# Entropy and Mutual Information for Markov Channels with General Inputs

Tim Holliday, Andrea Goldsmith, Peter Glynn
Stanford University

**Abstract**

We study new formulas based on Lyapunov exponents for entropy, mutual information, and capacity of finite state discrete time Markov channels. We also develop a method for directly computing mutual information and entropy using continuous state space Markov chains. Our methods allow for arbitrary input processes and channel dynamics, provided both have finite memory. We show that the entropy rate for a symbol sequence is equal to the primary Lyapunov exponent for a product of random matrices. We then develop a continuous state space Markov chain formulation that allows us to directly compute entropy rates as expectations with respect to the Markov chains' stationary distributions. We also show that the stationary distributions are continuous functions of the input symbol dynamics. This continuity facilitates optimization of the mutual information and allows the channel capacity to be written in terms of Lyapunov exponents.

# 1 Introduction

In this work we develop the theory to compute entropy and mutual information for Markov channels with general inputs. The results are extensions of the work by Goldsmith and Varaiya for finite-state Markov channels with i.i.d. inputs [3], although the methods by which we obtain our results are quite different than those used in [3]. We model the wireless 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).

Previous research on this case is essentially limited to [3], which extends the work in [6]. In [3] the authors restrict their analysis to channels where the channel transition probabilities do not depend on the input/output symbols (e.g. channels without ISI) and i.i.d. inputs. Under these restrictions the authors prove several key results. First, they show that two conditional probability distributions for the channel state – the channel distribution conditioned on all past outputs and the channel distribution conditioned on all past input/output pairs – are continuous state space DTMCs. Second, they show that these Markov chains posses unique stationary distributions that are continuous functions of the i.i.d. input distribution. Finally, the authors show that the symbol entropy rates and capacity of the channel can be computed using the stationary distributions for these Markov chains. We will review these results in more detail in the next section.

The approach in [3] does have limitations. In particular, the authors point out that the assumption of i.i.d. inputs cannot be easily relaxed (indeed, they provide an example where their methods fail using Markov rather than i.i.d. inputs). The framework in [3] also fails when channel transitions depend on the input/output sequences.

Our main contribution in this paper is to extend the results of [3] to allow computation of entropy and mutual information rates for Markov channels with symbol dependent dynamics and

general inputs. We prove these extensions using several results from chaotic dynamic systems and stochastic processes. The first, and perhaps most interesting, result is that the entropy rate of a symbol sequence generated by a Markov channel is equivalent to the primary Lyapunov exponent for a product of random matrices. Second, we show that while the probability distributions of the channel conditioned on past symbol sequences are not necessarily Markov chains, we can construct augmented processes from these probabilities that are Markov chains. We then use existence and uniqueness results from Lyapunov exponent theory, as well as theory for general state space Markov chains, to prove several theorems for the augmented chains. In particular we show that for the case of general inputs:

- the entropy rates of the input, output, and input/output sequences can be computed as expectations with respect to the stationary distributions of the augmented Markov chains,
- the mutual information rate is a continuous function of the dynamics driving the (non-i.i.d.) input sequences and can be optimized to compute capacity,
- while the augmented Markov chains may have multiple stationary measures, the entropy rates computed with respect to any of the stationary measures will be unique.

Due to length constraints we do not include detailed proofs of some of the more technical results. The details can be found in [4].

The rest of this paper is organized as follows. In the next section we discuss previous research and specify our channel and input/output symbol models. In Section 3 we develop the connection between entropy rates and Lyapunov exponents. In Section 4 we construct augmented Markov chains using the conditional channel probability distributions and show that we can compute entropy rates as a function of the Markov chains' stationary distributions. We conclude with a discussion of further research in Section 5.

# 2 Previous Research and Channel Model

Let $Z = (Z_n : n \geq 0)$ be a stationary finite-state irreducible Markov chain living on state space $\mathcal{Z}$. The random sequences of observed inputs and outputs will be denoted $X = (X_n : n \geq 0)$ and $Y = (Y_n : n \geq)$, and take values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. We will use the notation

$$S_m^n \triangleq (S_m, S_{m+1}, \ldots, S_{m+n}) \tag{1}$$

for the sequences $X_n$, $Y_n$, and $Z_n$. We will denote random variables or sequences using capital letters and deterministic variables or sequences using lower case letters. Deterministic matrices will be bold upper case letters and random matrices will be upper case Greek letters.

## 2.1 Previous Research

In [3] the authors study a simplified version of the problem we consider here. However, it is useful for us to review their results since they are similar to some of the results derived herein. In this sub-section we will review the setup and results of [3]. As we will see later, while the proofs cannot be easily generalized, many of the results in [3] can be extended to the general models considered here.

In [3] the authors consider the traditional Markov channel model with i.i.d. inputs. This Markov channel is defined by its conditional input/output probabilities, $p(y_n|x_n, z_n)$. By assumption, the channel state at time $n+1$ is independent of previous symbols given the channel state at time $n$, hence $p(z_{n+1}|z_n, x^n, y^n) = p(z_n, z_{n+1})$. Each of the states $z \in \mathcal{Z}$ corresponds to a memoryless channel and therefore $p(y_{n+1}|z_{n+1}, z^n, x^n, y^n) = p(y_{n+1}|z_{n+1}, x_{n+1})$. If the inputs

$x_n$ are i.i.d. then $p(y_{n+1}, x_{n+1}|z_{n+1}, z^n, x^n, y^n) = p(x_{n+1})p(y_{n+1}|x_{n+1}, z_{n+1})$. Finally, using these assumptions we can state that

$$p(x^N, y^N|z^N) = \prod_{n=1}^{N} p(x_n)p(y_n|x_n, z_n) \quad \text{and} \quad p(y_{n+1}|z_{n+1}, z^n, y^n) = p(y_{n+1}|z_{n+1}). \tag{2}$$

Under these assumptions the authors of [3] show several key results, which we will state here without proof. Define the vector $\pi_n$ as the conditional state distribution for the channel given all past input/output symbols, $\pi_n(z) = p(Z_n = z|X^{n-1}, Y^{n-1})$. Then $\pi_n$ is a continuous state space Markov chain. Furthermore, the evolution equation for $\pi_n$ can be written as

$$\pi_{n+1} = \frac{\pi_n \Lambda_{(X_n,Y_n)} P}{\pi_n \Lambda_{(X_n,Y_n)} e}, \tag{3}$$

where $e = [1, \cdots, 1]^T$, $P$ is the transition matrix for the channel, and $\Lambda_{(X_n,Y_n)}$ is an i.i.d. **random** diagonal matrix with $k$th diagonal term $p(Y_n|X_n, Z_n = k)$. Likewise, $\rho_n(k) = p(Z_n = k|Y^{n-1})$ is also a Markov chain with a similar evolution equation

$$\rho_{n+1} = \frac{\rho_n \Phi_{(X_n,Y_n)} P}{\rho_n \Phi_{(X_n,Y_n)} e}, \tag{4}$$

where $\Phi(k, k) = p(Y_n|Z_n = k)$.

Then according to [3], for the case of i.i.d. inputs $\pi_n$ and $\rho_n$ are Markov chains with unique stationary distributions. Furthermore, the following entropy rates can be written as functions of these Markov chains

$$H(Y_n|X_n, X^{n-1}, Y^{n-1}) = E\left[-\log \sum_{z\in\mathcal{Z}} p(y_n|x_n, Z_n = z)\pi_n(z)\right] = H(Y_n|X_n, \pi_n) \tag{5}$$

$$H(Y_n|Y^{n-1}) = E\left[-\log \sum_{z\in\mathcal{Z}} p(y_n|Z_n = z)\rho_n(z)\right] = H(Y_n|\rho_n). \tag{6}$$

Therefore, we can compute the mutual information rate $I(X, Y)$ as

$$I(X, Y) = \lim_{n\to\infty} \frac{1}{n} I(X^n, Y^n) = \lim_{n\to\infty} \frac{1}{n} \left(H(Y^n) - H(Y^n|X^n)\right) \tag{7}$$

$$= \lim_{n\to\infty} \frac{1}{n} \left(\sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X_i, Y^{i-1}, X^{i-1})\right) \tag{8}$$

$$= \lim_{n\to\infty} \frac{1}{n} \left(\sum_{i=1}^{n} H(Y_i|\rho_i) - H(Y_i|X_i, \pi_i)\right). \tag{9}$$

At this point, the authors of [3] stop to point out that this is a new formula for mutual information. The significance of this formula is that it can be easily computed since we have evolution equations for $\pi_n$ and $\rho_n$. The authors also show that the stationary distributions for $\pi_n$ and $\rho_n$ are continuous functions of the i.i.d. input distribution for $X_n$. This continuity implies that one can optimize the input distribution to maximize mutual information and find the channel capacity. Notice that the random matrices in the evolution equations for $\pi_n$ and $\rho_n$ must be i.i.d. for $\pi_n$ and $\rho_n$ to be Markov chains. Once the inputs are not i.i.d. the proof techniques from [3] can no longer be used since they cannot guarantee that $\pi_n$ and $\rho_n$ converge to appropriate random variables. If the convergence of $\pi_n$ and $\rho_n$ cannot be assured, then the expectations in (5) and (6) cannot be computed, rendering the convergence of (9) questionable.

However, if we write (5) and (6) in matrix form we find a much more compact representation and can shed some light on how to generalize the formulation to non-i.i.d. inputs. We have from (5) and (6) that

$$\lim_{n\to\infty} \frac{1}{n} \left(\sum_{i=1}^{n} H(Y_i|\rho_i) - H(Y_i|X_i, \pi_i)\right) =$$

$$\begin{aligned} & \lim_{n\to\infty} \frac{1}{n} \left( \sum_{i=1}^{n} E[-\log \pi_n \Lambda_{(X_n,Y_n)} e] - E[-\log \rho_n \Phi_{(X_n,Y_n)} e] \right) \\ = \; & E_{\pi_\infty}\left[-\log \pi \Lambda_{(X,Y)} e\right] - E_{\rho_\infty}\left[-\log \rho \Phi_{(X,Y)} e\right] \\ = \; & E_{\pi_\infty}\left[-\log ||\Lambda_{(X,Y)}||\right] - E_{\rho_\infty}\left[-\log ||\Phi_{(X,Y)}||\right], \end{aligned}$$

where the expectations are taken with respect to the stationary distributions of $\pi_n$ and $\rho_n$. We state this representation without proof since we will prove the general case in the next section.

Although the authors of [3] did not realize it, this matrix representation shows that the entropy rates $H(Y)$ and $H(Y|X)$ are equivalent to the Lyapunov exponents for two products of random matrices. This connection will allow us to extend the results of [3] since existence and uniqueness results for Lyapunov exponents do not require the random matrices in (3) and (4) to be i.i.d. or diagonal. In the following sections we will detail the equivalence between entropy rates and Lyapunov exponents, and then proceed to extend the Markov chain results from [3] to general channels. However, we must first define our model for general Markov channels with general inputs.

## 2.2 Channel Model

It is important to note that while we refer to the process $Z$ as "the channel", we are not restricted to using the states of $Z$ to model states in a wireless channel. Indeed, we must permit a general structure for $Z$ in order to model channels that are more complex than those considered in [3]. For example, $Z$ could conceptually represent two Markov chains: the first modeling a standard Gilbert-Elliot channel and the second modeling a Markov process that assigns probability distributions to the inputs (i.e. Markov modulated inputs). We can also model ISI channels by introducing symbol memory into $\mathcal{Z}$. That is, a state $z \in \mathcal{Z}$ could correspond to both a channel and a finite number of previous inputs or outputs. Our only restriction is that $|\mathcal{X}|$, $|\mathcal{Y}|$, and $|\mathcal{Z}|$ must be finite.

In order to model a sufficient level of generality we must use a non-standard formulation of the Markov channel. Readers are likely familiar with the traditional Markov channel model we discussed above, where each state in the channel Markov chain corresponds to a memoryless channel. That is, each state $z_n \in \mathcal{Z}$ corresponds to a probability distribution for the outputs given an input, $p(y_n|x_n, z_n)$. We will add two layers of generality to this model. First we will allow the input symbol probabilities to depend on the channel state, this changes the above probability to $p(x_n, y_n|z_n)$. Second, we will allow for non-causal dependence of the input/output symbols with respect to the channel. This adds one extra term such that for each *pair* of states $(z_n, z_{n+1}) \in \mathcal{Z}$ we have $p(x_n, y_n|z_n, z_{n+1})$. We introduce the non-causal dependence in order to facilitate our Lyapunov exponent formulation, it is unlikely that this dependence will be of any practical use when computing entropy or mutual information rates. Clearly the traditional channel model mentioned above can be formulated as a special case of what we propose here. Finally, by these assumptions, the input/output sequences $X$ and $Y$ have a joint distribution specified by

$$p(X_0 = x_0, Y_0 = y_0, \ldots, X_n = x_n, Y_n = y_n|z^{n+1}) = \tag{10}$$

$$\prod_{i=0}^{n} p(x_i, y_i|Z_i = z_i, Z_{i+1} = z_{i+1}). \tag{11}$$

# 3 Entropy Rates and Products of Random Matrices

Our goal is to compute the mutual information rate $I(X, Y)$ for the symbol sequences $X$ and $Y$. In this paper we will use the relation $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where

$$H(X) \;=\; -\lim_{n\to\infty} \frac{1}{n} \log p_{X,n}(X_0, \ldots, X_n) \tag{12}$$

$$H(Y) = -\lim_{n\to\infty}\frac{1}{n}\log p_{Y,n}(Y_0,\ldots,Y_n) \tag{13}$$

$$H(X,Y) = -\lim_{n\to\infty}\frac{1}{n}\log p_{(X,Y),n}((X_0,Y_0),\ldots,(X_n,Y_n)). \tag{14}$$

This reduces our mutual information problem to one of computing entropy rates. Likewise, the resulting capacity problem is

$$C = \lim_{n\to\infty}\max_{p(X^n)}\frac{1}{n}I(X^n,Y^n) = \lim_{n\to\infty}\max_{p(X^n)}\frac{1}{n}(H(X^n)+H(Y^n)-H(X^n,Y^n) \tag{15}$$

$$= \max_{p(X)}\left[H(X)+H(Y)-H(X,Y)\right], \tag{16}$$

where the last equality holds provided we can prove the entropy rates are continuous functions of the input dynamics $p(X)$.

In this section we show that the probability of observing a particular input, output, or input/output sequence can be represented as a product of random matrices. Then we show that the entropy rate of each of these sequences can be computed as the primary Lyapunov exponent for the random matrix product. Finally we present a relevant theorem from Lyapunov exponent theory regarding the existence and uniqueness of such quantities. The continuity issue and resulting capacity formula will be addressed in Section 4.

## 3.1 Entropy Rates and Random Matrices

Much of the work required to show the connection between entropy rates and Lyapunov exponents is simply manipulation of definitions. Let $p_{X,n}(x_0,\ldots,x_n) \stackrel{\Delta}{=} p(X_0=x_0,\ldots,X_n=x_n)$, $p_{Y,n}(y_0,\ldots,y_n) \stackrel{\Delta}{=} p(Y_0=y_0,\ldots,Y_n=y_n)$, and $p_{(X,Y),n}((x_0,y_0),\ldots,(x_n,y_n)) \stackrel{\Delta}{=} p(X_0=x_0,Y_0=y_0,\ldots,X_n=x_n,Y_n=y_n)$. Using our above assumptions we have

$$p_{(X,Y),n}(x_0,y_0,\ldots,x_n,y_n) = \sum_{z_0,\ldots z_{n+1}} p(z_0)\prod_{j=0}^{n} p(x_j,y_j|z_j,z_{j+1})p(z_j,z_{j+1})$$

$$p_{X,n}(x_0,\ldots,x_n) = \sum_{z_0,\ldots z_{n+1}} p(z_0)\prod_{j=0}^{n} p_X(x_j|z_j,z_{j+1})p(z_j,z_{j+1})$$

$$p_{Y,n}(y_0,\ldots,y_n) = \sum_{z_0,\ldots z_{n+1}} p(z_0)\prod_{j=0}^{n} p_Y(y_j|z_j,z_{j+1})p(z_j,z_{j+1}),$$

where $p_X(x,z_j,z_{j+1}) = \sum_{y\in\mathcal{Y}} p(x,y,z_j,z_{j+1})$, $p_Y(y,z_j,z_{j+1}) = \sum_{x\in\mathcal{X}} p(x,y,z_j,z_{j+1})$, and the $p(z_j,z_{j+1})$ are the transition probabilities for $Z$.

Now we can express $p_{(X,Y),n}$, $p_{X,n}$, and $p_{Y,n}$ in term of matrices. Set

$$\mathbf{G}_{(\mathbf{x_j},\mathbf{y_j})} = \left(\mathbf{G}_{(\mathbf{x_j},\mathbf{y_j})}(\mathbf{z_j},\mathbf{z_{j+1}}) : z_j, z_{j+1}\in\mathcal{Z}\right)$$

$$\mathbf{G}_{\mathbf{x_j}} = \left(\mathbf{G}_{\mathbf{x_j}}(\mathbf{z_j},\mathbf{z_{j+1}})) : z_j, z_{j+1}\in\mathcal{Z}\right)$$

$$\mathbf{G}_{\mathbf{y_j}} = \left(\mathbf{G}_{\mathbf{y_j}}(\mathbf{z_j},\mathbf{z_{j+1}})) : z_j, z_{j+1}\in\mathcal{Z}\right)$$

where,

$$\mathbf{G}_{(\mathbf{x_j},\mathbf{y_j})}(\mathbf{z_j},\mathbf{z_{j+1}}) = p(x_j,y_j|z_j,z_{j+1})p(z_j,z_{j+1})$$

$$\mathbf{G}_{\mathbf{x_j}}(\mathbf{z_j},\mathbf{z_{j+1}}) = p_X(x_j|z_j,z_{j+1})p(z_j,z_{j+1})$$

$$\mathbf{G}_{\mathbf{y_j}}(\mathbf{z_j},\mathbf{z_{j+1}}) = p_Y(y_j|z_j,z_{j+1})p(z_j,z_{j+1}).$$

That is, each input and output symbol, and input/output symbol pair has an associated matrix. The elements of each matrix are indexed by the states in $\mathcal{Z}$. The $(z_j,z_{j+1})$th entry of each matrix corresponds to the probability of observing the symbol(s) associated with that matrix given $(z_j,z_{j+1})$ multiplied by the probability of $(z_j,z_{j+1})$.

Let $e = [1, \ldots, 1]^T$ and let $\mu = (\mu(z) : z \in \mathcal{Z}, \mu > 0)$ be a row vector for the distribution of $Z_0$. Then we can write $p_{(X,Y),n}$, $p_{X,n}$, and $p_{Y,n}$ in terms of the following matrix products:

$$\begin{aligned} p_{(X,Y),n}((x_0, y_0), \ldots, (x_n, y_n)) &= \mu \mathbf{G}_{(\mathbf{x_0},\mathbf{y_0})} \cdots \mathbf{G}_{(\mathbf{x_n},\mathbf{y_n})} e \\ p_{X,n}(x_0, \ldots, x_n) &= \mu \mathbf{G}_{\mathbf{x_0}} \cdots \mathbf{G}_{\mathbf{x_n}} e \\ p_{Y,n}(y_0, \ldots, y_n) &= \mu \mathbf{G}_{\mathbf{y_0}} \cdots \mathbf{G}_{\mathbf{y_n}} e. \end{aligned}$$

Now we can make the connection between sample-path entropy rates and products of random matrices. We have

$$p_{(X,Y),n}((X_0, Y_0), \ldots, (X_n, Y_n)) = \mu \Gamma_{(X,Y)}(0)\Gamma_{(X,Y)}(1) \cdots \Gamma_{(X,Y)}(n) e, \quad (17)$$

$$p_{X,n}(X_0, \ldots, X_n) = \mu \Gamma_X(0)\Gamma_X(1) \cdots \Gamma_X(n) e, \quad (18)$$

$$p_{Y,n}(Y_0, \ldots, Y_n) = \mu \Gamma_Y(0)\Gamma_Y(1) \cdots \Gamma_Y(n) e, \quad (19)$$

where the $\Gamma_{(X,Y)}(j)$'s, $\Gamma_X(k)$'s, and $\Gamma_Y(l)$'s are random matrices: $\Gamma_{(X,Y)}(j) = \mathbf{G}_{(\mathbf{X_j},\mathbf{Y_j})}$, $\Gamma_X(k) = \mathbf{G}_{\mathbf{X_k}}$, and $\Gamma_Y(l) = \mathbf{G}_{\mathbf{Y_l}}$, which are generated by the sequences $(X, Y)$, $X$, and $Y$. Then the entropy quantities we wish to compute are:

$$\begin{aligned} H(X) &= -\lim_{n\to\infty} \frac{1}{n} \log \mu \Gamma_X(0)\Gamma_X(1) \cdots \Gamma_X(n-1) e \\ H(Y) &= -\lim_{n\to\infty} \frac{1}{n} \log \mu \Gamma_Y(0)\Gamma_Y(1) \cdots \Gamma_Y(n-1) e \\ H(X,Y) &= -\lim_{n\to\infty} \frac{1}{n} \log \mu \Gamma_{(X,Y)}(0)\Gamma_{(X,Y)}(1) \cdots \Gamma_{(X,Y)}(n-1) e. \end{aligned}$$

**Proposition 1:** The entropy rates $H(X)$, $H(Y)$, and $H(X, Y)$ are equivalent to the negative of the primary Lyapunov exponents for products of random matrices

$$H(X) = -\lim_{n\to\infty} \frac{1}{n} \log ||\Gamma_X(0)\Gamma_X(1) \cdots \Gamma_X(n-1)|| = -\lambda_X \quad (20)$$

$$H(Y) = -\lim_{n\to\infty} \frac{1}{n} \log ||\Gamma_Y(0)\Gamma_Y(1) \cdots \Gamma_Y(n-1)|| = -\lambda_Y \quad (21)$$

$$H(X,Y) = -\lim_{n\to\infty} \frac{1}{n} \log ||\Gamma_{(X,Y)}(0)\Gamma_{(X,Y)}(1) \cdots \Gamma_{(X,Y)}(n-1)|| = -\lambda_{XY}, \quad (22)$$

where the random matrices $\Gamma_X$, $\Gamma_Y$, and $\Gamma_{(X,Y)}$ are defined above. The Lyapunov exponent $\lambda$ for a product of random matrices is defined as $\lambda = \lim_{n\to\infty} \frac{1}{n} \log ||\Gamma(0)\Gamma(2) \cdots \Gamma(n-1)||$, where $\lambda$ may be deterministic or random. We will show in the next section that $\lambda$ is deterministic for our problem.

**Proof:** We have already provided the necessary definitions, all we need to show is that $\mu A e = ||A||$, for any matrix $A$. Notice that every element of our random matrices are non-negative and therefore the matrix products are also non-negative. Both $\mu$ and $e$ are strictly positive and therefore $\mu A e$ satisfies the conditions for a matrix norm. That is, $\mu A e \geq 0$ and $\mu A e = 0$ iff $A = 0$, $||\alpha A|| = \alpha ||A||$, and $||A + B|| \leq ||A|| + ||B||$.

## 3.2 Lyapunov Exponents

Some of the first significant results on random matrices were originally produced by Furstenberg and Kesten [2] and later refined by Oseledec [7]. This is still an active area of research, particularly within the statistical physics and mechanics communities. The most basic results in this area of research establish conditions under which a product of random matrices $\Gamma(0)\Gamma(1) \cdots \Gamma(n-1)$ satisfies

$$\frac{1}{n} \log ||\Gamma(0)\Gamma(1) \cdots \Gamma(n-1)|| \to \lambda \text{ a.s. as } n \to \infty, \quad (23)$$

where $\lambda$ is a constant known as the primary Lyapunov exponent. In the dynamic systems context, the Lyapunov exponent represents the exponential rate at which two initially close points diverge. In our setting, the constant $\lambda$ is precisely the entropy we wish to compute, which allows us to employ the entire theory of random matrix products in order to solve our problem. In this paper we will use one major result from this field, which states

**Theorem 1:** If $\Gamma(0), \Gamma(1) \dots \Gamma(n-1)$ form a stationary ergodic sequence and $E\left[\log^+ ||\Gamma(0)||\right] < \infty$ then

$$\lim_{n\to\infty} \frac{1}{n} \log ||\Gamma(0)\Gamma(1)\dots\Gamma(n-1)|| = \lim_{n\to\infty} \frac{1}{n} E[\log ||\Gamma(0)\Gamma(1)\dots\Gamma(n-1)||] = \lambda \text{ a.s..}$$

**Proof:** See [2].

This representation for entropy rates allows us to write a new expression for mutual information and capacity for a Markov channel.

$$I(X,Y) \;=\; \lambda_X + \lambda_Y - \lambda_{XY} \quad \text{and} \quad C = \max_{p(X)} \left[\lambda_X + \lambda_Y - \lambda_{XY}\right], \tag{24}$$

where the capacity formula holds if we can show that the Lyapunov exponents are continuous functions of the input dynamics $p(X)$. Much like the Asymptotic Equipartition Theorem for entropy rates, Theorem 1 provides an extremely useful result by guaranteeing the existence of a deterministic limit $\lambda$ independent of the initial distribution (indeed, the two theorems are related). There are many other results in this research area that provide some interesting insights into this problem as well as other connections between Lyapunov exponents and different forms of entropy. Due to space constraints we will not discuss these issues further, see [4] for details.

Another consequence of Theorem 1 is that we can use simulation to calculate our entropy rates via the following algorithm:

1. First simulate a long sequence for the channel $Z_n$.
2. Based on the simulated channel states simulate a sequence of input/output pairs $(x_n, y_n)$.
3. Compute the matrix products and entropy rates using (23).
4. Finally, compute mutual information using (24).

Simulation is a common method for computing Lyapunov exponents. However, simulation-based methods for computing Lyapunov exponents often have extremely poor convergence properties [1]. That is, while the Lyapunov exponents can be computed using the simulation method given above, there is no method to determine how long the simulation must run in order to achieve a particular confidence interval or error bound on the simulated estimate. Furthermore, optimization via simulation can be very time consuming. In the next section, we discuss a somewhat more elegant method that allows us to directly compute the entropy rates, mutual information, and capacity for a Markov channel.

# 4 Markov Chain Formulation

In [3] the authors show that the probability distribution of the channel conditioned on past inputs and outputs is a continuous state space Markov chain with a unique stationary distribution. They also show that entropy rates and capacity can be directly computed as functions of the stationary distributions for these Markov chains. However, these results rely heavily on the fact that the input sequence $X$ is i.i.d.

In this section we will show that we can use the stationary distributions for augmented versions of the Markov chains in [3] to directly compute the Lyapunov exponents (i.e. entropy

rates) for Markov channels with general inputs. We will show that the Lyapunov exponents are continuous functions of the parameters driving the (non-i.i.d.) input process $X$, which allows us to prove the capacity formula (24). Due to space constraints the existence and continuity results for our Markov chain formulation will be stated without proof. See [4] for these details.

## 4.1 Channel Conditional Probability Distributions

We begin by making the connection between our random matrix product formulation and the probability distribution of the channel conditioned on past observed symbols. Once we observe this connection the Markov chain formulation follows quickly. In order to conserve space and simplify our discussion we will only consider the computation of $\lambda_{XY} = H(X,Y)$ in this section. The results for $H(X)$ and $H(Y)$ are similar to those for $H(X,Y)$.

Consider the vector $W_n = \mu\Gamma_{(X,Y)}(0)\Gamma_{(X,Y)}(1)\dots\Gamma_{(X,Y)}(n-1)$ resulting from the product of the initial distribution $\mu$ with a sequence of random matrices. Using (17) we see that the vector $W_n$ is a random vector and $W_n(z) = p((X_0,Y_0),\dots,(X_{n-1},Y_{n-1}),Z_n = z)$. Suppose we normalize $W_n$ in each time slot so that we have a probability distribution, call this normalized vector $V_n$. Then

$$\begin{aligned} W_0 &= V_0 = \mu && (25)\\ W_n &= W_{n-1}\Gamma_{(X,Y)}(n-1) && (26)\\ V_n &= \frac{V_{n-1}\Gamma_{(X,Y)}(n-1)}{||V_{n-1}\Gamma_{(X,Y)}(n-1)||}, && (27) \end{aligned}$$

where $||\cdot||$ is the 1-norm. We will assume $||V_{n-1}\Gamma_{(X,Y)}(n-1)|| > \epsilon > 0$ for all possible values of $V$ and $\Gamma$. Since the $\Gamma$ can only be selected from a finite collection of matrices and the $V$ are probability distributions, this assumes the matrices $\mathbf{G}_{(\mathbf{X},\mathbf{Y})}$ have at least one positive element per column, which is equivalent to assuming any input/output pair can be observed in any channel state.

Notice that by normalizing $W_n$ at each time step we are simply dividing by the probability of observing the symbol sequence $(X_0,Y_0),\dots,(X_{n-1},Y_{n-1})$. Hence the sequence of vectors $V_n$ is

$$\begin{aligned} V_0 &= \mu = P(Z_0)\\ V_1 &= \frac{1}{\lambda_0}W_1 = \frac{p(X_0,Y_0,Z_1)}{\sum_{z\in\mathcal{Z}} p((X_0,Y_0),Z_1=z)} = p(Z_1|(X_0,Y_0))\\ &\vdots\\ V_n &= \frac{1}{\lambda_1\dots\lambda_{n-1}}W_n = \frac{p((X_0,Y_0),\dots,(X_{n-1},Y_{n-1}),Z_n)}{\sum_{z\in\mathcal{Z}} p((X_0,Y_0),\dots,(X_{n-1},Y_{n-1}),Z_n=z)} = p(Z_n|(X^n,Y^n)), \end{aligned}$$

namely the conditional probability distribution for the state of the channel given all past input/output symbol pairs.

In [3] the authors show that $V_n$ is a continuous state-space Markov chain with a unique stationary distribution that we can use to find entropy rates and capacity (provided the $X_n$ are i.i.d.). However, in our case $V_n$ is generally not Markov since the inputs $X_n$ are not i.i.d., hence we must use an augmented stochastic process in order to form a Markov chain.

## 4.2 An Augmented Markov Chain

Consider the process $U = (U_n : n \geq 0)$ where $U_n = (V_n, Z_{n+1}, Z_n)$. That is, rather than consider only the conditional distribution of the channel we will use the joint process consisting of the conditional distribution $V_n$ and the channel states $(Z_{n+1}, Z_n)$. (We must include both $Z_n$ and $Z_{n+1}$ due to our channel model (10).) At first this seems a bit awkward since we are using both a conditional distribution of the channel and the actual channel state to form our Markov

chain. However, keep in mind that the sequence $V_n$ is strictly determined by observation of the input/output process $(X, Y)$ and not $Z$.

We would like to show that $U$ is a Markov chain with at least one stationary distribution $\pi$. If this is the case, then we can compute $H(X, Y)$ as

$$\begin{aligned} H(X,Y) &= -\lim_{n\to\infty}\frac{1}{n}\log\left[\mu\Gamma_{(X,Y)}(0)\cdots\Gamma_{(X,Y)}(n-1)e\right] && (28)\\ &= -\lim_{n\to\infty}\frac{1}{n}\log W_n e = -\lim_{n\to\infty}\frac{1}{n}\log\nu_0\ldots\nu_{n-1}V_n e && (29)\\ &= -\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\log(\nu_k)\ \text{ (since } V_n e = 1) && (30)\\ &= -E_\pi\left[\log\nu\right] = -E_\pi\left[\log||V\Gamma_{(X,Y)}||\right], && (31)\\ &= -\lambda_{XY}\ \text{ a.s. by Theorem 1} && (32) \end{aligned}$$

where the $\nu_i$ are random scalars $\nu_i = ||V_i\Gamma_{(X,Y)}(i)||$. Notice the similarity between this expectation and those used in [3] (refer to section 2, equation 9). Indeed, the expectations are almost identical, except our random matrices are neither i.i.d nor diagonal. Likewise, our expectation is computed with respect to an augmented version of the Markov chain considered in [3].

The following propositions show that our Markov chain $U$ has a stationary distribution $\pi$ that is a continuous function of the input dynamics. This will allow us to compute the required Lyapunov exponents and state the capacity formula in terms of these quantities. Recall that we allowed our input probabilities to depend on the state of the channel. Therefore, our input dynamics are embedded in the random matrices $\Gamma$ rather than an external probability distribution (as is the case for i.i.d. inputs). Therefore, in order to prove this continuity property, we must show the Lyapunov exponents converge for a converging sequence of Markov chains, rather than a converging sequence of input distributions. Proofs of these propositions can be found in [4].

**Proposition 2:** Define a state-space for $U$ and its associated $\sigma$-algebra $(S, \mathcal{S})$ such that $S = \Delta^{|\mathcal{Z}|}$ x $\mathcal{Z}^2$. Then $U$ is a continuous state space discrete time Markov chain with state space $S$ and an associated transition kernel $K$ on $(S, \mathcal{S})$.

The kernel $K$ is defined in the usual fashion, if $\mu$ is a measure on $\mathcal{S}$, $\mu K(A)$ is the measure defined by $\mu K(A) = \int_S \mu(ds)K(s, A)$. If the function $g$ is bounded and $\mathcal{S}$-measurable, then $Kg$ is the function on $S$ defined by $Kg(s) = \int_S K(s, dt)g(t)$. For those not familiar with general state space Markov chains, the transition kernel $K$ is the continuous space equivalent of a transition matrix for a discrete space Markov chain. We can think of $\mu K$ as a one-step ahead probability measure. The function $Kg(x)$ is a one-step ahead conditional expectation of the function $g$ given current state $x \in S$.

**Proposition 3:** For any uniformly continuous function $g$ ($g \in UC$) and sequence of states $s_n^m \to s_n$ such that $s_m, s \in S$ we have $Kg(s_n^m) \to Kg(s_n)$.

**Proposition 4:** Suppose we enumerate the symbols pair $(x, y) \in (\mathcal{X}, \mathcal{Y})$ as $[1, \ldots, |\mathcal{X}, \mathcal{Y}|]$. As we defined in Section 3, we have a random matrix $\Gamma(n)$ such that for each $j \in (\mathcal{X}, \mathcal{Y})$ we have $\Gamma(n) = \mathbf{G_j} \in \mathcal{G}$ with probability $p(j|z_1, z_2)$. Now suppose we take a sequence of random matrices $\Gamma^m(n)$ such that $\Gamma^m(n) = G_j^m \in \mathcal{G}^m$ with probability $p^m(j|z_1, z_2)$, and each $\mathbf{G_j^m} \in \mathcal{G}^m \to \mathbf{G_j} \in \mathcal{G}$ and $p^m(j|z_1, z_2) \to p(j|z_1, z_2)$. Also define the corresponding sequence of transition kernels $K_m$. If we have $s_n^m \to s_n$, then $K_m g(s_n^m) \to Kg(s_n)$.

**Theorem 2:** Suppose we have a sequence of Markov chains with kernels $K, K_1, K_2, \ldots$ on $(S, \mathcal{S})$ as defined in Proposition 4, and let $\alpha, \alpha_1, \alpha_2, \ldots$ be initial distributions on $\mathcal{S}$ for each of the Markov chains. Then if

1. $\alpha_n \Rightarrow \alpha$

2. if $s_n^m \to s_n$ in $S$ and $g \in UC$, then $K_m g(s_n^m) \to Kg(s_n)$

3. every Markov kernel $K_m$ has an invariant measure $\pi^m$, and the family of probability measures $K_m(x, \cdot)$ is tight,

then the sequence $\pi^m$ is tight and every limit point of $\pi^m$ is an invariant measure for $K$. (See [5] for a proof of this theorem.)

This completes the steps needed to show that we can compute the entropy rate $H(X,Y) = \lambda_{XY} = -E_\pi[\log ||V\Gamma_{(X,Y)}||]$ as a function of the stationary distribution $\pi$ for the Markov chain $U$. Moreover, the stationary distribution is a continuous function of the dynamics for the input symbol process, which are determined by the matrices $\mathbf{G_{(X,Y)}}$. That is, if every matrix $\mathbf{G_j^m} \to \mathbf{G_j}$ and the probability mass functions $p^m(j|z_1, z_2) \to p(j|z_1, z_2)$, then every limit point of the sequence of stationary distributions $\pi_m$ is a stationary distribution for $K$. Hence, by Theorem 1, $E_{\pi^m}[\log ||V\Gamma_{(X,Y)}|| \to E_\pi[\log ||V\Gamma_{(X,Y)}||$.

Therefore, we can compute each Lyapunov exponent $\lambda_{XY}$, $\lambda_X$, and $\lambda_Y$, as an expectation with respect to the stationary distribution for a Markov chain. Furthermore, since the Lyapunov exponents are continuous functions of the input dynamics we can write the channel capacity as

$$C = \max_{p(X)} [\lambda_X + \lambda_Y - \lambda_{XY}] \tag{33}$$

# 5 Conclusions

We have presented a new representation for entropy rates, mutual information, and capacity based on Lyapunov exponents for products of random matrices. We showed the equivalence between the entropy rates for a Markov channel and Lyapunov exponents. Furthermore, we showed that the Lyapunov exponents are continuous functions of the dynamics for the input symbol process. These properties allow us to write mutual information in terms of Lyapunov exponents and to optimize the mutual information in order to find channel capacity.

Due to space constraints we are unable to discuss many other interesting possibilities that arise from this connection between entropy rates and Lyapunov exponents. In the journal version of this paper we plan to address these issues as well as provide more intuition for the results presented here.

# References


[1] G. Froyland, Rigorous Numerical Estimation of Lyapunov Exponents and Invariant Measures of Iterated Function Systems, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 2000.

[2] H. Furstenberg, H. Kesten, Products of Random Matrices, Ann. Math. Statist., 1960, 457-469.

[3] A. Goldsmith, P. Varaiya, Capacity, Mutual Information, and coding for Finite State Markov Channels, IEEE Trans. Information Theory, 1996.

[4] T. Holliday, A. Goldsmith, P. Glynn, Entropy and Mutual Information for Markov Channel with General Inputs, In preparation.

[5] A. Karr, Weak Convergence of a Sequence of Markov Chains, Z. Wahrscheinlichkeitstheorie verw. Geibite 33, 1975.

[6] M. Mushkin and I. Bar-David, Capacity and coding for the Gilbert–Elliot channel, IEEE Trans. Inform. Theory, Nov. 1989.

[7] V. Oseledec, A Multiplicative Ergodic Theorem, Trudy Moskov. Mat. Moskov. Mat. Obsc., 1968.