Asymptotics of Entropy Rate in Special Families of Hidden Markov Chains

Guangyue Han
University of Hong Kong
email: ghan@maths.hku.hk

Brian Marcus
University of British Columbia
email: marcus@math.ubc.ca

October 16, 2009

Abstract

We generalize a result in [8] and derive an asymptotic formula for entropy rate of a hidden Markov chain under certain parameterization. We also discuss applications of the asymptotic formula to the asymptotic behaviors of entropy rate of hidden Markov chains as outputs of certain channels.

Index Terms–entropy, entropy rate, hidden Markov chain, hidden Markov model, hidden Markov process

1 Introduction

Consider a discrete finite-valued stationary stochastic process $Y = \{Y_n, n \in \mathbb{Z}\}$. The entropy rate of Y is defined to be

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

here, $H(Y_{-n}^0)$ denotes the joint entropy of $Y_{-n}^0 := \{Y_{-n}, Y_{-n+1}, \cdots, Y_0\}$, and log is taken to mean the natural logarithm.

If Y is a Markov chain with alphabet $\mathcal{B} := \{1, 2, \cdots, B\}$ and transition probability matrix Δ, it is well known that $H(Y)$ can be explicitly expressed with the stationary vector of Y and Δ. Let $\mathcal{A}$ denote a finite alphabet $\{1, 2, \cdots, A\}$, and let Φ denote a function defined on alphabet $\mathcal{B}$, taking values in $\mathcal{A}$, then the stochastic process defined by $Z := \Phi(Y) = \{\Phi(Y_n), n \in \mathbb{Z}\}$ is called a *hidden Markov chain*; alternatively a hidden Markov chain can be defined as a Markov chain observed in noise. For a hidden Markov chain Z, $H(Z)$ turns out 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. In general it is very difficult to compute $H(Z)$; so far there is no simple and explicit formula for $H(Z)$.

Recently, the problem of computing entropy rate of a hidden Markov chain Z 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 [15] 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. [17], and Sharma and Singh [19]. 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 [16], [20], [21] and [8]. In [16] 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.

This paper is organized as follows. In Section 2 we give an asymptotic formula (Theorem 2.8) for the entropy rate of a hidden Markov chain around a weak Black Hole. The coefficients in the formula can be computed in principle (although explicit computations may be quite complicated in general). The formula can be viewed as a generalization of the Black Hole condition considered in [8]. The *weak* Black Hole case is important for hidden Markov chains obtained as output processes of noisy channels, corresponding to input processes, *for which certain sequences have probability zero*. Examples are given in Section 3. Example 3.1 was already treated in [9] for only the first few coefficients; but in this case, these coefficients were computed quite explicitly.

2 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 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 $a \in A$, 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.

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.

Remark 2.2. The term *Black Hole* comes from the fact that each f_a is defined on the whole simplex W and the image of f_a on W is a single point for 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", which contains Black Hole as special case, thus generalizing the corresponding result in [8].

Definition 2.3. Suppose that for every $a \in \mathcal{A}$, Δ_a is either an all zero matrix or a rank one matrix. We call this the *weak Black Hole* case.

Let Ω denote a compact subset of $\mathbb{R}$ and assume $0 \in \Omega$. For a real function $f(\varepsilon)$ defined on Ω, we say $f(\varepsilon)$ is analytic around $\varepsilon = 0$ if $f(\varepsilon)$ can be analytically continued to a neighborhood of $\varepsilon = 0$ in $\mathbb{C}$, thus admitting a Taylor series expansion around $\varepsilon = 0$, which converges to $f(\varepsilon)$ around $\varepsilon = 0$. For a given analytic function $f(\varepsilon)$ around $\varepsilon = 0$, let $\operatorname{ord}(f(\varepsilon))$ denote its order, i.e., the degree of the first non-zero term of its Taylor series expansion around $\varepsilon = 0$. We say the transition probability matrix $\Delta(\varepsilon)$ is *normally parameterized* (reader can jump to Section 3 for quick examples, e.g., Example 3.1) by ε ($\varepsilon \geq 0$) around $\varepsilon = 0$ if

1. each entry of $\Delta(\varepsilon)$ is an analytic function at $\varepsilon = 0$,
2. when $\varepsilon > 0$, $\Delta(\varepsilon)$ is (non-negative and) irreducible,
3. $\Delta(0)$ is a weak black hole, i.e., for every $a \in \mathcal{A}$, $\Delta_a(0)$ is either an all zero matrix or a rank one matrix.

In the following, expressions like $p_X(x)$ will be used to mean $P(X = x)$ and we drop the subscripts if the context is clear: $p(x), p(z)$ mean $P(X = x), P(Z = z)$, respectively, and further $p(y|x), p(z_0|z_{-n}^{-1})$ mean $P(Y = y|X = x), P(Z_0 = z_0|Z_{-n}^{-1} = z_{-n}^{-1})$, respectively.

Proposition 2.4. *Suppose that $\Delta(\varepsilon)$ is analytically parameterized by $\varepsilon \geq 0$ and when $\varepsilon > 0$, $\Delta(\varepsilon)$ is non-negative and irreducible. Then for any fixed hidden Markov sequence $z_{-n}^0 \in \mathcal{A}^{n+1}$,*

1. *$p(z_{-n}^{-1})$ is analytic around $\varepsilon = 0$;*
2. *$p(y_i = \cdot\ |z_{-n}^i) := (p(y_i = b|z_{-n}^i) : b = 1, 2, \cdots, B)$ is analytic around $\varepsilon = 0$, where $\cdot$ denotes B possible states of Markov chain Y,*
3. *$p(z_0|z_{-n}^{-1})$ is analytic around $\varepsilon = 0$.*

Proof. We first prove that when $\varepsilon > 0$, $\Delta(\varepsilon)$ has a unique positive stationary vector $\pi(\varepsilon)$, which can be continuously extended to an analytic function around $\varepsilon = 0$.

When $\varepsilon > 0$, $\Delta(\varepsilon)$ is non-negative and irreducible. By Perron-Frobenius theory [18], $\Delta(\varepsilon)$ 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 non-negative analytic functions of ε and the first non-zero term of every $\pi_i(\varepsilon)$'s Taylor series expansion has a positive coefficient. Then we conclude that for each i

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

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

1. Now

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

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

2. Let $x_{i,-n} = x_{i,-n}(z_{-n}^{i})$ denote $p(y_i = \cdot\, | z_{-n}^{i})$. Then one checks that $x_{i,-n}$ satisfies the following iteration:

$$x_{i,-n} = \frac{x_{i-1,-n}\Delta_{z_i}}{x_{i-1,-n}\Delta_{z_i}\mathbf{1}}, \qquad -n \leq i \leq -1, \quad (2)$$

starting with $x_{-n-1,-n} = p(y_{-n-1} = \cdot\,)$. Because Δ is analytically parameterized by ε ($\varepsilon \geq 0$) and $\Delta(\varepsilon)$ is non-negative and irreducible when $\varepsilon > 0$, inductively we can prove (the proof is similar to the proof of analyticity of $\pi(\varepsilon)$) 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 analytic functions around $\varepsilon = 0$. Note that for each i

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

The existence of the Taylor series expansion of $x_{i,-n}$ around $\varepsilon = 0$ (for any i) then follows.

3. One checks that

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

Analyticity of $p(z_0|z_{-n}^{-1})$ immediately follows from (3) and analyticity of $x_{-1,-n}$ around $\varepsilon = 0$, which has been shown in *2.*.

□

Lemma 2.5. *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 dependent only 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.

□

For a mapping $v = v(\varepsilon) : [0, \infty) \to W$ analytic at $\varepsilon = 0$ and a hidden Markov sequence z_{-n}^0, define

$$p_v(z_{-n}^{-1}) = v\Delta_{z_{-n}} \cdots \Delta_{z_{-1}} \mathbf{1}, \text{ and } p_v(z_0|z_{-n}^{-1}) = \frac{p_v(z_{-n}^0)}{p_v(z_{-n}^{-1})}.$$

Let $b_{v,j}(z_{-n}^0)$ denote the coefficient of ε^j in the Taylor series expansion of $p_v(z_0|z_{-n}^{-1})$ (note that $b_{v,j}(z_{-n}^0)$ does not depend on ε),

$$p_v(z_0|z_{-n}^{-1}) = \sum_{j=0}^{\infty} b_{v,j}(z_{-n}^0)\varepsilon^j.$$

We have the following lemma:

Lemma 2.6. *For two mappings $v = v(\varepsilon), \hat{v} = \hat{v}(\varepsilon) : [0, \infty) \to W$ analytic at $\varepsilon = 0$, if* $\operatorname{ord}(p_v(z_{-n}^{-1})), \operatorname{ord}(p_{\hat{v}}(z_{-n}^{-1})) \leq k$, *we then have*

$$b_{v,j}(z_{-n}^0) = b_{\hat{v},j}(z_{-n}^0), \quad 0 \leq j \leq n - 4k - 1.$$

Proof. Let $x_{v,i} = x_{v,i}(z_{-n}^i) = p_v(y_i = \cdot\ |z_{-n}^i)$ and $x_{\hat{v},i} = x_{\hat{v},i}(\hat{z}_{-\hat{n}}^i) = p_{\hat{v}}(y_i = \cdot\ |z_{-\hat{n}}^i)$, where $\cdot$ denotes the possible states of Markov chain Y. Consider the Taylor series expansion of $x_{v,i}, x_{\hat{v},i}$ around $\varepsilon = 0$,

$$x_{v,i} = a_{v,0}(z_{-n}^i) + a_{v,1}(z_{-n}^i)\varepsilon + a_{v,2}(z_{-n}^i)\varepsilon^2 + \cdots \tag{4}$$

$$x_{\hat{v},i} = a_{\hat{v},0}(z_{-n}^i) + a_{\hat{v},1}(z_{-n}^i)\varepsilon + a_{\hat{v},2}(z_{-n}^i)\varepsilon^2 + \cdots \tag{5}$$

We shall show that $a_{v,j}(z_{-n}^i) = a_{\hat{v},j}(z_{-n}^i)$ for j with

$$0 \leq j \leq n + i - \sum_{l=-n}^{i} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\},$$

where for any hidden Markov sequence z_{-n}^i,

$$J_v(z_{-n}^i) = \begin{cases} 1 + \operatorname{ord}(p_v(z_i|z_{-n}^{i-1})) & \text{if } \operatorname{ord}(p_v(z_i|z_{-n}^{i-1})) > 0 \\ 0 & \text{if } \operatorname{ord}(p_v(z_i|z_{-n}^{i-1})) = 0 \end{cases},$$

and $J_{\hat{v}}(z_{-n}^i)$ is similarly defined.

Note that

$$x_{v,i+1} = \frac{x_{v,i}\Delta_{z_{i+1}}(\varepsilon)}{x_{v,i}\Delta_{z_{i+1}}(\varepsilon)\mathbf{1}}. \tag{6}$$

Now with (4) and (5), we have

$$x_{v,i}\Delta_{z_{i+1}}(\varepsilon) = \sum_{j=0}^{\infty} a_{v,j}(z_{-n}^i) \sum_{k=0}^{\infty} \frac{\Delta_{z_{i+1}}^{(k)}(0)}{k!}\varepsilon^k = \sum_{l=0}^{\infty} c_{v,l}(z_{-n}^{i+1})\varepsilon^l, \tag{7}$$

where superscript ${}^{(k)}$ denotes the k-th order derivative with respect to ε.

We proceed by induction on i (from $-n$ to -1).

First consider the case when $i = -n$. When $\max\{J_v(z_{-n}), J_{\hat{v}}(z_{-n})\} > 0$, the statement is vacuously true; when $J_v(z_{-n}) = J_{\hat{v}}(z_{-n}) = 0$, necessarily $\Delta_{z_{-n}}(0)$ is a rank one matrix, $v(0)\Delta_{z_{-n}}(0)\mathbf{1} > 0$ and $\hat{v}(0)\Delta_{z_{-n}}(0)\mathbf{1} > 0$. Then we have

$$a_{v,0}(z_{-n}) = \frac{v(0)\Delta_{z_{-n}}(0)}{v(0)\Delta_{z_{-n}}(0)\mathbf{1}} \overset{(*)}{=} \frac{\hat{v}(0)\Delta_{z_{-n}}(0)}{\hat{v}(0)\Delta_{z_{-n}}(0)\mathbf{1}} = a_{\hat{v},0}(z_{-n}),$$

where $(*)$ follows from the fact that $\Delta_{z_{-n}}(0)$ is a rank one matrix. (Similarly as in Remark 2.2, when Δ_a is a rank one matrix, the mapping $f_a(w)$ will map every $w \in W$ with $w\Delta_a\mathbf{1} > 0$ to one single point.)

Now suppose $i \geq -n$ and that $a_{v,j}(z_{-n}^i) = a_{\hat{v},j}(z_{-n}^i)$ for j with $0 \leq j \leq n + i - \sum_{l=-n}^{i} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\}$.

If $\operatorname{ord}(p_v(z_{i+1}|z_{-n}^i)) > 0$, since the leading coefficient vector of the Taylor series expansion in (7) is non-negative, $c_{v,j}(z_{-n}^{i+1}) \equiv \mathbf{0}$ for all j with $0 \leq j \leq J_v(z_{-n}^{i+1})-2$ and $c_{v,J_v(z_{-n}^{i+1})-1}(z_{-n}^{i+1}) \not\equiv \mathbf{0}$. So applying Lemma 2.5 to the following expression

$$x_{v,i+1} = \frac{c_{v,0}(z_{-n}^{i+1}) + c_{v,1}(z_{-n}^{i+1})\varepsilon + \cdots + c_{v,l}(z_{-n}^{i+1})\varepsilon^l + \cdots}{c_{v,0}(z_{-n}^{i+1})\mathbf{1} + c_{v,1}(z_{-n}^{i+1})\mathbf{1}\varepsilon + \cdots + c_{v,l}(z_{-n}^{i+1})\mathbf{1}\varepsilon^l + \cdots} = \frac{\sum_{l=0}^{\infty} c_{v,l+J_v(z_{-n}^{i+1})-1}(z_{-n}^{i+1})\varepsilon^l}{\sum_{l=0}^{\infty} c_{l+J(z_{-m}^{i+1})-1}(z_{-m}^{i+1})\mathbf{1}\varepsilon^l}, \tag{8}$$

we conclude that for all j, $a_{v,j}(z_{-n}^{i+1})$ depends only on

$$c_{v,l}(z_{-n}^{i+1}), \qquad J_v(z_{-n}^{i+1}) - 1 \leq l \leq J_v(z_{-n}^{i+1}) - 1 + j,$$

implying that $a_{v,j}(z_{-n}^{i+1})$ depends only on (or some of)

$$a_{v,l}(z_{-n}^i), \qquad \Delta_{z_{i+1}}^{(l)}(0), \qquad 0 \leq l \leq J_v(z_{-n}^{i+1}) - 1 + j.$$

A completely parallel argument also applies to the case when $\operatorname{ord}(p_{\hat{v}}(z_{i+1}|z_{-n}^i)) > 0$. More specifically, the statements above for the case $\operatorname{ord}(p(z_{i+1}|z_{-m}^i)) > 0$ are still true if we replace v with $\hat{v}$, which implies that $a_{\hat{v},j}(z_{-n}^{i+1})$ depends only on (or some of)

$$a_{\hat{v},l}(z_{-n}^i), \qquad \Delta_{z_{i+1}}^{(l)}(0), \qquad 0 \leq l \leq J_{\hat{v}}(z_{-n}^{i+1}) - 1 + j.$$

Thus when $\max\{J_v(z_{-n}^{i+1}), J_{\hat{v}}(z_{-n}^{i+1})\} > 0$, we have $a_{v,j}(z_{-n}^{i+1}) = a_{\hat{v},j}(z_{-n}^{i+1})$ for j with

$$0 \leq j \leq n + i - \sum_{l=-n}^{i} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\} - \max\{J_v(z_{-n}^{i+1}) - 1, J_{\hat{v}}(z_{-n}^{i+1}) - 1\}$$

$$= n + (i+1) - \sum_{l=-n}^{i+1} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\}.$$

If $\text{ord}\,(p_v(z_{i+1}|z_{-n}^i)) = 0$, by (3) necessarily we have

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

Again by Lemma 2.5 applied to expression (8), for any j, $a_{v,j}(z_{-n}^{i+1})$ depends only on

$$a_{v,l}(z_{-n}^i), \qquad \Delta_{z_{i+1}}^{(l)}(0), \qquad 0 \leq l \leq j,$$

Similarly if $\text{ord}\,(p_{\hat{v}}(z_{i+1}|\hat{z}_{-n}^i)) = 0$, we deduce that for any j, $a_{\hat{v},j}(z_{-n}^{i+1})$ depends only on

$$a_{\hat{v},l}(z_{-n}^i), \qquad \Delta_{z_{i+1}}^{(l)}(0), \qquad 0 \leq l \leq j.$$

Thus if $\max\{J_v(z_{-n}^{i+1}), J_{\hat{v}}(z_{-n}^{i+1})\} = 0$, for any j with

$$0 \leq j \leq n + i - \sum_{l=-n}^{i} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\} = n + i - \sum_{l=-n}^{i+1} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\},$$

we have $a_{v,j}(z_{-n}^{i+1}) = a_{\hat{v},j}(z_{-n}^{i+1})$.

Now, let $t = n + (i+1) - \sum_{l=-n}^{i+1} \max\{J_v(z_{-n}^l), J_{\hat{v}}(\hat{z}_{-n}^l)\}$. Then one can show that

$$a_{v,t}(z_{-n}^{i+1}) = \frac{a_{v,t}(z_{-n}^i)\Delta_{z_{i+1}}(0)a_{v,0}(z_{-n}^i)\Delta_{z_{i+1}}(0)\mathbf{1} - a_{v,0}(z_{-n}^i)\Delta_{z_{i+1}}(0)a_{v,t}(z_{-n}^i)\Delta_{z_{i+1}}(0)\mathbf{1}}{(a_{v,0}(z_{-n}^i)\Delta_{z_{i+1}}(0)\mathbf{1})^2} + \text{other terms},$$

where the first term in the expression above is equal to $\mathbf{0}$ (since $\Delta_{z_{i+1}}(0)$ is a rank one matrix), and the "other terms" are functions of

$$a_{v,0}(z_{-n}^i), \cdots, a_{v,t-1}(z_{-n}^i), \Delta_{z_{i+1}}^{(0)}(0), \cdots, \Delta_{z_{i+1}}^{(t)}(0). \tag{9}$$

It follows that $a_{v,j}(z_{-n}^{i+1})$ is a function of the same quantities in (9). By a completely parallel argument as above, $a_{\hat{v},j}(z_{-n}^{i+1})$ is the same function of of the same quantities in (9). So we have $a_{v,j}(z_{-n}^{i+1}) = a_{\hat{v},j}(z_{-n}^{i+1})$ for j with

$$0 \leq j \leq n + (i+1) - \sum_{l=-n}^{i+1} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\}.$$

Notice that

$$\sum_{l=-n}^{-1} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\} \leq \sum_{l=-n}^{-1} (J_v(z_{-n}^l) + J_{\hat{v}}(z_{-n}^l)) \leq 4k.$$

The lemma then immediately follows from (3) and the proven fact that $a_{v,j}(z_{-n}^{-1}) = a_{\hat{v},j}(z_{-n}^{-1})$ for j with

$$0 \leq j \leq n - 1 - \sum_{l=-n}^{-1} \max\{J_v(z_{-n}^l), J_{\hat{v}}(z_{-n}^l)\}.$$

□

By Proposition 2.4, for any hidden Markov string z_{-m}^0 (or $z_{-\hat{m}}^0$), $p(y_{-n-1}=\cdot|z_{-m}^{-n-1})$ (or $p(y_{-n-1}=\cdot|z_{-\hat{m}}^{-n-1})$) is analytic. So for $n \le m, \hat{m}$, if $v(\varepsilon)$ (or $\hat{v}(\varepsilon)$) is equal to $p(y_{-n-1}=\cdot|z_{-m}^{-n-1})$ (or $p(y_{-n-1}=\cdot|z_{-\hat{m}}^{-n-1})$), then $p_v(z_{-n}^0)$ (or $p_{\hat{v}}(z_{-n}^0)$) will be equal to $p(z_{-n}^0|z_{-m}^{-n-1})$ (or $p(z_{-n}^0|z_{-\hat{m}}^{-n-1})$); and if for a Markov state y, $v(\varepsilon)$ (or $\hat{v}(\varepsilon)$) is equal to $p(y_{-n-1}=\cdot|z_{-m}^{-n-1}y)$ (or $p(y_{-n-1}=\cdot|z_{-\hat{m}}^{-n-1}y)$), then $p_v(z_{-n}^0)$ (or $p_{\hat{v}}(z_{-n}^0)$) will be equal to $p(z_{-n}^0|z_{-m}^{-n-1}y)$ (or $p(z_{-n}^0|z_{-\hat{m}}^{-n-1}y)$). In what follows, slightly abusing the notation, we use $b_j(z_{-m}^0)$ to represent the coefficient of ε^j in the expansion of $p(z_0|z_{-m}^{-1})$, namely

$$p(z_0|z_{-m}^{-1}) = b_0(z_{-m}^0) + b_1(z_{-m}^0)\varepsilon + b_2(z_{-m}^0)\varepsilon^2 + \cdots ; \tag{10}$$

and we use $b_j(z_{-m}^0 y_{-m-1})$ to represent the coefficient of ε^j in the expansion of $p(z_0|z_{-m}^{-1}y_{-m-1})$, namely

$$p(z_0|z_{-m}^{-1}y_{-m-1}) = b_0(z_{-m}^0 y_{-m-1}) + b_1(z_{-m}^0 y_{-m-1})\varepsilon + b_2(z_{-m}^0 y_{-m-1})\varepsilon^2 + \cdots . \tag{11}$$

It then immediately follows from Lemma 2.6 that

Proposition 2.7. *Given fixed sequences $z_{-m}^0, z_{-\hat{m}}^0, z_{-m}^0 y_{-m-1}, \hat{z}_{-\hat{m}}^0 y_{-\hat{m}-1}$ with $z_{-n}^0 = \hat{z}_{-n}^0$ such that*

$$\mathrm{ord}\,(p(z_{-n}^{-1}|z_{-m}^{-n-1})), \mathrm{ord}\,(p(\hat{z}_{-n}^{-1}|\hat{z}_{-\hat{m}}^{-n-1})), \mathrm{ord}\,(p(z_{-n}^{-1}|z_{-m}^{-n-1}y_{-m-1})), \mathrm{ord}\,(p(\hat{z}_{-n}^{-1}|\hat{z}_{-\hat{m}}^{-n-1}y_{-\hat{m}-1})) \le k,$$

for $n \le m, \hat{m}$ and some k, we have for j with $0 \le j \le n-4k-1$,

$$b_j(z_{-m}^0 y_{-m-1}) = b_j(\hat{z}_{-\hat{m}}^0 y_{-\hat{m}-1}) = b_j(z_{-m}^0) = b_j(\hat{z}_{-\hat{m}}^0). \tag{12}$$

Consider expression (10). 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 2.8. *For a hidden Markov chain Z with normally parameterized $\Delta(\varepsilon)$, we have for any $k \ge 0$,*

$$H(Z) = H(Z)|_{\varepsilon=0} + \sum_{j=1}^{k+1} f_j \varepsilon^j \log\varepsilon + \sum_{j=1}^{k} g_j \varepsilon^j + O(\varepsilon^{k+1}), \tag{13}$$

where f_j's and g_j's for $j = 1, 2, \cdots, k+1$ are functions (more specifically, elementary functions built from log and polynomials) of $\Delta^{(i)}(0)$ for $0 \le i \le 6k+6$ and can be computed from $H_{6k+6}(Z(\varepsilon))$.

The following theorem [3] states the Birch upper bound and lower bound of $H(Z)$, which we shall use in the proof of Theorem 2.8.

Theorem 2.9 (Birch, 1962)**.** *For any n,*

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

Proof of Theorem 2.8. First fix n such that $n \geq n_0 = 6k + 6$. 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 $j \geq k+2$,

$$\left| \sum_{\mathrm{ord}\,(p(z_{-n}^0))=j} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) \right| = O(\varepsilon^{k+1}). \tag{14}$$

(We used simplified notation above: $\sum_{z_{-n}^0}$ means summation over all $z_{-n}^0 \in \mathcal{A}^{n+1}$, while $\sum_{\mathrm{ord}\,(p(z_{-n}^0))=j}$ means summation over all $z_{-n}^0 \in \mathcal{A}^{n+1}$ with $\mathrm{ord}\,(p(z_{-n}^0)) = j$; the same notational convention will be followed in the rest of the proof.) So, in the following we only consider the sequences z_{-n}^0 with $\mathrm{ord}\,(p(z_{-n}^0)) \leq k+1$. For such sequences, since $\mathrm{ord}\,(p(z_0|z_{-n}^{-1})) \leq \mathrm{ord}\,(p(z_{-n}^0)) \leq k+1$, we have

$$|\log p(z_0|z_{-n}^{-1}) - \log p^{<2k+1>}(z_0|z_{-n}^{-1})| = O(\varepsilon^{k+1}); \tag{15}$$

and by Lemma 2.6, we have

$$p^{<2k+1>}(z_0|z_{-n}^{-1}) = p^{<2k+1>}(z_0|z_{-n_0}^{-1}). \tag{16}$$

Now for any fixed $n \geq n_0$,

$$\begin{aligned} H_n(Z) &= \sum_{z_{-n}^0} -p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) \\ &\overset{(a)}{=} \sum_{\mathrm{ord}\,(p(z_{-n}^0))\leq k+1} -p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) + O(\varepsilon^{k+1}) \\ &\overset{(b)}{=} \sum_{\mathrm{ord}\,(p(z_{-n}^0))\leq k+1} -p(z_{-n}^0) \log p^{<2k+1>}(z_0|z_{-n}^{-1}) + O(\varepsilon^{k+1}) \\ &\overset{(c)}{=} \sum_{\mathrm{ord}\,(p(z_{-n_0}^0))\leq k+1} -p(z_{-n}^0) \log p^{<2k+1>}(z_0|z_{-n_0}^{-1}) + O(\varepsilon^{k+1}) \\ &= \sum_{\mathrm{ord}\,(p(z_{-n_0}^0))\leq k+1} -p(z_{-n_0}^0) \log p^{<2k+1>}(z_0|z_{-n_0}^{-1}) + O(\varepsilon^{k+1}), \end{aligned} \tag{17}$$

where (a) follows from (14); (b) follows from (15); (c) follows from (16), (14) and the fact that

$$\{z_{-n}^0 : \mathrm{ord}\,(p(z_{-n_0}^0)) \leq k+1\}$$
$$= \{z_{-n}^0 : \mathrm{ord}\,(p(z_{-n}^0)) \leq k+1\} \cup \{z_{-n}^0 : \mathrm{ord}\,(p(z_{-n_0}^0)) \leq k+1, \mathrm{ord}\,(p(z_{-n}^0)) \geq k+2\}.$$

Expanding (17), we obtain:

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

where f_j's and g_j's for $j = 1, 2, \cdots, k+1$ are functions dependent only on $\Delta^{(i)}(0)$ for $0 \le i \le n_0$ and can be computed from $H_{n_0}(Z)$ (in fact for fixed j, f_j and g_j are functions dependent only on $\Delta^{(i)}(0)$ for $0 \le i \le 6j+6$ and can be computed from $H_{6j+6}(Z)$). In particular,

$$\sum_{\mathrm{ord}\,(p(z_{-n}^0)) \le k+1} \sum_{\mathrm{ord}\,(p(z_0|z_{-n_0}^{-1}))=0} -p(z_{-n_0}^0) \log p^{<2k+1>}(z_0|z_{-n_0}^{-1}) \tag{18}$$

will contribute to $H(Z)|_{\varepsilon=0}$ and the terms ε^j, and

$$\sum_{\mathrm{ord}\,(p(z_{-n}^0)) \le k+1} \sum_{\mathrm{ord}\,(p(z_0|z_{-n_0}^{-1}))>0} -p(z_{-n_0}^0) \log p^{<2k+1>}(z_0|z_{-n_0}^{-1}) \tag{19}$$

will contribute to the terms $\varepsilon^j \log \varepsilon$ and the terms ε^j.

Using Corollary 2.7, one can apply similar argument as above to the Birch lower bound

$$\tilde{H}_n(Z) := H(Z_0|Z_{-n}^{-1} Y_{-n-1}) = \sum_{z_{-n}^0, y_{-n-1}} -p(z_{-n}^0 y_{-n-1}) \log p(z_0|z_{-n}^{-1} y_{-n-1}).$$

For the same n_0, one can show that $\tilde{H}_n(Z)$ takes the same form (17) as $H_n(Z)$, which implies that $H_n(Z)$ and $\tilde{H}_n(Z)$ have exactly the same coefficients of ε^j for $j \le k$ and of $\varepsilon^j \log \varepsilon$ for $j \le k+1$ when $n \ge n_0$. We thus prove the theorem. □

Remark 2.10. Theorem 2.8 still holds if we assume each entry of $\Delta(\varepsilon)$ is merely a C^{6k+6} function of ε in a neighborhood of $\varepsilon = 0$: the proof still works if "analytic" is replaced by "C^{6k+6}", and the Taylor series expansions are replaced by Taylor polynomials with remainder. We assumed analyticity of the parametrization only for simplicity.

Remark 2.11. Note that at a Black Hole, we have $\mathrm{ord}\,(p(z_0|z_{-n}^{-1})) = 0$ for any hidden Markov symbol sequence z_{-n}^0. Thus, from the discussion surrounding expressions (18) and (19) above, we see that $f_j = 0$ for all j. By the proof of Theorem 2.8, Formula (13) is a Taylor polynomial with remainder; this is consistent with the Taylor series formula for a Black Hole in [8].

3 Applications to Finite-State Memoryless Channels at High Signal-to-Noise Ratio

Consider a finite-state memoryless channel with stationary input process. Here, $C = \{C_n\}$ is an i.i.d. channel state process over finite alphabet $\mathcal{C}$ with $p_C(c) = q_c$ for $c \in \mathcal{C}$, $X = \{X_n\}$ is a stationary input process, independent of C, over finite alphabet $\mathcal{X}$ and $Z = \{Z_n\}$ is the resulting (stationary) output process over finite alphabet $\mathcal{Z}$. Let $p(z_n|x_n, c_n) = P(Z_n = z_n|X_n = x_n, C_n = c_n)$ denote the probability that at time n, the channel output symbol is z_n given that the channel state is c_n and the channel input is x_n. The mutual information for such a channel is:

$$I(X, Z) := H(Z) - H(Z|X) \overset{(*)}{=} H(Z) - \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} p(x, z) \log p(z|x),$$

where $(*)$ follows from the memoryless property of the channel, and for $x \in \mathcal{X}, z \in \mathcal{Z}$,

$$p(x,z) = \sum_{c \in \mathcal{C}} p(z|x,c)p(x)p(c), \qquad p(z|x) = \sum_{c \in \mathcal{C}} p(z|x,c)p(c).$$

Now we introduce an alternative framework, using the concept of channel noise. As above, let C be an i.i.d. channel state process, and let X be a stationary input process, independent of C, over finite alphabets $\mathcal{C}$, $\mathcal{X}$. Let $\mathcal{E}$ (resp., $\mathcal{Z}$) be finite alphabets of abstract error events (resp. output symbols) and let $\Phi : \mathcal{X} \times \mathcal{C} \times \mathcal{E} \to \mathcal{Z}$ be a function. For each $x \in \mathcal{X}$ and $c \in \mathcal{C}$, let $p(\cdot|x,c)$ be a conditional probability distribution on $\mathcal{E}$. This defines a jointly distributed stationary process (X,C,E) over $\mathcal{X} \times \mathcal{C} \times \mathcal{E}$. If X is a first order Markov chain with transition probability matrix Π, then (X,C,E) is a Markov chain with transition probability matrix Δ, defined by

$$\Delta_{(x,c,e),(y,d,f)} = \Pi_{xy} \cdot q_d \cdot p(f|y,d)$$

and Φ, Δ define a hidden Markov chain, denoted $Z(\Delta, \Phi)$.

We claim that the output process Z, described in the first paragraph of this section, fits into this alternative framework (when X is a first order Markov chain). To see this, let $\mathcal{E} = \mathcal{X} \times \mathcal{C} \times \mathcal{Z}$, and define $p(e = (x,c,z)|x',c') = p(z|x,c)$ if $x = x'$ and $c = c'$, and 0 otherwise. Define $\Phi(x',c',(x,c,z)) = z$. Then, $Z = Z(\Delta, \Phi)$ is a hidden Markov chain. So, from hereon we adopt the alternative framework.

Now, we assume that X is an irreducible first order Markov chain and that the channel is parameterized by ε such that for each x, c, and e, $p(e|x,c)(\varepsilon)$ are analytic functions of $\varepsilon \geq 0$. For each $\varepsilon \geq 0$, let $\Delta(\varepsilon)$ denote the corresponding transition probability matrix on state set $\mathcal{X} \times \mathcal{C} \times \mathcal{E}$ and $\{Z(\varepsilon)\}$ denote the family of resulting output hidden Markov chains. We also assume that there is a one-to-one function from $\mathcal{X}$ into $\mathcal{Z}$, $z = z(x)$, such that for all c, $p(z(x)|x,c)(0) = 1$. In other words, ε behaves like a "composite index" indicating how good the channel is, and small ε corresponds to the high signal-to-noise ratio. Then one can verify that $\Delta(0)$ is a weak black hole and $\Delta(\varepsilon)$ is normally parameterized. Thus, by Theorem 2.8, we obtain an asymptotic formula for $H(Z(\varepsilon))$ around $\varepsilon = 0$. We remark that the above naturally generalizes to the case where X is a higher order irreducible Markov chain (through appropriately grouping matrices into blocks).

In the remainder of this section, we give three examples to illustrate the idea.

Example 3.1. [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. Note that this channel only has one channel state, and at $\varepsilon = 0$, $p_{Z|X}(1|1) = 1, p_{Z|X}(0|0) = 1$, so it fits in the alternative framework described in the beginning of Section 3.

Indeed, suppose X is a first order irreducible 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 (the column and row indices of the following matrix are ordered alphabetically):

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

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 $\varepsilon = 0$,

$$\Delta = \begin{bmatrix} \pi_{00} & 0 & \pi_{01} & 0 \\ \pi_{00} & 0 & \pi_{01} & 0 \\ \pi_{10} & 0 & \pi_{11} & 0 \\ \pi_{10} & 0 & \pi_{11} & 0 \end{bmatrix}, \Delta_0 = \begin{bmatrix} \pi_{00} & 0 & 0 & 0 \\ \pi_{00} & 0 & 0 & 0 \\ \pi_{10} & 0 & 0 & 0 \\ \pi_{10} & 0 & 0 & 0 \end{bmatrix}, \Delta_1 = \begin{bmatrix} 0 & 0 & \pi_{01} & 0 \\ 0 & 0 & \pi_{01} & 0 \\ 0 & 0 & \pi_{11} & 0 \\ 0 & 0 & \pi_{11} & 0 \end{bmatrix},$$

thus 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 $\varepsilon = 0$ [20, 8]; if π_{00} or π_{11} are zero, then this is a weak Black hole case with normal parameterization (of ε), for which Theorem 2.8 can be applied and an asymptotic formula for $H(Z)$ around $\varepsilon = 0$ can be derived.

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

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

where $0 \le p \le 1$, it has been shown [16] 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:

$$H(Z) = H(X) + f(X)\varepsilon \log(1/\varepsilon) + g(X)\varepsilon + O(\varepsilon^2 \log \varepsilon), \tag{20}$$

where X is the input Markov chain of any order m with transition probabilities $P(X_t = a_0 | X_{t-m}^{t-1} = a_{-m}^{-1})$, $a_{-m}^0 \in \mathcal{X}^m$, where $\mathcal{X} = \{0, 1\}$, Z is the output process obtained by passing X through a BSC(ε), and $f(X)$ and $g(X)$ can be explicitly computed. Theorem 2.8 can be used to generalize (20) to a formula with higher asymptotic terms. 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 define an augmented Markov chain $\{\tilde{X}_i, i \in \mathbb{Z}\}$ by

$$\tilde{X}_i = (X_{mi}, X_{mi+1}, \cdots, X_{mi+m-1});$$

correspondingly we have output process $\tilde{Z} = \{\tilde{Z}_i, i \in \mathbb{Z}\}$, where

$$\tilde{Z}_i = (Z_{mi}, Z_{mi+1}, \cdots, Z_{mi+m-1}).$$

Then one check that for this augmented hidden Markov chain, we have a Black Hole at $\varepsilon = 0$, which implies that the Taylor series expansions of $H(Z) = H(\tilde{Z})/m$ around $\varepsilon = 0$ can be explicitly computed (in principle); similarly 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 which case an asymptotic formula of $H(Z)$ around $\varepsilon = 0$ can be obtained.

Example 3.2. [Binary Markov Chains Corrupted by BEC(ε)]

Consider a binary erasure channel with fixed erasure rate ε (denoted by BEC(ε)). At time n the channel can be characterized by the following equation

$$Z_n = \begin{cases} X_n & \text{if } E_n = 0 \\ e & \text{if } E_n = 1 \end{cases},$$

where $\{X_n\}$ denotes the input process, e denotes the erasure, $\{E_n\}$ 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. Again this channel only has one channel state, and at $\varepsilon = 0$, $p_{Z|X}(1|1) = 1, p_{Z|X}(0|0) = 1$, so it fits in the alternative framework described in the beginning of Section 3.

If the input X is a first order irreducible Markov chain with transition probability matrix

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

and let Z denote the output process. Then $Y = (X, E)$ is jointly Markov with (the column and row indices of the following matrix are ordered alphabetically)

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

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

Now one checks that

$$\Delta_0 = \begin{bmatrix} \pi_{00}(1-\varepsilon) & 0 & 0 & 0 \\ \pi_{00}(1-\varepsilon) & 0 & 0 & 0 \\ \pi_{10}(1-\varepsilon) & 0 & 0 & 0 \\ \pi_{10}(1-\varepsilon) & 0 & 0 & 0 \end{bmatrix}, \Delta_1 = \begin{bmatrix} 0 & 0 & \pi_{01}(1-\varepsilon) & 0 \\ 0 & 0 & \pi_{01}(1-\varepsilon) & 0 \\ 0 & 0 & \pi_{11}(1-\varepsilon) & 0 \\ 0 & 0 & \pi_{11}(1-\varepsilon) & 0 \end{bmatrix}, \Delta_e = \begin{bmatrix} 0 & \pi_{00}\varepsilon & 0 & \pi_{01}\varepsilon \\ 0 & \pi_{00}\varepsilon & 0 & \pi_{01}\varepsilon \\ 0 & \pi_{10}\varepsilon & 0 & \pi_{11}\varepsilon \\ 0 & \pi_{10}\varepsilon & 0 & \pi_{11}\varepsilon \end{bmatrix}.$$

One checks that $\Delta(\varepsilon)$ is normally parameterized by ε and thus Theorem 2.8 can be applied. Furthermore, Theorem 2.8 can be applied to the case when the input is an m-th order irreducible Markov chain X to obtain asymptotic formula for $H(Z)$ around $\varepsilon = 0$.

Example 3.3. [Binary Markov Chains Corrupted by Special Gilbert-Elliot Channel]

Consider a binary Gilbert-Elliot channel, whose channel state (denoted by $C = \{C_n\}$) varies as an i.i.d. binary stochastic process with $p_C(0) = q_0, p_C(1) = q_1$ (here the channel state varies as an i.i.d. process, rather than a generic Markov process). 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 $p_{E|C}(0|0) = 1-\varepsilon_0$, $p_{E|C}(0|1) = 1-\varepsilon_1$, $p_{E|C}(1|0) = \varepsilon_0$, $p_{E|C}(1|1) = \varepsilon_1$ and $\{Z_n\}$ denotes the corrupted output. For such a channel, $p_{Z|(X,C)}(1|1,c) = 1, p_{Z|(X,C)}(0|0,c) =$

1 at $\varepsilon = 0$ for any channel state c. So it fits in the alternative framework described in the beginning of Section 3.

To see this in more detail, we consider the special case when the input X is a first order irreducible Markov chain with transition probability matrix

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

and let Z denote the output process. Then $Y = (X, C, E)$ is jointly Markov with transition probability matrix Δ, which can be concisely written using Kronecker product as follows (the column and row indices of the following matrix are ordered alphabetically)

$$\Delta = \begin{bmatrix} \pi_{00}q_0(1-\varepsilon_0) & \pi_{00}q_0\varepsilon_0 & \pi_{00}q_1(1-\varepsilon_1) & \pi_{00}q_1\varepsilon_1 & pi_{01}q_0(1-\varepsilon_0) & \pi_{01}q_0\varepsilon_0 & \pi_{01}q_1(1-\varepsilon_1) & \pi_{01}q_1\varepsilon_1 \\ \pi_{10}q_0(1-\varepsilon_0) & \pi_{10}q_0\varepsilon_0 & \pi_{10}q_1(1-\varepsilon_1) & \pi_{10}q_1\varepsilon_1 & \pi_{11}q_0(1-\varepsilon_0) & \pi_{11}q_0\varepsilon_0 & \pi_{11}q_1(1-\varepsilon_1) & \pi_{11}q_1\varepsilon_1 \end{bmatrix} \otimes \mathbf{1}_4,$$

where $\mathbf{1}_4$ stands for the all 1 column vector of length 4. $Z = \Phi(X, C, E)$ is hidden Markov with

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

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

For some positive k, let $\varepsilon_0 = \varepsilon, \varepsilon_1 = k\varepsilon$. If $\varepsilon = 0$, one checks that

$$\Delta_0 = \begin{bmatrix} \pi_{00}q_0 & 0 & \pi_{00}q_1 & 0 & 0 & 0 & 0 & 0 \\ \pi_{10}q_0 & 0 & \pi_{10}q_1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \otimes \mathbf{1}_4, \Delta_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & \pi_{01}q_0 & 0 & \pi_{01}q_1 & 0 \\ 0 & 0 & 0 & 0 & \pi_{11}q_0 & 0 & \pi_{11}q_1 & 0 \end{bmatrix} \otimes \mathbf{1}_4.$$

So, both Δ_0 and Δ_1 will be rank one matrices and one can check that $\Delta(\varepsilon)$ is normally parameterized by ε. Again, Theorem 2.8 can be applied to the case when the input is an m-th order irreducible Markov chain X to obtain an asymptotic formula for $H(Z)$ around $\varepsilon = 0$.

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) available at http://arxiv.org/abs/0803.3360.

[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] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1995.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] 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.