Title	Asymptotics of entropy rate of hidden Markov chains at weak Black Holes
Author(s)	Han, G; Marcus, B
Citation	The 2008 IEEE International Symposium on Information Theory, Toronto, ON., Canada, 6-11 July 2008. In Conference Proceedings, 2008, p. 2629-2633
Issued Date	2008
URL	http://hdl.handle.net/10722/124816
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Asymptotics of Entropy Rate of Hidden Markov Chains at Weak Black Holes

Guangyue Han*, Brian Marcus†
* University of Hong Kong, Email: ghan@maths.hku.hk
† University of British Columbia, Email: marcus@math.ubc.ca

For a more detailed version of this paper, please refer to http://hkumath.hku.hk/~ghan

***Abstract*— We generalize a result in [8] and derive an asymptotic formula for entropy rate of a hidden Markov chain around a "weak Black Hole". We also discuss applications of the asymptotic formula to certain channels.**

I. Introduction

Consider a finite-state stationary stochastic process $Y = Y_{-\infty}^{\infty}$. The entropy rate of Y is defined to be

$$H(Y) = \lim_{n\to\infty} H(Y_{-n}^{0})/(n+1);$$

here, H on finite length distributions is taken with the usual definition, with $\log$ taken to mean the natural logarithm.

If $Y = \{Y_{-\infty}^{\infty}\}$ is a Markov chain with alphabet $\{1, 2, \cdots, B\}$ and transition probability matrix Δ, it is well known that $H(Y)$ can be explicitly expressed with the stationary vector of Y and Δ. A function $Z = \{Z_{-\infty}^{\infty}\}$ 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 $\{1, 2, \cdots, B\}$, taking values in $\mathcal{A} := \{1, 2, \cdots, A\}$ (alternatively a hidden Markov chain is defined as a Markov chain observed in noise). For a hidden Markov chain Z, $H(Z)$ turns out (see Equation (1)) to be the integral of a certain function defined on a simplex with respect to a measure due to Blackwell [4]. However Blackwell's measure is somewhat complicated and the integral formula appears to be difficult to evaluate in most cases.

Recently, the problem of computing the entropy rate of a hidden Markov chain has drawn much interest, and many approaches have been adopted to tackle this problem. For instance, Blackwell's measure has been used to bound the entropy rate [14] and a variation on the Birch bound [3] was introduced in [5]. An efficient Monte Carlo method for computing the entropy rate of a hidden Markov chain was proposed independently by Arnold and Loeliger [1], Pfister et. al. [16], and Sharma and Singh [18]. The connection between the entropy rate of a hidden Markov chain and the top Lyapunov exponent of a random matrix product has been observed [10], [11], [12], [6]. In [7], it is shown that under mild positivity assumptions the entropy rate of a hidden Markov chain varies analytically as a function of the underlying Markov chain parameters.

Another recent approach is based on computing the coefficients of an asymptotic expansion of the entropy rate around certain values of the Markov and channel parameters. The first result along these lines was presented in [12], where for a binary symmetric channel with crossover probability ε (denoted by BSC(ε)), the Taylor expansion of $H(Z)$ around $\varepsilon = 0$ is studied for a binary hidden Markov chain of order one. In particular, the first derivative of $H(Z)$ at $\varepsilon = 0$ is expressed very compactly as a Kullback-Liebler divergence between two distributions on binary triplets, derived from the marginal of the input process X. Further improvements and new methods for the asymptotic expansion approach were obtained in [15], [19], [20] and [8]. In [15] the authors express the entropy rate for a binary hidden Markov chain where one of the transition probabilities is equal to zero as an asymptotic expansion including a $O(\varepsilon \log \varepsilon)$ term. The asymptotic expansion is further generalized in [9], [13].

This paper is organized as follows. In Section II we give an asymptotic formula for the entropy rate of a hidden Markov chain around a "Weak Black Hole". In Section III, we discuss applications of the formula to certain channels.

II. Asymptotic Formula for Entropy Rate

Let W be the simplex, comprising the vectors

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

and for $a \in \mathcal{A}$, let W_a be all $w \in W$ with $w_i = 0$ for $\Phi(i) \neq a$. For $a \in \mathcal{A}$, 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 $a \in \mathcal{A}$, 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

$$H(Z) = -\int \sum_a r_a(w) \log r_a(w) dQ(w), \qquad (1)$$

where Q is *Blackwell's measure* [4] on W. This measure, which satisfies an integral equation dependent on the parameters of the process, is however very hard to extract from the equation in any explicit way.

Definition 2.1: (see [8]) Suppose that for every $a \in \mathcal{A}$, Δ_a is a rank one matrix, and every column of Δ_a is either strictly positive or all zeros. We call this the *Black Hole* case.

It was shown [8] that $H(Z)$ is analytic around a Black Hole and the derivatives of $H(Z)$ can be exactly computed around a Black Hole. In this sequel, we consider weakened assumptions and prove an asymptotic formula for entropy rate of a hidden Markov chain around a "weak Black Hole", generalizing the corresponding result in [8].

Definition 2.2: Suppose that for every $a \in \mathcal{A}$, Δ_a is a rank one matrix. We call this the *weak Black Hole* case.

Remark 2.3: The weak Black Hole condition relaxes the Black Hole condition by eliminating the zero-positivity requirement. The differences and connections between these two conditions are illustrated briefly in simple examples (see section III-A).

In this paper we assume that Δ is affinely parameterized by ε ($\varepsilon \geq 0$), $\Delta(\varepsilon)$ is strictly positive when $\varepsilon > 0$, and $\varepsilon = 0$ corresponds to the weak Black Hole case. Namely the stochastic matrix Δ can be written as $\Delta^0 + \varepsilon\Delta^1$, where Δ^0 and Δ^1 are non-negative matrices independent of ε, Δ^0 corresponds to a weak Black Hole, and whenever some entry of Δ^0 is zero, the corresponding entry of Δ^1 is strictly positive. We use the standard notation: by $\alpha = \Theta(\beta)$, we mean there exist positive constants C_1, C_2 such that $C_1|\beta| \leq |\alpha| \leq C_2|\beta|$, while by $\alpha = O(\beta)$, we mean there exists a positive constant C such that $|\alpha| \leq C|\beta|$.

Proposition 2.4: For any fixed sequence $z_{-n}^{-1} \in \mathcal{A}^n$, $p(z_{-n}^{-1})$ is the quotient of two polynomials of ε. Moreover $p(z_{-n}^{-1})$ is analytic around $\varepsilon = 0$.

Proof:

When $\varepsilon > 0$, $\Delta(\varepsilon)$ is strictly positive. By Perron-Frobenius theory [17], $\Delta(\varepsilon)$ has a unique positive stationary vector, say $\pi(\varepsilon)$. Since

$$\mathrm{adj}\,(I - \Delta(\varepsilon))(I - \Delta(\varepsilon)) = \det(I - \Delta(\varepsilon))I = 0$$

(here $\mathrm{adj}\,(\cdot)$ denotes the adjugate operator on matrices), one can choose $\pi(\varepsilon)$ to be any normalized row vector of $\mathrm{adj}\,(I - \Delta(\varepsilon))$. So $\pi(\varepsilon)$ can be written as

$$\frac{(\pi_1(\varepsilon), \pi_2(\varepsilon), \cdots, \pi_B(\varepsilon))}{\pi_1(\varepsilon) + \pi_2(\varepsilon) + \cdots + \pi_B(\varepsilon)},$$

where $\pi_i(\varepsilon)$'s are polynomials of ε and the first non-zero term of every $\pi_i(\varepsilon)$ has a positive coefficient (we assume terms in the polynomails are increasingly ordered by the degree of ε unless otherwise specified). Let $\mathrm{mdeg}\,(\cdot)$ denote the degree of the first non-zero term of a polynomial, then we conclude that for each i

$$\mathrm{mdeg}\,(\pi_i(\varepsilon)) \geq \mathrm{mdeg}\,(\pi_1(\varepsilon) + \cdots + \pi_B(\varepsilon)),$$

and thus $\pi(\varepsilon)$, which is uniquely defined on $\varepsilon > 0$, can be continuously extended to $\varepsilon = 0$ via setting $\pi(0) = \lim_{\varepsilon \to 0} \pi(\varepsilon)$.

Now

$$p(z_{-n}^{-1}) = \pi(\varepsilon)\Delta_{z_{-n}} \cdots \Delta_{z_{-1}}\mathbf{1}$$

$$= \frac{(\pi_1(\varepsilon), \pi_2(\varepsilon), \cdots, \pi_B(\varepsilon))\Delta_{z_{-n}} \cdots \Delta_{z_{-1}}\mathbf{1}}{\pi_1(\varepsilon) + \pi_2(\varepsilon) + \cdots + \pi_B(\varepsilon)} =: \frac{f(\varepsilon)}{g(\varepsilon)}, \quad (2)$$

here $\mathrm{mdeg}\,(f(\varepsilon)) \geq \mathrm{mdeg}\,(g(\varepsilon))$. It then follows that $p(z_{-n}^{-1})$ is analytic around $\varepsilon = 0$. ■

Remark 2.5: It immediately follows from Proposition 2.4 that there exists $k \geq 0$ such that $p(z_{-n}^{-1}) = \Theta(\varepsilon^k)$.

Proposition 2.6: For any fixed sequence $z_{-n}^{0} \in \mathcal{A}^{n+1}$, $p(z_0|z_{-n}^{-1})$ is the quotient of two polynomials of ε. Moreover $p(z_0|z_{-n}^{-1})$ is analytic around $\varepsilon = 0$, furthermore either $p(z_0|z_{-n}^{-1}) = \Theta(1)$ or $p(z_0|z_{-n}^{-1}) = \Theta(\varepsilon)$.

Proof:

Let $x_{i,-n} = x_{i,-n}(z_{-n}^{i}) = p(y_i = \cdot\,|z_{-n}^{i})$, where $\cdot$ denotes the possible states of Markov chain Y. Then one checks that

$$p(z_0|z_{-n}^{-1}) = x_{-1,-n}\Delta_{z_0}\mathbf{1} \quad (3)$$

and

$$x_{i,-n} = \frac{x_{i-1,-n}\Delta_{z_i}}{x_{i-1,-n}\Delta_{z_i}, \mathbf{1}}, \qquad -n \leq i \leq -1. \quad (4)$$

Because Δ is affinely parameterized by ε ($\varepsilon \geq 0$) and $\Delta(\varepsilon)$ is strictly positive when $\varepsilon > 0$, inductively we can prove (the proof is similar to the proof of Proposition 2.4) that for any i, $x_{i,-n}$ can be written as follows:

$$x_{i,-n} = \frac{(f_1(\varepsilon), f_2(\varepsilon), \cdots, f_B(\varepsilon))}{f_1(\varepsilon) + f_2(\varepsilon) + \cdots + f_B(\varepsilon)},$$

where $f_i(\varepsilon)$'s are certain polynomials of ε such that for each i

$$\mathrm{mdeg}\,(f_i(\varepsilon)) \geq \mathrm{mdeg}\,(f_1(\varepsilon) + f_2(\varepsilon) + \cdots + f_B(\varepsilon)).$$

The existence of the Taylor series expansion of $x_{i,-n}$ around $\varepsilon = 0$ (for any i) then follows. Together with (3), we conclude that $p(z_0|z_{-n}^{-1})$ is analytic around $\varepsilon = 0$.

Now consider the Taylor series expansion of $x_{-1,-n}$ around $\varepsilon = 0$,

$$x_{-1,-n} = a_0(z_{-n}^{-1}) + a_1(z_{-n}^{-1})\varepsilon + a_2(z_{-n}^{-1})\varepsilon^2 + \cdots.$$

Since for any ε, $x_{-1,-n} \in W$, we conclude that $a_0(z_{-n}^{-1}) \geq \mathbf{0}$, but $a_0(z_{-n}^{-1}) \not\equiv \mathbf{0}$.

In the following we write for each i

$$\Delta_{z_i} = \Delta_{z_i}^0 + \varepsilon\Delta_{z_i}^1,$$

here $\Delta_{z_i}^0$ and $\Delta_{z_i}^1$ are non-negative matrices independent of ε. If $a_0(z_{-n}^{-1})\Delta_{z_0}^0\mathbf{1} > 0$, then

$$p(z_0|z_{-n}^{-1}) = x_{-1,-n}\Delta_{z_0}\mathbf{1} = a_0(z_{-n}^{-1})\Delta_{z_0}^0\mathbf{1} + O(\varepsilon) = \Theta(1).$$

Now consider the case when $a_0(z_{-n}^{-1})\Delta_{z_0}^0\mathbf{1} = 0$. Because $\Delta_{z_0}^0 + \varepsilon\Delta_{z_0}^1$ is strictly positive for $\varepsilon > 0$,

$$a_0(z_{-n}^{-1})(\Delta_{z_0}^0 + \varepsilon\Delta_{z_0}^1)\mathbf{1} > 0,$$

which implies $a_0(z_{-n}^{-1})\Delta_{z_0}^1\mathbf{1} > 0$. On the other hand, for $\varepsilon > 0$

$$x_{-1,-n}\Delta_{z_0}^0\mathbf{1} = (a_0(z_{-n}^{-1}) + a_1(z_{-n}^{-1})\varepsilon + \cdots)\Delta_{z_0}^0\mathbf{1} > 0,$$

which implies $a_1(z_{-n}^{-1})\Delta_{z_0}^0\mathbf{1} > 0$. So in this case,

$$p(z_0|z_{-n}^{-1}) = (a_0(z_{-n}^{-1})\Delta_{z_0}^1\mathbf{1} + a_1(z_{-n}^{-1})\Delta_{z_0}^0\mathbf{1})\varepsilon + O(\varepsilon^2).$$

It then follows that $p(z_0|z_{-n}^{-1}) = \Theta(\varepsilon)$, since the coefficient of ε has been proven to be strictly positive. ■

Lemma 2.7: Consider two formal series expansion $f(x), g(x) \in \mathbb{R}[[x]]$ such that $f(x) = \sum_{i=0}^{\infty} f_i x^i$ and $g(x) = \sum_{i=0}^{\infty} g_i x^i$, where $g_0 \neq 0$. Let $h(x) \in \mathbb{R}[[x]]$ be the quotient of $f(x)$ and $g(x)$ with $h(x) = \sum_{i=0}^{\infty} h_i x^i$. Then h_i is a function only dependent on $f_0, \cdots, f_i$ and $g_0, \cdots, g_i$.

Proof:

Comparing the coefficients of all the terms in the following identity:

$$\left(\sum_{i=0}^{\infty} h_i x^i\right)\left(\sum_{i=0}^{\infty} g_i x^i\right) = \sum_{i=0}^{\infty} f_i x^i,$$

we obtain that for any i,

$$h_0 g_i + h_1 g_{i-1} + \cdots + h_i g_0 = f_i.$$

The lemma then follows from an induction (on i) argument. ∎

Let i be a fixed non-positive integer and let s be a function defined on all hidden Markov strings z_{-m}^i for $-m \leq i$. We say that s *stabilizes* at a particular string z_{-n}^i if for all $m \geq n$ and hidden Markov strings $\hat{z}_{-m}^i$ such that $\hat{z}_{-n}^i = z_{-n}^i$, we have

$$s(\hat{z}_{-m}^i) = s(z_{-n}^i).$$

By Proposition 2.6, for any hidden Markov string z_{-m}^0, the Taylor series expansion of $p(z_0|z_{-m}^{-1})$ around $\varepsilon = 0$ exists. We use $b_j(z_{-m}^0)$ to represent the coefficient of ε^j in the expansion, namely

$$p(z_0|z_{-m}^{-1}) = b_0(z_{-m}^0) + b_1(z_{-m}^0)\varepsilon + b_2(z_{-m}^0)\varepsilon^2 + \cdots. \quad (5)$$

Lemma 2.8: Consider a hidden Markov chain Z at a weak Black Hole corresponding to $\varepsilon = 0$. Suppose that for a fixed sequence z_{-n}^0, $p(z_{-n}^{-1}) = \Theta(\varepsilon^k)$ for some $k \geq 0$. Then the first $n - 2k$ coefficients of $p(z_0|z_{-m}^{-1})$ are stabilized at z_{-n}^0, namely, for j with $0 \leq j \leq n-2k-1$, the coefficient $b_j(z_{-m}^0)$ is stabilized at z_{-n}^0.

Proof: Recall that $x_{i,-m} = x_{i,-m}(z_{-m}^i) = p(y_i = \cdot|z_{-m}^i)$, where $\cdot$ denotes the possible states of Markov chain Y. Consider the Taylor series expansion of $x_{i,-n}$ around $\varepsilon = 0$,

$$x_{i,-m} = a_0(z_{-m}^i) + a_1(z_{-m}^i)\varepsilon + a_2(z_{-m}^i)\varepsilon^2 + \cdots \quad (6)$$

Assume that $p(z_{-n}^i) = \Theta(\varepsilon^{k_i})$ $(0 \leq k_i \leq k)$, we shall show that $x_{i,-m}$ has the first $n+i+1-2k_i$ coefficients stabilized at z_{-n}^i, namely for j with $0 \leq j \leq n+i-2k_i$, $a_j(z_{-m}^i)$ is stabilized at z_{-n}^i.

We proceed by induction on i (from $-n$ to -1). The case when $i = -n$ is straightforward. Now we consider $i \geq -n$ and we suppose that $a_j(z_{-m}^i)$ $(0 \leq j \leq n+i-2k_i)$ is stabilized at z_{-n}^i. Recall that for each i

$$\Delta_{z_i} = \Delta_{z_i}^0 + \varepsilon\Delta_{z_i}^1,$$

where $\Delta_{z_i}^0$ and $\Delta_{z_i}^1$ are non-negative matrices independent of ε. Note that with this notation, we have

$$x_{i+1,-m} = \frac{x_{i,-m}\Delta_{z_{i+1}}}{x_{i,-m}\Delta_{z_{i+1}}\mathbf{1}}, \quad (7)$$

$$= \frac{(a_0(z_{-m}^i) + a_1(z_{-m}^i)\varepsilon + \cdots)(\Delta_{z_{i+1}}^0 + \varepsilon\Delta_{z_{i+1}}^1)}{(a_0(z_{-m}^i) + a_1(z_{-m}^i)\varepsilon + \cdots)(\Delta_{z_{i+1}}^0 + \varepsilon\Delta_{z_{i+1}}^1)\mathbf{1}}. \quad (8)$$

It follows from Proposition 2.6 that either $k_{i+1} = k_i + 1$ or $k_{i+1} = k_i$. If $k_{i+1} = k_i + 1$, necessarily we have for $m \geq n$

$$a_0(z_{-m}^i)\Delta_{z_{i+1}}^0 = a_0(z_{-n}^i)\Delta_{z_{i+1}}^0 = \mathbf{0}.$$

Applying Lemma 2.7 to expression (8), we conclude that for all j, $a_j(z_{-m}^{i+1})$ depends only on

$$a_1(z_{-m}^i), a_2(z_{-m}^i), \cdots, a_{j+1}(z_{-m}^i), \Delta_{z_{i+1}}^0, \Delta_{z_{i+1}}^1.$$

Thus for j with $0 \leq j \leq n+i-1-2k_i$, $a_j(z_{-m}^{i+1})$ is stabilized at z_{-n}^{i+1}. This implies that the first $n+i-2k_i = n+(i+1)+1-2k_{i+1}$ coefficients of $x_{i+1,-m}$ are stabilized at z_{-n}^{i+1}, namely for j with $0 \leq j \leq n+(i+1)-2k_{i+1}$, the coefficient $a_j(z_{-m}^{i+1})$ is stabilized at z_{-n}^{i+1}.

If $k_{i+1} = k_i$, necessarily we have

$$a_0(z_{-n}^i)\Delta_{z_{i+1}}^0\mathbf{1} \neq 0.$$

Again by Lemma 2.7 applied to expression (8), for any j, $a_j(z_{-m}^{i+1})$ depends only on

$$a_0(z_{-m}^i), a_1(z_{-m}^i), \cdots, a_j(z_{-m}^i), \Delta_{z_{i+1}}^0, \Delta_{z_{i+1}}^1.$$

Thus, for any j with $0 \leq j \leq n+i-2k_{i+1} = n+i-2k_i$, $a_j(z_{-m}^{i+1})$ stabilizes at z_{-n}^{i+1}. Now, let $j = n+i+1-2k_{i+1}$, then we have

$$a_j(z_{-m}^{i+1}) = \frac{*}{(a_0(z_{-m}^i)\Delta_{z_{i+1}}^0\mathbf{1})^2} + \text{ other terms },$$

here

$$* = a_{n+i-2k_i}(z_{-m}^i)\Delta_{z_{i+1}}^0 a_0(z_{-m}^i)\Delta_{z_{i+1}}^0\mathbf{1}$$
$$-a_0(z_{-m}^i)\Delta_{z_{i+1}}^0 a_{n+i-2k_i}(z_{-m}^i)\Delta_{z_{i+1}}^0\mathbf{1}.$$

Since $\Delta_{z_{i+1}}^0$ is a rank one matrix, we find $* = \mathbf{0}$ and "other terms" are functions of

$$a_0(z_{-m}^i), a_1(z_{-m}^i), \cdots, a_{n+i-1-2k_i}(z_{-m}^i), \Delta_{z_{i+1}}^0, \Delta_{z_{i+1}}^1.$$

So we conclude $a_j(z_{-m}^{i+1})$ is also stabilized at z_{-n}^{i+1}, and thus $x_{i+1,-m}$ has the first $n+i+2-2k_i = n+(i+1)+1-2k_{i+1}$ coefficients stabilized at z_{-n}^{i+1}.

The lemma then immediately follows from (3) and the proven fact that $x_{-1,-m}$ has the first $n-2k_{-1} = n-2k$ coefficients stabilized at z_{-n}^{-1}. ∎

Remark 2.9: Using the same iterative expression (7) and a completely parallel argument as in Lemma 2.8, one can show that the first $n-2k$ coefficients of $p(z_0|z_{-m}^{-1}y_{-m-1})$ are independent of y_{-m-1} and stabilized at z_{-n}^0 as well.

Consider expression (5). In the following, we use $p^{<l>}(z_0|z_{-n}^{-1})$ to denote the truncated (up to the $(l+1)$-st term) Taylor series expansion of $p(z_0|z_{-n}^{-1})$, i.e.,

$$p^{<l>}(z_0|z_{-n}^{-1}) = b_0(z_{-n}^0)+b_1(z_{-n}^0)\varepsilon+b_2(z_{-n}^0)\varepsilon^2+\cdots+b_l(z_{-n}^0)\varepsilon^l.$$

Theorem 2.10: For a hidden Markov chain Z around a weak Black Hole corresponding to $\varepsilon = 0$, we have for any $k \geq 0$,

$$H(Z) = H(Z)|_{\varepsilon=0} + \sum_{j=1}^{k} f_j\varepsilon^j + \sum_{j=1}^{k+1} g_j\varepsilon^j\log\varepsilon + O(\varepsilon^{k+1}), \quad (9)$$

where f_j's and g_j's are functions only dependent on Δ, the transition probability matrix of the underlying Markov chain Y.

Proof:

First we fix a large n and consider the Birch upper bound on $H(Z)$

$$H_n(Z) = H(Z_0|Z_{-n}^{-1}) = -\sum_{z_{-n}^0} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}).$$

Note that for $j \geq k+1$,

$$\left| \sum_{p(z_{-n}^{-1})=O(\varepsilon^j),} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) \right| = O(\varepsilon^{k+1}).$$

So, in the following we only consider the sequences z_{-n}^0 with $p(z_{-n}^{-1}) = \Theta(\varepsilon^j)$, $j \leq k$. For such sequences, by Lemma 2.8, as long as n is large enough, $p(z_0|z_{-m}^{-1})$ will have sufficiently many initial coefficients stabilized at z_{-n}^0. Moreover, for any sequence z_{-n}^0 with $n \geq 3k+1$, $p(z_{-n}^{-1}) = \Theta(\varepsilon^j)$, $p(z_0|z_{-n}^{-1}) = \Theta(1)$, we have

$$|\log p(z_0|z_{-n}^{-1}) - \log p^{<k>}(z_0|z_{-n}^{-1})| = O(\varepsilon^{k+1}); \qquad (10)$$

for any sequence z_{-n}^0 with $n \geq 3k+2$, $p(z_{-n}^{-1}) = \Theta(\varepsilon^j)$, $p(z_0|z_{-n}^{-1}) = \Theta(\varepsilon)$, we have

$$|\log p(z_0|z_{-n}^{-1}) - \log p^{<k+1>}(z_0|z_{-n}^{-1})| = O(\varepsilon^{k+1}). \qquad (11)$$

Now fix $n_0 \geq 3k+2$. For any $n \geq n_0$, using (10) and (11), we can show

$$H_n(Z) = -\sum_{p(z_{-n}^{-1})=\Theta(\varepsilon^j), j\leq k} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) + O(\varepsilon^{k+1})$$

$$= H(Z)|_{\varepsilon=0} + \sum_{j=1}^{k} f_j \varepsilon^j + \sum_{j=1}^{k+1} g_j \varepsilon^j \log \varepsilon + O(\varepsilon^{k+1}), \quad (12)$$

where f_j and g_j are functions dependent only on Y.

Using Remark 2.9, one can apply similar analysis to the Birch lower bound

$$\tilde{H}_n(Z) = H(Z_0|Z_{-n}^{-1} Y_{-n-1}).$$

For the same n_0, one can show that $\tilde{H}_n(Z)$ takes the same form as $H_n(Z)$ as in (12) with exactly the same coefficients of ε^j for $j \leq k$ and of $\varepsilon^j \log \varepsilon$ for $j \leq k+1$ when n is sufficiently large. We thus prove the theorem. ∎

Remark 2.11: Note that at a Black Hole, $g_j = 0$ for all j's. So Formula (9) is consistent with the Taylor series expansion of $H(Z)$ around a Black Hole.

III. Applications to Certain Channels

A. Binary Markov Chains Corrupted by BSC(ε)

Consider a binary symmetric channel with crossover probability ε. At time n the channel can be characterized by the following equation

$$Z_n = X_n \oplus E_n,$$

where X_n denotes the input process, $\oplus$ denotes binary addition, E_n denotes the i.i.d. binary noise with $p_E(0) = 1 - \varepsilon$ and $p_E(1) = \varepsilon$, and Z_n denotes the corrupted output.

Now, suppose $X = \{X_n\}$ is a first order Markov chain with the transition probability matrix

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

Then $Y = \{Y_n\} = \{(X_n, E_n)\}$ is jointly Markov with transition probability matrix:

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

and $Z = \Phi(Y)$ is a hidden Markov chain with $\Phi(0,0) = \Phi(1,1) = 0$, $\Phi(0,1) = \Phi(1,0) = 1$. When $\varepsilon = 0$, one checks that both Δ_0 and Δ_1 have rank one. If π_{ij}'s are all positive, then we have a Black Hole case, for which one can derive the Taylor series expansion of $H(Z)$ around $\varepsilon = 0$ [19], [8]; if some π_{ij}'s are zeros, then this is a weak Black hole case, for which Theorem 2.10 can be applied and an asymptotic formula for $H(Z)$ can be derived.

For instance, consider a first order Markov chain X with the following transition probability matrix

$$\begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix},$$

where $0 \leq p \leq 1$. It has been shown [15] that

$$H(Z) = H(X) - \frac{p(2-p)}{1+p} \varepsilon \log \varepsilon + O(\varepsilon)$$

as $\varepsilon \to 0$. This result has been further generalized [9], [13] to the following formula:

$$H(Z) = H(X) + f(X)\varepsilon \log(1/\varepsilon) + g(X)\varepsilon + O(\varepsilon^2 \log \varepsilon),$$

where X is the input Markov chain of any order, Z is the output process obtained by passing X through a BSC(ε), and $f(X)$ and $g(X)$ can be explicity computed.

Theorem 2.10 claims that higher order asymptotic terms together with their coefficients can be derived as well. Now suppose the input is an m-th order Markov chain X defined by the transition probabilities $P(X_t = a_0|X_{t-m}^{t-1} = a_{-m}^{-1})$, $a_{-m}^0 \in \mathcal{X}^m$, where $\mathcal{X} = \{0, 1\}$. For $i = 0, 1, \cdots$, let

$$\hat{X}_i = X_{-(i+1)m+1}^{-im}, \hat{E}_i = E_{-(i+1)m+1}^{-im}, \hat{Z}_i = Z_{-(i+1)m+1}^{-im}.$$

Then $\hat{X}$ is a first order Markov chain with state space $\mathcal{X}^m$, whose transition probability from state $\hat{x} \in \mathcal{X}^m$ to $\hat{y} \in \mathcal{X}^m$ (denoted by $P_{\hat{x}\hat{y}}$) can be easily computed. Obviously $(\hat{X}, \hat{E})$ is jointly Markov with state space $\mathcal{X}^m \times \mathcal{X}^m$, and the transition probability from state $(\hat{x}, \hat{d})$ to state $(\hat{y}, \hat{e})$ is

$$\hat{\Delta}_{(\hat{x},\hat{d})(\hat{y},\hat{e})} = P_{\hat{x}\hat{y}} \prod_{i=1}^{m} \delta(\hat{e}_i, \varepsilon),$$

where $\hat{\Delta}$ denotes the transition probability matrix of $(\hat{X}, \hat{E})$ and

$$\delta(0, \varepsilon) = 1 - \varepsilon, \qquad \delta(1, \varepsilon) = \varepsilon.$$

It is easy to check that when $\varepsilon = 0$, $\hat{\Delta}_{\hat{z}}$ is a rank one matrix for every $\hat{z} \in \mathcal{X}^m$. In particular, when $P(X_t = a_0 | X_{t-m}^{t-1} = a_{-m}^{-1}) > 0$ for $a_{-m}^0 \in \mathcal{X}^{m+1}$, we have a Black Hole. In this case, the derivatives of the entropy rate $H(\hat{Z})$ at $\varepsilon = 0$ can be exactly computed. Then together with

$$H(Z) = H(\hat{Z})/m,$$

we conclude that the derivatives of the entropy rate $H(Z)$ at $\varepsilon = 0$ can be exactly computed as well. When $P(X_t = a_0 | X_{t-m}^{t-1} = a_{-m}^{-1}) = 0$ for some $a_{-m}^0 \in \mathcal{X}^{m+1}$, we have a weak Black Hole. In this case, Theorem 2.10 can be applied and an asymptotical formula for $H(Z)$ around $\varepsilon = 0$ can be obtained.

B. Binary Markov Chains Corrupted by BEC(δ)

Consider a binary erasure channel with fixed erasure rate δ (denoted by BEC(δ)). We say the channel is in state 1 if the input digit is erased after passing through the channel (e will be used to denoted a erasure), otherwise we say the channel is in state 0. Let C denote the channel state, and let X denote the first order input Markov chain with transition probability matrix

$$\Pi = \left[\begin{array}{cc} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{array} \right],$$

and let Z denote the output process. Then $Y = (X, C)$ is jointly Markov with

$$\Delta = \left[\begin{array}{cccc} \pi_{00}(1-\delta) & \pi_{00}\delta & \pi_{01}(1-\delta) & \pi_{01}\delta \\ \pi_{00}(1-\delta) & \pi_{00}\delta & \pi_{01}(1-\delta) & \pi_{01}\delta \\ \pi_{10}(1-\delta) & \pi_{10}\delta & \pi_{11}(1-\delta) & \pi_{11}\delta \\ \pi_{10}(1-\delta) & \pi_{10}\delta & \pi_{11}(1-\delta) & \pi_{11}\delta \end{array} \right],$$

and $Z = \Phi(Y)$ is hidden Markov with $\Phi(0,1) = \Phi(1,1) = e$, $\Phi(0,0) = 0$ and $\Phi(1,0) = 1$.

Now one checks that if Π is a rank one matrix, $\Delta_0, \Delta_1, \Delta_e$ will all be rank one matrices, then we have a Black Hole case. In particular, if we use the following parameterization: $\pi_{00} = k/(k+1)+\varepsilon$, $\pi_{10} = k/(k+1)-\varepsilon$, $\pi_{01} = 1/(k+1)-\varepsilon$, $\pi_{11} = 1/(k+1)+\varepsilon$, where $k \geq 0$, then we have a Black Hole case corresponding to $\varepsilon = 0$.

C. Binary Markov Chains Corrupted by Gilbert-Elliot Channel

Consider a binary Gilbert-Elliot channel. The channel is said to be in state 0 if it behaves like a BSC(ε_0) and state 1 if it behaves like a BSC(ε_1). The channel state varies as a Markov chain with transition probability matrix:

$$C = \left[\begin{array}{cc} c_{00} & c_{01} \\ c_{10} & c_{11} \end{array} \right].$$

Let X denote the first order input Markov chain with transition probability matrix

$$\Pi = \left[\begin{array}{cc} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{array} \right],$$

and let Z denote the output process. Then $Y = (X, C, E)$ is jointly Markov with a 8×8 transition probability matrix Δ and $Z = \Phi(X, C, E)$ is hidden Markov with

$$\Phi(0,0,0) = \Phi(0,1,0) = \Phi(1,0,1) = \Phi(1,1,1) = 0,$$

$$\Phi(0,0,1) = \Phi(0,1,1) = \Phi(1,1,0) = \Phi(1,1,0) = 1.$$

If $\varepsilon_0 = \varepsilon_1 = 0$ and C is a rank one matrix, both Δ_0 and Δ_1 will be rank one matrices and we have a weak Black Hole. In particular, if we use the following parameterization: $(\varepsilon_0, \varepsilon_1) = (k_1\varepsilon, \varepsilon)$ and $c_{00} = k_2/(k_2+1)+\varepsilon$, $c_{10} = k_2/(k_2+1)-\varepsilon$, $c_{01} = 1/(k_2+1)-\varepsilon$, $c_{11} = 1/(k_2+1)+\varepsilon$, where $k_1, k_2 \geq 0$, then we have a Black Hole case corresponding to $\varepsilon = 0$.

References

[1] D. Arnold and H. Loeliger. The information rate of binary-input channels with memory. *Proc. 2001 IEEE Int. Conf. on Communications*, (Helsinki, Finland), pp. 2692–2695, June 11-14 2001.

[2] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavcic, W. Zeng, "Simulation-Based Computation of Information Rates for Channels With Memory," *IEEE Trans. Information Theory*, **52**, 3498–3508, 2006.

[3] J. Birch. Approximations for the entropy for functions of Markov chains. *Ann. Math. Statist.*, 33:930–938, 1962.

[4] D. Blackwell. The entropy of functions of finite-state Markov chains. *Trans. First Prague Conf. Information Thoery, Statistical Decision Functions, Random Processes*, pages 13–20, 1957.

[5] S. Egner, V. Balakirsky, L. Tolhuizen, S. Baggen and H. Hollmann. On the entropy rate of a hidden Markov model. *Proceedings of the 2004 IEEE International Symposium on Information Theory*, page 12, June 27-July 2, Chicago, U.S.A., 2004.

[6] R. Gharavi and V. Anantharam. An upper bound for the largest Lyapunov exponent of a Markovian product of nonnegative matrices. *Theoretical Computer Science*, Vol. 332, Nos. 1-3, pp. 543 -557, February 2005.

[7] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Transactions on Information Theory*, Volume 52, Issue 12, December, 2006, pages: 5251-5266.

[8] G. Han and B. Marcus. Derivatives of Entropy Rate in Special Families of Hidden Markov Chains. *IEEE Transactions on Information Theory*, Volume 53, Issue 7, July 2007, Page(s):2642 - 2652.

[9] G. Han and B. Marcus. Asymptotics of noisy constrained capacity. *Proc. ISIT 2007*, Nice, June 24-June 29, 2007, Pages:991-995. The journal version of this paper with slightly different title was submitted to Annals of Applied Probability.

[10] T. Holliday, A. Goldsmith and P. Glynn. Capacity of finite state markov channels with general inputs. *Proceedings of the 2003 IEEE International Symposium on Information Theory*, 29 June-4 July 2003, Page(s):289 - 289.

[11] T. Holliday, A. Goldsmith, and P. Glynn. Capacity of Finite State Channels Based on Lyapunov Exponents of Random Matrices. *IEEE Transactions on Information Theory*, Volume 52, Issue 8, Aug. 2006, Page(s):3509 - 3532.

[12] P. Jacquet, G. Seroussi, and W. Szpankowski. On the Entropy of a Hidden Markov Process (extended abstract). *Data Compression Conference*, 362–371, Snowbird, 2004.

[13] P. Jacquet, G. Seroussi, and W. Szpankowski. Noisy Constrained Capacity. *International Symposium on Information Theory*, 986-990, Nice, 2007.

[14] E. Ordentlich and T. Weissman. On the optimality of symbol by symbol filtering and denoising. *Information Theory, IEEE Transactions*, Volume 52, Issue 1, Jan. 2006 Page(s):19 - 40.

[15] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov process. *IEEE Information Theory Workshop*, San Antonio, Texas, 24-29 Oct. 2004, Page(s):117 - 122.

[16] H. Pfister, J. Soriaga and P. Siegel. The achievable information rates of finite-state ISI channels. *Proc. IEEE GLOBECOM*, (San Antonio, TX), pp. 2992–2996, Nov. 2001.

[17] E. Seneta. *Springer Series in Statistics*. Non-negative Matrices and Markov Chains. Springer-Verlag, New York Heidelberg Berlin, 1980.

[18] V. Sharma and S. Singh. Entropy and channel capacity in the regenerative setup with applications to Markov channels. *Proc. IEEE Intern. Symp. on Inform. Theory*, (Washington, D.C.), p. 283, June 24-29 2001.

[19] O. Zuk, I. Kanter and E. Domany. The entropy of a binary hidden Markov process. *J. Stat. Phys.*, 121(3-4): 343-360 (2005)

[20] O. Zuk, E. Domany, I. Kanter, and M. Aizenman. From Finite-System Entropy to Entropy Rate for a Hidden Markov Process. *Signal Processing Letters, IEEE*, Volume 13, Issue 9, Sept. 2006 Page(s):517 - 520.