Bounds on the Optimal Rate for Synchronization from Deletions and Insertions

Ramji Venkataramanan
Dept. of Electrical Engineering
Yale University, USA
Email: ramji.venkataramanan@yale.edu

Sekhar Tatikonda
Dept. of Electrical Engineering
Yale University, USA
Email: sekhar.tatikonda@yale.edu

Kannan Ramchandran
Dept. of EECS
University of California, Berkeley, USA
Email: kannanr@eecs.berkeley.edu

***Abstract*—Consider two remotely located binary sources X and Y, where Y is mis-synchronized from X due to deletions and insertions. The distribution of X is known, and Y is obtained from X through a process of i.i.d deletions and insertions. What is the minimum rate of information X needs to send in order to synchronize Y to X? This is a distributed source coding problem, so the optimal rate is the conditional entropy of X given Y. However, the optimal rate is difficult to compute due to the memory in the joint process (X, Y). The transformation from X to Y may be viewed in terms of runs as follows: some runs of X get shorter/longer, some runs of X get deleted, and some new runs are added. The optimal rate is difficult to compute mainly due to the last two phenomena: deleted runs, and new inserted runs. We start with this observation, and consider an augmented model where the decoder has additional information that indicates the positions of the deleted and inserted rounds. We compute the rate required to supply this information, and thereby obtain bounds on the optimal synchronization rate.**

I. Introduction

Consider Alice and Bob observing two distributed sources $\underline{X}$ and $\underline{Y}$, respectively. $\underline{Y}$ is an *edited* version of $\underline{X}$ sequence, where the edits consist of deletions and insertions. Under communication rate constraints, Bob would like to reconstruct Alice's sequence from $\underline{Y}$ using minimal communication between him and Alice. Here is an example:

$$\begin{cases} \underline{X} = abracadabradum\ldots \\ \underline{Y} = abacaddabadum\ldots. \end{cases} \tag{1}$$

In this case, $\underline{Y}$ is obtained from $\underline{X}$ by the deletion the of two 'r's, and an insertion of an an extra 'd' following the original seventh letter. Bob wants to reconstruct $\underline{\hat{X}}$ from $\underline{Y}$ using minimum communication between him and Alice when neither party knows what has been deleted or inserted or the locations of the edits. We will refer to this problem as *synchronization from deletions and insertions*.

There are many motivating scenarios where such a problem needs to be addressed. For instance, in file backup applications, the remotely located data sources often differ only by a small number of deletions and/or insertions. It is desirable to have a synchronization tool that achieves successful backup by transferring minimal information. The problem of synchronization also arises in other applications such as file sharing and online file editing. An interesting and important question to ask is: what is the minimal communication rate needed to achieve synchronization?

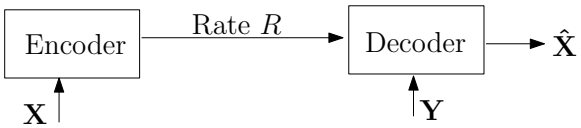

Fig. 1. Distributed source coding representation: $\mathbf{X}$ and $\mathbf{Y}$ are related through an edit model

Orlitsky [1] considered this question for a model in which the sequence $\underline{Y}$ is within a specified edit distance of $\underline{X}$, and Bob is allowed a limited amount of interactive communication with Alice. Depending on number of rounds of interaction allowed, bounds on the optimal rate of communication are derived for the zero-error reconstruction of $\underline{X}$.

In this work, we consider a stochastic edit model in which $\underline{X}$ and $\underline{Y}$ are related as follows: to generate $\underline{Y}$ from $\underline{X}$, each bit of $\underline{X}$ is independently deleted with probability d, or an extra bit inserted following it with probability i, or the bit is retained as is with probability $1-d-i$. We relax the requirement of zero-error reconstruction, and require instead that the probability of error go to zero as the length of $\underline{X}$ becomes large. Using this criterion, we derive bounds on the minimum rate required for synchronization.

The synchronization problem is an instance of source coding with side-information [2], depicted in Figure 1. The sequence $\underline{Y}$ available at the decoder (Bob) is related to $\underline{X}$ available at the encoder (Alice) through the stochastic edit model described above. Thus we expect the optimal synchronization rate to be the conditional entropy rate of $\underline{X}$ given $\underline{Y}$. While this is true and will be established in Section III, computing $H(\underline{X}|\underline{Y})$ is a challenging task due to the memory inherent in the joint distribution of $(\underline{X}, \underline{Y})$. Note that even if $\underline{X}$ is characterized by an i.i.d product distribution, the process of insertions and deletions induces memory in the joint distribution.

Our approach in this paper is to define a tractable memoryless problem that is close to the original synchronization problem. In a synchronization problem, it is useful to think of the source sequence in terms of *runs* of symbols rather than individual symbols. For example, let $\underline{X} = 00011\mathit{1}1$, and suppose that the bits shown in italics are deleted so that $\underline{Y} = 00111$. To reconstruct $\underline{X}$, we can insert the deleted 0 anywhere in the first run, and the one anywhere in the second run. This suggests that the synchronization problem could be characterized by a single-letter distribution describing

how a run in $\underline{X}$ gets transformed into a run (of a different length) in $\underline{Y}$. Then, if we had a one-to-one correspondence between runs in $\underline{X}$ and runs in $\underline{Y}$, computing the optimal synchronization rate would be straightforward. However, such a correspondence is not possible since deletions can lead to some runs being lost, and insertions to new runs being inserted.

The main idea of the paper is to augment the decoder with auxiliary sequences which indicate where runs were deleted and inserted. To obtain bounds on the optimal synchronization rate, we compute the synchronization rate given these auxiliary sequences as well as the additional rate needed to provide these sequences. We consider binary first-order Markov sources $\underline{X}$ and an i.i.d process of deletions and insertions to obtain $\underline{Y}$. Markov sources are used to model text, images, and video, where the synchronization problem often appears in applications such as distributed editing and file sharing. Further, first-order Markov processes have independently distributed run-lengths which leads to analytical bounds on the optimal rate; besides, they include i.i.d sources as a special case.

The bounds obtained here can be used to evaluate the performance of practical synchronization protocols, e.g., [3]–[6]. Though these protocols all use interaction, our bounds are still valid. This is because that the optimal rate for a problem of source coding with side information, even with interaction, is the conditional entropy rate of $\underline{X}$ given $\underline{Y}$. We note that this is true only when the requirement of the synchronization code is a vanishing probability of error, which is the case with the above protocols. For zero-error synchronization, interaction can strictly decrease the optimal synchronization rate [1].

Since source coding with side-information and channel coding are dual problems [7], [8], the synchronization problem is closely related to the problem of communicating over a channel with synchronization errors. In particular, the bounding techniques presented here can be used to yield lower bounds on the capacity of channels with synchronization errors. There is a large body of work on synchronization channels [9], [10], and on the deletion channel in particular [11]–[18]. We mention that the idea of augmenting the channel output with locations of deleted runs was used to obtain an upper bound to the deletion channel capacity in [14].

II. Preliminaries

We use uppercase letters to denote random variables, underlined letters for random vectors, and bold-face letters for random processes. The source sequence available at the encoder has length n and is denoted $\underline{X} = (X_1, X_2, \ldots, X_n)$. $\underline{Y}$ is the corresponding sequence at the decoder which needs to be synchronized to $\underline{X}$. Its length is denoted m. We note that m is random, and determined by the realization of the deletion/insertion process.

To keep the exposition simple, we assume that the source process $\mathbf{X}$ is binary valued and symmetric, i.e., $P(X_j = 0) = P(X_j = 1) = 0.5$, $\forall j$. The ideas presented here can be generalized in a fairly straightforward manner to sources which are asymmetric or have larger alphabet. Logarithms are with base 2, and entropy and mutual information are measured in bits. $h(.)$ denotes the binary entropy function. For any $0 < \alpha \leq 1$, $\bar{\alpha} \triangleq 1 - \alpha$.

$\mathbf{X}$ is a first-order Markov source with parameter γ, i.e.,

$$P(X_j = 0) = P(X_j = 1) = 0.5,$$
$$P(X_j = 1|X_{j-1} = 1) = P(X_j = 0|X_{j-1} = 0) = \gamma, \ \forall j. \tag{2}$$

We note that the $\mathbf{X}$ is i.i.d when $\gamma = 0.5$. We consider three different edit models relating $\mathbf{X}$ and $\mathbf{Y}$:

1) *Deletion Model*: $\mathbf{Y}$ is generated from $\mathbf{X}$ by independently deleting each bit with probability d, or retaining the bit with probability $1 - d$.
2) *Insertion Model*: $\mathbf{Y}$ is generated as follows. After each bit of $\mathbf{X}$, one bit may be inserted with probability i. When a bit is inserted after X_j, the inserted bit is equal to X_j with probability α, and equal to $\bar{X}_j$ with probability $1 - \alpha$. When $\alpha = 1$, this is the 'sticky' insertion model [19].
3) *Deletion + Insertion Model*: This general case combines the two above models. To generate $\mathbf{Y}$, each bit of $\mathbf{X}$ is independently deleted with probability d, *or* an extra bit inserted after it with probability i, *or* the bit is retained as is with probability $1 - d - i$. When a bit is inserted after X_j, the inserted bit is equal to X_j with probability α, and equal to $\bar{X}_j$ with probability $1 - \alpha$.

The edit model determines the joint distribution of $(\mathbf{X}, \mathbf{Y})$. For a given edit model, we now define a synchronization code.

Definition 1: An $(n, 2^{NR})$ synchronization code with block length n and rate R consists of

1) An encoder mapping
$$e : \{0,1\}^n \to \{1, \ldots, 2^{nR}\},$$
2) A decoder mapping
$$g : \{0,1\}^* \times \{1, \ldots, 2^{nR}\} \to \{0,1\}^n$$
where
$$\{0,1\}^* \triangleq \begin{cases} \cup_{k=0}^{n} \{0,1\}^k & \text{for the deletion model,} \\ \cup_{k=n}^{2n} \{0,1\}^k & \text{for the insertion model,} \\ \cup_{k=0}^{2n} \{0,1\}^k & \text{for the del. + ins. model.} \end{cases}$$

The probability of error of an $(n, 2^{nR})$ synchronization code is

$$P_{e,n} = \Pr(g(\underline{Y}, e(\underline{X})) \neq \underline{X})$$

A synchronization rate R is achievable if there exists a sequence of $(n, 2^{nR})$ codes such that $P_{e,n} \to 0$ as $n \to \infty$. The infimum of all achievable rates is the optimal synchronization rate R^*.

III. The Optimal Synchronization Rate

As explained in Section I, the synchronization problem is a distributed source coding problem. If the joint process $(\mathbf{X}, \mathbf{Y})$ is ergodic, then the optimal synchronization rate is the conditional entropy rate of $\mathbf{X}$ given $\mathbf{Y}$ [20]. The result is also true if is the joint process is information stable [21].

When $\mathbf{X}$ is a first-order Markov source and $\mathbf{Y}$ is generated according to either of the three edit models, the joint process $(\mathbf{X}, \mathbf{Y})$ is information stable [10]. We thus have the following characterization of the optimal synchronization rate.

Proposition 1: Let $\mathbf{X}$ be a first-order Markov source, and let $\mathbf{Y}$ be generated according to one of the edit models described in Section II. Then the optimal rate for synchronizing $\mathbf{Y}$ to $\mathbf{X}$ is

$$R^* = \lim_{n\to\infty} \frac{1}{n} H(\underline{X}|\underline{Y}).$$

Proof: The information stability of the joint process $(\mathbf{X}, \mathbf{Y})$ can be shown using methods analogous to the proof of Theorem 1 in [10]. Then the limit of $\frac{1}{n}H(\underline{X}|\underline{Y})$ exists, and is the optimal synchronization rate, as explained above.

The remainder of the paper is devoted to computing the optimal synchronization rate for each of the three edit models. We will see that in each of these problems, it is difficult to compute $\frac{1}{n}H(\underline{X}|\underline{Y})$ exactly. So we develop a bounding technique that yields good bounds on the optimal synchronization rate.

In the sequel, we will often think of the source as a sequence of alternating runs of zeros and ones. More precisely, a binary source sequence may be represented by a sequence of positive integers representing the lengths of its runs, and the value of the first bit (to indicate whether the first run has zeros or ones). For example, the sequence 0001100000 can be represented as $(3, 2, 5)$ if we know that the first bit is 0. The value of the first bit of $\underline{X}$ can be communicated to the decoder with vanishing rate, and we will assume this has been done at the outset. Hence, denoting the length of the jth run of $\underline{X}$ by L_{Xj} we have the following equivalence:

$$\underline{X} \leftrightarrow (L_{X1}, L_{X2}, \ldots). \tag{3}$$

For the first-order Markov binary source of (2), the run-lengths are independent and geometrically distributed, i.e.,

$$\Pr(L_{Xj} = r) = \gamma^{r-1}(1-\gamma), \qquad r = 1, 2, \ldots \tag{4}$$

Thus a first-order binary Markov source is equivalent to a memoryless source with alphabet $\mathbb{N}$, with symbols drawn independently according to (4).

IV. Deletion Model

The joint distribution of $(\underline{X}, \underline{Y})$ for the deletion model can be expressed as follows [15].

$$P(\underline{X}, \underline{Y}) = P(\underline{X}) \cdot P(\underline{Y}|\underline{X}) \text{ with}$$

$$P(\underline{X}) = \prod_{j=1}^{n} P(X_j|X^{j-1}),\ P(\underline{Y}|\underline{X}) = S(\underline{X}, \underline{Y})(1-d)^m d^{n-m} \tag{5}$$

where d is the deletion probability, m is the length of $\underline{Y}$, and $S(\underline{X}, \underline{Y})$ is the number of times $\underline{Y}$ appears as a subsequence in $\underline{X}$. For large n, the length m of $\underline{Y}$ will be close to $n(1-d)$ due to the law of large numbers. The term $S(\underline{X}, \underline{Y})$ represents the main challenge in obtaining an analytical expression for $H(\underline{X}|\underline{Y})$ for the deletion model.

To motivate our bounding technique, consider the following $(\underline{X}, \underline{Y})$ pair, generated according to the deletion model.

$$\begin{aligned} \underline{X} &= 000111000 \\ \underline{Y} &= 0010. \end{aligned} \tag{6}$$

For this pair $(\underline{X}, \underline{Y})$, we can associate each *run* of $\underline{Y}$ uniquely with a run in $\underline{X}$. Denote by (L_{X1}, L_{X2}, L_{X3}) and (L_{Y1}, L_{Y2}, L_{Y3}) the lengths of the three runs of $\underline{X}$ and $\underline{Y}$, respectively. Recall that

$$\underline{X} \leftrightarrow L_{X1}\ L_{X2}\ L_{X3}, \quad \underline{Y} \leftrightarrow L_{Y1}\ L_{Y2}\ L_{Y3}. \tag{7}$$

Therefore, for the $(\underline{X}, \underline{Y})$ pair in (6), we can write

$$\begin{aligned} &P(\underline{Y} = 0010|\underline{X} = 000111000) \\ &= P(L_{Y1} = 2\ L_{Y2} = 1\ L_{Y3} = 1|L_{X1} = 3\ L_{X2} = 3\ L_{X3} = 3) \\ &= P(L_{Y1} = 2|L_{X1} = 3)P(L_{Y2} = 1|L_{X2} = 3)P(L_{Y3} = 1|L_{X3} = 3). \end{aligned} \tag{8}$$

We make the following observation: If no runs in $\underline{X}$ are completely deleted, then the conditional distribution of $\underline{Y}$ given $\underline{X}$ may be written as a product distribution of run-length transformations:

$$P(\underline{Y}|\underline{X}) = P(L_{Y1}|L_{X1})P(L_{Y2}|L_{X2})P(L_{Y3}|L_{X3})\ldots \tag{9}$$

Since the run-lengths of the source sequence $\underline{X}$ are independently distributed according to (4), the joint distribution can also be written in product form:

$$P(\underline{X}, \underline{Y}) = P(L_{X1})P(L_{Y1}|L_{X1}) \cdot P(L_{X2})P(L_{Y2}|L_{X2})\ldots$$

where for all runs j,

$$\begin{aligned} &P(L_{Xj} = r, L_{Yj} = s) = P(L_{Xj} = r) \cdot P(L_{Yj} = s|L_{Xj} = r) \\ &= \gamma^{r-1}(1-\gamma) \cdot \binom{r}{s} d^{r-s}(1-d)^s,\ r = 1, 2, \ldots;\ 1 \le s \le r. \end{aligned} \tag{10}$$

Thus, if the deletion process acting on $\underline{X}$ to generate $\underline{Y}$ did not completely delete any runs of $\underline{X}$, the joint distribution of $(\underline{X}, \underline{Y})$ can be characterized in terms of a single-letter distribution of run-lengths given by (10).

Of course, in reality, we do have runs of $\underline{X}$ that are completely deleted. For example, consider $\underline{X} = 000111000$, and $\underline{Y} = 000$. $\underline{Y}$ has only one run, and we cannot associate it uniquely with a run of $\underline{X}$: $\underline{Y}$ could have been obtained from just the first run of $\underline{X}$, or from just the third run, or a combination of the first and third runs.

Now suppose that in addition to $\underline{Y}$, the decoder is also given an auxiliary sequence $\underline{S} = (S_1, \ldots, S_m)$, where $S_j \in 0, 1, \ldots$ is the number of runs *completely* deleted in $\underline{X}$ between the bits corresponding to Y_j and Y_{j+1}. For example, if $\underline{X} = 00\underbrace{\textit{011100}}0$ and the bits shown in italics were deleted to give $\underline{Y} = 000$, then $\underline{S} = (0, 1, 0)$. On the other hand, if the last six bits were all deleted, i.e., $\underline{X} = 000\underbrace{\textit{111000}}$, then $\underline{S} = (0, 0, 2)$.[1]

[1] $\underline{Y}$ has length m. S_m corresponds to the number of runs deleted after the bit in $\underline{X}$ corresponding to Y_m.

The auxiliary sequence $\underline{S}$ enables us to augment $\underline{Y}$ with the positions of missing runs. Consider $\underline{X} = 000111000$, as before. If the decoder were given $\underline{Y} = 000$ and $\underline{S} = (0, 1, 0)$, it can form the augmented sequence $\underline{Y}' = 00 - 0$, where a $-$ denotes a missing run, or equivalently a 'run of length 0' in $\underline{Y}$. Similarly, if the decoder were given $Y = 000$ and $\underline{S} = (0, 0, 2)$, the augmented sequence would be $\underline{Y}' = 000 - -$.

With the "$-$" markers indicating deleted runs, the number of runs in the augmented sequence $\underline{Y}'$ is equal to the number of runs in $\underline{X}$. We can therefore associate each run of the augmented sequence $\underline{Y}'$ uniquely with a run in $\underline{X}$. Denote by $L_{Y'1}, L_{Y'2}, \ldots$ the run-lengths of the augmented sequence $\underline{Y}'$, where $L_{Y'j} = 0$ if the run is a $-$. Then, we have

$$P(\underline{X}, \underline{Y}') = P(L_{X1})P(L_{Y'1}|L_{X1}) \cdot P(L_{X2})P(L_{Y'2}|L_{X2}) \ldots \tag{11}$$

where $\forall j$:

$$\begin{aligned} P(L_{Xj} = r) &= \gamma^{r-1}(1-\gamma), \quad r = 1, 2, \ldots \\ P(L_{Y'j} = s | L_{Xj} = r) &= \binom{r}{s} d^{r-s}(1-d)^s, \ 0 \le s \le r. \end{aligned} \tag{12}$$

Due to this product-form decomposition of $P(\underline{X}, \underline{Y}')$, and because the number of runs in $\underline{X}$ is close to $n(1-\gamma)$, the conditional entropy rate is

$$\frac{1}{n} H(\underline{X}|\underline{Y}') = (1 - \gamma + \epsilon_n) H(L_{X1}|L_{Y'1}), \tag{13}$$

where ϵ_n is a sequence that $\to 0$ as $n \to \infty$.

Lower Bound: The above argument shows that we can obtain a computable lower bound to the optimal synchronization rate by augmenting the decoder with the sequence $\underline{S}$. Since conditioning reduces entropy, we have

$$H(\underline{X}|\underline{Y}) \ge H(\underline{X}|\underline{Y}\ \underline{S}) = H(\underline{X}|\underline{Y}') \tag{14}$$

where the last equality holds because $(\underline{Y}, \underline{S})$ is equivalent to $\underline{Y}'$. Thus, using (13), we obtain the following lower bound to the optimal rate:

$$R^*(d) \ge \lim_{n\to\infty} \frac{1}{n} H(\underline{X}|\underline{Y}\ \underline{S}) = (1-\gamma) H(L_{X1}|L_{Y'1}).$$

Upper Bound: We can obtain an upper bound to the optimal rate by calculating the extra rate required to provide the decoder with the auxiliary sequence $\underline{S}$. We have

$$H(\underline{X}|\underline{Y}) \le H(\underline{X}, \underline{S}|\underline{Y}) = H(\underline{S}|\underline{Y}) + H(\underline{X}|\underline{S}\ \underline{Y}). \tag{15}$$

The term $\frac{1}{n} H(\underline{S}|\underline{Y})$ represents the additional rate needed to convey the auxiliary sequence to the decoder. To compute $H(\underline{S}|\underline{Y})$, we have

$$\begin{aligned} H(\underline{S}|\underline{Y}) &= \sum_{j=1}^{m} H(S_j | S^{j-1}\ \underline{Y}) \\ &\overset{(a)}{=} \sum_{j=1}^{m} H(S_j | Y_j\ Y_{j+1}) \\ &\overset{(b)}{=} n(1 - d \pm \epsilon_n) H(S_1 | Y_1\ Y_2). \end{aligned} \tag{16}$$

where $\epsilon_n \to 0$ as $n \to \infty$. Here, (a) holds because $\underline{X}$ is first order Markov, and the deletion process is i.i.d. If $Y_j = X_k$, and $Y_{j+1} = X_{k'}$ for some indices (k, k'), the values of the deleted bits between X_k and $X_{k'}$ are independent of all other variables given the values of X_k and X'_k. (b) holds because the length of $\underline{Y}$ concentrates around $n(1-d)$ due to the law of large numbers.

To calculate $H(S_1|Y_1Y_2)$, first observe that if $Y_1 = Y_2$, then S_1, the number of deleted runs between Y_1 and Y_2 belongs to the set $\{0, 1, 3, 5, \ldots\}$. If $Y_1 \ne Y_2$, then S_1 belongs to the set $\{0, 2, 4, 6, \ldots\}$. It can be shown [15] that for the deletion model, $\underline{Y}$ is also a first-order Markov sequence with parameter q, where

$$q = \frac{\gamma + d - 2\gamma d}{1 + d - 2\gamma d}. \tag{17}$$

Hence,

$$H(S_1|Y_1Y_2) = qH(S_1|Y_1 = Y_2) + (1-q)H(S_1|Y_1 \ne Y_2).$$

Each of the terms in the above equation can be calculated and substituted in (16) to obtain

$$\begin{aligned} \lim_{n\to\infty} \frac{H(\underline{S}|\underline{Y})}{n} &= (1-d) H(S_1|Y_1Y_2) \\ &= \frac{\gamma(1-d)^2}{1-\gamma d} \log_2 \frac{q(1-\gamma d)}{\gamma(1-d)} + \frac{\beta\theta(1-d)}{(1-\theta)^2} \log_2 \frac{1}{\theta} \\ &\quad + \frac{\beta\theta(1-d)}{1-\theta^2} \log_2 \frac{q}{\beta} + \frac{\beta(1-d)}{1-\theta^2} \log_2 \frac{1-q}{\beta}, \end{aligned} \tag{18}$$

where

$$q \triangleq \frac{\gamma + d - 2\gamma d}{1 + d - 2\gamma d}, \quad \theta \triangleq \frac{(1-\gamma)d}{1-\gamma d}, \quad \beta \triangleq \frac{(1-\gamma)(1-d)}{(1-\gamma d)^2}.$$

We omit the details of the calculation, and state the lower and upper bounds in the following proposition.

Proposition 2: Let $\mathbf{X}$ be a binary first-order Markov source with parameter γ, and let $\mathbf{Y}$ be generated according to the deletion model with deletion probability d. Then the optimal synchronization rate $R^*(d)$ can be bounded as

$$\lim_{n\to\infty} \frac{1}{n} H(\underline{X}|\underline{Y}\ \underline{S}) \le R^*(d) \le \lim_{n\to\infty} \frac{1}{n} \left(H(\underline{S}|\underline{Y}) + H(\underline{X}|\underline{Y}\ \underline{S}) \right)$$

where $\lim_{n\to\infty} \frac{1}{n} H(\underline{S}|\underline{Y})$ is given by (18) and

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} H(\underline{X}|\underline{Y}\ \underline{S}) &= \frac{(1-d)(2-\gamma-\gamma d) \log_2 \frac{1}{1-\gamma d}}{(1-\gamma d)} \\ &\quad + \frac{d(1-\gamma)^2 h(d\gamma)}{(1-d\gamma)^2} + \left(d - \frac{d(1-\gamma)^2}{(1-\gamma d)^2} \right) \log_2 \frac{1}{\gamma d} + \frac{(1-\gamma)^2}{\gamma} S, \end{aligned}$$

with

$$S \triangleq -\sum_{k=1}^{\infty} \sum_{j=1}^{\infty} ((1-d)\gamma)^k \, (d\gamma)^j \binom{j+k}{k} \log_2 \binom{j+k}{k}.$$

Figure 2 shows the upper and lower bounds of Proposition 2 for a first-order Markov source with $\gamma = 0.75$ for various values of d. We see that the gap between the upper and lower bounds grows with d. This is because as d increases from 0, more runs are deleted, and we need a larger rate to augment

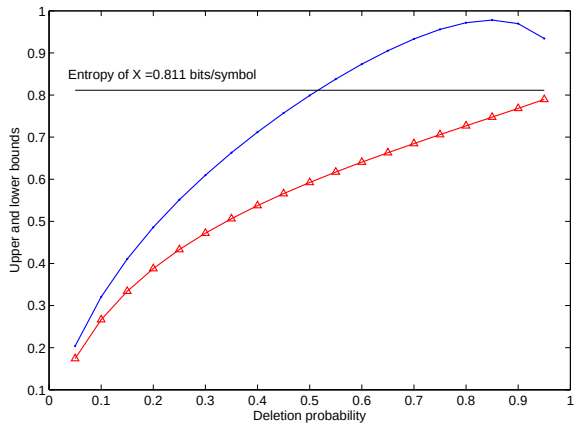

Fig. 2. Upper and lower bounds on the optimal rate for the deletion model for a first-order Markov source with parameter $\gamma = 0.75$. The entropy/bit of $\mathbf{X} = h(\gamma)$ is also shown.

the decoder with $\underline{S}$. Around $d = 0.5$, the upper bound starts exceeding the per-symbol entropy of the source $h(\gamma)$. Hence, for $d > 0.5$, it is not worth spending the additional rate to equip the decoder with the sequence $\underline{S}$ of deleted runs; it is more efficient to just send the entire $\underline{X}$ sequence.

The price we pay (over the optimal rate) for augmenting the decoder with the sequence $\underline{S}$ is

$$\frac{1}{n}(H(\underline{X}|\underline{Y}) - H(\underline{X}\ \underline{S}|\underline{Y})) = \frac{1}{n}H(\underline{S}|\underline{X}\underline{Y}).$$

In words, $H(\underline{S}|\underline{X}\ Y)$ represents the uncertainty in the positions of the deleted runs given both $\underline{X}$ and $\underline{Y}$. As an example, consider

$$\underline{X} = 001100 \quad \underline{Y} = 00.$$

Given both $\underline{X}$ and $\underline{Y}$, we know that the run of ones was deleted, but we do not know where the deleted run markers should be inserted. It is possible to bound $H(\underline{S}|\underline{X}\underline{Y})$ by calculating the entropy of such events, and thereby improve the upper bound of Proposition 2. This will be discussed in an extended version of this paper.

V. Insertion Model

In the insertion model, an extra bit may be inserted after each bit of $\underline{X}$ with probability i, where $0 < i < 1$. When a bit is inserted after X_j, the inserted bit is equal to X_j with probability α, and equal to $\bar{X}_j$ with probability $1 - \alpha$. We call the former a duplication, and the latter a complementary insertion. For large n, from the law of large numbers, the length m of $\underline{Y}$ will be close to $n(1 + i)$.

First consider the case of $\alpha = 1$, i.e., we only have duplications. Here, we can associate each run of $\underline{Y}$ with a unique run in $\underline{X}$, which leads to a product-form representation for the joint distribution $P(\underline{X}, \underline{Y})$. As before, denoting the runs of $\underline{X}$ and $\underline{Y}$ by $L_{X1}, L_{X2}, \ldots$ and $L_{Y1}, L_{Y2}, \ldots$, respectively, we have:

$$\begin{aligned} P(\underline{X}, \underline{Y}) &= P(L_{X1}L_{X2}\ldots) \cdot P(L_{Y1}L_{Y2}\ldots|L_{X1}L_{X2}\ldots) \\ &= P(L_{X1})P(L_{Y1}|L_{X1}) \cdot P(L_{X2})P(L_{Y2}|L_{X2})\ldots \end{aligned} \tag{19}$$

where $\forall j$:

$$P(L_{Yj} = s|L_{Xj} = r) = \binom{r}{s-r} i^{s-r}(1-i)^{2r-s},\ r \le s \le 2r. \tag{20}$$

Therefore, the optimal synchronization rate when $\alpha = 1$ is

$$\lim_{n\to\infty} \frac{1}{n} H(\underline{X}|\underline{Y}) = (1-\gamma)H(L_{X1}|L_{Y1}),$$

which is straightforward to compute using (20). We note that the channel coding problem corresponding to $\alpha = 1$, dubbed the sticky insertion channel, was studied in [19].

For $0 < \alpha < 1$, the inserted bits may create new runs, and so we cannot associate each run of $\underline{Y}$ with a run in $\underline{X}$. To see this, consider the following example. Let

$$\underline{X} = 000111000, \quad \underline{Y} = 00\mathit{1}011\mathit{1}00000\mathit{0}, \tag{21}$$

where the inserted bits are indicated in italics. There is one duplication - in the third run, and two complementary insertions - in the first and second runs. While a duplication never introduces a new run, a complementary insertion introduces a new run (e.g., the 1 inserted in the first run), except when it occurs at the end of a run of $\underline{X}$ (the 0 inserted at the end of the second run).

For $0 < \alpha < 1$, it appears difficult to even write a succinct expression for the joint distribution $P(\underline{X}, \underline{Y})$, so calculating the conditional entropy rate exactly is not feasible.

Suppose now that an auxiliary sequence $\underline{T} = (T_1, \ldots, T_m)$ of length m is available at the decoder, where $T_j = 1$ if bit Y_j is a complementary insertion, and $T_j = 0$ otherwise. For the $(\underline{X}, \underline{Y})$ pair in (21), $\underline{T} = (0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0)$ since only the first two insertions are complementary.

Using the auxiliary sequence $\underline{T}$, the decoder can remove the complementary insertions from $\underline{Y}$ (the bits indicated by the ones in $\underline{T}$) to form an updated sequence $\underline{Y}'$. The runs of $\underline{Y}'$ will be in one-to-one correspondence with the runs of $\underline{X}$. The auxiliary sequence $\underline{T}$ can be used to obtain a computable upper bound on the optimal synchronization rate for the insertion model as follows.

$$\begin{aligned} H(\underline{X}|\underline{Y}) &\le H(\underline{X}, \underline{T}|\underline{Y}) \\ &= H(\underline{T}|\underline{Y}) + H(\underline{X}|\underline{T}\ \underline{Y}) \\ &\le H(\underline{T}|\underline{Y}) + H(\underline{X}|\underline{Y}') \end{aligned} \tag{22}$$

where the last inequality holds because $\underline{Y}'$ is a function of $\underline{Y}$ and $\underline{T}$, obtained by removing just the complementary insertions from $\underline{Y}$. We now compute each of the terms in (22).

Computing $H(\underline{T}|\underline{Y})$: The term $\frac{1}{n}H(\underline{T}|\underline{Y})$, which represents the additional rate needed to convey the auxiliary sequence to

the decoder, can be upper bounded as follows. We have

$$\begin{aligned} H(\underline{T}|\underline{Y}) &= \sum_{j=1}^{m} H(T_j|T^{j-1}\ \underline{Y}) \\ &\leq \sum_{j=1}^{m} H(T_j|T_{j-1}\ Y_{j-1}\ Y_j) \\ &= n(1+i\pm\epsilon_n)H(T_j|T_{j-1}\ Y_{j-1}\ Y_j). \end{aligned} \tag{23}$$

Observe that that $T_j = 0$ if $T_{j-1} = 1$ since the model allows only one insertion after each bit. Also note that $T_j = 0$ if the triple (Y_{j-1}, Y_j) is not equal to either $(1,0)$ or $(0,1)$. Using the notation $(y, \bar{y})$ to represent either a $(1,0)$ or $(0,1)$, we need to consider three different ways in which $(Y_{j-1} = y, Y_j = \bar{y})$ can occur with $T_{j-1} = 0$: (original bit, inserted bit), (original, original), and (inserted, original). Out of these three ways, only the first corresponds to $T_j = 1$. In a typical $\underline{Y}$ sequence of length close to $n(1+i)$, the number of times each of the above patterns appears can be calculated and is given below:

Pattern $(Y_{j-1} = y, Y_j = \bar{y}$	No. of occurrences
(original, inserted)	$ni\bar{\alpha}$
(original, original)	$n(1-\gamma)(1-i)$
(inserted, original)	$ni\alpha(1-\gamma)$

Hence the total number of times the pattern $(y, \bar{y})$ appears in the sequence $\underline{Y}$ is the sum of the quantities in (V). This is equal to $n(1-\gamma+\gamma i\bar{\alpha})$, where The conditional entropy $H(T_j|T_{j-1}, Y_{j-1}, Y_j)$ in (23) is thus equal to

$$H(T_j|T_{j-1}\ Y_{j-1}Y_j) = \frac{n(1-\gamma+\gamma i\bar{\alpha})}{n(1+i)} h\left(\frac{i\bar{\alpha}}{1-\gamma+\gamma i\bar{\alpha}}\right). \tag{24}$$

Using this in (23), we obtain

$$\lim_{n\to\infty} \frac{1}{n} H(\underline{T}|\underline{Y}) \leq (1-\gamma+\gamma i\bar{\alpha})\, h\left(\frac{i\bar{\alpha}}{1-\gamma+\gamma i\bar{\alpha}}\right). \tag{25}$$

Computing $H(\underline{X}|\underline{Y}')$: Recall that $\underline{Y}'$ is the sequence formed from $\underline{Y}$ by removing the complementary insertions. The runs of $\underline{X}$ and $\underline{Y}'$ are in one to one correspondence, and so the joint distribution of $(\underline{X}, \underline{Y}')$ can be expressed as a product of $P(L_{Xj}, L_{Y'j})$, $j = 1, 2, \ldots$, where L_{Xj} ($L_{Y'j}$) represents the length of the jth run of $\underline{X}$ ($\underline{Y}$). Thus we have

$$H(\underline{X}|\underline{Y}') = n(1-\gamma\pm\epsilon_n)H(L_{X1}|L_{Y'1})$$

where $\epsilon_n \to 0$ as $n \to \infty$. Since only duplications occur in the $L_{Xj} \to L_{Yj}$ transformation, the joint distribution $P(L_{Xj}, L_{Y'j})$ for all runs j can be written as

$$\begin{aligned} P(L_{Xj} = r) &= \gamma^{r-1}(1-\gamma),\ r = 1, 2, \ldots \\ P(L_{Y'j} = s|L_{Xj} = r) &= \binom{r}{s-r} (i\alpha)^{s-r}(1-i\alpha)^{2r-s},\ r \leq s \leq 2r. \end{aligned} \tag{26}$$

Using this, we can compute $H(L_{X1}|L_{Y'1})$ and obtain

$$\lim_{n\to\infty} \frac{1}{n} H(\underline{X}|\underline{Y}') = (1-\gamma)H(L_{X1}|L_{Y'1}).$$

The upper bound on the optimal rate for the insertion model is given by the following proposition.

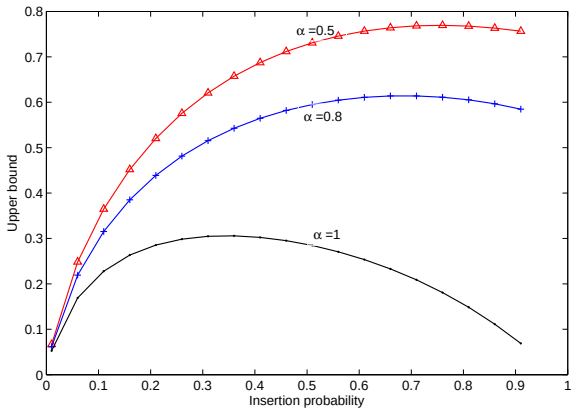

Fig. 3. Upper bound on the optimal rate for the insertion model for first-order Markov source with parameter $\gamma = 0.75$. Curves for $\alpha = 0.5, 0.8$, and 1 are shown.

Proposition 3: Let $\mathbf{X}$ be a binary first-order Markov source with parameter γ, and let $\mathbf{Y}$ be generated according to the insertion model with duplication probability $i\alpha$, and complementary insertion probability $i\bar{\alpha}$. Then the optimal synchronization rate $R^*(i,\alpha)$ can be upper bounded as

$$R^*(i,\alpha) \leq (\bar{\gamma}+\gamma i\bar{\alpha})\, h\left(\frac{i\bar{\alpha}}{\bar{\gamma}+\gamma i\bar{\alpha}}\right) + \bar{\gamma} H(L_{X1}|L_{Y'1})$$

where $H(L_{X1}|L_{Y'1})$ is computed using the joint distribution in (26).

Remarks: To obtain the upper bound, we need a rate of $\frac{1}{n}H(\underline{T}|\underline{Y})$ to augment the decoder with the auxiliary sequence $\underline{T}$. Given $\underline{T}$, the decoder needs an additional rate of $\frac{1}{n}H(\underline{X}|\underline{TY})$. This is reflected in the second line of (22). The inequality in third line of (22) may be interpreted as follows. Instead of using the sequences $(\underline{T}, \underline{Y})$ directly, the decoder only uses a function $\underline{Y}'$ of these sequences. (Recall that $\underline{Y}'$ is formed by removing the complementary insertions in $\underline{Y}$, which correspond to the ones in $\underline{T}$.) We do this because the joint distribution of $(\underline{X}, \underline{Y}')$ can be decomposed as a product of runs, with a single-letter distribution given by (26).

However, we pay a price for this tractability because using $\underline{Y}'$ instead of $(\underline{T}, \underline{Y})$ is sub-optimal in general. To see this, consider the extreme case of $i = 1$ with some $0 < \alpha < 1$. This means there is an insertion after *every* bit of $\underline{X}$. The synchronization rate $H(\underline{X}|\underline{Y})$ is zero because the decoder can just discard all the even bits of $\underline{Y}$ to recover $\underline{X}$. In contrast, $H(\underline{X}|\underline{Y}')$ is strictly positive since the decoder cannot determine the duplications exactly once the complementary insertions are removed.

This is illustrated in Figure 3, where the upper bound of Proposition 3 is plotted for a first-order Markov source $\mathbf{X}$ with $\gamma = 0.75$. The entropy of this source is 0.811 bits/source symbol. For $\alpha = 1$, there are only duplications, and there is no auxiliary sequence. Therefore the upper bound is tight and is equal to the optimal synchronization rate. Accordingly, for $\alpha = 1$, the synchronization rate goes to zero as i gets close to 1. However, the upper bound is clearly not tight for $\alpha = 0.5$

and 0.8 since it does not go to 0 as i gets large. The smaller the α, the greater the fraction of complementary insertions, and larger the overhead for large i.

Computing the quantity $H(\underline{X}|\underline{TY})$ exactly will significantly improve the upper bound for large i, as well provide a lower bound on the optimal synchronization rate. This is an interesting direction for future work.

VI. Deletion + Insertion Model

In this model, to generate $\underline{Y}$ from $\underline{X}$, each bit of $\underline{X}$ is deleted with probability d, or an extra bit may be inserted with probability i, or the bit is retained as is with probability $1-d-i$. When a bit is inserted, it is a duplication with probability α, and a complementary insertion with probability $1-\alpha$. For large n, from the law of large numbers, the length m of $\underline{Y}$ will be close to $n(1+i-d)$.

We can think of $\underline{Y}$ as being generated from $\underline{X}$ in two steps: First generate an intermediate sequence $\underline{Z}$ from $\underline{X}$ by deleting each bit with probability d. Then, after each bit of $\underline{Z}$, an extra bit is inserted with probability $i' \triangleq \frac{i}{1-d}$. Inserted bits are duplications with probability α, or complementary insertions with probability $1-\alpha$.

One way to synchronize $\underline{Y}$ to $\underline{X}$ is first recover $\underline{Z}$ with the rate specified by Proposition 3, and then recover $\underline{X}$ from $\underline{Z}$ using the rate specified by Proposition 2. However, this is not an efficient way of synchronizing because it does not take advantage of the fact that deletions and duplications within the same run cancel each other out. We propose a better way to synchronize by considering two auxiliary sequences at the decoder - one indicating complementary insertions, and the other indicating deleted runs.

Suppose the decoder is given two auxiliary sequences, $\underline{T}$ and $\underline{S}$. $\underline{T}$ is a sequence of the same length as $\underline{Y}$ indicating the complementary insertions in $\underline{Y}$. $\underline{Y}$ has length approximately $n(1-d+i)$, out of which there are close to $ni\bar{\alpha}$ complementary insertions. As in the insertion model, the decoder uses $\underline{T}$ to eliminate the complementary insertions from $\underline{Y}$ to form the sequence $\underline{Y}'$, which has length approximately $n(1-d+i\alpha)$. Now another, auxiliary sequence $\underline{S}$ indicates the positions where new runs need to be inserted in $\underline{Y}'$. Using this, we can create an augmented sequence $\underline{Y}''$ in which missing runs are indicated by "-" markers, as in Section IV. The runs of the augmented sequence $\underline{Y}''$ are in one to one correspondence with the runs of $\underline{X}$.

The auxiliary sequences $(\underline{S},\underline{T})$ can be used to obtain a computable upper bound on the optimal synchronization rate for the deletion+insertion model as follows. We have

$$\begin{aligned} H(\underline{X}|\underline{Y}) &\le H(\underline{X},\underline{T},\underline{S}|\underline{Y}) \\ &= H(\underline{T}|\underline{Y}) + H(\underline{S}|\underline{T}\ \underline{Y}) + H(\underline{X}|\underline{S}\ \underline{T}\ \underline{Y}) \\ &\le H(\underline{T}|\underline{Y}) + H(\underline{S}|\underline{Y}') + H(\underline{X}|\underline{Y}'') \end{aligned} \tag{27}$$

where the last inequality holds because (a) $\underline{Y}'$ is a function of $\underline{Y}$ and $\underline{T}$, obtained by removing the complementary insertions from $\underline{Y}$, and (b) $\underline{Y}''$ is equivalent to $(\underline{S},\underline{Y}')$. The term $H(\underline{S}\ \underline{T}|\underline{Y})$ represents the bits needed to convey the auxiliary sequences to the decoder. In order to obtain a single-letter characterization, this term has been bounded by $H(\underline{T}|\underline{Y}) + H(\underline{S}|\underline{Y}')$. We now compute each of the terms in (27).

Computing $H(\underline{T}|\underline{Y})$: We have

$$\begin{aligned} H(\underline{T}|\underline{Y}) &= \sum_{j=1}^{m} H(T_j|T^{j-1}\ \underline{Y}) \\ &\le \sum_{j=1}^{m} H(T_j|T_{j-1},Y_{j-1},Y_j) \\ &= n(1+i-d\pm\epsilon_n)H(T_j|T_{j-1}\ Y_{j-1}Y_j) \end{aligned} \tag{28}$$

The conditional entropy $H(T_j|T_{j-1},Y_{j-1},Y_j)$ in (28) can be calculated in manner similar to Section V is equal to

$$H(T_j|T_{j-1}Y_{j-1}Y_j) = \left(\frac{\bar{q}+qi'\bar{\alpha}}{1+i'}\right) h\left(\frac{i'\bar{\alpha}}{\bar{q}+qi'\bar{\alpha}}\right) \tag{29}$$

where $i' \triangleq \frac{i}{(1-d)}$ and $q \triangleq \frac{\gamma+d-2\gamma d}{1+d-2\gamma d}$. Using this in (23), we obtain

$$\lim_{n\to\infty}\frac{1}{n}H(\underline{T}|\underline{Y}) \le (\bar{q}(1-d)+qi\bar{\alpha})h\left(\frac{i\bar{\alpha}}{\bar{q}(1-d)+qi\bar{\alpha}}\right). \tag{30}$$

Computing $H(\underline{S}|\underline{Y}')$: Given $(\underline{T},\underline{Y})$, we can remove the complementary insertions from $\underline{Y}$ to form $\underline{Y}'$, whose length is approximately $n(1-d+i\alpha)$. We have

$$\begin{aligned} H(\underline{S}|\underline{Y}') &= \sum_{j} H(S_j|S^{j-1}\ \underline{Y}) \\ &\le n(1-d+i\alpha)H(S_j|Y_j'\ Y_{j+1}'). \end{aligned} \tag{31}$$

$H(S_j|Y_j'\ Y_{j+1}')$ can be calculated to be

$$\begin{aligned} H(S_j|Y_j'\ Y_{j+1}') = \frac{1}{1+i'\alpha}\Bigg[&\left(i'\alpha+\frac{\gamma(1-d)}{1-\gamma d}\right)\log_2\frac{i'\alpha+q}{i'\alpha+\frac{\gamma(1-d)}{1-\gamma d}} \\ &+\frac{\beta\theta}{1-\theta^2}\log_2\frac{i'\alpha+q}{\beta}+\frac{\beta\theta}{(1-\theta)^2}\log_2\frac{1}{\theta}+\frac{\beta}{1-\theta^2}\log_2\frac{1-q}{\beta}\Bigg] \end{aligned} \tag{32}$$

where

$$i' \triangleq \frac{i}{1-d},\ q \triangleq \frac{\gamma+d-2\gamma d}{1+d-2\gamma d},\ \theta \triangleq \frac{(1-\gamma)d}{1-\gamma d},\ \beta \triangleq \frac{(1-\gamma)(1-d)}{(1-\gamma d)^2}.$$

Computing $H(\underline{X}|\underline{Y}'')$: The runs of $\underline{X}$ are in one to one correspondence with the runs of $\underline{Y}$. The joint distribution of $(\underline{X},\underline{Y}'')$ can therefore be expressed as a product of $P(L_{Xj},L_{Y''j})$, $j=1,2,\ldots$, where L_{Xj} ($L_{Y''j}$) represents the length of the jth run of $\underline{X}$ ($\underline{Y}$). The conditional distribution $p_{s|r} \triangleq P(L_{Y'j}=s|L_{Xj}=r)$ can be written as

$$\begin{aligned} p_{s|r} = \sum_{n_i\in\mathcal{I}}\binom{r}{n_i, r+n_i-s}(i\alpha)^{n_i}d^{r+n_i-s}(1-d-i\alpha)^{s-2n_i}, \\ 0\le s\le 2r. \end{aligned} \tag{33}$$

where $\mathcal{I}$, the set of possible values for the number of insertions, is given by

$$\mathcal{I} = \begin{array}{ll} \{0,1,\ldots,\lfloor\frac{s}{2}\rfloor\} & s\le r, \\ \{s-r,\ldots,\lfloor\frac{s}{2}\rfloor\} & s>r, \end{array} \tag{34}$$

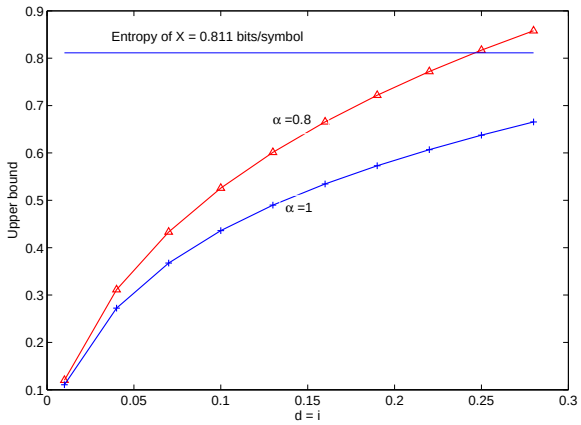

Fig. 4. Upper bound on the optimal rate for the deletion+insertion model with $d = i$ for first-order Markov source with parameter $\gamma = 0.75$. Curves for $\alpha = 0.8$ and $\alpha = 1$ are shown.

Using this, we can compute $H(L_{X1}|L_{Y''1})$, and obtain

$$\lim_{n\to\infty} \frac{1}{n} H(\underline{X}|\underline{Y}'') = (1-\gamma)H(L_{X1}|L_{Y''1}).$$

The upper bound on the optimal rate for the deletion+insertion model is given by the following proposition.

Proposition 4: Let $\mathbf{X}$ be a binary first-order Markov source with parameter γ, and let $\mathbf{Y}$ be generated according to the deletion+insertion model with deletion probability d, duplication probability $i\alpha$, and complementary insertion probability $i\bar{\alpha}$. Then the optimal synchronization rate $R^*(d, i, \alpha)$ can be upper bounded as

$$R^*(d, i, \alpha) \leq \bar{\gamma} H(L_{X1}|L_{Y''1}) + T_1 + T_2$$

where $H(L_{X1}|L_{Y''1})$ is computed using the joint distribution in (33) and

$$T_1 = (\bar{q}(1-d) + qi\bar{\alpha}) h\left(\frac{i\bar{\alpha}}{\bar{q}(1-d) + qi\bar{\alpha}}\right),$$

$$T_2 = (1-d)\left[\left(i'\alpha + \frac{\gamma(1-d)}{1-\gamma d}\right)\log_2 \frac{i'\alpha + q}{i'\alpha + \frac{\gamma(1-d)}{1-\gamma d}} + \frac{\beta\theta}{1-\theta^2}\log_2 \frac{i'\alpha+q}{\beta} + \frac{\beta\theta}{(1-\theta)^2}\log_2 \frac{1}{\theta} + \frac{\beta}{1-\theta^2}\log_2 \frac{1-q}{\beta}\right],$$

$$i' \triangleq \frac{i}{1-d},\ q \triangleq \frac{\gamma + d - 2\gamma d}{1 + d - 2\gamma d},\ \theta \triangleq \frac{(1-\gamma)d}{1-\gamma d},\ \beta \triangleq \frac{(1-\gamma)(1-d)}{(1-\gamma d)^2}.$$

The upper bound is plotted in Figure 4 for a first-order Markov source $\underline{X}$ with $\gamma = 0.75$ with $d = i$. The two curves correspond to duplication probability $\alpha = 0.8$ and $\alpha = 1$.

VII. Conclusion

We considered the problem of determining the minimal rate for synchronizing two sequences which differ from one another by a process of i.i.d deletions and insertions. Though this is essentially a distributed source coding problem, the optimal rate is difficult to compute due to the memory in the joint distribution of the two sources. Our approach was to augment the decoder with minimal extra information to reduce it to a tractable memoryless problem. This extra information indicated the locations of deleted and inserted runs. One obvious question is: are there other choices for the auxiliary sequences which result in lower overhead, but still result in tractable problem? As discussed at the end of Sections IV and V, even with the current choice of auxiliary sequences, we can improve the bounds if we can precisely compute the additional rate to supply these sequences. Another interesting direction is to use these techniques to obtain good lower bounds on the capacity of channels with synchronization errors. This will be addressed in an upcoming paper.

References

[1] A. Orlitsky, "Interactive communication of balanced distributions and of correlated files," *SIAM J. Discrete Math.*, vol. 6, no. 4, pp. 548–564, 1993.

[2] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, pp. 471–480, July 1973.

[3] A. Tridgell and P. Mackerras, "The rsync algorithm." http://rsync.samba.org/, Nov 1998.

[4] H. Zhang, C. Yeo, and K. Ramchandran, "VSYNC: a novel video file synchronization protocol," in *ACM Multimedia*, pp. 757–760, 2008.

[5] S. Agarwal, V. Chauhan, and A. Trachtenberg, "Bandwidth efficient string reconciliation using puzzles," *IEEE Trans. Parallel Distrib. Syst.*, vol. 17, no. 11, pp. 1217–1225, 2006.

[6] R. Venkataramanan, H. Zhang, and K. Ramchandran, "Interactive low-complexity codes for synchronization from deletions and insertions," in *Proc. 48th Annual Allerton Conference on Communication, Control, and Computing*, 2010.

[7] A. Wyner, "Recent results in the Shannon theory," *IEEE Trans. Inf. Theory*, vol. 20, pp. 2–10, Jan 1974.

[8] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 626–643, 2003.

[9] R. G. Gallager, "Sequential decoding for binary channels with noise and synchronization errors," October 1961. Lincoln Lab Group Report.

[10] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 18–36, 1967.

[11] S. N. Diggavi and M. Grossglauser, "On information transmission over a finite buffer channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1226–1237, 2006.

[12] E. Drinea and M. Mitzenmacher, "Improved lower bounds for the capacity of i.i.d. deletion and duplication channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 8, pp. 2693–2714, 2007.

[13] E. Drinea and A. Kirsch, "Directly lower bounding the information capacity for channels with i.i.d.deletions and duplications," *IEEE Trans. Inf. Theory*, vol. 56, pp. 86–102, January 2010.

[14] S. Diggavi, M. Mitzenmacher, and H. Pfister, "Capacity upper bounds for the deletion channel," in *Proc. Int. Symp. on Inf. Theory*, 2007.

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

[16] D. Fertonani and T. M. Duman, "Novel bounds on the capacity of the binary deletion channel," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2753–2765, June 2010.

[17] A. Kalai, M. Mitzenmacher, and M. Sudan, "Tight asymptotic bounds for the deletion channel with small deletion probabilities," in *Proc. Int. Symp. on Inf. Theory*, 2010.

[18] Y. Kanoria and A. Montanari, "On the deletion channel with small deletion probability," in *Proc. Int. Symp. on Inf. Theory*, 2010.

[19] M. Mitzenmacher, "Capacity bounds for sticky channels," *IEEE Trans. on Inf. Theory*, p. 2008, 72-77.

[20] T. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. on Inf. Theory,*, vol. 21, pp. 226 – 228, March 1975.

[21] R. L. Dobrushin, "General formulation of shannon's main theorem in information theory," *Uspekhi Mat. Nauk*, vol. 14, no. 6, pp. 3–104, 1959. English transl. in Amer. Math. Soc. Transl. (2) 33, 1963.