



Title	Derivatives of entropy rate in special families of hidden Markov chains
Author(s)	Han, G; Marcus, B
Citation	IEEE Transactions On Information Theory, 2007, v. 53 n. 7, p. 2642-2652
Issued Date	2007
URL	http://hdl.handle.net/10722/135158
Rights	Creative Commons: Attribution 3.0 Hong Kong License

- [12] R. Ahlswede, "The capacity region of a channel with two senders and two receivers," *Ann. Prob.*, pp. 805–814, Oct. 1974.
- [13] A. B. Carliel, "A case where interference does not reduce capacity," *IEEE Trans. Inf. Theory*, vol. 21, pp. 569–570, Sep. 1975.

Derivatives of Entropy Rate in Special Families of Hidden Markov Chains

Guangyue Han and Brian Marcus, *Fellow, IEEE*

Abstract—Consider a hidden Markov chain obtained as the observation process of an ordinary Markov chain corrupted by noise. Recently Zuk *et al.* showed how, in principle, one can explicitly compute the derivatives of the entropy rate of at extreme values of the noise. Namely, they showed that the derivatives of standard upper approximations to the entropy rate actually stabilize at an explicit finite time. We generalize this result to a natural class of hidden Markov chains called "Black Holes." We also discuss in depth special cases of binary Markov chains observed in binary-symmetric noise, and give an abstract formula for the first derivative in terms of a measure on the simplex due to Blackwell.

Index Terms—Analyticity, entropy, entropy rate, hidden Markov chain, hidden Markov model, hidden Markov process.

I. INTRODUCTION

Let $Y = \{Y_\infty\}$ be a stationary Markov chain with a finite state alphabet $\{1, 2, \dots, B\}$. A function $Z = \{Z_\infty\}$ of the Markov chain Y with the form $Z = \Phi(Y)$ is called a hidden Markov chain; here Φ is a finite-valued function defined on $\{1, 2, \dots, B\}$, taking values in $\{1, 2, \dots, A\}$. Let Δ denote the probability transition matrix for Y ; it is well known that the entropy rate $H(Y)$ of Y can be analytically expressed using the stationary vector of Y and Δ . Let W be the simplex, comprising the vectors

$$\{w = (w_1, w_2, \dots, w_B) \in \mathbb{R}^B : w_i \geq 0, \sum_i w_i = 1\}$$

and let W_a be all $w \in W$ with $w_i = 0$ for $\Phi(i) \neq a$. For $a \in A$, let Δ_a denote the $B \times B$ matrix such that $\Delta_a(i, j) = \Delta(i, j)$ for j with $\Phi(j) = a$, and $\Delta_a(i, j) = 0$ otherwise. For $a \in A$, define the scalar-valued and vector-valued functions r_a and f_a on W by

$$r_a(w) = w \Delta_a \mathbf{1}$$

and

$$f_a(w) = w \Delta_a / r_a(w).$$

Note that f_a defines the action of the matrix Δ_a on the simplex W .

If Y is irreducible, it turns out that the entropy rate

$$H(Z) = - \int \sum_a r_a(w) \log r_a(w) dQ(w) \quad (1.1)$$

Manuscript received April 14, 2006; revised January 20, 2007. The material in this correspondence was presented at the IEEE International Symposium on Information Theory, Seattle, WA, July 2006.

The authors are with the Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada (e-mail: ghan@math.ubc.ca; marcus@math.ubc.ca).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2007.899467

where Q is Blackwell's measure [1] on W . This measure is defined as the limiting distribution $p(y_0 = \cdot | z_{-\infty}^0)$.

Recently, there has been a great deal of work on the entropy rate of a hidden Markov chain. Jacquet *et al.* [6] considered entropy rate of the hidden Markov chain Z , obtained by passing a binary Markov chain through a binary-symmetric channel with crossover probability ε , and computed the derivative of $H(Z)$ with respect to ε at $\varepsilon = 0$. For the same channel, Ordentlich and Weissman used Blackwell's measure to bound the entropy rate [11] and obtained an asymptotical formula for entropy rate [12]. For certain more general channels, Zuk *et al.* [16], [17] proved a "stabilizing" property of the derivatives of entropy rate of a hidden Markov chain and computed the Taylor series expansion for a special case. Several authors have observed that the entropy rate of a hidden Markov chain can be viewed as the top Lyapunov exponent of a random matrix product [5], [6], [3]. Under mild positivity assumptions, Han and Marcus [4] showed the entropy rate of a hidden Markov chain varies analytically as a function of the underlying Markov chain parameters.

In Section II, we establish a "stabilizing" property for the derivatives of the entropy rate in a family we call "Black Holes." Using this property, one can, in principle, explicitly calculate the derivatives of the entropy rate for this case, generalizing the results of [16], [17].

In Section III, we consider binary Markov chains corrupted by binary-symmetric noise. For this class, we obtain results on the support of Blackwell's measure, and for a special case, that we call the "nonoverlapping" case, we express the first derivative of the entropy rate as the sum of terms, involving Blackwell's measure, which have meaningful interpretations.

II. STABILIZING PROPERTY OF DERIVATIVES IN BLACK HOLE CASE

Suppose that for every $a \in A$, Δ_a is a rank one matrix, and every column of Δ_a is either strictly positive or all zeros. In this case, the image of f_a is a single point and each f_a is defined on the whole simplex W . Thus, we call this the *Black Hole* case. Analyticity of the entropy rate at a Black Hole follows from Theorem 1.1 of [4].

As an example, consider a binary-symmetric channel with crossover probability ε . Let $\{X_n\}$ be the input Markov chain with the transition matrix

$$\Pi = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}. \quad (2.2)$$

At time n the channel can be characterized by the following equation:

$$Z_n = X_n \oplus E_n$$

where \oplus denotes binary addition, E_n denotes the independent and identically distributed (i.i.d.) binary noise with $p_E(0) = 1 - \varepsilon$ and $p_E(1) = \varepsilon$, and Z_n denotes the corrupted output. Then $Y_n = (X_n, E_n)$ is jointly Markov, so $\{Z_n = \Phi(Y_n)\}$ is a hidden Markov chain with the corresponding

$$\Delta = \begin{bmatrix} \pi_{00}(1 - \varepsilon) & \pi_{00}\varepsilon & \pi_{01}(1 - \varepsilon) & \pi_{01}\varepsilon \\ \pi_{00}(1 - \varepsilon) & \pi_{00}\varepsilon & \pi_{01}(1 - \varepsilon) & \pi_{01}\varepsilon \\ \pi_{10}(1 - \varepsilon) & \pi_{10}\varepsilon & \pi_{11}(1 - \varepsilon) & \pi_{11}\varepsilon \\ \pi_{10}(1 - \varepsilon) & \pi_{10}\varepsilon & \pi_{11}(1 - \varepsilon) & \pi_{11}\varepsilon \end{bmatrix};$$

here, Φ maps states 1 and 4 to 0 and maps states 2 and 3 to 1 (the reader should not confuse Π with the 4×4 matrix Δ , which defines the hidden Markov chain via a deterministic function). When $\varepsilon = 0$

$$\Delta = \begin{bmatrix} \pi_{00} & 0 & \pi_{01} & 0 \\ \pi_{00} & 0 & \pi_{01} & 0 \\ \pi_{10} & 0 & \pi_{11} & 0 \\ \pi_{10} & 0 & \pi_{11} & 0 \end{bmatrix}.$$

In this case, the nonzero entries of Δ_0 and Δ_1 are restricted to a single column and so both Δ_0 and Δ_1 have rank one. If π_{ij} 's are all positive, then this is a Black Hole case.

For this channel, Zuk, *et al.* [17] and Ordentlich and Weissman [13] proved the ‘‘stabilizing’’ property of the derivatives of $H(Z)$ with respect to ε at $\varepsilon = 0$. We remark that some other special kinds of channels, which are Black Holes, are treated too in [17] and Taylor series expansion of $H(Z)$ around $\varepsilon = 0$ were computed in [16]. In this section, we show that, in general, the coefficients of a Taylor series expansion, centered at a Black Hole, can be explicitly computed.

Suppose that Δ is analytically parameterized by a vector variable $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m)$. For any smooth function f of ε and $\vec{n} = (n_1, n_2, \dots, n_m) \in \mathbb{Z}_+^m$, define

$$f^{(\vec{n})} = \frac{\partial^{|\vec{n}|} f}{\partial \varepsilon_1^{n_1} \partial \varepsilon_2^{n_2} \dots \partial \varepsilon_m^{n_m}};$$

here $|\vec{n}|$ denotes the order of the \vec{n} th derivative of f with respect to ε , and is defined as

$$|\vec{n}| = n_1 + n_2 + \dots + n_m.$$

For $\vec{n} = \vec{l}_1 + \vec{l}_2 + \dots + \vec{l}_k$, where $\vec{l}_j = (l_{j,1}, l_{j,2}, \dots, l_{j,m}) \in \mathbb{Z}_+^m$, define

$$\frac{\vec{n}!}{\vec{l}_1! \vec{l}_2! \dots \vec{l}_k!} = \prod_i \frac{n_i!}{l_{1,i}! l_{2,i}! \dots l_{k,i}!}.$$

We say $\vec{l} \preceq \vec{n}$, if every component of \vec{l} is less or equal to the corresponding one of \vec{n} , and $\vec{l} \prec \vec{n}$ if $\vec{l} \preceq \vec{n}$ and at least one component of \vec{l} is strictly less than the corresponding one of \vec{n} . For $\vec{l} \preceq \vec{n}$, define

$$C_{\vec{n}}^{\vec{l}} = \frac{\vec{n}!}{\vec{l}!(\vec{n} - \vec{l})!}.$$

Let $H_n(Z) = H(Z_0|Z_{-n}^-)$. It is well known that $H(Z) = \lim_{n \rightarrow \infty} H_n(Z)$. The following theorem says that at a Black Hole, one can calculate the derivatives of $H(Z)$ with respect to ε by taking the derivatives of $H_n(Z)$ with respect to ε for large enough n .

Theorem 2.1: If at $\varepsilon = \hat{\varepsilon}$, for every $a \in A$, Δ_a is a rank-one matrix, and every column of Δ_a is either a positive or a zero column, then for $\vec{n} = (n_1, n_2, \dots, n_m)$

$$H(Z)^{(\vec{n})} \Big|_{\varepsilon=\hat{\varepsilon}} = H_{|\vec{n}|}(Z)^{(\vec{n})} \Big|_{\varepsilon=\hat{\varepsilon}}.$$

In fact, we give a stronger result, Theorem 2.5, later in this section.

Proof: For simplicity we drop ε when the implication is clear from the context.

We shall first prove that for all sequences $z_{-\infty}^0$, the \vec{n} th derivative of $p(z_0|z_{-\infty}^-)$ stabilizes

$$p^{(\vec{n})}(z_0|z_{-\infty}^-) = p^{(\vec{n})}(z_0|z_{-|\vec{n}|-1}^-), \quad \text{at } \varepsilon = \hat{\varepsilon}. \quad (2.3)$$

Since $p(z_0|z_{-\infty}^-) = p(y_{-1} = \cdot | z_{-\infty}^-) \Delta_{z_0} \mathbf{1}$ (here \cdot represent the states of the Markov chain Y , thus $p(y_{-1} = \cdot | z_{-\infty}^-)$ is a row vector), it suffices to prove that for the \vec{n} th derivative of $x_i = p(y_i = \cdot | z_{-\infty}^-)$, we have

$$x_i^{(\vec{n})} = p^{(\vec{n})}(y_i = \cdot | z_{-\infty}^-) = p^{(\vec{n})}(y_i = \cdot | z_{i-|\vec{n}|}^-), \quad \text{at } \varepsilon = \hat{\varepsilon}. \quad (2.4)$$

Consider the iteration

$$x_i = \frac{x_{i-1} \Delta_{z_i}}{x_{i-1} \Delta_{z_i} \mathbf{1}}.$$

In other words, x_i can be viewed as a function of x_{i-1} and Δ_{z_i} . Let g denote this function. Since at $\varepsilon = \hat{\varepsilon}$, Δ_{z_i} is a rank-one matrix, we conclude that g is constant as a function of x_{i-1} . Thus, at $\varepsilon = \hat{\varepsilon}$

$$\begin{aligned} x_i &= p(y_i = \cdot | z_{-\infty}^-) \\ &= \frac{x_{i-1} \Delta_{z_i}}{x_{i-1} \Delta_{z_i} \mathbf{1}} = \frac{p(y_{i-1} = \cdot) \Delta_{z_i}}{p(y_{i-1} = \cdot) \Delta_{z_i} \mathbf{1}} = p(y_i = \cdot | z_i) \end{aligned} \quad (2.5)$$

where the third equality follows from the fact that g is a constant as a function of x_{i-1} .

When $|\vec{n}| = 1$, we have at $\varepsilon = \hat{\varepsilon}$

$$x_i^{(\vec{n})} = \frac{\partial g}{\partial \Delta_{z_i}} \Big|_{\varepsilon=\hat{\varepsilon}} (x_{i-1}, \Delta_{z_i}) \Delta_{z_i}^{(\vec{n})} + \frac{\partial g}{\partial x_{i-1}} \Big|_{\varepsilon=\hat{\varepsilon}} (x_{i-1}, \Delta_{z_i}) x_{i-1}^{(\vec{n})}.$$

Since at $\varepsilon = \hat{\varepsilon}$, g is a constant as a function of x_{i-1} , we have

$$\frac{\partial g}{\partial x_{i-1}} \Big|_{\varepsilon=\hat{\varepsilon}} (x_{i-1}, \Delta_{z_i}) = \frac{\partial(\text{a constant vector})}{\partial x_{i-1}} = \mathbf{0}.$$

It then follows from (2.5) that at $\varepsilon = \hat{\varepsilon}$

$$x_i^{(\vec{n})} = p^{(\vec{n})}(y_i = \cdot | z_{-\infty}^-) = p^{(\vec{n})}(y_i = \cdot | z_{i-1}^-).$$

When $|\vec{n}| > 1$, we have

$$x_i^{(\vec{n})} = \frac{\partial g}{\partial x_{i-1}} \Big|_{\varepsilon=\hat{\varepsilon}} (x_{i-1}, \Delta_{z_i}) x_{i-1}^{(\vec{n})} + \text{other terms}$$

where ‘‘other terms’’ involve only lower order (than $|\vec{n}|$) derivatives of x_{i-1} . By induction, we conclude that

$$x_i^{(\vec{n})} = p^{(\vec{n})}(y_i = \cdot | z_{-\infty}^-) = p^{(\vec{n})}(y_i = \cdot | z_{i-|\vec{n}|}^-).$$

at $\varepsilon = \hat{\varepsilon}$. We then have (2.4) and therefore (2.3) as desired.

By the proof of Theorem 1.1 of [4], the complexified $H_n(Z)$ uniformly converges to the complexified $H(Z)$, and so we can switch the limit operation and the derivative operation.

Thus, when $|\vec{l}| = 1$, at all ε

$$\begin{aligned} H^{(\vec{l})}(Z) &= \left(\lim_{k \rightarrow \infty} \sum_{z_{-k}^0} (p(z_{-k}^0) \log p(z_0|z_{-k}^-)) \right)^{(\vec{l})} \\ &= - \lim_{k \rightarrow \infty} \sum_{z_{-k}^0} \left(p^{(\vec{l})}(z_{-k}^0) \log p(z_0|z_{-k}^-) + p(z_{-k}^0) \frac{p^{(\vec{l})}(z_0|z_{-k}^-)}{p(z_0|z_{-k}^-)} \right). \end{aligned}$$

Since

$$\sum_{z_0} p(z_{-k}^0) \frac{p^{(\vec{l})}(z_0|z_{-k}^-)}{p(z_0|z_{-k}^-)} = \sum_{z_0} p(z_{-k}^-) p^{(\vec{l})}(z_0|z_{-k}^-) = 0$$

we have for all ε

$$H^{(\vec{l})}(Z) = - \lim_{k \rightarrow \infty} \sum_{z_{-k}^0} (p^{(\vec{l})}(z_{-k}^0) \log p(z_0|z_{-k}^-)). \quad (2.6)$$

At $\varepsilon = \hat{\varepsilon}$, we obtain

$$\begin{aligned} H^{(\vec{l})}(Z) &= - \lim_{k \rightarrow \infty} \sum_{z_{-k}^0} (p^{(\vec{l})}(z_{-k}^0) \log p(z_0|z_{-1})) \\ &= - \sum_{z_{-1}^0} (p^{(\vec{l})}(z_{-1}^0) \log p(z_0|z_{-1})) = H_1^{(\vec{l})}(Z) \end{aligned}$$

which establishes the theorem in the case when $|\vec{n}| = 1$.

When $|\vec{n}| > 1$, we fix any $\vec{l} \preceq \vec{n}$ with $|\vec{l}| = 1$. Using the fact that we can interchange the order of limit and derivative operations and using (2.6) and multivariate Leibnitz formula, we have for all ε

$$\begin{aligned} H^{(\vec{n})}(Z) &= (H^{(\vec{l})})^{(\vec{n}-\vec{l})}(Z) \\ &= -\lim_{k \rightarrow \infty} \sum_{z_{-k}^0} \sum_{\vec{j} \preceq \vec{n}-\vec{l}} C_{\vec{n}-\vec{l}}^{\vec{j}} p^{(\vec{l}+\vec{j})}(z_{-k}^0) (\log p(z_0|z_{-k}^{-1}))^{(\vec{n}-\vec{l}-\vec{j})}. \end{aligned}$$

Note that the term $(\log p(z_0|z_{-k}^{-1}))^{(\vec{n}-\vec{l}-\vec{j})}$ involves only the lower order (less than or equal to $|\vec{n}| - 1$) derivatives of $p(z_0|z_{-k}^{-1})$, which are already “stabilizing” in the sense of (2.3); so, we have at $\varepsilon = \hat{\varepsilon}$

$$\begin{aligned} H^{(\vec{n})}(Z) &= -\sum_{z_{-|\vec{n}|}^0} \sum_{\vec{j} \preceq \vec{n}-\vec{l}} C_{\vec{n}-\vec{l}}^{\vec{j}} p^{(\vec{l}+\vec{j})}(z_{-|\vec{n}|}^0) (\log p(z_0|z_{-|\vec{n}|}^{-1}))^{(\vec{n}-\vec{l}-\vec{j})} \\ &= H_{|\vec{n}|}^{(\vec{n})}(Z). \end{aligned}$$

We thus prove the theorem. \square

Remark 2.2: It follows from (2.5) that a hidden Markov chain at a Black Hole is, in fact, a Markov chain. Note that in the argument above the proof of the stabilizing property of the first derivative (as opposed to higher derivatives) requires only that the hidden Markov chain is Markov and that we can interchange the order of limit and derivative operations (instead of the stronger Black Hole property). Therefore, if a hidden Markov chain Z defined by $\hat{\Delta}$ and Φ is in fact a Markov chain, and the complexified $H_n(Z)$ uniformly converges to $H(Z)$ on some neighborhood of $\hat{\Delta}$ (e.g., if the conditions of Theorem 1.1, 6.1, or 7.5 of [4] hold), then at $\hat{\Delta}$, we have

$$H'(Z) = H'_1(Z), \quad (2.7)$$

For instance, consider the following hidden Markov chain Z defined by

$$\hat{\Delta} = \begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 0 & 1/6 & 5/6 \\ 7/8 & 1/8 & 0 \end{bmatrix}$$

with $\Phi(1) = 0$ and $\Phi(2) = \Phi(3) = 1$. Z is in fact a Markov chain (see [7, p. 134]), and one checks that $\hat{\Delta}$ satisfies the conditions in Theorem 7.5 in [4]. We conclude that for this example, (2.7) holds.

In the cases studied in [16], [17], [13], the authors obtained, using a finer analysis, a shorter “stabilizing length.” This shorter length can be derived for the Black Hole case as well, as shown in Theorem 2.5 below, even though the proof in [17] does not seem to work.

We need some preliminary lemmas for the proof of Theorem 2.5.

We say $\vec{l} < \vec{n}$ if either $(|\vec{l}| < |\vec{n}|)$ or $(|\vec{l}| = |\vec{n}|)$ and \vec{l} is less than \vec{n} lexicographically. By induction, one can prove that the formal derivative of $y \log y$ takes the following form of the first equation at the bottom of the page, where $E_{[\vec{a}_1, \vec{a}_2, \dots, \vec{a}_{m+1}]}$ is a real number, denoting the corresponding coefficient. Let $q_{\vec{i}}[y]$ denote the “coefficient” of $y^{(\vec{i})}$, which is a function of y and its formal derivatives (up to the $|\vec{i}|$ th order derivative). Thus, we have

$$(y \log y)^{(\vec{N})} = \sum_{i=1}^{\vec{N}} q_{\vec{i}}[y] y^{(\vec{i})} = \text{High}_{\vec{N}}[y] + \text{Low}_{\vec{N}}[y]$$

where

$$\text{High}_{\vec{N}}[y] = \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} q_{\vec{i}}[y] y^{(\vec{i})}$$

and

$$\text{Low}_{\vec{N}}[y] = \sum_{|\vec{i}| \leq \lceil (|\vec{N}|-1)/2 \rceil} q_{\vec{i}}[y] y^{(\vec{i})}.$$

For a sequence (without order)

$$\mathbf{a} = \underbrace{a_1, \dots, a_1}_{k_1 \text{ times}} \underbrace{a_2, \dots, a_2}_{k_2 \text{ times}} \dots \underbrace{a_l, \dots, a_l}_{k_l \text{ times}},$$

where a_j 's are pairwise distinct, let $\langle \mathbf{a} \rangle = k_1! k_2! \dots k_l!$.

Lemma 2.3: When $\lceil (|\vec{N}| + 1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|$

$$q_{\vec{i}}[y] = C_{\vec{N}}^{\vec{i}} (\log y + 1)^{(\vec{N}-\vec{i})}.$$

Proof: In this proof, we use $f(y)$ to denote the function $y \log y$. Then by multivariate Faa Di Bruno formula [2], [8], we have the second equation at the bottom of the page, where we used the fact that

$$\langle a_1, a_2, \dots, a_m \rangle = \langle a_2, \dots, a_m \rangle$$

if $|\vec{a}_1| \geq \lceil (|\vec{a}_1 + \dots + \vec{a}_m| + 1)/2 \rceil$. Bearing in mind that $f^{(1)}(y) = \log y + 1$, we prove the lemma. \square

$$\begin{aligned} (y \log y)^{(\vec{N})} &= \sum_{\vec{a}_1 \geq \vec{a}_2 \geq \dots \geq \vec{a}_{m+1}; |\vec{a}_1 + \vec{a}_2 + \dots + \vec{a}_{m+1}| = \vec{N}} E_{[\vec{a}_1, \vec{a}_2, \dots, \vec{a}_{m+1}]} \frac{y^{(\vec{a}_1)} y^{(\vec{a}_2)} \dots y^{(\vec{a}_{m+1})}}{y^m} + y^{(\vec{N})} (\log y + 1) \\ &= \sum_{\vec{i} \prec \vec{N}} y^{(\vec{a}_1 = \vec{i})} \sum_{\vec{a}_2 \geq \vec{a}_3 \geq \dots \geq \vec{a}_{m+1}} E_{[\vec{a}_1, \vec{a}_2, \dots, \vec{a}_{m+1}]} \frac{y^{(\vec{a}_2)} y^{(\vec{a}_3)} \dots y^{(\vec{a}_{m+1})}}{y^m} + y^{(\vec{N})} (\log y + 1) \end{aligned}$$

$$\begin{aligned} f(y)^{(\vec{N})} &= \sum_{\vec{a}_1 \geq \vec{a}_2 \geq \dots \geq \vec{a}_m; |\vec{a}_1 + \vec{a}_2 + \dots + \vec{a}_m| = \vec{N}} f^{(m)}(y) \frac{1}{\langle \vec{a}_1, \vec{a}_2, \dots, \vec{a}_m \rangle} \frac{\vec{N}!}{\vec{a}_1! \vec{a}_2! \dots \vec{a}_m!} y^{(\vec{a}_1)} y^{(\vec{a}_2)} \dots y^{(\vec{a}_m)} \\ &= \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{a}_1| \leq |\vec{N}|} C_{\vec{N}}^{\vec{a}_1} y^{(\vec{a}_1)} \sum_{\vec{a}_2 \geq \dots \geq \vec{a}_m; |\vec{a}_2 + \dots + \vec{a}_m| = \vec{N} - \vec{a}_1} (f^{(1)})^{(m-1)}(y) \frac{1}{\langle \vec{a}_2, \dots, \vec{a}_m \rangle} \frac{(\vec{N} - \vec{a}_1)!}{\vec{a}_2! \dots \vec{a}_m!} y^{(\vec{a}_2)} \dots y^{(\vec{a}_m)} + \text{Low}_{\vec{N}}[y] \\ &= \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{a}_1| \leq |\vec{N}|} C_{\vec{N}}^{\vec{a}_1} y^{(\vec{a}_1)} (f^{(1)}(y))^{\vec{N} - \vec{a}_1} + \text{Low}_{\vec{N}}[y] \end{aligned}$$

Lemma 2.4:

$$\text{Low}_{\vec{N}}[ax] = \sum_{|\vec{i}| \leq \lceil (|\vec{N}|-1)/2 \rceil} r_{\vec{i}}[a]x^{(\vec{i})} + \sum_{|\vec{j}| \leq \lceil (|\vec{N}|-1)/2 \rceil} s_{\vec{j}}[x]a^{(\vec{j})}$$

where $r_{\vec{i}}[a]$ is a function of a and its derivatives (up to order $\lceil (|\vec{N}| - 1)/2 \rceil$), and $s_{\vec{j}}[x]$ is a function of x and its derivatives (up to order $\lceil (|\vec{N}| - 1)/2 \rceil$). Also

$$s_0[x] = \text{Low}_{\vec{N}}[x].$$

Proof: By multivariate Leibnitz formula, we have

$$\begin{aligned} (ax \log(ax))^{(\vec{N})} &= \sum_{\vec{i}+\vec{j} \leq \vec{N}} \frac{\vec{N}!}{\vec{i}!\vec{j}!(\vec{N}-\vec{i}-\vec{j})!} a^{(\vec{i})}x^{(\vec{j})}(\log(ax))^{(\vec{N}-\vec{i}-\vec{j})} \\ &= \sum_{\vec{i}+\vec{j} \leq \vec{N}} \frac{\vec{N}!}{\vec{i}!\vec{j}!(\vec{N}-\vec{i}-\vec{j})!} a^{(\vec{i})}x^{(\vec{j})}(\log a + \log x)^{(\vec{N}-\vec{i}-\vec{j})}. \end{aligned}$$

Thus, there exist a function of a and its derivatives $t_{\vec{i}}[a]$, and a function of x and its derivatives $w_{\vec{j}}[x]$ such that

$$((ax) \log(ax))^{(\vec{N})} = \sum_{\vec{i} \leq \vec{N}} t_{\vec{i}}[a]x^{(\vec{i})} + \sum_{\vec{j} \leq \vec{N}} w_{\vec{j}}[x]a^{(\vec{j})}$$

with $w_0[x] = (x \log x)^{(\vec{N})}$.

By Lemma 2.3, we have

$$\begin{aligned} \text{High}_{\vec{N}}[ax] &= \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} q_{\vec{i}}ax^{(\vec{i})} \\ &= \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} C_{\vec{N}}^{\vec{i}}(\log a + \log x + 1)^{(\vec{N}-\vec{i})}(ax)^{(\vec{i})}. \end{aligned}$$

Thus, we conclude that there exist a function of a and its derivatives $u_{\vec{i}}[a]$, and a function of x and its derivatives $v_{\vec{j}}[x]$ such that

$$\text{High}_{\vec{N}}[ax] = \sum_{\vec{i} \leq \vec{N}} u_{\vec{i}}[a]x^{(\vec{i})} + \sum_{\vec{j} \leq \vec{N}} v_{\vec{j}}[x]a^{(\vec{j})}$$

with $v_0[x] = \text{High}_{\vec{N}}[x]$. Since

$$\text{Low}_{\vec{N}}[ax] = ((ax) \log(ax))^{(\vec{N})} - \text{High}_{\vec{N}}[ax]$$

the existence of $r_{\vec{i}}[a]$ and $s_{\vec{j}}[x]$ then follows, and they depend on the derivatives only up to $\lceil (|\vec{N}| - 1)/2 \rceil$, and $s_0[x] = \text{Low}_{\vec{N}}[x]$. \square

Theorem 2.5: If at $\varepsilon = \hat{\varepsilon}$, for every $a \in A$, Δ_a is a rank-one matrix, and every column of Δ_a is either a positive or a zero column, then for $\vec{n} = (n_1, n_2, \dots, n_m)$

$$H(Z)^{(\vec{n})} \Big|_{\varepsilon=\hat{\varepsilon}} = H_{\lceil (|\vec{n}|+1)/2 \rceil}(Z)^{(\vec{n})} \Big|_{\varepsilon=\hat{\varepsilon}}.$$

Proof: For simplicity, we drop ε when the implication is clear from the context. Recall that

$$\begin{aligned} H_n(Z) &= - \sum_{z_{-n}^0} p(z_{-n}^0) \log p(z_0 | z_{-n}^{-1}) \\ &= - \left(\sum_{z_{-n}^0} p(z_{-n}^0) \log p(z_{-n}^0) - \sum_{z_{-n}^{-1}} p(z_{-n}^{-1}) \log p(z_{-n}^{-1}) \right). \end{aligned}$$

With slight abuse of notation (by replacing the formal derivative with the derivative with respect to ε , we can define $\text{High}_{\vec{N}}[p(z_{-n}^0)] = \text{High}_{\vec{N}}[p^\varepsilon(z_{-n}^0)]$. Similarly, for $\text{Low}_{\vec{N}}[p(z_{-n}^0)]$, etc.), we have

$$\begin{aligned} (p(z_{-n}^0) \log p(z_{-n}^0))^{(\vec{N})} &= \text{High}_{\vec{N}}[p(z_{-n}^0)] + \text{Low}_{\vec{N}}[p(z_{-n}^0)], \\ (p(z_{-n}^{-1}) \log p(z_{-n}^{-1}))^{(\vec{N})} &= \text{High}_{\vec{N}}[p(z_{-n}^{-1})] + \text{Low}_{\vec{N}}[p(z_{-n}^{-1})]. \end{aligned}$$

Note that by Lemma 2.3, we have the first equation at the bottom of the page, and

$$\begin{aligned} \text{High}_{\vec{N}}[p(z_{-n}^{-1})] &= \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} C_{\vec{N}}^{\vec{i}}(\log p(z_{-n}^{-1}) + 1)^{(\vec{N}-\vec{i})} p(z_{-n}^{-1})^{(\vec{i})}. \end{aligned}$$

Thus, we get the second array of equations at the bottom of the page. So the higher derivative part stabilizes at $\lceil (|\vec{N}| + 1)/2 \rceil$, namely, for any $|\vec{n}| \geq \lceil (|\vec{N}| + 1)/2 \rceil$

$$\begin{aligned} \sum_{z_{-n}^0} \text{High}_{\vec{N}}[p(z_{-n}^0)] - \sum_{z_{-n}^{-1}} \text{High}_{\vec{N}}[p(z_{-n}^{-1})] &= \sum_{z_{-n}^0} \text{High}_{\vec{N}}[p(z_{-n}^0 | z_{-n}^{-1})] \\ &\quad - \sum_{z_{-n}^{-1}} \text{High}_{\vec{N}}[p(z_{-n}^{-1} | z_{-n}^0)]. \end{aligned}$$

And by Lemma 2.4, we have

$$\begin{aligned} \text{Low}_{\vec{N}}[p(z_{-n}^0)] &= \sum_{|\vec{i}| \leq \lceil (|\vec{N}|-1)/2 \rceil} r_{\vec{i}}[p(z_0 | z_{-n}^{-1})] p(z_{-n}^{-1})^{(\vec{i})} \\ &\quad + \sum_{|\vec{j}| \leq \lceil (|\vec{N}|-1)/2 \rceil} s_{\vec{j}}[p(z_{-n}^{-1})] p(z_0 | z_{-n}^{-1})^{(\vec{j})} \end{aligned}$$

$$\text{High}_{\vec{N}}[p(z_{-n}^0)] = \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} C_{\vec{N}}^{\vec{i}}(\log p(z_0 | z_{-n}^{-1}) + \log p(z_{-n}^{-1}) + 1)^{(\vec{N}-\vec{i})} p(z_{-n}^0)^{(\vec{i})}$$

$$\begin{aligned} \sum_{z_{-n}^0} \text{High}_{\vec{N}}[p(z_{-n}^0)] - \sum_{z_{-n}^{-1}} \text{High}_{\vec{N}}[p(z_{-n}^{-1})] &= \sum_{z_{-n}^0} \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} C_{\vec{N}}^{\vec{i}}(\log p(z_0 | z_{-n}^{-1}) + \log p(z_{-n}^{-1}) - \log p(z_{-n}^{-1}))^{(\vec{N}-\vec{i})} p(z_{-n}^0)^{(\vec{i})} \\ &= \sum_{z_{-n}^0} \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} C_{\vec{N}}^{\vec{i}}(\log p(z_0 | z_{-n}^{-1}))^{(\vec{N}-\vec{i})} p(z_{-n}^0)^{(\vec{i})} \\ &= \sum_{z_{-n}^0} \sum_{\lceil (|\vec{N}|+1)/2 \rceil \leq |\vec{i}| \leq |\vec{N}|} C_{\vec{N}}^{\vec{i}}(\log p(z_0 | z_{-n}^{-1} | z_{-n}^0))^{(\vec{N}-\vec{i})} p(z_{-n}^0 | z_{-n}^{-1} | z_{-n}^0)^{(\vec{i})}. \end{aligned}$$

with $s_0[p(z_{-n}^{-1})] = \text{Low}_{\tilde{N}}[p(z_{-n}^{-1})]$. Thus

$$\begin{aligned} & \sum_{z_{-n}^0} \text{Low}_{\tilde{N}}[p(z_{-n}^0)] - \sum_{z_{-n}^{-1}} \text{Low}_{\tilde{N}}[p(z_{-n}^{-1})] \\ &= \sum_{z_{-n}^0} \sum_{|\vec{i}| \leq \lceil (|\tilde{N}|-1)/2 \rceil} r_{\vec{i}}[p(z_0|z_{-n}^{-1})] p(z_{-n}^{-1})^{(\vec{i})}. \\ &= \sum_{z_{-n}^0} \sum_{|\vec{i}| \leq \lceil (|\tilde{N}|-1)/2 \rceil} r_{\vec{i}}[p(z_0|z_{-\lceil (|\tilde{N}+1)/2 \rceil}^{-1})] p(z_{-\lceil (|\tilde{N}+1)/2 \rceil}^{-1})^{(\vec{i})}. \end{aligned}$$

Consequently, the lower derivative part stabilizes at $\lceil (|\tilde{N}| + 1)/2 \rceil$ as well, namely, for any $n \geq \lceil (|\tilde{N}| + 1)/2 \rceil$

$$\begin{aligned} & \sum_{z_{-n}^0} \text{Low}_{\tilde{N}}[p(z_{-n}^0)] - \sum_{z_{-n}^{-1}} \text{Low}_{\tilde{N}}[p(z_{-n}^{-1})] \\ &= \sum_{z_{-\lceil (|\tilde{N}+1)/2 \rceil}^0} \text{Low}_{\tilde{N}}[p(z_{-\lceil (|\tilde{N}+1)/2 \rceil}^0)] \\ & \quad - \sum_{z_{-\lceil (|\tilde{N}+1)/2 \rceil}^{-1}} \text{Low}_{\tilde{N}}[p(z_{-\lceil (|\tilde{N}+1)/2 \rceil}^{-1})]. \end{aligned}$$

The theorem then follows. \square

Remark 2.6: For an irreducible stationary Markov chain Y with probability transition matrix Δ , let Y^{-1} denote its reverse Markov chain. It is well known that the probability transition matrix of Y^{-1} is $\text{diag}(\pi_1^{-1}, \pi_2^{-1}, \dots, \pi_B^{-1}) \Delta^t \text{diag}(\pi_1, \pi_2, \dots, \pi_B)$, where Δ^t denotes the transpose of Δ and $(\pi_1, \pi_2, \dots, \pi_B)$ is the stationary vector of Y . Therefore, if Δ^t is a Black Hole case, the derivatives of $H(Z^{-1})$ (here, Z^{-1} is the reverse hidden Markov chain defined by $Z^{-1} = \Phi(Y^{-1})$) also stabilize. It then follows from $H(Z) = H(Z^{-1})$ that the derivatives of $H(Z)$ also stabilize.

III. BINARY MARKOV CHAINS CORRUPTED BY BINARY-SYMMETRIC NOISE

In this section, we further study hidden Markov chains obtained by passing binary Markov chains through binary-symmetric channels with crossover probability ε (described in the beginning of Section II). We take a concrete approach to study $H(Z)$, and we will “compute” $H'(Z)$ in terms of Blackwell’s measure.

Recall that the Markov chain is defined by a 2×2 stochastic matrix $\Pi = [\pi_{ij}]$. When $\det(\Pi) = 0$, the rows of Π are identical, and so Y is an i.i.d. random sequence with distribution (π_{00}, π_{01}) . Thus, Z is an i.i.d. random sequence with distribution $(\pi, 1 - \pi)$ where $\pi = \pi_{00}(1 - \varepsilon) + \pi_{01}\varepsilon$. So

$$H(Z) = -\pi \log \pi - (1 - \pi) \log(1 - \pi).$$

From now through the end of Section III-B, we **assume** the following:

- $\det(\Pi) > 0$ – and –
- all $\pi_{ij} > 0$ – and –
- $\varepsilon > 0$.

We remark that the condition $\det(\Pi) > 0$ is purely for convenience. Results in this section will hold with the condition $\det(\Pi) < 0$ through similar arguments, unless specified otherwise.

The integral formula (1.1) expresses $H(Z)$ in terms of the measure Q on the four-dimensional simplex; namely, Q is the distribution of $p((y_0, e_0)|z_{-\infty}^0)$. However, in the case under consideration, $H(Z)$ can be expressed as an integral on the real line [11], which we review as follows.

From the chain rule of probability theory

$$\begin{aligned} p(z_1^i, y_i) &= p(z_1^{i-1}, z_i, y_{i-1} = 0, y_i) + p(z_1^{i-1}, z_i, y_{i-1} = 1, y_i) \\ &= p(z_i, y_i | z_1^{i-1}, y_{i-1} = 0) p(z_1^{i-1}, y_{i-1} = 0) \\ & \quad + p(z_i, y_i | z_1^{i-1}, y_{i-1} = 1) p(z_1^{i-1}, y_{i-1} = 1), \end{aligned}$$

and

$$\begin{aligned} p(z_i, y_i | z_1^{i-1}, y_{i-1} = 0) \\ &= p(z_i | z_1^{i-1}, y_i, y_{i-1} = 0) p(y_i | z_1^{i-1}, y_{i-1} = 0) \\ &= p(z_i | y_i) p(y_i | y_{i-1} = 0) = p_E(e_i) p(y_i | y_{i-1} = 0). \end{aligned}$$

Let $a_i = p(z_1^i, y_i = 0)$ and $b_i = p(z_1^i, y_i = 1)$. The pair (a_i, b_i) satisfies the following dynamical system:

$$\begin{cases} a_i = p_E(z_i) \pi_{00} a_{i-1} + p_E(z_i) \pi_{10} b_{i-1} \\ b_i = p_E(\bar{z}_i) \pi_{01} a_{i-1} + p_E(\bar{z}_i) \pi_{11} b_{i-1}. \end{cases}$$

Let $x_i = a_i/b_i$, we have a dynamical system with just one variable

$$x_{i+1} = f_{z_{i+1}}(x_i),$$

where

$$f_z(x) = \frac{p_E(z) \pi_{00} x + \pi_{10}}{p_E(\bar{z}) \pi_{01} x + \pi_{11}}, \quad z = 0, 1$$

starting with

$$x_0 = \pi_{10}/\pi_{01}.$$

We are interested in the invariant distribution of x_n , which is closely related to Blackwell’s distribution of $p((y_0, e_0)|z_{-\infty}^0)$. Now

$$\begin{aligned} p(y_i = 0 | z_1^{i-1}) \\ &= p(y_i = 0, y_{i-1} = 0 | z_1^{i-1}) + p(y_i = 0, y_{i-1} = 1 | z_1^{i-1}) \\ &= \pi_{00} p(y_{i-1} = 0 | z_1^{i-1}) + \pi_{10} p(y_{i-1} = 1 | z_1^{i-1}) \\ &= \pi_{00} \frac{a_{i-1}}{a_{i-1} + b_{i-1}} + \pi_{10} \frac{b_{i-1}}{a_{i-1} + b_{i-1}} \\ &= \pi_{00} \frac{x_{i-1}}{1 + x_{i-1}} + \pi_{10} \frac{1}{1 + x_{i-1}}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} p(y_i = 1 | z_1^{i-1}) \\ &= p(y_i = 1, y_{i-1} = 0 | z_1^{i-1}) + p(y_i = 1, y_{i-1} = 1 | z_1^{i-1}) \\ &= \pi_{01} \frac{x_{i-1}}{1 + x_{i-1}} + \pi_{11} \frac{1}{1 + x_{i-1}}. \end{aligned}$$

Further computation leads to

$$\begin{aligned} p(z_i = 0 | z_1^{i-1}) \\ &= p(y_i = 0, e_i = 0 | z_1^{i-1}) + p(y_i = 1, e_i = 1 | z_1^{i-1}) \\ &= p(e_i = 0) p(y_i = 0 | z_1^{i-1}) + p(e_i = 1) p(y_i = 1 | z_1^{i-1}) \\ &= ((1 - \varepsilon) \pi_{00} + \varepsilon \pi_{01}) \frac{x_{i-1}}{1 + x_{i-1}} + ((1 - \varepsilon) \pi_{10} + \varepsilon \pi_{11}) \frac{1}{1 + x_{i-1}} \\ &= r_0(x_{i-1}), \end{aligned}$$

where

$$r_0(x) = \frac{((1 - \varepsilon) \pi_{00} + \varepsilon \pi_{01}) x + ((1 - \varepsilon) \pi_{10} + \varepsilon \pi_{11})}{x + 1}. \quad (3.8)$$

Similarly, we have

$$\begin{aligned} p(z_i = 1 | z_1^{i-1}) \\ &= p(y_i = 0, e_i = 1 | z_1^{i-1}) + p(y_i = 1, e_i = 0 | z_1^{i-1}) \\ &= p(e_i = 1) p(y_i = 0 | z_1^{i-1}) + p(e_i = 0) p(y_i = 1 | z_1^{i-1}) \\ &= ((\varepsilon \pi_{00} + (1 - \varepsilon) \pi_{01}) \frac{x_{i-1}}{1 + x_{i-1}} + (\varepsilon \pi_{10} + (1 - \varepsilon) \pi_{11}) \frac{1}{1 + x_{i-1}} \\ &= r_1(x_{i-1}) \end{aligned}$$

where

$$r_1(x) = \frac{(\varepsilon\pi_{00} + (1-\varepsilon)\pi_{01})x + (\varepsilon\pi_{10} + (1-\varepsilon)\pi_{11})}{x+1}. \quad (3.9)$$

Now we write

$$p(z_i \in E | x_{i-1}) = \sum_{\{a | f_a(x_{i-1}) \in E\}} p(z_i = a | x_{i-1}).$$

Note that

$$\begin{aligned} p(z_i = 0 | x_{i-1}) &= p(z_i = 0 | z_1^{i-1}) = r_0(x_{i-1}) \\ p(z_i = 1 | x_{i-1}) &= p(z_i = 1 | z_1^{i-1}) = r_1(x_{i-1}). \end{aligned}$$

The analysis above leads to

$$\begin{aligned} p(x_i \in E) &= \int_{f_0^{-1}(E)} r_0(x_{i-1}) dp(x_{i-1}) + \int_{f_1^{-1}(E)} r_1(x_{i-1}) dp(x_{i-1}). \end{aligned}$$

Abusing notation, we let Q denote the limiting distribution of x_i (the limiting distribution exists due to the martingale convergence theorem) and obtain

$$Q(E) = \int_{f_0^{-1}(E)} r_0(x) dQ(x) + \int_{f_1^{-1}(E)} r_1(x) dQ(x). \quad (3.10)$$

We may now compute the entropy rate of Z_i in terms of Q . Note that

$$\begin{aligned} E(\log p(z_i | z_1^{i-1})) &= E(p(z_i = 0 | z_1^{i-1}) \log p(z_i = 0 | z_1^{i-1}) \\ &\quad + p(z_i = 1 | z_1^{i-1}) \log p(z_i = 1 | z_1^{i-1})) \\ &= E(r_0(x_{i-1}) \log r_0(x_{i-1}) + r_1(x_{i-1}) \log r_1(x_{i-1})). \end{aligned}$$

Thus, (1.1) becomes

$$H(Z) = - \int (r_0(x) \log r_0(x) + r_1(x) \log r_1(x)) dQ(x). \quad (3.11)$$

A. Properties of Q

Since $\det(\Pi) > 0$, f_0 and f_1 are increasing continuous functions bounded from above, and $f_0(0)$ and $f_1(0)$ are positive; therefore, they each have a unique positive fixed point p_0 and p_1 . Since f_1 is dominated by f_0 , we conclude $p_1 \leq p_0$. Let

- I denote the interval $[p_1, p_0]$ – and –
- $L = \bigcup_{n=1}^{\infty} L_n$ where

$$L_n = \{f_{i_1} \circ f_{i_2} \cdots \circ f_{i_n}(p_j) | i_1, i_2, \dots, i_n \in \{0, 1\}, j = 0, 1\}.$$

Let $I_{i_1 i_2 \dots i_n}$ denote $f_{i_n} \circ f_{i_{n-1}} \circ \dots \circ f_{i_1}(I)$, and $p_{i_1 i_2 \dots i_n}$ denote $p(z_1 = i_1, z_2 = i_2, \dots, z_n = i_n)$. The *support* of a probability measure Q , denoted $\text{supp}(Q)$, is defined as the smallest closed subset with measure one.

Theorem 3.1: $\text{supp}(Q) = \bar{L}$.

Proof: First, by straightforward computation, one can check that $f'_0(p_0)$ and $f'_1(p_1)$ are both less than 1. Thus, p_0 and p_1 are attracting fixed points. Since p_i is the unique positive fixed point of f_i , it follows that the entire positive half of the real line is in the domain of attraction of each f_i , i.e., for any $p > 0$, $f_i^{(n)}(p)$ approaches p_i (here the superscript (n) denotes the composition of n copies of the function).

We claim that both p_0 and p_1 are in $\text{supp}(Q)$. If p_0 is not in the support, then there is a neighborhood I_{p_0} containing p_0 with Q -measure 0. For any point $p > 0$, for some n , $f_0^{(n)}(p) \in I_{p_0}$. Thus,

by (3.10) there is a neighborhood of p with Q -measure 0. It follows that $Q([0, \infty)) = 0$. On the other hand, Q is the limiting distribution of $x_i > 0$ and so $Q([0, \infty)) = 1$. This contradiction shows that $p_0 \in \text{supp}(Q)$. Similarly, $p_1 \in \text{supp}(Q)$.

By (3.10), we deduce

$$f_i(\text{supp}(Q)) \subseteq \text{supp}(Q).$$

It follows that $L \subseteq \text{supp}(Q)$. Thus, $\bar{L} \subseteq \text{supp}(Q)$.

Since $f_i((0, \infty))$ is contained in a compact set, we may assume f_i is a contraction mapping (otherwise, compose f_0 or f_1 enough many times to make the composite mapping a contraction as we argued in [4]). In this case, the set of accumulation points of $\{f_{i_n} \circ f_{i_{n-1}} \cdots \circ f_{i_1}(p) | i_1, i_2, \dots, i_n \in \{0, 1\}, p > 0\}$ does not depend on p . Since any point in $\text{supp}(Q)$ has to be an accumulation point of $\{f_{i_n} \circ f_{i_{n-1}} \cdots \circ f_{i_1}(\pi_{10}/\pi_{01}) | i_1, i_2, \dots, i_n \in \{0, 1\}\}$, it has to be an accumulation point of L as well, which implies $\text{supp}(Q) \subseteq \bar{L}$. \square

It is easy to see the following.

Lemma 3.2: The following statements are equivalent.

1. $f_0(I) \cup f_1(I) \subsetneq I$.
2. $f_0(I) \cap f_1(I) = \emptyset$.
3. $f_1(p_0) < f_0(p_1)$.

Theorem 3.3: $\text{supp}(Q)$ is either a Cantor set or a closed interval. Specifically

1. $\text{supp}(Q)$ is a Cantor set if $f_0(I) \cup f_1(I) \subsetneq I$;
2. $\text{supp}(Q) = I$ if equivalently $f_0(I) \cup f_1(I) = I$.

Proof: Suppose that $f_0(I) \cup f_1(I) \subsetneq I$. If $(i_1, i_2, \dots, i_n) \neq (j_1, j_2, \dots, j_n)$, then

$$I_{i_1 i_2 \dots i_n} \cap I_{j_1 j_2 \dots j_n} = \emptyset.$$

Define

$$I_{\langle n \rangle} = \bigcup_{i_1, i_2, \dots, i_n} I_{i_1 i_2 \dots i_n}.$$

Alternatively we can construct $I_{\langle n \rangle}$ as follows: let

$$I^d = (f_1(p_0), f_0(p_1))$$

then

$$I_{\langle n+1 \rangle} = I_{\langle n \rangle} \setminus \bigcup_{i_1, i_2, \dots, i_n} f_{i_n} \circ f_{i_{n-1}} \circ \dots \circ f_{i_1}(I^d).$$

Let $I_{\langle \infty \rangle} = \bigcap_{n=1}^{\infty} I_{\langle n \rangle}$. It follows from the way it is constructed that I_{∞} is a Cantor set (think of I^d as a “deleted” interval), and $\bar{L} = I_{\langle \infty \rangle}$. Thus, by Theorem 3.1, $\text{supp}(Q) = \bar{L}$ is a Cantor set.

Suppose $f_0(I) \cup f_1(I) = I$. In this case, for any point $p \in I$, and for all n , there exists i_1, i_2, \dots, i_n such that

$$p \in I_{i_1 i_2 \dots i_n}.$$

From the fact that f_0 and f_1 are both contraction mappings (again, otherwise compose f_0 or f_1 enough many times to make the composite mapping a contraction as we argued in [4]), we deduce that the length of $I_{i_1 i_2 \dots i_n}$ is exponentially decreasing with respect to n . It follows that \bar{L} is dense in I , and therefore, $\text{supp}(Q) = \bar{L} = I$. \square

Theorem 3.4: Q is a continuous measure, namely for any point $p \in \text{supp}(Q)$, and for any $\eta > 0$, there exists an interval I_p containing p with $Q(I_p) < \eta$ (or equivalently Q has no point mass).

Proof: Assume that there exists $p \in I$ such that for any interval containing p , $Q(I_p) > \eta_0$, where η_0 is a positive constant. Let $\xi = \max\{r_0(x), r_1(x) : x \in I\}$. One checks that $0 < \xi < 1$. By (3.10), we have

$$\frac{1}{\xi} Q(I_p) \leq Q(f_0^{-1}(I_p)) + Q(f_1^{-1}(I_p)).$$

Iterating, we obtain

$$\left(\frac{1}{\xi}\right)^n \eta_0 \leq \sum_{i_1, i_2, \dots, i_n} Q(f_{i_1}^{-1} \circ f_{i_2}^{-1} \circ \dots \circ f_{i_n}^{-1}(I_p)).$$

For fixed n , if we choose I_p small enough, then

$$f_{i_1}^{-1} \circ f_{i_2}^{-1} \circ \dots \circ f_{i_n}^{-1}(I_p) \cap f_{j_1}^{-1} \circ f_{j_2}^{-1} \circ \dots \circ f_{j_n}^{-1}(I_p) = \emptyset$$

for $(i_1, i_2, \dots, i_n) \neq (j_1, j_2, \dots, j_n)$. It follows in this case that

$$Q(I) \geq \sum_{i_1, i_2, \dots, i_n} Q((f_{i_1}^{-1} \circ f_{i_2}^{-1} \circ \dots \circ f_{i_n}^{-1}(I_p))) \geq \left(\frac{1}{\xi}\right)^n \eta_0.$$

Therefore, for large n , we deduce

$$Q(I) > 1$$

which contradicts the fact that Q is a probability measure. \square

By virtue of Lemma 3.2, it makes sense to refer to Case 1 in Theorem 3.3 as the *nonoverlapping* case. We now focus on this case. Note that this is the case whenever ε is sufficiently small; also, it turns out that for some values of π_{ij} 's, the nonoverlapping case holds for all ε .

Starting with $x_0 = \pi_{10}/\pi_{01}$, and iterating according to $x_n = f_{z_n}(\varepsilon, x_{n-1})$, each word $z = z_1, z_2, \dots, z_n$ determines a point $x_n = x_n(z)$ with probability $p(z_1, z_2, \dots, z_n)$. In the nonoverlapping case, the map $z \mapsto x_n$ is one-to-one. We order the distinct points $\{x_n\}$ from left to right as

$$x_{n,1}, x_{n,2}, \dots, x_{n,2^n}$$

with the associated probabilities

$$p_{n,1}, p_{n,2}, \dots, p_{n,2^n}.$$

This defines a sequence of distribution Q_n which converge weakly to Q . In particular, by the continuity of Q , $Q_n(J) \rightarrow Q(J)$ for any interval J .

Theorem 3.5: In the nonoverlapping case

$$Q(I_{i_1 i_2 \dots i_n}) = Q_n(I_{i_1 i_2 \dots i_n}) = p_{i_1 i_2 \dots i_n}.$$

Proof: We have

$$Q_n(I_{i_1 i_2 \dots i_n}) = p(z_1 = i_1, z_2 = i_2, \dots, z_n = i_n).$$

Furthermore

$$\begin{aligned} Q_{n+1}(I_{i_1 i_2 \dots i_n}) &= Q_{n+1}(I_{0 i_1 i_2 \dots i_n}) + Q_{n+1}(I_{1 i_1 i_2 \dots i_n}) \\ &= p(z_0 = 0, z_1 = i_1, z_2 = i_2, \dots, z_n = i_n) \\ &\quad + p(z_0 = 1, z_1 = i_1, z_2 = i_2, \dots, z_n = i_n) \\ &= p(z_1 = i_1, z_2 = i_2, \dots, z_n = i_n). \end{aligned}$$

Iterating one shows that for $m \geq n$

$$Q_m(I_{i_1 i_2 \dots i_n}) = Q_n(I_{i_1 i_2 \dots i_n}) = p_{i_1 i_2 \dots i_n}.$$

By the continuity of Q (Theorem 3.4)

$$Q(I_{i_1 i_2 \dots i_n}) = p_{i_1 i_2 \dots i_n}. \quad \square$$

From this, as in [11], [12] we can derive bounds for the entropy rate. Let

$$r(x) = -(r_0(x) \log r_0(x) + r_1(x) \log r_1(x)).$$

Using (3.11) and Theorem 3.5, we obtain the following.

Theorem 3.6: In the nonoverlapping case

$$\sum_{i_1 i_2 \dots i_n} r_{i_1 i_2 \dots i_n}^m p_{i_1 i_2 \dots i_n} \leq H(Z) \leq \sum_{i_1 i_2 \dots i_n} r_{i_1 i_2 \dots i_n}^M p_{i_1 i_2 \dots i_n}$$

where

$$r_{i_1 i_2 \dots i_n}^m = \min_{x \in I_{i_1 i_2 \dots i_n}} r(x)$$

and

$$r_{i_1 i_2 \dots i_n}^M = \max_{x \in I_{i_1 i_2 \dots i_n}} r(x).$$

Proof: This follows immediately from the formula for the entropy rate $H(Z)$ (3.11). \square

B. Computation of the First Derivative in Nonoverlapping Case

To emphasize the dependence on ε , we write $p_{n,i}(\varepsilon) = p_{n,i}$, $x_{n,i}(\varepsilon) = x_{n,i}$, $p_0(\varepsilon) = p_0$, $p_1(\varepsilon) = p_1$, and $Q_n(\varepsilon) = Q_n$. Let $F_n(\varepsilon, x)$ denote the cumulative distribution function of $Q_n(\varepsilon)$. Let $H_n^\varepsilon(Z)$ be the finite approximation to $H^\varepsilon(Z)$. It can be easily checked that

$$H_n^\varepsilon(Z) = \int_I r(\varepsilon, x) dQ_n(\varepsilon)$$

and we can rewrite (3.11) as

$$H^\varepsilon(Z) = \int_I r(\varepsilon, x) dQ(\varepsilon).$$

In Theorem 3.7, we express the derivative of the entropy rate, with respect to ε , as the sum of four terms which have meaningful interpretations. Essentially, we are differentiating $H^\varepsilon(Z)$ with respect to ε under the integral sign, but care must be taken since $Q(\varepsilon)$ is generally singular and varies with ε .

Rewriting this using the Riemann–Stieltjes integral and applying integration by parts, we obtain

$$\begin{aligned} H_n^\varepsilon(Z) &= \int_I r(\varepsilon, x) dF_n(\varepsilon, x) \\ &= F_n(\varepsilon, x) r(\varepsilon, x) \Big|_{p_1(\varepsilon)}^{p_0(\varepsilon)} - \int_I F_n(\varepsilon, x) g(\varepsilon, x) dx \end{aligned}$$

where $g(\varepsilon, x) = \frac{\partial r(\varepsilon, x)}{\partial x}$.

From now on $'$ denotes the derivative with respect to ε . Now

$$H_n^\varepsilon(Z)' = r(\varepsilon, p_0(\varepsilon))' - D_n(\varepsilon)$$

where

$$D_n(\varepsilon) = \lim_{h \rightarrow 0} \frac{\int_I F_n(\varepsilon+h, x) g(\varepsilon+h, x) dx - \int_I F_n(\varepsilon, x) g(\varepsilon, x) dx}{h}.$$

We can decompose $D_n(\varepsilon)$ into two terms

$$D_n(\varepsilon) = D_n^1(\varepsilon) + D_n^2(\varepsilon),$$

$$D_n^1(\varepsilon) = \lim_{h \rightarrow 0} \int_I \frac{F_n(\varepsilon+h, x) - F_n(\varepsilon, x)}{h} g(\varepsilon, x) dx,$$

and

$$D_n^2(\varepsilon) = \int_I F_n(\varepsilon, x) g'(\varepsilon, x) dx.$$

In order to compute $D_n^1(\varepsilon)$, we partition I into two pieces: 1) small intervals $(x_{n,i}(\varepsilon), x_{n,i}(\varepsilon + h))$ and 2) the complement of the union of these neighborhoods, to yield

$$D_n^1(\varepsilon) = \lim_{h \rightarrow 0} \int_I \frac{F_n(\varepsilon + h, x) - F_n(\varepsilon, x)}{h} g(\varepsilon, x) dx = \\ - \sum_i p_{n,i}(\varepsilon) x_{n,i}(\varepsilon)' g(\varepsilon, x_{n,i}(\varepsilon)) + \int_I F_n'(\varepsilon, x) g(\varepsilon, x) dx.$$

Combining the foregoing expressions, we arrive at an expression for $H_n^\varepsilon(Z)'$

$$H_n^\varepsilon(Z)' = r(\varepsilon, p_0(\varepsilon))' + \sum_i p_{n,i}(\varepsilon) x_{n,i}'(\varepsilon) g(\varepsilon, x_{n,i}(\varepsilon)) \\ - \int_I F_n'(\varepsilon, x) g(\varepsilon, x) dx - \int_I F_n(\varepsilon, x) g'(\varepsilon, x) dx.$$

Write $H^\varepsilon(Z) = H(Z)$, $Q(\varepsilon) = Q$, and let $F(\varepsilon, x)$ be the cumulative distribution function of $Q(\varepsilon)$.

We then show that $H_n^\varepsilon(Z)$ converges uniformly to $H^\varepsilon(Z)$ and $H_n^\varepsilon(Z)'$ converges uniformly to some function; it follows that this function is $H^\varepsilon(Z)'$. This requires showing that the integrands in the second and third terms of the previous expression converge to well-defined functions.

We think of the $x_{n,i}(\varepsilon)$ as *locations* of point masses. So, we can think of $x_{n,i}(\varepsilon)'$ as an instantaneous location change.

1. **Second Term, Instantaneous Location Change (See Appendix C):** For $x \in \text{supp}(Q(\varepsilon))$ and any sequence of points $x_{n_1, i_1}(\varepsilon), x_{n_2, i_2}(\varepsilon), \dots$ approaching x ,

$$K_1(\varepsilon, x) = \lim_{j \rightarrow \infty} x_{n_j, i_j}'(\varepsilon)$$

is a well-defined continuous function.

2. **Third Term, Instantaneous Probability Change (See Appendix D):** Recall that $\text{supp}(Q(\varepsilon))$ is a Cantor set defined by a collection of “deleted” intervals: namely, $I^d \equiv (f_0(p_1), f_1(p_0))$, and all intervals of the form $f_{i_1} \circ f_{i_2} \circ \dots \circ f_{i_n}(I^d)$ (called deleted intervals on level n). For x belonging to a deleted interval on level n , define $K_2(\varepsilon, x) = F_n'(\varepsilon, x)$. Since the union of deleted intervals is dense in I , we can extend $K_2(\varepsilon, x)$ to a function on all $x \in I$, and we show that $K_2(\varepsilon, x)$ is a well-defined continuous function.

Using the boundedness of the instantaneous location change and probability change (established in Appendix A and Appendix B) and the Arzela–Ascoli theorem (note that Appendix C and Appendix D imply pointwise convergence of $H_n^\varepsilon(Z)'$ and Appendix A and Appendix B imply equicontinuity of $H_n^\varepsilon(Z)'$), we obtain uniform convergence of $H_n^\varepsilon(Z)'$ to $H^\varepsilon(Z)'$, which gives the result:

Theorem 3.7: In the nonoverlapping case

$$H^\varepsilon(Z)' = r(\varepsilon, p_0(\varepsilon))' + \int_{\text{supp}(Q(\varepsilon))} K_1(\varepsilon, x) g(\varepsilon, x) dF(\varepsilon, x) \\ - \int_I K_2(\varepsilon, x) g(\varepsilon, x) dx - \int_I F(\varepsilon, x) g'(\varepsilon, x) dx.$$

Note that the second term in this expression is a weighted mean of the instantaneous location change and the third term in this expression is a weighted mean of the instantaneous probability change.

Remark 3.8: Using the same technique, we can give a similar formula for the derivative of $H^\varepsilon(Z)$ with respect to π_{ij} 's when $\varepsilon > 0$. We can also give such formulas for higher derivatives in a similar way.

Remark 3.9: The techniques in this section can be applied to give an expression for the derivative of the entropy rate in the special overlapping case where $f_0(p_1) = f_1(p_0)$.

C. Derivatives in Other Cases

1. If any two of the π_{ij} 's are equal to 0, then

$$H^\varepsilon(Z) = -\varepsilon \log \varepsilon - (1 - \varepsilon) \log(1 - \varepsilon)$$

$H^\varepsilon(Z)$ is not differentiable with respect to ε at $\varepsilon = 0$.

2. Of more interest, it was shown in [12] that $H(Z)$ is not differentiable with respect to ε at $\varepsilon = 0$ when **exactly one of the π_{ij} 's are equal to 0**.

3. Consider the case that $\varepsilon = 0$ and all the π_{ij} 's are **positive**. As discussed in Example 4.1 of [4], the entropy rate is analytic as a function of ε and π_{ij} 's.

In [6] (and more generally in [16], [17]), an explicit formula was given for $H'(Z)$ at $\varepsilon = 0$ in this case. We briefly indicate how this is related to our results in Section III-B.

Instead of considering the dynamics of x_n on the real line, we consider those of (a_n, b_n) on the one-dimensional simplex

$$W = \{(w_1, w_2) : w_1 + w_2 = 1, w_i \geq 0\}.$$

Let Q denote the limiting distribution of (a_n, b_n) on W , the entropy $H(Z)$ can be computed as follows:

$$H(Z) = \int_W -(r_0(w) \log r_0(w) + r_1(w) \log r_1(w)) dQ$$

where

$$r_0(w) = ((1 - \varepsilon)\pi_{00} + \varepsilon\pi_{01})w_1 + ((1 - \varepsilon)\pi_{10} + \varepsilon\pi_{11})w_2 \\ r_1(w) = ((\varepsilon\pi_{00} + (1 - \varepsilon)\pi_{01})w_1 + (\varepsilon\pi_{10} + (1 - \varepsilon)\pi_{11})w_2.$$

In order to calculate the derivative, we split the region of integration into two disjoint parts $W = W^0 \cup W^1$ with

$$W^0 = \{t(0, 1) + (1 - t)(1/2, 1/2) : 0 \leq t \leq 1\} \\ W^1 = \{t(1/2, 1/2) + (1 - t)(1, 0) : 0 \leq t \leq 1\}.$$

Let $r(w) = -(r_0(w) \log r_0(w) + r_1(w) \log r_1(w))$ and $H^i(Z) = \int_{W^i} r(w) dQ$, then

$$H(Z) = H^0(Z) + H^1(Z).$$

For W^0 , we represent every point (w_1, w_2) using the coordinate w_1/w_2 . For W^1 , we represent every point (w_1, w_2) using the coordinate w_2/w_1 . Computation shows that $H_n^\varepsilon(Z)$ uniformly converge to $H^\varepsilon(Z)$ on $[0, 1/2]$. Note that expressions in Theorem 3.7 are not computable for $\varepsilon > 0$, however, we can apply similar uniform convergence ideas in each of these regions to recover the formula given in [6] for $\varepsilon = 0$.

4. **(Low signal-to-noise ratio (SNR) regime, $\varepsilon = 1/2$)** In Corollary 6 of [11], it was shown that in the symmetric case (i.e., $\pi_{01} = \pi_{10}$), the entropy rate approaches zero at rate $(1/2 - \varepsilon)^4$ as ε approaches $1/2$. It can be shown that the entropy rates at ε and $1 - \varepsilon$ are the same, and so all odd-order derivatives vanish at $\varepsilon = 1/2$. It follows that this result of [11] is equivalent to the statement that in the symmetric case $H''(Z)|_{\varepsilon=1/2} = 0$. We generalize this result to the nonsymmetric case as follows:

$$H''(Z)|_{\varepsilon=1/2} = -4 \left(\frac{\pi_{10} - \pi_{01}}{\pi_{10} + \pi_{01}} \right)^2.$$

For more details, see Appendix E.

APPENDIX A

PROOF OF BOUNDEDNESS OF INSTANTANEOUS LOCATION CHANGE

Claim: For any fix $0 < \eta < 1/2$, $x_{n,i}^{(k)}(\varepsilon) \leq C_1(k, \eta)$, $\eta \leq \varepsilon \leq 1/2$, C_1 is a positive constant only depending on k, η .

Proof: We only prove the case when $k = 1$. Consider the iteration

$$x_{n+1} = f_{z_{n+1}}(\varepsilon, x_n).$$

Taking the derivative with respect to ε , we obtain

$$x'_{n+1} = \frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, x_n) + \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, x_n)x'_n.$$

Note that $\frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, x_n)$ is uniformly bounded by a constant and $\frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, x_n)$ is bounded by ρ with $0 < \rho < 1$, we conclude x'_n is uniformly bounded too. \square

APPENDIX B

PROOF OF BOUNDEDNESS OF INSTANTANEOUS PROBABILITY CHANGE

Claim: For $x \notin \{x_{n,i}\}$ and $0 \leq \varepsilon \leq 1/2$, $F_n^{(k)}(\varepsilon, x) \leq C_2(k)$, where C_2 is a positive constant only depending on k .

Proof: We only prove the case when $k = 1$. For x with $x_{n,2i} < x < x_{n,2i+1}$, we have $F_n(\varepsilon, x) = F_{n-1}(\varepsilon, x)$, and consequently

$$\frac{\partial F_n(\varepsilon, x)}{\partial \varepsilon} = \frac{\partial F_{n-1}(\varepsilon, x)}{\partial \varepsilon}.$$

For x with $x_{n,2i-1} < x < x_{n,2i}$, $\frac{\partial F_n(\varepsilon, x)}{\partial \varepsilon} - \frac{\partial F_{n-1}(\varepsilon, x)}{\partial \varepsilon}$ is bounded by $C\rho_1^n$, here C is a positive constant and $0 < \rho_1 < 1$ (see proof that K_2 is well defined in Appendix D). Therefore, we conclude that the instantaneous probability change is uniformly bounded. \square

APPENDIX C

PROOF THAT K_1 IS WELL DEFINED

Proof: We need to prove that if two points x_{n_k, i_k} and x_{n_l, i_l} are close, then x'_{n_k, i_k} and x'_{n_l, i_l} are also close. Note that for nonoverlapping case, if x_{n_k, i_k} and x_{n_l, i_l} are very close, their corresponding symbolic sequences must share a long common tail. We shall prove that the asymptotical dynamics of x_n does not depend on the starting point as long as they have the same common long tail. Without loss of generality, we assume that z, \hat{z} have common tail z_1, z_2, \dots, z_n . In this case, the two dynamical systems start with different values x_0, \hat{x}_0 along the same path. Now the two iterations produce

$$\begin{aligned} x'_{n+1} &= \frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, x_n) + \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, x_n)x'_n \\ \hat{x}'_{n+1} &= \frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, \hat{x}_n) + \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, \hat{x}_n)\hat{x}'_n. \end{aligned}$$

Taking the difference, we have

$$\begin{aligned} x'_{n+1} - \hat{x}'_{n+1} &= \frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, x_n) - \frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, \hat{x}_n) \\ &\quad + \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, x_n)x'_n - \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, \hat{x}_n)\hat{x}'_n \\ &= \frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, x_n) - \frac{\partial f_{z_{n+1}}}{\partial \varepsilon}(\varepsilon, \hat{x}_n) \\ &\quad + \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, x_n)x'_n - \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, \hat{x}_n)\hat{x}'_n \\ &\quad + \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, \hat{x}_n)x'_n - \frac{\partial f_{z_{n+1}}}{\partial x}(\varepsilon, \hat{x}_n)\hat{x}'_n. \end{aligned}$$

Since

- when $n \rightarrow \infty$, x_n and \hat{x}_n are getting close uniformly with respect to ε – and –
- $\frac{\partial f_i}{\partial \varepsilon}(\varepsilon, \cdot)$ and $\frac{\partial f_i}{\partial x}(\varepsilon, \cdot)$ ($i = 0, 1$) are Lipschitz – and –
- $f_i(\varepsilon, \cdot)$ ($i = 0, 1$) are ρ -contraction mappings,

we conclude that x'_n and \hat{x}'_n are very close uniformly with respect to ε . The well definedness of K_1 then follows. \square

APPENDIX D

PROOF THAT K_2 IS WELL DEFINED

Proof: Every deleted interval corresponds to a finite sequence of binary digits and K_2 is well defined on these intervals. We order the deleted intervals on level n from left to right

$$I_{n,1}^d, I_{n,2}^d, \dots, I_{n,2^n-1}^d.$$

We need to prove that if two deleted intervals $I_{m,i}^d, I_{n,j}^d$ are close, then $F_m(\varepsilon, I_{m,i}^d)$ (which is defined as $F_m(\varepsilon, x)$ with $x \in I_{m,i}^d$) and $F_m(\varepsilon, I_{n,j}^d)$ are close. Assume $m \leq n$, then the points $x_{n,k}$'s in between $I_{m,i}^d$ and $I_{n,j}^d$ must have a long common tail. Suppose that the common tail is the path z_1, z_2, \dots, z_n , let q_i denote the sum of the probabilities associated with these points. Note that as long as the sequences have long common tail, the corresponding values of K_2 are getting closer and closer. For simplicity we only track one path for the time being. Then we have

$$\begin{aligned} a_{i+1} &= p_E(z_{i+1})(\pi_{00}a_i + \pi_{10}b_i) \\ b_{i+1} &= p_E(\bar{z}_{i+1})(\pi_{01}a_i + \pi_{11}b_i). \end{aligned}$$

It follows that

$$(a_{i+1} + b_{i+1}) \leq \rho(a_i + b_i);$$

here $0 < \rho < 1$ and ρ is defined as

$$\rho = \max\{(1 - \varepsilon)\pi_{00} + \varepsilon\pi_{01}, (1 - \varepsilon)\pi_{10} + \varepsilon\pi_{11}, \varepsilon\pi_{00} + (1 - \varepsilon)\pi_{01}, \varepsilon\pi_{10} + (1 - \varepsilon)\pi_{11}\}.$$

Immediately, we have

$$(a_n + b_n) \leq \rho^n.$$

Taking the derivative, we have

$$\begin{aligned} a'_{n+1} &= -(\pi_{00}a_n + \pi_{10}b_n) + (1 - \varepsilon)(\pi_{00}a'_n + \pi_{10}b'_n) \\ b'_{n+1} &= (\pi_{01}a_n + \pi_{11}b_n) + \varepsilon(\pi_{10}a'_n + \pi_{11}b'_n). \end{aligned}$$

In this case, we obtain

$$|a'_{n+1}| + |b'_{n+1}| \leq \rho(|a'_n| + |b'_n|) + \rho^n$$

which implies that there is a positive constant C and ρ_1 with $\rho < \rho_1 < 1$ such that

$$a'_n + b'_n \leq C\rho_1^n.$$

Then we conclude that $|a'_n + b'_n| \rightarrow 0$ as $n \rightarrow \infty$. Exactly the same derivation can be applied to multiple path; it follows that

$$q_n \leq \rho^n, \quad q'_n \leq C\rho_1^n.$$

So no matter which level we started from the deleted intervals, as long as they have long common tails, the corresponding values of K_2 function are close. Therefore, K_2 is well defined. \square

APPENDIX E
COMPUTATION OF $H''(Z)|_{\varepsilon=1/2}$

In this appendix, we basically follow the framework of [6]. Let

$$\mathbf{p}_n = [p(Z_1^n, E_n = 0), p(Z_1^n, E_n = 1)]$$

and

$$\mathbf{M}(Z_{n-1}, Z_n) = \begin{bmatrix} (1-\varepsilon)p_X(Z_n|Z_{n-1}) & \varepsilon p_X(\bar{Z}_n|Z_{n-1}) \\ (1-\varepsilon)p_X(Z_n|\bar{Z}_{n-1}) & \varepsilon p_X(\bar{Z}_n|\bar{Z}_{n-1}) \end{bmatrix}.$$

Then we have

$$\mathbf{p}_n = \mathbf{p}_{n-1} \mathbf{M}(Z_{n-1}, Z_n).$$

Immediately, we obtain

$$p_Z(Z_1^n) = \mathbf{p}_1 \mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n) \mathbf{1}.$$

We consider the case when the channel is operating on the low-SNR region. For convenience, we let

$$\delta = \frac{1}{2} - \varepsilon.$$

Thus, when the SNR is very low, namely, $\varepsilon \rightarrow \frac{1}{2}$, correspondingly we have $\delta \rightarrow 0$. Since $H(Z)$ is an even function at $\delta = 0$, the odd-order derivatives at $\delta = 0$ are all equal to 0. In the sequel, we shall compute the second derivative of $H(Z)$ at $\delta = 0$.

In this case, we can rewrite the random matrix $\mathbf{M}_i = \mathbf{M}(z_i z_{i+1})$ in the following way:

$$\mathbf{M}_i = \frac{1}{2} \begin{bmatrix} p_X(z_{i+1}|z_i) & p_X(\bar{z}_{i+1}|z_i) \\ p_X(z_{i+1}|\bar{z}_i) & p_X(\bar{z}_{i+1}|\bar{z}_i) \end{bmatrix} + \delta \begin{bmatrix} p_X(z_{i+1}|z_i) & -p_X(\bar{z}_{i+1}|z_i) \\ p_X(z_{i+1}|\bar{z}_i) & -p_X(\bar{z}_{i+1}|\bar{z}_i) \end{bmatrix}.$$

For the special case when $i = 0$, we have

$$\mathbf{M}_0 = \frac{1}{2} [p_X(z_1), p_X(\bar{z}_{i+1})] + \delta [p_X(z_1), -p_X(\bar{z}_1)].$$

Then

$$p_Z(z_1^n) = \left(\frac{1}{2}\mathbf{M}_0^{(0)} + \delta\mathbf{M}_0^{(1)}\right) \left(\frac{1}{2}\mathbf{M}_1^{(0)} + \delta\mathbf{M}_1^{(1)}\right) \cdots \left(\frac{1}{2}\mathbf{M}_{n-1}^{(0)} + \delta\mathbf{M}_{n-1}^{(1)}\right) \mathbf{1}.$$

Now define the function

$$\mathbf{R}_n(\delta) = \sum_{z_1^n} p_Z(z_1^n) \log(p_Z(z_1^n)).$$

Then according to the definition of $H(Z)$

$$H(Z) = - \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{R}_n(\delta).$$

It can be checked that

$$\frac{\partial \mathbf{R}_n(\delta)}{\partial \delta} = \sum_{z_1^n} \frac{\partial p_Z(z_1^n)}{\partial \delta} (\log p_Z(z_1^n) + 1).$$

Now

$$\begin{aligned} & \left. \frac{\partial p_Z(z_1^n)}{\partial \delta} \right|_{\delta=0} \\ &= \left(\frac{1}{2}\right)^{n-1} \sum_{i=0}^{n-1} \mathbf{M}_0^{(0)} \mathbf{M}_1^{(0)} \cdots \mathbf{M}_{i-1}^{(0)} \mathbf{M}_i^{(1)} \mathbf{M}_{i+1}^{(0)} \cdots \mathbf{M}_{n-1}^{(0)} \mathbf{1} \\ &= \left(\frac{1}{2}\right)^{n-1} \sum_{i=1}^n (p_X(z_i) - p_X(\bar{z}_i)). \end{aligned}$$

Again, simple calculations will lead to the first equation at the bottom of the page. Since $\left. \frac{\partial^2 p_Z(z_1^n)}{\partial \delta^2} \right|_{\delta=0}$ is defined in the second equation at the bottom of the page, we have

$$\left. \frac{\partial^2 \mathbf{R}_n(\delta)}{\partial \delta^2} \right|_{\delta=0} = \sum_{z_1^n} 2^n \left(\left(\frac{1}{2}\right)^{n-1} \sum_{i=1}^n (p_X(z_i) - p_X(\bar{z}_i)) \right)^2.$$

Let x, y temporarily denote the stationary distribution

$$p_X(0) = \frac{\pi_{10}}{\pi_{01} + \pi_{10}}, \quad p_X(1) = \frac{\pi_{01}}{\pi_{01} + \pi_{10}}$$

respectively. Then we get the equation at the top of the following page.

Using the following two combinatoric identities:

$$\sum_{i=0}^n i C_n^i = n 2^{n-1}$$

and

$$\sum_{i=0}^n i^2 C_n^i = n(n-1)2^{n-2} + n 2^{n-1},$$

we derive

$$\begin{aligned} \left. \frac{\partial^2 \mathbf{R}_n(\delta)}{\partial \delta^2} \right|_{\delta=0} &= \frac{1}{2^{n-2}} ((x-y)^2 (n(n-1)2^n + n 2^{n+1}) \\ &\quad + n^2 2^n (2y-1)^2 + 2(x-y)(2y-1)n 2^n) \\ &= 4n(x-y)^2. \end{aligned}$$

$$\frac{\partial^2 \mathbf{R}_n(\delta)}{\partial \delta^2} = \sum_{z_1^n} \left(\frac{\partial^2 p_Z(z_1^n)}{\partial \delta^2} \log p_Z(z_1^n) + \frac{1}{p_Z(z_1^n)} \left(\frac{\partial p_Z(z_1^n)}{\partial \delta} \right)^2 + \frac{\partial^2 p_Z(z_1^n)}{\partial \delta^2} \right).$$

$$\begin{aligned} \left. \frac{\partial^2 p_Z(z_1^n)}{\partial \delta^2} \right|_{\delta=0} &= \left(\frac{1}{2}\right)^{n-2} \sum_{i \neq j} \mathbf{M}_0^{(0)} \mathbf{M}_1^{(0)} \cdots \mathbf{M}_{i-1}^{(0)} \mathbf{M}_i^{(1)} \mathbf{M}_{i+1}^{(0)} \cdots \mathbf{M}_{j-1}^{(0)} \mathbf{M}_j^{(1)} \mathbf{M}_{j+1}^{(0)} \cdots \mathbf{M}_{n-1}^{(0)} \mathbf{1} \\ &= \left(\frac{1}{2}\right)^{n-2} [p_X(z_{i+1}), -p_X(\bar{z}_{i+1})] \begin{bmatrix} p_X(z_{j+1}|z_{i+1}) & -p_X(\bar{z}_{j+1}|z_{i+1}) \\ p_X(z_{j+1}|\bar{z}_{i+1}) & -p_X(\bar{z}_{j+1}|\bar{z}_{i+1}) \end{bmatrix} \\ &= \left(\frac{1}{2}\right)^{n-2} \sum_{i \neq j} (p_X(z_{j+1}, z_{i+1}) - p_X(z_{j+1}, \bar{z}_{i+1}) - p_X(\bar{z}_{j+1}, z_{i+1}) + p_X(\bar{z}_{j+1}, \bar{z}_{i+1})) \end{aligned}$$

$$\begin{aligned} \left. \frac{\partial^2 \mathbf{R}_n(\delta)}{\partial \delta^2} \right|_{\delta=0} &= \frac{1}{2^{n-2}} \sum_{i=0}^n C_n^i (2ix + 2(n-i)y - n)^2 \\ &= \frac{1}{2^{n-2}} \sum_{i=0}^n C_n^i ((2x-2y)i + 2ny - n)^2 \\ &= (2x-2y)^2 \sum_{i=0}^n C_n^i i^2 + (2ny-n)^2 \sum_{i=0}^n 1 + 2(2x-2y)(2ny-n) \sum_{i=0}^n C_n^i i. \end{aligned}$$

From the fact that the derivatives of $H(Z)$ with respect to ε are uniformly bounded on $[0, 1/2]$ (see [6], also implied by Theorem 1.1 of [4] and the computation of $H^\varepsilon(Z)|_{\varepsilon=0}$), we draw the conclusion that the second coefficient of $H(Z)$ is equal to

$$H''(Z)|_{\varepsilon=1/2} = -4 \left(\frac{\pi_{10} - \pi_{01}}{\pi_{10} + \pi_{01}} \right)^2.$$

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewer for pointing out the Faa Di Bruno formula, which greatly simplified the proof of Lemma 2.3.

REFERENCES

- [1] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Trans. 1st Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, Prague, Czechoslovakia, 1957, pp. 13–20.
- [2] G. Constantine and T. Savits, "A multivariate Faa Di Bruno formula with applications," *Trans. Amer. Math. Soc.*, vol. 348, no. 2, pp. 503–520, Feb. 1996.
- [3] R. Gharavi and V. Anantharam, "An upper bound for the largest Lyapunov exponent of a Markovian product of nonnegative matrices," *Theor. Comp. Sci.*, vol. 332, no. 1–3, pp. 543–557, Feb. 2005.
- [4] G. Han and B. Marcus, "Analyticity of entropy rate of hidden Markov chains," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5251–5266, Dec. 2006.
- [5] T. Holliday, A. Goldsmith, and P. Glynn, "Capacity of finite state channels based on Lyapunov exponents of random matrices," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3509–3532, Aug. 2006.
- [6] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the entropy of a hidden Markov process," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 2004, pp. 362–371.
- [7] J. Kemeny and J. Snell, *Finite Markov Chains*. Princeton, N.J.: Van Nostrand, 1960.
- [8] R. Leipnik and T. Reid, "Multivariable Faa Di Bruno formulas," in *Electronic Proc 9th Annu. Int. Conf. Technology in Collegiate Mathematics* [Online]. Available: <http://archives.math.utk.edu/ICTCM/EP-9.html#C23>
- [9] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*. Cambridge, U.K.: Cambridge Univ. Press, 1995.
- [10] B. Marcus, K. Petersen, and S. Williams, "Transmission rates and factors of Markov chains," *Contemp. Math.*, vol. 26, pp. 279–294, 1984.
- [11] E. Ordentlich and T. Weissman, "On the optimality of symbol by symbol filtering and denoising," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 19–40, Jan. 2006.
- [12] E. Ordentlich and T. Weissman, "New bounds on the entropy rate of hidden Markov process," in *Proc. Information Theory Workshop*, San Antonio, TX, Oct. 2004, pp. 117–122.
- [13] E. Ordentlich and T. Weissman, Personal Communication.
- [14] Y. Peres, "Analytic dependence of Lyapunov exponents on transition probabilities," in *Lyapunov's Exponents, Proceedings of a Workshop (Lecture Notes in Mathematics)*. Berlin, Germany: Springer-Verlag, 1990, vol. 1486.
- [15] Y. Peres, "Domains of analytic continuation for the top Lyapunov exponent," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 28, no. 1, pp. 131–148, 1992.

[16] O. Zuk, I. Kanter, and E. Domany, "Asymptotics of the entropy rate for a hidden Markov process," *J. Statist. Phys.*, vol. 121, no. 3–4, pp. 343–360, 2005.

[17] O. Zuk, E. Domany, I. Kanter, and M. Aizenman, "From finite-system entropy to entropy rate for a hidden Markov process," *IEEE Signal Process. Lett.*, vol. 13, no. 9, pp. 517–520, Sep. 2006.

The Fading Number of Memoryless Multiple-Input Multiple-Output Fading Channels

Stefan M. Moser, *Member, IEEE*

Abstract—In this correspondence, we derive the fading number of multiple-input multiple-output (MIMO) flat-fading channels of general (not necessarily Gaussian) regular law without temporal memory. The channel is assumed to be noncoherent, i.e., neither receiver nor transmitter have knowledge about the channel state, but they only know the probability law of the fading process. The fading number is the second term, after the double-logarithmic term, of the high signal-to-noise ratio (SNR) expansion of channel capacity. Hence, the asymptotic channel capacity of memoryless MIMO fading channels is derived exactly. The result is then specialized to the known cases of single-input–multiple-output (SIMO), multiple-input single-output (MISO), and single-input–single-output (SISO) fading channels, as well as to the situation of Gaussian fading.

Index Terms—Channel capacity, fading number, Gaussian fading, general flat fading, high signal-to-noise ratio (SNR), multiple antenna, multiple-input multiple-output (MIMO), noncoherent.

I. INTRODUCTION

It has been recently shown in [1], [2] that, whenever the matrix-valued fading process is of finite differential entropy rate (a so-called *regular* process), the capacity of noncoherent multiple-input multiple-output (MIMO) fading channels typically grows only double-logarithmically in the signal-to-noise ratio (SNR).

This is in stark contrast to both, the *coherent* fading channel where the receiver has *perfect* knowledge about the channel state, and to the noncoherent fading channel with *nonregular* channel law, i.e., the differential entropy rate of the fading process is not finite. In the former case the capacity grows logarithmically in the SNR with a

Manuscript received June 1, 2006; revised March 12, 2007. This work was supported by the Industrial Technology Research Institute (ITRI), Zhudong, Taiwan, under Contract G1-95003.

The author is with the Department of Communication Engineering, National Chiao Tung University (NCTU), Hsinchu, Taiwan (e-mail: stefan.moser@ieec.org).

Communicated by K. Kobayashi, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2007.899512