On the Information Rates of Channels with Insertion/Deletion/Substitution Errors

Jun Hu  Tolga M. Duman  
Arizona State University  
Tempe, AZ 85287-5706, USA  
Email: {hujun, duman}@asu.edu

M. Fatih Erden  
Seagate Technology  
Pittsburgh, PA 15203, USA  
Email: {fatih.erdem}@seagate.com

Abstract—In this work, we propose to use trellis structures to characterize channels with insertions, deletions and substitutions. We start with binary-input binary-output channels with either insertions or deletions, and develop a simulation based approach to estimate the corresponding information rates. This approach is then generalized to the case where there exist both insertions and deletions. We show that, the proposed algorithm is an efficient and flexible technique to closely estimate the information rates for channels with insertions, deletions and substitutions.

Index Terms—Information rates, insertion/deletion/substitution channels, trellis.

I. INTRODUCTION

For communication channels with synchronization errors, insertions or deletions may occur, i.e., symbols may be randomly inserted into or deleted from the data sequence. For such channels, exact capacity limits are generally not known. Over the years, several researchers [1]–[6] have provided various analytical lower and upper bounds to characterize the capacity limits for different types of insertion/deletion models. Most of the recent work derives analytical lower bounds by following Shannon’s approach, i.e., they choose a random codebook, define “typical” outputs and use suitable decoding algorithms to find the matching codeword. In this work, we consider a different approach, namely a trellis-based technique, to compute the information rates.

In the past few years, simulation based approaches have been proposed for information rate calculation for channels with memory [7]–[10]. The basic idea is to generate a trellis based on the channel model, transmit a long input sequence, and then apply the forward recursion of the BCJR algorithm [11] to calculate the joint probability of the output sequence. This joint probability can then be used to determine the entropy rate of the channel output and the mutual information between the input and output processes. This approach is shown to be flexible and efficient for many channel models. It has also been exploited in the calculation of capacity bounds for channels with synchronization errors. Specifically, [13] proposes a reduced-state technique to obtain a lower bound on information rates for insertion/deletion channels, while [14] employs the simulation-based approach to derive bounds on information rates for intersymbol interference (ISI) channels with timing errors. In this approach, the trellis to compute the conditional entropy is based on the knowledge of the input sequence. Here, we develop a trellis structure that is independent of the channel inputs and is flexible and easy to extend to different channel models. Also, since the trellis is not conditional on the input, the approach may also be useful for practical decoding purposes.

The organization of the paper is as follows: In the next section, we introduce the specific channel models formally. Section III presents a detailed discussion on the trellis construction and information rate calculations for binary insertion/deletion channels. Then Section IV extends the technique to channels with both insertions and deletions. Finally, Section V concludes this paper.

II. CHANNEL MODEL

The basic model we consider is an i.i.d. binary insertion (deletion) channel characterized by a parameter \( P_I (P_D) \) (insertion (deletion) probability). The input sequence is \( x^N = (x_1, \ldots, x_N) \) where each bit takes on the values 0 or 1. During the transmission, random bits are inserted independently with probability \( P_I \) (for the insertion channel) and the transmitted bits are deleted independently with probability \( P_D \) (for the deletion channel). At the receiver end, the binary output sequence is denoted as \( y^N = (y_1, \ldots, y_N) \), where \( N' \) may be larger/smaller than \( N \) due to insertions/deletions.

The capacity of general insertion/deletion channels is defined by [12]

\[
C = \lim_{N \to \infty} \frac{1}{N} \sup_{p(x^N_1)} I(x^N_1; y^N_{1+N'}),
\]

where the mutual information between the input sequence and the output sequence is maximized over the joint distribution of the input sequence. Here, we are interested in computing the information rates for binary insertion/deletion channels with independent and uniformly distributed (i.u.d.) inputs, given by

\[
C_{iud} = \lim_{N \to \infty} \frac{1}{N} I(x^N_1; y^N_{1+N'}),
\]

where \( x^N_1 \) is a binary sequence with each element drawn independently from a uniform distribution. In general, since the input is constrained, \( C_{iud} \) is smaller than or equal to \( C \). Note that, different input distributions give different information rates. Although we only consider i.u.d. inputs in this work, other types of inputs, e.g. Markovian inputs, can also be incorporated using a similar approach.
If there are possible input and output combinations at the each pair of consecutive transmitted bits and an upper limit \( C \) and each state has at most \( x \) the transmission of \( A \) is not satisfied and \( A \) represents its complementary event. \( P_A = P(A) \) is directly related to \( P_I \), \( M_I \) and \( B_I \). In many application scenarios, \( P_I \) has a very small value, thus \( P_A \) can be made arbitrarily small if we choose large enough values of \( M_I \) and \( B_I \) for a certain \( N \). Therefore, the mutual information \( I(x^N_1; y^N_1) \) can be well approximated by \( I(x^N_1; y^N_1|A) \). In the following, we do not differentiate between these two terms for notational simplicity. To proceed further, we note that

\[
I(x^N_1; y^N_1) = H(x^N_1) + H(y^N_1) - H(x^N_1, y^N_1),
\]

\[
= N - E[\log P(y^N_1)] + E[\log P(x^N_1, y^N_1)],
\]

where \( E[\cdot] \) denotes the expectation operator and \( H(x^N_1) = N \) due to the use of an i.i.d. input sequence. The only issue that remains is the calculation of \( E[\log P(y^N_1)] \) and \( E[\log P(x^N_1, y^N_1)] \). To accomplish this, we take a Monte Carlo based approach [7], [8] where we generate a large number of channel simulations (based on the given statistics), calculate \( P(y^N_1) \) and \( P(x^N_1, y^N_1) \) for all these realizations, and take their average to estimate the expected values. Let us demonstrate now these computations can be done in an efficient manner for a given input and output sequence.

3) Calculation of \( P(y^N_1) \): To calculate \( P(y^N_1) \), we first define the following forward recursion

\[
\alpha_k(j) = P(y^k_1, S_k = j),
\]

\[
= \sum_i P(y^k_{i-1}, S_{k-1} = i)P(y^k_j, S_k = j|S_{k-1} = i),
\]

\[
= \sum_i \alpha_{k-1}(i)\gamma_k(i, j),
\]

where

\[
\gamma_k(i, j) = P(y^k_{j+1}, S_k = j|S_{k-1} = i),
\]

\[
= P(y^k_{j+1}|S_{k-1} = i, S_k = j)P(S_k = j|S_{k-1} = i),
\]

\[
= P(y^k_{j+1}|S_{k-1} = i, S_k = j)P(S_k = j|S_{k-1} = i) \cdot P(y^k_{j+1}|S_{k-1} = i, S_k = j). \tag{6}
\]

The third equation follows the fact that: Given the state transition \( (S_{k-1} = i \rightarrow S_k = j) \), \( y^k_{j+1} \) is the observation of the original desired bit, \( y^k_{j+1} \) are the observations of inserted bits and they are independent.

Let us consider the first term of \( \gamma_k(i, j) \), calculated as

\[
P(y^k_{j+1}|S_{k-1} = i, S_k = j) = \sum_{\hat{x}_k} P(y^k_{j+1}|S_{k-1} = i, S_k = j, \hat{x}_k)P(\hat{x}_k),
\]

\[
= \frac{1}{2}[(1 - P_e) + P_e] = \frac{1}{2}, \tag{7}
\]

where \( \hat{x}_k \) is the input corresponding to the transition \( (S_{k-1} = i \rightarrow S_k = j) \). For the second term, \( y^k_{j+1} \) correspond to the \( j-x \) insertions between \( (k-1) \)th data bit and \( k \)th data bit. We assume that the inserted bits are independent and 0 or 1 with probability 1/2 each. Thus, we have

\[
P(y^k_{j+1}|S_{k-1} = i, S_k = j) = \frac{1}{2^j-i}. \tag{8}
\]
Now let us consider the term \( P(S_k = j|S_{k-1} = i) \), which depends on the state transition probability. Under the truncation condition, Eq. (3) is revised as
\[
P(S_k = j|S_{k-1} = i) = \frac{P^{(u)}(S_k = j|S_{k-1} = i)}{1 - P^Q_{j+1} T_i},
\]
\[i = 0, 1, \ldots, B_t, \quad j = i, i + 1, \ldots, i + Q_i,
\]
where \( Q_i = \min(M_t, B_t - i) \) and \((1 - P^Q_{j+1})\) is the normalization factor which appears due to the implicit conditioning on the event \( \tilde{A} \).

Therefore, we obtain
\[
P(y'N) = \alpha_N(N' - N),
\]
where we have implicitly assumed that we know the length of the received sequence \( N' \) at each simulation, i.e., the ending state is the state that we have \( N' - N \) insertions. We also note that, due to trellis termination, the state transition probability at the last time instant need to be revised accordingly.

4) Calculation of \( P(x_N^k, y_N'^k) \): Now let us consider \( P(x_N^k, y_N'^k) \). We define another forward recursion
\[
\varphi_k(j) = P(x_N^k, y_N'^k, S_k = j) = \sum_i \varphi_{k-1}(i) \theta_k(i, j),
\]
where
\[
\theta_k(i, j) = P(y_{k+i+1}, x_{k+i}|S_{k-1} = i, S_k = j)P(S_k = j|S_{k-1} = i),
\]
\[
= P(y_{k+i+1}, x_{k+i}|S_{k-1} = i, S_k = j)P(S_k = j|S_{k-1} = i)
\cdot P(y_{k+i}+1|S_{k-1} = i, S_k = j).
\]

Based on a similar reasoning as in the previous derivations, \( P(S_k = j|S_{k-1} = i) \) is given by Eq. (9), \( P(y_{k+i+1}+1|S_{k-1} = i, S_k = j) = 1/2^{j-i} \), and the first term is given by
\[
P(y_{k+i}, x_{k+i}|S_{k-1} = i, S_k = j) = \begin{cases} \frac{1}{2}(1 - P_e) & \text{if } y_{k+i} = x_k \\ \frac{1}{2} P_e & \text{if } y_{k+i} \neq x_k \end{cases},
\]
and, we obtain
\[
P(x_N^k, y_N'^k) = \varphi_N(N' - N).
\]

To summarize, to approximate the achievable information rates for i.i.d binary insertion channels, we first generate a long realization of input sequence with length \( N \) with a resulting output sequence with length \( N' \) using a simple channel simulation, use Eq. (5)-Eq. (10) and Eq. (11)-Eq. (14) to compute two joint probabilities and their logarithms. After repeating this process many times, we average the results to obtain a close approximation of \( C_{\text{ SSD}} \).

B. Deletion Channels

For the deletion channels, the channel model is similar to the insertion case, except that the parameter \( P_i \) is replaced with \( P_D \) (deletion probability). In this case, we consider the dual approach to the one used for the insertion channels.

A suitable channel trellis is illustrated in Fig. 2. In this case, each trellis segment corresponds to a received symbol and \( S_k = i \) denotes the state where there are \( i \) deletions before the reception of \( y_k \). If there are \( m \) deletions \((m = 0, 1, \ldots)\) between \( y_{k-1} \) and \( y_k \), a transition from the state \( S_{k-1} = i \) to \( S_k = j = i + m \) with the corresponding input segment of \( x_{k+i+m} \) will occur. We also consider a “truncation condition” where \( M_D \) is the upper limit for the number of deletions between each pair of consecutive received bits and \( B_D \) is the upper limit for the total number of deletions in the whole sequence. Based on this representation, we can use a similar forward recursion technique as in the previous section. In the following, we only summarize the results and omit the details:
\[
P(y_N'^k) = \beta_N(N' - N'),
\]
\[
\beta_k(j) = P(y_k, S_k = j) = \sum_i \beta_{k-1}(i) \lambda_k(i, j),
\]
\[
\lambda_k(i, j) = P(y_k|S_{k-1} = i, S_k = j)P(S_k = j|S_{k-1} = i),
\]
\[
= \frac{1}{2} P(S_k = j|S_{k-1} = i),
\]
\[
P(S_k = j|S_{k-1} = i) = P_D^j_{D-1}(1 - P_D)/(1 - P_D^T_{D+1}),
\]
\[i = 0, 1, \ldots, B_D, \quad j = i, i + 1, \ldots, i + T_i,
\]
where \( T_i = \min(M_D, B_D - i) \).

\[
P(x_N^k, y_N'^k) = \xi_N(N' - N'),
\]
\[
\xi_k(j) = P(x_N^k, y_N'^k, S_k = j) = \sum_i \xi_{k-1}(i) \pi_k(i, j),
\]
\[
\pi_k(i, j) = P(y_k, x_{k+i+1}+1|S_{k-1} = i, S_k = j)P(S_k = j|S_{k-1} = i),
\]
\[
= \frac{P(y_k, x_{k+i+1}+1|S_{k-1} = i, S_k = j)P(S_k = j|S_{k-1} = i)}{2^{j-1}},
\]
where
\[
P(y_k, x_{k+i+1}+1|S_{k-1} = i, S_k = j) = \begin{cases} \frac{1}{2}(1 - P_e) & \text{if } y_k = x_{k+i+1}+1 \\ \frac{1}{2} P_e & \text{if } y_k \neq x_{k+i+1}+1 \end{cases}
\]
and \( P(S_k = j|S_{k-1} = i) \) is given by Eq. (15).

C. Examples

Let us illustrate the proposed techniques via two examples. We show the information rate results for an i.i.d. binary deletion channel with no substitution errors \( P_e = 0 \) in Fig. 3.
For this binary deletion channel, the exact characterization of the channel capacity is generally not possible, and only lower and upper bounds were available (several of the well-known bounds [2]–[6] are also included in the figure). Clearly, our results are consistent with the earlier bounds. We stress that these information rates with constrained inputs serve as lower bounds for the actual unconstrained channel capacity.

We note again that our result is the information rate under the i.i.d. input constraints. As pointed out in [5], [6], by introducing memory into the input sequence, the lower bounds of Shannon capacity can be improved. Similarly, by incorporating different input constraints (e.g., Markovian inputs) into our approach, we may achieve higher information rates.

Several other comments are in order. First, to generate the results, we have used \( N = 100, M_D = 10 \) and \( B_D = 100 \). The choice of \( B_D \) provides a tradeoff between the complexity of calculation and the accuracy of results. For small values of \( B_D \), the computational complexity is lower, however, the results may not be very accurate. On the other hand, the choice of \( B_D \) also depends on the range of \( P_D \). For small \( P_D \), a small value for \( B_D \) is sufficient (as \( P_A \) is very small) and the results converge to the actual information rates very rapidly. For large values of \( P_D \), a large \( B_D \) is required. In many application scenarios for insertion/deletion channels, \( P_I/P_D \) is usually very small. Therefore, our technique is efficient.

As another example, we provide the information rates for the insertion channels with different levels of substitution errors in Fig. 4. We observe that, combination of insertion and substitution errors greatly degrades the information rates for successful communication. We also would like to emphasize that, although most of the earlier results are usually for channels with deletions or insertions only (without substitutions), our technique can easily incorporate substitution errors. Therefore, it is more flexible than the existing bounds.

IV. INFORMATION RATES FOR CHANNELS WITH BOTH INSERTIONS AND DELETIONS

A more general problem is the computation of channel capacity when there are both insertions and deletions during the transmission. The channel is now characterized by three parameters: \( P_I, P_D \) and \( P_e \). We study the case where insertions and deletions occur independently. At each time instant, each information bit is either transmitted or deleted (with probability \( P_D \)). Between a pair of transmitted bits, there may also exist random insertions (with each insertion happening with probability \( P_I \)). Furthermore, all the transmitted bits go through an imperfect channel where substitutions may happen with error probability \( P_e \). We will consider the extension of the approach in the previous section to this case and develop suitable algorithms to compute the corresponding information rates.

First, we develop a different trellis whose state is defined by a two-element vector \( i = (i_1, i_2) \), where \( i_2 \) denotes the total number of insertions and \( i_2 \) denotes the total number of deletions until the current time instant. The trellis is illustrated in Fig. 5. For each trellis segment, only one bit from the original input sequence is actually transmitted, i.e., there may be insertions and deletions before this transmission. Let us consider the transition from state \( S_{k-1} = i = (i_1, i_2) \) to state \( S_k = j = (j_1, j_2) \) where \( j_1 = i_1 + i_1 + m_1 \) and \( j_2 = i_2 + m_2 \). At the \( k \)th time instant, \( x_{k+j_2} \) is transmitted and \( m_2 \) deletions (input segment \( x_{k+j_2-1}^{k+j_2-1} \)) occur before this transmission. At the receiver side, \( y_{k+j_1} \) are the corresponding output sequence segment (including \( m_1 \) insertions).
The state transition probabilities are now given by
\[ P^{(n)}(S_k = j | S_{k-1} = i) = P^{(1)}(j | S_{k-1} = i) P^{(2)}(i | S_{k-1} = i) (1 - P_I) (1 - P_D), \]
\[ i_1 = 0, 1, \ldots, \]
\[ i_2 = 0, 1, \ldots, \]
\[ j_1 = i_1, i_1 + 1, \ldots, \]
\[ j_2 = i_2, i_2 + 1, \ldots. \]
(16)

By applying the “truncation condition”, they are modified as
\[ P(S_k = j | S_{k-1} = i) = \frac{P^{(n)}(S_k = j | S_{k-1} = i) (1 - P_I) (1 - P_D)}{(1 - P_I^{Q_{i_1} + 1}) (1 - P_D^{Q_{i_2} + 1})}, \]
\[ i_1 = 0, 1, \ldots, B_1, \]
\[ i_2 = 0, 1, \ldots, B_D, \]
\[ j_1 = i_1, i_1 + 1, \ldots, Q_{i_1}, \]
\[ j_2 = i_2, i_2 + 1, \ldots, Q_{i_2}. \]
(17)

where \( Q_{i_1} = \min(M_I, B_1 - i_1) \) and \( Q_{i_2} = \min(M_D, B_D - i_2) \).

As we can see, this new trellis is a combination of the ideas used in the trellis for both the insertion channel and the deletion channel. Therefore, the computation of the achievable information rates will also be the integration of the proposed algorithms for those two.

The two forward recursions are now defined as
\[ \alpha_k(j) = P(y_{1}^{k+j_1}, S_k = j) = \sum_{i_1} \sum_{i_2} \alpha_{k-1}((i) \gamma_k(i,j), \]
\[ \varphi_k(j) = P(x_{1}^{k+j_2}, y_{1}^{k+j_1}, S_k = j) = \sum_{i_1} \sum_{i_2} \varphi_{k-1}((i) \theta_k(i,j), \]
\[ \gamma_k(i,j) = P(y_{k+j_1}^{k+i_1} | S_{k-1} = i, S_k = j) P(S_k = j | S_{k-1} = i), \]
\[ \gamma_0(i) = \frac{1}{2} \cdot P(y_k^{k+i_1} | S_{k-1} = i, S_k = j) P(S_k = j | S_{k-1} = i), \]
\[ \theta_k(i,j) = P(y_{k+j_1}^{k+i_1}, x_{k+i_2}^{k+j_2}, S_k = j | S_{k-1} = i), \]
\[ \theta_0(i,j) = \frac{1}{2} \cdot \frac{1}{2} P(y_k^{k+i_1}, x_{k+i_2}^{k+j_2} | S_{k-1} = i, S_k = j) P(S_k = j | S_{k-1} = i), \]
\[ P(y_{k+j_1}, x_{k+j_2} | S_{k-1} = i, S_k = j) = \begin{cases} \frac{1}{2} (1 - P_e) & \text{if } y_{k+j_1} = x_{k+j_2} \\ \frac{1}{2} P_e & \text{if } y_{k+j_1} \neq x_{k+j_2}. \end{cases} \]

The two joint probabilities are then given by
\[ P(y_1^{N'}) = \sum_{j \in \Omega} \alpha_{N'}(j), \]
(18)
\[ P(x_1^{N'}, y_1^{N'}) = \sum_{j \in \Omega} \varphi_{N'}(j), \]
(19)

where \( N'' = \min(N, N'), \) and \( \Omega \) is defined as the set that include all the combinations of insertion and deletion patterns which could result in a received sequence with length \( N'. \)

Let us now give an example. Fig. 6 shows the information rates for binary channels with both insertions and deletions. In this 3-D plot, we can see that, the information rates decrease rapidly when the insertion and deletion error probabilities are increased, and therefore powerful coding techniques are necessary to ensure successful data recovery for insertion and deletion channels.

V. CONCLUSIONS

We have considered information rate computation for insertion/deletion/substitution channels. We have proposed to use a trellis structure to characterize the insertions or deletions introduced, and developed a forward recursion approach to estimate the information rates of such channels. By comparing with the existing upper and lower bounds, we show that this technique can provide very close approximations to the exact capacity limits, and by modifying the trellis representations, it also provides efficient and flexible solutions for different types of insertion/deletion channels.

REFERENCES