# State Learning and Mixing in Entropy of Hidden Markov Processes and the Gilbert–Elliott Channel

Bertrand M. Hochwald, *Member, IEEE*, and Predrag R. Jelenković, *Associate Member, IEEE*

***Abstract*—Hidden Markov processes such as the Gilbert–Elliott channel have an infinite dependency structure. Therefore, entropy and channel capacity calculations require knowledge of the infinite past. In practice, such calculations are often approximated with a finite past. It is commonly assumed that the approximations require an unbounded amount of the past as the memory in the underlying Markov chain increases. We show that this is not necessarily true. We derive an exponentially decreasing upper bound on the accuracy of the finite-past approximation that is much tighter than existing upper bounds when the Markov chain mixes well. We also derive an exponentially decreasing upper bound that applies when the Markov chain does not mix at all. Our methods are demonstrated on the Gilbert–Elliott channel, where we prove that a prescribed finite-past accuracy is quickly reached, independently of the Markovian memory. We conclude that the past can be used either to learn the channel state when the memory is high, or wait until the states mix when the memory is low. Implications for computing and achieving capacity on the Gilbert–Elliott channel are discussed.**

***Index Terms*—Birkhoff contraction coefficient, fading channel, function of a Markov chain, Markov-modulated random walk, Markovian memory.**

## I. Introduction

Hidden Markov processes, or equivalently, Markov-modulated random walks, are used to model many diverse systems ranging from image and speech recognizers [9] to communication channels with memory [6]. They have the advantage of being flexible and simple.

The Gilbert–Elliott channel shown in Fig. 1 is an example of a hidden Markov process that is used to model a digital channel whose errors appear in bursts due to, for example, a random fading process. In the "good" state, the channel causes errors with probability $p_g$, while in the "bad" state the channel makes errors with probability $p_b > p_g$. In this model, the underlying Markov chain state is hidden since we cannot necessarily tell the channel state by observing the error process.

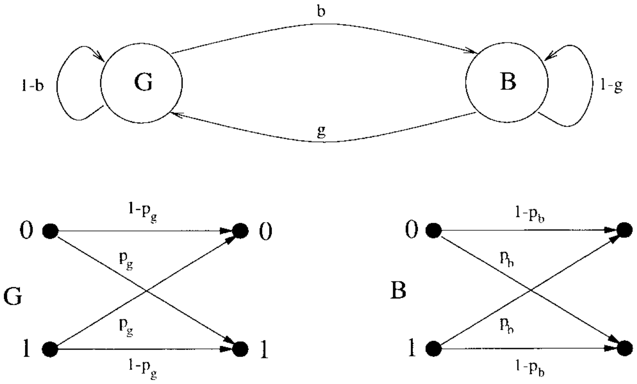

Fig. 1. Hidden Markov model of Gilbert–Elliott channel. The binary- symmetric channel is either in the "good" state or "bad" state, and switches states according to a Markov chain with the indicated transition probabilities.

Since its inception in [6], this model has been extended to models with more states and transitions between the states [4], [7], [11]. Of importance to achieving capacity on hidden Markov channels is the dependence of future errors on past ones, but even for the relatively simple two-state Gilbert–Elliott channel, this dependence is poorly understood. We show how, in entropy calculations, hidden Markov processes have dependence structures that are due to state "mixing" and "learning."

Loosely speaking, mixing is the act of switching states, which the underlying Markov chain does with frequency that depends on the off-diagonal elements of its transition matrix. Markov chains that mix well are said to have low memory. We prove that when the chain has low memory, the dependence structure is weak and entropy approximations are easily made. Conversely, when the memory is large, entropy approximations are more difficult. We show, however, that the effects of a large memory are limited because the underlying Markov chain tends to stay in any state for a long time, and this state can be learned from the observed process.

A Markov-modulated random walk is defined as a process $\{(S_n, Z_n),\ n = 0, 1, \cdots\}$ such that

$$\mathbb{P}[S_n = j, Z_n = k | S_{n-1}, Z_{n-1}, \cdots, S_0, Z_0] = \mathbb{P}[S_n = j, Z_n = k | S_{n-1}]$$

i.e., $\{(S_n, Z_n),\ n \geq 0\}$ is a Markov chain with past dependence on only the first coordinate. The transition probabilities for this chain are

$$\mathbb{P}[S_n = j, Z_n = k | S_{n-1} = i] = q_{ij} a_{jk}$$

where

$$\begin{aligned} q_{ij} &\overset{\text{def}}{=} \mathbb{P}[S_n = j | S_{n-1} = i] \\ a_{jk} &\overset{\text{def}}{=} \mathbb{P}[Z_n = k | S_n = j, S_{n-1} = i] \end{aligned} \tag{1.1}$$

Manuscript received November 20, 1997; revised July 2, 1998. The material in this paper was presented in part at the 35th Allerton Conference on Communication, Control, and Computing, Monticello, IL, October 1997.

B. M. Hochwald is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: hochwald@bell-labs.com).

P. R. Jelenković was with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA. He is now with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: predrag @ee.columbia.edu).

Communicated by S. Shamai, Associate Editor for Shannon Theory.
Publisher Item Identifier S 0018-9448(99)00050-4.

$i, j = 1, \cdots, K$ and $k = 1, \cdots, D$; the underlying state $S_n$ takes on $K$ possible values, and $Z_n$ takes on $D$ possible values. In this paper, we assume for simplicity that $a_{jk}$ does not depend on $i$ and that the number of states is finite. Abusing notation, we refer to $Z_n$ as a Markov-modulated random walk if the accompanying underlying Markov chain $S_n$ is self-understood.

For example, the error process $Z_n$ of the Gilbert–Elliott channel is a Markov-modulated random walk. For notational convenience, we denote the event of a channel error at time $n$ by $Z_n = 2$, and no error by $Z_n = 1$. Therefore, $Z_n$ is two with probability $p_g$ and one with probability $1 - p_g$ when $S_n$ is the good state, and is two with probability $p_b$ and one with probability $1 - p_b$ when $S_n$ is the bad state. Letting the good state be denoted by a one and the bad state by a two, we obtain $q_{11} = 1 - b$, $q_{12} = b$, $q_{21} = g$, $q_{22} = 1 - g$, $a_{11} = 1 - p_g$, $a_{12} = p_g$, $a_{21} = 1 - p_b$, and $a_{22} = p_b$. Thus the underlying Markov chain has transition matrix

$$Q = \begin{pmatrix} 1-b & b \\ g & 1-g \end{pmatrix} \tag{1.2}$$

and has stationary probabilities $\pi_1 = g/(g+b)$ and $\pi_2 = b/(g+b)$. The transition matrix $P$ for the $(S_n, Z_n)$ pair for the states $(1, 1), (2, 1), (1, 2)$, and $(2, 2)$ is shown in (1.3) at the bottom of this page.

Throughout this paper, we assume that the transition probabilities $b$ and $g$ are parameterized by $\varepsilon \in (0, 1]$ such that

$$g = \pi_1\varepsilon, \;\; b = \pi_2\varepsilon, \qquad \pi_1, \pi_2 > 0, \;\; \pi_1 + \pi_2 = 1 \tag{1.4}$$

where $\pi_1$ (good state) and $\pi_2$ (bad state) are the stationary probabilities for the underlying Markov chain matrix $Q$ given in (1.2). The Markovian memory is defined as $1 - b - g$, and because $\pi_1 + \pi_2 = 1$, the memory is zero when $\varepsilon = 1$, whence the channel errors $Z_0, Z_1, \cdots$, are independent and identically distributed. When $\varepsilon = 0$ the underlying chain is decomposable, but we assume that $\pi_1$ and $\pi_2$ are always defined. For simplicity, we do not consider negative memory $\varepsilon > 1$, but our arguments can easily be generalized to include this case.

Let $\{X_n, n \geq 0\}$ be a stationary Markov chain with transition probability $P = \{p_{ij}\}$ and finite state space $E = \{1, \cdots, A\}$. If $f$ is a function that maps $E \to \{1, \cdots D\}$, then the process $\{Z_n \overset{\text{def}}{=} f(X_n), n \geq 0\}$ is called a *function of the Markov chain* $X_n$.

Functions of a Markov chain and Markov-modulated random walks are equivalent. That a Markov-modulated random walk is a function of a Markov chain follows immediately from the definition, since $Z_n = g(S_n, Z_n)$, where $g(\cdot, \cdot)$ is the function that takes the second argument. On the other hand, suppose that $Z_n$ is a function of a Markov chain; that is, $Z_n = f(X_n)$ for some function $f(\cdot)$ of the Markov chain $X_n$. Then, the pair $(X_n, Z_n)$ form a Markov-modulated random walk with $a_{jk} = 1$ when $k = f(j)$, and $a_{jk} = 0$ otherwise.

Let $Z_n = f(X_n)$ be a function of a stationary Markov chain $X_n$. The entropy of the sequence $Z_0, Z_1, \cdots$ is defined as

$$H(Z) \overset{\text{def}}{=} \lim_{n\to\infty} (1/n) H(Z^{(n)})$$

(whenever this limit exists), where

$$Z^{(n)} \overset{\text{def}}{=} Z_1, \cdots, Z_n.$$

Among other roles, $H(Z)$ is fundamental towards determining the capacity of channels with hidden Markov structure [5]. This limit is generally difficult to compute and is often approximated by the following upper and lower bounds:

$$H(Z_n|Z^{(n-1)}, X_0) \leq H(Z) \leq H(Z_n|Z^{(n-1)}) \tag{1.5}$$

where

$$\lim_{n\to\infty} H(Z_n|Z^{(n-1)}, X_0) = H(Z) = \lim_{n\to\infty} H(Z_n|Z^{(n-1)}). \tag{1.6}$$

The proof of these results can be found in [1] or [3, pp. 69–71]. For numerical computation, one can use either $H(Z_n|Z^{(n-1)}, X_0)$ or $H(Z_n|Z^{(n-1)})$ for some finite $n$ to approximate $H(Z)$. It is of practical importance to know the error of either approximation, but rather than work with these quantities separately, we combine them and define the conditional mutual information

$$I(Z_n; X_0|Z^{(n-1)}) \overset{\text{def}}{=} H(Z_n|Z^{(n-1)}) - H(Z_n|Z^{(n-1)}, X_0) \geq 0.$$

Equation (1.6) says that $I(Z_n; X_0|Z^{(n-1)}) \to 0$ as $n \to \infty$; that is, $Z_n$ and $X_0$ become independent as $n \to \infty$, conditionally on the $(n-1)$th-order history. This paper is concerned primarily with the rate with which this happens.

In the case of a Markov-modulated random walk $Z_n = g(X_n)$, where $X_n = (S_n, Z_n)$ and $g(\cdot, \cdot)$ is the function that takes its second argument, we have

$$I(Z_n; X_0|Z^{(n-1)}) = I(Z_n; S_0|Z^{(n-1)})$$

for all $n$. It is commonly believed that $I(Z_n; S_0|Z^{(n-1)})$ converges to zero with a rate that goes to zero as $q_{ij}$ $(i \neq j)$ tends to zero (for example, as $\varepsilon \to 0$ in the Gilbert–Elliott channel and the underlying Markov chain becomes decomposable). Theorem 1, given in the next section and first derived in [1], strengthens this belief by finding an upper bound on $I(Z_n; X_0|Z^{(n-1)})$ for functions of a Markov chain; when applied to a Markov-modulated random walk, the theorem upper-bounds $I(Z_n; S_0|Z^{(n-1)})$ with $\zeta^n$ for some $\zeta$ that, in general, approaches one as $q_{ij}$ $(i \neq j)$ tends to zero. Define $-\log \zeta$ to be the *rate* of the bound and define

$$\lim_{n\to\infty} -(1/n) \log I(Z_n; S_0|Z^{(n-1)})$$

$$P = \begin{pmatrix} (1-b)(1-p_g) & b(1-p_b) & (1-b)p_g & bp_b \\ g(1-p_g) & (1-g)(1-p_b) & gp_g & (1-g)p_b \\ (1-b)(1-p_g) & b(1-p_b) & (1-b)p_g & bp_b \\ g(1-p_g) & (1-g)(1-p_b) & gp_g & (1-g)p_b \end{pmatrix}. \tag{1.3}$$

to be the true *convergence rate* (assuming that this exists). Then Theorem 1 gives a convergence rate lower bound that approaches zero as $q_{ij} \to 0$ $(i \neq j)$.

However, in this paper we argue that Theorem 1 gives a very loose bound and there are, instead, two exponential rates of convergence, in general. One is due to mixing and dominates when $q_{ij} \neq 0 (i \neq j)$, and the other is due to learning when $q_{ij} = 0 (i \neq j)$. *Neither rate is zero.* We demonstrate our arguments explicitly on the Gilbert–Elliott channel.

There are three main theorems derived in this paper. In Theorem 2 we derive an upper bound on the exponential rate of convergence that is much tighter than Theorem 1, especially when $q_{ij}$ $(i \neq j)$ are not too small or the underlying Markov chain mixes well. When applied to the Gilbert–Elliott model, Theorem 2 gives a rate that goes to infinity as $\varepsilon \to 1$, whereas Theorem 1 gives a finite rate. (The actual rate does go to infinity as $\varepsilon \to 1$ because $I(Z_n; S_0|Z^{(n-1)}) \to 0$ for all $n$.) The proof of Theorem 2 relies on a novel application to conditional entropies of a Hilbert pseudometric and Birkhoff contraction coefficient.

Theorem 2, although much tighter than Theorem 1, turns out to be loose as $\varepsilon \to 0$ in the Gilbert–Elliott channel. When applied to this channel, Theorem 2 yields a convergence rate lower bound that is $O(\varepsilon)$, where as Theorem 1 yields $O(\varepsilon^4)$. It turns out that both are pessimistic and we show, in Theorem 3, that there is a universal "mixing" rate that does not go to zero as $\varepsilon \to 0$. Theorem 4 shows that there is a "learning" rate that applies when $q_{ij} = 0$ $(i \neq j)$ $(\varepsilon = 0$ in the Gilbert–Elliot channel). We then show that this learning rate provides a uniform bound for all $\varepsilon \geq 0$.

## II. Bibliographical Note on the Convergence Rate

The best known bound on the rate of convergence of $I(Z_n; X_0|Z^{(n-1)})$ for functions of a Markov chain is due to Birch in 1962 [1].

*Theorem 1:* Let $X_n$, $n = 0, 1, \cdots$, be a Markov chain with transition probabilities $p_{ij} > 0$, $i, j = 1, \cdots, A$, and $Z_n = f(X_n) \in \{1, \cdots, D\}$. Then

$$I(Z_n; X_0|Z^{(n-1)}) \leq B\zeta^{n-1} \tag{2.1}$$

where

$$\begin{aligned} \zeta &= 1 - \frac{N_1}{(N_2)^2} \min_{1 \leq i, j, k, m, n \leq A} \left(\frac{p_{ik}p_{kn}}{p_{ij}p_{jm}}\right)^2 \\ B &= (N_2 \log e) \bigg/ \left(N_1 \min_{1 \leq i, j \leq A} p_{ij}\right) \end{aligned}$$

and $N_1$ and $N_2$ are the minimum and maximum cardinalities of the sets $f^{-1}(1), \cdots, f^{-1}(D)$ that are nonempty.

While exponential rates of convergence are often desirable, the rate $-\log \zeta$ in Theorem 1 is, in general, very small and suggests that a very large history is needed. We demonstrate this with the Gilbert–Elliott channel.

In Section I this channel is described as the function that takes the second coordinate of the Markov chain with transition matrix given by (1.3). Since the second argument can be either one or two, $N_1 = N_2 = 2$. Let the transition probabilities $b$ and $g$ of the Markov chain $S_n$ be parameterized as in (1.4) [see (2.2) at the bottom of this page]. As $\varepsilon \to 0$ (implying that $b, g \to 0$), we therefore have $\zeta = 1 - O(\varepsilon^4)$ and

$$I(Z_n; S_0|Z^{(n-1)}) = I(Z_n; X_0|Z^{(n-1)}) \leq B[1 - O(\varepsilon^4)]^n$$

where $B \to \infty$ as $\varepsilon \to 0$. Hence, the rate is

$$-\log \zeta = -\log (1 - O(\varepsilon^4)) = O(\varepsilon^4)$$

which rapidly goes to zero as the Markovian memory in $S_n$ rises. With its dependency on $\varepsilon^4$, this lower bound on the convergence rate turns out to be extremely loose. In Section IV, we show that the true convergence rate for small $\varepsilon$ is, in fact, *uniformly bounded* from below.

At the other extreme, as $\varepsilon \to 1$ the Markov chain states $S_0, S_1, \cdots$, and the channel errors $Z_0, Z_1, \cdots$, become, within themselves, independent and identically distributed processes. Therefore, $I(Z_n; S_0|Z^{(n-1)}) \to 0$ as $\varepsilon \to 1$ for every $n$, and the true convergence rate should go to infinity. However, $\zeta$ in Theorem 1 does not go to zero ($-\log \zeta$ does not go to infinity) as $\varepsilon \to 1$, and thus the bound equation (2.1) again becomes arbitrarily loose.

In the next section we derive a new much tighter bound that is as widely applicable as Theorem 1. We show that the rate of our new bound approaches infinity as $\varepsilon \to 1$ in the Gilbert–Elliott channel. The bound is derived with the help of a contraction coefficient first considered by Birkhoff. Our novel application of this contraction coefficient to conditional entropy is, we believe, of independent interest.

## III. Hilbert's Projective Metric and Birkhoff's Contraction Coefficient

In order to state and prove our results we need to introduce the notion of Hilbert's pseudometric [10, p. 80] $d$ which is defined for any two vectors $x = (x_1, \cdots, x_n)$, $y = (y_1, \cdots, y_n)$ with positive elements as

$$d(x, y) = \max_{i, j} \log \left(\frac{x_i y_j}{x_j y_i}\right). \tag{3.1}$$

This pseudometric has the property that $d(x, y) = 0$ if and only if $x = \lambda y$ for some scalar $\lambda$. The metric turns out to be ideally suited for analytical manipulations involving conditional entropy.

$$P = \begin{pmatrix} (1-\pi_2\varepsilon)(1-p_g) & \pi_2\varepsilon(1-p_b) & (1-\pi_2\varepsilon)p_g & \pi_2\varepsilon p_b \\ \pi_1\varepsilon(1-p_g) & (1-\pi_1\varepsilon)(1-p_b) & \pi_1\varepsilon p_g & (1-\pi_1\varepsilon)p_b \\ (1-\pi_2\varepsilon)(1-p_g) & \pi_2\varepsilon(1-p_b) & (1-\pi_2\varepsilon)p_g & \pi_2\varepsilon p_b \\ \pi_1\varepsilon(1-p_g) & (1-\pi_1\varepsilon)(1-p_b) & \pi_1\varepsilon p_g & (1-\pi_1\varepsilon)p_b \end{pmatrix}. \tag{2.2}$$

For any matrix $T$ with nonnegative elements, Birkhoff's contraction coefficient is defined as

$$\tau(T) = \sup_{\substack{x,\, y > 0 \\ x \neq \lambda y}} \frac{d(xT,\, yT)}{d(x,\, y)}.$$

In [10] it is shown that $0 \le \tau(T) \le 1$ and, for any two nonnegative matrices $T_1, T_2$

$$d(xT_1T_2,\, yT_1T_2) \le \tau(T_2) d(xT_1,\, yT_1) \le \tau(T_2)\tau(T_1) d(x,\, y).$$

An explicit formula for $\tau(T)$ is given as

$$\tau(T) = \frac{1 - \sqrt{\phi(T)}}{1 + \sqrt{\phi(T)}} \tag{3.2}$$

where

$$\phi(T) = \min_{i,\, j,\, k,\, l} \frac{t_{ik} t_{jl}}{t_{jk} t_{il}}.$$

This is surprisingly tedious to prove [10]. It is clear that $\tau(T) = 0$ if and only if $T$ has rank one.

### A. Convergence Rate

To reduce the notational complexity, we consider the special case where the function $f$ of the Markov chain partitions the state space $E$ into sets all with the same cardinality $K$; i.e., we assume that the sets

$$E_i = f^{-1}(i) \overset{\text{def}}{=} \{j: f(j) = i\}, \qquad i = 1, \cdots, D$$

all have $K$ elements, where $KD = A$ (recall that $X$ takes on $A$ distinct values). Without loss of generality, suppose that the state space is labeled such that $f(j) = i$, for all $j = (i-1)K+1, \cdots, iK$, i.e., $E_i = \{(i-1)K+1, \cdots, iK\}$. Then we can partition the transition matrix $P$ into blocks $P_{ij}$

$$P_{ij} = \{p_{kl}: k \in E_i,\, l \in E_j\}, \qquad i, j = 1, \cdots, D. \tag{3.3}$$

Let $\nu$ be the stationary probability vector for the transition matrix $P$, and let

$$\nu^i = (\nu_{(i-1)K+1}, \cdots, \nu_{iK}), \qquad i = 1, \cdots, D \tag{3.4}$$

be the $D$ subvectors corresponding to the subsets $E_i$. Similarly, let us break the $j$th row of the matrix $P$ into $D$ blocks

$$p_j^i = (p_{j,\,(i-1)K+1}, \cdots, p_{j,\,iK}), \qquad i = 1, \cdots, D \tag{3.5}$$

each of length $K$.

*Theorem 2:* Let $X_n$, $n = 0, 1, \cdots$, be a Markov chain with transition probabilities $p_{ij} > 0$, $i, j = 1, \cdots, A$, and $Z_n = f(X_n) \in \{1, \cdots, D\}$, where $f^{-1}(1), \cdots, f^{-1}(D)$ all have the same cardinality. Then

$$I(Z_n; X_0 | Z^{(n-1)}) \le C\rho^{n-2}, \qquad n \ge 2 \tag{3.6}$$

where

$$\rho = \max_{1 \le i,\, j \le D} \tau(P_{ij}), \qquad C = \max_{1 \le j \le A,\, 1 \le i \le D} d(p_j^i,\, \nu^i)$$

and $P_{ij}$, $\nu^i$, and $p_j^i$ are defined in (3.3)–(3.5).

*Proof:* We have that

$$I(Z_n; X_0 | Z^{(n-1)}) = \mathbb{E} \log \frac{\mathbb{P}[Z_n | Z_{n-1}, \cdots, Z_1, X_0]}{\mathbb{P}[Z_n | Z_{n-1}, \cdots, Z_1]}. \tag{3.7}$$

We wish to uniformly upper-bound the argument of the logarithm. Observe that

$$\begin{aligned}\mathbb{P}[Z_n = i_n,\, Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1,\, X_0 = i_0]& \\ = \nu_{i_0} p_{i_0}^{i_1} P_{i_1 i_2} \cdots P_{i_n-1} i_n e&\end{aligned}$$

where $e$ is a column vector of ones. Similarly,

$$\begin{aligned}\mathbb{P}[Z_n = i_n,\, Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1]& \\ = \nu^{i_1} P_{i_1 i_2} \cdots P_{i_{n-1} i_n} e.&\end{aligned}$$

To simplify the notation, let

$$\Pi(i_1, \cdots, i_k) \overset{\text{def}}{=} P_{i_1 i_2} \cdots P_{i_{k-1} i_k}.$$

Observe that $\Pi(i_1, \cdots, i_k)$ is the product of $k-1$ matrices. Then

$$\begin{aligned}&\frac{\mathbb{P}[Z_n = i_n | Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1,\, X_0 = i_0]}{\mathbb{P}[Z_n = i_n | Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1]} \\ &\quad = \frac{p_{i_0}^{i_1} \Pi(i_1, \cdots, i_n) e}{\nu^{i_1} \Pi(i_1, \cdots, i_n) e} \frac{\nu^{i_1} \Pi(i_1, \cdots, i_{n-1}) e}{p_{i_0}^{i_1} \Pi(i_1, \cdots, i_{n-1}) e} \\ &\quad \le \max_{1 \le k \le K} \left( \frac{[p_{i_0}^{i_1} \Pi(i_1, \cdots, i_{n-1})]_k \nu^{i_1} \Pi(i_1, \cdots, i_{n-1}) e}{[\nu^{i_1} \Pi(i_1, \cdots, i_{n-1})]_k p_{i_0}^{i_1} \Pi(i_1, \cdots, i_{n-1}) e} \right)\end{aligned} \tag{3.8}$$

where $[\cdot]_k$ denotes the $k$th element of the vector argument, and the inequality follows from

$$\frac{\sum\limits_j a_j x_j}{\sum\limits_j a_j y_j} \le \max_j \frac{x_j}{y_j}, \qquad a_j,\, x_j,\, y_j > 0 \tag{3.9}$$

with $a_j = [P_{i_{n-1} i_n} e]_j$. By applying (3.9) once more to (3.8) with $a_j = 1$, we obtain

$$\begin{aligned}&\frac{\mathbb{P}[Z_n = i_n | Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1,\, X_0 = i_0]}{\mathbb{P}[Z_n = i_n | Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1]} \\ &\quad \le \max_{1 \le k,\, l \le K} \left( \frac{[p_{i_0}^{i_1} \Pi(i_1, \cdots, i_{n-1})]_k [\nu^{i_1} \Pi(i_1, \cdots, i_{n-1})]_l}{[\nu^{i_1} \Pi(i_1, \cdots, i_{n-1})]_k [p_{i_0}^{i_1} \Pi(i_1, \cdots, i_{n-1})]_l} \right)\end{aligned} \tag{3.10}$$

and taking the logarithm of (3.10) yields

$$\begin{aligned}&\log \left( \frac{\mathbb{P}[Z_n = i_n | Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1,\, S_0 = i_0]}{\mathbb{P}[Z_n = i_n | Z_{n-1} = i_{n-1}, \cdots, Z_1 = i_1]} \right) \\ &\quad \le d(p_{i_0}^{i_1} \Pi(i_1, \cdots, i_{n-1}),\, \nu^{i_1} \Pi(i_1, \cdots, i_{n-1})) \\ &\quad \le \tau(\Pi(i_1, \cdots, i_{n-1})) d(p_{i_0}^{i_1},\, \nu^{i_1}) \\ &\quad \le \rho^{n-2} C\end{aligned} \tag{3.11}$$

which follows because

$$\Pi(i_1, \cdots, i_{n-1}) = P_{i_1 i_2} \cdots P_{i_{n-2} i_{n-1}}$$

is the product of $n-2$ matrices. Finally, using inequality equations (3.11) in (3.7) concludes the proof. □

Suppose $\{(S_n, Z_n), n = 0, 1, \cdots\}$ is a Markov-modulated random walk with transition matrix $P$ that can be partitioned into blocks

$$P_{1k} = \cdots = P_{Dk} = \{q_{ij}a_{jk}: i, j = 1, \cdots, K\}, \qquad k = 1, \cdots, D$$

with positive $q_{ij}$ and $a_{jk}$ as defined in (1.1). Then from (3.2) it follows that

$$\tau(P_{lk}) = \tau(Q). \tag{3.12}$$

The stationary probabilities for $(S_n, Z_n)$ are $\nu_{(i-1)K+j} = \mathbb{P}[S_n = j, Z_n = i] = \pi_j a_{ji}$, where $\pi = (\pi_1, \cdots, \pi_K)$ is the stationary probability for the underlying Markov chain $S_n$. In particular

$$\nu^i = (\pi_1 a_{1i}, \cdots, \pi_K a_{Ki}) \tag{3.13}$$

where $\nu^i$ is defined in (3.4). Next, observe that only the first $K$ rows of the matrix $P$ are distinct, and that

$$p_j^i = (q_{j1}a_{1i}, \cdots, q_{jK}a_{Ki}) \tag{3.14}$$

where $p_j^i$ is defined in (3.5). Equations (3.13) and (3.14) and the definition of Hilbert's pseudometric equation (3.1) therefore imply that

$$d(p_j^i, \nu^i) = d(q^j, \pi) \tag{3.15}$$

where $q^j = (q_{j1}, \cdots, q_{jK})$ is the $j$th row of the matrix $Q$. By combining (3.12) and (3.15) with Theorem 2, we arrive at the following result.

*Corollary 1:* Let $Z_n$, $n = 0, 1, \cdots$, be a Markov-modulated random walk whose underlying chain $S_n$ has a transition matrix $Q$ with stationary probability $\pi$. Then

$$I(Z_n; S_0 | Z^{(n-1)}) \le C\rho^{n-2}, \qquad n \ge 2 \tag{3.16}$$

where

$$\rho = \tau(Q), \qquad C = \max_{1 \le j \le K} d(q^j, \pi)$$

where $q^j = (q_{j1}, \cdots, q_{jK})$ denotes the $j$th row of the matrix $Q$.

The extension of Theorem 2 to functions that do not partition the state space $E$ into sets with the same cardinality is also possible but is omitted; see [10, p. 147] for comments on the application of the contraction coefficient to nonsquare matrices.

### B. Application to Gilbert–Elliott Channel

In the Gilbert–Elliott channel, the underlying Markov chain has transition matrix $Q$ given in (1.2). Applying Corollary 1 and (3.2) to this matrix yields

$$\rho = \frac{1-\sqrt{\phi}}{1+\sqrt{\phi}}$$
$$\phi = \min\left(\frac{bg}{(1-b)(1-g)}, \frac{(1-b)(1-g)}{bg}\right) \tag{3.17}$$

and

$$C = \log \max\left(\frac{b^2}{g(1-b)}, \frac{g(1-b)}{b^2}, \frac{g^2}{b(1-g)}, \frac{b(1-g)}{g^2}\right).$$

Observe that $\rho$ and $C$ do not depend on $p_g$ or $p_b$.

Let us now compare the bounds (2.1) and (3.16) when $b$ and $g$ are parameterized as in (1.4). As $\varepsilon \to 1$, we already know from Section II that the true convergence rate goes to infinity whereas the rate given by Theorem 1 remains bounded. From (3.17), we obtain

$$\phi = \frac{1}{1 + \frac{1-\varepsilon}{\pi_2\pi_1\varepsilon^2}}, \qquad \rho = \frac{\sqrt{1 + \frac{1-\varepsilon}{\pi_2\pi_1\varepsilon^2}} - 1}{\sqrt{1 + \frac{1-\varepsilon}{\pi_2\pi_1\varepsilon^2}} + 1}.$$

As $\varepsilon \to 1$, we see that $\rho = O(1-\varepsilon)$ and hence $-\log\rho \to \infty$. Thus Theorem 2 gives a much tighter bound (that also happens to be independent of $p_g$ and $p_b$) than Theorem 1.

As $\varepsilon \to 0$, we have $\phi = O(\varepsilon^2)$ and $\rho = 1 - O(\varepsilon)$. Therefore, the rate $-\log\rho = O(\varepsilon)$ depends linearly on the Markovian memory and the bound given in Theorem 2 is again much tighter than the bound given in Theorem 1 (which behaves as $O(\varepsilon^4)$). The linear dependence on $\varepsilon$ also seems intuitively reasonable since the Markovian memory is $1-b-g = 1-O(\varepsilon)$. However, as we now show, there exists a uniform bound on the convergence rate that is independent of $\varepsilon$ and, therefore, the bound in Theorem 2 also inevitably becomes loose as $\varepsilon \to 0$.

### C. Explicit Analysis of Gilbert–Elliott Channel for $p_g = 0$

We consider the case $p_g = 0$ since it is readily analyzed. Define

$$x_{ij}^{(n)} = \mathbb{P}[S_0 = i, Z_1 = 1, \cdots, Z_{n-1} = 1, Z_n = j], \qquad i, j \in \{1, 2\}. \tag{3.18}$$

It is readily seen that $x_{ij}^{(n)}$ has the form

$$x_{ij}^{(n)} = c_{ij}\lambda_1^{n-1} + d_{ij}\lambda_2^{n-1} \tag{3.19}$$

for some $c_{ij}$ and $d_{ij}$, where $\lambda_1$ and $\lambda_2$ are the eigenvalues of the upper left-hand $2 \times 2$ submatrix of $P$ in (1.3)

$$P_{11} = \begin{bmatrix} 1-b & bq_b \\ g & (1-g)q_b \end{bmatrix}, \qquad q_b \overset{\text{def}}{=} 1 - p_b.$$

Hence (see (3.20) at the bottom of this page) $0 < \lambda_1 < \lambda_2 < 1$. By analyzing $x_{ij}^{(n)}$ in Appendix A, we prove the following theorem.

$$\lambda_{1,2} = \frac{1 - b + q_b(1-g) \mp \sqrt{(1-b+q_b(1-g))^2 - 4(1-b-g)q_b}}{2} \tag{3.20}$$

*Theorem 3:* Let $p_g = 0$ and $\varepsilon > 0$, and define $\lambda_1$ and $\lambda_2$ as in (3.20). Then the conditional mutual information for the Gilbert–Elliott channel obeys

$$I(Z_n;\, S_0|Z^{(n-1)}) \sim C'\left(\frac{(\lambda_1)^2}{\lambda_2}\right)^{n-1}$$

as $n \to \infty$. The constant $C'$ is given explicitly in (A.2).

*Proof:* See Appendix A. □

Let $b$ and $g$ be parameterized as in (1.4). It is straightforward to show that $(\lambda_1)^2/\lambda_2 \nearrow (1-p_b)^2$ as $\varepsilon \searrow 0$. Hence, when $\varepsilon > 0$, or equivalently, when the underlying Markov chain mixes, the convergence rate of $I(Z_n;\, S_0|Z^{(n-1)})$ is uniformly bounded by $-2\log(1-p_b)$, *independently of the Markovian memory.*

Thus although always much tighter than Theorem 1, the bound in Theorem 2 becomes loose as $\varepsilon \to 0$. We therefore seek a bound that is tight when $\varepsilon$ is small by deriving a bound that applies when $\varepsilon = 0$.

## IV. Learning Bound

While Theorem 2 provides a tight bound when $q_{ij} (i \neq j)$ are not too small ($\varepsilon$ not too small in Gilbert–Elliott channel), we now show that the rate of convergence *is nonzero* even when $q_{ij} = 0 (i \neq j)$. Roughly speaking, Theorem 2 captures the rate due to state mixing, but when $q_{ij} = 0$ $(i \neq j)$, the convergence is instead due to state learning.

We assume that $Z_n$ is a Markov-modulated random walk such that $q_{ij} = 0$ for all $i \neq j$, and $(a_{j1}, \cdots, a_{jD}) \neq (a_{j'1}, \cdots, a_{j'D})$ for all $j \neq j'$, where $q_{ij}$ and $a_{jk}$ are defined in (1.1). That is, the underlying Markov chain does not mix, and the state $S_n = S_0$ uniquely determines the distribution of $Z_n$. For example, these requirements become $\varepsilon = 0$ (or $b = g = 0$) and $p_g \neq p_b$ when applied to the Gilbert–Elliott channel. We obtain an upper bound on $I(Z_n;\, S_0|Z^{(n-1)})$ by using the relation

$$\begin{aligned} I(Z_n;\, S_0|Z^{(n-1)}) &= H(S_0|Z^{(n-1)}) - H(S_0|Z^{(n)}) \\ &\leq H(S_0|Z^{(n-1)}) \end{aligned} \tag{4.1}$$

and finding an upper bound on $H(S_0|Z^{(n-1)})$.

To find the bound, we introduce the notion of a state estimate. Because $q_{ij} = 0$ $(i \neq j)$, the state of the underlying Markov chain is fixed for all time, and standard parameter estimation theory suggests that there exists an estimate of $S_0$ from the sequence of observations $Z^{(n-1)} = Z_1, \cdots, Z_{n-1}$ that converges exponentially quickly with some rate $r$ in probability to the true state. Let $\hat{S}_0 = \hat{S}_0(Z^{(n-1)})$ be such an estimate (we provide an example in Section IV-A for the Gilbert–Elliott channel) and let $P_e^{n-1} = \mathbb{P}(\hat{S}_0 \neq S_0)$ be its probability of error as a function of $n$. By assumption

$$\limsup_{n\to\infty} (\log P_e^{n-1})/n \leq -r.$$

By Fano's inequality [3, p. 39]

$$H(S_0|Z^{(n-1)}) \leq h(P_e^{n-1}) + P_e^{n-1} \cdot \log(K-1)$$

where

$$h(p) \overset{\text{def}}{=} -p\log p - (1-p)\log(1-p) \tag{4.2}$$

is the binary entropy function, and $K$ is the number of states in the underlying Markov chain. It follows that

$$\limsup_{n\to\infty} \frac{\log H(S_0|Z^{(n-1)})}{n} \leq -r. \tag{4.3}$$

We, therefore, have the following theorem.

*Theorem 4:* Let $Z_n$ be a Markov-modulated random walk such that $q_{ij} = 0$ for all $i \neq j$, and $(a_{j1}, \cdots, a_{jD}) \neq (a_{j'1}, \cdots, a_{j'D})$ for all $j \neq j'$. Let $\hat{S}_0 = \hat{S}_0(Z^{(n-1)})$ be an estimate of $S_0$ that converges in probability exponentially quickly to $S_0$ with some rate $r > 0$. Then

$$\limsup_{n\to\infty} \frac{\log I(Z_n;\, S_0|Z^{(n-1)})}{n} \leq -r.$$

*Remark:* The estimate $\hat{S}_0$ that gives the largest $r$ (fastest learning) gives the tightest bound.

### A. Universal Learning Bound for Gilbert–Elliott Channel

The previous section argues that $I(Z_n;\, S_0|Z^{(n-1)})$ decays exponentially even when the underlying Markov chain does not mix at all. The decay is due to the effects of state learning that are overlooked by Theorems 1 and 2. We now apply the concept of learning to the Gilbert–Elliott channel. We show that the learning bound for this channel is asymptotically tight and bounds the rate of convergence of $I(Z_n;\, S_0|Z^{(n-1)})$ for every $\varepsilon \geq 0$. Thus we can bound the speed with which $I(Z_n;\, S_0|Z^{(n-1)})$ converges, uniformly in $\varepsilon$, implying that a fixed amount of history is needed to approximate $I(Z_n;\, S_0|Z^{(n-1)})$, no matter how large the Markovian memory (or how small $\varepsilon$) is.

Define

$$D(p_1,\, p_2) \overset{\text{def}}{=} p_1 \log \frac{p_1}{p_2} + (1-p_1)\log\frac{1-p_1}{1-p_2}$$

and let $\varepsilon = 0$, and $p_g,\, p_b > 0$. We now prove that

$$\lim_{n\to\infty} \frac{-\log I(Z_n;\, S_0|Z^{(n-1)})}{n} = D(p^*,\, p_g) = D(p^*,\, p_b) \tag{4.4}$$

where

$$p^* \overset{\text{def}}{=} \frac{\log \frac{1-p_g}{1-p_b}}{\log \frac{p_b}{p_g} + \log \frac{1-p_g}{1-p_b}}, \qquad p_g < p^* < p_b. \tag{4.5}$$

To show this, we first apply Theorem 4 to obtain an upper bound on $I(Z_n;\, S_0|Z^{(n-1)})$, and then derive an analytic lower bound that is asymptotically tight. We apply Theorem 4 to an estimate of $S_0$ from the observations $Z^{(n-1)}$ that converges exponentially with rate $D(p^*,\, p_g)$. The estimate counts the number of twos (error indicators) in the sequence $Z_1, \cdots, Z_{n-1}$, divides by $n$, and compares this value to a threshold $p$. Clearly, any $p$ such that $p_g < p < p_b$ will yield an estimate that converges in probability to the correct state as $n \to \infty$. The convergence rate of the estimate is optimized when $p = p^*$ defined in (4.5).

To see this, let $n_z$ be the number of twos in $Z_1, \cdots, Z_{n-1}$, and define the estimate $\hat{S}_0$ as

$$\hat{S}_0 = \begin{cases} 1 \text{ (good state)}, & \text{if } n_z/n \le p \\ 2 \text{ (bad state)}, & \text{if } n_z/n > p. \end{cases}$$

Then the probability that the estimate fails is

$$\begin{aligned} \mathbb{P}[\hat{S}_0 \ne S_0] &= \mathbb{P}[\hat{S}_0 = 2|S_0 = 1]\mathbb{P}[S_0 = 1] \\ &\quad + \mathbb{P}[\hat{S}_0 = 1|S_0 = 2]\mathbb{P}[S_0 = 2] \\ &= \pi_1 \mathbb{P}[n_z > np|S_0 = 1] \\ &\quad + \pi_2 \mathbb{P}[n_z \le np|S_0 = 2]. \end{aligned}$$

If $S_0 = 1$, the random variable $n_z$ has a binomial distribution with parameters $n$ and $p_g$; and if $S_0 = 2$, $n_z$ is binomial with parameters $n$ and $p_b$. An application of Cramér's theorem in large deviation theory (see, for example, [2] or [12]) yields

$$\begin{aligned} \lim_{n\to\infty} \frac{-\log \mathbb{P}[n_z > np|S_0 = 1]}{n} &= D(p, p_g) \\ \lim_{n\to\infty} \frac{-\log \mathbb{P}[n_z \le np|S_0 = 2]}{n} &= D(p, p_b). \end{aligned}$$

Therefore,

$$\lim_{n\to\infty} \frac{-\log \mathbb{P}[\hat{S}_0 \ne S_0]}{n} = \min(D(p, p_g), D(p, p_b)).$$

As functions of $p$, $D(p, p_g)$ is monotonically increasing for $p \ge p_g$, and $D(p, p_b)$ is monotonically decreasing for $p \le p_b$. The convergence rate of the estimate to the correct answer is therefore maximized by choosing $p$ so that $D(p, p_g) = D(p, p_b)$. Thus $p = p^*$ and the learning rate is $D(p^*, p_g) = D(p^*, p_b)$. By Theorem 4

$$\limsup_{n\to\infty} \frac{\log I(Z_n;\, S_0|Z^{(n-1)})}{n} \le -D(p^*, p_g). \tag{4.6}$$

It remains to prove the lower bound

$$\liminf_{n\to\infty} \frac{\log I(Z_n;\, S_0|Z^{(n-1)})}{n} \ge -D(p^*, p_g). \tag{4.7}$$

From (4.3) and (4.1), we may prove (4.7) by showing that

$$\liminf_{n\to\infty} \frac{\log H(S_0|Z^{(n)})}{n} \ge -D(p^*, p_g). \tag{4.8}$$

Simple algebra yields

$$H(S_0|Z^{(n)}) = F_n(\pi_1, \pi_2, p_g, p_b) + F_n(\pi_2, \pi_1, p_b, p_g) \tag{4.9}$$

where

$$\begin{aligned} F_n(x, y, u, v) \overset{\text{def}}{=} x \sum_{k=0}^{n} \binom{n}{k} u^k (1-u)^{n-k} \\ \cdot \log\left[1 + \frac{y}{x}\left(\frac{v}{u}\right)^k \left(\frac{1-v}{1-u}\right)^{n-k}\right]. \end{aligned} \tag{4.10}$$

Thus

$$H(S_0|Z^{(n)}) \ge F_n(\pi_1, \pi_2, p_g, p_b). \tag{4.11}$$

Choosing the summand $k = \lfloor np^* \rfloor$ in $F_n(\pi_1, \pi_2, p_g, p_b)$ gives

$$\begin{aligned} &F_n(\pi_1, \pi_2, p_g, p_b) \\ &\quad \ge \pi_1 \binom{n}{\lfloor np^* \rfloor} p_g^{\lfloor np^* \rfloor} (1-p_g)^{n-\lfloor np^* \rfloor} \\ &\quad \cdot \log\left[1 + \frac{\pi_2}{\pi_1}\left(\frac{p_b}{p_g}\right)^{\lfloor np^* \rfloor} \left(\frac{1-p_b}{1-p_g}\right)^{n-\lfloor np^* \rfloor}\right] \end{aligned}$$

where $\lfloor x \rfloor$ represents the integer part of $x$. By Stirling's approximation,

$$\lim_{n\to\infty} \frac{\log \binom{n}{\lfloor np^* \rfloor}}{n} = h(p^*)$$

where $h(\cdot)$ is the binary entropy function defined in (4.2). The choice of $p^*$ ensures that

$$\lim_{n\to\infty} \left(\frac{p_b}{p_g}\right)^{\lfloor np^* \rfloor} \left(\frac{1-p_b}{1-p_g}\right)^{n-\lfloor np^* \rfloor} = 1.$$

Hence

$$\begin{aligned} &\liminf_{n\to\infty} \frac{\log F(\pi_1, \pi_2, p_g, p_b)}{n} \\ &\quad \ge h(p^*) + p^* \log p_g + (1-p^*) \log(1-p_g) \\ &\quad = -D(p^*, p_g). \end{aligned} \tag{4.12}$$

Equation (4.8) now follows from (4.11) and (4.12), and the proof of (4.4) is concluded.

*Remark:* When $p_g = 0$ (error-free good state) the expression for $I(Z_n;\, S_0|Z^{(n-1)})$ is especially simple and we may obtain its exact asymptotic form. In this case, we get

$$\begin{aligned} &I(Z_n;\, S_0|Z^{(n-1)}) \\ &\quad = \pi_1(1-p_b)^n \log \frac{(1-p_b)(\pi_1(1-p_b)^{n-1} + \pi_2)}{\pi_1(1-p_b)^n + \pi_2} \\ &\qquad + \pi_1(1-p_b)^{n-1} p_b \log \frac{\pi_1(1-p_b)^{n-1} + \pi_2}{\pi_1(1-p_b)^{n-1}} \\ &\qquad + \pi_2 \log \frac{\pi_1(1-p_b)^{n-1} + \pi_2}{\pi_1(1-p_b)^n + \pi_2} \\ &\quad \sim \pi_1 p_b \log\left(\frac{1}{1-p_b}\right) n(1-p_b)^{n-1}. \end{aligned} \tag{4.13}$$

Hence the convergence rate of $I(Z_n;\, S_0|Z^{(n-1)})$ for $\varepsilon = p_g = 0$ is $-\log(1-p_b)$. This concurs with (4.4) since $D(p^*, p_b) \to -\log(1-p_b)$ as $p_g \to 0$.

We show in Section III-B that the convergence rate of $I(Z_n;\, S_0|Z^{(n-1)})$ due to mixing when $p_g = 0$ is given by a quantity that approaches $-2\log(1-p_b)$ as $\varepsilon \to 0$. From this, we might be tempted to infer that the convergence rate is exactly $-2\log(1-p_b)$ when $\varepsilon = 0$. However, because there is no state mixing, the analysis in Section III-B is invalid when $\varepsilon = 0$, and, in fact, (4.13) demonstrates that the convergence rate is $-\log(1-p_b)$ rather than $-2\log(1-p_b)$. Clearly, $-\log(1-p_b) < -2\log(1-p_b)$, and this slower rate is due to learning rather than mixing. Therefore, $-\log(1-p_b)$ is a lower bound on the convergence rate, universally valid for all $\varepsilon \ge 0$.

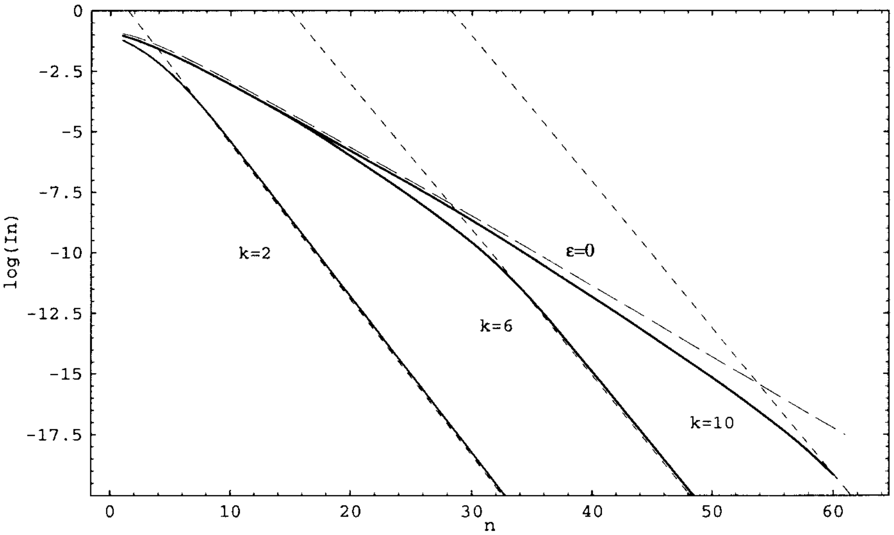

Fig. 2. Graph of $\log_{10} I(Z_n;\, S_0|Z^{(n-1)})$ versus $n$ for $p_g = 0$. Solid lines represent $I(Z_n;\, S_0|Z^{(n-1)})$ for $p_b = 0.5$, $g = 5 * 10^{-k}$, $b = 10^{-k}$ for $k = 2,\, 6$, and $10$. Long-dashed line is $\varepsilon = 0$ asymptote (learning rate $-\log(1 - p_b)$ given by (4.4)), while short-dashed lines are given by Theorem 3.

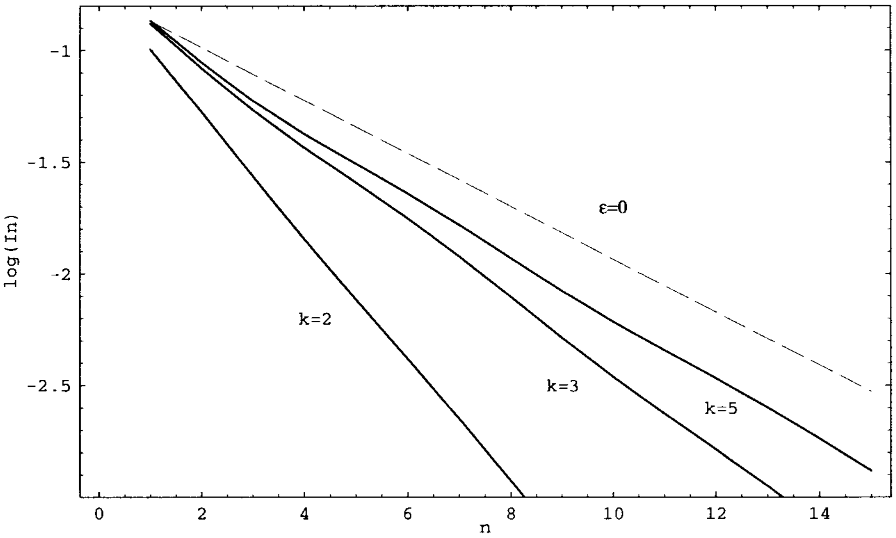

Fig. 3. Graph of $\log_{10} I(Z_n;\, S_0|Z^{(n-1)})$ versus $n$ for $p_g = 0.01$. Solid lines represent $I(Z_n;\, S_0|Z^{(n-1)})$ for $p_b = 0.5$, $g = 5 * 10^{-k}$, $b = 10^{-k}$ for $k = 2,\, 3$, and $5$. Long-dashed line is $\varepsilon = 0$ asymptote (learning rate $D(p^*,\, p_g) = D(p^*,\, p_b)$ given by (4.4)).

## V. Implications for Gilbert–Elliott Channel

The capacity of the Gilbert–Elliott channel is often computed by using (1.5) to approximate $H(Z)$ to some prescribed accuracy. Suppose that we wish to estimate how large $n$ needs to be so that $H(Z_n|Z^{(n-1)}) - H(Z) \leq 0.01$. A large estimate for $n$ would be discouraging since the computational complexity of either $H(Z_n|Z^{(n-1)})$ or $H(Z_n|Z^{(n-1)},\, S_0)$ grows as $2^n$. Let the Gilbert–Elliott parameters be, for example, $b = 0.1$, $g = 0.5$, $p_g = 0.01$, and $p_b = 0.5$. Then $B \approx 289$ and $\zeta \approx 1 - 10^{-8}$ in Theorem 1, and thus the estimate according to this theorem is $n \approx 10^9$ for $I(Z_n;\, S_0|Z^{(n-1)}) \approx 0.01$. This means that, according to Theorem 1, approximately $2^{10^9}$ floating-point operations are needed, seemingly putting an accurate estimate of $H(Z)$ well beyond our reach. Furthermore, observe that this estimate of $n$ goes to infinity as $p_g \to 0$ because $\zeta \to 1$.

However, our bound in Theorem 2 (Corollary 1) yields $I(Z_n;\, S_0|Z^{(n-1)}) \leq 5.49{\cdot}0.5^{n-2}$ independently of $p_g$. Therefore, $n = 9$ suffices. This much more reasonable value of $n$ is well within our computational ability!

Fig. 2 demonstrates the mixing and learning rates for $p_g = 0$. For small $n$, these curves have decay rates that are dominated by learning and are well approximated by the $\varepsilon = 0$ learning asymptote (given by long-dashed line and (4.4)), while for larger $n$, $I(Z_n;\, S_0|Z^{(n-1)})$ decays at the mixing rate given in Theorem 3. Observe the rather abrupt change in rate (knees in the curves) as $n$ increases. From the figure, we see that the mutual information is always bounded by the $\varepsilon = 0$ learning asymptote. Therefore, for $n = 20$, $I(Z_n;\, S_0|Z^{(n-1)})$ is approximately $10^{-5}$, independently of $\varepsilon$.

Fig. 3 demonstrates the mixing and learning rates for $p_g = 0.01$. Again, these curves have decay rates that are dominated by the $\varepsilon = 0$ learning asymptote (given by long-dashed line

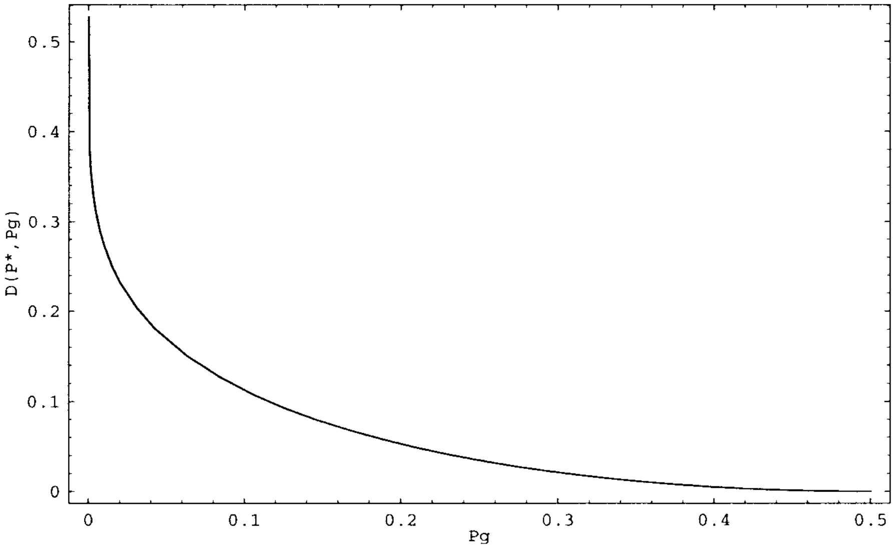

Fig. 4. Graph of the learning rate $D(p^*, p_g) = D(p^*, p_b)$ as a function of $p_g$ for $p_b = 0.5$. The curve is very steep near $p_g = 0$.

and (4.4)). Unlike the $p_g = 0$ case, we have not been able to explicitly identify the mixing rate.

The learning rates in Fig. 2 ($p_g = 0$) and Fig. 3 ($p_g = 0.01$) are quite different, even though they only differ slightly in their value of $p_g$. We plot in Fig. 4 the learning rate $D(p^*, p_g) = D(p^*, p_b)$ as a function of $p_g$ for $p_b = 0.5$. We see, in fact, that $D$ changes very rapidly with $p_g$ in the neighborhood of $p_g = 0$. Hence, learning rapidly becomes more difficult as $p_g$ increases.

To achieve capacity on the Gilbert–Elliott channel, Mushkin and Bar-David in [8] propose a scheme involving a $J \times N$ dimensional interleaver and similar deinterleaver (where $J$ represents depth, and $N$ length) and a metric calculator that effectively convert the Gilbert–Elliott channel into a single-input $J$-output essentially memoryless channel with the same channel capacity. At the transmitter, the interleaver is filled by row and transmitted by column, and at the receiver the deinterleaver is filled by column and read by row. Generally, $J$ is chosen large enough so that $(1 - b - g)^J$ is small and the channel errors within each row consequently happen approximately independently of one another.

Mushkin and Bar-David prove that the capacity achievable on the $j$th output of the $J$-output channel is given by

$$C^{(j)} = 1 - H(Z_j | Z^{(j-1)}), \qquad j = 1, \cdots, J.$$

The capacity of the Gilbert–Elliott channel itself is given by

$$C^* \overset{\text{def}}{=} \lim_{j \to \infty} C^{(j)}.$$

Hence, if $J$ is chosen sufficiently large, the capacity of the $J$-output channel is the same as the original Gilbert–Elliott channel. In [8] it is empirically observed that $C^{(j)}$ generally approaches $C^*$ quickly with increasing $j$. Our theoretical results explain this observation and show that $J$ does not need to be large for $C^{(J)}$ to be close to $C^*$. We have shown that

$$I(Z_j; S_0 | Z^{(j-1)}) = H(Z_j | Z^{(j-1)}) - H(Z_j | Z^{(j-1)}, S_0)$$

decays exponentially quickly and with a minimum rate that is independent of $b$ and $g$ when $p_g = 0$. Numerical evidence suggests that this is also true when $p_g > 0$. Therefore, $H(Z_j | Z^{(j-1)})$ must approach $H(Z)$ exponentially quickly with at least the minimum rate. We may conclude that if $J$ is chosen large enough to ensure that $(1 - b - g)^J$ is small, then $J$ will generally also be large enough to ensure that the capacity of the construction in [8] will be very close to the Gilbert–Elliott channel capacity.

## VI. Conclusion

We have argued that, in general, $I(Z_n; S_0 | Z^{(n-1)})$ decreases exponentially with $n$ and therefore $H(Z_n | Z^{(n-1)})$ and $H(Z_n | Z^{(n-1)}, S_0)$ both approach the limiting entropy $H(Z)$ exponentially quickly. We identified mixing and learning rates for functions of a Markov chain and showed that the worst-case convergence for the Gilbert–Elliott channel with $p_g = 0$ was given by the explicitly identified learning rate. Learning rates for other functions of Markov chains can also be computed using the techniques we have outlined. These rates are generally nonzero as long as the observed process has a distribution that is uniquely determined by the underlying state.

Based on the evidence we have given, we conjecture that, in general, the worst case convergence of $I(Z_n; S_0 | Z^{(n-1)})$ is the learning rate obtained when $q_{ij} = 0$ $(i \neq j)$. Hence the worst case convergence is still exponential, but at the learning rate. The amount of past needed to obtain a prescribed accuracy in (1.5) can thus be chosen independently of $q_{ij}$. Proving (or disproving) this conjecture will probably require a detailed analysis of the behavior of $I(Z_n; S_0 | Z^{(n-1)})$ as a function of $Q$.

## Appendix
## Proof of Theorem 3

Define $\alpha = \sqrt{(1 - b + q_b(1 - g))^2 - 4(1 - b - g)q_b}$. By evaluating (3.19) for $n = 1, 2$, after some straightforward algebra, one finds that

$$
\begin{aligned}
c_{11} &= \lambda_1 \frac{b(1-b+g(q_b-2)-q_b+\alpha)}{2\alpha(b+g)}\\
d_{11} &= \lambda_2 \frac{b(b-1-g(q_b-2)+q_b+\alpha)}{2\alpha(b+g)}\\
c_{21} &= \lambda_1 \frac{g(b-1+q_b(1-g-2b)+\alpha)}{2\alpha(b+g)}\\
d_{21} &= \lambda_2 \frac{g(1-b-q_b(1-g-2b)+\alpha)}{2\alpha(b+g)}\\
c_{22} &= \frac{-bg(q_b-1)(b-1+(g-1)q_b+\alpha)}{2\alpha(b+g)}\\
d_{22} &= \frac{-bg(q_b-1)(1-b+q_b(1-g)+\alpha)}{2\alpha(b+g)}\\
c_{12} &= \frac{b(q_b-1)(b(1+g)+(g-1)(1+(g-1)q_b+\alpha))}{2\alpha(b+g)}\\
d_{12} &= \frac{b(q_b-1)(-(b(1+g))+(g-1)(-1+q_b(1-g)+\alpha))}{2\alpha(b+g)}.
\end{aligned}
$$

The Markov property states that $S_0$ is independent of $Z_n$, $n \geq 1$, given $S_k$ for any $k = 1, \cdots, n-1$. Since $p_g = 0$, given that $Z_k = 2$ (a channel error has occurred) we are also given that $S_k = 2$ (the underlying chain is in the "bad" state). Hence

$$
\begin{aligned}
I(Z_n;\, S_0 | Z_1 = i_1, \cdots, Z_{k-1} = i_{k-1}, Z_k = 2,&\\
Z_{k+1} = i_{k+1}, \cdots, Z_{n-1} = i_{n-1}) &= 0
\end{aligned}
$$

and it follows that $I(Z_n;\, S_0|Z^{(n-1)})$ comprises terms only of the form $x_{ij}^{(n)}$ defined in (3.18)

$$
\begin{aligned}
&I(Z_n;\, S_0|Z^{(n-1)})\\
&\quad = x_{11}^{(n)} \log \left[\frac{x_{11}^{(n)}(x_{11}^{(n)}+x_{12}^{(n)}+x_{21}^{(n)}+x_{22}^{(n)})}{(x_{11}^{(n)}+x_{12}^{(n)})(x_{21}^{(n)}+x_{11}^{(n)})}\right]\\
&\qquad + x_{22}^{(n)} \log \left[\frac{x_{22}^{(n)}(x_{11}^{(n)}+x_{12}^{(n)}+x_{21}^{(n)}+x_{22}^{(n)})}{(x_{21}^{(n)}+x_{22}^{(n)})(x_{22}^{(n)}+x_{12}^{(n)})}\right]\\
&\qquad + x_{12}^{(n)} \log \left[\frac{x_{12}^{(n)}(x_{11}^{(n)}+x_{12}^{(n)}+x_{21}^{(n)}+x_{22}^{(n)})}{(x_{11}^{(n)}+x_{12}^{(n)})(x_{22}^{(n)}+x_{12}^{(n)})}\right]\\
&\qquad + x_{21}^{(n)} \log \left[\frac{x_{21}^{(n)}(x_{11}^{(n)}+x_{12}^{(n)}+x_{21}^{(n)}+x_{22}^{(n)})}{(x_{21}^{(n)}+x_{22}^{(n)})(x_{21}^{(n)}+x_{11}^{(n)})}\right]. \qquad \text{(A.1)}
\end{aligned}
$$

By using the identities

$$
\begin{aligned}
1 &= \frac{d_{11}(d_{11}+d_{12}+d_{21}+d_{22})}{(d_{11}+d_{12})(d_{21}+d_{11})}\\
&= \frac{d_{22}(d_{11}+d_{12}+d_{21}+d_{22})}{(d_{21}+d_{22})(d_{22}+d_{12})}\\
&= \frac{d_{12}(d_{11}+d_{12}+d_{21}+d_{22})}{(d_{11}+d_{12})(d_{22}+d_{12})}\\
&= \frac{d_{21}(d_{11}+d_{12}+d_{21}+d_{22})}{(d_{21}+d_{22})(d_{21}+d_{11})}
\end{aligned}
$$

the first summand can be written as

$$
\begin{aligned}
&x_{11}^{(n)} \log \left[\frac{x_{11}^{(n)}(x_{11}^{(n)}+x_{12}^{(n)}+x_{21}^{(n)}+x_{22}^{(n)})}{(x_{11}^{(n)}+x_{12}^{(n)})(x_{21}^{(n)}+x_{11}^{(n)})}\right]\\
&\quad = (c_{11}\lambda_1^{n-1} + d_{11}\lambda_2^{n-1})\\
&\qquad \cdot \left[\log\left(1+\frac{c_{11}}{d_{11}}\left(\frac{\lambda_1}{\lambda_2}\right)^{n-1}\right)\right.\\
&\qquad + \log\left(1+\frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\left(\frac{\lambda_1}{\lambda_2}\right)^{n-1}\right)\\
&\qquad - \log\left(1+\frac{c_{11}+c_{12}}{d_{11}+d_{12}}\left(\frac{\lambda_1}{\lambda_2}\right)^{n-1}\right)\\
&\qquad \left. - \log\left(1+\frac{c_{11}+c_{21}}{d_{11}+d_{21}}\left(\frac{\lambda_1}{\lambda_2}\right)^{n-1}\right)\right].
\end{aligned}
$$

If we repeat the preceding representation for each of the summands in (A.1), expand all of the logarithms using $\log(1+x) = x + o(x)$ as $x \to 0$, and use the identity

$$
\begin{aligned}
0 = {} & d_{11}\left(\frac{c_{11}}{d_{11}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{11}+c_{12}}{d_{11}+d_{12}} - \frac{c_{11}+c_{21}}{d_{11}+d_{21}}\right)\\
& + d_{12}\left(\frac{c_{12}}{d_{12}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{11}+c_{12}}{d_{11}+d_{12}} - \frac{c_{12}+c_{22}}{d_{12}+d_{22}}\right)\\
& + d_{21}\left(\frac{c_{21}}{d_{21}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{11}+c_{21}}{d_{11}+d_{21}} - \frac{c_{21}+c_{22}}{d_{21}+d_{22}}\right)\\
& + d_{22}\left(\frac{c_{22}}{d_{22}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{12}+c_{22}}{d_{12}+d_{22}} - \frac{c_{21}+c_{22}}{d_{21}+d_{22}}\right)
\end{aligned}
$$

the theorem is proven with

$$
\begin{aligned}
C' = {} & c_{11}\left(\frac{c_{11}}{d_{11}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{11}+c_{12}}{d_{11}+d_{12}} - \frac{c_{21}+c_{11}}{d_{21}+d_{11}}\right)\\
& + c_{12}\left(\frac{c_{12}}{d_{12}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{11}+c_{12}}{d_{11}+d_{12}} - \frac{c_{12}+c_{22}}{d_{12}+d_{22}}\right)\\
& + c_{21}\left(\frac{c_{21}}{d_{21}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{21}+c_{11}}{d_{21}+d_{11}} - \frac{c_{21}+c_{22}}{d_{21}+d_{22}}\right)\\
& + c_{22}\left(\frac{c_{22}}{d_{22}} + \frac{c_{11}+c_{12}+c_{21}+c_{22}}{d_{11}+d_{12}+d_{21}+d_{22}}\right.\\
&\left. - \frac{c_{22}+c_{12}}{d_{22}+d_{12}} - \frac{c_{22}+c_{21}}{d_{22}+d_{21}}\right). \qquad \text{(A.2)}
\end{aligned}
$$

## ACKNOWLEDGMENT

The authors thank one referee in particular for suggesting several improvements, including the elegant proof of Theorem 4.

## References

[1] J. J. Birch, "Approximations for the entropy for functions of Markov chains," *Ann. Math. Stat.*, vol. 33, pp. 930–938, 1962.

[2] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation.* New York: Wiley, 1990.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[4] B. D. Fritchman, "A binary channel characterization using partitioned Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 221–226, Apr. 1967.

[5] R. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[6] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–265, Sept. 1960.

[7] L. N. Kanal and A. R. K. Sastry, "Models for channels with memory and their applications to error control," *Proc. IEEE*, vol. 66, July 1978.

[8] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert–Elliot channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1277–1290, Nov. 1989.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[10] E. Seneta, *Non-Negative Matrices and Markov Chains.* New York: Springer-Verlag, 1981.

[11] W. Turin and M. M. Sondhi, "Modeling error sources in digital channels," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 340–347, Apr. 1993.

[12] A. Weiss and A. Shwartz, *Large Deviations for Performance Analysis: Queues, Communications, and Computing.* New York: Chapman & Hall, 1995.