\tilde{g}\left(p_0^{\infty}\right)-\mathrm{E}\tilde{g}^2\left(p_0^{\infty}\right);\end{aligned}$$

this proves the first of our two FCLTs.

The second FCLT is proved in the same way. The martingale difference D_j is replaced by $\hat{D}_j$, where

$$\begin{aligned}\hat{D}_j &= \tilde{g}\left(p_{j-1}^{\infty}, X_j\right)-\tilde{g}\left(p_{j-1}^{\infty}\right)+k\left(p_j^{\infty}\right)-\mathrm{E}\left[k\left(p_j^{\infty}\right)\middle| p_{j-1}^{\infty}\right]\\ &= \tilde{g}\left(p_{j-1}^{\infty}, X_j\right)+k\left(p_j^{\infty}\right)-k\left(p_{j-1}^{\infty}\right);\end{aligned}$$

the remainder of the proof is similar to that for $\hat{H}_n(X)$ and is therefore omitted. □

I. Proof of Proposition 8

Let $(p_n^{\infty} : n \geq 0)$ be a stationary version of the channel $\mathcal{P}$-chain, and recall that

$$p_n^{\infty}=\frac{p_0^{\infty}G_{X_1}^{X}\cdots G_{X_n}^{X}}{\left\|p_0^{\infty}G_{X_1}^{X}\cdots G_{X_n}^{X}\right\|_1} \tag{86}$$

where $(X_n: n \geq 1)$ is stationary. Define $\tilde{p}_n^{c}$ as in (43), where the X_i's appearing in (43) are precisely those of (86).

We claim that $p_n^{\infty} \in \Gamma_n$ for $n \geq 0$. The result is obvious if $n=0$. Suppose that $p_n^{\infty} \in \Gamma_n$, so that

$$p_n^{\infty}=\sum_c \nu(c)\tilde{p}_n^{c} \tag{87}$$

for some $\nu \in \mathcal{P}$. Then

$$\begin{aligned}p_{n+1}^{\infty} &= \frac{p_n^{\infty}G_{X_{n+1}}^{X}}{\left\|p_n^{\infty}G_{X_{n+1}}^{X}\right\|_1}\\ &= \sum_c \nu(c)\frac{\tilde{p}_n^{c}G_{X_{n+1}}^{X}}{\left\|\tilde{p}_n^{c}G_{X_{n+1}}^{X}\right\|_1}\frac{\left\|\tilde{p}_n^{c}G_{X_{n+1}}^{X}\right\|}{\left\|p_n^{\infty}G_{X_{n+1}}^{X}\right\|_1}\\ &= \sum_c \nu'(c)\tilde{p}_{n+1}^{c},\end{aligned}$$

where

$$\nu'(c)=\left\|\nu(c)\tilde{p}_n^{c}G_{X_{n+1}}^{X}\right\|_1 \Big/ \left\|p_n^{\infty}G_{X_{n+1}}^{X}\right\|_1. \tag{88}$$

Since $\nu' \in \mathcal{P}$, the required induction is complete. Now

$$\mathrm{E}g(\tilde{p}_n^{w})-H(X)=\mathrm{E}g\left(\tilde{p}_n^{w}\right)-\mathrm{E}g\left(p_n^{\infty}\right) \tag{89}$$

and

$$\left|g\left(\tilde{p}_n^{w}\right)-g\left(p_n^{\infty}\right)\right|=\left|\nabla g(\zeta)\left(\tilde{p}_n^{w}-p_n^{\infty}\right)\right| \tag{90}$$

for some $\zeta \in \Gamma_n$. So

$$\begin{aligned}\left|g\left(\tilde{p}_n^{w}\right)-g\left(p_n^{\infty}\right)\right| &\leq \|\nabla g(\zeta)\|_2\left\|\tilde{p}_n^{w}-p_n^{\infty}\right\|_2\\ &\leq \sup\{\|\nabla g(x)\|_2 : x \in \Gamma_n\}\left\|\tilde{p}_n^{w}-p_n^{\infty}\right\|_2.\end{aligned} \tag{91}$$

The random vector $\tilde{p}_n^{w} \in \Gamma_n$; the proof is identical to that for p_n^{∞}. So

$$\begin{aligned}\tilde{p}_n^{\infty} &= \sum_c \nu_1(c)\tilde{p}_n^{c}\\ p_n^{\infty} &= \sum_c \nu_2(c)\tilde{p}_n^{c}\end{aligned}$$

for some $\nu_1, \nu_2 \in \mathcal{P}$. Observe that

$$\begin{aligned}\left\|\tilde{p}_n^{w}-p_n^{\infty}\right\|_2 &= \left\|\sum_{c_0}\nu_1(c_0)\tilde{p}_n^{c_0}-\sum_{c_1}\nu_2(c_1)\tilde{p}_n^{c_1}\right\|_2\\ &\leq \sum_{c_0}\sum_{c_1}\nu_1(c_0)\nu_2(c_1)\max_{c_0,c_1}\left\|\tilde{p}_n^{c_0}-\tilde{p}_n^{c_1}\right\|_2\\ &= \max_{c_0,c_1}\left\|\tilde{p}_n^{c_0}-\tilde{p}_n^{c_1}\right\|_2.\end{aligned}$$

Combining this bound with (91) completes the proof. □

J. Proof of Theorem 7

We will apply the corollary to Theorem 6 of Karr [30]. Our Theorem 5 establishes the required uniqueness for the stationary distribution of the channel $\mathcal{P}$-chain. Furthermore, the compactness of $\mathcal{P}$ yields the necessary tightness. The corollary also demands that one establish that if h is continuous on $\mathcal{P}$ and $w_n \to w_\infty \in \mathcal{P}$, then

$$\mathrm{E}[h(p_{n,1})\,|\,p_{n,0}=w_n] \to \mathrm{E}[h(p_1)\,|\,p_0=w_\infty]. \tag{92}$$

Observe that

$$\begin{aligned}\mathrm{E}[h(p_{n,1})\,|\,p_{n,0}=w_n] &= \sum_{j=1}^{n}\mathrm{E}[h(w_{nj})I(p_1 \in w_{nj})\,|\,p_0=\tilde{w}_{ni}]\\ &= \mathrm{E}[h_n(p_1)\,|\,p_0=\tilde{w}_{ni}]\end{aligned}$$

where $\tilde{w}_{ni} \in \mathcal{P}_n$ is the representative point associated with the set $\tilde{E}_{ni}$ of which w_n is a member, and

$$h_n(w)=\sum_{i=1}^{n}h(w_{ni})I(w \in E_{ni}). \tag{93}$$

Since $\mathcal{P}$ is compact, h is uniformly continuous on $\mathcal{P}$. Thus,

$$|h_n(w) - h(w)| \leq \sup\{|h(x) - h(y)| : x, y \in E_{ni}, 1 \leq i \leq n\} \to 0 \quad (94)$$

as $n \to \infty$. It follows that

$$|\mathrm{E}[h(p_{n,i})|p_{n,0} = w_n] - \mathrm{E}[h(p_1)\,|\,p_0 = \tilde{w}_{ni}]\,| \to 0 \quad (95)$$

as $n \to \infty$. But p is a Feller chain (see the Proof of Theorem 5), so $\mathrm{E}[h(p_1)\,|\,p_0 = w]$ is continuous in w. Since $\tilde{w}_{ni} \to w_\infty$, the proof of (92) is complete. The corollary of [30] therefore guarantees that $\pi_n \Rightarrow \pi$ as $n \to \infty$, where π is the stationary distribution of p.

Finally, note that B6–B7 forces each G_x^X to be row-admissible. As a consequence, $\|wG_x^X\|_1 > 0$ for $w \in \mathcal{P}$, and thus, g is bounded and continuous over $\mathcal{P}$. It follows that $\hat{H}_n(X) \to H(X)$ as $n \to \infty$, as desired. □

K. Proof of Theorem 6

We use an argument similar to that employed in the proof of Theorem 7. Let $(p_{n,i} : i \geq 0)$ be the channel $\mathcal{P}$-chain associated with (R_n, q_n), and let $\{G_x^n : x \in \mathcal{X}\}$ be the associated family of matrices $\{G_x^X : x \in \mathcal{X}\}$ corresponding to model n. Note that for n sufficiently large, (R_n, q_n) satisfies the conditions B1–B4 and B6–B7, so that Theorem 5 applies to $(p_{n,i} : i \geq 0)$ for n large. Let π_n and K_n be, respectively, the unique stationary distribution of $(p_{n,i} : i \geq 0)$ and the K-set guaranteed by Theorem 5. For any continuous function $h : \mathcal{P} \to \Re$ and sequence $w_n \to w \in \mathcal{P}$,

$$\mathrm{E}[h(p_{n,1})\,|\,p_{n,0} = w_n] = \sum_{x \in \mathcal{X}} h\left(\frac{w_n G_x^n}{\|w_n G_x^n\|_1}\right) \|w_n G_x^n\|_1 \to \sum_{x \in \mathcal{X}} h\left(\frac{wG_x^X}{\|wG_x^X\|_1}\right) \|wG_x^X\|_1$$

as $n \to \infty$ (because of the fact that the G_x^X's are row-allowable, so $\|wG_x^X\|_1 > 0$ for $x \in \mathcal{X}$ and $w \in \mathcal{P}$). As in the Proof of Theorem 9, we may therefore conclude that $\pi_n \Rightarrow \pi$ as $n \to \infty$, where π is the unique stationary distribution of the channel $\mathcal{P}$-chain associated with (R, q).

Note that

$$H_n(X) = \int_{\tilde{K}} g_n(w)\pi_n(dw) \quad (96)$$

where

$$g_n(w) = \sum_{x \in \mathcal{X}} \log\left(1 / \|wG_x^n\|_1\right) \|wG_x^n\|_1 \quad (97)$$

and $\tilde{K} \subset \mathcal{P}^+$ is a compact set containing all the K_n's for n sufficiently large. Since $g_n \to g$ as $n \to \infty$ uniformly on $\tilde{K}$, it follows from (96) and $\pi_n \Rightarrow \pi$ as $n \to \infty$ that $H_n(X) \to H(X)$ as $n \to \infty$. □

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their significant efforts and detailed feedback. We would also like to thank Prof. Tsachy Weissman for several helpful discussions.

REFERENCES

[1] D. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *Proc. 2001 IEEE Int. Conf. Communications*, Helsinki, Finland, Jun. 2001, pp. 2692–2695.
[2] D. Arnold, H. A. Loeliger, and P. Vontobel, "Computation of information rates from finite-state source/channel models," in *Proc. Allerton Conf. Communications, Control and Computing*, Monticello, IL, Sep. 2002.
[3] D. W. Arnold, *Computing Information Rates of Finite-State Models with Application to Magnetic Recording*. Konstanz, Germany: Hartung-Gorre Verlag, 2002, ETH-Diss. no. 14760.
[4] L. Arnold, L. Demetrius, and M. Gundlach, "Evolutionary formalism for products of positive random matrices," *Ann. Appl. Probab.*, vol. 4, pp. 859–901, 1994.
[5] S. Asmussen, P. Glynn, and H. Thorisson, "Stationarity detection in the initial transient problem," *ACM Trans. Modeling and Computer Simulation*, vol. 2, pp. 130–157, Apr. 1992.
[6] S. Asmussen, *Applied Probability and Queues*. New York: Wiley, 1987.
[7] R. Atar and O. Zeitouni, "Lyapunov exponents for finite state nonlinear filtering problems," *SIAM J. Cont. Optimiz. 35*, pp. 36–55, 1997.
[8] R. Bellman, "Limit theorems for noncommutative operations," *Duke Math. J.*, 1954.
[9] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.
[10] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Trans. First Prague Conf. Information Theory, Random Processes*, Prague, Czechoslovakia, 1957, pp. 13–20.
[11] S. Boyd, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA: Soc. Ind. Applied Math., 1994.
[12] P. Brately, B. Fox, and L. Schrage, *A Guide to Simulation*, 2nd ed. New York: Springer-Verlag, 1986.
[13] H. Cohn, O. Nerman, and M. Peligrad, "Weak ergodicity and products of random matrices," *J. Theor. Probab.*, vol. 6, pp. 389–405, Jul. 1993.
[14] J. E. Cohen, "Subadditivity, generalized products of random matrices and operations research," *SIAM Rev.*, vol. 30, pp. 69–86, 1988.
[15] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[16] R. Darling, "The Lyapunov exponent for product of infinite-dimensional random matrices," in *Lyapunov Exponents (Lecture Notes in Mathematics)*. New York: Springer-Verlag, 1991, vol. 1486, Proceedings of a conference held in Oberwolfach, Germany.
[17] P. Diaconis and D. A. Freedman, "Iterated random functions," *SIAM Rev.*, vol. 41, pp. 45–67, 1999.
[18] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1518–1569, Jun. 2002.
[19] S. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. New York: Wiley, 1986.
[20] G. Froyland, "Rigorous numerical estimation of Lyapunov exponents and invariant measures of iterated function systems," *Int. J. Bifur. Chaos Appl. Sci. Engrg.*, vol. 10, pp. 103–122, 2000.
[21] H. Furstenberg and H. Kesten, "Products of random matrices," *Ann. Math. Statist.*, pp. 457–469, 1960.
[22] H. Furstenberg and Y. Kifer, "Random matrix products and measures on projective spaces," *Isreal J. Math.*, vol. 46, pp. 12–32, 1983.
[23] R. Gharavi and V. Anantharam, "An upper bound for the largest Lyapunov exponent of a Markovian product of nonnegative matrices," *Theor. Comp. Sci.*, vol. 332, no. 1–3, pp. 543–557, Feb. 2005.
[24] A. Goldsmith and P. Varaiya, "Capacity, mutual information, and coding for finite state Markov channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 868–886, May 1996.
[25] P. Glynn and S. Meyn, "A Lyapunov bound for solutions of Poisson's equation," *Ann. Probab.*, pp. 916–931, 1996.
[26] P. Hall and C. Heyde, *Martingale Limit Theory and Its Application*. New York: Academic, 1980.
[27] G. Han and B. Marcus, "Analyticity of entropy rate of a hidden Markov chain," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 2193–2197.
[28] A. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd ed. New York: McGraw-Hill, 2000.
[29] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the entropy of a hidden Markov process," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jul. 2004, p. 10.
[30] A. Karr, "Weak convergence of a sequence of Markov chains," *Z. Wahrscheinlichkeitstheorie verw. Geibite 33*, vol. 33, pp. 41–48, 1975.

[31] A. Kavčić, "On the capacity of Markov sources over noisy channels," in *Proc. IEEE Globecom 2001*, San Antonio, TX, Nov. 2001, pp. 2997–3001.
[32] E. S. Key, "Lower bounds for the maximal Lyapunov exponent," *J. Theor. Probab.*, vol. 3, pp. 447–487, 1990.
[33] J. Kingman, "Subadditive ergodic theory," *Ann. Probab.*, vol. 1, pp. 883–909, 1973.
[34] F. LeGland and L. Mevel, "Exponential forgetting and geometric ergodicity in hidden Markov models," *Math. Control, Signals and Syst.*, vol. 13, no. 1, pp. 63–93, 2000.
[35] T. Lindvall, "Weak convergence of probability measures and random functions in the function space D[0; 1)," *J. Appl. Probab.*, vol. 10, pp. 109–121, 1973.
[36] N. Maigret, "Théoóréme de limite centrale fonctionnel pour une chaine de Markov récurrente au sens de Harris et positive," *Ann. Inst. H. Poincaré*, vol. 14, pp. 425–440, 1979, Sect. B (N. S.).
[37] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*. New York: Springer-Verlag, 1994.
[38] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliot channel," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1277–1290, Nov. 1989.
[39] E. Ordentlich and T. Weissman, "New bounds on the entropy rate of hidden Markov processes," in *Proc. San Antonio Information Theory Workshop*, San Antonio, TX, Oct. 2004.
[40] E. Ordentlich and T. Weissman, "Approximations for the Entropy Rate of a Hidden Markov Process," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1838–1842.
[41] V. Oseledec, "A multiplicative ergodic theorem," *Trudy Moskov. Mat. Obsc.*, vol. 19, pp. 179–210, 1968.
[42] Y. Peres, "Domains of analytic continuation for the top Lyapunov exponent," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 29, pp. 131–148, 1992.
[43] Y. Peres, K. Simon, and B. Solomyak, Absolute Continuity for Random Iterated Function Systems with Overlaps, 2005, preprint.
[44] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite-state ISI channels," in *Proc. IEEE Globecom 2001*, San Antonio, TX, Nov. 2001, pp. 2992–2996.
[45] K. Ravishankar, "Power law scaling of the top Lyapunov exponent of a product of random matrices," *J. Statist. Phys.*, vol. 54, pp. 531–537, 1989.
[46] E. Seneta, *Non-Negative Matrices and Markov Chains*, 2nd ed. New York: Springer-Verlag, 1981.
[47] V. Sharma and S. K. Singh, "Entropy and channel capacity in the regenerative setup with applications to Markov channels," in *Proc. 2001 IEEE Int. Symp.nformation Theory*, Washington, DC, Jun. 2001, p. 283.
[48] G. Stuber, *Principles of Mobile Communication*. Norwell, MA: Kluwer Academic, 1999.
[49] N. Stokey and R. Lewis, *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard Univ. Press, 2001.
[50] J. Tsitsiklis and V. Blondel, "The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—When not impossible—To compute and to approximate," *Math. Control, Signals, and Syst.*, vol. 10, pp. 31–40, 1997, correction in vol. 10, no. 4, p. 381, 2000.