Efficient File Synchronization: a Distributed Source Coding Approach[1]

Nan Ma, Kannan Ramchandran and David Tse

Wireless Foundations, Dept. of Electrical Engineering and Computer Sciences

University of California at Berkeley

Abstract—**The problem of reconstructing a source sequence with the presence of decoder side-information that is mis-synchronized to the source due to deletions is studied in a distributed source coding framework. Motivated by practical applications, the deletion process is assumed to be bursty and is modeled by a Markov chain. The minimum rate needed to reconstruct the source sequence with high probability is characterized in terms of an information theoretic expression, which is interpreted as the amount of information of the deleted content and the locations of deletions, subtracting "nature's secret", that is, the uncertainty of the locations given the source and side-information. For small bursty deletion probability, the asymptotic expansion of the minimum rate is computed.**

I. INTRODUCTION

In distributed file backup or file sharing systems, different source nodes may have different versions of the same file differing by a small number of edits including deletions and insertions. The edits usually appear in bursts, for example, a paragraph of text is deleted, or several consecutive frames of video are inserted. An important question is: how to efficiently send a file to a remote node that has a different version of it? Further, what is the fundamental limit of the number of bits that needs to be sent to achieve this goal?
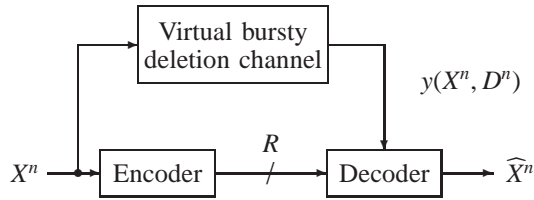

Fig. 1. *Synchronizing source sequences based on deletion side-information*

In this paper, we study the problem of reconstructing a source sequence with the help of decoder side-information using a distributed source coding framework (see Figure 1 for an illustration of the system). In this paper we focus on a simple case where the side-information is a deleted version of the source sequence. Consider a binary sequence of length n denoted by $X^n = (X_1, \ldots, X_n)$. Consider another binary sequence of length n called deletion pattern, denoted by

$D^n = (D_1, \ldots, D_n)$, which determines how X^n is to be deleted. The outcome of the deletion process, denoted by $y(X^n, D^n)$, is derived from X^n by deleting the bits at those locations where the deletion pattern is 1. Here is an example:

$$
\begin{aligned}
X^n &= (0, 1, 0, 1, 1, 0, 1, 0, 1, 0) \\
D^n &= (0, 1, 1, 0, 0, 0, 1, 1, 1, 0) \\
y(X^n, D^n) &= (0, 1, 1, 0, 0).
\end{aligned}
$$

Note that the deletion pattern D^n tends to have bursts of consecutive 1's, which lead to bursty deletions. The original files X^n and the deleted files $y(X^n, D^n)$ are available to the encoder and the decoder, respectively. The encoder sends a message to the decoder, so that the latter can reconstruct (synchronize) the original files X^n with an error probability that is vanishing when n goes to infinity. The objective of this work is to characterize the minimum rate of the message defined as the minimum number of bits per source bit.

The problem of synchronizing edited sequences has been studied by [1], [2] under the assumptions (1) the decoder is not allowed to make any error, and (2) the number of edits is a constant that does not increase with the length of the sequence. Upper and lower bounds of the minimum number of communication bits were provided as functions of the number of edits and the length of the sequence. In [3], an interactive, low-complexity and asymptotically optimal scheme was proposed. In comparison, in this paper, we consider on information theoretic formulation allowing a positive probability of error that vanishes as n increases. This assumption allows us to use additional techniques like random binning to improve the minimum rate. Unlike in assumption (2), we consider the case that a vanishing fraction of source bits, rather than a constant number of bits, is deleted, to get which makes the problem harder and more realistic.

In this paper, we characterize the minimum rate in terms of the limit of the conditional entropy of the source sequence given the side-information. We interpret the minimum rate as the amount of information in the deleted content and the locations of the deletions, subtracting the uncertainty of the locations given the source and side-information. We refer to the latter as "nature's secret". This is the information that the decoder will never find out even if it knows the source sequence and the side-information exactly; it represents the over-counting of information in the locations of the deletions. For example, if $X^n = (0, 0)$ and $y(X^n, D^n) = (0)$, the decoder

will never know and never needs to know whether the first bit or the second bit is deleted. Therefore the information about the precise location of the deleted bit is over-counted and should be subtracted. For small deletion rate and geometrically distributed burst length, the minimum rate is computed up to the precision of two leading terms.

If the deletion pattern D^n is independent and identically distributed (iid), X^n and $y(X^n, D^n)$ are the input and output of a binary iid deletion channel (see [4] and references therein). In this case, the problem of characterizing the minimum rate to reconstruct iid uniform source sequences in the distributed source coding problem is closely related to the evaluation of the mutual information across the deletion channel with iid uniform input distribution. For small deletion probability, the second and third order terms[2] of the channel capacity are achieved by iid uniform input distribution and are computed in [5, Lemma III.1]. In this paper we consider the asymptotic expansion of the minimum rate for the general bursty deletion process where the deletions are correlated over time. In the special case of iid deletion process, the expansion in Theorem 1 reduces to [5, Lemma III.1]. Note that in the source coding problem, the constant term becomes zero, which means that the second and third order terms of the channel capacity correspond to the first and second order terms of the minimum rate. Therefore, although it is mathematically equivalent to evaluate the these terms for the source coding and channel coding problems, from the practical point of view, the evaluation is more important for the source coding problem than for the channel coding problem. See Remark 3 for detailed discussions.

When we generalize the iid deletion process to bursty deletion process, new techniques are introduced. The most interesting technique is the generalization of the usual concept of a "run". We view the sequence $(1, 0, 1, 0, 1, 0)$ as a run with respect to deletion bursts of length two, because deleting two consecutive bits from that sequence always results in the same outcome sequence $(1, 0, 1, 0)$.

The rest of this paper is organized as follows. In Section II we formally setup the problem and provide a preview of the main result. In Section III we provide information theoretic expressions of the minimum rate for general parameters of the deletion pattern. In Section IV we focus on the asymptotics when the deletion rate is small and compute the two leading terms of the minimum rate. All the proofs are provided in the appendices.

Notation: With the exception of the symbols R, E, C, and J, random quantities are denoted in upper case and their specific instantiations in lower case. For $i, j \in \mathbb{Z}$, V_i^j denotes the sequence $(V_i, \ldots, V_j)$ and V^i denotes V_1^i. The binary entropy function is denoted by $h_2(\cdot)$. All logarithms are base 2. The notation $\{0, 1\}^n$ denotes the n-fold Cartesian product of $\{0, 1\}$, and $\{0, 1\}^*$ denotes $\left(\bigcup_{k \in \mathbb{Z}^+} \{0, 1\}^k \right) \bigcup \{\emptyset\}$.

II. Problem Formulation and Main Result

A. Problem formulation

The source sequence $X^n = (X_1, \ldots, X_n) \in \{0, 1\}^n$ is iid Bernoulli(1/2). Let $\alpha, \beta \in (0, 1)$. The deletion pattern $(D_0, D_1, \ldots, D_{n+1})$ is a two-state stationary Markov chain illustrated in Figure 2 with the initial distribution $p_{D_0} \sim$ Bernoulli(d), where $d := \beta/(\alpha + \beta)$ and transition probabilities $\mathbb{P}(D_i = 0 | D_{i-1} = 1) = 1 - \mathbb{P}(D_i = 1 | D_{i-1} = 1) = \alpha$ and $\mathbb{P}(D_i = 1 | D_{i-1} = 0) = 1 - \mathbb{P}(D_i = 0 | D_{i-1} = 0) = \beta$, for all $i = 1, 2, \ldots, n + 1$. Note that the initial distribution p_{D_0} is the stationary distribution of the Markov chain. The deleted sequence $y(X^n, D^n) \in \{0, 1\}^*$ is a subsequence of X^n, which is derived from X^n by deleting all those X_i's with $D_i = 1$ [3]. The length of $y(X^n, D^n)$, denoted by L_y, is a random variable taking values in $\{0, 1, \ldots, n\}$. For $i < L_y$, Y_i denotes the i-th bit in the $y(X^n, D^n)$ sequence. A run of consecutive 1's in the deletion pattern is called a burst of deletion. Since β is the probability to initiate a burst of deletion, it is called the deletion rate.
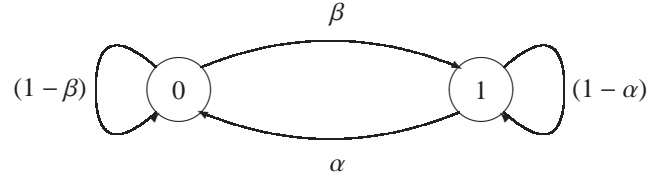
Fig. 2. *Markov model for the deletion pattern process $\{D_i\}_{i \geq 0}$. $D_i = 1$ means X_i is deleted; $D_i = 0$ means X_i is not deleted.*

The source sequence X^n is available to the encoder and the deleted sequence $y(X^n, D^n)$ is available only to the decoder as side-information. The deletion patterns D^n is available to neither the encoder nor the decoder. The encoder encodes X^n and sends a message to the decoder so that the decoder can reproduce the source with high probability.

Remark 1: If $\beta = 1 - \alpha = d$, D^n becomes iid, and the relation between X^n and $y(X^n, D^n)$ can be modeled as an iid deletion channel with deletion probability d. In this paper we consider the Markov deletion pattern to emphasize the bursty nature of the deletion process in the source coding problem.

The formal definitions of a code and an achievable rate are as follows.

Definition 1: A distributed source code for deletion side-information with parameters $(n, |\mathcal{M}_n|)$ is the tuple (f_n, g_n) consisting of an encoding function $f_n : \{0, 1\}^n \to \mathcal{M}_n$ and a decoding function $g_n : \mathcal{M}_n \times \{0, 1\}^* \to \{0, 1\}^n$.

Definition 2: A real number R is called an achievable rate if, there exists a sequence of distributed source codes $\{(f_n, g_n)\}_{n \geq 1}$ for deletion side-information with parameters $(n, |\mathcal{M}_n|)$ satisfying $\lim_{n \to \infty} \mathbb{P}(X^n \neq g_n(f_n(X^n), y(X^n, D^n))) = 0$ and $\limsup_{n \to \infty} (1/n) \log |\mathcal{M}_n| \leq R$.

The set of all achievable rates is necessarily closed and hence the minimum exists. The minimum achievable rate is

---

[2]For small deletion probability d, the first order term of the channel capacity is 1, the second order term is $\Theta(d \log d)$, and the third order term is $\Theta(d)$.

[3]D_0 and D_{n+1} do not determine the deletion of any source bit and do not play a role in the problem formulation. However, they are used in the information theoretic expressions in Sections III and IV.

denoted by R_{min}. The focus of this paper is to characterize R_{min}, especially for small β.

B. Main result

In Section III we express R_{min} using information theoretic quantities when the parameters α and β take arbitrary values. Unfortunately, we cannot provide an explicit expression of R_{min} as a function of α and β. Hence we focus on asymptotic regimes in Section IV when β is small.

Since the main difference between the erasure process and the deletion process is that the locations of the erasures are explicit but those of the deletions are not, it is interesting to focus on a regime where the amount of information to describe the locations of the deletions should play a significant role in the minimum rate. When α is vanishing and the length of bursts of deletions is increasing, for each burst, the number of bits to describe the deleted content increases linearly with respect to the length of the burst, but the number of bits to describe the location and length of the burst increases logarithmly. Therefore the regime with a vanishing α is not interesting. On the contrary, when α is fixed, the length of a burst is of order $\Theta(1)$ and we have an interesting regime. In this case, we evaluate $R_{min}(\alpha, \beta)$ as follows.

Theorem 1: When α is fixed, for any $\epsilon > 0$, we have

$$R_{min}(\alpha, \beta) = -\beta \log \beta + \beta \left(\frac{1 + h_2(\alpha)}{\alpha} + \log e - C \right) + O(\beta^{2-\epsilon}),$$
(2.1)

where $C = \sum_{l=1}^{\infty} 2^{-l-1} l \log l \approx 1.29$.

The proof of Theorem 1 based on Lemmas 1 and 2, and is provided in Appendix C. Detailed discussions about the proof techniques are given in Section IV-B.

Remark 2: The dominating term on the right side of (2.1) is $-\beta \log \beta$, and the second leading term is of order $\Theta(\beta)$. Since $-\log \beta$ tends to infinity slowly as β decreases to zero, in practice these two terms are often in the same order of magnitude. Therefore we need to evaluate both of them.

Remark 3: In [5], the authors evaluated the mutual information across the iid deletion channel with iid Bernoulli(1/2) input as

$$\lim_{n \to \infty} \frac{1}{n} I(X^n; y(X^n, D^n)) = 1 + d \log d - d(\log 2e - C) + O(d^{2-\epsilon}),$$

which implies that

$$\lim_{n \to \infty} \frac{1}{n} H(y(X^n, D^n)|X^n) = -d \log d + d(\log 2e - C) + O(d^{2-\epsilon}).$$

This expression should be compared with (2.1) in the special case that the deletion process is iid, which requires $\beta = 1 - \alpha = d$. Under this condition, (2.1) also has the same two leading terms $-d \log d + d(\log 2e - C)$. Therefore in the special case of iid deletion process, (2.1) is consistent with the result in [5].

Remark 4: Theorem 1 implies that when the input distribution is iid Bernoulli(1/2), the mutual information across the

bursty deletion channel is

$$\lim_{n \to \infty} \frac{1}{n} I(X^n; y(X^n, D^n))$$
$$= 1 + \beta \log \beta - \beta \left(\frac{1 + h_2(\alpha)}{\alpha} + \log e - C \right) + O(\beta^{2-\epsilon}) \quad (2.2)$$

In [6], Dobrushin showed that the channel capacity of the iid deletion channel is $\lim_{n \to \infty} (1/n) \max_{p_{X^n}} I(X^n; y(X^n, D^n))$. If this expression can be extended to the bursty deletion channel where the deletion pattern process is a Markov chain, then (2.2) provides an asymptotic lower bound for the capacity of the bursty deletion channel for small values of β.

III. INFORMATION THEORETIC EXPRESSION FOR GENERAL α AND β

We can write the minimum achievable rate R_{min} as the following information theoretic expression.

Lemma 1:

$$R_{min} = \lim_{n \to \infty} \frac{1}{n} H(X^n | y(X^n, D^n), D_0, D_{n+1}).$$

The proof of Lemma 1 is given in Appendix A. The structure of the proof is as follows: (1) we show that the limit $\lim_{n \to \infty} (1/n) H(X^n | y(X^n, D^n), D_0, D_{n+1})$ exists, (2) using the information-spectrum method [7, Section 7.2], we have $R_{min} = \overline{H}(X^n | y(X^n, D^n)) := \text{p-lim sup}_{n \to \infty} (1/n) \log(1/p_{X^n|y(X^n,D^n)}(X^n | y(X^n, D^n)))$, which is the conditional spectral sup-entropy, (3) we show that $\overline{H}(X^n | y(X^n, D^n)) = \lim_{n \to \infty} (1/n) H(X^n | y(X^n, D^n), D_0, D_{n+1})$. The techniques we use in step (3) are similar to those Dobrushin used in [6], where the capacity of the iid deletion channel is characterized by $\lim_{n \to \infty} (1/n) \max_{p_{X^n}} I(X^n; y(X^n, D^n))$.

In Lemma 2, the information theoretic expression of the minimum rate is written in another way, which has a more intuitive interpretation as explained in Remark 5.

Lemma 2:

$$R_{min} = d + H(D_1|D_0) - E_\infty, \quad (3.3)$$

where $E_\infty := \lim_{n \to \infty} E_n$, and $E_n := H(D_1|D_0, X^n, y(X^n, D^n), D_{n+1})$.

The proof of Lemma 2 is given in Appendix B.

Remark 5: Lemma 2 expresses R_{min} in terms of three parts, which can be intuitively interpreted as follows. The first term d is the fraction of deleted bits in X^n. It represents the amount of information per source bit in the deleted content, and thus the rate needed to send the deleted content. The second term is the entropy rate of the deletion pattern process, which is the rate needed to describe the locations of deletions. If the encoder knew the locations and sent them together with the deleted content, the decoder could reproduce X^n. However, this is excessive information. In fact, even if the decoder can correctly reproduce X^n, it can never know the exact deletion pattern. Therefore the uncertainty of the deletion pattern D^n, given X^n and $y(X^n, D^n)$, is not required to be revealed in order to reproduce X^n.

The uncertainty in the deletion pattern, given the source sequence and side-information is the *nature's secret*, which is known only to an imaginary third party (nature) who generates

the deletion pattern. Since nature's secret is not required to reproduce X^n, it should be subtracted from the message rate. Lemma 2 shows that nature's secret per source bit, which is the uncertainty in the whole deletion pattern D^n normalized by n, can be expressed as E_∞, which is the uncertainty in only D_1. An intuitive explanation is that, the uncertainty in each bit in D^n is approximately the same, therefore the uncertainty can be represented by the uncertainty in only D_1.

IV. ASYMPTOTIC BEHAVIOR OF R_{min} FOR SMALL VALUES OF β

In typical settings the number of edits is often much less than the file size. Since β is the probability to start a burst of deletions, the asymptotic behavior of R_{min} for small β is of special interest.

A. Case 1: Few number of long bursts of deletion: $\alpha \ll 1, \beta \ll 1$, and α/β is fixed

When $\alpha \ll 1, \beta \ll 1$ and α/β is fixed, the number of bursts are much smaller than the length of the sequence, and each burst is so long that the overall fraction of deletion $d = \beta/(\alpha + \beta)$ is a constant.

On the right side of (3.3), the first term d is a constant. For any $\epsilon > 0$, the second term $H(D_1|D_0) = dh_2(\alpha) + (1-d)h_2(\beta) = O(\beta^{1-\epsilon})$, and the third term $E_\infty \leq H(D_1|D_0) = O(\beta^{1-\epsilon})$. According to Lemma 2, we have

$$R_{min}(\alpha, \beta) = d + O(\beta^{1-\epsilon}).$$

Intuitively speaking, if we have a small number of long bursts of deletion, the amount of information of the locations of deletions is orderwise less than the amount of information of the content of deletion. Therefore R_{min} is dominated by the rate needed to deliver the deleted content.

A more interesting case is when all three terms of (3.3) are comparable.

B. Case 2: Few number of short bursts of deletion: α is fixed and $\beta \ll 1$

When α is fixed and $\beta \ll 1$, the number of bursts is much smaller than the length of the sequence. Since the length of a burst is drawn from a geometric distribution with parameter α, the expected length is of order $\Theta(1)$. The overall proportion of deleted bits is $d = \beta/(\alpha + \beta) = \beta/\alpha + \Theta(\beta^2)$. In this case, unlike in Case 1, the location information and "nature's secret" are comparable to the content information. Therefore we need to evaluate all three terms for this case. The three terms on the right side of (3.3) are evaluated as follows. For any $\epsilon > 0$, we have

$$d = \beta/\alpha + \Theta(\beta^2), \tag{4.4}$$

$$H(D_1|D_0) = -\beta \log \beta + \frac{\beta h_2(\alpha)}{\alpha} + \beta \log e + O(\beta^{2-\epsilon}), \tag{4.5}$$

$$-E_\infty = -C\beta + O(\beta^{2-\epsilon}), \tag{4.6}$$

where $C = \sum_{l=1}^\infty 2^{-l-1} l \log l \approx 1.29$. Combining (4.4) through (4.6) gives Theorem 1.

The proofs of (4.4) and (4.5) are trivial. The proof of (4.6) is highly nontrivial and is the essence of the proof of Theorem 1.

The complete proof of (4.6) is given in Appendix C. In this subsection we explain only the intuition of (4.6).

Let us first consider the case that the deletion is not bursty ($\alpha = 1$), i.e., no consecutive bits are deleted. In order to evaluate nature's secret E_∞ we need to estimate the uncertainty in D_1 given $X^n, y(X^n, D^n), D_0$ and D_{n+1}. The uncertainty is significant if the first run of X^n is different from the first run of $y(X^n, D^n)$. For example, if $X^n = (0, 0, 0, 1)$ and $y(X^n, D^n) = (0, 0, 1)$, we know that one bit is deleted in the first run (first three bits) of X^n, but do not know which bit is deleted. The true identity of the deleted bit is nature's secret. Since there are three equally likely possible deletion patterns and only one leads to $D_1 = 1$, the conditional entropy of D_1 is $h_2(1/3)$. The length of the first run of X^n is L, a geometrically distributed random variable with parameter $1/2$. If one bit is deleted in the first run, the conditional entropy is $h_2(1/L)$. The probability that any bit in L bits is deleted is roughly $L\beta$, therefore the average uncertainty is $\mathbb{E}[h_2(1/L)L\beta] = \left(\sum_{l=1}^\infty h_2(1/l)2^{-l}l \right)\beta = \left(\sum_{l=1}^\infty 2^{-l-1}l \log l \right)\beta = C\beta.$[4]

Let us now extend the discussion in the previous paragraph to the case of bursty deletions ($\alpha < 1$). First, we need to generalize the usual definition of "run" to b-run.

Definition 3: For any b and $l \in \mathbb{Z}^+$, a sequence $(x_1, \ldots, x_{b+l-1})$ is called a b-run of extent l if for all i, j satisfying $(i \equiv j \mod b)$, $x_i = x_j$ holds.

For example, $(1, 1, 1, 1, 1)$ is a 1-run of extent 5, and 1-run is the usual definition of a run. The sequence $(1, 0, 1, 0, 1)$ is a 2-run of extent 4. Note that there are l different ways to delete b consecutive bits in a sequence of length $l + b - 1$. A special property of a b-run of extent l is that, all the l ways of deletion result in the same outcome. For example, all four ways of deleting two consecutive bits in $(1, 0, 1, 0, 1)$ lead to the same outcome $(1, 0, 1)$. This observation is formally stated in the following fact.

Fact 1: Let x^{b+l-1} be a b-run of extent l. Let $\mathbf{d}_{i,b}$ denote the sequence of $(i - 1)$ 0's followed by b 1's, then followed by $(l - i)$ 0's. Then $y(x^{b+l-1}, \mathbf{d}_{i,b})$ is the same for all $i = 1, \ldots, l$.

Definition 4: For any $b \in \mathbb{Z}^+$, the first b-run of a sequence $(x_1, \ldots, x_n)$ is the longest segment starting from x_1 that is a b-run.

For example, the first 2-run of $(0, 1, 0, 1, 1)$ is $(0, 1, 0, 1)$.

Now let us consider the uncertainty in D_1 given $X^n, y(X^n, D^n), D_0$ and D_{n+1} through an example. If we know that a burst of 2 bits is deleted in $X^n = (0, 1, 0, 1, 1)$ to produce $y(X^n, D^n) = (0, 1, 1)$, we know that the deletion occurs within the first 2-run, i.e., $(0, 1, 0, 1)$. Since there are three indistinguishable deletion patterns, $(1, 1, 0, 0, 0)$, $(0, 1, 1, 0, 0)$, and $(0, 0, 1, 1, 0)$, among which only the first one satisfies $D_1 = 1$, the conditional entropy of D_1 is $h_2(1/3)$.

For any b, the extent of the first b-run, L, is a geometrically distributed random variable with parameter $1/2$, as in the non-

---

[4]In this section we only provide an intuitive explanation using a simplified case that there is only one burst of deletion. In a rigorous proof it is shown that with high probability the first burst of deletion can be isolated from the other bursts so that the general case is reduced to the simplified case. See Appendix C for details.

bursty case. This fact can be seen by sequentially generating $X_1, X_2, \ldots$. For arbitrary realization of $X^b = x^b$, X^b always belongs to the first b-run. If the first b-run has been extended to the $(i-1)$-th bit, it will be extended to the i-th bit if $X_i = x_{i-b}$, which occurs with probability $\frac{1}{2}$. Therefore the extent of the first b-run is a geometrically distributed variable. If one burst of b is deleted in the first b-run, the conditional entropy of D_1 is $h_2(1/L)$. Since given the length of burst b, the probability that any deletion pattern among all L possible deletion patterns occurs is roughly $L\beta$, the average uncertainty of $\mathbb{E}[h_2(1/L)L\beta] = C\beta$. Note that the result is the same for all b. In other words, nature's secret is always $C \approx 1.29$ bits *per burst*, regardless of the length of burst.

Remark 6: Since nature's secret is $C\beta + O(\beta^{2-\epsilon})$ for *any given value of the length of burst* $b \in \mathbb{Z}^+$, the fact that nature's secret averaged across different possible values of b is $C\beta + O(\beta^{2-\epsilon})$, regardless of the distribution of the length of a burst of deletions. This implies that Theorem 1 may generalize to more general deletion processes beyond the two-state Markov chains. In order to draw a rigorous statement, however, one has to revisit Lemmas 1 and 2 and prove them for the general setup.

V. Concluding Remarks

We studied the distributed source coding problem of synchronizing source sequences based on bursty deletion side-information. We evaluated the two leading terms of the minimum achievable rate for small deletion rate. Directions for future work include considering insertions in addition to deletions, and evaluating the leading terms of the capacity of the bursty deletion channel.

Appendix A
Proof of Lemma 1

(1) We first show that $R_n := (1/n)H(X^n|y(X^n, D^n), D_0, D_{n+1})$ converges as $n \to \infty$, so that the limit in the statement of Lemma 1 is well defined.

For all $m \in \{1, \ldots, n-1\}$, we have

$$
\begin{aligned}
nR_n &= H(X^n|y(X^n, D^n), D_0, D_{n+1}) \\
&\overset{(a)}{\geq} H(X^n|y(X^m, D^m), y(X^n_{m+1}, D^n_{m+1}), D_0, D_{n+1}) \\
&\geq H(X^m|y(X^m, D^m), y(X^n_{m+1}, D^n_{m+1}), D_0, D_{n+1}, D_{m+1}) \\
&\quad + H(X^n_{m+1}|y(X^m, D^m), y(X^n_{m+1}, D^n_{m+1}), D_0, D_{n+1}, D_m) \\
&\overset{(b)}{=} H(X^m|y(X^m, D^m), D_0, D_{m+1}) \\
&\quad + H(X^n_{m+1}|y(X^n_{m+1}, D^n_{m+1}), D_{n+1}, D_m) \\
&= H(X^m|y(X^m, D^m), D_0, D_{m+1}) \\
&\quad + H(X^{n-m}|y(X^{n-m}, D^{n-m}), D_0, D_{n-m})
\end{aligned}
$$

where step (a) holds because the tuple $(y(X^m, D^m), y(X^n_{m+1}, D^n_{m+1}))$ determines $y(X^n, D^n)$, and step (b) holds because the Markov chains $(y(X^n_{m+1}, D^n_{m+1}), D_{n+1}) - D_{m+1} - (X^m, y(X^m, D^m), D_0)$ and $(y(X^m, D^m), D_0) - D_m - (X^n_{m+1}, y(X^n_{m+1}, D^n_{m+1}), D_{n+1})$ hold. Therefore the sequence $\{nR_n\}_{n \in \mathbb{N}}$ is superadditive. By Fekete's lemma [8], the limit $\lim_{n \to \infty} R_n$ exists.

(2) Using the information-spectral version of the Slepian-Wolf theorem [7, Section 7.2], we have $R_{min} = \overline{H}(X^n|y(X^n, D^n)) := \text{p-}\lim\sup_{n \to \infty}(1/n)\log(1/p_{X^n|y(X^n,D^n)}(X^n|y(X^n, D^n)))$. In the rest of this appendix, for any random variables A, B, we abbreviate $p_A(A)$ and $p_{A|B}(A|B)$ to $p(A)$ and $p(A|B)$, respectively, to avoid cumbersome notations.

(3) Now we show that the sequence of random variables $(1/n)\log(1/p(X^n|y(X^n, D^n)))$ converges in probability to the limit $\lim_{n \to \infty} R_n$.

We introduce a segmented deletion process as follows. Let $k \geq 3$ be the length of a segment. Let $g := \lfloor n/k \rfloor$ be the number of complete segments and $l := n - gk$ be the length of the remainder. Consider the outcome of a segmented deletion process as follows: let $z(X^n, D^n) := (Z_{1L}, Z_{1M}, Z_{1R}, \ldots, Z_{gL}, Z_{gM}, Z_{gR}, Z_{remainder})$ be a vector with $(3g + 1)$ components, where $\forall i = 1, \ldots, g$, $Z_{iL} := y(X_{(i-1)k+1}, D_{(i-1)k+1})$, $Z_{iM} := y(X^{ik-1}_{(i-1)k+2}, D^{ik-1}_{(i-1)k+1})$, $Z_{iR} := y(X_{ik}, D_{ik})$, and $Z_{remainder} := y(X^n_{gk+1}, D^n_{gk+1})$. From $z(X^n, D^n)$ we can find out how many source bits are deleted in each segment and the remainder, and whether the first and last bits of each segment are deleted. The sequence $y(X^n, D^n)$ can be obtained by merging all the $(3g + 1)$ components of $z(X^n, D^n)$. Therefore the sequence $z(X^n, D^n)$ contains more information than $y(X^n, D^n)$. We will first fix k and let n go to infinity. Then we increase k to prove the final result.

The statement to be proved is based on the following three facts.

Fact 2: For any $k \geq 3$, n and any $\delta > 0$, there exists a function $\epsilon_1(k)$ satisfying $\lim_{k \to \infty} \epsilon_1(k) = 0$, so that

$$
\mathbb{P}\left(\frac{1}{n}\left|\log\frac{1}{p(X^n|y(X^n, D^n))} - \log\frac{1}{p(X^n|z(X^n, D^n))}\right| > \delta\right) \leq \frac{\epsilon_1(k)}{\delta}.
$$

Fact 3: For any k and any $\delta > 0$, there exists a function $\epsilon_2(k)$ satisfying $\lim_{k \to \infty} \epsilon_2(k) = 0$, so that as $n \to \infty$,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\log\frac{1}{p(X^n|z(X^n, D^n))}\right.\right. \\
&\left.\left. -\frac{1}{k}H(X^{k-1}_2|y(X^{k-1}_2, D^{k-1}_2), D_1, D_k)\right| > \delta\right) \\
&\leq \frac{\epsilon_2(k)}{\delta}.
\end{aligned}
$$

Fact 4:

$$
\begin{aligned}
&\lim_{k \to \infty}\frac{1}{k}H(X^{k-1}_2|y(X^{k-1}_2, D^{k-1}_2), D_1, D_k) \\
&= \lim_{k \to \infty}\frac{1}{k}H(X^k|y(X^k, D^k), D_0, D_{k+1}).
\end{aligned}
$$

Proof of Fact 2:

Since $y(X^n, D^n)$ can be determined by $z(X^n, D^n)$, there exists a function ϕ_n such that $y(X^n, D^n) = \phi_n(z(X^n, D^n))$. For any realization of $z(X^n, D^n) = z$, we have $\mathbb{P}(z(X^n, D^n) = z) \leq \mathbb{P}(y(X^n, D^n) = \phi_n(z))$, which implies that $(1/n)\log\mathbb{P}(z(X^n, D^n) = z) - (1/n)\log\mathbb{P}(y(X^n, D^n) = \phi_n(z)) \leq 0$ always holds. Let L_Z be the vector of $(3g+1)$ components representing the lengths of all the components of $z(X^n, D^n)$. Then we have

$$\mathbb{E} \left| \frac{1}{n} \log p(y(X^n, D^n)) - \frac{1}{n} \log p(z(X^n, D^n)) \right|$$

$$= \mathbb{E} \left[\frac{1}{n} \log p(y(X^n, D^n)) \right] - \mathbb{E} \left[\frac{1}{n} \log p(z(X^n, D^n)) \right]$$

$$= \frac{1}{n} (-H(y(X^n, D^n)) + H(z(X^n, D^n)))$$

$$= \frac{1}{n} H(z(X^n, D^n) | y(X^n, D^n))$$

$$= \frac{1}{n} H(L_Z | y(X^n, D^n))$$

$$\leq \frac{1}{n} H(L_Z)$$

$$\leq \frac{1}{n} (3g + 1) \log k$$

$$\leq \frac{4 \log k}{k}.$$

By Markov's inequality,

$$\mathbb{P} \left(\frac{1}{n} \left| \log p(y(X^n, D^n)) - \log p(z(X^n, D^n)) \right| > \delta \right) \leq \frac{4 \log k}{k\delta}.$$

Using the same argument we also have

$$\mathbb{P} \left(\frac{1}{n} \left| \log p(X^n, y(X^n, D^n)) - \log p(X^n, z(X^n, D^n)) \right| > \delta \right) \leq \frac{4 \log k}{k\delta}.$$

Combining the last two inequalities completes the proof of Fact 2. ∎

Proof of Fact 3:

Let $Z_B := (Z_{1L}, Z_{1R}, \ldots, Z_{gL}, Z_{gR}, Z_{remainder})$. Then

$$\frac{1}{n} \log p(z(X^n, D^n))$$

$$\overset{(c)}{=} \frac{1}{n} \log p(Z_B) + \sum_{i=1}^{g} \frac{1}{n} \log p(Z_{iM} | Z_B)$$

$$\overset{(d)}{=} \frac{1}{n} \log p(Z_B) + \sum_{i=1}^{g} \frac{1}{n} \log p(Z_{iM} | Z_{iL}, Z_{iR}), \quad \text{(A.1)}$$

where step (c) holds because given Z_B, $Z_{1M}, \ldots, Z_{gM}$ are conditionally independent, and step (d) holds because D^n is a Markov chain.

Since the expectation of the first term of (A.1) is equal to $(1/n) H(Z_B) \leq (2g + l)/n \log 3$, by Markov's inequality we have $\mathbb{P}((1/n) \log p(Z_B) > \delta) < (2g + l) \log 3 / (n\delta)$.

Due to the law of large number, as $n \to \infty$, which implies $g \to \infty$, the second term of (A.1) converges to $(1/k) H(y(X_2^{k-1}, D_2^{k-1}) | D_1, D_k)$ in probability.

Therefore we have: for any k and $n \to \infty$,

$$\mathbb{P} \left(\left| \frac{1}{n} \log \frac{1}{p(z(X^n, D^n))} - \frac{1}{k} H(y(X_2^{k-1}, D_2^{k-1}), D_1, D_k) \right| > \delta \right) \leq \frac{\epsilon_2'(k)}{\delta}$$

for some $\epsilon_2'(k)$ which vanishes as k increases.

Using the same argument we also have

$$\mathbb{P} \left(\left| \frac{1}{n} \log \frac{1}{p(X^n, z(X^n, D^n))} - \frac{1}{k} H(X_2^{k-1}, y(X_2^{k-1}, D_2^{k-1}), D_1, D_k) \right| > \delta \right)$$

$$\leq \frac{\epsilon_2''(k)}{\delta}.$$

Combining the last two inequalities completes the proof of Fact 3 ∎

Proof of Fact 4: Fact 4 holds because (i) $p_{X_2^{k-1}, D_2^{k-1}} = p_{X^{k-2}, D^{k-2}}$ and (ii) $(k - 2)/k \to 1$ as $k \to \infty$. ∎

Combining Facts 2 and 3, we have: for any fixed k and δ, as $n \to \infty$,

$$\mathbb{P} \left(\left| \frac{1}{n} \log \frac{1}{p(X^n | y(X^n, D^n))} - \frac{1}{k} H(X_2^{k-1} | y(X_2^{k-1}, D_2^{k-1}), D_1, D_k) \right| > \delta \right)$$

$$\leq \frac{\epsilon_3(k)}{\delta} \quad \text{(A.2)}$$

for some $\epsilon_3(k)$ which vanishes as k increases. By choosing a large enough k, the right hand side of (A.2) can be made arbitrarily small. Combining (A.2) and Fact 4, the sequence of random variables $(1/n) \log(1/p(X^n | y(X^n, D^n)))$ is shown to be converging in probability to the limit $\lim_{n \to \infty} R_n$.

Combining (1), (2) and (3) we have $R_{min} = \lim_{n \to \infty} (1/n) H(X^n | y(X^n, D^n), D_0, D_{n+1})$.

APPENDIX B
PROOF OF LEMMA 2

We will first introduce a sequence $\{J_n\}_{n \in \mathbb{N}}$ and show that $\lim_{n \to \infty} J_n = R_{min}$.

Lemma 3: For all $n \in \mathbb{Z}^+$, let $J_n := d + (1/n) H(y(X^n, D^n) | X^n, D_0, D_{n+1})$. Then we have $\lim_{n \to \infty} J_n = R_{min}$.

Proof: We have

$$R_{min} = \lim_{n \to \infty} \frac{1}{n} H(X^n | y(X^n, D^n), D_0, D_{n+1})$$

$$= \lim_{n \to \infty} \frac{1}{n} [H(X^n | D_0, D_{n+1}) + H(y(X^n, D^n) | X^n, D_0, D_{n+1})$$

$$- H(y(X^n, D^n) | D_0, D_{n+1})]$$

$$= 1 + \lim_{n \to \infty} \frac{1}{n} H(y(X^n, D^n) | X^n, D_0, D_{n+1})$$

$$- \lim_{n \to \infty} \frac{1}{n} (H(L_y | D_0, D_{n+1}) + H(y(X^n, D^n) | L_y, D_0, D_{n+1})).$$

Since

$$0 \leq \lim_{n \to \infty} \frac{1}{n} H(L_y | D_0, D_{n+1}) \leq \lim_{n \to \infty} \frac{1}{n} \log(n + 1) = 0,$$

we have $\lim_{n \to \infty} \frac{1}{n} H(L_y | D_0, D_{n+1}) = 0$. Since given $L_y = l$ and given (D_0, D_{n+1}) the sequence $y(X^n, D^n)$ is an iid

Bernoulli(1/2) sequence, $H(y(X^n, D^n)|L_y = l, D_0, D_{n+t}) = l$ holds. Therefore $H(y(X^n, D^n)|L_y, D_0, D_{n+t}) = \mathbb{E}(L_y)$ and hence

$$\lim_{n\to\infty} \frac{1}{n} H(y(X^n, D^n)|L_y, D_0, D_{n+1}) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}[L_y] = \frac{\alpha}{\alpha + \beta} = 1-d.$$

In conclusion,

$$\begin{aligned}
R_{min} &= 1 + \lim_{n\to\infty} \frac{1}{n} H(y(X^n, D^n)|X^n, D_0, D_{n+1}) - (1 - d) \\
&= \lim_{n\to\infty} \left(d + \frac{1}{n} H(y(X^n, D^n)|X^n, D_0, D_{n+1}) \right) \\
&= \lim_{n\to\infty} J_n,
\end{aligned}$$

which completes the proof of Lemma 3. ∎

Now let us use Lemma 3 to prove Lemma 2.

Expanding $I(D_1; y(X^n, D^n)|X^n, D_0, D_{n+1})$ in two ways, we have

$$H(D_1|X^n, D_0, D_{n+1}) - H(D_1|X^n, y(X^n, D^n), D_0, D_{n+1})$$
$$= H(y(X^n, D^n)|X^n, D_0, D_{n+1}) - H(y(X^n, D^n)|X^n, D_0, D_1, D_{n+1}). \quad \text{(B.3)}$$

The first term on the left side of (B.3) is equal to $H(D_1|D_0, D_{n+1})$. The second term on the left side of (B.3) is denoted by E_n. The first term on the right side of (B.3) is equal to $n(J_n - d)$. The second term on the right side of (B.3) is:

$$\begin{aligned}
&H(y(X^n, D^n)|X^n, D_0, D_1, D_{n+1}) \\
&= H(y(X^n, D^n)|X^n, D_1, D_{n+1}) \\
&= H(y(X^n, D^n)|X^n, D_1 = 1, D_{n+1})p_{D_1}(1) \\
&\quad + H(y(X^n, D^n)|X^n, D_1 = 0, D_{n+1})p_{D_1}(0) \\
&= H(y(X_2^n, D_2^n)|X_1, X_2^n, D_1 = 1, D_{n+1})p_{D_1}(1) \\
&\quad + H(X_1, y(X_2^n, D_2^n)|X_1, X_2^n, D_1 = 0, D_{n+1})p_{D_1}(0) \\
&\overset{(e)}{=} H(y(X_2^n, D_2^n)|X_2^n, D_1 = 1, D_{n+1})p_{D_1}(1) \\
&\quad + H(y(X_2^n, D_2^n)|X_2^n, D_1 = 0, D_{n+1})p_{D_1}(0) \\
&= H(y(X_2^n, D_2^n)|X_2^n, D_1, D_{n+1}) \\
&= H(y(X^{n-1}, D^{n-1})|X^{n-1}, D_0, D_n) \\
&= (n - 1)(J_{n-1} - d),
\end{aligned}$$

where step (e) holds because X_1 is independent of $(D^{n+1}, X_2^n, y(X_2^n, D_2^n))$. Therefore (B.3) becomes

$$H(D_1|D_0, D_{n+1}) - E_n = n(J_n - J_{n-1}) + J_{n-1} - d. \quad \text{(B.4)}$$

Now let us take the limit as $n \to \infty$ on both sides of (B.4). Because of mixing of the Markov chain $\{D_i\}_{i\geq 0}$, the distribution $p_{D_{n+1}|D_0, D_1}(\cdot|d_0, d_1)$ converges to the stationary distribution regardless of the initial values (d_0, d_1) as n goes to infinity. Therefore $\lim_{n\to\infty} H(D_1|D_0, D_{n+1}) = H(D_1|D_0)$. For the second term on the left side of (B.4), Lemma 4 guarantees the convergence of $\{E_n\}_{n\geq 1}$.

Lemma 4: (1) The sequence $\{E_n\}_{n\geq 1}$ is nondecreasing. (2) $\lim_{n\to} E_n$ exists.

Proof: (1) For all $n \geq 2$, we have

$$\begin{aligned}
E_n &= H(D_1|X^n, y(X^n, D^n), D_0, D_{n+1}) \\
&\geq H(D_1|X^n, y(X^n, D^n), D_0, D_n, D_{n+1}) \\
&= H(D_1|X^n, y(X^n, D^n), D_0, D_n) \\
&= H(D_1|X^n, y(X^n, D^n), D_0, D_n = 1)p_{D_n}(1) \\
&\quad + H(D_1|X^n, y(X^n, D^n), D_0, D_n = 0)p_{D_n}(0) \\
&= H(D_1|X^{n-1}, X_n, y(X^{n-1}, D^{n-1}), D_0, D_n = 1)p_{D_n}(1) \\
&\quad + H(D_1|X^{n-1}, X_n, y(X^{n-1}, D^{n-1}), D_0, D_n = 0)p_{D_n}(0) \\
&= H(D_1|X^{n-1}, y(X^{n-1}, D^{n-1}), D_0, D_n) \\
&= E_{n-1}.
\end{aligned}$$

Therefore $\{E_n\}_{n\geq 1}$ is nondecreasing.

(2) Since for all n, $E_n \geq 1$ holds and $\{E_n\}_{n\geq 1}$ is nondecreasing, $E_\infty = \lim_{n\to} E_n$ exists. ∎

By Lemma 4, the left side of (B.4) converges to $H(D_1|D_0) - E_\infty$ as $n \to \infty$. Since (B.4) holds, the right side also converges and the limit is $(\lim_{n\to\infty} n(J_n - J_{n-1})) + R_{min} - d$. Since $\{J_n\}_{n\geq 1}$ is a converging sequence and the $\lim_{n\to\infty} n(J_n - J_{n-1})$ exists, $\lim_{n\to\infty} n(J_n - J_{n-1}) = 0$. Therefore in the limit as $n \to \infty$, (B.4) becomes

$$H(D_1|D_0) - E_\infty = R_{min} - d,$$

which completes the proof of Lemma 2.

APPENDIX C
PROOF OF THEOREM 1

When α is a fixed constant and $\beta \ll 1$, it is easy to verify that the first two terms of (3.3) are

$$\begin{aligned}
d + H(D_1|D_0) &= \frac{\beta}{\alpha + \beta} + \frac{\alpha h_2(\beta)}{\alpha + \beta} + \frac{\beta h_2(\alpha)}{\alpha + \beta} \\
&= -\beta \log\beta + \beta \left(\frac{1 + h_2(\alpha)}{\alpha} + \log e \right) + O(\beta^{2-\epsilon}),
\end{aligned}$$

for any $\epsilon > 0$. We will show that the third term of (3.3) $E_\infty = C\beta + O(\beta^{2-\epsilon})$.

Let us first define "typicality" of the deletion pattern. Since E_∞ is the conditional entropy of D_1, which is more relevant to the first a few bits of D_0^n, the typicality of the D_0^n concerns about only the first a few bits.

Definition 5: Let $k = \max\{6, 6/(\log(1 - \alpha))\}$. For $n > -k\log\beta$, the deletion pattern D_0^n is typical if the following two conditions hold.

1) There is at most one run of 1's in $(D_0, \ldots, D_{-k\log\beta})$.
2) There are no more than $(-k/3 \log\beta)$ 1's in $(D_0, \ldots, D_{-k\log\beta})$.

Lemma 5 states that the deletion pattern is typical with high probability.

Lemma 5: For any $\epsilon > 0$, the probability that D_0^n is typical is at least $1 - O(\beta^{2-\epsilon})$.

Proof: Since any deletion pattern that has r runs of 1's in $(D_0, \ldots, D_{-k\log\beta})$ occurs with probability $O(\beta^r)$ and there are no more than $(-k\log\beta)^{2r}$ such patterns, $\mathbb{P}((D_0, \ldots, D_{-k\log\beta})$ contains r runs of 1's$) = O(\beta^{r-\epsilon})$ for any $\epsilon > 0$. Hence

condition 1) of Definition 5 holds with probability $1 - O(\beta^{2-\epsilon})$. Given that condition 1) holds, condition 2) is violated if there is a burst of deletion longer than $(-k/3 \log \beta)$, which occurs with the probability $O((1-\alpha)^{-k/3 \log \beta}) = O(\beta^2)$. In conclusion, $\mathbb{P}(D_0^n$ is typical $) = 1 - O(\beta^{2-\epsilon})$ for any $\epsilon > 0$. ∎

Let the indicator random variable $T := 1$ if D_0^n is typical and $T := 0$ otherwise. Lemma 5 implies that $p_T(0) = O(\beta^{2-\epsilon})$, $\forall \epsilon > 0$. Lemma 6 states that we can focus on the typical case $T = 1$ in order to evaluate E_∞ to the precision of $O(\beta^{2-\epsilon})$.

Lemma 6:

$$E_\infty = \lim_{n \to \infty} H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T = 1) p_T(1) + O(\beta^{2-\epsilon}).$$

Proof: For all $n > -k \log \beta$, we have the following lower bound of E_n

$$
\begin{aligned}
E_n &\geq H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T) \\
&\geq H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T = 1) p_T(1),
\end{aligned}
$$

and the following upper bound

$$
\begin{aligned}
E_n &\leq H(D_1, T | X^n, y(X^n, D^n), D_0, D_{n+1}) \\
&= H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T) \\
&\quad + H(T | X^n, y(X^n, D^n), D_0, D_{n+1}) \\
&\leq H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T = 1) p_T(1) \\
&\quad + H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T = 0) p_T(0) + H(T) \\
&\leq H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T = 1) p_T(1) \\
&\quad + p_T(0) + H(T) \\
&= H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T = 1) p_T(1) + O(\beta^{2-\epsilon}).
\end{aligned}
$$

Taking the limit as $n \to \infty$ completes the proof. ∎

For all $n > -k \log \beta$, we have

$$
\begin{aligned}
&H(D_1 | X^n, y(X^n, D^n), D_0, D_{n+1}, T = 1) p_T(1) \\
&= H(D_1 | X^n, y(X^n, D^n), D_0 = 1, D_{n+1}, T = 1) p_{D_0,T}(1, 1) \\
&\quad + H(D_1 | X^n, y(X^n, D^n), D_0 = 0, D_{n+1}, T = 1) p_{D_0,T}(0, 1).
\end{aligned}
$$

We will separately analyze the following two cases: (1) $D_0 = 1, T = 1$ and (2) $D_0 = 0, T = 1$.

- Case (1): $D_0 = 1, T = 1$. In this case we check whether $X^{-k \log \beta} = Y^{-k \log \beta}$. Let $M_1 := 1$ if they match and $M_1 := 0$ otherwise. Note that M_1 is determined by X^n and $y(X^n, D^n)$.

  - Case (1.1): $D_0 = 1, T = 1, M_1 = 0$. There exists at least one 1 in $D_1^{-k \log \beta}$. Since $D_0 = 1$ and there is at most one run of 1 in $D_0^{-k \log \beta}$ in a typical deletion pattern, $D_1 = 1$ must hold. Therefore $H(D_1 | D_0 = 1, T = 1, M_1 = 0) = 0$.

  - Case (1.2): $D_0 = 1, T = 1, M_1 = 1$. In this case, both $D_1 = 0$ and $D_1 = 1$ are possible. Given $D_0 = 1, T = 1$, if $D_1 = 0$, then for all $i = 2, \ldots, -k \log \beta$, $D_i = 0$, which implies that $X^{-k \log \beta} = Y^{-k \log \beta}$. If $D_1 = 1$, then for all $i = 1, \ldots, -k \log \beta$, X_i and Y_i are independently generated fair bits, hence the event $X_i = Y_i$ occurs with probability $1/2$. Since events $\{X_i = Y_i\}_i$ are independent across i, $\mathbb{P}(M_1 = 1 | D_1 =$

$1, D_0 = 1, T = 1) = (1/2)^{k \log \beta} = O(\beta^6)$. Since $\mathbb{P}(D_1 = 1 | D_0 = 1, T = 1) = \Theta(1)$ and $\mathbb{P}(D_1 = 0 | D_0 = 1, T = 1) = \Theta(1)$, by Bayes' rule, we have $\mathbb{P}(D_1 = 1 | D_0 = 1, T = 1, M_1 = 1) = O(\beta^6)$. Therefore $H(D_1 | D_0 = 1, T = 1, M_1 = 1) = O(\beta^{6-\epsilon})$, $\forall \epsilon > 0$.

In conclusion, the contribution of Case (1) to E_∞ is

$$
\begin{aligned}
&H(D_1 | X^n, y(X^n, D^n), (D_0, T) = (1, 1), D_{n+1}) p_{D_0,T}(1, 1) \\
&= H(D_1 | X^n, y(X^n, D^n), (D_0, T) = (1, 1), D_{n+1}, M_1) \\
&\quad \times p_{D_0,T}(1, 1) \\
&= O(\beta^{6-\epsilon}).
\end{aligned}
$$

- Case (2): $D_0 = 0, T = 1$. In this case we will first check whether $X^{-k/3 \log \beta} = Y^{-k/3 \log \beta}$. Let $M_2 := 1$ if they match and $M_2 := 0$ otherwise.

  - Case (2.1): $D_0 = 0, T = 1, M_2 = 1$. By the same argument as in Case 1 for $M_1 = 1$, we have $\mathbb{P}(D_1 = 1 | D_0 = 0, T = 1, M_2 = 1) = O(\beta^2)$, and $H(D_1 | D_0 = 0, T = 1, M_2 = 1) p_{D_0,T,M_2}(0, 1, 1) = O(\beta^{2-\epsilon})$, $\forall \epsilon > 0$.

  - Case (2.2): $D_0 = 0, T = 1, M_2 = 0$. We try to find a length-$(-k/3 \log \beta)$ segment in $Y^{-k \log \beta}$ that matches $X_{-2k/3 \log \beta+1}^{-k \log \beta}$. Since (i) $M_2 = 0$ implies that at least one bit in the first $-k/3 \log \beta$ bits is deleted and (ii) a burst of deletion in a typical deletion pattern is no longer than $-k/3 \log \beta$, there must be no deletion in $D_{-2k/3 \log \beta+1}^{-k \log \beta}$, which implies that there must be at least one segment in $Y^{-k \log \beta}$ that matches $X_{-2k/3 \log \beta+1}^{-k \log \beta}$. Define $B := 0$ if there are two or more segments that match $X_{-2k/3 \log \beta}^{-k \log \beta}$; and for $b \in \mathbb{Z}^+$, define $B := b$ if there is a unique segment $Y_{-2k/3 \log \beta+1-b}^{-k \log \beta-b}$ that matches $X_{-2k/3 \log \beta+1}^{-k \log \beta}$ with an offset b.

    * Case (2.2.1): $D_0 = 0, T = 1, M_2 = 0, B = 0$. The condition $B = 0$ requires at least $(-k/3 \log \beta)$ independent bit-wise matches, each of which occurs with probability $(1/2)$. Hence $B = 0$ occurs with probability at most $(1/2)^{-k/3 \log \beta} = O(\beta^2)$. Therefore the contribution of Case (2.2.1) is $H(D_1 | D_0 = 0, T = 1, M_2 = 0, B = 0) p_{D_0,T,M_2,B}(0, 1, 0, 0) = O(\beta^2)$.

    * Case (2.2.2): $D_0 = 0, T = 1, M_2 = 0, B = b \in \mathbb{Z}^+$. There must be a burst of deletion of length b taking place in $D_1^{-2k/3 \log \beta}$ which causes the offset of b between $X_{-2k/3 \log \beta+1}^{-k \log \beta}$ and the matching segment in $y(X^n, D^n)$. Since the length of the burst is bounded by $(-k/3 \log \beta)$ in a typical deletion pattern, $b \leq (-k/3 \log \beta)$ must hold. Since we can find a correct correspondence between a segment of X^n to its outcome of deletion, the deletion process to the left of the segment is conditionally independent to the deletion process to the right. Therefore in order to evaluate the conditional entropy of D_1 we need to focus on the process to the left of the segment only. Hence the contribution of this case to E_∞ is: $\sum_b H(D_1 | X^n, y(X^n, D^n), D_{n+1}, T = 1, D_0 = 0, M_2 = 0, B = b) p_{T,D_0,M_2,B}(1, 0, 0, b) =$

$\sum_b H(D_1|X^{n'}, y(X^{n'}, D^{n'}), T = 1, D_0 = 0, M_2 = 0, B = b)p_{T,D_0,M_2,B}(1,0,0,b)$, where $n' := -2k/3\log\beta$. Lemma 7 will show that the contribution of Case (2.2.2) is $C\beta + O(\beta^{2-\epsilon})$. This is the only case that is responsible for the leading term $C\beta$ in E_∞.

As a summary, the contribution of all the cases (1.1), (1.2), (2.1), (2.2.1) to E_∞ is of order $O(\beta^{2-\epsilon})$. Lemma 7 will show that the contribution of Case (2.2.2) is $C\beta + O(\beta^{2-\epsilon})$, which will complete the proof of Theorem 1.

Lemma 7: For $n' := -2k/3\log\beta$, we have $\sum_{b=1}^{-k/3\log\beta} H(D_1|X^{n'}, y(X^{n'}, D^{n'}), (T, D_0, M_2, B) = (1,0,0,b)) \times p_{T,D_0,M_2,B}(1,0,0,b) = C\beta + O(\beta^{2-\epsilon})$.

Proof: Using the abbreviation $Y := y(X^{n'}, D^{n'})$, we have

$$\sum_{b=1}^{-k/3\log\beta} H(D_1|X^{n'}, Y, (T, D_0, M_2, B) = (1,0,0,b))$$
$$\times p_{T,D_0,M_2,B}(1,0,0,b)$$
$$= \sum_{b=1}^{-k/3\log\beta} \sum_{x^{n'},y} H(D_1|X^{n'} = x^{n'}, Y = y,$$
$$(T, D_0, M_2, B) = (1,0,0,b))$$
$$\times p_{X^{n'},Y,T,D_0,M_2,B}(x^{n'}, y, 1, 0, 0, b) \qquad \text{(C.5)}$$
$$\overset{(f)}{=} \sum_{b=1}^{-k/3\log\beta} \sum_{x^{n'}} H(D_1|X^{n'} = x^{n'}, Y = x^{n'}_{b+1},$$
$$(T, D_0, M_2, B) = (1,0,0,b))$$
$$\times p_{X^{n'},Y,T,D_0,M_2,B}(x^{n'}, x^{n'}_{b+1}, 1, 0, 0, b)$$
$$\overset{(g)}{=} \sum_{b=1}^{-k/3\log\beta} \sum_{x^{n'}} H(D_1|X^{n'} = x^{n'}, Y = x^{n'}_{b+1}, (T, D_0, B) = (1,0,b))$$
$$\times p_{X^{n'},Y,T,D_0,B}(x^{n'}, x^{n'}_{b+1}, 1, 0, b) + O(\beta^2), \qquad \text{(C.6)}$$

where step (f) holds because of the following reason. Given $(T, D_0, M_2, B) = (1,0,0,b)$, if $D_1 = 1$, then $D_1^b = \mathbf{1}$ and $D_{b+1}^{n'} = \mathbf{0}$ hold, which imply that $Y = X^{n'}_{b+1}$. Therefore the conditional entropy in (C.5) is nonzero only if $y = x^{n'}_{b+1}$. Step (g) holds because given $y = x^{n'}_{b+1}$, the probability that $M_2 = 0$ is of order $O(\beta^2)$.

Define $l_b(\cdot) : \{0,1\}^* \to \mathbb{Z}^+$ to be the length of the first b-run of x^n (c.f. Definitions 3 and 4). In other words, for $l = 1, 2, \ldots$, $l_b(x^n) := l$ if (i) $\forall b \le i < b + l$, $x_i = x_{i-b}$ and (ii) $x_{b+l} \ne x_l$. Let $\mathbf{d}_{i,b}$ denote the sequence $d_1^{n'} \in \{0,1\}^{n'}$ satisfying that if $j = i, \ldots, i+b-1$, then $d_j = 1$, otherwise $d_j = 0$. Due to Fact 1, if $l_b(x^{n'}) = l$, then $y(x^{n'}, \mathbf{d}_{i,b}) = x^{n'}_{b+1}$ holds for all $i = 1, \ldots, l$, but does not hold for any $i > l$. Since given $D_0 = 0$ and $D_{n+1} = 0$ all l deletion patterns $\{\mathbf{d}_{i,b}\}_{i=1}^l$ occurs with the same probability $\alpha(1-\alpha)^{b-1}\beta(1-\beta)^{n'-b}$, and only one of them, $\mathbf{d}_{1,b}$, satisfies $D_1 = 1$, we have $H(D_1|X^{n'} = x^{n'}, Y = x^{n'}_{b+1}, l(X^{n'}) = l, (T, D_0, B) = (1,0,b)) = h_2(1/l)$.

For a sequence $x^{n'}$ satisfying $l_b(x^{n'}) = l$, we have

$$p_{X^{n'},Y,T,D_0,B}(x^{n'}, x^{n'}_{b+1}, 1, 0, b)$$
$$= p_{X^{n'}}(x^{n'}) \sum_{i=1}^l p_{D_1^{n'}|D_0,D_{n'+1}}(\mathbf{d}_{i,b}|0,0)$$
$$= p_{X^{n'}}(x^{n'})l\alpha(1-\alpha)^{b-1}\beta(1-\beta)^{n'-b}$$
$$= p_{X^{n'}}(x^{n'})l\alpha(1-\alpha)^{b-1}\beta(1-O(\beta^{1-\epsilon})),$$

for any $\epsilon > 0$.

Therefore we continue (C.6) as

$$= \sum_{b=1}^{-k/3\log\beta} \sum_{l=1}^{n'-b} \sum_{x^{n'}:l_b(x^{n'})=l} h_2\left(\frac{1}{l}\right)p_{X^{n'}}(x^{n'})l\alpha(1-\alpha)^{b-1}\beta(1-O(\beta^{1-\epsilon}))$$
$$+ O(\beta^2)$$
$$= \sum_{b=1}^{-k/3\log\beta} \sum_{l=1}^{n'-b} h_2\left(\frac{1}{l}\right)2^{-l}l\alpha(1-\alpha)^{b-1}\beta(1-O(\beta^{1-\epsilon})) + O(\beta^2)$$
$$\overset{(h)}{=} \sum_{b=1}^{\infty} \sum_{l=1}^{\infty} h_2\left(\frac{1}{l}\right)2^{-l}l\alpha(1-\alpha)^{b-1}\beta(1-O(\beta^{1-\epsilon})) + O(\beta^2)$$
$$= \sum_{l=1}^{\infty} h_2\left(\frac{1}{l}\right)2^{-l}l\beta + O(\beta^{2-\epsilon})$$
$$\overset{(i)}{=} C\beta + O(\beta^{2-\epsilon}),$$

where step (h) holds because $k = \max\{6, 6/(\log(1-\alpha))\}$ and $n' = -2k/3\log\beta$, which guarantee that changing the limits of summations to infinity only leads to a change of order $O(\beta^2)$, and step (i) holds because $\sum_{l=1}^{\infty} h_2(1/l)2^{-l}l = \sum_{l=1}^{\infty} 2^{-l-1}l\log l$. ∎

REFERENCES

[1] V. L. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Doklady Akademii Nauk SSSR*, vol. 163, no. 4, pp. 845–848, 1965.

[2] A. Orlitsky and K. Viswanathan, "One-way communication and error-correcting codes," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1781–1788, 2003.

[3] R. Venkataramanan, H. Zhang, and K. Ramchandran, "Interactive Low-complexity Codes for Synchronization from Deletions and Insertions."

[4] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, vol. 6, pp. 1–33, 2009.

[5] Y. Kanoria and A. Montanari, "On the deletion channel with small deletion probability," in *Proc. IEEE Int. Symp. Information Theory*, Austin, Texas, Jul. 13–18, 2010, pp. 1002–1006.

[6] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Problems of Information Transmission*, vol. 3, no. 4, pp. 11–26, 1967, translated from Problemy Peredachi Informatsii, vol. 3, no. 4, pp. 18 –36, 1967.

[7] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer, 2003.

[8] A. Schrijver, *Combinatorial Optimization*. Springer, 2003.