Noisy Constrained Capacity for BSC Channels

Philippe Jacquet
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
Email: philippe.jacquet@inria.fr

Wojciech Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
Email: spa@cs.purdue.edu

***Abstract*— 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. As stated in [4] "... while calculation of the noise-free capacity of constrained sequences is well known, the computation of the capacity of a constraint in the presence of noise ... has been an unsolved problem in the half-century since Shannon's landmark paper ...." We first express the constrained capacity of a binary symmetric channel with (d, k)-constrained input as a limit of the top Lyapunov exponents of certain matrix random processes. Then, we compute asymptotic approximations of the noisy constrained capacity for cases where the noise parameter ε is small. In particular, we show that when $k \leq 2d$, the error term with respect to the constraint capacity is $O(\varepsilon)$, whereas it is $O(\varepsilon \log \varepsilon)$ when $k > 2d$. In both cases, we compute the coefficient of the error term. In the course of establishing these findings, we also extend our previous results on the entropy of a hidden Markov process to higher-order finite memory processes. These conclusions are proved by a combination of analytic and combinatorial methods.**

I. Introduction

We consider a binary symmetric channel (BSC) with crossover probability ε, and a constrained set of inputs. More precisely, let $\mathcal{S}_n$ denote the set of binary sequences of length n satisfying a given (d, k)-RLL constraint [18], i.e., no sequence in $\mathcal{S}_n$ contains a run of zeros of length shorter than d or longer than k (we assume that the values d and k, $d \leq k$, are understood from the context). We write $X_1^n \in \mathcal{S}_n$ for $X_1^n = X_1 \dots X_n$. Furthermore, we denote $\mathcal{S} = \bigcup_{n>0} \mathcal{S}_n$. We assume that the input to the channel is a stationary process $X = \{X_k\}_{k \geq 1}$ supported on $\mathcal{S}$. We regard the BSC channel as emitting a Bernoulli noise sequence $E = \{E_k\}_{k \geq 1}$, independent of X, with $P(E_i = 1) = \varepsilon$. The channel output is

$$Z_i = X_i \oplus E_i.$$

where $\oplus$ denotes addition modulo 2 (exclusive-or).

For ease of notation, we identify the BSC channel with its parameter ε. Let $C(\varepsilon)$ denote conventional BSC channel capacity (over unconstrained binary sequences), namely, $C(\varepsilon) = 1 - H(\varepsilon)$, where $H(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon) \log(1-\varepsilon)$. We use natural logarithms throughout. Entropies are correspondingly measured in nats. The entropy of a random variable or process X will be denoted $\mathbf{H}(X_1^n)$, and the entropy rate by $H(X)$. The *noisy constrained capacity* $C(\mathcal{S}, \varepsilon)$ is defined [4] by

$$C(\mathcal{S}, \varepsilon) = \sup_{X \in \mathcal{S}} I(X; Z) = \lim_{n \to \infty} \frac{1}{n} \sup_{X_1^n \in \mathcal{S}_n} I(X_1^n, Z_1^n), \quad (1)$$

where the supreme are over all stationary processes supported on $\mathcal{S}$ and $\mathcal{S}_n$, respectively. The *noiseless capacity* of the constraint is $C(\mathcal{S}) \triangleq C(\mathcal{S}, 0)$. This quantity has been extensively studied, and several interpretations and methods for its explicit derivation are known (see, e.g., [18] and extensive bibliography therein). 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., [26], [23], [1] and references therein). These methods allow for fairly precise numerical approximations of the capacity for given constraints and channel parameters.

Our approach to the noisy constrained capacity $C(\mathcal{S}, \varepsilon)$ is different. We first consider the corresponding mutual information,

$$I(X; Z) = H(Z) - H(Z|X). \quad (2)$$

Since $H(Z|X) = H(\varepsilon)$, the problem reduces to finding $H(Z)$, the entropy rate of the output process. If we restrict our attention to constrained processes X that are generated by Markov sources, the output process Z can be regarded as a *hidden Markov process* (HMP), and the problem of computing $I(X; Z)$ reduces to that of computing the entropy rate of this HMP. The noisy constrained capacity follows provided we find the maximizing distribution $P^{\max}$ of X, as it turns out.

It is well known (see, e.g., [18]) that we can regard the (d, k) constraint as the output of a kth-order finite memory (Markov) stationary process, uniquely defined by conditional probabilities $P(x_t | x_{t-k}^{t-1})$, where for any sequence $\{x_i\}_{i \geq 1}$, we denote by x_i^j, $j \geq i$, the sub-sequence $x_i, x_{i+1}, \dots, x_j$. For nontrivial constraints, some of these conditional probabilities must be set to zero in order to enforce the constraint (for example, the probability of a zero after seeing k consecutive zeros, or of a one after seeing less than d consecutive zeros). When the remaining free probabilities are assigned so that the entropy of the process is maximized, we say that the process is

*Preliminary version of this paper was presented at ISIT, Nice, 2007. Work of W. Szpankowski was supported in part by the NSF Grants CCF-0513636, DMS-0503742, DMS-0800568, and CCF -0830140, NSA Grant H98230-08-1-0092, EU Project No. 224218 through Poznan University of Technology, and the AFOSR Grant FA8655-08-1-3018.

maxentropic, and we denote it by $P^{\max}$. The noiseless capacity $C(\mathcal{S})$ is equal to the entropy of $P^{\max}$ [18].

The Shannon entropy (or, simply, *entropy*) of a HMP was studied as early as [2], where the analysis suggests the intrinsic complexity of the HMP entropy as a function of the process parameters. Blackwell [2] showed an expression of the entropy 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, we have seen a resurgence of interest in estimating HMP entropies [7], [8], [14], [19], [20], [27]. In particular, one 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 [14], where the Taylor expansion around $\varepsilon = 0$ is studied for a binary HMP of order one. In particular, the first derivative of the entropy rate at $\varepsilon = 0$ is expressed very compactly as a Kullback-Liebler divergence between two distributions on binary triplets, derived from the marginals of the input process X. It is also shown in [14], [15] that the entropy rate of a HMP can be expressed in terms of the top Lyapunov exponent of a random process of 2×2 matrices (cf. also [11], where the capacity of certain channels with memory is also shown to be related to top Lyapunov exponents). Further improvements, and new methods for the asymptotic expansion approach were obtained in [19], [27], and [8]. In [20] the authors express the entropy rate for a binary HMP where one of the transition probabilities is equal to zero as an asymptotic expansion including a $O(\varepsilon \log \varepsilon)$ term. As we shall see in the sequel, this case is related to the $(1,\infty)$ (or the equivalent $(0,1)$) RLL constraint. Analyticity of the entropy as a function of ε was studied in [7].

In Section II of this paper we extend the results of [14], [15] on HMP entropy to higher order Markov processes. We show that the entropy of a rth-order HMP can be expressed as the top Lyapunov exponent of a random process of matrices of dimensions $2^r \times 2^r$ (cf. Theorem 1). As an additional result of this work, of interest on its own, we derive the asymptotic expansion of the HMP entropy rate around $\varepsilon = 0$ for the case where all transition probabilities are positive (cf. Theorem 2). In particular, we derive an expression for the first derivative of the entropy rate as the Kullback-Liebler divergence between two distributions on $2r{+}1$-tuples, again generalizing the formula for $r{=}1$ [14].The results of Section II are applied, in Section III, to express the noisy constrained capacity as a limit of top Lyapunov exponents of certain matrix processes. These exponents, however, are notoriously difficult to compute [25]. Hence, as in the case of the entropy of HMPs, it is interesting to study asymptotic expansions of the noisy constrained capacity. In Section III-B, we study the asymptotics of the noisy constrained capacity, and we show that for (d,k) constraints with $k \leq 2d$, we have $C(\mathcal{S},\varepsilon) = C(\mathcal{S}) + K\,\varepsilon + O(\varepsilon^2 \log \varepsilon)$, where K is a well characterized constant. On the other hand, when $k > 2d$, we have $C(\mathcal{S},\varepsilon) = C(\mathcal{S}) + L\,\varepsilon \log \varepsilon + O(\varepsilon)$, where, again, L is an explicit constant. The latter case covers the $(0,1)$ constraint (and also the equivalent $(1,\infty)$ constraint). Our formula for the constant L in this case is consistent with the one derived from the results of [20]. Preliminary results of this paper were presented in [16].

We also remark that recently Han and Marcus [9] reached similar conclusions and obtained some generalizations using different methodology.

II. Entropy of Higher Order HMPs

Let $X = \{X_i\}_{i\geq1}$ be an rth-order stationary *finite memory (Markov) process* over a binary alphabet $\mathcal{A}{=}\{0,1\}$. The process is defined by the set of conditional probabilities $P(X_t = 1|X_{t-r}^{t-1} = a_1^r)$, $a_1^r \in A^r$. The process is equivalently interpreted as the Markov chain of its *states* $s_t = X_{t-r}^{t-1}$, $t > 0$ (we assume X_{-r+1}^0 is defined and distributed according to the stationary distribution of the process).[1] Clearly, a transition from a state $u{\in}A^r$ to a state $v{\in}A^r$ can have positive probability only if u and v satisfy $u_2^r{=}v_1^{r-1}$, 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 $P(E_i{=}1) = \varepsilon$. Finally, the HMP is

$$Z{=}\{Z_i\}_{i\geq1}, \quad Z_i = X_i{\oplus}E_i, \;\; i \geq 1\,. \tag{3}$$

Let $\tilde{Z}_i = (Z_i, Z_{i+1}, \ldots, Z_{i+r-1})$ and $\widetilde{E}_i = (E_i, \ldots, E_{i+r-1})$. Also, for $e{\in}\{0,1\}$, let $\widetilde{E}_i^e = (e, E_i, \ldots, E_{i+r-1})$. We next compute[2] $P(\tilde{Z}_1^n)$ (equivalently, $P(Z_1^{n+r-1})$). From the definitions of X and E, we have

$$\begin{aligned}
P(\tilde{Z}_1^n, \widetilde{E}_n) &= \sum_{e\in\mathcal{A}} P(\tilde{Z}_1^n, \widetilde{E}_n, E_{n-1} = e) \qquad (4)\\
&= \sum_{e\in\mathcal{A}} P(\tilde{Z}_1^{n-1}, Z_{n+r-1}, E_{n-1} = e, \widetilde{E}_n)\\
&= \sum_{e\in\mathcal{A}} P(Z_{n+r-1}, E_{n+r-1}|\tilde{Z}_1^{n-1}, \widetilde{E}_{n-1}^e) P(\tilde{Z}_1^{n-1}, \widetilde{E}_{n-1}^e)\\
&= \sum_{e\in\mathcal{A}} P(E_{n+r-1}) P_X(\tilde{Z}_n{\oplus}\widetilde{E}_n|\tilde{Z}_{n-1}{\oplus}\widetilde{E}_{n-1}^e) P(\tilde{Z}_1^{n-1}, \widetilde{E}_{n-1}^e).
\end{aligned}$$

Observe that in the last line the transition probabilities $P_X(\cdot|\cdot)$ are with respect to the original Markov chain.

We next derive, from (4) , an expression for $P(\tilde{Z}_1^n)$ as a product of matrices extending our earlier work [14], [15]. In what follows, vectors are of dimension 2^r, and matrices are of dimensions $2^r \times 2^r$. We denote *row* vectors by bold lowercase letters, matrices by bold uppercase letters, and we let $\mathbf{1} = [1, \ldots, 1]$; superscript t denotes transposition. Entries in vectors and matrices are indexed by vectors in A^r, according to some fixed order, so that $A^r = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{2^r}\}$. Let

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

[1] We generally use the term "finite memory process" for the first interpretation, and "Markov chain" for the second.

[2] In general, the measures governing probability expressions will be clear from the context. In cases when confusion is possible, we will explicitly indicate the measure, e.g., P_X.

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

$$\mathbf{M}_{\mathbf{e}_{n-1},\mathbf{e}_n}(\tilde{Z}_n|\tilde{Z}_{n-1}) = P_X(\tilde{Z}_n \oplus \mathbf{e}_n|\tilde{Z}_{n-1} \oplus \mathbf{e}_{n-1})P(\widetilde{E}_n = \mathbf{e}_n). \tag{5}$$

All other entries are zero. Clearly, $\mathbf{M}(\tilde{Z}_n|\tilde{Z}_{n-1})$ is a random matrix, drawn from a set of 2^{r+1} possible realizations.

With these definitions, it follows from (4) that

$$\mathbf{p}_n = \mathbf{p}_{n-1}\mathbf{M}(\tilde{Z}_n|\tilde{Z}_{n-1}). \tag{6}$$

Since $P_Z(\tilde{Z}_1^n) = \mathbf{p}_n\mathbf{1}^t = \sum_{\mathbf{e}\in A^r} P_Z(\tilde{Z}_1^n, \widetilde{E}_n = \mathbf{e})$, after iterating (6), we obtain

$$P_Z(\tilde{Z}_1^n) = \mathbf{p}_1\mathbf{M}(\tilde{Z}_2|\tilde{Z}_1)\cdots\mathbf{M}(\tilde{Z}_n|\tilde{Z}_{n-1})\mathbf{1}^t. \tag{7}$$

The joint distribution $P_Z(Z_1^n)$ of the HMP, presented in (7), has the form $\mathbf{p}_1\mathbf{A}_n\mathbf{1}^t$, where $\mathbf{A}_n$ is the product of the first $n-1$ random matrices of the process

$$\mathcal{M} = \mathbf{M}(\tilde{Z}_2|\tilde{Z}_1), \mathbf{M}(\tilde{Z}_3|\tilde{Z}_2), \ldots, \mathbf{M}(\tilde{Z}_n|\tilde{Z}_{n-1}), \ldots \tag{8}$$

Applying a subadditive ergodic theorem, and noting that $\mathbf{p}_1\mathbf{A}_n\mathbf{1}^t$ is a norm of $\mathbf{A}_n$, it is readily proved that $n^{-1}\mathbf{E}[-\log P_Z(Z_1^n)]$ must converge to a constant ξ known as the *top Lyapunov exponent* of the random process $\mathcal{M}$ (cf. [5], [21], [25]). This leads to the following theorem.

Theorem 1: The entropy rate of the HMP Z of (3) satisfies

$$H(Z) = \lim_{n\to\infty} \mathbf{E}\left[-\frac{1}{n}\log P_Z(Z_1^{n+r})\right]$$
$$= \lim_{n\to\infty}\frac{1}{n}\mathbf{E}\left[-\log\left(\mathbf{p}_1\mathbf{M}(\tilde{Z}_2|\tilde{Z}_1)\cdots\mathbf{M}(\tilde{Z}_n|\tilde{Z}_{n-1})\mathbf{1}^t\right)\right] = \xi,$$

where ξ is the top Lyapunov exponent of the process $\mathcal{M}$ of (8).

Theorem 1 and its derivation generalize the results, for $r = 1$, of [14], [15], [27], [28]. It is known that computing top Lyapunov exponents is hard (maybe infeasible), as shown in [25]. Therefore, we shift our attention to asymptotic approximations.

We consider the entropy rate $H(Z)$ for the HMP Z as a function of ε for *small* ε. In order to derive expressions for the entropy rate, we resort to the following formal definition (which was also used in entropy computations in [13] and [15]):

$$R_n(s,\varepsilon) = \sum_{z_1^n \in \mathcal{A}^n} P_Z^s(z_1^n), \tag{9}$$

where s is a real (or complex) variable, and the summation is over all binary n-tuples. It is readily verified that

$$\mathbf{H}(Z_1^n) = \mathbf{E}\left[-\log P_Z(Z_1^n)\right] = -\left.\frac{\partial}{\partial s}R_n(s,\varepsilon)\right|_{s=1}. \tag{10}$$

The entropy of the underlying Markov sequence is

$$\mathbf{H}(X_1^n) = -\left.\frac{\partial}{\partial s}R_n(s,0)\right|_{s=1}.$$

Furthermore, let $\mathbf{P} = [p_{\mathbf{e}_i,\mathbf{e}_j}]_{\mathbf{e}_i,\mathbf{e}_j\in\mathcal{A}^r}$ be the transition matrix of the underlying rth order Markov chain, and let $\boldsymbol{\pi} = [\pi_\mathbf{e}]_{\mathbf{e}\in\mathcal{A}^r}$ be the corresponding stationary distribution . Define also $\mathbf{P}(s) = [p^s_{\mathbf{e}_i,\mathbf{e}_j}]_{\mathbf{e}_i,\mathbf{e}_j\in\mathcal{A}^r}$ and $\boldsymbol{\pi}(s) = [\pi^s_\mathbf{e}]_{\mathbf{e}\in\mathcal{A}^r}$. Then

$$R_n(s,0) = \sum_{z^n} P_X^s(z_1^n) = \boldsymbol{\pi}(s)\mathbf{P}(s)^{n-1}\mathbf{1}^t. \tag{11}$$

Using a formal Taylor expansion near $\varepsilon = 0$, we write

$$R_n(s,\varepsilon) = R_n(s,0) + \varepsilon\left.\frac{\partial}{\partial\varepsilon}R_n(s,\varepsilon)\right|_{\varepsilon=0} + O(g(n)\varepsilon^2), \tag{12}$$

where $g(n)$ is the second derivative of $R_n(s,\varepsilon)$ with respect to ε, computed at some ε', provided these derivatives exist (the dependence on n stems from (9)).

Using analyticity at $\varepsilon = 0$ (cf. [7], [15]), we find

$$\mathbf{H}(Z_1^n) = \mathbf{H}(X_1^n) - \varepsilon\left.\frac{\partial^2}{\partial s\partial\varepsilon}R_n(s,\varepsilon)\right|_{\varepsilon=0,\, s=1} + O(g(n)\varepsilon^2)$$
$$= \mathbf{H}(X_1^n) - \varepsilon\frac{\partial}{\partial s}\frac{\partial}{\partial\varepsilon}\left.\sum_{z_1^n} P_Z^s(z_1^n)\right|_{\varepsilon=0,\, s=1} + O(g(n)\varepsilon^2). \tag{13}$$

To compute the linear term in the Taylor expansion (13), we differentiate with respect to s, and evaluate at $s = 1$. Proceeding in analogy to the derivation in [14], we obtain the following result basically proved in [15], so we omit details here.

Theorem 2: If the conditional symbol probabilities in the finite memory (Markov) process X satisfy $P(a_{r+1}|a_1^r) > 0$ for all $a_1^{r+1} \in \mathcal{A}^{r+1}$, then the entropy rate of Z for small ε is

$$H(Z) = \lim_{n\to\infty}\frac{1}{n}H_n(Z^n) = H(X) + f_1(P_X)\varepsilon + O(\varepsilon^2), \tag{14}$$

where, denoting by $\bar{z}_i$ the Boolean complement of z_i, and $\check{z}^{2r+1} = z_1\ldots z_r\bar{z}_{r+1}z_{r+2}\ldots z_{2r+1}$, we have

$$\begin{aligned} f_1(P_X) &= \sum_{z_1^{2r+1}} P_X(z_1^{2r+1})\log\frac{P_X(z_1^{2r+1})}{P_X(\check{z}_1^{2r+1})} \\ &= \mathbb{D}\left(P_X(z_1^{2r+1})||P_X(\check{z}_1^{2r+1})\right). \end{aligned} \tag{15}$$

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

A question arises about the asymptotic expansion of the entropy $H(Z)$ when some of the conditional probabilities are zero. Clearly, when some transition probabilities are zero, then certain sequences x_1^n are not reachable by the Markov process, which provides the link to constrained sequences. For example, consider a Markov chain with the following transition probabilities

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix} \tag{16}$$

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 transition matrix $\mathbf{P}$ of (16) in [20], where it is shown that

$$H(Z) = H(X) - \frac{p(2-p)}{1+p}\varepsilon \log \varepsilon + O(\varepsilon) \tag{17}$$

as $\varepsilon \to 0$.

In this paper, in Section IV and Appendix A we prove the following generalization of Theorem 2 for (d,k) sequences.

Theorem 3: Let Z be a HMM representing a (d,k) sequence. Then

$$H(Z) = H(X) - f_0(P_X)\varepsilon \log \varepsilon + f_1(P_X)\varepsilon + O(\varepsilon^2 \log \varepsilon) \tag{18}$$

for some $f_0(P_X)$ and $f_1(P_X)$. If all transition probabilities are positive, then $f_0(P_X) = 0$ and the coefficient $f_1(P_X)$ at ε is presented in Theorem 2. The coefficient $f_0(P_X)$ is derived in Section IV, and for the maximizing distribution is presented in Theorems 5 and 6.

Recently, Han and Marcus [9] showed that in general for any HMM

$$H(Z) = H(X) - f_0(P_X)\varepsilon \log \varepsilon + O(\varepsilon)$$

which is further generalized in [10] to

$$H(Z) = H(X) - f_0(P_X)\varepsilon \log \varepsilon + f_1(P)\varepsilon + O(\varepsilon^2 \log \varepsilon)$$

when at least one of the transition probabilities in the Markov chain is zero.

III. Capacity of the Noisy Constrained System

We now apply the results on HMPs to the problem of noisy constrained capacity.

A. Capacity as a Lyapunov Exponent

Recall that $I(X;Z) = H(Z) - H(\varepsilon)$ and, by Theorem 1, when X is a Markov process, we have $H(Z) = \xi(P_X)$ where $\xi(P_X)$ is the top Lyapunov exponent of the process $\{\mathbf{M}(\tilde{Z}_i|\tilde{Z}_{i-1})\}_{i>0}$. In [3] it is proved that the process optimizing the mutual information can be approached by a sequence of Markov representations of increasing order. Therefore, as a direct consequence of this fact and Theorem 1 we conclude the following.

Theorem 4: The noisy constrained capacity $C(\mathcal{S},\varepsilon)$ for a (d,k) constraint through a BSC channel of parameter ε is given by

$$C(\mathcal{S},\varepsilon) = \lim_{r\to\infty} \sup_{P_X^{(r)}} \xi(P_X^{(r)}) - H(\varepsilon) \tag{19}$$

where $P_X^{(r)}$ denotes the probability law of an rth-order Markov process generating the (d,k) constraint $\mathcal{S}$.

In the next subsection, we turn our attention to asymptotic expansions of $C(\mathcal{S},\varepsilon)$ near $\varepsilon = 0$.

B. Asymptotic Behavior

A nontrivial constraint will necessarily have some zero-valued conditional probabilities. Therefore, the associated HMP will not be covered by Theorem 2, but rather by Theorem 3. For (d,k) sequences we have

$$H(Z) = H(X) - f_0(P_X)\varepsilon \log \varepsilon + f_1(P_X)\varepsilon + o(\varepsilon) \tag{20}$$

for some $f_0(P_X)$ and $f_1(P_X)$ where P_X is the underlying Markov process. As discussed in (1) of the introduction,

$$C(\mathcal{S},\varepsilon) = \sup_{X\in\mathcal{S}} H(Z) - H(\varepsilon)$$

where $H(\varepsilon) = -\varepsilon \log \varepsilon + \varepsilon - O(\varepsilon^2)$ for small ε. In [9], [10] Han and Marcus prove that the maximizing distribution in (1) is the maxentropic distribution $P^{\max}$ with the error term $O(\varepsilon^2 \log^2 \varepsilon)$ (cf. Theorem 3.2 of [10]), thus exceeding the error term $O(\varepsilon^2 \log \varepsilon)$ of the entropy estimation of Theorem 3. We establish the same error term in Section IV using different methodology. In summary, we are led to

$$\begin{aligned} C(\mathcal{S},\varepsilon) &= C(\mathcal{S}) - (1 - f_0(P_X^{\max}))\varepsilon \log \varepsilon + (f_1(P_X^{\max}) - 1)\varepsilon \\ &+ O(\varepsilon^2 \log^2 \varepsilon) \end{aligned} \tag{21}$$

where $C(\mathcal{S})$ is the capacity of noiseless RLL system. Various methods exist to derive $C(\mathcal{S})$ [18]. In particular, one can write [18], [24] $C(\mathcal{S}) = -\log \rho_0$, where ρ_0 is the smallest real root of

$$\sum_{\ell=d}^{k} \rho_0^{\ell+1} = 1. \tag{22}$$

Our goal is to derive explicit expressions for $f_0(P_X^{\max})$ and $f_1(P_X^{\max})$ for (d,k) sequences. For example, we will show in Theorem 5 below that for some RLL constraints, we have $f_0(P_X^{\max}) = 1$ in (21), hence the noisy constrained capacity is of the form $C(\mathcal{S},\varepsilon) = C(\mathcal{S}) + O(\varepsilon)$. In Theorem 6 below we derive also $f_1(P_X^{\max})$.

We apply the same approach as in previous section, that is, we use the auxiliary function $R_n(s,\varepsilon)$ defined in (9). To start, we find a simpler expression for $P_Z(z_1^n)$. Summing over the number of errors introduced by the channel, we find

$$P_Z(z_1^n) = P_X(x_1^n)(1-\varepsilon)^n + \varepsilon(1-\varepsilon)^{n-1}\sum_{i=1}^{n} P_X(x_1^n \oplus e_i)$$

plus the error term $O(\varepsilon^2)$ (resulting from two or more errors), where $e_j = (0,\ldots,0,1,0,\ldots,0) \in \mathcal{A}^n$ with a 1 at position j. Let $B_n \subseteq \mathcal{A}^n$ denote the set of sequence z_1^n at Hamming distance one from $\mathcal{S}_n$, and $C_n = \mathcal{A}^n \setminus (\mathcal{S}_n \cup B_n)$. Notice that sequences in C_n are at distance at least two from $\mathcal{S}_n$, and contribute the $O(\varepsilon^2)$ term. From the above, we conclude

$$R_n(s,\varepsilon) = \sum_{z_1^n} P_Z(z_1^n) \tag{23}$$

$$\sum_{z_1^n\in\mathcal{S}_n} P_Z(z_1^n)^s + \sum_{z_1^n\in B_n} P_Z(z_1^n)^s + \sum_{z_1^n\in C_n} P_Z(z_1^n).$$

We observe that

$$\sum_{z_1^n\in\mathcal{S}_n} P_Z(z_1^n)^s = O(1), \quad \sum_{z_1^n\in B_n} P_Z(z_1^n)^s = O(\varepsilon^s),$$
$$\sum_{z_1^n\in B_n} P_Z(z_1^n)^s = O(\varepsilon^{2s}), \qquad \varepsilon\to 0.$$

Defining

$$\begin{aligned}\phi_n(s) &= \sum_{z_1^n\in\mathcal{S}_n} P_X(z_1^n)^{s-1}\sum_{i=1}^n P_X(z_1^n)\\ Q_n(s) &= \sum_{z_1^n\in B_n}\left(\sum_{i=1}^n P_X(z_1^n\oplus e_i)\right)^s\end{aligned}$$

we arrive at the following expression for $R_n(s,\varepsilon)$

$$\begin{aligned}R_n(s,\varepsilon) &= (1-\varepsilon)^{ns}R_n(s,0)+\varepsilon(1-\varepsilon)^{ns-1}\phi_n(s) \quad (24)\\ &+ \varepsilon^s(1-\varepsilon)^{(n-1)s}Q_n(s)+O(\varepsilon^2+\varepsilon^{1+s}+\varepsilon^{2s}).\end{aligned}$$

Notice that $\phi_n(1)+Q_n(1)=\sum_{z_1^n}\sum_{i=1}^n P_X(z_1^n\oplus e_i)=n$.

We now derive $\mathbf{H}(Z_1^n)=-\frac{\partial}{\partial s}R_n(1,\varepsilon)$ using the fact that $R_n(1,\varepsilon)=1$. Since all the functions involved are analytic, we obtain

$$\begin{aligned}\mathbf{H}(Z_1^n) &= \mathbf{H}(X_1^n)(1-n\varepsilon)+n\varepsilon-\varepsilon(\phi_n(1)+\phi_n'(1))\\ &- \varepsilon\log\varepsilon Q_n(1)-\varepsilon Q_n'(1)+O(n\varepsilon^2\log\varepsilon), \quad (25)\end{aligned}$$

where the error term is derived in Appendix A. In the above, $\phi_n'(1)$ and $Q_n'(1)$ are, respectively, the derivative of $\phi_n(s)$ and $Q_n(s)$ at $s=1$. Notice also that the term $n\mathbf{H}(X_1^n)\varepsilon$ of order $n^2\varepsilon$ is cancelled by $(\phi_n'(1)+Q_n'(1))\varepsilon=(\mathbf{H}(X_1^n)n+O(n))\varepsilon$ and only $n\varepsilon$ term remains.

The case $k\le 2d$ is interesting: one-bit flip in a (d,k) sequence is guaranteed to violate the constraint, and consequently $\forall z_1^n\in\mathcal{S}_n$ and $\forall i$: $P_X(z_1^n\oplus e_i)=0$. Therefore $\phi_n(s)=0$ in this case, leaving $Q_n(1)=n$. Thus, in the case $k\le 2d$, we have $f_0(P)=1$, and the term $O(\varepsilon\log\varepsilon)$ in (21) cancels out.

Further considerations are required to compute $Q_n'(1)$ and obtain the coefficient of ε in (25) . Here, we provide the necessary definitions, and state our result that are proved in Section IV. Ignoring border effects (which do not affect asymptotics, as easy to see[3]), we restrict our analysis to (d,k) sequences over the *extended* alphabet (of *phrases*) [18]

$$\mathcal{B}=\{\,0^d1,0^{d+1}1,\ldots,0^k1\,\}.$$

In other words, we consider only (d,k) sequences that end with a "1". For such sequences, we assume that they are generated by a memoryless process over the super-alphabet. This is further discussed in Section IV.

Let p_ℓ denote the probability of the super-symbol $0^\ell 1$. The maxentropic distribution $P^{\max}$ corresponds to the case of

$$p_\ell = P_X^{\max}(0^\ell 1), \quad d\le\ell\le k\,. \qquad (26)$$

[3]Indeed, in general a (d,k) sequence may have at most k starting and ending zeros of total length $n+O(1)$ that cannot affect the entropy rate.

Note that in this case $p_\ell=\rho_0^{\ell+1}$, with ρ_0 as in (22). The expected length of a super-symbol in $\mathcal{B}$ is $\lambda=\sum_{\ell=d}^k(\ell+1)p_\ell$. We also introduce the generating function

$$r(s,z)=\sum_\ell p_\ell^s z^{\ell+1}.$$

By $\rho(s)$ we denote the smallest root in z of $r(s,z)=1$, that is $r(s,\rho(s))=1$. Clearly, $\rho(1)=\rho_0$ and

$$\rho'(1)=-\frac{\sum_\ell p_\ell\log p_\ell}{\lambda}$$

is the entropy rate *per bit* of the super-alphabet, and $\rho'(1)=H(X)$. Furthermore, we define

$$\lambda(s)=\frac{\partial}{\partial z}r(s,z)\,\big|_{z=\rho(s)}$$

and notice that $\lambda(1)=\lambda$.

Finally, to present succinctly our results, we introduce some additional notation. Let

$$\alpha(s,z)=\sum_\ell(2d-\ell)p_\ell^s z^{\ell+1}.$$

For integers ℓ_1,ℓ_2, $d\le\ell_1,\ell_2\le k$, let $\mathcal{I}_{\ell_1,\ell_2}$ denote the interval

$$\mathcal{I}_{\ell_1,\ell_2}=$$
$$\{\ell\colon -\min{}_+\{\ell_1-d,k-\ell_2-1\}\le\ell\le\min{}_+\{\ell_2-d,k-\ell_1-1\}\},$$

where $\min_+\{a,b\}=\max\{\min\{a,b\},0\}$. We shall write $\mathcal{I}^*_{\ell_1,\ell_2}=\mathcal{I}_{\ell_1,\ell_2}\setminus\{0\}$. At last, we define $\tau(s,z)=\tau_1(s,z)+\tau_2(s,z)+\tau_3(s,z)$ where

$$\begin{aligned}\tau_1(s,z) &= \sum_{\ell_1,\ell_2}2\max\{0,\ell_1+\ell_2-k-d\}p_{\ell_1}^s p_{\ell_2}^s z^{\ell_2+\ell_2+2}\\ \tau_2(s,z) &= \sum_{\ell_1=d}^k\sum_{\ell_2=d}^k\sum_{\theta\in\mathcal{I}^*_{\ell_1,\ell_2}}\frac{1}{2}(p_{\ell_1}p_{\ell_2}+p_{\ell_1+\theta}p_{\ell_2-\theta})^s z^{\ell_2+\ell_2+2}\\ \tau_3(s,z) &= \sum_{\ell_1=d}^k\sum_{\ell_2=d}^k\frac{1}{2\min\{k,\ell_1+\ell_2-d\}-(\ell_1+\ell_2)+1}\\ &\times\left(\sum_{\theta\in I_{\ell_1,\ell_2}}p_{\ell_1+\theta}p_{\ell_2-\theta}\right)^s z^{\ell_2+\ell_2+2}\end{aligned}$$

Now we are in a position to present our main results. The proofs are delayed till the next section. The following theorem summarizes our findings for the case $k\le 2d$.

Theorem 5: Consider the constrained (d,k) system $\mathcal{S}$ with $k\le 2d$. Then,

$$C(\mathcal{S},\varepsilon)=C(\mathcal{S})-(1-f_0(P_X^{\max}))\varepsilon+O(\varepsilon^2\log^2\varepsilon),$$

where

$$\begin{aligned}f_0(P_X^{\max}) &= \log\lambda+2\frac{\lambda'(1)}{\lambda}+\frac{\frac{\partial}{\partial s}\tau(1,1)+\frac{\partial}{\partial s}\alpha(1,1)}{\lambda}\\ &+ \rho'(1)(\frac{\partial^2}{\partial s\partial z}\alpha(1,1)+\frac{\partial^2}{\partial s\partial z}\tau(1,1))\\ &+ \frac{\rho'(1)}{\lambda}(\frac{\partial}{\partial z}\alpha(1,1)+\frac{\partial}{\partial z}\tau(1,1))-1\end{aligned}$$

for $\varepsilon \to 0$ and $\lambda(s)$, $\alpha(s,z)$ and $\tau(s,z)$ are defined above.

In the complementary case $k > 2d$, the term $\phi_n(s)$ in (23) does not vanish, and thus the $O(\varepsilon \log \varepsilon)$ term in (21) is generally nonzero. For this case, using techniques similar to the ones leading to Theorem 5, we obtain the following result.

Theorem 6: Consider the constrained (d,k) system $\mathcal{S}$ with $k \geq 2d$. Define

$$\gamma = \sum_{\ell > 2d} (\ell - 2d) p_\ell \,, \qquad \delta = \sum_{d \leq \ell_1 + \ell_2 + 1 \leq k} p_{\ell_1} p_{\ell_2} \,,$$

and $\lambda = \sum_{\ell=d}^{k} = (\ell+1) p_\ell$ where p_ℓ is from (26) Then,

$$C(\mathcal{S}, \varepsilon) = C(\mathcal{S}) - (1 - f_0(P_X^{\max}))\, \varepsilon \log \varepsilon^{-1} + O(\varepsilon) \,, \quad (27)$$

where

$$f_0(P_X^{\max}) = 1 - \frac{\gamma + \delta}{\lambda}$$

for $\varepsilon \to 0$.

Example. We consider the $(1,\infty)$ constraint with transition matrix $\mathbf{P}$ as in (16) . Computing the quantities called for in Theorem 6 for $d=1$ and $k=\infty$, we obtain $p_\ell = (1-p)^{\ell-1} p$, $\lambda = \frac{1+p}{p}$, $\gamma = \frac{(1-p)^2}{p}$, and $\delta = 1$. Thus,

$$f_0(P_X) = 1 - \frac{\gamma + \delta}{\lambda} = \frac{p(p-2)}{p-1} \,,$$

consistent with the calculation of the same quantity in [20]. The noisy constrained capacity is obtained when $P = P^{\max}$, i.e., $p = 1/\varphi^2$, where $\varphi = (1+\sqrt{5})/2$, the golden ratio. Then, $f_0(P^{\max}) = 1/\sqrt{5}$, and by Theorem 6

$$C(\mathcal{S}, \varepsilon) = \log \varphi - (1 - 1/\sqrt{5}) \varepsilon \log(1/\varepsilon) + O(\varepsilon)$$

for $\varepsilon \to 0$.

IV. Analysis

In this section, we derive explicit expression for the coefficients $f_0(P_X)$ and $f_1(P_X)$ of Theorem 3, as well as $f_0(P^{\max})$ and $f_1(P^{\max})$ of Theorems 5 and 6. We also establish the error term in Theorem 5.

Throughout, we consider the super-alphabet approach. Recall that a super-symbol is a text $0^\ell 1$ for $d \leq \ell \leq k$ which is drawn from a memoryless source. This model is equivalent to a Markov process with renewals at symbols ”1”. As before, p_ℓ is the probability of the super symbol $0^\ell 1$. The entropy rate per symbol is $-\sum_\ell p_\ell \log p_\ell$. It is not difficult to see that the maximal entropy rate is attained at $p_\ell = \rho_0^{\ell+1}$ where ρ_0 is defined in (22). This in fact corresponds to the case when all (d,k) sequences of length n are equiprobable. Furthermore, we consider (d,k) sequences generated by super symbols under the assumption that they are of length n. This is equivalent to consider a Markovian (d,k)-sequence of length n under the restriction that it ends with a “1”.

Let x_1^n be a sequence of length n made of m super-symbols: $x^n = 0^{\ell_1} 1 0^{\ell_2} 1 \ldots 0^{\ell_m} 1$. We shall call such (d,k) sequences *reduced* (d,k) sequences. The actual length of such sequences is $L(x_1^n) = n$. We also write $\lambda = \sum_\ell (\ell+1) p_\ell$.

In the sequel, we only consider reduced (d,k) sequences, and therefore define

$$\tilde{P}(x_1^n) = \prod_{i=1}^{m} p_{\ell_i}.$$

Notice that $\tilde{P}(x_1^n) = 0$ if x_1^n is not a reduced (d,k) sequence (i.e., it doesn't end on a 1). In view of this we have

$$P_X(x_1^n) = \frac{\tilde{P}(x_1^n)}{P_n}$$

where

$$P_n = \sum_{x_1^n} \tilde{P}(x_1^n).$$

Observe that P_n is the probability that the n-th symbol is exactly a “1” (in other words, x_1^n is built from a finite number of super symbols).

Recalling the definition $r(s,z) = \sum_\ell p_\ell^s z^{\ell+1}$, we find

$$\sum_n P_n z^n = \frac{1}{1 - r(1,z)}.$$

Indeed, every reduced (d,k) sequence consists of an empty string, one super symbol, two super symbols or more, thus $\sum_n P_n z^n = \sum_k r^k(1,z) = 1/(1 - r(1,z))$ (cf. [24]). By the Cauchy formula [24] we obtain

$$\begin{aligned} P_n &= \frac{1}{2\pi i} \oint \frac{1}{1 - r(1,z)} \frac{dz}{z^{n+1}} \\ &= \frac{1}{\frac{\partial}{\partial z} r(1,1)} + O(\mu^{-n}) = \frac{1}{\lambda} + O(\mu^{-n}) \end{aligned}$$

for some $\mu > 1$, since 1 is the largest root of $1 = r(1,z)$ and $\frac{\partial}{\partial z} r(1,1) = \lambda$.

Let $\tilde{A}_m$ be the set of (d,k) reduced sequences made of exactly m super-symbols with no restriction on its length. We call it the variable-length model. Let $\tilde{A}_* = \bigcup_m \tilde{A}_m$. Let $\tilde{B}_m$ be the set of such sequences that are exactly at Hamming distance 1 from a sequence in $\tilde{A}_m$. By our convention, if $x \in \tilde{A}_m$ for some m, (*i.e.* if $x = 0^{\ell_1} 1 0^{\ell_2} 1 \ldots 0^{\ell_m} 1$), then $\tilde{P}(x) = \prod_{i=1}^{i=m} p_\ell$; otherwise $\tilde{P}(x) = 0$). We denote by $L(x)$ the length of x.

To derive $\mathbf{H}(Z_1^n)$ found in (25) we need to evaluate $\phi_n'(1)$ and $Q_n'(1)$. We estimate these quantities in the variable-length model as described above and then re-interpret them in the original model. Define

$$\phi(s,z) = \sum_n P_n^s \phi_n(s) z^n, \quad (28)$$

$$Q(s,z) = \sum_n P_n^s Q_n(s) z^n \quad (29)$$

which we re-write as

$$\phi(s,z) = \sum_m \tilde{\phi}_m(s,z), \quad (30)$$

$$Q(s,z) = \sum_m \tilde{Q}_m(s,z), \quad (31)$$

where

$$\begin{aligned}\tilde{\phi}_m(s,z) &= \sum_{x\in\tilde{A}_m} \tilde{P}^{s-1}(x) \sum_{i=1}^{L(x)} \tilde{P}(x\oplus e_i) z^{L(x)},\\ \tilde{Q}_m(s,z) &= \sum_{x\in\tilde{B}_m} \left(\sum_{i=1}^{L(x)} \tilde{P}(x\oplus e_i)\right)^s z^{L(x)}.\end{aligned}$$

Notice that

$$\tilde{\phi}_m(1,z)+\tilde{Q}_m(1,z) = \mathbf{E}[L(x)z^{L(x)}] = z\frac{\partial}{\partial z} r^m(1,z),$$

and $\tilde{\phi}_m(1,1)+\tilde{Q}_m(1,1)=m\lambda$. We next evaluate $\tilde{\phi}(s,z)$ and $\tilde{Q}(s,z)$.

A. Computation of $\tilde{\phi}_m(s,z)$

The case $k\le 2d$ is easy since $x\oplus e_j\notin\tilde{A}_m$ when $x\in\tilde{A}_m$. Thus $\tilde{\phi}_m(s,z)=0$. In the sequel we concentrate on $k>2d$. The following result is easy to prove.

Theorem 7: For reduced (d,k) sequences consisting of m super symbols, we have

$$\tilde{\phi}_m(s,z) = mb_1(s,z)r^{m-1}(s,z)+(m-1)b_2(s,z)r^{m-2}(s,z),$$

where

$$\begin{aligned}b_1(s,z) &= \sum_{\ell=d}^{k} p_\ell^{s-1} \sum_{j=1}^{\ell} p_{j-1}p_{\ell-j} z^{\ell+1},\\ b_2(s,z) &= \sum_{d\le \ell_1+\ell_2\le k} p_{\ell_1}^{s-1}p_{\ell_2}^{s-1} p_{\ell_1+\ell_2+1} z^{\ell_1+\ell_2+2}.\end{aligned}$$

In particular,

$$\begin{aligned}b_1(1,1) &= \sum_{\ell=d}^{\ell=k}\sum_j p_{j-1}p_{\ell-j},\\ b_2(1,1) &= \sum_\ell \max\{0,\ell-2d\}p_\ell.\end{aligned}$$

Proof. We need to consider two cases: one in which the error changes a 0 to a 1, and the other one when the error occurs on a 1. In the first case, $m-1$ super symbols are not changed and each contributes $r(s,z)$. The corrupted super symbol is divided into two and its contribution is summarized in $b_1(s,z)$.

In the second case, an ending 1 is changed into a 0 so two super symbols (except the last one) collapsed into a one super symbol. This contribution is summarized by $b_2(s,z)$ while the other $m-2$ super symbols, represented by $r(s,z)$ are unchanged. ∎

B. Computation of $\tilde{Q}_m(s,z)$

We recall the following definitions. For integers ℓ_1,ℓ_2, $d\le \ell_1,\ell_2\le k$, let $\mathcal{I}_{\ell_1,\ell_2}$ denote the interval

$$\mathcal{I}_{\ell_1,\ell_2} = \\ \{\ell\colon -\min{}_+\{\ell_1-d,k-\ell_2-1\}\le\ell\le\min{}_+\{\ell_2-d,k-\ell_1-1\}\},$$

where $\min_+\{a,b\}=\max\{\min\{a,b\},0\}$. We shall write $\mathcal{I}^*_{\ell_1,\ell_2}=\mathcal{I}_{\ell_1,\ell_2}\setminus\{0\}$.

Observe first that $\tilde{Q}_m(s,z)$ can be rewritten as

$$\begin{aligned}\tilde{Q}_m(s,z) &= \sum_{x\in\tilde{A}_m}\sum_{j=1}^{j=L(x)} \frac{1_{x\oplus e_j\notin\tilde{A}_*}}{|B(x\oplus e_j)\cap\tilde{A}_*|}\\ &\quad\times\left(\sum i=1^{i=L(x)}\tilde{P}(x\oplus e_j\oplus e_i)\right)^s z^{L(x)}\end{aligned}$$

where $B(y)$ is the set of all sequences of the same length as y and within Hamming distance 1 from y.

Theorem 8: For reduced (d,k) sequences consisting of m super symbols, the following holds

$$\tilde{Q}_m(s,z) = m\alpha(s,z)r^{m-1}(s,z)+(m-1)\tau(s,z)r^{m-2}(s,z)$$

where

$$\alpha(s,z)=\sum_\ell \max\{0,2d-\ell\}p_\ell^s z^{\ell+1}$$

and $\tau(s,z)=\tau_1(s,z)+\tau_2(s,z)+\tau_3(s,z)$ where

$$\begin{aligned}\tau_1(s,z) &= \sum_{\ell_1=d}^{k}\sum_{\ell_2=d}^{k}(\max\{0,d(\ell_1)+\ell_2-k\}\\ &\quad+\max\{0,d(\ell_2)+\ell_1-k\})\,p_{\ell_1}^s p_{\ell_2}^s z^{\ell_1+\ell_2+2},\\ \tau_2(s,z) &= \sum_{\ell_1=d}^{k}\sum_{\ell_2=d}^{k}\sum_{\theta\in\mathcal{I}^*_{\ell_1,\ell_2}} \frac{1_{|\theta|\le d}}{2}\left(p_{\ell_1}p_{\ell_2}p_{\ell_1+\theta}p_{\ell_2-\theta}\right)^s z^{\ell_1+\ell_2+2}\\ \tau_3(s,z) &= \sum_{\ell_1=d}^{k}\sum_{\ell_2=d}^{k}\frac{1_{\ell_1+\ell_2+1>k}}{2\min\{k,\ell_1+\ell_2-d\}-(\ell_1+\ell_2)+1}\\ &\quad\left(\sum_{\theta\in I_{\ell_1,\ell_2}} p_{\ell_1+\theta}p_{\ell_2-\theta}\right)^s z^{\ell_1+\ell_2+2},\end{aligned}$$

with $d(\ell)=\min\{d,\ell-d\}$.

In particular, for $k\le 2d$ we have the following simplifications:

$$\alpha(s,z)=\sum_\ell(2d-\ell)p_\ell^s z^{\ell+1},$$

and

$$\begin{aligned}\tau_1(s,z) &= \sum_{\ell_1,\ell_2} 2\max\{0,\ell_1+\ell_2-k-d\}p_{\ell_1}^s p_{\ell_2}^s z^{\ell_2+\ell_2+2},\\ \tau_2(s,z) &= \sum_{\ell_1=d}^{k}\sum_{\ell_2=d}^{k}\sum_{\theta\in\mathcal{I}^*_{\ell_1,\ell_2}}\frac{1}{2}(p_{\ell_1}p_{\ell_2}+p_{\ell_1+\theta}p_{\ell_2-\theta})^s z^{\ell_2+\ell_2+2},\\ \tau_3(s,z) &= \sum_{\ell_1=d}^{k}\sum_{\ell_2=d}^{k}\frac{1}{2\min\{k,\ell_1+\ell_2-d\}-(\ell_1+\ell_2)+1}\\ &\quad\times\left(\sum_{\theta\in I_{\ell_1,\ell_2}} p_{\ell_1+\theta}p_{\ell_2-\theta}\right)^s z^{\ell_2+\ell_2+2}.\end{aligned}$$

Proof. As in the previous proof, the main idea is to enumerate all possible ways a sequence x leaves the status of (d,k) after a bit corruption and returns to (d,k) status after a second bit corruption. In other words, $x\in\tilde{A}_*$, $x\oplus e_j\notin\tilde{A}_*$, and $x\oplus e_j\oplus e_i\in\tilde{A}_*$. We consider several cases:

a) Property 1: Let x be a single super-symbol: $x = 0^\ell 1$. Consider now $x \oplus e_j$. First, suppose $\ell \le 2d$ and the error e_j falls on a zero of x. If e_j falls on a zero between $\ell - d$ and d, then

$$0^\ell 1 \oplus e_j = 0^{\ell_1} 1 0^{\ell_2} 1,$$

and at least one of ℓ_1, ℓ_2 is smaller than d. Therefore, $x \oplus e_j$ is not a (d,k) sequence. The only way e_i can produce a (d,k) sequence is when it is equal to e_j: $|B(x \oplus e_j) \cap \tilde{A}_*| = 1$. Assume now $\ell > 2d$. If e_j falls at distance greater than d from both ends, then $x \oplus e_j \in \tilde{A}_*$ and does not leave $\tilde{A}_*$.

b) Property 2: If the error e_j falls on a symbol $0^{\ell_1}1$ in $x = 0^{\ell_1} 1 0^{\ell_2} 1$, on the last $\min\{d, \ell_1 - d\}$ zeros, then with $\theta \le \min\{d, \ell_2 - d\}$

$$0^{\ell_1} 1 0^{\ell_2} 1 \oplus e_j = 0^{\ell_1 - \theta} 1 0^{\theta - 1} 1 0^{\ell_2} 1,$$

and $x \notin \tilde{A}_*$. We have:

- if it falls also on the last $\min\{d, \ell_1 - d, k - \ell_2\}$ zeros, *i.e* $\theta \le \min\{d, \ell_1 - d, k - \ell_2\}$, then the only e_i that moves $x \oplus e_i \oplus e_j$ back a (d,k) sequence is either $e_j = e_i$ or e_j such that it falls on the 1 of $0^{\ell_1}1$, and $|B(x \oplus e_j) \cap \tilde{A}_*| = 2$,
- otherwise, the only acceptable j is i, so that $|B(x \oplus e_j) \cap \tilde{A}_*| = 1$ and $x \oplus e_j \notin \tilde{A}_*$.

c) Property 2bis: If the error e_j in $x = 0^{\ell_1} 1 0^{\ell_2} 1$ falls on the first $\min\{d, \ell_2 - d\}$ zeros of $0^{\ell_2}1$, then

- if it falls also on the first $\min\{d, \ell_2 - d, k - \ell_1\}$ zeros, then the only e_j that moves $x \oplus e_i \oplus e_j$ back a (d,k) sequence is either $e_j = e_i$ or e_j such that it falls on the 1 of $0^{\ell_1}1$, and $|B(x \oplus e_j) \cap \tilde{A}_*| = 2$,
- otherwise, the only acceptable j is i so that $|B(x \oplus e_j) \cap \tilde{A}_*| = 1$ and $x \oplus e_j \notin \tilde{A}_*$.

d) Property 3: We still consider $x = 0^{\ell_1} 1 0^{\ell_2} 1$. If the error falls on the "1" of $0^{\ell_1}1$, then the only e_j that moves $x \oplus e_j \oplus e_i$ back (d,k) sequences are those that either fall back on the 1, or on the $\min\{\ell_2 - d, k - \ell_1\}$ first zeros of $0^{\ell_2}1$, or on the $\min\{\ell_1 - d, k - \ell_2\}$ last zeros of $)0^{\ell_1}1$, and then

$$\begin{aligned} |B(x \oplus e_j) \cap \tilde{A}_*| &= 1 + \min\{\ell_1 - d, k - \ell_2\} \\ &\quad + \min\{\ell_2 - d, k - \ell_1\} \\ &= 1 + 2\min\{k, \ell_1 + \ell_2 - d\} - \ell_1 - \ell_2. \end{aligned}$$

Clearly, then we must have $\ell_1 + \ell_2 + 1 > k$ in order $x \oplus e_j \notin \tilde{A}_m$.

Given these four properties we can define the following quantities

$$\alpha(s,z) = \sum_\ell \max\{0, 2d - \ell\} p_\ell^s z^{\ell+1}$$

and $\tau(s,z) = \tau_1(s,z) + \tau_2(s,z) + \tau_3(s,z)$ with the convention that $\alpha(s,z)$ corresponds to Property 1, $\tau_1(s,z)$ to Property 2 and 2*bis* (second bullet), $\tau_2(s,z)$ to Property 2 and Property 2*bis* (first bullet), $\tau_3(s,z)$ to Property 3. This completes the proof. ∎

C. Asymptotic analysis

Finally, we can re-interpret our results for reduced (d,k) sequences of the variable-length model in terms of the original (d,k) sequences of fixed length. Our aim is to provide an asymptotic evaluation of $\phi_n(1)$, $Q_n(1)$, $\phi_n'(1)$ and $Q_n'(1)$ as $n \to \infty$. To this end, we will present an asymptotic evaluation of $\phi_n(s)$ and $Q_n(s)$.

From (30) and (31) we easily find

$$\begin{aligned} \phi(s,z) &= \sum_m \tilde{\phi}_m(s,z) = \frac{b_1(s,z) + b_2(s,z)}{(1 - r(s,z))^2}, \\ Q(s,z) &= \sum_m \tilde{Q}_m(s,z) = \frac{\alpha(s,z) + \tau(s,z)}{(1 - r(s,z))^2}. \end{aligned}$$

Then by Cauchy formula applied to (28) and (29)

$$\begin{aligned} P_n^s \phi_n(s,z) &= \frac{1}{2i\pi} \oint \phi(s,z) \frac{dz}{z^{n+1}}, \\ P_n^s Q_n(s,z) &= \frac{1}{2i\pi} \oint Q(s,z) \frac{dz}{z^{n+1}}. \end{aligned}$$

A simple application of the residue analysis leads to

$$\begin{aligned} P_n^s \phi_n(s) &= \frac{\rho^{-n-1}(s)}{\lambda(s)^2} \Big((n+1)(b_1(s,\rho(s)) + b_2(s,\rho(s))) \\ &\quad - \frac{\partial}{\partial z} b_1(s,\rho(s)) - \frac{\partial}{\partial z} b_1(s,\rho(s)) \Big) + O(\mu^{-n}), \\ P_n^s Q_n(s) &= \frac{\rho^{-n-1}(s)}{\lambda(s)^2} \Big((n+1)(\alpha(s,\rho(s)) + \tau(s,\rho(s))) \\ &\quad - \frac{\partial}{\partial z} \alpha(s,\rho(s)) - \frac{\partial}{\partial z} \tau(s,\rho(s)) \Big) + O(\mu^{-n}). \end{aligned}$$

Since functions involved are analytic and uniformly bounded in s in a compact neighborhood, the asymptotic estimates of $\phi_n'(1)$ and $Q_n'(1)$ can be easily derived.

In summary, we find

$$\begin{aligned} \phi_n'(1) + Q_n'(1) &= -(n+1)\rho'(1)(\phi_n(1) + Q_n(1)) + O(n) \\ &= -n\mathbf{H}(X_1^n) + O(n), \end{aligned}$$

which cancels the coefficient $n\varepsilon\mathbf{H}(X_1^n)$ in the expansion of $\mathbf{H}(Z_1^n)$ in (25). More precisely,

$$\begin{aligned} \phi_n'(1) + Q_n'(1) &= -n\mathbf{H}(X_1^n) + n\log\lambda - 2\frac{\lambda'(1)}{\lambda} \\ &+ \frac{n}{\lambda}\Big(\frac{\partial}{\partial s} b_1(1,1) + \frac{\partial}{\partial s} b_2(1,1) \\ &+ \frac{\partial}{\partial s}\alpha(1,1) + \frac{\partial}{\partial s}\tau(1,1) \\ &\quad \rho'(1)\Big(\frac{\partial^2}{\partial s \partial z} b_1(1,1) + \frac{\partial^2}{\partial s \partial z} b_2(1,1) \\ &+ \frac{\partial^2}{\partial s \partial z}\alpha(1,1) + \frac{\partial^2}{\partial s \partial z}\tau(1,1) \Big)\Big) \\ &+ n\frac{\rho'(1)}{\lambda}\Big(\frac{\partial}{\partial z} b_1(1,1) + \frac{\partial}{\partial z} b_2(1,1) \\ &+ \frac{\partial}{\partial z}\alpha(1,1) + \frac{\partial}{\partial z}\tau(1,1) \Big) + O(1). \end{aligned} \quad (32)$$

The expression for $f_0(P^{\max})$ in Theorem 5 follows directly from the expression (32) since the coefficient at ε is exactly $n\mathbf{H}(X_1^n) + \phi'_n(1) + Q'_n(1) + \phi_n(1)$ and $\phi_n(1) = 0$ when $k \le 2d$. The proof of Theorem 6 is even easier since

$$f_0(P^{\max}) = \frac{Q_n(1)}{n} = 1 - \frac{\phi_n(1)}{n}.$$

We have from (32):

$$\phi_n(1) = n\left(\frac{b_1(1,1) + b_2(1,1)}{\lambda}\right).$$

Observe that $b_1(1,1)$ exactly matches γ and $b_2(1,1)$ matches δ in Theorem 6.

D. Error Term in Theorem 5

To complete the proof of Theorem 5, we establish here that the dominating error term of the capacity $C(S,\varepsilon)$ estimation is $O(\varepsilon^2 \log^2 \varepsilon)$. For this we need to show that the maximizing distribution $P_X^{\max}(\varepsilon)$ $H(Z)$ introduces error of order $O(\varepsilon^2 \log^2 \varepsilon)$. Recall that $P^{\max}$ maximizes $H(X)$.

In Appendix A we show that

$$\frac{\partial}{\partial \epsilon} H(Z) = O(\log \varepsilon)$$

uniformly in P_X. As a consequence $H(Z)$ converges to $H(X)$ uniformly in P_X as $\varepsilon \to 0$. We also prove in the Appendix that

$$H(Z) = H(X) + f_0(P_X)\varepsilon \log \varepsilon + f_1(P_X)\varepsilon + g(P_X) O(\varepsilon^2 \log \varepsilon),$$

where the functions f_0, f_1 and g of P_X are in C_∞. Let $P_X^{\max}(\varepsilon)$ be the distribution that maximizes $H(Z)$, hence the capacity $C(S,\varepsilon)$. Let $\alpha > 0$ and let K_α be a compact set of distributions that are at topological distance smaller than or equal to α from $P_X^{\max}$. Since $H(Z)$ converges to $H(X)$ uniformly, there exists $\varepsilon' > 0$ such that $\forall \varepsilon < \varepsilon', \varepsilon > 0$ we have $P_X^{\max} \in K_\alpha$.

Let now $\beta = \max_{P_X \in K_\alpha}\{g(P_X)\}$. Clearly, $\beta \to g(P^{\max})$ as $\alpha \to 0$. Let also

$$F(P_X, \varepsilon) = H(X) + f_0(P_X)\varepsilon \log \varepsilon + f_1(P_X)\varepsilon,$$

and

$$F_\alpha(\varepsilon) = \max_{P_X \in K_\alpha}\{F(P_X, \varepsilon)\}.$$

The following inequality for $\varepsilon < 1$ follows from our analysis in Appendix A

$$F_\alpha(\varepsilon) + \beta\varepsilon^2 \log \varepsilon \le H(P_X^{\max}(\varepsilon)) \le F_\alpha(\varepsilon) - \beta\varepsilon^2 \log \varepsilon.$$

We will prove here that $F_\alpha(\varepsilon) = F(P_X^{\max}, \varepsilon) + O(\varepsilon^2 \log^2 \varepsilon)$.

Let $\tilde{P}_X^{\max} = \arg\max\{F(P_X, \varepsilon)\}$. We have $\nabla F(\tilde{P}_X^{\max}, \varepsilon) = 0$, where ∇F denotes the gradient of F with respect to P_X. Defining $dP_X = \tilde{P}_X^{\max} - P_X^{\max}$ we find

$$\begin{aligned} \nabla F(\tilde{P}_X^{\max}, \varepsilon) &= \nabla^2 F(P_X^{\max}, \varepsilon) dP_X \\ &+ \nabla f_0(P_X^{\max})\varepsilon \log \varepsilon + \nabla f_1(P_X^{\max})\varepsilon \\ &+ O(\|dP_X\|^2 + \|dP_X\|\varepsilon \log \varepsilon), \end{aligned}$$

where $\nabla^2 F$ is the second derivative matrix (i.e., Hessian) of F and $\|v\|$ is the norm of vector v. Denoting $F_2 = \nabla^2 F(P_X^{\max})$ and its inverse matrix as F_2^{-1}, we arrive at

$$\begin{aligned} dP_X &= -F_2^{-1} \cdot (\nabla f_0(P_X^{\max})\varepsilon \log \varepsilon + \nabla f_1(P_X^{\max})\varepsilon) \\ &+ O(\varepsilon^2 \log^2 \varepsilon). \end{aligned}$$

Since

$$\begin{aligned} F(\tilde{P}_X^{\max}, \varepsilon) &= F(P_X^{\max}, \varepsilon) + \frac{1}{2} dP_X \cdot F_2 \cdot dP_X \\ &+ \nabla f_0(P_X^{\max}) dP_X \varepsilon \log \varepsilon \\ &+ \nabla f_1(P_X^{\max}) dP_X \varepsilon + O(\varepsilon^3 \log^3 \varepsilon), \end{aligned}$$

we obtain for $\|dP_X\| \le \alpha$ (for sufficiently small ε) :

$$\begin{aligned} F_\alpha(\varepsilon) &= F(P_X^{\max}, \varepsilon) \\ &- \frac{1}{2}\nabla f_0(P_X^{\max}) \cdot F_2^{-1} \cdot \nabla f_0(P_X^{\max})\varepsilon^2 \log^2 \varepsilon \\ &- \nabla f_0(P_X^{\max}) \cdot F_2^{-1} \cdot \nabla f_1(P_X^{\max})\varepsilon^2 \log \varepsilon \\ &- \frac{1}{2}\nabla f_1(P_X^{\max}) \cdot F_2^{-1} \cdot \nabla f_1(P_X^{\max})\varepsilon^2 \\ &+ O(\varepsilon^3 \log^3 \varepsilon). \end{aligned}$$

This completes the proof.

V. Conclusion

We study the capacity of the constrained BSC channel in which the input is a (d,k) sequence. After observing that a (d,k) sequence can be generated by a k-order Markov chain, we reduce the problem to estimating the entropy rate of the underlying hidden Markov process (HMM). In our previous paper [14], [15], we established that the entropy rate for a HMM process is equal to a Lyapunov exponent. After realizing that such an exponent is hard to compute, theoretically and numerically, we obtained an asymptotic expansion of the entropy rate when the error rate ε is small (cf. also [27]).

In this paper, we extend previous results in several directions. First, we present asymptotic expansion of the HMM when some of the transition probabilities of the underlying Markov are zero. This adds additional term of order $\varepsilon \log \varepsilon$ to the asymptotic expansion. Then, we return to the noisy constrained capacity and prove that the exact capacity is related to supremum of Lyapunov exponents over increasing order Markov processes. Finally, for (d,k) sequences we obtain an asymptotic expansion for the noisy capacity when the noise $\varepsilon \to 0$. In particular, we prove that for $k \le 2d$ the noisy capacity is equal to the noiseless capacity plus a term $O(\varepsilon)$. In the case $k > 2d$, the correction term is $O(\varepsilon \log \varepsilon)$. We should point out that recently Han and Marcus [9], [10] reached similar conclusions (and obtained some generalizations) using quite different methodology.

Appendix A: Proof of Theorem 3

In this Appendix we prove the error term of (18) in Theorem 3 using the methodology developed by us in [15]. We need to prove that for $\varepsilon < 1/2$

$$\mathbf{H}(Z_1^n) = \mathbf{H}(X_1^n) + n f_1(P_X)\varepsilon + n f_0(P_X)\varepsilon \log \varepsilon + O(n\varepsilon^2 \log \varepsilon) \quad (33)$$

for some $f_1(P_X)$ and $f_0(P_X)$. We start with

$$\mathbf{H}(Z_1^n) = \mathbf{H}(X_1^n) - \varepsilon \frac{\partial}{\partial \epsilon}\mathbf{H}(Z_1^n) + G_n \quad (34)$$

and show at the end of this section that $G_n = O(n\varepsilon^2 \log \varepsilon)$.

We first concentrate on proving that

$$\frac{\partial}{\partial \epsilon}\mathbf{H}(Z_1^n) = n f_1(P_X) + n f_0(P_X)\log\varepsilon \quad (35)$$

for some $f_0(P_X)$ and $f_1(P_X)$. We use equation (48) from [15] which we reproduce below

$$\frac{\partial}{\partial \epsilon}P_Z(z) = \frac{1}{1-2\varepsilon}\sum_i \left(P_Z(z\oplus e_i) - P_Z(z)\right)$$

for any sequence z of length n (hereafter, we simply write x for x_1^n and z for z_1^n). Consequently,

$$\frac{\partial}{\partial \epsilon}\mathbf{H}(Z_1^n) = -\frac{1}{1-2\varepsilon}\sum_z\sum_i (P_Z(z\oplus e_i) - P_Z(z))\log P_Z(z)$$

that can be rewritten as

$$\frac{\partial}{\partial \epsilon}\mathbf{H}(Z_1^n) = -\frac{1}{1-2\varepsilon}\sum_x\sum_i P_Z(z)\log\frac{P_Z(z\oplus e_i)}{P_Z(z)}.$$

In order to estimate the ratio of $P_Z(z\oplus e_i)$ and $P_Z(z)$, we observe that

$$P_Z(z) = (1-\varepsilon)^n \sum_x P_X(x)\left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(x,z)},$$

where $d_H(,x,z)$ is the Hamming distance between x and z. Similarly,

$$P_Z(z\oplus e_i) = (1-\varepsilon)^n \sum_x P_X(x)\left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(x,z\oplus e_i)}.$$

The following inequality is easy to prove

$$\begin{aligned}\min_i \left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(x,z\oplus e_i)-d_H(x,z)} &\leq \frac{P_Z(z\oplus e_i)}{P_Z(z)} \\ &\leq \max_i \left(\frac{\varepsilon}{1-\varepsilon}\right)^{d_H(x,z\oplus e_i)-d_H(x,z)}.\end{aligned}$$

Since $d_H(x, z\oplus e_i) = d_H(x,z)\pm 1$ we conclude that

$$\frac{\varepsilon}{1-\varepsilon} \leq \frac{P_Z(z\oplus e_i)}{P_Z(z)} \leq \frac{1-\varepsilon}{\varepsilon}.$$

Thus

$$\left|\sum_z\sum_i P_Z(z)\log\frac{P_Z(z\oplus e_i)}{P_Z(z)}\right| \leq -n\log(1-\varepsilon) - n\log\varepsilon$$

and this completes the proof of (33).

To finish the proof of Theorem 3, it remains to show that that $G_n = O(n\varepsilon^2\log\varepsilon)$, that is, uniformly in n and $\varepsilon > 0$

$$\mathbf{H}(Z_1^n) = \mathbf{H}(X_1^n) - \varepsilon\frac{\partial}{\partial\epsilon}\mathbf{H}(Z_1^n) + O(n\varepsilon^2\log\varepsilon). \quad (36)$$

To this end, we make use of the Taylor expansion:

$$\begin{aligned}\mathbf{H}(Z_1^n) &= \mathbf{H}(X_1^n) - \varepsilon\frac{\partial}{\partial\epsilon}\mathbf{H}(Z_1^n) \\ &\quad - \int_0^\varepsilon \theta \frac{\partial^2}{\partial\epsilon^2}\mathbf{H}(Z_1^n)|_{\varepsilon=\theta}d\theta,\end{aligned}$$

and prove that for ε small enough we have uniformly in n and $\varepsilon > 0$

$$\frac{\partial^2}{\partial\epsilon^2}\mathbf{H}(Z_1^n) = O(n\log\varepsilon), \quad (37)$$

from which the error term $O(n\varepsilon^2\log\varepsilon)$ follows immediately.

In [15] we proved that for all sequences z

$$\frac{\partial^2}{\partial\epsilon^2}P_Z(z) = -\frac{2}{1-2\varepsilon}\frac{\partial}{\partial\epsilon}P_Z(z) - \frac{1}{(1-2\varepsilon)^2}\sum_{i,j}$$
$$\times\left(P_Z(z\oplus e_i\oplus e_j) - P_Z(z\oplus e_i) - P_Z(z\oplus e_j) + P_Z(z)\right),$$

which led to equation (49) of [15] repeated below

$$\frac{\partial^2}{\partial\epsilon^2}\mathbf{H}(Z_1^n) = -\frac{2}{1-2\varepsilon}\frac{\partial}{\partial\epsilon}\mathbf{H}(Z_1^n) - \frac{1}{(1-2\varepsilon)^2}(D_1+D_2),$$

where

$$\begin{aligned}D_1 &= \sum_z\sum_{i,j} P_Z(z\oplus e_i\oplus e_j) - P_Z(z\oplus e_i) \\ &\quad - P_Z(z\oplus e_j) + P_Z(z)\log P_Z(z),\end{aligned}$$

and

$$\begin{aligned}D_2 &= \sum_z\sum_{ij}(P_Z(z\oplus e_i)) - P_Z(z)) \\ &\quad \times(P_Z(z\oplus e_j)) - P_Z(z))\frac{1}{P_Z(z)}.\end{aligned}$$

We will prove that $D_1 = O(n\log\varepsilon)$ and $D_2 = O(n)$.

Let first deal with D_1. We can write it as

$$D_1 = \sum_z\sum_{i,j} P_Z(z)\log\frac{P_Z(z\oplus e_i\oplus e_j)P_Z(z)}{P_Z(z\oplus e_i)P(z\oplus e_j)}.$$

We now split $D_1 = D_1' + D_1''$ where D_1' involves the pairs (i,j) such that $|i-j| \leq k+1$ and D_1'' deals with such pairs that $|j-i| > k+1$. For all z and all i and j such that $|i-j|\leq k+1$, we have

$$\frac{\varepsilon^2}{(1-\varepsilon)^2} < \frac{P_Z(z\oplus e_i\oplus e_j)P_Z(z)}{P_Z(z\oplus e_i)P_Z(z\oplus e_j)} < \frac{(1-\varepsilon)^2}{\varepsilon^2}. \quad (38)$$

Therefore,

$$\begin{aligned}|D_1'| &\leq \sum_z \sum_{|j-i|\leq k+1} P_Z(z)(-2\log(1-\varepsilon) - 2\log\varepsilon) \\ &\leq (k+1)n(-2\log(1-\varepsilon) - \log\varepsilon).\end{aligned}$$

For $|j-i| > k+1$, we observe, as in [15], that there exists $\mu < 1$ such that for all z

$$\begin{aligned}\frac{P_Z(z\oplus e_i\oplus e_j)P_Z(z)}{P_Z(z\oplus e_i)P_Z(z\oplus e_j)} &= 1 + O(\mu^i) + O(\mu^j) + O(\mu^{|j-i|}) \\ &\quad + O(\mu^{n-i}) + O(\mu^{n-i}).\end{aligned}$$

Thus we find

$$\begin{aligned} D_1'' &= \sum_z \sum_{|j-i|>k+1} P_Z(z) \log\big(1 + O(\rho^i) + O(\mu^j) \\ &\quad + O(\mu^{|j-i|}) + O(\mu^{n-i}) + O(\mu^{n-i})\big) \\ &= \sum_z P_Z(z) O(n/(1-\mu)) = O(n). \end{aligned}$$

Now we turn our attention to D_2, and similarly we split $D_2 = D_2' + D_2''$ with D_2' involving only i, j such that $|i-j| \leq k+1$ and D_2'' involving i, j such that $|i-j| > k+1$. We easily see that $|D_2'| \leq n(k+1)$, and then

$$\begin{aligned} D_2'' &= \sum_z \sum_{|i-j|>k} P_Z(z) - P_Z(z \oplus e_i) \\ &- P_Z(z \oplus e_j) + P_Z(z \oplus e_i \oplus e_j) \\ &+ \left(\frac{P_Z(z \oplus e_i) P_Z(z \oplus e_j)}{P_Z(z \oplus e_i \oplus e_j) P_Z(z)} - 1 \right) P_Z(z \oplus e_i \oplus e_j, \varepsilon). \end{aligned}$$

We now notice that

$$\sum_z \sum_{i,j} P_Z(z) - P_Z(z \oplus e_i) - P_Z(z \oplus e_j) + P_Z(z \oplus e_i \oplus e_j) = 0.$$

Restricting this sum to $|i-j| > k+1$ we observe that it gives the opposite of the sum for $|i-j| \leq k+1$. Therefore, the total contribution is $O((k+1)n$. Furthermore,

$$\begin{aligned} &\sum_z \sum_{|i-j|>k+1} \left(\frac{P_Z(z \oplus e_i) P_Z(z \oplus e_j)}{P_Z(z \oplus e_i \oplus e_j) P_Z(z)} - 1 \right) \\ &\times P_Z(z \oplus e_i \oplus e_j) = \sum_z P_Z(z) O(n/(1-\mu)) = O(n), \end{aligned}$$

and this completes the proof of Theorem 3.

Acknowledgment

The authors thank B. Marcus for very helpful discussions during the summer of 2006 when this research was shaping up. We also thank G. Seroussi for participating in the initial stage of this research.

References

[1] 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.

[2] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, (Prague, Czechoslovakia), pp. 13–20, Pub. House Chechoslovak Acad. Sci., 1957.

[3] J. Chen and P. Siegel, "Markov processes asymptotically achieve the capacity of finite-state intersymbol interference channels," *IEEE Trans. Information Theory*, **54**, 1295-1303, 2008.

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

[5] H. Furstenberg and H. Kesten, "Products of random matrices," *Ann. Math. Statist*, pp. 457–469, 1960.

[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, 2005.

[7] G. Han and B. Marcus, Analyticity of Entropy Rate of Hidden Markov Chains,*IEEE Trans. Information Theory*, **52**, 5251–5266, 2006.

[8] G. Han and B. Marcus, Analyticity of Entropy Rate in Families of Hidden Markov Chains (II), ,*IEEE Trans. Information Theory*, **52**, 103-107, 2006.

[9] G. Han and B. Marcus, "Capacity of noisy constrained channels, *Proc. ISIT 2007*, 991-995, Nice, 2007.

[10] G. Han and B. Marcus, Asymptotics of the input-constrained binary symmetric channel capacity, *Annals of Applied Probability*, 2009, to appear.

[11] T. Holliday, A. Goldsmith, and P. Glynn, "Capacity of Finite State Channels Based on Lyapunov Exponents of Random Matrices," *IEEE Trans. Information Theory*, **52**, 3509- 3532, 2006.

[12] K. A. Schouhammer Immink, *Codes for Mass Data Storage Systems*, Shannon Foundation Publishers, Eindhoven, 2004.

[13] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *IEEE Trans. Inform. Theory*, **45**, 1072–1081, 1999.

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

[15] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the Entropy of a Hidden Markov Process" (full version) *Theoretical Computer Science*, 395, 203-219, 2008.

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

[17] S. Karlin and H. Taylor, *A First Course in Stochastic Processes*. New York: Academic Press, 1975.

[18] 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.

[19] E. Ordentlich and T. Weissman, "New Bounds on th Entropy Rate of Hidden Markov Process," *Information Theory Workshop*, 117–122, San Antonio, 2004.

[20] E. Ordentlich and T. Weissman, "On the optimality of symbol by symbol filtering and denoising," *IEEE Trans. Information Theory*, **52**, 19-40, 2006.

[21] V. Oseledec, "A multiplicative ergodic theorem," *Trudy Moskov. Mat. Obshch.*, 1968.

[22] E. Seneta, *Non-Negative Matrices*. New York: John Wiley & Sons, 1973.

[23] 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.

[24] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*. New York: John Wiley & Sons, Inc., 2001.

[25] J. Tsitsiklis and V. Blondel, "The Lyapunov exponent and joint spectral radius of pairs of matrices are hard - when not impossible – to compute and to approximate," *Mathematics of Control, Signals, and Systems*, **10**, 31–40, 1997.

[26] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Trans. Information Theory*, **34**, 45–54, 1988.

[27] O. Zuk, I. Kanter and E. Domany, "Asymptotics of the Entropy Rate for a Hidden Markov Process," *J. Stat. Phys.*, **121**, 343–360, 2005.

[28] 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*, 13, 517 - 520, 2006.