Submitted to the Annals of Applied Probability

ASYMPTOTICS OF INPUT-CONSTRAINED BINARY SYMMETRIC CHANNEL CAPACITY

By Guangyue Han, Brian Marcus

University of Hong Kong, University of British Columbia

In this paper, we study the classical problem of noisy constrained capacity in the case of the binary symmetric channel (BSC), namely, the capacity of a BSC whose input is a sequence from a constrained set. Motivated by a result of Ordentlich and Weissman in [26], we derive an asymptotic formula (when the noise parameter is small) for the entropy rate of a hidden Markov chain, observed when a Markov chain passes through a binary symmetric channel. Using this result we establish an asymptotic formula for the capacity of a binary symmetric channel with input process supported on an irreducible finite type constraint, as the noise parameter tends to zero.

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 stationary finite-state Markov chain, then $H(X)$ has a simple analytic form. A *hidden Markov chain* Z 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 Z, the entropy rate $H(Z)$ was studied by Blackwell [6] as early as 1957, where the analysis suggested the intrinsic complexity of $H(Z)$ as a function of the process parameters. He gave an expression for $H(Z)$ in terms of a measure Q, obtained by solving an integral equation dependent on the parameters of the process. The measure is hard to extract from the equation in any explicit way.

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

AMS 2000 subject classifications: Primary 60K99, 94A15; secondary 60J10
Keywords and phrases: hidden Markov chain, entropy, constrained capacity

bound the entropy rate [25] and a variation on the Birch bound [5] was introduced in [10]. 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. [31], and Sharma and Singh [34]. The connection between the entropy rate of a hidden Markov chain and the top Lyapunov exponent of a random matrix product has been observed [12, 17–19]. Results in [4, 29, 30, 32] show that under certain conditions the top Lyapunov exponent of a random matrix product varies analytically as either the underlying Markov process varies analytically or as the matrix entries vary analytically, but not both. In [14], 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 [19], 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 [26], [37], and [15]. In [26] 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.

Let $\mathcal{W}$ denote all the finite length binary words, and $\mathcal{W}_n$ denote all the binary words with length n. For a binary stationary distribution X (with length possibly infinite) and 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 $\mathcal{A}_n(X) = \mathcal{A}(X) \cap \mathcal{W}_n$. Consider a binary irreducible finite type constraint (defined in Section 3; for more details, see [22]) $\mathcal{S}$. 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($\varepsilon$) with input sequences in $\mathcal{S}$, the *noisy constrained capacity* $C(\mathcal{S}, \varepsilon)$ is defined as

$$C(\mathcal{S}, \varepsilon) = \lim_{n \to \infty} \frac{1}{n+1} \sup_{\mathcal{A}(X_{-n}^0) \subseteq \mathcal{A}_{n+1}(\mathcal{S})} I(X_{-n}^0, Z_{-n}^0).$$

Using the approach in Section 12.4 of [13], one can show that

(1.1)
$$C(\mathcal{S},\varepsilon)=\lim_{n\to\infty}\frac{1}{n+1}\sup_{X_{-n}^0\in\mathbb{P}_{n+1},\mathcal{A}(X_{-n}^0)\subseteq\mathcal{A}_{n+1}(\mathcal{S})}I(X_{-n}^0,Z_{-n}^0)=\sup_{X\in\mathbb{P},\mathcal{A}(X)\subseteq\mathcal{A}(\mathcal{S})}I(X;Z).$$

Generally speaking, it is very difficult to calculate the capacity of a generic channel. For a discrete memoryless channel (DMC), the Blahut-Arimoto algorithm ([3, 7]) 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 [28]. As for $C(\mathcal{S},\varepsilon)$, the best results in the literature have been in the form of bounds and numerical simulations based on producing random (and, hopefully, typical) channel output sequences (see, e.g., [36], [35], [2] and references therein). These methods allow for fairly precise numerical approximations of the capacity for given constraints and channel parameters.

This paper is organized as follows. In section 2 we give asymptotic formulas 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, as the noise tends to zero. In section 2.1, we review, from [20], the result when the transition probabilities are strictly positive. In section 2.2, we develop the formula when some transition probabilities are zero, thereby generalizing the result from [26]. In section 3, we consider the binary symmetric channel with input sequences supported on an irreducible finite type constraint, and we derive an asymptotic formula for capacity (again as the noise tends to zero). In section 4, we consider the special case of the (d,k)-RLL constraint, and compute the coefficients of the asymptotic formulas for capacity of the constrained binary symmetric channel.

2. Asymptotics of Entropy Rate. Let $\mathcal{X}=\{0,1\}$. Let X be an m-th order binary irreducible Markov process. The process is defined by the set of conditional probabilities $P(X_t=1|X_{t-m}^{t-1}=a_1^m)$, $a_1^m\in\mathcal{X}^m$. The process is equivalently interpreted as a first-order Markov chain on *states* $s_t=X_{t-m}^{t-1}$, $t>0$ (we assume X_{-m+1}^0 is defined and distributed according to the stationary distribution of the process). Clearly, a transition from a state $u\in\mathcal{X}^m$ to a state $v\in\mathcal{X}^m$ can have positive probability only if u and v satisfy $u_2^m{=}v_1^{m-1}$, in which case we say that (u,v) is an *overlapping* pair.

The *noise process* $E=\{E_i\}_{i\geq1}$ is Bernoulli (binary i.i.d.), independent of X, with $P(E_i{=}1)=\varepsilon$. Let $Z=Z_\varepsilon$ denote the function of the Markov chain $X\times E$ defined by:

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

and

$$Z_i = \overline{X_i} \text{ if } E_i = 1,$$

where $\overline{X_i}$ denotes the Boolean complement of X_i. So, Z is the hidden Markov chain obtained by observing X over BSC(ε).

2.1. *When transition probabilities of X are all positive.* This case is treated in [20]:

THEOREM 2.1. *([20] (Theorem 3)) If the conditional symbol probabilities in a m-th order binary Markov process X satisfy $P(a_{m+1}|a_1^m) > 0$ for all $a_1^{m+1} \in \mathcal{X}^{m+1}$, then the entropy rate of Z for small ε is*

$$H(Z) = H(X) + g(X)\varepsilon + O(\varepsilon^2), \tag{2.1}$$

where, denoting by $\bar{z}_i$ the Boolean complement of z_i, and

$$\check{z}^{2m+1} = z_1 \dots z_m \bar{z}_{m+1} z_{m+2} \dots z_{2m+1},$$

we have

$$\begin{aligned} g(X) &= \sum_{z_1^{2m+1}} P_X(z_1^{2m+1}) \log \frac{P_X(z_1^{2m+1})}{P_X(\check{z}_1^{2m+1})} \\ &= \mathbb{D}\left(P_X(z_1^{2m+1}) || P_X(\check{z}_1^{2m+1})\right). \end{aligned} \tag{2.2}$$

Here, $\mathbb{D}(\cdot||\cdot)$ is the Kullback-Liebler divergence, applied here to distributions on $\mathcal{X}^{2m+1}$ derived from the marginals of X. □

In [20] a complete proof is given for first-order Markov chains, as well as the sketch for the generalization to higher order Markov chains.

We discuss other perspectives on this result.

Let $\widetilde{Z}_i = (Z_i, Z_{i+1}, \dots, Z_{i+m-1})$ and $\widetilde{E}_i = (E_i, \dots, E_{i+m-1})$. An expression for $P(\widetilde{Z}_1^n)$ can be given in terms of a product of matrices, as follows. Here, vectors are of dimension 2^m, and matrices are of dimensions $2^m \times 2^m$. We denote *row* vectors by bold lowercase letters, matrices by bold uppercase letters, and we let $\mathbf{1} = [1, \dots, 1]$; superscript t denotes transposition. Entries in vectors and matrices are indexed by vectors in $\mathcal{X}^m$, according to some fixed order, so that $\mathcal{X}^m = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{2^m}\}$. Let

$$\mathbf{p}_n = [P(\widetilde{Z}_1^n, \widetilde{E}_n{=}\mathbf{a}_1), P(\widetilde{Z}_1^n, \widetilde{E}_n{=}\mathbf{a}_2) \dots P(\widetilde{Z}_1^n, \widetilde{E}_n{=}\mathbf{a}_{2^m})]$$

and let $\mathbf{M}(\widetilde{Z}_n|\widetilde{Z}_{n-1})$ be a $2^m \times 2^m$ matrix defined as follows: if $(\mathbf{e}_{n-1}, \mathbf{e}_n) \in \mathcal{X}^m \times \mathcal{X}^m$ is an overlapping pair, then

$$\mathbf{M}_{\mathbf{e}_{n-1},\mathbf{e}_n}(\widetilde{Z}_n|\widetilde{Z}_{n-1}) = P_X(\widetilde{Z}_n \oplus \mathbf{e}_n | \widetilde{Z}_{n-1} \oplus \mathbf{e}_{n-1}) P(\widetilde{E}_n{=}\mathbf{e}_n). \tag{2.3}$$

All other entries are zero. With these definitions, it follows that

$$P(\widetilde{Z}_1^n) = \mathbf{p}_1 \mathbf{M}(\widetilde{Z}_2|\widetilde{Z}_1)\cdots\mathbf{M}(\widetilde{Z}_n|\widetilde{Z}_{n-1})\mathbf{1}^t. \tag{2.4}$$

Note that at $\varepsilon = 0$, the matrices $\mathbf{M}(\widetilde{Z}_n|\widetilde{Z}_{n-1})$ are all rank one and every column of these matrices is either all positive or all zero. This is exactly the condition needed to apply [15] (Theorem 2.5), which shows that the derivatives of all orders of $H(Z)$ with respect to ε at $\varepsilon = 0$ "stabilize" in the sense that:

$$H^{(n)}(Z)\Big|_{\varepsilon=0} = H^{(n)}_{\lceil (n+1)/2\rceil m}(Z)\Big|_{\varepsilon=0},$$

where the superscript n denotes the n-th order derivative with respect to ε. This means that "in principle" one can compute the derivatives of all orders. Theorem 2.1 does this explicitly for the first derivative.

2.2. *When transition probabilities of X are not necessarily all positive.* Consider a first order Markov chain X with the following probability transition matrix

$$\begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix} \tag{2.5}$$

where $0 \le p \le 1$. This process generates sequences satisfying the $(1,\infty)$ constraint (or, under a different interpretation of rows and columns, the equivalent $(0,1)$ constraint). The output sequence Z, however, will generally not satisfy the constraint. The probability of the constraint-violating sequences at the output of the channel is polynomial in ε, which will generally contribute a term $O(\varepsilon\log\varepsilon)$ to the entropy rate $H(Z)$ when ε is small. This was already observed for the probability transition matrix (2.5) in [26], where it is shown that

$$H(Z) = H(X) - \frac{p(2-p)}{1+p}\varepsilon\log\varepsilon + O(\varepsilon) \tag{2.6}$$

as $\varepsilon \to 0$.

In the following, we shall generalize this result and derive an formula for entropy rate of any hidden Markov chain Z, obtained when passing a Markov chain X of any order m through a BSC(ε). By the Birch bound [5], for $n \ge m$, we have:

$$H(Z_0|Z_{-n+m}^{-1}, X_{-n}^{-n+m-1}, E_{-n}^{-n+m-1}) \le H(Z) \le H(Z_0|Z_{-n}^{-1}). \tag{2.7}$$

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

$$H(Z_0|Z_{-n+m}^{-1}, X_{-n}^{-n+m-1}),$$

since Z_{-n+m}^{0}, if conditioned on X_{-n}^{-n+m-1}, is independent of E_{-n}^{-n+m-1} .

LEMMA 2.2. *For a stationary input distribution $X = X_{-n}^{0} \in \mathbb{P}_{n+1}$ and the corresponding output distribution $Z = Z_{-n}^{0}$ through BSC(ε) and $0 \le k \le n$,*

$$H(Z_0|Z_{-n+k}^{-1}, X_{-n}^{-n+k-1}) = H(X_0|X_{-n}^{-1}) + f_n^k(X_{-n}^0)\varepsilon\log(1/\varepsilon) + g_n^k(X_{-n}^0)\varepsilon + O(\varepsilon^2\log\varepsilon),$$

where $f_n^k(X_{-n}^0)$ and $g_n^k(X_{-n}^0)$ are the functions defined on $\mathbb{P}_{n+1}$ given by (2.8) and (2.9), respectively.

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:

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

$$p_{XZ}(wv) = p(X_{-n}^{-n+k-1} = w_{-n}^{-n+k-1}, Z_{-n+k}^{-1} = w_{-n+k}^{-1}, Z_0 = v),$$

and

$$p_{XZ}(v|w) = p(Z_0 = v|Z_{-n+k}^{-1} = w_{-n+k}^{-1}, X_{-n}^{-n+k-1} = w_{-n}^{-n+k-1}).$$

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 $H(Z_0|Z_{-n+k}^{-1}, X_{-n}^{-n+k-1})$ into five terms:

$$H(Z_0|Z_{-n+k}^{-1}, X_{-n}^{-n+k-1}) = \sum_{wv\in\mathcal{A}(X)} -p_{XZ}(wv)\log(p_{XZ}(v|w))$$

$$+ \sum_{w\in\mathcal{A}(X), wv\notin\mathcal{A}(X)} -p_{XZ}(wv)\log(p_{XZ}(v|w)) + \sum_{p_{XZ}(w)=\Theta(\varepsilon), p_{XZ}(wv)=\Theta(\varepsilon)} -p_{XZ}(wv)\log(p_{XZ}(v|w))$$

$$+ \sum_{p_{XZ}(w)=\Theta(\varepsilon), p_{XZ}(wv)=\Theta(\varepsilon^2)} -p_{XZ}(wv)\log(p_{XZ}(v|w)) + \sum_{p_{XZ}(w)=O(\varepsilon^2)} -p_{XZ}(wv)\log(p_{XZ}(v|w)),$$

here by $\alpha = \Theta(\beta)$, we mean there exist positive constant C_1, C_2 such that $C_1|\beta| \le |\alpha| \le C_2|\beta|$, while by $\alpha = O(\beta)$, we mean there exists positive constant C such that $|\alpha| \le C|\beta|$; note that $p_{XZ}(w) = \Theta(\varepsilon)$ is equivalent to the statement that the Hamming distance of w from $\mathcal{A}(X)$ is 1 and achieved by flipping one of the bits in w_{-n+k}^{-1}.

For the fourth term, we have

$$\sum_{p_{XZ}(w)=\Theta(\varepsilon), p_{XZ}(wv)=\Theta(\varepsilon^2)} -p_{XZ}(wv)\log(p_{XZ}(v|w)) = O(\varepsilon^2\log\varepsilon).$$

For the fifth term, we have

$$\sum_{p_{XZ}(w)=O(\varepsilon^2)} -p_{XZ}(wv)\log(p_{XZ}(v|w)) = \sum_{p_{XZ}(w)=O(\varepsilon^2)} -p_{XZ}(w)\sum_v p_{XZ}(v|w)\log(p_{XZ}(v|w))$$

$$\leq (\log 2)\sum_{p_{XZ}(w)=O(\varepsilon^2)} p_{XZ}(w) = O(\varepsilon^2),$$

where we use $-\sum_v p_{XZ}(v|w)\log(p_{XZ}(v|w)) \leq \log 2$ for any w. We conclude that the sum of the fourth term and the fifth term is $O(\varepsilon^2 \log \varepsilon)$.

For a binary sequence u_{-n}^{-1}, define $h_n^k(u_{-n}^{-1})$ to be:

$$h_n^k(u_{-n}^{-1}) = \sum_{j=1}^{n-k} p_X(u_{-n}^{-j-1}\bar{u}_{-j}u_{-j+1}^{-1}) - (n-k)p_X(u_{-n}^{-1}).$$

Note that with this notation, $h_n^k(w)$ and $h_{n+1}^k(wv)$ can be expressed as derivatives with respect to ε at $\varepsilon = 0$:

$$h_n^k(w) = p'_{XZ}(w)|_{\varepsilon=0}, \qquad h_{n+1}^k(wv) = p'_{XZ}(wv)|_{\varepsilon=0}.$$

Then for the first term, we have

$$\sum_{wv\in\mathcal{A}(X)} -p_{XZ}(wv)\log(p_{XZ}(v|w))$$

$$= -\sum_{wv\in\mathcal{A}(X)} (p_X(wv)+h_{n+1}^k(wv)\varepsilon+O(\varepsilon^2))\log(p_X(v|w)+\frac{h_{n+1}^k(wv)p_X(w)-h_n^k(w)p_X(wv)}{p_X^2(w)}\varepsilon+O(\varepsilon^2))$$

$$= H(X_0|X_{-n}^{-1}) - \sum_{wv\in\mathcal{A}(X)} \left(h_{n+1}^k(wv)\log p_X(v|w) + \frac{h_{n+1}^k(wv)p_X(w)-h_n^k(w)p_X(wv)}{p_X(w)}\right)\varepsilon+O(\varepsilon^2).$$

For the second term, it is easy to check that for $w \in \mathcal{A}(X)$ and $wv \notin \mathcal{A}(X)$, $p_{XZ}(v|w) = \Theta(\varepsilon)$ and

$$p_{XZ}(wv) = h_{n+1}^k(wv)\varepsilon + O(\varepsilon^2),$$

we obtain

$$\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} -p_{XZ}(wv)\log(p_{XZ}(v|w)) = -\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} h_{n+1}^k(wv)\varepsilon\log\frac{h_{n+1}^k(wv)\varepsilon+O(\varepsilon^2)}{p_X(w)}+O(\varepsilon^2)\log\Theta(\varepsilon)$$

$$= \sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} h_{n+1}^k(wv)\varepsilon\log(1/\varepsilon) - \left(\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} h_{n+1}^k(wv)\log\frac{h_{n+1}^k(wv)}{p_X(w)}\right)\varepsilon+O(\varepsilon^2\log\varepsilon).$$

For the third term, we have

$$\sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon)} -p_{XZ}(wv)\log(p_{XZ}(v|w))$$

$$= -\sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon)} (h^k_{n+1}(wv)\varepsilon + O(\varepsilon^2))\log\left(\frac{h^k_{n+1}(wv)}{h^k_n(w)} + O(\varepsilon)\right)$$

$$= -\left(\sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon)} h^k_{n+1}(wv)\log\left(\frac{h^k_{n+1}(wv)}{h^k_n(w)}\right)\right)\varepsilon + O(\varepsilon^2).$$

In summary, $H(Z_0|Z_{-n+k}^{-1}, X_{-n}^{-n+k-1})$ can be rewritten as

$$H(Z_0|Z_{-n+k}^{-1}, X_{-n}^{-n+k-1}) = H(X_0|X_{-n}^{-1}) + f^k_n(X^0_{-n})\varepsilon\log(1/\varepsilon) + g^k_n(X^0_{-n})\varepsilon + O(\varepsilon^2\log\varepsilon),$$

where

$$f^k_n(X^0_{-n}) = \sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} h^k_{n+1}(wv), \tag{2.8}$$

and

$$g^k_n(X^0_{-n}) = -\sum_{wv\in\mathcal{A}(X)}\left(h^k_{n+1}(wv)\log p_X(v|w) + \frac{h^k_{n+1}(wv)p_X(w) - h^k_n(w)p_X(wv)}{p_X(w)}\right)$$

$$-\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} h^k_{n+1}(wv)\log\frac{h^k_{n+1}(wv)}{p_X(w)} - \sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon)} h^k_{n+1}(wv)\log\left(\frac{h^k_{n+1}(wv)}{h^k_n(w)}\right). \tag{2.9}$$

□

REMARK 2.3. For any $\delta > 0$ and fixed n, the constant in $O(\varepsilon^2\log\varepsilon)$ in Lemma 2.2 can be chosen uniformly on $S_{n,\delta}$, where $S_{n,\delta}$ denotes the collection of stationary distributions $X \in \mathbb{P}_{n+1}$, such that for all $w \in \mathcal{A}_{n+1}(X)$, $p(w) \geq \delta$.

THEOREM 2.4. *For an m-th order Markov chain X passing through a BSC(ε), with Z as the output hidden Markov chain,*

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

where $f(X) = f^0_{2m}(X^0_{-2m}) = f^m_{2m}(X^0_{-2m})$ and $g(X) = g^0_{3m}(X^0_{-3m}) = g^m_{3m}(X^0_{-3m})$.

PROOF. We apply Lemma 2.2 to the Birch upper and lower bounds (eqn. (2.7)) of $H(Z)$. For the upper bound, $k = 0$, we have, for all n,

$$H(Z_0|Z_{-n}^{-1}) = H(X_0|X_{-n}^{-1}) + f_n^0(X_{-n}^0)\varepsilon \log(1/\varepsilon) + g_n^0(X_{-n}^0)\varepsilon + O(\varepsilon^2 \log \varepsilon).$$

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

$$H(Z_0|Z_{-n+m}^{-1}, X_{-n}^{-n+m-1}) = H(X_0|X_{-n}^{-1}) + f_n^m(X_{-n}^0)\varepsilon log(1/\varepsilon) + g_n^m(X_{-n}^0)\varepsilon + O(\varepsilon^2 \log \varepsilon).$$

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 $n \geq 2m$,

$$f_n^m(X_{-n}^0) = f_n^0(X_{-n}^0) = f_{2m}^0(X_{-2m}^0) = f_{2m}^m(X_{-2m}^0).$$

Now consider $g_n^k(X_{-n}^0)$. When $0 \leq k \leq m$, we have (for detailed derivation of (2.10)-(2.12), see Appendix A)

$$\text{if } wv \in \mathcal{A}(X), p_X(v|w) = p_X(v|w_{-m}^{-1}), \text{ for } n \geq m, \tag{2.10}$$

$$\text{if } w \in \mathcal{A}(X), wv \notin \mathcal{A}(X), \frac{h_{n+1}^k(wv)}{p_X(w)} \text{ is constant (as function of } n \text{ and } k) \text{ for } n \geq 2m, 0 \leq k \leq m, \tag{2.11}$$

$$\text{if } p_{XZ}(w) = \Theta(\varepsilon), p_{XZ}(wv) = \Theta(\varepsilon), \frac{h_{n+1}^k(wv)}{h_n^k(w)} \text{ is constant for } n \geq 3m, 0 \leq k \leq m. \tag{2.12}$$

It then follows from the "stabilizing" property of the quantities above that (for detailed derivation of (2.13)-(2.15), see Appendix A)

$$\sum_{wv \in \mathcal{A}(X)} \frac{h_{n+1}^k(wv)p_X(w) - h_n^k(w)p_X(wv)}{p_X(w)} \text{ is constant (as a function of } n) \text{ for } n \geq 2m, 0 \leq k \leq m, \tag{2.13}$$

$$\sum_{w \in \mathcal{A}(X), wv \notin \mathcal{A}(X)} h_{n+1}^k(wv) \log \frac{h_{n+1}^k(wv)}{p_X(w)} \text{ is constant for } n \geq 2m, 0 \leq k \leq m, \tag{2.14}$$

and

(2.15)

$$\sum_{wv\in\mathcal{A}(X)} h^k_{n+1}(wv)\log p_X(v|w)+\sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon)} h^k_{n+1}(wv)\log\frac{h^k_{n+1}(wv)}{h^k_n(w)} \text{ is constant for } n\geq 3m, 0\leq k\leq m.$$

Consequently, we have

$$g^m_n(X^0_{-n}) = g^0_n(X^0_{-n}) = g^0_{3m}(X^0_{-3m}) = g^m_{3m}(X^0_{-3m}).$$

Let $f(X) = f^0_{2m}(X^0_{-2m})$ and $g(X) = g^0_{3m}(X^0_{-3m})$, then the theorem follows. □

REMARK 2.5. Note that this result applies in particular to the case when the transition probabilities are all positive; thus in this case the formula should reduce to that of Theorem 2.1, i.e., f is zero and g reduces to the Kullback-Leibler divergence expression in Theorem 2.1.

3. Asymptotics of Capacity. A *finite type* constraint [22] $\mathcal{S}$ is defined by a finite set (denoted by $\mathcal{F}$) of forbidden words. A prominent example is the (d,k)-RLL constraint $\mathcal{S}(d,k)$ which forbids any sequence with fewer than d or more than k consecutive zeros in between two 1's. Let $\mathcal{A}(\mathcal{S})$ denote the set of all allowable words in $\mathcal{S}$, and $\mathcal{A}_n(\mathcal{S}) = \mathcal{A}(\mathcal{S})\cap\mathcal{W}_n$. A finite type constraint is *irreducible* if for any $u,v\in\mathcal{A}(\mathcal{S})$, there is a $w\in\mathcal{A}(\mathcal{S})$ such that $uwv\in\mathcal{A}(\mathcal{S})$.

Consider a binary irreducible finite type constraint $\mathcal{S}$ defined by $\mathcal{F}$, which consists of forbidden words with length $\hat{m}+1$. In general, 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 $\mathcal{S}$. Note that the order of $\mathcal{S}(d,k)$ is k.

The conventional BSC channel capacity (with unconstrained binary input sequences) $C(\varepsilon)$ can be easily computed as follows:

$$C(\varepsilon) = 1 - H(\varepsilon),$$

where $H(\varepsilon) = -\varepsilon\log\varepsilon - (1-\varepsilon)\log(1-\varepsilon)$. For a constrained BSC($\varepsilon$) with input sequences X^0_{-n} in $\mathcal{S}$ and with the corresponding output $Z^0_{-n}(\varepsilon)$, the capacity $C(\mathcal{S},\varepsilon)$ can be written as:

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

$$= \lim_{n\to\infty} \sup_{X^0_{-n}\in\mathbb{P}_{n+1},\mathcal{A}(X^0_{-n})\subseteq\mathcal{A}_{n+1}(\mathcal{S})} H(Z^0_{-n}(\varepsilon))/(n+1) - H(\varepsilon)$$

$$= \lim_{n\to\infty} \sup_{X^0_{-n}\in\mathbb{P}_{n+1},\mathcal{A}(X^0_{-n})\subseteq\mathcal{A}_{n+1}(\mathcal{S})} H(Z_0(\varepsilon)|Z^{-1}_{-n}(\varepsilon)) - H(\varepsilon),$$

where the last equality follows from the the fact that

$$H(Z^0_{-n}(\varepsilon)) = \sum_{j=0}^{n} H(Z_0(\varepsilon)|Z^{-1}_{-j}(\varepsilon)),$$

and

$$H(Z_0(\varepsilon)|Z^{-1}_{-j_1}(\varepsilon)) \geq H(Z_0(\varepsilon)|Z^{-1}_{-j_2}(\varepsilon)) \text{ for } j_1 \leq j_2.$$

Alternatively

$$C(\mathcal{S},\varepsilon) = \sup_{X\in\mathbb{P},\mathcal{A}(X)\subseteq\mathcal{A}(\mathcal{S})} H(Z_\varepsilon) - H(\varepsilon), \tag{3.1}$$

where Z_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

$$\mathcal{H}_n(\mathcal{S},\varepsilon) = \sup_{X^0_{-n}\in\mathbb{P}_{n+1},\mathcal{A}(X^0_{-n})\subseteq\mathcal{A}_{n+1}(\mathcal{S})} H(Z_0(\varepsilon)|Z^{-1}_{-n}(\varepsilon)),$$

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

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

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

$$C_m(\mathcal{S},\varepsilon) = h_m(\mathcal{S},\varepsilon) - H(\varepsilon), \tag{3.2}$$

and we have bounds on $C(\mathcal{S},\varepsilon)$:

$$h_m(\mathcal{S},\varepsilon) - H(\varepsilon) \leq C(\mathcal{S},\varepsilon) \leq \mathcal{H}_n(\mathcal{S},\varepsilon) - H(\varepsilon). \tag{3.3}$$

Noting that

$$\sup_{X^0_{-n}\in\mathbb{P}_{n+1},\mathcal{A}(X^0_{-n})\subsetneq\mathcal{A}_{n+1}(\mathcal{S})} H(X_0|X^{-1}_{-n}) < \sup_{X^0_{-n}\in\mathbb{P}_{n+1},\mathcal{A}(X^0_{-n})=\mathcal{A}_{n+1}(\mathcal{S})} H(X_0|X^{-1}_{-n}),$$

$$\sup_{X\in\mathcal{M}_m,\mathcal{A}(X)\subsetneq\mathcal{A}(\mathcal{S})} H(X) < \sup_{X\in\mathcal{M}_m,\mathcal{A}(X)=\mathcal{A}(\mathcal{S})} H(X),$$

and $H(Z_0(\varepsilon)|Z_{-n}^{-1}(\varepsilon))$, $H(Z_\varepsilon)$ are continuous at $\varepsilon = 0$, we conclude that for ε sufficiently small ($\varepsilon < \varepsilon_0$), one may choose $\delta > 0$ (here, δ depends on n and m) such that

$$\mathcal{H}_n(\mathcal{S},\varepsilon) = \sup_{X_{-n}^0\in\mathbb{P}_{n+1},\mathcal{A}(X_{-n}^0)=\mathcal{A}_{n+1}(\mathcal{S}),X_{-n}^0\in S_{n,\delta}} H(Z_0(\varepsilon)|Z_{-n}^{-1}(\varepsilon)),$$

and

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

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 $\mathcal{A}_{n+1}(\mathcal{S})$),

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

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 $Z_{\vec{p},\varepsilon}$ denote the output process obtained by passing $X_{\vec{p}}$ through BSC(ε). Similarly, we use $X_{-n}^0(\vec{p})$ to denote the stationary input distribution X_{-n}^0, and let $Z_{-n}^0(\vec{p},\varepsilon)$ denote the output distribution obtained by passing $X_{-n}^0(\vec{p})$ through BSC(ε).

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

PROOF. Note that

$$H(X_0(\vec{p})|X_{-n}^{-1}(\vec{p})) = -\sum_{x_{-n}^0\in\mathcal{A}(\mathcal{S})} p(x_{-n}^0)\log p(x_0|x_{-n}^{-1}).$$

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 $H(X_0(\vec{r}(t))|X_{-n}^{-1}(\vec{r}(t)))$ has a strictly negative second derivative with respect to t. Now consider a single term in $H(X_0(\vec{p})|X_{-n}^{-1}(\vec{p}))$:

$$-(tp(x_{-n}^0) + (1-t)q(x_{-n}^0))\log\frac{tp(x_{-n}^0) + (1-t)q(x_{-n}^0)}{tp(x_{-n}^{-1}) + (1-t)q(x_{-n}^{-1})}.$$

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:

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

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

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

which is equivalent to

$$p(x_0|x^{-1}_{-n}) = q(x_0|x^{-1}_{-n}). \tag{3.4}$$

Since $\mathcal{S}$ is an irreducible finite type constraint and $\mathcal{A}(X^0_{-n}(\vec{p})) = \mathcal{A}(X^0_{-n}(\vec{q})) = \mathcal{A}_{n+1}(\mathcal{S})$, the expression (3.4) can't hold true for every x^0_{-n} unless $\vec{p} = \vec{q}$. So we conclude that the second derivative of $H(X_0(\vec{r}(t))|X^{-1}_{-n}(\vec{r}(t)))$ (with respect to t) is strictly negative. Thus $H(X_0(\vec{p})|X^{-1}_{-n}(\vec{p}))$ 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 $\vec{p}^{max}_m$ (see [22, 27]). Moreover $X_{\vec{p}^{max}_m}$ doesn't depend on m and is an $\hat{m}$-th order Markov chain, so we will drop the subscript m and use $X_{\vec{p}^{max}}$ instead to denote $X_{\vec{p}^{max}_m}$ for any $m \geq \hat{m}$. The same idea shows that over all stationary distributions X^0_{-n} ($n \geq \hat{m}$) with $\mathcal{A}(X^0_{-n}) = \mathcal{A}_{n+1}(\mathcal{S})$, $H(X_0(\vec{p})|X^{-1}_{-n}(\vec{p}))$ is maximized at $\vec{p}^{max}_n$, which corresponds to $X_{\vec{p}^{max}}$ as well.

Let $C(\mathcal{S}) = C(\mathcal{S}, 0)$ denote the *noiseless capacity* of the constraint $\mathcal{S}$. This quantity has been extensively studied, and several interpretations and methods for its explicit derivation are known (see, e.g., [24] and extensive bibliography therein). It is well known that $C(\mathcal{S}) = H(X_{\vec{p}^{max}})$ (see [22, 27]).

THEOREM 3.2. *1. If $n \geq 3\hat{m}(\mathcal{S})$,*

$$\mathcal{H}_n(\mathcal{S}, \varepsilon) = C(\mathcal{S}) + f(X_{\vec{p}^{max}})\varepsilon \log(1/\varepsilon) + g(X_{\vec{p}^{max}})\varepsilon + O(\varepsilon^2 \log^2 \varepsilon).$$

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

$$h_m(\mathcal{S},\varepsilon) = C(\mathcal{S}) + f(X_{\vec{p}^{max}})\varepsilon\log(1/\varepsilon) + g(X_{\vec{p}^{max}})\varepsilon + O(\varepsilon^2\log^2\varepsilon).$$

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

$$\mathcal{H}_n(\mathcal{S},\varepsilon) = \sup_{\vec{p}: X^0_{-n}(\vec{p})\in\mathbb{P}_{n+1}, \mathcal{A}(X^0_{-n}(\vec{p}))=\mathcal{A}_{n+1}(\mathcal{S}), X^0_{-n}(\vec{p})\in S_{n,\delta}} H(Z_0(\vec{p},\varepsilon)|Z^{-1}_{-n}(\vec{p},\varepsilon)).$$

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

For a distribution p on words with length $n+1$, define

$$f_n(\vec{p}) = f^0_n(X^0_{-n}(\vec{p})),$$

and

$$g_n(\vec{p}) = g^0_n(X^0_{-n}(\vec{p})).$$

In Lemma 2.2, we have proved that

$$H(Z_0(\vec{p},\varepsilon)|Z^{-1}_{-n}(\vec{p},\varepsilon)) = H(X_0(\vec{p})|X^{-1}_{-n}(\vec{p})) + f_n(\vec{p})\varepsilon\log(1/\varepsilon) + g_n(\vec{p})\varepsilon + O(\varepsilon^2\log\varepsilon).$$

Moreover, by Remark 2.3, for any $\delta > 0$, $O(\varepsilon^2\log\varepsilon)$ is uniform on $S_{n,\delta}$, i.e., there is a constant C (depending on n) such that for all X with $\mathcal{A}(X^0_{-n}) = \mathcal{A}_{n+1}(\mathcal{S})$ and $X^0_{-n}(\vec{p}) \in S_{n,\delta}$,

$$|H(Z_0(\vec{p},\varepsilon)|Z^{-1}_{-n}(\vec{p},\varepsilon)) - H(X_0(\vec{p})|X^{-1}_{-n}(\vec{p})) - f_n(\vec{p})\varepsilon\log(1/\varepsilon) - g_n(\vec{p})\varepsilon| \leq C\varepsilon^2\log\varepsilon.$$

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

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

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

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

Now we expand $f_n(\vec{p})$ and $g_n(\vec{p})$ around $\vec{p}^{max}_n$:

$$f_n(\vec{p}) = f_n(\vec{p}^{max}_n) + K_2\cdot\vec{q} + O(|\vec{q}|^2),$$

$$g_n(\vec{p}) = g_n(\vec{p}^{max}_n) + K_3\cdot\vec{q} + O(|\vec{q}|^2),$$

(here, K_2 and K_3 are vectors of first order partial derivatives). Then, for $|\vec{q}|$ sufficiently small, we have

$$f_n(\vec{p}) \le f_n(\vec{p}_n^{max}) + 2\sum_j |K_{2,j}||\vec{q}_j|,$$

$$g_n(\vec{p}) \le g_n(\vec{p}_n^{max}) + 2\sum_j |K_{3,j}||\vec{q}_j|.$$

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 $\ell \ge 1$, $|q_j| > 4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon) + 4|K_{3,j}/K_{1,j}|\varepsilon$ for $j \le \ell - 1$, and $|q_j| \le 4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon)+4|K_{3,j}/K_{1,j}|\varepsilon$ for $j \ge \ell$. Then for each $j \le l-1$, we have $(1/2)K_{1,j}q_j^2 + 2|K_{2,j}||q_j|\varepsilon\log(1/\varepsilon) + 2|K_{3,j}||\vec{q}_j|\varepsilon < 0$. Thus,

$$H(Z_0(\vec{p},\varepsilon)|Z_{-n}^{-1}(\vec{p},\varepsilon)) < H(X_{\vec{p}^{max}}) + f_n(\vec{p}_n^{max})\varepsilon\log(1/\varepsilon) + g_n(\vec{p}_n^{max})\varepsilon$$

$$+\sum_j((1/2)K_{1,j}q_j^2 + 2|K_{2,j}||q_j|\varepsilon\log(1/\varepsilon) + 2|K_{3,j}||\vec{q}_j|\varepsilon) + C\varepsilon^2\log\varepsilon$$

$$< H(X_{\vec{p}^{max}})+f_n(\vec{p}_n^{max})\varepsilon\log(1/\varepsilon)+g_n(\vec{p}_n^{max})\varepsilon+\sum_{j\ge l}(1/2)K_{1,j}(4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon)+4|K_{3,j}/K_{1,j}|\varepsilon)^2$$

$$+\sum_{j\ge l} 2|K_{2,j}|(4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon) + 4|K_{3,j}/K_{1,j}|\varepsilon)\varepsilon\log(1/\varepsilon)$$

$$+\sum_{j\ge l} 2|K_{3,j}|(4|K_{2,j}/K_{1,j}|\varepsilon\log(1/\varepsilon) + 4|K_{3,j}/K_{1,j}|\varepsilon)\varepsilon + C\varepsilon^2\log\varepsilon.$$

Collecting terms, we eventually reach:

$$H(Z_0(\vec{p},\varepsilon)|Z_{-n}^{-1}(\vec{p},\varepsilon)) < H(X_{\vec{p}^{max}})+f_n(\vec{p}_n^{max})\varepsilon\log(1/\varepsilon)+g_n(\vec{p}_n^{max})\varepsilon+O(\varepsilon^2\log^2\varepsilon),$$

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

$$\mathcal{H}_n(\mathcal{S},\varepsilon) \le C(\mathcal{S}) + f_n(\vec{p}_n^{max})\varepsilon\log(1/\varepsilon) + g_n(\vec{p}_n^{max})\varepsilon + O(\varepsilon^2\log^2\varepsilon).$$

As discussed in Theorem 2.4, we have

$$f_n(\vec{p}_n^{max}) = f(X_{\vec{p}^{max}}), \quad n \ge 2\hat{m}, \tag{3.5}$$

and

$$g_n(\vec{p}_n^{max}) = g(X_{\vec{p}^{max}}), \quad n \ge 3\hat{m}. \tag{3.6}$$

So eventually we reach

$$\mathcal{H}_n(\mathcal{S},\varepsilon) \le C(\mathcal{S}) + f(X_{\vec{p}^{max}})\varepsilon\log(1/\varepsilon) + g(X_{\vec{p}^{max}})\varepsilon + O(\varepsilon^2\log^2\varepsilon).$$

The reverse inequality follows trivially from the definition of $\mathcal{H}_n(\varepsilon)$. We now prove the statement for $h_m(\mathcal{S},\varepsilon)$. First, observe that

$$\mathcal{H}_{3m}(\mathcal{S},\varepsilon) \ge h_m(\mathcal{S},\varepsilon) \ge h_{\hat{m}}(\mathcal{S},\varepsilon) \ge H(X_{\vec{p}^{max}})$$

By part 1, $\mathcal{H}_{3m}(\mathcal{S},\varepsilon)$ is of the form $C(\mathcal{S})+f(X_{\vec{p}^{max}})\varepsilon\log(1/\varepsilon)+g(X_{\vec{p}^{max}})\varepsilon+O(\varepsilon^2\log^2\varepsilon)$. By Theorem 2.4, $H(X_{\vec{p}^{max}})$ is of the same form. Thus, $h_m(\mathcal{S},\varepsilon)$ is also of the same form, as desired.

□

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

$$C(\mathcal{S}) + (f(X_{\vec{p}^{max}})-1)\varepsilon\log(1/\varepsilon) + (g(X_{\vec{p}^{max}})-1)\varepsilon + O(\varepsilon^2\log^2\varepsilon).$$

PROOF. This follows from Theorem 3.2, (3.2) and (3.3), and the fact that

$$H(\varepsilon) = \varepsilon\log 1/\varepsilon + (1-\varepsilon)\log 1/(1-\varepsilon) = \varepsilon\log 1/\varepsilon + \varepsilon + O(\varepsilon^2).$$

□

4. Binary Symmetric Channel with (d,k)-RLL Constrained Input. We now apply the results of the preceding section to compute asymptotics for the the noisy constrained BSC channel with inputs restricted to the RLL constraint $\mathcal{S}(d,k)$. Expressions (2.8) and (2.9) allow us to explicitly compute $f(\vec{p}^{max})$ and $g(\vec{p}^{max})$. In this section, as an example, we derive the explicit expression for $f(\vec{p}^{max})$, omitting the computation of $g(\vec{p}^{max})$ due to tedious derivation. We remark that for a BSC(ε) with (d,k)-RLL constrained input, similar expressions have been independently obtained in [21].

It is first shown in [21] that in the case $k \le 2d$, in fact for any Markov chain X, of any order, supported on $\mathcal{S}(d,k)$, $f(X)=1$, and so, in this case, $C(\mathcal{S}(d,k),\varepsilon) = C(\mathcal{S}(d,k)) + O(\varepsilon)$, i.e., the noisy constrained capacity differs from the noiseless capacity by $O(\varepsilon)$, rather than $O(\varepsilon\log\varepsilon)$. In the following, we take a look at this using different approach. For this, first note that for any d,k, $f(X)$ takes the form:

(4.1)
$$f(X) = \sum_{l_1+l_2\le k-1, 0\le l_2\le d-1, l_1\ge d} p(10^{l_1+l_2+1}1) + \sum_{l_1+l_2=k, l_1\ge d} p(10^{l_1}10^{l_2}) + \sum_{1\le l\le d} p(10^l).$$

Now, when $k \le 2d$,

$$\sum_{l_1+l_2=k, l_1 \ge d} p(10^{l_1}10^{l_2}) = \sum_{d \le l_1 \le k} p(10^{l_1}1) = p(1),$$

and

$$\sum_{l_1+l_2 \le k-1, 0 \le l_2 \le d-1, l_1 \ge d} p(10^{l_1+l_2+1}1) = p(10^{d+1}) + p(10^{d+2}) + \cdots + p(10^k).$$

So

$$f(X) = p(1) + p(10) + \cdots + p(10^d) + p(10^{d+1}) + \cdots + p(10^k) = 1,$$

as desired.

Now we consider the general RLL constraint $\mathcal{S}(d,k)$. By Corollary 3.3, we have

(4.2)

$$C(\mathcal{S}(d,k),\varepsilon) = C(\mathcal{S}(d,k)) + (f(p_{\max}) - 1)\varepsilon \log 1/\varepsilon + (g(p_{\max}) - 1)\varepsilon + o(\varepsilon),$$

where $C(\mathcal{S}(d,k))$ is the capacity of noiseless (d,k)-RLL system.

For any irreducible finite type constraint, the noiseless capacity and Markov process of maximal entropy rate can be computed in various ways (which all go back to Shannon; see [24] or [22] (p. 444)). Let A denote the adjacency matrix of the standard graph presentation, with $k+1$ states, of $\mathcal{S}(d,k)$. Let ρ denote the reciprocal of the largest eigenvalue. One can write $C(\mathcal{S}(d,k)) = -\log \rho_0$, and in this case ρ_0 is the real root of

$$\sum_{\ell=d}^{k} \rho_0^{\ell+1} = 1. \tag{4.3}$$

In the following we compute $f(p_{\max})$ explicitly in terms of ρ_0. Let $\vec{w} = (w_0, w_1, \cdots, w_k)$ and $\vec{v} = (v_0, v_1, \cdots, v_k)$ denote the left and right eigenvectors of A. Assume that $\vec{w}$ and $\vec{v}$ are scaled such that $\vec{w} \cdot \vec{v} = 1$. Then one checks that with $X = X_{\vec{p}^{max}}$,

$$p(1) = w_0 v_0 = \frac{1}{(k+1) - \sum_{j=d+1}^{k} \sum_{l=0}^{j-d-1} 1/\rho_0^{l-j}},$$

$$p(10^{l_1+l_2+1}1) = p(1)\rho_0^{l_1+l_2+2}, p(10^k1) = p(1)\rho_0^{k+1},$$

$$p(10^{l_1}10^{l_2}) = p(10^{l_1}10^{l_2}1) + p(10^{l_1}10^{l_2+1}1) + \cdots + p(10^{l_1}10^k1)$$

$$= p(1)\rho_0^{l_1+l_2+2}(1+\rho_0+\cdots+\rho_0^{k-l_2}) = p(1)\rho_0^{l_1+l_2+2}\frac{1-\rho_0^{k-l_2+1}}{1-\rho_0}$$

and

$$p(10^l) = p(10^l1) + p(10^{l+1}1) + \cdots + p(10^k1)$$

$$= p(1)\rho_0^{l+1}(1+\rho_0+\cdots+\rho_0^{k-l}) = p(1)\rho_0^{l+1}\frac{1-\rho_0^{k-l+1}}{1-\rho_0}.$$

So we obtain an explicit expression:

$$f(\vec{p}^{max}) = \sum_{l_1+l_2\le k-1, 0\le l_2\le d-1, l_1\ge d} p(10^{l_1+l_2+1}1) + \Big(\sum_{l_1=k, l_2=0} + \sum_{l_1+l_2=k, k-1\ge l_1\ge d}\Big) p(10^{l_1}10^{l_2}) + \sum_{1\le l\le d} p(10^l)$$

$$= p(1)\rho_0^{k+1} + \sum_{l_1+l_2\le k-1, 0\le l_2\le d-1, l_1\ge d} p(1)\rho_0^{l_1+l_2+2}$$

$$+ \sum_{l_1+l_2=k, k-1\ge l_1\ge d} p(1)\rho_0^{l_1+l_2+2}\frac{1-\rho_0^{k-l_2+1}}{1-\rho_0} + \sum_{1\le l\le d} p(1)\rho_0^{l+1}\frac{1-\rho_0^{k-l+1}}{1-\rho_0}.$$

We remark similar computations apply to the calculation of g, which takes more complicated form.

EXAMPLE 4.1. Consider a first order input Markov chain X supported on RLL$(1,\infty)$ constraint $\mathcal{S}$, transmitted over BSC(ε) with corresponding output Z, 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. Tedious computations also leads to

$$g(\vec{p}) = (1+\pi_{01})^{-1}(2\pi_{01}-\pi_{01}^2-2\pi_{01}^3+3\pi_{01}^4-\pi_{01}^5+(-2\pi_{01}+4\pi_{01}^3-2\pi_{01}^4)\ln(2)$$

$$+(-1+3\pi_{01}-\pi_{01}^2-2\pi_{01}^3+5\pi_{01}^4-3\pi_{01}^5)\ln(\pi_{01})+(2-6\pi_{01}+7\pi_{01}^3-8\pi_{01}^4+3\pi_{01}^5)\ln(1-\pi_{01})$$

$$+(2\pi_{01}+\pi_{01}^2-3\pi_{01}^3+\pi_{01}^4)\ln(2-\pi_{01}))$$

Thus,

$$H(Z) = H(X) + (\pi_{01}(2-\pi_{01})/(1+\pi_{01}))\varepsilon\log(1/\varepsilon) + (g(\vec{p})-1)\varepsilon + O(\varepsilon^2\log\varepsilon).$$

This asymptotic formula was originally proven in [26], with the less precise result that replaces $(g(\vec{p})-1)\varepsilon + o(\varepsilon)$ by $O(\varepsilon)$.

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:

$$C(\varepsilon) = \log\lambda - ((2\lambda+2)/(4\lambda+3))\varepsilon\log(1/\varepsilon) + (g(\varepsilon)|_{\pi_{01}} - 1)\varepsilon + O(\varepsilon^2\log^2\varepsilon).$$

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.

Appendices.

APPENDIX A: DERIVATIONS

We first prove (2.10)-(2.12).

- (2.10) follows trivially from the fact that X is an m-th order Markov chain.
- Now consider (2.11). For $w \in \mathcal{A}(X)$ and $wv \notin \mathcal{A}(X)$,

$$h^k_{n+1}(wv) = \sum_{j=1}^{n-k} p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v) + p_X(w_{-n}^{-1}\bar{v}) = \sum_{j=1}^{m} p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v) + p_X(w_{-n}^{-1}\bar{v}).$$

So

$$\frac{h^k_{n+1}(wv)}{p_X(w)} = \frac{\sum_{j=1}^m p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v) + p_X(w_{-n}^{-1}\bar{v})}{p_X(w_{-n}^{-1})}$$

$$= \frac{(\sum_{j=1}^m p_X(w_{-m}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v|w_{-2m}^{-m-1}) + p_X(w_{-m}^{-1}\bar{v}|w_{-2m}^{-m-1}))p_X(w_{-n}^{-m-1})}{p_X(w_{-m}^{-1}|w_{-2m}^{-m-1})p_X(w_{-n}^{-m-1})}$$

$$= \frac{\sum_{j=1}^m p_X(w_{-2m}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v) + p_X(w_{-2m}^{-1}\bar{v})}{p_X(w_{-2m}^{-1})}.$$

- For (2.12), there are two cases. If $p_X(w_{-n}^{-m-1})=0$,

$$\frac{h_{n+1}^k(wv)}{h_n^k(w)}=\frac{\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)}{\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1})}=\frac{\sum_{j=m+1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)}{\sum_{j=m+1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1})}=p_X(v|w_{-m}^{-1}).$$

If $p_X(w_{-n}^{-m-1})>0$,

$$\frac{h_{n+1}^k(wv)}{h_n^k(w)}=\frac{\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)}{\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1})}=\frac{\sum_{j=1}^{2m}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)}{\sum_{j=1}^{2m}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1})}=\frac{\sum_{j=1}^{2m}p_X(w_{-3m}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)}{\sum_{j=1}^{2m}p_X(w_{-3m}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1})}.$$

Using (2.10)-(2.12), we now proceed to prove (2.13)-(2.15).

- For (2.13), we have

$$\sum_{wv\in\mathcal{A}(X)}\frac{h_{n+1}^k(wv)p_X(w)-h_n^k(w)p_X(wv)}{p_X(w)}=\sum_{wv\in\mathcal{A}(X)}h_{n+1}^k(wv)-\sum_{wv\in\mathcal{A}(X)}h_n^k(w)p_X(v|w_{-m}^{-1})$$

$$=\sum_{wv\in\mathcal{A}(X)}(\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)+p_X(w_{-n}^{-1}\bar{v}))-(n+1-k)\sum_{wv\in\mathcal{A}(X)}p_X(wv)$$

$$-\sum_{wv\in\mathcal{A}(X)}\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1})p_X(v|w_{-m}^{-1})+(n-k)\sum_{wv\in\mathcal{A}(X)}p_X(wv)$$

$$=\sum_{wv\in\mathcal{A}(X)}\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)-\sum_{w\in\mathcal{A}(X)}\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1})+\sum_{wv\in\mathcal{A}(X)}p_X(w_{-n}^{-1}\bar{v})-1$$

$$=\sum_{wv\in\mathcal{A}(X)}\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)-\sum_{w\in\mathcal{A}(X)}(\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}0)$$

$$+\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}1))+\sum_{w_{-m}^{-1}v\in\mathcal{A}(X)}p_X(w_{-m}^{-1}\bar{v})-1$$

$$=\sum_{wv\in\mathcal{A}(X)}\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)-\sum_{wv\in\mathcal{A}(X)}\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)$$

$$-\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)}\sum_{j=1}^{n-k}p_X(w_{-n}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)+\sum_{w_{-m}^{-1}v\in\mathcal{A}(X)}p_X(w_{-m}^{-1}\bar{v})-1$$

$$=-\sum_{w_{-2m}^{-1}\in\mathcal{A}(X),w_{-2m}^{-1}v\notin\mathcal{A}(X)}\sum_{j=1}^{m}p_X(w_{-2m}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)+\sum_{w_{-m}^{-1}v\in\mathcal{A}(X)}p_X(w_{-m}^{-1}\bar{v})-1.$$

- For (2.14), we have

$$\sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} h^k_{n+1}(wv)\log\frac{h^k_{n+1}(wv)}{p_X(w)} = \sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)} h^k_{n+1}(wv)\log\frac{h^0_{2m+1}(w^{-1}_{-2m}v)}{p_X(w^{-1}_{-2m})}$$

$$= \sum_{w\in\mathcal{A}(X),wv\notin\mathcal{A}(X)}\sum_{j=1}^{n-k} p_X(w^{-j-1}_{-n}\bar{w}_{-j}w^{-1}_{-j+1}v)\log\frac{h^0_{2m+1}(w^{-1}_{-2m}v)}{p_X(w^{-1}_{-2m})}$$

$$= \sum_{w^{-1}_{-2m}\in\mathcal{A}(X),w^{-1}_{-2m}v\notin\mathcal{A}(X)}\sum_{j=1}^{m} p_X(w^{-j-1}_{-2m}\bar{w}_{-j}w^{-1}_{-j+1}v)\log\frac{h^0_{2m+1}(w^{-1}_{-2m}v)}{p_X(w^{-1}_{-2m})}.$$

- For (2.15), we have

$$\sum_{wv\in\mathcal{A}(X)} h^k_{n+1}(wv)\log p_X(v|w) + \sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon)} h^k_{n+1}(wv)\log\frac{h^0_{n+1}(wv)}{h^0_n(w)}$$

$$= \sum_{wv\in\mathcal{A}(X)}(\sum_{j=1}^{n-k} p_X(w^{-j-1}_{-n}\bar{w}_{-j}w^{-1}_{-j+1}v)+p_X(w^{-1}_{-n}\bar{v})-(n+1-k)p_X(wv))\log p_X(v|w^{-1}_{-m})$$

$$+ \sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon),p_X(w^{-m-1}_{-n})=0} h^k_{n+1}(wv)\log\frac{h^k_{n+1}(wv)}{h^k_n(w)}$$

$$+ \sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon),p_X(w^{-m-1}_{-n})>0} h^k_{n+1}(wv)\log\frac{h^k_{n+1}(wv)}{h^k_n(w)}$$

$$= (\sum_{wv\in\mathcal{A}(X)} + \sum_{p_{XZ}(w)=\Theta(\varepsilon),p_{XZ}(wv)=\Theta(\varepsilon),p_X(w^{-m-1}_{-n})=0})(\sum_{j=1}^{n-k} p_X(w^{-j-1}_{-n}\bar{w}_{-j}w^{-1}_{-j+1}v)+p_X(w^{-1}_{-n}\bar{v}))\log p_X(v|w^{-1}_{-m})$$

$$-(n+1-k)\sum_{w^{-1}_{-m}v\in\mathcal{A}(X)} p_X(w^{-1}_{-m}v)\log p_X(v|w^{-1}_{-m})$$

$$+ \sum_{p_{XZ}(w^{-1}_{-3m})=\Theta(\varepsilon),p_{XZ}(w^{-1}_{-3m}v)=\Theta(\varepsilon),p_X(w^{-m-1}_{-3m})>0} h^0_{3m+1}(wv)\log\frac{h^0_{3m+1}(wv)}{h^0_{3m}(w)}$$

$$= (n-k-m)\sum_{w^{-1}_{-m}v\in\mathcal{A}(X)} p_X(w^{-1}_{-m}v)\log p_X(v|w^{-1}_{-m})$$

$$+ \sum_{wv\in\mathcal{A}(X)}(\sum_{j=1}^{m} p(w^{-j-1}_{-n}\bar{w}_{-j}w^{-1}_{-j+1}v)+p_X(w^{-1}_{-n}\bar{v}))\log p_X(v|w^{-1}_{-m})$$

$$-(n+1-k)\sum_{w_{-m}^{-1}v\in\mathcal{A}(X)} p_X(w_{-m}^{-1}v)\log p_X(v|w_{-m}^{-1})$$

$$+\sum_{p_{XZ}(w_{-3m}^{-1})=\Theta(\varepsilon),p_{XZ}(w_{-3m}^{-1}v)=\Theta(\varepsilon),p_X(w_{-3m}^{-m-1})>0} h_{3m+1}^0(wv)\log\frac{h_{3m+1}^0(wv)}{h_{3m}^0(w)}$$

$$=(-m-1)\sum_{w_{-m}^{-1}v\in\mathcal{A}(X)} p_X(w_{-m}^{-1}v)\log p_X(v|w_{-m}^{-1})$$

$$+\sum_{w_{-2m}^{-1}v\in\mathcal{A}(X)}(\sum_{j=1}^{m} p_X(w_{-2m}^{-j-1}\bar{w}_{-j}w_{-j+1}^{-1}v)+p_X(w_{-2m}^{-1}\bar{v}))\log p_X(v|w_{-m}^{-1})$$

$$+\sum_{p_{XZ}(w_{-3m}^{-1})=\Theta(\varepsilon),p_{XZ}(w_{-3m}^{-1}v)=\Theta(\varepsilon),p_X(w_{-3m}^{-m-1})>0} h_{3m+1}^0(wv)\log\frac{h_{3m+1}^0(wv)}{h_{3m}^0(w)}.$$

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] 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.

[4] L. Arnold, V. M. Gundlach and L. Demetrius. Evolutionary formalism for products of positive random matrices. *Annals of Applied Probability*, 4:859–901, 1994.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] J. Fan, T. Poo, and B. Marcus. Constraint Gain *IEEE Trans. Information Theory*, **50**, 1989–1999, 2001.

[12] 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.

[13] Robert Gray. *Entropy and Information Theory.* 2007. Available at http://ee.stanford.edu/~gray/it.pdf

[14] 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.

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

[16] G. Han and B. Marcus. Asymptotics of noisy constrained capacity. *International Symposium on Information Theory*, pages: , Nice, France, 2007.

[17] 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.

[18] 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.

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

[20] P. Jacquet, G. Seroussi, and W. Szpankowski. On the Entropy of a Hidden Markov Process (full version). to appear in *Theoretical Computer Science*, 2007.

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

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

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

[24] B. Marcus, R. Roth and P. Siegel. Constrained Systems and Coding for Recording Channels. Chap. 20 in *Handbook of Coding Theory* (eds. V. S. Pless and W. C. Huffman), Elsevier Science, 1998.

[25] 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.

[26] 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.

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

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

[29] Y. Peres. *Analytic dependence of Lyapunov exponents on transition probabilities*, volume 1486 of *Lecture Notes in Mathematics, Lyapunov's exponents, Proceedings of a Workshop.* Springer Verlag, 1990.

[30] Y. Peres. Domains of analytic continuation for the top Lyapunov exponent. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(1):131–148, 1992.

[31] 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.

[32] D. Ruelle. Analyticity properties of the characteristic exponents of random matrix products. *Adv. Math.*, 32:68–80, 1979.
[33] E. Seneta. *Non-Negative Matrices.* New York: John Wiley & Sons, 1973.
[34] 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.
[35] S. Shamai (Shitz) and Y. Kofman. On the capacity of binary and Gaussian channels with run-length limited inputs. *IEEE Trans. Commun.*, **38**, 584–594, 1990.
[36] E. Zehavi and J. Wolf. On runlength codes. *IEEE Trans. Information Theory*, **34**, 45–54, 1988.
[37] O. Zuk, I. Kanter and E. Domany. The entropy of a binary hidden Markov process. *J. Stat. Phys.*, 121(3-4): 343-360 (2005)
[38] 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.

Department of Mathematics
University of Hong Kong
Pokfulam Road, Pokfulam, Hong Kong
E-mail: ghan@maths.hku.hk

Department of Mathematics
University of British Columbia
Vancouver, B.C. Canada, V6T 1Z2
E-mail: marcus@math.ubc.ca