Mutual Information for a Deletion Channel

Michael Drmota
TU Wien
A-1040 Wien, Austria
Email: michael.drmota@tuwien.ac.at

Wojciech Szpankowski
Purdue University
West Lafayette, IN, USA
Email: spa@cs.purdue.edu

Krishnamurthy Viswanathan
Hewlett-Packard Laboratories
Palo Alto, CA, USA
Email: krishnamurthy.viswanathan@hp.com

***Abstract*—We study the binary deletion channel where each input bit is independently deleted according to a fixed probability. We relate the conditional probability distribution of the output of the deletion channel given the input to the *hidden pattern matching* problem. This yields a new characterization of the mutual information between the input and output of the deletion channel. Through this characterization we are able to comment on the the deletion channel capacity, in particular for deletion probabilities approaching 0 and 1.**

I. Introduction

A deletion channel with parameter d takes a binary sequence $x := x_1^n = x_1 \cdots x_n$ where $x_i \in \mathcal{A} = \{0,1\}$ as input and deletes each symbol in the sequence independently with probability d. The output of such a channel is then a *subsequence* $Y = Y(x) = x_{i_1}...x_{i_M}$ of x, where M follows the binomial distribution $\mathrm{Bi}(n,(1-d))$, and the indices $i_1,...,i_M$ correspond to the bits that are *not* deleted. Despite significant effort [2], [3], [5], [9], [10], [11], [12], [14] the mutual information between the input and output of the deletion channel and its capacity are still unknown. Our goal is to provide a more detailed characterization of the mutual information for memoryless sources (extensions to strongly mixing sources or Markovian sources seem likely). Through this characterization we are able to comment on the channel capacity for two special cases: $d \to 1$ and $d \to 0$. The latter case was already discussed in [10], [9]. We derive our results by relating the the conditional probability distribution of the output of the deletion channel given the input to the so called *hidden pattern matching* analyzed recently in [1], [7].

Following [4], the channel capacity of the deletion channel with deletion probability d is

$$C(d) = \lim_{n\to\infty} \frac{1}{n} \sup_{P_{X_1^n}} I(X_1^n; Y(X_1^n)),$$

where $P_{X_1^n}$ is the distribution of X_1^n, and $I(X_1^n;Y(X_1^n))$ is the mutual information between the input and output of the deletion channel. Many bounds have been derived for the capacity (see the survey article by Mitzenmacher [11]).

Let $x = x_1^n \in \{0,1\}^n$ and $w = w_1 w_2 \dots w_m \in \{0,1\}^m$, $m \le n$, be binary sequences. Let $\Omega_x(w)$ denote the number of occurrences of w as a *subsequence* (i.e., not consecutive symbols) of x, that is,

$$\Omega_x(w) = \sum_{1\le i_1<i_2<\cdots<i_m\le n} \mathbf{I}_{[x_{i_1}=w_1]}\mathbf{I}_{[x_{i_2}=w_2]}\cdots\mathbf{I}_{[x_{i_m}=w_m]}, \quad (1)$$

where $\mathbf{I}_A = 1$ if A is true and zero otherwise. The problem of counting subsequences in a text is known as the hidden pattern matching problem and was studied in [1], [7]. In this paper, to derive our results we first represent the mutual information between the input and output of a deletion channel in terms of the count $\Omega_X(w)$ for a random sequence X.

Theorem 1. *For any random input X_1^n, the mutual information satisfies*

$$I(X_1^n;Y(X_1^n)) = \sum_w d^{n-|w|}(1-d)^{|w|}\big(\mathbb{E}[\Omega_{X_1^n}(w)\log\Omega_{X_1^n}(w)] - \mathbb{E}[\Omega_{X_1^n}(w)]\log\mathbb{E}[\Omega_{X_1^n}(w)]\big), \quad (2)$$

where the sum is over all binary sequences of length $\le n$.

From Theorem 1, we have $I(X_1^n;Y(X_1^n)) = S_1(X_1^n,Y(X_1^n)) - S_2(X_1^n,Y(X_1^n)) := S_1 - S_2$ where

$$S_1 = \sum_w d^{n-|w|}(1-d)^{|w|}\mathbb{E}[\Omega_{X_1^n}(w)\log\Omega_{X_1^n}(w)], \quad (3)$$

$$S_2 = \sum_w d^{n-|w|}(1-d)^{|w|}\mathbb{E}[\Omega_{X_1^n}(w)]\log\mathbb{E}[\Omega_{X_1^n}(w)]. \quad (4)$$

In this paper, we focus on memoryless distributions on X_1^n, however, it appears that most of our results extend to larger classes (Markovian). Suppose that $X_1X_2\dots$ is an *i.i.d.* sequence of Bernoulli random variables with parameter p. For such sequences, let $I(d,p) = \lim_{n\to\infty}\frac{1}{n}I(X_1^n;Y(X_1^n))$, and $\lambda(d,p) = \lim_{n\to\infty}\frac{1}{n}S_1(X_1^n,Y(X_1^n))$.

Theorem 2. *For all $0 \le d \le 1$, and $0 \le p \le 1$, the limit $I(d,p)$ as well as the non-negative limits $\lambda(d,p)$ and*

$$\lim_{n\to\infty}\frac{1}{n}S_2(X_1^n,Y(X_1^n)) = H(1-d) - (1-d)H(p)$$

exist, and

$$I(d,p) = \lambda(d,p) + (1-d)H(p) - H(1-d)$$

where, $H(\cdot)$ is the binary entropy function. Furthermore, $I(d,p) = \inf_{n\ge1}\frac{1}{n}I(X_1^n;Y(X_1^n))$, and $\lambda(d,p) = \sup_{n\ge1}\frac{1}{n}S_1(X_1^n,Y(X_1^n))$.

From Theorem 2, $I(d,p) \le I(X_1^1;Y(X_1^1)) = H(p)(1-d)$. When optimized over p, this upper bound matches the capacity asymptotically for $d \to 0$ but not for $d \to 1$, as our next result (Theorem 3) shows. This also implies that $\lambda(d,p) \le H(1-d)$. Note that for $d \to 1$ it is just known that $C(d) = \Theta(1-d)$ [2],

[11], [12]. Our next result is a bound on $I(d,p)$ that implies that, in contrast to the case $d\to 0$, *i.i.d.* distributions over the inputs X_1^n do not asymptotically achieve capacity as $d\to 1$.

Theorem 3. *For all $p\ge 0$, as $d\to 1$*

$$I(d,p)\le K(1-d)^{4/3}\log\frac{1}{1-d}$$

where the constant $K>0$ is absolute.

Finally we demonstrate the strength of our method by reproving Kanoria and Montanari's [10] expansion for $I(d,p)$ for $d\to 0$ leading to $C(d)=I(d,1/2)+O(d^{3/2-\varepsilon})=1+d\log d-Ad+O(d^{3/2-\varepsilon})$ (cf. Theorem 4), where $A=\log(2e)-\sum_{\ell\ge 1}2^{-\ell-1}\ell\log\ell$. Note that the symmetric memoryless distribution is asymptotically optimal in this regime.

II. Proof of Theorem 1 and Capacity Bound

In this section, we first prove Theorem 1 and then present a simple proof of the fact that $C(d)\le 1-d$.

A. Proof of Theorem 1

To prove Theorem 1, we relate hidden pattern matching to the deletion channel through the following observation. For all $x_1^n\in\mathcal{A}^n$

$$P(Y(X_1^n)=w|X_1^n=x_1^n)=\Omega_{x_1^n}(w)d^{n-|w|}(1-d)^{|w|}. \quad (5)$$

We use X and Y to abbreviate X_1^n and $Y(X_1^n)$ respectively. Using (5), we will compute $H(Y)$ and $H(Y|X)$ and use $I(X;Y)=H(Y)-H(Y|X)$ to prove the theorem. We first compute $H(Y)$. Observe that, from (5) $P(Y=w)=\sum_{x\in\mathcal{A}^n}P(X=x)\Omega_x(w)d^{n-|w|}(1-d)^{|w|}$ which leads to

$$H(Y)=-\sum_w d^{n-|w|}(1-d)^{|w|}\left(\mathbb{E}[\Omega_{X_1^n}(w)]\log\mathbb{E}[\Omega_X(w)]\right.$$
$$\left.+\,\mathbb{E}[\Omega_X(w)]\log(d^{n-|w|}(1-d)^{|w|})\right). \quad (6)$$

Next, we compute the conditional entropy $H(Y|X)$. Notice that for $x\in\mathcal{A}^n$ and $y\in\mathcal{A}^m$ we have $P(x,y)=P(x)\Omega_x(y)d^{n-m}(1-d)^m$. Combining this with (5) we obtain

$$H(Y|X)=-\sum_w d^{n-|w|}(1-d)^{|w|}\left(\mathbb{E}[\Omega_X(w)\log\Omega_X(w)]\right.$$
$$\left.+\,\mathbb{E}[\Omega_X(w)]\log d^{n-|w|}(1-d)^{|w|}\right). \quad (7)$$

The theorem follows from (6) and (7).

B. Upper Bound for the Capacity

It is well known that the capacity $C(d)$ of a deletion channel with deletion probability d can be bounded from above by the capacity of an erasure channel with the erasure probability d (e.g., see [3]). We provide a direct proof of this fact. To do so, we first compute the expectation of $\Omega_X(w)$.

Lemma 1. *For any random X_1^n, and all binary sequences w*

$$\mathbb{E}\left[\Omega_{X_1^n}(w)\right]=\binom{n}{|w|}\overline{P}_n(w),$$

where

$$\overline{P}_n(w)=\frac{1}{\binom{n}{|w|}}\sum_{i_1<\cdots<i_m}P(X_{i_1}=w_1,X_{i_2}=w_2,...,X_{i_m}=w_m)$$

with $\sum_{|w|=m}\overline{P}_n(w)=1$. In particular, if X is memoryless, then $\overline{P}_n(w)=P(w)$ where $P(w)$ is the probability that $X_1X_2...X_{|w|}=w$ (see [1] for dynamic X).

Proof: Taking expectation on both sides of (1) we have

$$\sum_{1\le i_1<\cdots<i_m\le n}P(X_{i_1}=w_1,\cdots,X_{i_m}=w_m)=\binom{n}{|w|}\overline{P}_n(w).$$

proving the lemma. ■

Lemma 2. *For any distribution on the input binary random sequence X_1^n, and and deletion probability $d\ge 0$, $I(X_1^n;Y(X_1^n))\le n(1-d)$.*

Proof: Following Theorem 1, we can write $I(X_1^n;Y(X_1^n))=S_1-S_2$ where S_1 and S_2 are defined in (3)–(4). Since $\Omega_X(w)\le\binom{n}{|w|}$ we first have

$$S_1\le\sum_w d^{n-|w|}(1-d)^{|w|}\log\binom{n}{|w|}\mathbb{E}[\Omega_X(w)]$$

and this in combination with Lemma 1 gives us

$$I(X_1^n;Y(X_1^n))$$
$$\le-\sum_w d^{n-|w|}(1-d)^{|w|}\binom{n}{|w|}\overline{P}_n(w)\log\overline{P}_n(w)$$
$$=-\sum_{m=0}^n d^{n-m}(1-d)^m\binom{n}{m}\sum_{|w|=m}\overline{P}_n(w)\log(\overline{P}_n(w)). \quad (8)$$

Since for all $m\ge 0$, $\overline{P}_n(w)$ is a probability distribution over $w\in\mathcal{A}^m$, we have $\sum_{|w|=m}\overline{P}_n(w)\log(1/\overline{P}_n(w))\le\log 2^m=m$, and consequently

$$\sum_{m=0}^n\sum_{|w|=m}d^{n-m}(1-d)^m\binom{n}{m}m=n\cdot(1-d).$$

Substituting this in (8) completes the proof, and also establishes an upper bound of $C(d)\le 1-d$ for the capacity. ■

III. Memoryless Input Distributions

We now restrict the channel input distributions to be memoryless over $\mathcal{A}$ with p denoting the probability of "0". We prove Theorems 2 and 3 in this section.

A. Proof of Theorem 2

The next lemma follows from the definition $\Omega_X(w)$.

Lemma 3. *For all binary sequences w, and all $x^{n+k}\in\mathcal{A}^{n+k}$*

$$\Omega_{x_1^{n+k}}(w)=\sum_{w_1w_2=w}\Omega_{x_1^n}(w_1)\Omega_{x_{n+1}^{n+k}}(w_2), \quad (9)$$

where the sum is taken over all pairs w_1,w_2 such that their concatenation w_1w_2 equals w.

We also require the following lemma.

Lemma 4. *Let z_m and a_m, $1\le m\le M$, be non-negative numbers. Then we have*

$$\sum_{m=1}^M z_m\log\frac{\sum_{m=1}^M z_m}{\sum_{m=1}^M a_m}\le\sum_{m=1}^M z_m\log\frac{z_m}{a_m}. \quad (10)$$

Proof: Apply the inequality $\log x \le x-1$. ∎

Lemma 5. *Let $X_1X_2\dots$ be a memoryless random binary sequence. Then*

$$I(X_1^{n+k};Y(X_1^{n+k})) \le I(X_1^n;Y(X_1^n)) + I(X_1^k;Y(X_1^k)).$$

Proof: We abbreviate $\Omega_{X_1^n}(w_1)$ by $\alpha(w_1)$ and $\Omega_{X_{n+1}^{n+k}}(w_2)$ by $\beta(w_2)$. Applying (9) and (10) we obtain

$$\begin{aligned}
&\Omega_{X_1^{n+k}}(w)\log\Omega_{X_1^{n+k}}(w) - \Omega_{X_1^{n+k}}(w)\log\mathbb{E}\left[\Omega_{X_1^{n+k}}(w)\right]\\
&= \sum_{w_1w_2=w}\alpha(w_1)\beta(w_2)\log\frac{\sum_{w_1w_2=w}\alpha(w_1)\beta(w_2)}{\sum_{w_1w_2=w}\mathbb{E}\left[\alpha(w_1)\beta(w_2)\right]}\\
&\le \sum_{w_1w_2=w}\alpha(w_1)\beta(w_2)\log\frac{\alpha(w_1)\beta(w_2)}{\mathbb{E}\left[\alpha(w_1)\beta(w_2)\right]}\\
&= \sum_{w_1w_2=w}\alpha(w_1)\beta(w_2)\left(\log\frac{\alpha(w_1)}{\mathbb{E}\left[\alpha(w_1)\right]} + \log\frac{\beta(w_2)}{\mathbb{E}\left[\beta(w_2)\right]}\right)
\end{aligned} \tag{11}$$

where the last equality follows holds as $\alpha(w_1)$ and $\beta(w_2)$ are independent. Let now $c_n = I(X_1^n;Y(X_1^n))$. Then, by Theorem 1

$$\begin{aligned}
c_{n+k} = \sum_w d^{n+k-|w|}(1-d)^{|w|}\Big(&\mathbb{E}\left[\Omega_{X_1^{n+k}}(w)\log\Omega_{X_1^{n+k}}(w)\right]\\
&-\mathbb{E}\left[\Omega_{X_1^{n+k}}(w)\right]\log\mathbb{E}\left[\Omega_{X_1^{n+k}}(w)\right]\Big).
\end{aligned}$$

Hence by taking expectations of (11) and using the relation

$$\begin{aligned}
1 &= \sum_{w_1} d^{n-|w_1|}(1-d)^{|w_1|}\mathbb{E}[\Omega_{X_1^n}(w_1)]\\
&= \sum_w d^{n-|w|}(1-d)^{|w|}\binom{n}{|w|}\overline{P}_n(w) = \sum_{\ell=0}^{n} d^{n-\ell}(1-d)^{\ell}\binom{n}{\ell}
\end{aligned}$$

(and a similar relation for the sum over w_2) we immediately derive $c_{n+k}\le c_n + c_k$. Note that we have used the property that X_1^n and X_{n+1}^{n+k} are independent and that X_{n+1}^{n+k} has the same distribution as X_1^k. ∎

By Fekete's lemma [13] the following corollary follows.

Corollary 1. $I(d,p) = \inf_{n\ge1}\frac{1}{n}I(X_1^n;Y(X_1^n))$.

In particular, $I(d,p)\le\frac{1}{n}I(X_1^n;Y(X_1^n))$ for all $n\ge1$. If we apply this for $n=1,2$ we find

$$\begin{aligned}
&I(d,p)\le(1-d)H(p), \text{ and}\\
&I(d,p)\le d(1-d)(H(p)+p^2+q^2-1)+(1-d)^2H(p),
\end{aligned}$$

where $q=1-p$. For example, by looking at the second bound we observe that $\sup_{0\le p\le1}I(d,p)\le\frac{1-d}{2}+(1-d)^2$ which implies that memoryless input distributions do not meet the general upper bound $1-d$ when $d\to1$. Actually we will show that $\sup_{0\le p\le1}I(d,p)$ is much smaller as $d\to1$ (Theorem 3).

We now prove Theorem 2. As above, we write $I(X_1^n;Y(X_1^n)) = S_1 - S_2$. Also, given two sequences a_n and b_n, $a_n\sim b_n$ if $a_n/b_n\to1$ as $n\to\infty$.

Lemma 6. *If X_1^n is a memoryless binary sequence with parameter p, then $S_2\sim n\cdot(H(1-d)-(1-d)H(p))$ as $n\to\infty$.*

Proof: By Theorem 1 and Lemma 1, and by the trivial observation $\sum_{|w|=m}P(w)=1$, we have

$$\begin{aligned}
S_2 &= \sum_w d^{n-|w|}(1-d)^{|w|}\binom{n}{|w|}P(w)\log\binom{n}{|w|}\\
&\quad+\sum_w d^{n-|w|}(1-d)^{|w|}\binom{n}{|w|}P(w)\log P(w)\\
&= \sum_{m=0}^{n} d^{n-m}(1-d)^m\binom{n}{m}\log\binom{n}{m}\\
&\quad+\sum_{m=0}^{n} d^{n-m}(1-d)^m\binom{n}{m}\sum_{|w|=m}P(w)\log P(w).
\end{aligned}$$

The second term above can be computed directly. By the definition of the entropy we have $\sum_{|w|=m}P(w)\log P(w) = -mH(p)$. Consequently,

$$\begin{aligned}
&\sum_{m=0}^{n} d^{n-m}(1-d)^m\binom{n}{m}\sum_{|w|=m}P(w)\log P(w)\\
&= -\sum_{m=0}^{n} d^{n-m}(1-d)^m\binom{n}{m}mH(p) = -n(1-d)H(p).
\end{aligned}$$

In order to evaluate the first term we apply the results of [6], [8] about the so called *binomial sums*. Notice that

$$\sum_{m=0}^{n} d^{n-m}(1-d)^m\binom{n}{m}\log\binom{n}{m}\sim nH(1-d).$$

This completes the proof of the lemma. ∎

The next step is to show a similar property for S_1, namely that $S_1\sim n\cdot\lambda(d,p)$, where $\lambda(d,p)$ is a non-negative constant. The problem is to obtain some information about $\lambda(d,p)$, but for this we would need precise information about the behavior of $\Omega_X(w)$.

Lemma 7. *Suppose that $X_1X_2\dots$ is a binary memoryless sequence and $a_n = S_1(X_1^n, Y(X_1^n))$. Then $a_{n+k}\ge a_n + a_k$.*

Proof: We have

$$\begin{aligned}
\Omega_{X_1^{n+k}}(w)\log\Omega_{X_1^{n+k}}(w) &= \left(\sum_{w_1w_2=w}\Omega_{X_1^n}(w_1)\Omega_{X_{n+1}^{n+k}}(w_2)\right)\\
&\quad\times\log\left(\sum_{\tilde w_1\tilde w_2=w}\Omega_{X_1^n}(\tilde w_1)\Omega_{X_{n+1}^{n+k}}(\tilde w_2)\right)\\
&\ge\sum_{w_1w_2=w}\Omega_{X_1^n}(w_1)\Omega_{X_{n+1}^{n+k}}(w_2)\log\left(\Omega_{X_1^n}(w_1)\Omega_{X_{n+1}^{n+k}}(w_2)\right)\\
&=\sum_{w_1w_2=w}\Omega_{X_1^n}(w_1)\Omega_{X_{n+1}^{n+k}}(w_2)\log\Omega_{X_1^n}(w_1)\\
&\quad+\sum_{w_1w_2=w}\Omega_{X_1^n}(w_1)\Omega_{X_{n+1}^{n+k}}(w_2)\log\Omega_{X_{n+1}^{n+k}}(w_2)
\end{aligned}$$

and consequently

$$a_{n+k} = \sum_w d^{n+k-|w|}(1-d)^{|w|}\,\mathbb{E}\left[\Omega_{X_1^{n+k}}(w)\log\Omega_{X_1^{n+k}}(w)\right]$$
$$\ge \sum_w \sum_{w_1w_2=w} d^{n+k-|w_1|-|w_2|}(1-d)^{|w_1|+|w_2|}$$
$$\times\,\mathbb{E}\left[\Omega_{X_1^n}(w_1)\Omega_{X_{n+1}^{n+k}}(w_2)\log\Omega_{X_1^n}(w_1)\right]$$
$$+\sum_w \sum_{w_1w_2=w} d^{n+k-|w_1|-|w_2|}(1-d)^{|w_1|+|w_2|}$$
$$\times\,\mathbb{E}\left[\Omega_{X_1^n}(w_1)\Omega_{X_{n+1}^{n+k}}(w_2)\log\Omega_{X_{n+1}^{n+k}}(w_2)\right].$$

Hence, as in Lemma 5, we obtain that $a_{n+k} \ge a_k + a_n$. ∎

The superadditivity property of Lemma 7 provides the following convergence result.

Lemma 8. *If $X = X_1^n$ is a binary memoryless sequence, then there exists a non-negative constant $\lambda(d,p) \le H(1-d)$ such that $S_1 \sim n\cdot\lambda(d,p)$ as $n\to\infty$.*

Proof: Since $\Omega_X(w)$ is a non-negative integer we certainly have $S_1 \ge 0$. Furthermore, since $\Omega_X(w) \le \binom{n}{|w|}$ it follows (as in the proof of Lemma 6) that

$$S_1 \le \sum_w d^{n-|w|}(1-d)^{|w|}\,\mathbb{E}[\Omega_X(w)]\log\binom{n}{|w|} \sim nH(1-d).$$

Hence (using the notation $a_n = S_1(X_1^n, Y(X_1^n))$)

$$0 \le\ \lambda(d,p) := \sup_{n\ge 1}\frac{a_n}{n}\ \le H(1-d).$$

By another application of Fekete's lemma [13] the sequence a_n/n has a limit that equals the supremum $\sup(a_n/n)$. We have used the property $a_{n+k} \ge a_n + a_k$ here. ∎

The proof of Theorem 2 is a combination of Lemma 6 and Lemma 8. The lower bound on $\lambda(d,p)$ follows from the fact that $I(d,p)\ge 0$.

Remark (Extension to Mixing Sources): Most results of this section hold for more general distributions. For example, from the proof of Lemma 6 we conclude that

$$S_2 \sim n\cdot\left(H(1-d) - (1-d)H(\bar{P})\right)$$

where $\bar{P}$ is the limit of $\bar{P}_n$ which was defined in Lemma 1 (provided the limit exists). A distribution $P(X_1^n)$ is said to correspond to a strongly mixing source [13] if for all $m\le n$, there exist constants c_1, c_2 such that

$$c_1 P(X_1^m)P(X_{m+1}^n) \le P(X_1^n) \le c_2 P(X_1^m)P(X_{m+1}^n).$$

For such distributions, Lemma 7 generalizes to $a_{n+k} \ge a_n + a_k + K_1$ for some constant K_1, hence Lemma 8 holds as well.

B. Proof of Theorem 3: $d\to 1$

We consider the expression in (2). We first note that the empty word does not contribute to the sum (2). Next we consider words of length 1. If $w=0$ and if $X = X_1^n$ consists of m zeroes and $n-m$ ones then $\Omega_X(w) = m$. The situation is completely symmetric if $w=1$. Hence the contribution of words of length 1 to $I(X^n;Y(X^n))$ is

$$T_1 := d^{n-1}(1-d)\left(\sum_{m=1}^n m\log m\binom{n}{m}\left(p^m q^{n-m} + p^{n-m}q^m\right)\right)$$
$$-\,d^{n-1}(1-d)\left(np\log(np) + nq\log(nq)\right)$$

where $q = 1-p$. By using the inequality

$$\log m = \log(np) + \log\left(1 + \frac{m-np}{np}\right) \le \log(np) + \frac{m-np}{np}$$

we obtain that

$$\sum_{m=1}^n m\log m\binom{n}{m}p^m q^{n-m}$$
$$\le \sum_{m=1}^n m\left(\log np + \frac{m-np}{np}\right)\binom{n}{m}p^m q^{n-m}$$
$$= \log(np)np + \frac{npq}{np} = np\log(np) + q.$$

Putting all parts together we obtain that $T_1 \le d^{n-1}(1-d) \le (1-d)$.

Let T_2 denote the subsum of (2) corresponding to those terms with $|w|\ge 2$. By using the trivial estimate $\Omega_X(w) \le \binom{n}{|w|}$ and taking absolute values we obtain the upper bound

$$T_2 \le 2\sum_{\ell=2}^n d^{n-\ell}(1-d)^\ell\binom{n}{\ell}\log\binom{n}{\ell}$$
$$\le 2\sum_{\ell=2}^n d^{n-\ell}(1-d)^\ell\frac{n^\ell}{\ell!}\log n^\ell$$
$$= 2d^n\log n\sum_{\ell\ge 2}\left(\frac{n(1-d)}{d}\right)^\ell\frac{1}{(\ell-1)!}$$
$$\le 2d^n\log n\frac{n(1-d)}{d}\left(e^{n(1-d)/d} - 1\right).$$

If $n(1-d) = o(1)$ this leads to $T_2 \le C_1 n^2(1-d)^2\log n$ for some absolute constant $C_1 > 0$. Summing up and using Corollary 1, we obtain that

$$I(d,p) \le \frac{1}{n}I(X_1^n;Y(X_1^n)) \le \frac{1-d}{n} + C_1 n(1-d)^2\log n.$$

Finally by choosing $n = \lfloor(1-d)^{-1/3}\rfloor$ we derive the upper bound

$$I(d,p) \le K\,(1-d)^{4/3}\log\frac{1}{1-d}$$

for an absolute constant $K > 0$.

C. Lower Bound for $d\to 0$

Finally, we comment on the case $d\to 0$ that has been already solved in [10] and [9] where it is shown that $I(d,0.5) = 1 + d\log d - Ad + O(d^{2-\varepsilon})$ as $d\to 0$ and $C(d) = I(d,0.5) + O(d^{3/2-\varepsilon})$. The approach presented in [10] is quite different from ours. However, we can use our methods to obtain corresponding bounds. In particular, we easily obtain the following lower bound for $I(d,p)$.

Theorem 4. *As $d \to 0$,*

$$I(d,p) \geq (1-d)H(p) + d\log d - d\log(e) + d(q^2 f(p) + p^2 f(p)) + O\left(d^{2-\varepsilon}\right) \qquad (12)$$

for every $\varepsilon > 0$, where $f(x)$ denotes the function $f(x) = \sum_{\ell \geq 2} x^\ell \, \ell \log \ell$ and $q = 1-p$. Furthermore, as $d \to 0$,

$$I(d,p) \leq H(p) + d\log d + O(d \log\log(1/d)). \qquad (13)$$

Proof: The lower bound for $I(d,p)$ follows from ideas similar to those in the proof of Theorem 2. Instead of taking the limit of a_n/n defined in Lemma 7 we derive lower bounds for a_n/n for certain n. We will only consider words w with $|w| = n-1$. Then

$$a_n \geq d(1-d)^{n-1} \sum_{|w|=n-1} \mathbb{E}[\Omega_X(w) \log \Omega_X(w)].$$

Suppose for the moment that w has the form $w = 0^{i_1}1^{j_1}0^{i_2}1^{j_2}\cdots 0^{i_K}1^{j_K}$, where $i_r, j_r \geq 1$; this means that $w_1 = 0$ and $w_{n-1} = 1$ (the other cases can be handled in completely the same way). If $|w| = n-1$, then we have $\Omega_X(w) = \ell$ (for some $\ell > 2$) if and only if there exists r with

$$i_r = \ell - 1 \quad \text{and} \quad X = 0^{i_1}1^{j_1}\cdots 1^{j_{r-1}}0^{i_r+1}1^{j_r}\cdots 0^{i_K}1^{j_K}$$

or there exists r with

$$j_r = \ell - 1 \quad \text{and} \quad X = 0^{i_1}1^{j_1}\cdots 0^{i_r}1^{j_r+1}0^{i_{r+1}}\cdots 0^{i_K}1^{j_K}.$$

Hence, by expanding $\mathbb{E}[\Omega_X(w)\log\Omega_X(w)]$,

$$\sum_{|w|=n-1} \mathbb{E}[\Omega_X(w)\log\Omega_X(w)] = \sum_{\ell \geq 2} \ell \log \ell \sum_{|w|=n-1} P(w) \sum_{r \geq 1} (p\mathbf{I}_{[i_r(w)=\ell-1]} + q\mathbf{I}_{[j_r(w)=\ell-1]}),$$

where $i_r(w)$ denotes the length of the r-th 0-run in w and $j_r(w)$ the length of the r-th 1-run in w. Now let Z be a new random variable defined on words w of length $n-1$ as $Z = Z(w) = \sum_{r\geq 1}(p\mathbf{I}_{[i_r(w)=\ell-1]} + q\mathbf{I}_{[j_r(w)=\ell-1]})$. Then we just have to compute the expected value

$$\mathbb{E}[Z] = \sum_{r \geq 1} (p\mathbb{P}[i_r = \ell-1] + q\mathbb{P}[j_r = \ell - 1]).$$

Recall that the expected value $\mathbb{E}\left[\mathbf{I}_{[i_r=\ell-1]}\right] = \mathbb{P}[i_r = \ell - 1]$ has to be computed according the probability distribution of word W (of length $n-1$).

Next, note that the probability distribution of the length-k 0-run is given by $p^k q/(1-q) = p^{k-1}q$ and that the number of runs in a string of length n is approximately pqn. Consequently

$$\mathbb{E}[Z] \sim npq\left(pp^{\ell-2}q + qq^{\ell-2}p\right)$$

and finally

$$\sum_{|w|=n-1} \mathbb{E}[\Omega_X(w)\log\Omega_X(w)] \sim n\sum_{\ell\geq 2} \ell\log\ell\left(p^\ell q^2 + q^\ell p^2\right) = n\left(q^2 f(p) + p^2 f(q)\right).$$

Now we choose $n = \lfloor d^{-\varepsilon} \rfloor$ which ensures that $(1-d)^{n-1} = 1 + O(d^{1-\varepsilon})$. From the definition of $\lambda(d,p)$ and (3), this implies that

$$\lambda(d,p) \geq d(q^2 f(p) + p^2 f(q)) + O(d^{2-\varepsilon}).$$

Since $H(1-d) = -d\log d - (1-d)\log(1-d) = -d\log d + d\log(e) + O(d^2)$ we obtain the lower bound (12).

For the upper bound we proceed as in the proof of Theorem 3. We start with S_1. Let $S_{1,n-1}$ denote the subsum of S_1 corresponding to words of length $n-1$. Then it follows from the above calculations that $S_{1,n-1} = O(nd)$ (actually we can be much more precise). Furthermore, it follows as in the proof of Theorem 3 that $S_1 - S_{1,n-1} = O\left(\log n \, d^2 n^2\right)$ if $dn \to 0$. Finally, for S_2 we have (see Lemma 6)

$$S_2 = -n(1-d)H(p) + d(1-d)^{n-1} n\log n + O\left(\log n\, d^2 n^2\right).$$

Consequently, we obtain for $n = n(d) = \lfloor d^{-1}/\log d^{-1} \rfloor$

$$\begin{aligned} I(d,p) &\leq \frac{S_1 - S_2}{n} \\ &= (1-d)H(p) - (1-d)^{n-1} d\log n + O(d) + O\left(\log n\, d^2 n\right) \\ &= H(p) + d\log d + O(d\log\log(1/d)). \end{aligned}$$

This completes the proof of the theorem. ∎

Acknowledgment

M. Drmota was supported in part by the Austrian Science Foundation FWF Grant No. S9604. W. Szpankowski was supported in part by NSF Science and Technology Center on Science of Information Grant CCF-0939370, NSF Grants DMS-0800568, CCF-0830140, AFOSR Grant FA8655-11-1-3076, NSA Grant H98230-11-1-0141.

References

[1] J. Bourdon, and B. Vallée, "Generalized pattern matching statistics," *Mathematics and Computer Science II*, Trends. Math., 249–26, 2002.

[2] M. Dalai, "A new bound for the capacity of the deletion channel with high deletion probabilities," arXiv:1004.0400.

[3] S. Diggavi and M. Grossglauser, "Information transmission over finite buffer channels," *IEEE Trans. Info. Th.*, 52, 1226-1237, 2006.

[4] R.L. Dobrushin, "Shannon's theorem for channles with synchronization errors." *Prob. Info. Trans.*, 18-36, 1967.

[5] A. Iyengar, P. Siegel, J. Wolf, "Modeling and information rates for synchronization error channels," arXiv:1106.0070.

[6] P. Flajolet, "Singularity analysis and asymptotics of Bernoulli sums," *Theoretical Computer Science*, 215, 371–381, 1999.

[7] P. Flajolet, W. Szpankowski, and B. Vallée, "Hidden word statistics," *J. ACM*, **53**(1), 147–183, 2006.

[8] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *IEEE Trans. Info. Th.*, 45, 1072-1081, 1999.

[9] A. Kalai, M. Mitzenmacher, and M. Sudan, "Tight asymptotic bounds for the deletion channel with small deletion probabilities," *ISIT*, Austin, 2010.

[10] Y. Kanoria and A. Montanari, "On the deletion channel with small deletion probability," *ISIT*, Austin, 2010; see arXiv:1104.5546 for an extension.

[11] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probab. Surveys*, 1-33, 2009.

[12] M. Mitzenmacher and E. Drinea, "A simple lower bound for the capacity of the deletion channel," *IEEE Trans. Info. Th.*, 4657-4660, 2006.

[13] W. Szpankowski, "Average case analysis of algorithms on sequences," Wiley, New York, 2001.

[14] R. Venkataramanan, S. Tatikonda, and K. Ramchandran, "Achievable rates for channels with deletions and insertions," *ISIT*, St. Petersburg, Russia, 2011.