Asymptotics of Noisy Constrained Channel Capacity

Guangyue Han, Brian Marcus
Department of Mathematics
University of British Columbia
Vancouver, B.C., V6T 1Z2
e-mail: ghan, marcus@math.ubc.ca

January 14, 2007

Abstract

In this paper, we generalize a result in [17] and derive an asymptotic formula for the entropy rate of a hidden Markov chain, observed when a Markov chain passes through a binary symmetric channel. And we prove an asymptotic formula for the capacity of a binary symmetric channel with input process supported on an irreducible finite type constraint.

1 Introduction

Let $\mathbb{P}$ denote the set of all the stationary stochastic processes on the binary alphabet, and let $\mathbb{P}_n$ denote the set of all the stationary distributions (again binary) with length n. Consider $X = X_{-\infty}^{\infty} \in \mathbb{P}$. The entropy rate of X is defined to be

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

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

If X is a finite-state Markov chain, then $H(X)$ has a simple analytic form. A *hidden Markov chain* Y can be defined as a deterministic function of a Markov chain. Alternatively a hidden Markov chain is defined as a Markov chain observed in noise. It is well known that the two definitions are equivalent. For a hidden Markov chain Y, Blackwell [4] proved that $H(Y)$ is the integral of a certain function defined on a simplex with respect to certain measure. However Blackwell's measure is somewhat complicated, and it appears difficult to evaluate the integral.

Recently 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 [16] and a variation on the Birch bound [3] was introduced in [8]. 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. [20], and Sharma and Singh [21]. The connection between the entropy rate of a

hidden Markov chain and the top Lyapunov exponent of a random matrix product has been observed [11, 12, 10]. Several authors have studied [12, 16, 22] how the entropy rate varies as parameters of the underlying Markov chain vary. In [9], 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.

The capacity of a finite state channel is defined to be:

$$C = \lim_{n\to\infty} \sup_{X^0_{-n}\in\mathbb{P}_{n+1}} I(X^0_{-n}, Y^0_{-n})/(n+1);$$

here, Y^0_{-n} is the output distribution corresponding to X^0_{-n}. Alternatively capacity can be written as

$$C = \sup_{X\in\mathbb{P}} I(X, Y),$$

where Y is the output process corresponding to X. It is shown in [6] that the capacity of this type of channel can be achieved by a stationary input process. As shown in [7], for a finite state intersymbol interference channel one can approximate C using Markov input processes.

Generally speaking, it is very difficult to calculate the capacity. For a discrete memoryless channel (DMC), the Blahut-Arimoto algorithm ([2, 5]) can be applied to compute the capacity numerically. A generalized Blahut-Arimoto algorithm has been proposed to numerically compute the local maximum mutual information rate of a finite state machine channel [19].

This paper is organized as follows. In section 2 we generalize a result in [17] and derive an asymptotic formula for the entropy rate of a hidden Markov chain, obtained by observing a binary Markov chain, of arbitrary order, passed through a binary symmetric channel; this is of particular interest when the Markov chain has some zero transition probabilities. In Section 3, we discuss asymptotics of a binary symmetric channel with input sequences supported on an irreducible finite type constraint, and we derive an asymptotic formula for capacity.

We remark that recently Jacquet et. al. [13] reached similar conclusions and obtained higher order asymptotics for certain special cases.

2 Asymptotics of Entropy Rate

Let $\mathcal{W}$ denote all the binary words, and $\mathcal{W}_n$ denote all the binary words with length n. Let X denote any binary stationary distribution (with length possibly infinite). For a binary word $w \in \mathcal{W}$, we say that w is *allowed* in X if $p_X(w) > 0$. Let $\mathcal{A}(X)$ denote the set of all allowed words in X, and $\mathcal{A}_n(X) = \mathcal{A}(X) \cap \mathcal{W}_n$.

Now let X be an m-th order binary irreducible Markov process. Let E denote the i.i.d. $(\varepsilon, 1-\varepsilon)$ error process. Let $Y = Y_\varepsilon$ denote the function of the Markov chain $X \times E$ defined by:

$$Y_i = X_i \text{ if } E_i = 0,$$

and

$$Y_i = \overline{X_i} \text{ if } E_i = 1.$$

So, Y is the hidden Markov chain obtained by observing X over a binary symmetric channel with crossover probability ε (denoted by BSC(ε)).

By the Birch bound [3], for $n \geq m$, we have:

$$H(Y_0|Y_{-n+m}^{-1}, X_{-n}^{-n+m-1}, E_{-n}^{-n+m-1}) \leq H(Y) \leq H(Y_0|Y_{-n}^{-1}). \tag{1}$$

Note that each of these quantities is a function of ε, and the lower bound is really just

$$H(Y_0|Y_{-n+m}^{-1}, X_{-n}^{-n+m-1}).$$

Lemma 2.1. *For a stationary input distribution $X = X_{-n}^0 \in \mathbb{P}_{n+1}$ and the corresponding output distribution $Y = Y_{-n}^0$ through BSC(ε) and $0 \leq k \leq n$,*

$$H(Y_0|Y_{-n+k}^{-1}, X_{-n}^{-n+k-1}) = H(X_0|X_{-n}^{-1}) + f_n^k(X_{-n}^0)\varepsilon \log(1/\varepsilon) + O(\varepsilon),$$

where $f_n^k(X_{-n}^0)$ is the function defined on $\mathbb{P}_{n+1}$ given by (2).

Proof. In this proof, $w = w_{-n}^{-1}$, where w_{-j} is a single bit, and we let v denote a single bit. And we use the notation for probability:

$$p_{XY}(w) = p(X_{-n}^{-n+k-1} = w_{-n}^{-n+k-1}, Y_{-n+k}^{-1} = w_{-n+k}^{-1}),$$

$$p_{XY}(wv) = p(X_{-n}^{-n+k-1} = w_{-n}^{-n+k-1}, Y_{-n+k}^{-1} = w_{-n+k}^{-1}, Y_0 = v),$$

and

$$p_{XY}(v|w) = p(Y_0 = v|Y_{-n+k}^{-1} = w_{-n+k}^{-1}, X_{-n}^{-n+k-1} = w_{-n}^{-n+k-1}).$$

We remark that the definition of p_{XY} does depend on how we partition w_{-n}^{-1} according to k, however we keep the dependence implicit for notational convenience.

We split $H(Y_0|Y_{-n+k}^{-1}, X_{-n}^{-n+k-1})$ into three terms:

$$H(Y_0|Y_{-n+k}^{-1}, X_{-n}^{-n+k-1}) = \sum_{wv \in \mathcal{A}(X)} -p_{XY}(wv)\log(p_{XY}(v|w))$$

$$+ \sum_{w \in \mathcal{A}(X), wv \notin \mathcal{A}(X)} -p_{XY}(wv)\log(p_{XY}(v|w)) + \sum_{w \notin \mathcal{A}(X), v} -p_{XY}(wv)\log(p_{XY}(v|w)).$$

For the third term, we have

$$\sum_{w \notin \mathcal{A}(X), v} -p_{XY}(wv)\log(p_{XY}(v|w)) = \sum_{w \notin \mathcal{A}(X)} -p_{XY}(w)\sum_v p_{XY}(v|w)\log(p_{XY}(v|w)) \leq (\log 2)\sum_{w \notin \mathcal{A}(X)} p_{XY}(w),$$

where we use $-\sum_v p_{XY}(v|w)\log(p_{XY}(v|w)) \leq \log 2$ for any w. We conclude that the third term is $O(\varepsilon)$.

Since the function $x\log(x)$ is a smooth function of x when $x > 0$, the first term is a smooth function of ε, and we have

$$\sum_{wv \in \mathcal{A}(X)} -p_{XY}(wv)\log(p_{XY}(v|w)) = H(X_0|X_{-n}^{-1}) + O(\varepsilon).$$

For the second term, it is easy to check that for $w \in \mathcal{A}(X)$ and $wv \notin \mathcal{A}(X)$, $p_{XY}(v|w)/\varepsilon$ is bounded from above and away from 0. And together with

$$\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} p_{XY}(wv) = \left(\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X),j=1,\cdots,n-k} p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v) + \sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} p_X(w\bar{v}) \right) \varepsilon + o(\varepsilon),$$

we obtain

$$\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} -p_{XY}(wv)\log(p_{XY}(v|w)) = \left(\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X),j=1,\cdots,n-k} p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v) \right.$$

$$\left. + \sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} p_X(w\bar{v}) \right) \varepsilon\log(1/\varepsilon) + O(\varepsilon).$$

In short, $H(Y_0|Y_{-n+k}^{-1}, X_{-n}^{-n+k-1})$ can be rewritten as

$$H(Y_0|Y_{-n+k}^{-1}, X_{-n}^{-n+k-1}) = H(X_0|X_{-n}^{-1}) + f_n^k(X_{-n}^0)\varepsilon\log(1/\varepsilon) + O(\varepsilon),$$

where

$$f_n^k(X_{-n}^0) = \left(\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X),j=1,\cdots,n-k} p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v) + \sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} p_X(w\bar{v}) \right). \tag{2}$$

□

Remark 2.2. For any $\delta > 0$ and fixed m, the constant in $O(\varepsilon)$ in Lemma 2.1 can be chosen uniformly on $S_{m,\delta}$, where $S_{m,\delta}$ denotes the collection of stationary distributions $X \in \mathbb{P}_{m+1}$, such that for all $w \in \mathcal{A}_{m+1}(X)$, $p(w) \geq \delta$.

Theorem 2.3. *For an m-th order Markov chain X passing through a BSC(ε), with Y as the output hidden Markov chain,*

$$H(Y) = H(X) + f(X)\varepsilon\log(1/\varepsilon) + O(\varepsilon),$$

where $f(X) = f_{2m}^0(X_{-2m}^0) = f_{2m}^m(X_{-2m}^0)$.

Proof. We apply Lemma 2.1 to the Birch upper and lower bounds (eqn. (1)) of $H(Y)$. For the upper bound, $k = 0$, we have, for all n,

$$H(Y_0|Y_{-n}^{-1}) = H(X_0|X_{-n}^{-1}) + f_n^0(X_{-n}^0)\varepsilon\log(1/\varepsilon) + O(\varepsilon).$$

And for the lower bound, $k = m$, we have, for $n \geq m$,

$$H(Y_0|Y_{-n+m}^{-1}, X_{-n}^{-n+m-1}) = H(X_0|X_{-n}^{-1}) + f_n^m(X_{-n}^0)\varepsilon log(1/\varepsilon) + O(\varepsilon).$$

The first term always coincides for the upper and lower bounds. When $n \geq m$, since X is an m-th order Markov chain,

$$H(X_0|X_{-n}^{-1}) = H(X_0|X_{-m}^{-1}) = H(X).$$

Again let $w = w_{-n}^{-1}$, where w_{-j} is a single bit, and v denotes a single bit. If $w \in \mathcal{A}(X)$ and $wv \notin \mathcal{A}(X)$, then $p(w_{-m}^{-1}v) = 0$. It then follows that for an m-th order Markov chain, when $n \geq 2m$, $f_n^m(X_{-n}^0) = f_n^0(X_{-n}^0) = f_{2m}^0(X_{-2m}^0) = f_{2m}^m(X_{-2m}^0)$. Let $f(X) = f_{2m}^0(X_{-2m}^0)$, then the theorem follows. □

3 Asymptotics of Capacity

Consider a binary irreducible finite type constraint [14] $\mathcal{S}$ defined by the set (denoted by $\mathcal{F}$) of forbidden words with length $\hat{m}+1$. There are many such $\mathcal{F}$'s corresponding to the same $\mathcal{S}$ with different lengths; here we may choose $\mathcal{F}$ to be the one with the smallest length $\hat{m}+1$. And $\hat{m}=\hat{m}(\mathcal{S})$ is defined to be the topological order of the constraint $\mathcal{S}$. Let $\mathcal{A}(\mathcal{S})$ denote the set of all allowable words in $\mathcal{S}$, and $\mathcal{A}_n(\mathcal{S})=\mathcal{A}(\mathcal{S})\cap\mathcal{W}_n$.

For a constrained BSC(ε) with input sequences in $\mathcal{S}$, the capacity $C(\varepsilon)$ can be written as:

$$C(\varepsilon)=\lim_{n\to\infty}\sup_{X_{-n}^0\in\mathbb{P}_{n+1},\mathcal{A}(X_{-n}^0)\subset\mathcal{A}(\mathcal{S})}\frac{H(Y_{-n}^0(\varepsilon))-H(Y_{-n}^0(\varepsilon))|H(X_{-n}^0)}{n+1}$$

$$=\lim_{n\to\infty}\sup_{X_{-n}^0\in\mathbb{P}_{n+1},\mathcal{A}(X_{-n}^0)\subset\mathcal{A}(\mathcal{S})}H(Y_{-n}^0(\varepsilon))/(n+1)-H(\varepsilon)=\lim_{n\to\infty}\sup_{X_{-n}^0\in\mathbb{P}_{n+1},\mathcal{A}(X_{-n}^0)\subset\mathcal{A}(\mathcal{S})}H(Y_0(\varepsilon)|Y_{-n}^{-1}(\varepsilon))-H(\varepsilon).$$

Here, $Y_{-n}^0(\varepsilon)$ is the output corresponding to X_{-n}^0. Alternatively

$$C(\varepsilon)=\sup_{X\in\mathbb{P},\mathcal{A}(X)\subset\mathcal{A}(\mathcal{S})}H(Y_\varepsilon)-H(\varepsilon), \tag{3}$$

where Y_ε is the output process corresponding to X. We shall derive in this section an asymptotic formula for capacity of this noisy constrained channel as $\varepsilon\to 0$.

Now let

$$\mathcal{H}_m(\varepsilon)=\sup_{X_{-m}^0\in\mathbb{P}_{m+1},\mathcal{A}(X_{-m}^0)\subset\mathcal{A}(\mathcal{S})}H(Y_0(\varepsilon)|Y_{-m}^{-1}(\varepsilon)),$$

and letting $\mathcal{M}_m$ denote the set of all m-th order binary irreducible Markov chains, we define

$$h_m(\varepsilon)=\sup_{X\in\mathcal{M}_m,\mathcal{A}(X)\subset\mathcal{A}(\mathcal{S})}H(Y_\varepsilon).$$

Now let $C_m(\varepsilon)$ denote the maximum mutual information rate over all m-th order input Markov chains supported on $\mathcal{S}$ transmitted over BSC(ε); then

$$C_m(\varepsilon)=h_m(\varepsilon)-H(\varepsilon), \tag{4}$$

and we have bounds on $C(\varepsilon)$:

$$h_m(\varepsilon)-H(\varepsilon)\le C(\varepsilon)\le\mathcal{H}_m(\varepsilon)-H(\varepsilon). \tag{5}$$

For ε sufficiently small ($\varepsilon<\varepsilon_0$), one may choose $\delta>0$ (here, δ depends on m) such that

$$\mathcal{H}_m(\varepsilon)=\sup_{X_{-m}^0\in\mathbb{P}_{m+1},\mathcal{A}(X_{-m}^0)=\mathcal{A}_{m+1}(\mathcal{S}),X_{-m}^0\in S_{m,\delta}}H(Y_0(\varepsilon)|Y_{-m}^{-1}(\varepsilon)),$$

and

$$h_m(\varepsilon)=\sup_{X\in\mathcal{M}_m,\mathcal{A}(X)=\mathcal{A}(\mathcal{S}),X\in S_{m,\delta}}H(Y_\varepsilon).$$

So from now on we only consider stationary distributions and Markov chains whose allowed words coincide with those of $\mathcal{S}$. Let $\vec{p}$ denote the joint probability vector (indexed by $\mathcal{A}_{m+1}(\mathcal{S})$),

$$\vec{p}=(p(w):w\in\mathcal{A}_{m+1}(\mathcal{S})).$$

In the following, the input and output of a BSC(ε) will be parameterized by $\vec{p}$. More specifically, we use $X_{\vec{p}}$ to denote the binary irreducible Markov chain. Let $Y_{\vec{p},\varepsilon}$ denote the output process obtained by passing $X_{\vec{p}}$ through BSC(ε). Similarly, we use $X^0_{-m}(\vec{p})$ to denote the stationary input distribution X^0_{-m}, and let $Y^0_{-m}(\vec{p},\varepsilon)$ denote the output distribution obtained by passing $X^0_{-m}(\vec{p})$ through BSC(ε).

Lemma 3.1. *$H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p}))$, as a function of $\vec{p}$ in the space of distributions $X^0_{-m}(\vec{p}) \in \mathbb{P}_{m+1}$ with $\mathcal{A}(X^0_{-m}(\vec{p})) = \mathcal{A}_{m+1}(\mathcal{S})$, has a negative definite Hessian matrix.*

Proof. Note that

$$H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p})) = -\sum_{x^0_{-m}\in\mathcal{A}(\mathcal{S})} p(x^0_{-m}) \log p(x_0|x^{-1}_{-m}).$$

For two different probability vectors $\vec{p}$ and $\vec{q}$, consider the convex combination

$$\vec{r}(t) = t\vec{p} + (1-t)\vec{q},$$

where $0 \le t \le 1$. It suffices to prove that $H(X_0(\vec{r}(t))|X^{-1}_{-m}(\vec{r}(t)))$ has a strictly negative second derivative with respect to t. Now consider a single term in $H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p}))$:

$$-(tp(x^0_{-m}) + (1-t)q(x^0_{-m})) \log \frac{tp(x^0_{-m}) + (1-t)q(x^0_{-m})}{tp(x^{-1}_{-m}) + (1-t)q(x^{-1}_{-m})}.$$

Note that for two symbols α and β, if we assume $\alpha'' = 0$ and $\beta'' = 0$, the second order formal derivative of $\alpha \log \frac{\alpha}{\beta}$ can be computed as:

$$\left(\alpha \log \frac{\alpha}{\beta}\right)'' = \left(\frac{\alpha'}{\sqrt{\alpha}} - \sqrt{\alpha}\frac{\beta'}{\beta}\right)^2.$$

It then follows that the second derivative of this term (with respect to t) can be calculated as:

$$-\left(\frac{p(x^0_{-m}) - q(x^0_{-m})}{\sqrt{tp(x^0_{-m}) + (1-t)q(x^0_{-m})}} - \sqrt{tp(x^0_{-m}) + (1-t)q(x^0_{-m})}\frac{p(x^0_{-(m-1)}) - q(x^0_{-(m-1)})}{tp(x^0_{-(m-1)}) + (1-t)q(x^0_{-(m-1)})}\right)^2.$$

That is, the expression above is always non-positive, and is equal to 0 only if

$$\frac{p(x^0_{-m}) - q(x^0_{-m})}{tp(x^0_{-m}) + (1-t)q(x^0_{-m})} = \frac{p(x^0_{-(m-1)}) - q(x^0_{-(m-1)})}{tp(x^0_{-(m-1)}) + (1-t)q(x^0_{-(m-1)})},$$

which is equivalent to

$$p(x_0|x^{-1}_{-m}) = q(x_0|x^{-1}_{-m}).$$

Note that the expression above can't hold true for every x^0_{-m} unless $\vec{p} = \vec{q}$, so we conclude that the second derivative of $H(X_0(\vec{r}(t))|X^{-1}_{-m}(\vec{r}(t)))$ (with respect to t) is strictly negative. Thus $H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p}))$ has a strictly negative definite Hessian as a function of $\vec{p}$.

□

For $m \geq \hat{m}$, over all m-th order Markov chains X with $\mathcal{A}(X) = \mathcal{A}(\mathcal{S})$, $H(X_{\vec{p}})$ is maximized at some unique value $\vec{p}_{max}$ (see [18, 14]), moreover $X_{\vec{p}_{max}}$ is an $\hat{m}$-th order Markov chain. Note that, by abuse of notation, we are using $\vec{p}_{max}$ to mean for any $m \geq \hat{m}$ the stationary distribution on words of length $m+1$ induced by the $\hat{m}$-th order Markov chain.

The same idea shows that over all stationary distributions X^0_{-m} ($m \geq \hat{m}$) with $\mathcal{A}(X^0_{-m}) = \mathcal{A}_{m+1}(\mathcal{S})$, $H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p}))$ is maximized at the same value $\vec{p}_{max}$.

Let $C(\mathcal{S})$ denote the noiseless capacity of the constraint $\mathcal{S}$; then $H(X_{\vec{p}_{max}}) = C(\mathcal{S})$ (see also [18, 14]).

Theorem 3.2. *1. If $m \geq 2\hat{m}(\mathcal{S})$,*

$$\mathcal{H}_m(\varepsilon) = C(\mathcal{S}) + f(\vec{p}_{max})\varepsilon \log(1/\varepsilon) + O(\varepsilon)$$

where $f(\vec{p}_{max}) = f^0_{2\hat{m}}(X^0_{-2\hat{m}}(\vec{p}_{max}))$.

2. If $m \geq \hat{m}(\mathcal{S})$,

$$h_m(\varepsilon) = C(\mathcal{S}) + f(\vec{p}_{max})\varepsilon \log(1/\varepsilon) + O(\varepsilon).$$

Proof. We first prove the statement for $\mathcal{H}_m(\varepsilon)$. As mentioned before, for ε sufficiently small ($\varepsilon < \varepsilon_0$), $\mathcal{H}_m(\varepsilon)$ is achieved by X^0_{-m} with $\mathcal{A}(X^0_{-m}) = \mathcal{A}_{m+1}(\mathcal{S})$; and one may choose δ such that

$$\mathcal{H}_m(\varepsilon) = \sup_{\vec{p}: X^0_{-m}(\vec{p}) \in \mathbb{P}_{m+1}, \mathcal{A}(X^0_{-m}(\vec{p})) = \mathcal{A}(\mathcal{S}), X^0_{-m}(\vec{p}) \in S_{m,\delta}} H(Y_0(\vec{p},\varepsilon)|Y^{-1}_{-m}(\vec{p},\varepsilon)).$$

Below, we assume $\varepsilon < \varepsilon_0$, $\mathcal{A}(X^0_{-m}(\vec{p})) = \mathcal{A}_{m+1}(\mathcal{S})$ and $X^0_{-m}(\vec{p}) \in S_{m,\delta}$.

In Lemma 2.1, we have proved that

$$H(Y_0(\vec{p},\varepsilon)|Y^{-1}_{-m}(\vec{p},\varepsilon)) = H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p})) + f^0_m(X^0_{-m}(\vec{p}))\varepsilon \log(1/\varepsilon) + O(\varepsilon),$$

where one can check $f^0_m(X^0_{-m}(\vec{p})) = f^0_{2\hat{m}}(X^0_{-2\hat{m}}(\vec{p})) = f(\vec{p})$, since $\mathcal{A}(X^0_{-m}(\vec{p})) = \mathcal{A}(\mathcal{S})$ and $m \geq 2\hat{m}$. Moreover, by Remark 2.2, for any $\delta > 0$, $O(\varepsilon)$ is uniform on $S_{m,\delta}$, i.e., there is a constant C (depending on m) such that for all X with $\mathcal{A}(X^0_{-m}) = \mathcal{A}_{m+1}(\mathcal{S})$ and $X^0_{-m}(\vec{p}) \in S_{m,\delta}$,

$$|H(Y_0(\vec{p},\varepsilon)|Y^{-1}_{-m}(\vec{p},\varepsilon)) - H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p})) - f(\vec{p})\varepsilon \log(1/\varepsilon)| \leq C\varepsilon.$$

Let $\vec{q} = \vec{p} - \vec{p}_{max}$. Since $H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p}))$ is maximized at $\vec{p}_{max}$, we can expand $H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p}))$ around $\vec{p}_{max}$:

$$H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p})) = H(X_0(\vec{p}_{max})|X^{-1}_{-m}(\vec{p}_{max})) + \vec{q}^t K_1 \vec{q} + O(|\vec{q}|^3) = H(X_{\vec{p}_{max}}) + \vec{q}^t K_1 \vec{q} + O(|\vec{q}|^3),$$

where K_1 is a negative definite matrix by Lemma 3.1 (the second equality follows from the fact that $X_{\vec{p}_{max}}$ is an $\hat{m}$-th order Markov chain). So for $|\vec{q}|$ sufficiently small, we have

$$H(X_0(\vec{p})|X^{-1}_{-m}(\vec{p})) < H(X_{\vec{p}_{max}}) + (1/2)\vec{q}^t K_1 \vec{q}.$$

Now we expand $f(\vec{p})$ around $\vec{p}_{max}$:

$$f(\vec{p}) = f(\vec{p}_{max}) + K_2 \cdot \vec{q} + O(|\vec{q}|^2).$$

(here, K_2 is a vector of first order partial derivatives). Then, for $|\vec{q}|$ sufficiently small, we have

$$f(\vec{p}) \leq f(\vec{p}_{max}) + 2\sum_j |K_{2,j}||\vec{q}_j|.$$

With a change of coordinates, if necessary, we may assume K_1 is a diagonal matrix with strictly negative diagonal elements $K_{1,j}$. In the following we assume $0 < \varepsilon < \varepsilon_0$. And we may further assume that for some $\ell \geq 1$, $|q_j| > 4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon)$ for $j \leq \ell - 1$, and $|q_j| \leq 4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon)$ for $j \geq \ell$. Then for each $j \leq l-1$, we have $(1/2)K_{1,j}q_j^2 + 2|K_{2,j}||q_j|\varepsilon\log(1/\varepsilon) < 0$. Thus,

$$H(Y_0(\vec{p},\varepsilon)|Y_{-m}^{-1}(\vec{p},\varepsilon)) < H(X_{\vec{p}_{max}}) + f(\vec{p}_{max})\varepsilon\log(1/\varepsilon) + \sum_j((1/2)K_{1,j}q_j^2 + 2|K_{2,j}||q_j|\varepsilon\log(1/\varepsilon)) + C\varepsilon$$

$$< H(X_{\vec{p}_{max}}) + f(\vec{p}_{max})\varepsilon\log(1/\varepsilon) + \sum_{j\geq l}(1/2)K_{1,j}(4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon))^2$$

$$+ \sum_{j\geq l} 2|K_{2,j}|4|K_{2,j}/K_{1,j}|(\varepsilon\log(1/\varepsilon))^2 + C\varepsilon.$$

Since $(\varepsilon\log(1/\varepsilon))^2 = O(\varepsilon)$,

$$H(Y_0(\vec{p},\varepsilon)|Y_{-m}^{-1}(\vec{p},\varepsilon)) < H(X_{\vec{p}_{max}}) + f(\vec{p}_{max})\varepsilon\log(1/\varepsilon) + O(\varepsilon),$$

and since $\mathcal{H}_m(\varepsilon)$ is the sup of the left hand side expression, together with $H(X_{\vec{p}_{max}}) = C(\mathcal{S})$, we have

$$\mathcal{H}_m(\varepsilon) \leq C(\mathcal{S}) + f(\vec{p}_{max})\varepsilon\log(1/\varepsilon) + O(\varepsilon).$$

The reverse inequality follows trivially from the definition of $\mathcal{H}_m(\varepsilon)$.

We now prove the statement for $h_m(\varepsilon)$. First, observe that

$$\mathcal{H}_{2m}(\varepsilon) \geq h_m(\varepsilon) \geq h_{\hat{m}}(\varepsilon) \geq H(X_{\vec{p}_{max}})$$

By part 1, $\mathcal{H}_{2m}(\varepsilon)$ is of the form $C(\mathcal{S}) + f(\vec{p}_{max})\varepsilon\log(1/\varepsilon) + O(\varepsilon)$. By Theorem 2.3, $H(X_{\vec{p}_{max}})$ is of the same form. Thus, $h_m(\varepsilon)$ is also of the same form, as desired. □

Corollary 3.3. *$C_m(\varepsilon)$ ($m \geq \hat{m}(\mathcal{S})$) and $C(\varepsilon)$ are of the form*

$$C(\mathcal{S}) + (f(\vec{p}_{max}) - 1)\varepsilon\log(1/\varepsilon) + O(\varepsilon).$$

Proof. This follows from Theorem 3.2, (4), and (5). □

Example 3.4. Consider a first order input Markov chain X supported on RLL$(1,\infty)$ constraint $\mathcal{S}$, transmitted over BSC(ε) with corresponding output Y, a hidden Markov chain. In this case, $\vec{p}$ takes the form:

$$\vec{p} = (p(00), p(01), p(10)).$$

Note that $\hat{m}(\mathcal{S}) = 1$, and the only sequence $w_{-2}w_{-1}v$, which satisfies the requirement that $w_{-2}w_{-1}$ is allowable in $\mathcal{S}$ and $w_{-2}w_{-1}v$ is disallowable in $\mathcal{S}$, is 011. It then follows that

$$f(\vec{p}) = p(01\bar{1}) + p(0\bar{1}1) + p(\bar{0}11) = \pi_{01}(2 - \pi_{01})/(1 + \pi_{01}),$$

where π_{01} denotes the transition probability from 0 to 1 in X. Thus,

$$H(Y) = H(X) + (\pi_{01}(2 - \pi_{01})/(1 + \pi_{01}))\varepsilon \log(1/\varepsilon) + O(\varepsilon),$$

which was originally proven in [17].

It is well known (see, for example, the general formula on p. 444 of [14]) that the maximum entropy Markov chain on $S = \mathrm{RLL}(1, \infty)$ is defined by the transition matrix:

$$\begin{bmatrix} 1/\lambda & 1/\lambda^2 \\ 1 & 0 \end{bmatrix}$$

and

$$C(\mathcal{S}) = H(X_{\vec{p}_{max}}) = \log \lambda,$$

where λ is the golden mean. Thus, in this case $\pi_{01} = 1/\lambda^2$ and so by Corollary 3.3, we obtain:

$$C(\varepsilon) = \log \lambda - ((2\lambda + 2)/(4\lambda + 3))\varepsilon \log(1/\varepsilon) + O(\varepsilon).$$

Acknowledgements: We are grateful to Wojciech Szpankowski, who raised the problem addressed in this paper and suggested a version of the result in Corollary 3.3. We also thank Erik Ordentlich and Tsachy Weismann whose papers [16], [17] provided much motivation for this work.

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] S. Arimoto. An algorithm for computing the capacity of arbitrary memoryless channels. *IEEE Trans. on Inform. Theory*, vol. IT-18, no. 1, pp. 14-20, 1972.

[3] J. 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] R. E. Blahut. Computation of channel capacity and rate distortion functions. *IEEE Trans. on Inform. Theory*, vol. IT-18, no. 4, pp. 460-473, 1972.

[6] L. Breiman. On achieving channel capacity in finite-memory channels. *Illinois J. Math.* 4 1960 246–252.

[7] J. Chen and P. Siegel. Markov Processes Asymptotically Achieve the Capacity of Finite-State Intersymbol Interference Channels. *Proceedings of the 2004 IEEE International Symposium on Information Theory*, page 346, June 27-July 2, Chicago, U.S.A., 2004.

[8] 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.

[9] 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.

[10] 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.

[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. *Data Compression Conference*, 362-371, Snowbird, 2004.

[13] P. Jacquet, G. Seroussi and W. Szpankowski. Noisy Constrained Capacity. Preprint. December, 2006.

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

[15] B. Marcus, K. Petersen and S. Williams. Transmission rates and factors of Markov chains. *Contemporary Mathematics*, 26:279–294, 1984.

[16] 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.

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

[18] W. Parry. Intrinsic Markov chains. *Trans. Amer. Math. Soc.* 112 (1964), 55-66.

[19] P. Vontobel, A. Kavcic, D. Arnold and Hans-Andrea Loeliger. Capacity of Finite-State Machine Channels. Submitted to *IEEE Transactions on Information Theory*, November 29, 2004.

[20] 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.

[21] 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.

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

[23] O. Zuk, E. Domany, I. Kanter, and M. Aizenman. Taylor series expansions for the entropy rate of hidden Markov Processes. ICC 2006, Istanbul.