Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Shannon Meets Lyapunov: Connections between Information Theory and Dynamical Systems

Tim Holliday
Princeton/Bell Labs

Peter Glynn
Stanford University

Andrea Goldsmith
Stanford University

Abstract— **This paper explores connections between Information Theory, Lyapunov exponents for products of random matrices, and hidden Markov models. Specifically, we will show that entropies associated with finite-state channels are equivalent to Lyapunov exponents. We use this result to show that the traditional prediction filter for hidden Markov models is *not* an irreducible Markov chain in our problem framework. Hence, we do not have access to many well-known properties of irreducible continuous state space Markov chains (e.g. a unique and continuous stationary distribution). However, by exploiting the connection between entropy and Lyapunov exponents and applying proof techniques from the theory of random matrix products we can solve a broad class of problems related to capacity and hidden Markov models. Our results provide strong regularity results for the non-irreducible prediction filter as well as some novel theoretical tools to address problems in these areas.**

I. Introduction

In this paper we explore connections between Shannon entropy and Lyapunov exponents for products of random matrices. Specifically, we will examine some of the unique problems, solution techniques, and insights that arise from the connection between these seemingly disparate bodies of theory. In [10] we focused on using this connection to compute entropy and Lyapunov exponents for finite-state channels. In this paper we will address some of the surprising theoretical problems that arise when examining the connections between these areas. Perhaps most significantly, we show that the standard prediction filter associated with hidden Markov models (HMMs) is *not* an irreducible Markov chain. Hence, much of the standard theory for continuous state space Markov chains cannot be applied to ensure results that are often taken for granted. For example, lack of irreducibility prevents automatic access to a unique stationary distribution, rates of convergence to stationarity, or continuity of the stationary distribution for the HMM prediction filter.

We then show that the connection between Lyapunov exponents and Shannon entropy gives us access to a new set of tools that allows to address these non-irreducibility problems. In particular, we show that many convergence results for products of random matrices can be applied to ensure strong convergence results for the HMM prediction filter (even though the filter is not irreducible). Many of these strong convergence results require a fair amount of technical detail for which we will refer to [9]. Our goal in this paper is to summarize the theoretical results and to present some of the tools we use to solve these problems.

In Section II we present our first result that shows the symbol entropies associated with a finite state channel are equivalent to the Lyapunov exponents associated with a particular class of random matrix products.

Section III presents our second set of results that show the entropy associated with finite-state Markov chains (with no channel state information) may be computed as an expectation with respect to the stationary distribution of a particular Markov chain. This Markov chain is closely related to the well-known "projective product" in chaotic dynamic systems or the prediction filter in HMMs. In Section III.B we show the rather startling conclusion that the prediction filter is typically an *extremely* poorly behaved process. Specifically, the filter does not satisfy even the weakest form of irreducibility (i.e. Harris recurrence), and possesses infinite memory.

In Sections IV and V we present our third set of results to show that we can combine theory from random matrix products and Markov chains to resolve many of the difficulties described above. Specifically, by exploiting the contraction property of positive matrices we are able to show conditions under which the prediction filter has a unique stationary distribution, exponential rates of convergence to steady-state, and continuity of the stationary distribution with respect to the channel input distribution and channel transition probabilities.

II. Markov Channels with Ergodic Inputs

Consider a communication channel with (channel) state sequence $C = (C_n : n \geq 0)$, input symbol sequence $X = (X_n : n \geq 0)$, and output symbol sequence $Y = (Y_n : n \geq 0)$. 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 $s = (s_n : n \geq 0)$ 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 $m + n$.

A. Channel Model Assumptions

In this section (and throughout the rest of this paper), we will assume that:

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

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

0-7803-9568-9/05/$20.00 ©2005 IEEE

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/output symbol pairs $\{(X_i, Y_i) : i \geq 0\}$ 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 $x_0^n \in \mathcal{X}^{n+1}, y_0^n \in \mathcal{Y}^{n+1}$.

A3: 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

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

The non-causal 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 A1-A3, in particular channels with ISI, dependent inputs, or any other finite-memory chain (see [9] for some specific examples).

B. Entropy as a Lyapunov Exponent

Let the stationary distribution of the channel be represented as a row vector $r = (r(c) : c \in \mathcal{C})$, and let e be a column vector in which every entry is equal to one. For $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, let

$$G_{(x,y)}^{(X,Y)} = (G_{(x,y)}^{(X,Y)}(c_0, c_1) : c_0, c_1 \in \mathcal{C})$$

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

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

Observe that

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

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

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

where

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

The quantity $\lambda(X, Y)$ is known as the largest Lyapunov exponent (or, simply, Lyapunov exponent) associated with the sequence of random matrix products

$$\left(G_{(X_0, Y_0)}^{(X,Y)} G_{(X_1, Y_1)}^{(X,Y)} \cdots G_{(X_n, Y_n)}^{(X,Y)} : n \geq 0 \right).$$

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 $(B_n : n \geq 0)$ 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

$$\frac{1}{n} \log ||B_1 B_2 \cdots B_n|| \to \lambda \text{ a.s.} \quad (3)$$

as $n \to \infty$. Furthermore,

$$\begin{aligned} \lambda &= \lim_{n \to \infty} \frac{1}{n} \mathrm{E} \log ||B_1 \cdots B_n|| \quad (4) \\ &= \inf_{n \geq 1} \frac{1}{n} \mathrm{E} \log ||B_1 \cdots B_n||. \quad (5) \end{aligned}$$

The standard proof of Theorem 1 is based on the sub-additive ergodic theorem due to Kingman [11].

Note that for $||A||_\infty \stackrel{\Delta}{=} \max\{\sum_{c_1} |A(c_0, c_1)| : c_0 \in \mathcal{C}\}$,

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

The positivity of r therefore guarantees that

$$\begin{aligned} &\frac{1}{n} \mathrm{E} \log (r G_{(X_0, Y_0)} G_{(X_1, Y_1)} \cdots G_{(X_n, Y_n)} e) \quad (6) \\ &- \; \frac{1}{n} \mathrm{E} \log ||G_{(X_0, Y_0)} G_{(X_1, Y_1)} \cdots G_{(X_n, Y_n)}||_\infty \to 0 \end{aligned}$$

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).

With the channel model described by A1-A3, 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 $G_x^X = (G_x^X(c_0, c_1) : c_0, c_1 \in \mathcal{C})$, $G_y^Y = (G_y^Y(c_0, c_1) : c_0, c_1 \in \mathcal{C})$, and $G_{(x,y)}^{(X,Y)} = (G_{(x,y)}^{(X,Y)}(c_0, c_1) : c_0, c_1 \in \mathcal{C})$ 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 A1-A3. 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:

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

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

C. Entropy, Lyapunov Exponents, and Typical Sequences

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

$$P(X_1,\ldots,X_n) = rG^X_{X_1}G^X_{X_2}\cdots G^X_{X_n}e,$$

where r is the stationary distribution for the channel C. Using Proposition 1 and (6) we can interpret the Lyapunov exponent λ_X as the average exponential rate of growth for a random dynamic system that describes the probability of the sequence X. Since $P(X_1,\ldots,X_n)\to 0$ as $n\to\infty$ for any non-trivial 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 [3], the typical set A^n_ϵ is the set of sequences $x_1,\ldots,x_n$ satisfying

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

and $P(A^n_\epsilon) > 1-\epsilon$ 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 probability growth rate intuition will be useful in understanding the results presented in the next section where we show that $\lambda(X)$ can also be viewed as an expectation rather than an asymptotic quantity.

III. A Markov Chain Representation for Lyapunov Exponents

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 particular 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 [7]. Since our matrices are not i.i.d. we will use a slightly modified version of this Markov chain, namely

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

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 $wG^X_{X_1}\cdots G^X_{X_n}$ 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

$$\tilde{p}_n = \frac{wG^X_{X_1}\cdots G^X_{X_n}}{||wG^X_{X_1}\cdots G^X_{X_n}||}$$

rather than the vector itself because

$$wG^X_{X_1}\cdots G^X_{X_n} \to 0 \text{ as } n\to\infty,$$

but we expect some sort of non-trivial steady-state behavior for the normalized version.

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

$$Z_n = (\tilde{p}_n, C_n, C_{n+1}) \Rightarrow Z_\infty \stackrel{\Delta}{=} (\tilde{p}_\infty, C_\infty, \tilde{C}_\infty) \qquad (7)$$

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

$$\tilde{p}_n \stackrel{\Delta}{=} \frac{wG^X_{X_1}\cdots G^X_{X_n}}{||wG^X_{X_1}\cdots G^X_{X_n}||} = \frac{\tilde{p}_{n-1}G^X_{X_n}}{||\tilde{p}_{n-1}G^X_{X_n}||} \qquad (8)$$

for $n\ge 1$. If w is positive, the same argument as that leading to (6) shows that

$$\frac{1}{n}\log||G^X_{X_1}\cdots G^X_{X_n}|| - \frac{1}{n}\log||wG^X_{X_1}\cdots G^X_{X_n}|| \to 0 \text{ a.s.} \qquad (9)$$

as $n\to\infty$, which implies

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

Furthermore, it is easily verified that

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

Relations (9) and (10) together guarantee that

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

In view of (7), this suggests that

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

where $q(x|c_0,c_1) \stackrel{\Delta}{=} \sum_y q(x,y|c_0,c_1)$. Recall the above discussion regarding the intuitive interpretation of Lyapunov exponents and entropy and suppose we apply the 1-norm, given by $||w||_1 \stackrel{\Delta}{=} \sum_c |w(c)|$, 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 (7) (it turns out that existence of a stationary distribution is not automatic), computing the Lyapunov exponent effectively amounts to computing the stationary distribution of the Markov chain Z.

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

A. The Connection to Hidden Markov Models

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 and assume that $\tilde{p}_0 = r$, it turns out that the first component of Z can be viewed as the standard prediction filter for the channel given the input symbol sequence X, and this prediction filter is itself Markov. We state these results in the following propositions that are proved in [9].

Proposition 2: Assume A1-A3, and let $w = r$, the stationary distribution of the channel C. Then, for $n \geq 0$ and $c \in \mathcal{C}$,

$$\tilde{p}_n(c) = P(C_{n+1} = c|X_1^n). \tag{13}$$

Proposition 3 Assume A1-A3 and suppose $w = r$. Then, the sequence $\tilde{p} = (\tilde{p}_n : n \geq 0)$ is a Markov chain taking values in the continuous state space $\mathcal{P} = \{w : w \geq 0, ||w||_1 = 1\}$. Furthermore,

$$||\tilde{p}_n G_x^X||_1 = P(X_{n+1} = x|X_1^n). \tag{14}$$

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

$$-\mathrm{E}\log(||\tilde{p}_{j-1}G_{X_j}^X||_1) = H(X_{j+1}|X_1^j),$$

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;

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

In addition, an expected value representation of $H(X)$, similar in spirit to (12), is a well known result in the hidden Markov model literature [5]. 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 $\log(||p_{j-1}G_{X_j}^X||_1)$ 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.

B. Pathologies of the Prediction Filter

It turns out that the prediction filter $\tilde{p}_n$ with $w = r$ and the general process Z can be extremely ill-behaved Markov chains. In general, these filter processes are NOT irreducible in even the weakest sense (i.e. Harris recurrence). [See [13] 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 Z defined on the space $\mathcal{P}$ x $\mathcal{C}$ x $\mathcal{C}$ with Borel sets $\mathcal{B}(\mathcal{P}$ x $\mathcal{C}$ x $\mathcal{C})$. Define τ_A as the first return time to the set $A \in \mathcal{P}$ x $\mathcal{C}$ x $\mathcal{C}$. Then, the Markov chain Z is ϕ-irreducible if there exists a non-trivial measure ϕ on $\mathcal{B}(\mathcal{P}$ x $\mathcal{C}$ x $\mathcal{C})$ such that for every state $z \in \mathcal{P}$ x $\mathcal{C}$ x $\mathcal{C}$

$$\phi(A) > 0 \Rightarrow P_z(\tau_A < \infty) > 0. \tag{15}$$

However, the Markov chain Z is never irreducible, as illustrated by the following example. Suppose that the output symbol process Y is binary, hence the random matrices $G_{Y_n}^Y$ can only take two values, say G_0^Y and G_1^Y corresponding to output symbols 0 and 1, respectively. Suppose we initialize $\tilde{p}_0 = r$ and examine the possible values for $\tilde{p}_n^r$. Notice that for any n, the random vector $\tilde{p}_n^r$ can take on only 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 another initial vector belonging to $\mathcal{P}$, call it $w \neq r$ for which the support of the corresponding $\tilde{p}_n^w$'s are disjoint from the support for the $\tilde{p}_n^r$'s for all $n \geq 0$. This contradicts (15). Hence, the Markov chain Z has infinite memory and is not irreducible.

This technical difficulty means that we cannot apply the vast number of results available from the standard theory of Harris recurrent Markov chains. For example, we cannot automatically state

1) the Markov chain Z has a steady-state distribution;
2) the rate of convergence to steady-state;
3) if the steady-state distribution is unique;
4) if the steady-state distribution and entropy are continuous functions of the input symbol distribution and channel transition probabilities.

All of the above issues are critical if we wish to use the filter Markov chain in proof techniques and for the purposes of computing entropy and mutual information. In the following sections we will address the above problems by exploiting the connection between Lyapunov exponents, products of random matrices, and Shannon entropy.

Before we move on, we should note that the authors of [12] point out an important exception to this irreducibility 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 Z is in fact irreducible and standard theory applies. However, since we wish to consider any finite state channel, including those with finite symbol sets, we cannot appeal to existing Harris chain theory.

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. We address the rate of convergence and continuity issues in Section V.

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 $p_{n+1} = \frac{p_n G_{\tilde{X}}^X}{||p_n G_{\tilde{X}}^X||_1}$.
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 $p^w = (p_n^w : n \geq 0)$ denote the output of Algorithm A when $p_0 = w$. Proposition 3 proves that for $w = r$, p^r coincides with the sequence $\tilde{p}^r = (\tilde{p}_n^r : n \geq 0)$, where $\tilde{p}^w = (\tilde{p}_n^w : n \geq 0)$ for $w \in \mathcal{P}$ is defined by the recursion (also known as the forward Baum equation)

$$\tilde{p}_n^w = \frac{w G_{X_1}^X \cdots G_{X_n}^X}{||w G_{X_1}^X \cdots G_{X_n}^X||_1}, \tag{16}$$

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 (see [9] for a proof).

Proposition 4: Assume A1-A3. Then, if $w \in \mathcal{P}$,

$$p_n^w = \frac{w G_{X_1(w)}^X \cdots G_{X_n(w)}^X}{||w G_{X_1(w)}^X \cdots G_{X_n(w)}^X||_1} \tag{17}$$

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,

$$P(X_1(w) = x_1, \ldots, X_n(w) = x_n) = w G_{x_1}^X \cdots G_{x_n}^X e. \tag{18}$$

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 *non-stationary 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 "hidden Markov models". 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 the 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

$$H(X) = -\sum_{x\in\mathcal{X}} \mathrm{E}\log(||p_\infty G_x^X||_1)||p_\infty G_x^X||_1, \qquad (19)$$

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

Theorem 2: Assume A1-A3 and let $\mathcal{P}^+ = \{w \in \mathcal{P} : w(c) > 0, c \in \mathcal{C}\}$. Then,

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

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

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

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

Proof: See [9] for details. The result follows from a straightforward application of Birkhoff's ergodic theorem and the Monotone Convergence Theorem.

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 A1-A3.

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

$$R = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad (20)$$

and

$$G_1^X = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \qquad G_2^X = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}. \qquad (21)$$

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

$$\pi_1((\frac{1}{2}, \frac{1}{2})) = 1 \qquad (22)$$

and

$$\pi_2((0, 1)) = \pi_2((1, 0)) = \frac{1}{2}. \qquad (23)$$

Theorem 2 leaves open the possibility that stationary distributions with support on the boundary of $\mathcal{P}$ will fail to satisfy (19). Furstenberg and Kifer [7] discuss the behavior of $p = (p_n : n \geq 0)$ when p has multiple stationary distributions, some of which violate (19) (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:

A4: $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$.

A5: 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 A1-A5. Then, $p = (p_n : n \geq 0)$ possesses a stationary distribution π.

Proof: See [9] for details. The row allowability of $G_{(x,y)}^{(X,Y)}$ allows us to show that p is a Feller chain on a compact space, and therefore it possesses a stationary distribution.

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. For example, Peres [14], [15], provides examples of discontinuity and lack of smoothness in the Lyapunov exponent as a function of the input symbol distribution when the random matrices have a sparsity structure like that of this section. These examples suggest strongly that entropy can be discontinuous in the presence of A1-A5. We will alleviate these problems in the next section through additional assumptions on the aperiodicity 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 is rather technical. Hence we will first summarize the results of this section and then explain 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 $q(x, y|c_0, c_1)$ 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 $G_{X_1(w)}^X \cdots G_{X_n(w)}^X$ can be split into a product of *strictly positive* random matrices. We then exploit the fact that strictly positive matrices are strict contractions on $\mathcal{P}^+ = \{w \in \mathcal{P} : w(c) > 0, c \in \mathcal{C}\}$ 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 $\lambda(X)$ is a continuous function of both the transition matrix R and the symbol probabilities $q(x, y|c_0, c_1)$.

A. The Contraction Property of Positive Matrices

We assume here that:

A6: The transition matrix R is aperiodic.

A7: For each $(c_0, c_1, x, y) \in \mathcal{C}^2$ x $\mathcal{X}$ x $\mathcal{Y}$, $q(x, y|c_0, c_1) > 0$ whenever $R(c_0, c_1) > 0$.

Under A6-A7, all the matrices $\{G_x^X, G_y^Y, G_{(x,y)}^{(X,Y)} : x \in \mathcal{X}, y \in \mathcal{Y}\}$ exhibit the same (aperiodic) sparsity pattern as R. That is, the matrices have the same pattern of zero and non-zero elements. Note that under A1 and A6, R^l is strictly positive for some finite value of l. So,

$$G_{X_{(j-1)l+1}}^X \cdots G_{X_{jl}}^X \tag{24}$$

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

$$d(v, w) = \log \left(\frac{\max_c (v(c)/w(c))}{\min_c (v(c)/w(c))} \right). \tag{25}$$

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 [16]. For any non-negative matrix T, let

$$\tau(T) = \frac{1 - \theta(T)^{-1/2}}{1 + \theta(T)^{-1/2}}, \tag{26}$$

where

$$\theta(T) = \max_{c_0, c_1, c_2, c_3} \left(\frac{T(c_0, c_3) T(c_1, c_4)}{T(c_0, c_4) T(c_1, c_3)} \right). \tag{27}$$

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,

$$d(vT, wT) \leq \tau(T) d(v, w). \tag{28}$$

For a proof, see pages 100-110 of Seneta [16]. 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 $n \geq l$, $\tilde{p}_n^r$ and $\tilde{p_n}^w$ both lie in $\mathcal{P}^+$, so $d(\tilde{p}_n^r, \tilde{p}_n^w)$ is well-defined for $n \geq l$. Proposition 5 will allow us to show that $\tilde{p}^w = (\tilde{p}_n^w : n \geq 0)$ has a unique stationary distribution. Proposition 6 will allow us to show that $p^w = (p_n^w : n \geq 0)$ must have the same stationary distribution as $\tilde{p}^w$.

Proposition 5: Assume A1-A4 and A6-A7. If $w \in \mathcal{P}$, then

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

where $\alpha \stackrel{\Delta}{=} -(\log \beta)/l$ and

$$\beta \stackrel{\Delta}{=} \max\{\tau(G_{x_1}^X \cdots G_{x_l}^X) : p(X_1 = x_1, \ldots, X_l = x_l) > 0\}.$$

Proposition 6: Assume B1-B4 and B6-B7. For $w \in \mathcal{P}$, there exists a probability space upon which $d(p_n^w, \tilde{p}_n^r) = \mathrm{O}(e^{-\alpha n})$ a.s. as $n \to \infty$.

The proofs of Propositions 5 and 6 rely on Proposition 4 and a coupling argument that we will summarize here (see [9] for the details). Recall from Proposition 4 that we can view $\tilde{p}_n^r$ and p_n^w as being generated by a stationary and non-stationary version of the channel C, respectively. The key idea is that along each sample path the non-stationary version of the channel will eventually couple with the stationary version. Once the channels couple then the non-stationary version of the symbol sequence $X(w)$ will also couple with the stationary version X. When this coupling occurs, say at time $T < \infty$, the symbol sequences $(X_n(w) : n > T)$ and $(X_n : n > T)$ will be identical. This means that for all $n > T$ the matrices applied to $\tilde{p}_n^r$ and p_n^w will also be identical. This allows us to apply the contraction result from Theorem 4 and complete the proofs. Note that without the contraction property of positive matrices we would not be able to prove this convergence argument due to the infinite memory problem discussed earlier.

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 $w \in \mathcal{P}$, $p_n^w \Rightarrow p_\infty^*$ 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 [4]. We leave the technical details of the argument for [9] and present a summary of the proof technique here.

Let $X = (X_n : -\infty < n < \infty)$ be a doubly-infinite stationary version of the input symbol sequence, and put $\chi_n = X_{-n}$ for $n \in \mathcal{Z}$. Then,

$$r G_{X_1}^X \cdots G_{X_n}^X \stackrel{\mathcal{D}}{=} r G_{\chi_n}^X \cdots G_{\chi_1}^X, \tag{29}$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. Put $p_0^* = r$ and

$$p_n^* = \frac{r G_{\chi_n}^X \cdots G_{\chi_1}^X}{||r G_{\chi_n}^X \cdots G_{\chi_1}^X||},$$

for $n \geq 0$.

The process $(p_n^* : n \geq 0)$ is a time-reversed version of $\tilde{p}^r$, such that $\tilde{p}_n^r \stackrel{\mathcal{D}}{=} p_n^*$ for all $n \geq 0$. Take careful note of the order in which matrices are applied to p_n^* (they are applied in reverse). Hence, the matrix observed at time n is applied *in front* of the matrix products rather than at the end (as is the case with $\tilde{p}_n^r$). Then, if we look at consecutive values of p_n^* we can show that it is a.s. a Cauchy sequence (through the contraction property). Hence, there exists a random variable p_∞^* such that $p_n^* \to p_\infty^*$ a.s. as $n \to \infty$. This guarantees weak convergence of the forward process $\tilde{p}^r$ to a random variable with the same distribution as p_∞^*. This reversal argument is remarkably useful since we have no means of directly showing that the forward process converges. In [9] we provide the detail to prove the following theorem.

Theorem 5: Assume A1-A4 and A6-A7. Then,

i.) $p = (p_n : n \geq 0)$ has a unique stationary distribution π.
ii.) There exists a set $K \in \mathcal{P}^+$ such that $\pi(K) = 1$ 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 $w \in K$.

Applying Theorem 2, we may conclude that under A1-A4 and A6-A7, the channel $\mathcal{P}$-chain has a unique stationary distribution π on $\mathcal{P}^+$ satisfying

$$H(X) = -\sum_{x \in \mathcal{X}} \int_{\mathcal{P}} \log(||wG^X_x||_1)||wG^X_x||_1 \pi(dw). \quad (30)$$

We can also use our Markov chain machinery to establish continuity of the entropy $H(X)$ as a function of R and q. Such a continuity result is of theoretical importance in optimizing the mutual information between X and Y, which is necessary to compute channel capacity. The following theorem generalizes a continuity result of [8] obtained in the setting of i.i.d. input symbol sequences.

Theorem 6: Assume A1-A4 and A6-A7. Suppose that $(R_n : n \geq 1)$ is a sequence of transition matrices on $\mathcal{C}$ for which $R_n \to R$ as $n \to \infty$. Also, suppose that for $n \geq 1$, $q_n(\cdot|c_0, c_1)$ is a probability mass function on $\mathcal{X}$ x $\mathcal{Y}$ for each $(c_0, c_1) \in \mathcal{C}^2$ and that $q_n \to q$ as $n \to \infty$. If $H_n(X)$ is the entropy of X associated with the channel model characterized by (R_n, q_n), then $H_n(X) \to H(X)$ as $n \to \infty$.

Proof: See [9] for the details. With Theorem 5 in hand we need a standard weak convergence argument for the expected value of a continuous function on a compact set.

Our final result presents a simple and intuitive connection between the stationary prediction filter p^*_∞, the Lyapunov exponent (or entropy), and the random matrix process G^X_X. It turns out, that these three quantities are related through the following *random* eigenvector-eigenvalue relation:

$$pG^X_X \stackrel{\mathcal{D}}{=} \gamma p', \quad (31)$$

where $p \stackrel{\mathcal{D}}{=} p' \stackrel{\mathcal{D}}{=} p^*_\infty$, and $E_\pi \log \gamma = -H(x)$, where π is the stationary distribution from Theorem 5. Hence, the stationary filter random variable is a random eigenvector for the random matrix G^X_X. The associated random eigenvalue γ then determines the Lyapunov exponent. A proof of this result follows directly from [1].

VI. Conclusions

This work examines some of the theoretical issues that arise when investigating the connections between Lyapunov exponents, Shannon entropy, and hidden Markov models. We first demonstrated that entropies for finite state channels are equivalent to Lyapunov exponents for products of random matrices. We then showed that entropy can be expressed as an expectation with respect to the stationary distribution of a Markov chain that is closely related to the hidden Markov prediction filter. As a consequence, we proved that the prediction filter is a non-irreducible Markov chain. By applying tools from the theory of random matrix products we were able to provide new regularity results for the prediction filter, even though the filter Markov chain is not Harris recurrent. The theoretical tools presented here may also be used to compute entropy and mutual information for finite-state Markov channels. The non-irreducibility problems discussed above create many problems in simulation-based algorithms for computation, and we also address these issues in [9].

Finally, we should note that this connection between Lyapunov exponents and Information Theory is just a first step. We now have access to a wide range of tools from statistical mechanics that can be applied to problems in Information Theory and hidden Markov models. Clearly, a significant amount of translation work between the languages of the different fields needs to occur, but the mathematical connections and potential results appear to be promising. Indeed, it might now be possible to fully extend the notion of a "Thermodynamic Formalism" [2] to Shannon's Theory of Information – thereby permitting information theoretic interpretations of many quantities in quantum physics and dynamic systems.

References

[1] L. Arnold, L. Demetrius, and M. Gundlach, "Evolutionary formalism for products of positive random matrices", *Annals of Applied Probability 4*, pp.859-901, 1994.
[2] C. Beck and F. Schlgl, "Thermodynamics of Chaotic Systems: An Introduction", *Cambridge Nonlinear Science Series* , vol. 4, Cambridge UP, 1993.
[3] T. Cover, J. Thomas, "Elements of Information Theory", John Wiley & Sons, 1991.
[4] P. Diaconis, D.A. Freedman, "Iterated random functions", *SIAM Review* , vol. 41, pp. 45-67, 1999.
[5] Y. Ephraim and N. Merhav, "Hidden Markov processes", *IEEE Trans. Information Theory* , vol. 48, pp. 1518-1569, June 2002.
[6] H. Furstenberg, H. Kesten, "Products of Random Matrices", *Ann. Math. Statist.*, pp.457-469, 1960.
[7] H. Furstenberg and Y. Kifer, "Random matrix products and measures on projective spaces", *Isreal J. Math.* , vol.46, pp.12-32, 1983.
[8] A. Goldsmith, P. Varaiya, "Capacity, mutual information, and coding for finite state Markov channels", *IEEE Trans. Information Theory* , vol. 42, pp. 868-886, May 1996.
[9] T. Holliday, P. Glynn, A. Goldsmith, "On Entropy and Lyapunov Exponents for Finite-State Channels", *submitted to IEEE Trans. on Information Theory*, available at http://wsl.stanford.edu .
[10] T. Holliday, A. Goldsmith, P. Glynn, Capacity of Finite State Markov Channels with General Inputs, *IEEE ISIT 2003* , Yokohama, Japan, 2003.
[11] J. Kingman, "Subadditive ergodic theory", *Annals of Probability* , 1:883–909, 1973.
[12] F. LeGland, L. Mevel, "Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models", *Mathematics of Control, Signals and Systems* , 13, 1, pp. 63-93, 2000.
[13] S. Meyn, R. Tweedie, "Markov Chains and Stochastic Stability", Springer Verlag Press, 1994.
[14] Y. Peres, "Domains of analytic continuation for the top Lyapunov exponent", *Ann. Inst. H. Poincar Probab. Statist.* , vol. 29, pp.131-148, 1992.
[15] Y. Peres, K. Simon, B. Solomyak, "Absolute continuity for random iterated function systems with overlaps", /em preprint /em, 2005.
[16] E. Seneta, "Non–negative matrices and Markov chains", Springer-Verlag Press, second edition, 1981.