

Directly Lower Bounding the Information Capacity for Channels with I.I.D. Deletions and Duplications

Adam Kirsch and Eleni Drinea

Abstract—We directly lower bound the information capacity for channels with i.i.d. deletions and duplications. Our approach differs from previous work in that we focus on the information capacity using ideas from renewal theory, rather than focusing on the transmission capacity by analyzing the error probability of some randomly generated code using a combinatorial argument. Of course, the transmission and information capacities are equal, but our change of perspective allows for a much simpler analysis that gives more general theoretical results. We then apply these results to the binary deletion channel to improve existing lower bounds on its capacity.

I. INTRODUCTION

THIS work gives a new approach to lower bounding the asymptotic capacity of channels with i.i.d. deletions and duplications with arbitrary finite alphabets. Specifically, we consider channels that send an i.i.d. number (possibly zero) of copies of each transmitted symbol. These channels are a subset of the class of channels with synchronization errors, first analyzed by Dobrushin [2], who generalized Shannon’s channel coding theorem to show that for any such channel, the information and transmission capacities are equal. We apply Dobrushin’s result to obtain lower bounds on the transmission capacity of our channels by directly lower bounding their information capacities. Our techniques are substantially different from those in prior works, which typically lower bound the transmission capacity directly through a more combinatorial approach. Using our techniques, which are based on elementary facts from renewal theory, we are able to achieve more general theoretical results than prior work with a much cleaner analysis, along with improved lower bounds on the transmission capacities of the channels that we consider.

We start by giving a very brief overview of previous work along these lines and its connections to this paper. For concreteness, we temporarily restrict our attention to the case of a binary alphabet. After Dobrushin’s work [2], for the specific case of the deletion channel, a series of theoretical works [1], [3], [4] (and to some extent [6]) gave successively improved lower bounds on the transmission capacity. All of these works adhere to the following basic paradigm. First, a

set of 2^{nR} codewords of some length n is randomly chosen according to some symmetric first order Markov chain; here R is the rate of the code. (In other words, the codewords are obtained by generating alternating blocks of zeros and ones; the block lengths are i.i.d. samples from some geometric distribution.) Second, a decoding algorithm is proposed and analyzed. The analysis of the decoding algorithm results in an upper bound for the probability of an incorrect decoding. The supremum of the set of values for R that results in this upper bound vanishing as $n \rightarrow \infty$ is then a lower bound on the transmission capacity of the channel. (As an aside, [6] follows this paradigm indirectly by deriving a relationship between the deletion channel and another channel and then lower bounding the transmission capacity of the other channel using the basic paradigm.)

We take a different approach. By Dobrushin’s result [2], we can lower bound the transmission capacity by lower bounding the information capacity. That is, rather than generating codewords randomly from some distribution X over strings of length n , introducing an explicit decoding algorithm, and then deriving asymptotic error probability bounds, we simply compute $\lim_{n \rightarrow \infty} \frac{1}{n} I(X; Y)$, where Y denotes the received sequence when X is transmitted across the channel. For the commonly studied case where X is just the first n steps of some symmetric first order Markov chain on $\{0, 1\}$, this limit has a natural stochastic interpretation. Intuitively, as we scan Y from left to right, it *restarts* after every block. Thus it is natural to think of $\frac{1}{n} I(X; Y) = \mathbf{E}[Z]$ for $Z = \frac{1}{n} \log \frac{\Pr(X, Y)}{\Pr(X) \Pr(Y)}$ and then apply ideas from renewal theory to analyze Z . (In fairness, this intuition is implicit in the combinatorial analyses in [3], [4].) This results in an exact expression for $\lim_{n \rightarrow \infty} \frac{1}{n} I(X; Y)$, whereas prior work only gives lower bounds. As an added bonus, our analysis allows for a much wider variety of channels and block length distributions for X than prior work (e.g. [3], [4]). Specifically, our results apply to i.i.d. channels governed by an arbitrary deletion/duplication distribution with finite entropy and sources with i.i.d. block lengths generated from any distribution with finite mean and entropy, whereas the analyses in prior works force artificial restrictions on these distributions.

Furthermore, since the lower bounds on the transmission capacity from [4] can be interpreted as lower bounds on $\lim_{n \rightarrow \infty} \frac{1}{n} I(X; Y)$, we can exactly identify the improvement of our bounds over those in [4]. However, the expression for this improvement is too computationally challenging to evaluate to obtain numerical bounds. Even obtaining lower bounds for this expression is a very time-consuming process. In fact, an exact numerical computation seems too slow and so we resort to simulations. On the positive side, the addition

A preliminary version of this work was presented at the 2007 IEEE International Symposium on Information Theory.

Adam Kirsch is in the School of Engineering and Applied Sciences at Harvard University, Cambridge, MA 02138. He was supported in part by an NSF Graduate Research Fellowship and NSF grant CCF-0634923. Email: kirsch@eecs.harvard.edu

Eleni Drinea is in the School of Computer and Communication Sciences at EPFL, Lausanne, Switzerland. Part of this work was done while she was at the New England Complex Systems Institute (NECSI) and in the School of Engineering and Applied Sciences at Harvard University, Cambridge, MA 02138. While at Harvard, she was supported by NSF grant CCF-0634923. Email: eleni.drinea@epfl.ch

of this term to the previous bounds in [4] does seem to incur a non-negligible improvement for all but very small deletion probabilities, and therefore its evaluation appears worthwhile.

In effect, this work exhibits the best possible (theoretical) lower bounds on the transmission capacity that can be obtained by using random codes where codewords are generated with i.i.d. block lengths. The analyses in most prior works (e.g. [1], [3], [4]) intrinsically rely on this model. Indeed, if codewords are generated independently from some distribution X over strings of length n , then the best possible resulting lower bound on the transmission capacity is $\lim_{n \rightarrow \infty} \frac{1}{n} I(X; Y)$, which we determine exactly in the case where X has i.i.d. block lengths. Thus, for further improvements, one must consider codewords with dependencies among the block lengths (e.g., codewords generated by higher order Markov chains), and so the analysis in this paper no longer applies.

II. PRELIMINARIES

We consider finite alphabet channels with i.i.d. deletions and duplications. Formally, we fix some finite alphabet Σ with $|\Sigma| \geq 2$, and assume without loss of generality that $\Sigma = \{0, \dots, |\Sigma| - 1\}$. We also fix some distribution D on $\mathbb{Z}_{\geq 0}$. For convenience, we let $d_i = \Pr(D = i)$ for $i \in \mathbb{Z}_{\geq 0}$, and we assume that $d_0 < 1$ and $H(D) < \infty$. For concreteness, we measure all entropies and related functions in bits and we use $\log(\cdot)$ to denote the base 2 logarithm function. We consider the channel which, for some input string $x_1 \dots x_n \in \Sigma^n$, chooses D_1, \dots, D_n independently from D and then outputs the string consisting of D_1 copies of x_1 , followed by D_2 copies of x_2 , etc.

We let L be an arbitrary positive integer probability distribution with $\mathbf{E}[L], H(L) < \infty$. Let L_1, L_2, \dots be independent random variables with common distribution L . Let $A'_1 = 0$ and let A'_2, A'_3, \dots be i.i.d. random variables distributed uniformly on $\{1, \dots, |\Sigma| - 1\}$. Define the sequence of symbols A_1, A_2, \dots by $A_1 = A'_1$ and $A_i = (A_{i-1} + A'_i) \bmod |\Sigma|$ for $i > 1$. Note that for $i > 1$, the conditional distribution of A_i given A_1, \dots, A_{i-1} is uniform on $\Sigma - \{A_{i-1}\}$. We now define the infinite source $X = A_1^{L_1} A_2^{L_2} \dots$. That is, X is an infinite string consisting of L_1 copies of A_1 , followed by L_2 copies of $A_2 \neq A_1$, followed by L_3 copies of $A_3 \neq A_2$, etc.

Now we define a random variable Y that corresponds to the sequence received when we send X through the channel. We do this by sending X through the channel block-by-block. For each block sent, we look at the corresponding received sequence and see whether this string is entirely contained in the current block Y_j of Y , or whether this string is the prefix of a new block Y_{j+1} of Y (one of these cases must occur since the channel does not allow arbitrary insertions, just duplications). In this way, we can not only build the blocks Y_1, Y_2, \dots of Y , but we can also associate each block of Y with the group X_i of blocks in X from which it arose. This procedure is expressed formally in Figure 1.

The procedure in Figure 1 allows us to define the random variables Y_1, Y_2, \dots and X_1, X_2, \dots in a natural way. For any $k \geq 1$, once the variable j in the procedure is strictly greater than k , the variables Y_k and X_k remain constant, and we define

```

1: Set  $X_1 = Y_1 = \epsilon$  (the empty string).
2:  $j \leftarrow 1$ 
3: for  $i = 1$  to  $\infty$  do
4:    $B \leftarrow A_i^{L_i}$ 
5:    $\mathcal{B} \leftarrow$  the sequence received when  $B$  is sent.
       $\triangleright \mathcal{B}$  is a (possibly empty) string of all  $A_i$ 's.
       $\triangleright Y_j$  is a (possibly empty) string of all  $A_j$ 's.
6:   if  $A_i = A_j$  or  $|\mathcal{B}| = 0$  then
       $\triangleright \mathcal{B}$  is contained in the current block  $Y_j$  of  $Y$ .
7:      $Y_j \leftarrow Y_j \mathcal{B}$ 
8:      $X_j \leftarrow X_j B$ 
9:   else  $\triangleright \mathcal{B}$  is a prefix of the next block  $Y_{j+1}$  of  $Y$ .
10:     $j \leftarrow j + 1$ 
11:     $Y_j \leftarrow \mathcal{B}$ 
12:     $X_j \leftarrow B$ 

```

Fig. 1. The procedure for generating Y_1, Y_2, \dots , and X_1, X_2, \dots given X .

the value of the random variables Y_k and X_k to be those values. It is not hard to see that with probability 1, Line 10 is executed infinitely often and the lengths of the variables B and \mathcal{B} are always finite, and therefore X_k and Y_k are well-defined random variables with $\Pr(|X_k| < \infty) = \Pr(|Y_k| < \infty) = 1$. (We use the notation $|\cdot|$ to denote both the length of a string and the absolute value of a real number, with the meaning always being clear from context.) Thus, we may write $Y = Y_1 Y_2 \dots$ and $X = X_1 X_2 \dots$, where $|X_j|, |Y_j| < \infty$ and $|X_j| > 0$ for all j , and $|Y_j| > 0$ for $j > 1$.

We note that since the channel treats each symbol of X independently, Y really does correspond to the received sequence that results from sending X through the channel symbol-by-symbol. That is, it does not matter whether we generate Y by sending X through the channel symbol-by-symbol or block-by-block, as we do in the procedure in Figure 1.

Next we define some new notation. First, define $\mathcal{L}_1, \mathcal{L}_2, \dots$ and $\mathcal{A}_1, \mathcal{A}_2, \dots$ so that $Y_j = \mathcal{A}_j^{\mathcal{L}_j}$. Let $\mathcal{A}'_1 = \mathcal{A}_1$ and for $j > 1$, let $\mathcal{A}'_j = (\mathcal{A}_j - \mathcal{A}_{j-1}) \bmod |\Sigma|$. Clearly, the \mathcal{A}'_j 's are independent, and for $j \geq 2$, they are distributed uniformly on $\{1, \dots, |\Sigma| - 1\}$.

For symbols $t_1, t_2 \in \Sigma$, define $\sigma(t_1, t_2) = (t_1 - t_2) \bmod |\Sigma|$. For a string $s = s_1 \dots s_k \in \Sigma^*$ and symbol $t \in \Sigma$, define $\sigma(s, t) = \sigma(s_1, t) \dots \sigma(s_k, t)$ (where we are concatenating the symbols, not multiplying them). For the empty string ϵ , we set $\sigma(\epsilon, t) = \epsilon$. For $j \geq 1$, we let $G_j = \sigma(X_j, \mathcal{A}_j)$. We think of G_j as the group of blocks of X that gives rise to Y_j , normalized to remove any information about \mathcal{A}_j . The G_j 's are useful to us because the channel is symmetric with respect to transmitted symbols, and so we expect G_2, G_3, \dots to be identically distributed (G_1 is special since Y_1 is generated in a slightly different way than the other Y_j 's). Indeed, we have the following key lemma, which forms the basis of our subsequent analysis.

Lemma 1:

- 1) The $(G_j, \mathcal{L}_j, \mathcal{A}'_j)$'s are independent and for $j > 2$, they have some common distribution $(G, \mathcal{L}, \mathcal{A}')$, where \mathcal{A}' is uniformly distributed on $\{1, \dots, |\Sigma| - 1\}$ and is

independent of (G, \mathcal{L}) .

2) $\mathbf{E}[|G_1|], H(G_1, \mathcal{L}_1), \mathbf{E}[|G|], H(G, \mathcal{L}) < \infty$.

Proof: That the $(G_j, \mathcal{L}_j, \mathcal{A}'_j)$'s are independent is obvious from an examination of Figure 1. Similarly, it is easy to see that the randomness used between successive invocations of Line 10 is identically distributed, and therefore, for $j \geq 2$, the $(G_j, \mathcal{L}_j, \mathcal{A}'_j)$'s are identically distributed. Furthermore, it is also clear that \mathcal{A}' is uniformly distributed on $\{1, \dots, |\Sigma| - 1\}$ and is independent of (G, \mathcal{L}) .

The rest of the proof is just a sequence of calculations. We start with some technical preliminaries. Suppose that we send L symbols across the channel. Let L' denote the conditional distribution of L given that all L symbols are deleted. Let L'_1, L'_2, \dots denote i.i.d. samples from L' , and let L'' denote the distribution of $\sum_{i=1}^{\text{Geom}(1/(|\Sigma|-1))} L'_i$, where the geometric random variable is independent of the L'_i 's. (Throughout this paper, $\text{Geom}(p)$ denotes the geometric distribution with $\Pr(\text{Geom}(p) = j) = p(1-p)^{j-1}$ for $j \geq 1$.) Then

$$\mathbf{E}[L'] = \sum_{j=1}^{\infty} j \frac{\Pr(L = j)d_0^j}{\mathbf{E}[d_0^L]} = \frac{\mathbf{E}[Ld_0^L]}{\mathbf{E}[d_0^L]} < \infty.$$

and by Wald's equation,

$$\mathbf{E}[L''] = (|\Sigma| - 1) \mathbf{E}[L'] = (|\Sigma| - 1) \frac{\mathbf{E}[Ld_0^L]}{\mathbf{E}[d_0^L]} < \infty.$$

Next, let S denote the longest suffix of G that does not contain the symbol 0, and let J denote the number of blocks in S . (Note that $J = 0$ if $|\Sigma| = 2$.) To describe the distributions of S and J , consider sending $T = A_2^{L_2} A_3^{L_3} \dots$ across the channel. Recall that A_2 is uniformly distributed on $\Sigma - \{0\}$ and that for $i \geq 2$, the conditional distribution of A_i given A_1, \dots, A_{i-1} is uniform on $\Sigma - \{A_{i-1}\}$. Suppose we send T across the channel block-by-block, stopping just after the first block where a symbol comes out of the channel. Let T' denote the prefix of T that we send, and construct T'' from T' by dropping the last block (the first block in T resulting in a symbol coming out of the channel). Then the distribution of S is the same as the conditional distribution of T'' given that T' does not contain the symbol 0. It follows that $|S|$ has the same distribution as $\sum_{i=1}^J L'_i$, where J is independent of the L'_i 's. Furthermore, for any $j \geq 1$, conditioned on any particular values for the first j blocks of T' , the probability that the next block sent across the channel is not a block of 0's and is entirely deleted by the channel is $p \triangleq \mathbf{E}[d_0^L] (|\Sigma| - 2) / (|\Sigma| - 1)$. Thus, the distribution of J is stochastically dominated by the distribution $\text{Geom}(1-p)$. By Wald's equation, it follows that

$$\mathbf{E}[|S|] = \mathbf{E}[J] \mathbf{E}[L'] \leq \frac{\mathbf{E}[Ld_0^L]}{(1-p) \mathbf{E}[d_0^L]} < \infty.$$

We are now ready to commence the calculations in earnest. Define independent random variables W_0, W_1, \dots and Z_1, Z_2, \dots and N and S' so that

- $\Pr(W_0 = j) = \frac{\Pr(L=j)(1-d_0^j)}{1-\mathbf{E}[d_0^L]}$,
- Z_1 has distribution L'' ,
- W_1 has distribution L ,
- N has distribution $\text{Geom}\left(\mathbf{E}\left[1 - d_0^{L''}\right]\right) - 1$,

- S' has distribution S , and
- the (W_i, Z_i) 's are identically distributed for $i \geq 1$.

For a string $s = s_1 \dots s_k \in \Sigma^*$, define s''_1, \dots, s''_k so that $s''_i = \mathbf{1}(s_i \neq s_1)$, where $\mathbf{1}(\cdot)$ denotes the indicator function. Let $\sigma'(s) = s''_1 \dots s''_k$. For the empty string ϵ , let $\sigma'(\epsilon) = \epsilon$. For $j \geq 2$, let $G'_j = \sigma'(G_j) = \sigma'(X_j)$. It is easy to see that G'_2 is distributed as

$$0^{W_0} 1^{Z_1} 0^{W_1} 1^{Z_2} 0^{W_2} \dots 1^{Z_N} 0^{W_N} 1^{|S'|}.$$

In other words, the first block of G'_2 is a block of zeros whose length is distributed according to the conditional distribution of L given that at least one symbol in a block of length L is transmitted through the channel. The blocks of G'_2 then alternate between 1's and 0's, where a block length of 0's is just determined from L , and a block length of 1's represents all of the symbols that the channel deletes between subsequent occurrences of the symbol A_2 in X_2 . This alternation between blocks of 0's and blocks of 1's continues until the 0 in G_2 corresponding to the last occurrence of A_2 in G_2 . Then there is one last block of 1's, possibly empty, corresponding to all of the symbols deleted between this last occurrence of A_2 and the next block of X that is not entirely deleted by the channel.

Using Wald's equation, we can compute

$$\begin{aligned} \mathbf{E}[|G|] &= \mathbf{E}[|G'_2|] \\ &= \mathbf{E}[W_0] + \mathbf{E}[|S|] + \mathbf{E}\left[\sum_{i=1}^N (Z_i + W_i)\right] \\ &= \mathbf{E}[W_0] + \mathbf{E}[|S|] + \mathbf{E}[N] (\mathbf{E}[L''] + \mathbf{E}[L]) \\ &= \frac{\mathbf{E}[L(1 - d_0^L)]}{1 - \mathbf{E}[d_0^L]} + \mathbf{E}[|S|] \\ &\quad + \frac{\mathbf{E}[d_0^{L''}]}{1 - \mathbf{E}[d_0^{L''}]} \left((|\Sigma| - 1) \frac{\mathbf{E}[Ld_0^L]}{\mathbf{E}[d_0^L]} + \mathbf{E}[L] \right) \\ &< \infty. \end{aligned}$$

Next, we write

$$\begin{aligned} H(G'_2) &= H(|S'|, W_0, Z_1, W_1, \dots, W_N, Z_N, N) \\ &= H(|S'|) + H(W_0) + H(N) \\ &\quad + H(Z_1, W_1, \dots, W_N, Z_N | N). \end{aligned}$$

Now,

$$\begin{aligned} &H(Z_1, W_1, \dots, W_N, Z_N | N) \\ &= \mathbf{E}\left[\mathbf{E}[-\log \Pr(Z_1, W_1, \dots, W_N, Z_N | N)]\right] \end{aligned}$$

and for $j \geq 0$,

$$\begin{aligned} &\mathbf{E}[-\log \Pr(Z_1, W_1, \dots, W_N, Z_N | N = j)] \\ &= H(Z_1, W_1, \dots, W_j, Z_j | N = j) \\ &= j(H(Z_1) + H(W_1)), \end{aligned}$$

and so

$$\begin{aligned} H(Z_1, W_1, \dots, W_N, Z_N | N) &= \mathbf{E}[N](H(Z_1) + H(W_1)) \\ &= \mathbf{E}[N](H(L'') + H(L)). \end{aligned}$$

Similarly,

$$H(|S|) \leq H(J, L'_1, \dots, L'_J) = \mathbf{E}[J]H(L') \quad \text{and} \\ H(L'') \leq (|\Sigma| - 1)H(L').$$

Therefore,

$$H(G'_2) \leq \mathbf{E}[J]H(L') + H(W_0) + H(N) \\ + \mathbf{E}[N]((|\Sigma| - 1)H(L') + H(L)).$$

Now let D_1, D_2, \dots denote the samples of D used by the channel in transmitting X . Then

$$H(W_0), H(L') \\ \leq H(L_1, D_1, \dots, D_{L_1}) \\ = H(L) + H(D_1, \dots, D_{L_1} | L_1) \\ = H(L) + \mathbf{E}[L]H(D) < \infty,$$

because W_0 and L' are each distributed as the conditional distribution of something determined by the process of sending L zeros through the channel, given that the process satisfies some condition; specifically, that at least one 0 comes out of the channel (for W_0), or that all symbols are deleted (for L'). Also, it is easy to see that $H(N), \mathbf{E}[N], \mathbf{E}[J] < \infty$, and therefore $H(G'_2) < \infty$. Thus,

$$H(G) = H(G_2) \\ = H(G'_2, G_2) \\ = H(G'_2) + H(G_2 | G'_2) \\ = H(G'_2) + \mathbf{E}[\mathbf{E}[-\log \Pr(G_2 | G'_2) | G'_2]] \\ \leq H(G'_2) + \mathbf{E}[|G'_2| \log |\Sigma|] < \infty,$$

where we have used the fact that the conditional distribution of G_2 given G'_2 has support at most $|\Sigma|^{|G'_2|}$, and therefore has entropy at most $|G'_2| \log |\Sigma|$.

Now let $R = (D_{|G_1|+1}, \dots, D_{|G_1|+|G_2|})$. Then

$$H(\mathcal{L} | G) = H(\mathcal{L}_2 | G_2) \\ \leq H(R | G_2) \\ = \mathbf{E}[\mathbf{E}[-\log \Pr(R | G_2) | G_2]],$$

where the inequality follows from the fact that R and G_2 determine \mathcal{L}_2 . For any g in the support of G ,

$$\mathbf{E}[-\log \Pr(R | G_2 = g) | G_2 = g] \\ = H(R | G_2 = g) \leq |g|H(D),$$

where we have used the fact that the conditional distribution of R given $G_2 = g$ can be thought of as the conditional distribution of $|g|$ independent samples from D given that they satisfy some condition; since conditioning does not increase entropy, this distribution has entropy at most $|g|H(D)$. Thus, $H(\mathcal{L} | G) \leq \mathbf{E}[|G|]H(D) < \infty$, and so

$$H(G, \mathcal{L}) = H(G) + H(\mathcal{L} | G) < \infty.$$

We have now shown that $\mathbf{E}[|G|], H(G, \mathcal{L}) < \infty$.

The analysis of G_1 is similar. Let M denote the number of blocks that are completely transmitted across the channel

before the first symbol is received. Then X_1 has the same distribution as

$$A_1^{L_1} A_2^{L_2} \dots A_M^{L_M} \sigma(G'', (-A_{M+1}) \bmod |\Sigma|)$$

where G'' is a random variable that is the empty string if $A_{M+1} \neq A_1$ and whose conditional distribution given $A_{M+1} = A_1$ and L_1, \dots, L_M is the same as G . It is clear that $M+1$ has distribution $\text{Geom}(1 - \mathbf{E}[d_0^L])$ and that $M+1$ is a stopping time for L_1, L_2, \dots . Thus, Wald's equation gives

$$\mathbf{E}[|G_1|] = \mathbf{E}[|X_1|] \\ \leq \mathbf{E}[|G|] + \mathbf{E}\left[\sum_{i=1}^{M+1} L_i\right] \\ = \mathbf{E}[|G|] + \mathbf{E}[M+1] \mathbf{E}[L] < \infty.$$

Furthermore,

$$H(G_1) \leq H(L_1, A_1, \dots, L_M, A_M, G'', A_{M+1}) \\ \leq H(M) + H(G'' | A_M, A_{M+1}) \\ + H(L_1, \dots, L_M | M) \\ + H(A_1, \dots, A_{M+1} | M),$$

and the conditional distribution of L_1, \dots, L_M given M is the same as M i.i.d. random variables whose common distribution L' is the conditional distribution of L given that sending L symbols across the channel does not result in any received symbols. Thus, $H(L_1, \dots, L_M | M) \leq \mathbf{E}[M]H(L') < \infty$ (we saw that $H(L') < \infty$ earlier in the proof). Also,

$$H(A_1, \dots, A_{M+1} | M) \\ = \mathbf{E}[\mathbf{E}[-\log \Pr(A_1, \dots, A_{M+1} | M) | M]] \\ \leq \mathbf{E}[(M+1) \log |\Sigma|] < \infty,$$

and so $H(G_1) < \infty$. It is easy to see that $H(M) < \infty$, and we have

$$H(G'' | A_M, A_{M+1}) \leq H(G) < \infty$$

since we can think of G'' as being determined by a sample from G and the random variable $\mathbf{1}(A_M = A_{M+1})$. Finally, using the same technique as for $H(\mathcal{L} | G)$ yields $H(\mathcal{L}_1 | G_1) \leq \mathbf{E}[|G_1|]H(D) < \infty$, implying that

$$H(G_1, \mathcal{L}_1) = H(G_1) + H(\mathcal{L}_1 | G_1) < \infty,$$

which completes the proof. \blacksquare

III. THE MAIN RESULT

We are now ready to state the main result of this work. Let $X(n)$ denote the first n symbols of X , and let $Y(n)$ denote the prefix of Y that is received if $X(n)$ is sent across the channel. Then we have the following theorem, whose proof is subject of this section.

Theorem 1:

$$\lim_{n \rightarrow \infty} \frac{I(X(n); Y(n))}{n} \\ = \frac{H(L) + \log(|\Sigma| - 1)}{\mathbf{E}[L]} - \frac{H(G | \mathcal{L})}{\mathbf{E}[|G|]} + \frac{h}{\mathbf{E}[|G|]}$$

where $h = \lim_{m \rightarrow \infty} h_m$ for

$$h_m = H(|X_2| | X_2 \cdots X_m, Y_2, \dots, Y_m)$$

for $m \geq 2$. Furthermore, the sequence $\{h_m\}_{m \geq 2}$ is non-decreasing and bounded.

The term $h/\mathbf{E}[|G|]$ in Theorem 1 is the (theoretical) improvement of our bounds over the bounds in [4]. Furthermore, as noted in the introduction, the expression in Theorem 1 exactly achieves the best possible bounds obtainable in the model of that work. Also, since h is the limit of a non-decreasing sequence, we can lower bound it simply by lower bounding any term in the sequence. (Indeed, $h_2 = 0$, corresponding to the lower bound in [4].) We use this observation in our simulations in Section IV.

The rest of this section is devoted to the proof of Theorem 1. Let $Y''(n)$ denote the last block of $Y(n)$ and write

$$\begin{aligned} Y(n) &= Y_1 \cdots Y_{M(n)} Y''(n) \\ X(n) &= X_1 \cdots X_{M(n)} X''(n) \\ K(n) &= (|X_1|, \dots, |X_{M(n)}|). \end{aligned}$$

Therefore $Y''(n) = Y_{M(n)+1}$ and $X''(n) = X_{M(n)+1}$, but we devote special notation to them because they will play an important role in our subsequent proofs.

For brevity, we define

$$\begin{aligned} X'(n) &= X_1 \cdots X_{M(n)} \\ Y'(n) &= Y_1 \cdots Y_{M(n)}. \end{aligned}$$

Also, we let $X'''(n)$ denote the last block of $X(n)$ (which is also the last block of $X''(n)$) and write

$$X(n) = A_1^{L_1} \cdots A_{B(n)}^{L_{B(n)}} X'''(n).$$

We now express $I(X(n); Y(n))$ as follows.

$$\begin{aligned} I(X(n); Y(n)) &= H(X(n)) - H(X(n) | Y(n)) \\ &= H(X(n)) - H(X(n), K(n) | Y(n)) \\ &\quad + H(K(n) | X(n), Y(n)). \end{aligned} \quad (1)$$

We prove Theorem 1 by analyzing each of the terms in (1) separately. The full proof is somewhat technical, so we start by giving heuristic arguments for each of the terms in (1).

The first term is fairly easy to think about (and formally analyze). Each block length of X has entropy $H(L)$, and is, on average, $\mathbf{E}[L]$ symbols long. Furthermore, for each $j > 1$ the conditional distribution of the symbol A_j used in the j th block of X given A_1, \dots, A_{j-1} is uniform over a set of size $|\Sigma| - 1$. Therefore, we have

$$\lim_{n \rightarrow \infty} \frac{H(X(n))}{n} = \frac{H(L) + \log(|\Sigma| - 1)}{\mathbf{E}[L]}.$$

The second term is slightly more challenging. As $n \rightarrow \infty$, we expect $(X(n), K(n))$ to behave like $(X_2 X_3 \cdots, |X_2|, |X_3|, \dots)$, which is essentially the same as (G_2, G_3, \dots) , up to the A_j 's. (We neglect X_1 since it is special and should not have any effect on the asymptotics.) Similarly, as $n \rightarrow \infty$, we expect $Y(n)$ to behave like $(\mathcal{L}_2, \mathcal{L}_3, \dots)$, up to the A_j 's. Now, since the (G_j, \mathcal{L}_j) 's are independent and

have common distribution (G, \mathcal{L}) for $j \geq 2$, the conditional entropy of $(X_2, X_3, \dots, X_{m+1})$ given $(Y_2, Y_3, \dots, Y_{m+1})$ is $mH(G | \mathcal{L})$ for any $m \geq 1$. Since the average length of each X_j is $\mathbf{E}[|G|]$ for $j \geq 2$, we have that, conditioned on Y , we accumulate roughly $H(G | \mathcal{L})$ bits of conditional entropy for every $\mathbf{E}[|G|]$ symbols of X . Thus,

$$\lim_{n \rightarrow \infty} \frac{H(X(n), K(n) | Y(n))}{n} = \frac{H(G | \mathcal{L})}{\mathbf{E}[|G|]}.$$

The third term is the most challenging. As before, we neglect (X_1, Y_1) because it is special and should have no effect on the asymptotics. Now, for very large n , we have that $M(n) \approx n/\mathbf{E}[|G|] \triangleq m(n)$, since the X_j 's have average length $\mathbf{E}[|G|]$. Then, letting $m_i(n) = m(n) - (i-2)$, we obtain

$$\begin{aligned} &\frac{H(K(n) | X(n), Y(n))}{n} \\ &\approx \frac{H(|X_2|, \dots, |X_{m(n)}| | X_2 \cdots X_{m(n)}, Y_2, \dots, Y_{m(n)})}{n} \\ &= \frac{1}{n} \sum_{i=2}^{m(n)} H(|X_i| | X_i \cdots X_{m(n)}, Y_i, \dots, Y_{m(n)}) \\ &= \frac{1}{n} \sum_{i=2}^{m(n)} H(|X_2| | X_2 \cdots X_{m_i(n)}, Y_2, \dots, Y_{m_i(n)}) \\ &= \frac{1}{n} \sum_{i=2}^{m(n)} h_{m_i(n)}, \end{aligned} \quad (2)$$

where we have used the chain rule for entropy and the fact that the $(G_j, \mathcal{L}_j, A'_j)$'s are independent and identically distributed for $j \geq 2$. Now we examine the terms in (2).

Lemma 2: The sequence $\{h_m\}_{m \geq 2}$ is non-decreasing and bounded.

Proof: By Lemma 1, each $h_m \leq H(|X_j|) < \infty$. Thus, it suffices to show that the sequence is non-decreasing. Indeed, for any $m \geq 2$,

$$\begin{aligned} h_m &= H(|X_2| | X_2 \cdots X_m, Y_1, \dots, Y_m) \\ &= H(|X_2| | X_2 \cdots X_m, Y_2, \dots, Y_m, (G_{m+1}, \mathcal{L}_{m+1}, A'_{m+1})) \\ &\leq H(|X_2| | X_2 \cdots X_{m+1}, Y_2, \dots, Y_{m+1}) = h_{m+1}, \end{aligned}$$

where the second step follows from the fact that the $(G_i, \mathcal{L}_i, A'_i)$'s are independent, and the third step follows from the fact that

$$(X_2 \cdots X_m, Y_1, \dots, Y_m, (G_{m+1}, \mathcal{L}_{m+1}, A'_{m+1}))$$

determines $(X_2 \cdots X_{m+1}, Y_1, \dots, Y_{m+1})$. \blacksquare

Returning to (2), Lemma 2 tells us that each term of the sum converges to h , and assuming that the convergence occurs sufficiently quickly, then we should obtain $\frac{1}{n} H(K(n) | X(n), Y(n)) \rightarrow h$ as $n \rightarrow \infty$. Indeed, by Lemma 2, we have that for any n ,

$$\frac{1}{n} \sum_{i=2}^{m(n)} h_{m_i(n)} \leq \frac{1}{n} \sum_{i=2}^{m(n)} h \rightarrow \frac{h}{\mathbf{E}[|G|]}.$$

For the corresponding lower bound, we fix any $\epsilon > 0$, let $a(n, \epsilon) = \lceil \epsilon m(n) \rceil$ and write

$$\begin{aligned} \frac{1}{n} \sum_{i=2}^{m(n)} h_{m_i(n)} &\geq \frac{1}{n} \sum_{i=a(n, \epsilon)}^{m(n)} h_{m_i(n)} \\ &\geq \frac{1}{n} \sum_{i=a(n, \epsilon)}^{m(n)} h_{a(n, \epsilon)} \\ &\rightarrow (1 - \epsilon) \frac{h}{\mathbf{E}[|G|]}, \end{aligned}$$

where we have used Lemma 2. Taking the limit as $\epsilon \rightarrow 0$ then completes the analysis of the third term in (1).

We now give the formal proof of Theorem 1. First, we formalize the claim that $M(n) \sim n / \mathbf{E}[|G|]$ and $B(n) \sim n / \mathbf{E}[L]$ using basic arguments from renewal theory.

Lemma 3: $M(n)/n \rightarrow 1 / \mathbf{E}[|G|]$ and $B(n)/n \rightarrow 1 / \mathbf{E}[L]$ almost surely as $n \rightarrow \infty$.

Proof: Let $S_0 = 0$ and let $S_m = \sum_{j=1}^m |G_j|$ for $m \geq 1$. For $n \geq 1$, let $N(n) = \sup \{m : S_m \leq n\}$. By Lemma 1, the $|G_i|$'s are independent and $|G_2|, |G_3|, \dots$ have common distribution $|G|$. Therefore, $N(n)$ is a delayed renewal process. Since $N(n) - 1 \leq M(n) \leq N(n)$, we may apply a standard result from renewal theory (e.g. [7, Proposition 3.14]) to $N(n)$ to obtain $N(n)/n \rightarrow 1 / \mathbf{E}[|G|]$ almost surely as $n \rightarrow \infty$. Thus, $M(n)/n \rightarrow 1 / \mathbf{E}[|G|]$ almost surely. The proof that $B(n)/n \rightarrow 1 / \mathbf{E}[L]$ almost surely is entirely similar. ■

We can now sketch the formal asymptotic analyses of the first two terms in (1). The analyses rely heavily on the notion of uniform integrability, and so we encourage readers who are unfamiliar with that concept to read the Appendix before continuing for the definition, standard results, and intuition that is extremely useful in understanding our proofs.

We define the following notation for jointly distributed discrete random variables (W, Z) . We let $\text{Supp}(W)$ denote the support of W . For any $z \in \text{Supp}(Z)$, we let $\text{Supp}(W | Z = z)$ denote the support of the conditional distribution of W given $Z = z$. Finally, we let $\text{Supp}(W | Z)$ denote the random variable whose value is the support of the conditional distribution of W given Z .

We start by stating two general technical lemmas.

Lemma 4: Let $\{(W_n, Z_n) : n \geq 1\}$ be any family of discrete random variables with $|\text{Supp}(W_n | Z_n)| \leq c^n$ for some constant $c \geq 1$. Then $\sup_n \mathbf{E} \left[\left(\frac{1}{n} \log \Pr(W_n | Z_n) \right)^2 \right] < \infty$. In particular, the family $\left\{ -\frac{1}{n} \log \Pr(W_n | Z_n) : n \geq 1 \right\}$ is uniformly integrable.

Proof: We write

$$\begin{aligned} &\mathbf{E} \left[\left(-\log \Pr(W_n | Z_n) \right)^2 \right] \\ &\leq \mathbf{E} \left[\left(\log \left(\frac{1}{\Pr(W_n | Z_n)} + e \right) \right)^2 \right] \\ &\leq \left(\log \left(\mathbf{E} \left[\frac{1}{\Pr(W_n | Z_n)} \right] + e \right) \right)^2 \\ &= \left(\log \left(\mathbf{E}[|\text{Supp}(W_n | Z_n)|] + e \right) \right)^2 \\ &\leq \left(\log(c^n + e) \right)^2 \end{aligned}$$

$$= O(n^2) \quad \text{as } n \rightarrow \infty,$$

where we have used Jensen's inequality and the concavity of $f(x) = (\log(x + e))^2$ on $[0, \infty)$ (which is easy to check by examining the second derivative of f). It follows that

$$\sup_n \mathbf{E} \left[\left(\frac{1}{n} \log \Pr(W_n | Z_n) \right)^2 \right] < \infty,$$

and therefore the $-\frac{1}{n} \log \Pr(W_n | Z_n)$'s are uniformly integrable (by, e.g. [5, Example 7.10.5]). ■

Lemma 5: Let $\{Z_n\}_{n \geq 1}$ be a sequence of independent discrete random variables where $H(Z_n) < \infty$ for all n and, for sufficiently large n , the Z_n 's have some common distribution Z . Let $\{N_n\}_{n \geq 1}$ be a sequence of positive integral random variables defined on the same probability space as the Z_n 's, and suppose that $\lim_{n \rightarrow \infty} \frac{1}{n} N_n = x$ almost surely for some constant x . Then, almost surely,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr(Z_1, \dots, Z_{N_n}) = H(Z)x.$$

Proof: Fix some $k \in \mathbb{Z}^+$. Since $\lim_{n \rightarrow \infty} N_n = x$ almost surely, we have

$$N_n \geq \left\lfloor n \left(x - \frac{1}{k} \right) \right\rfloor \triangleq a(n, k)$$

for sufficiently large n , almost surely. It follows that, almost surely, for sufficiently large n ,

$$\begin{aligned} &-\frac{1}{n} \log \Pr(Z_1, \dots, Z_{N_n}) \\ &\geq -\frac{1}{n} \log \Pr(Z_1, \dots, Z_{a(n, k)}) \\ &= \frac{1}{n} \sum_{i=1}^{a(n, k)} -\log \Pr(Z_i) \\ &= \frac{a(n, k)}{n} \cdot \frac{1}{a(n, k)} \sum_{i=1}^{a(n, k)} -\log \Pr(Z_i) \\ &\rightarrow \left(x - \frac{1}{k} \right) \mathbf{E}[-\log \Pr(Z)] \\ &= H(Z)x - \frac{H(Z)}{k} \end{aligned} \tag{3}$$

by the strong law of large numbers. A union bound over all $k \in \mathbb{Z}^+$ now tells us that (3) holds almost surely for all k simultaneously, and therefore we may take the limit as $k \rightarrow \infty$ to obtain

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log \Pr(Z_1, \dots, Z_{N_n}) \geq H(Z)x$$

almost surely. The proof of the corresponding upper bound is entirely similar. ■

We are now ready to analyze the first term in (1).

Lemma 6:

$$\lim_{n \rightarrow \infty} \frac{H(X(n))}{n} = \frac{H(L) + \log(|\Sigma| - 1)}{\mathbf{E}[L]}$$

Proof: We start by noting that $\frac{1}{n} H(X(n)) = \mathbf{E} \left[-\frac{1}{n} \log \Pr(X(n)) \right]$ and that $|\text{Supp}(X(n))| \leq |\Sigma|^n$. Thus, by Lemma 4 and Theorem 2 (in Appendix), it suffices to

show that $-\frac{1}{n} \log \Pr(X(n)) \rightarrow H(L)/\mathbf{E}[L]$ almost surely as $n \rightarrow \infty$. To this end, we write

$$\begin{aligned} & \Pr(X(n)) \\ &= \Pr(A_1, \dots, A_{B(n)+1}, L_1, \dots, L_{B(n)}, |X'''(n)|) \\ &= \Pr(L_1, \dots, L_{B(n)}, |X'''(n)|) \left(\frac{1}{|\Sigma| - 1} \right)^{B(n)}, \end{aligned}$$

where we have used the fact that $A_1 = 0$ and that for $i \geq 2$, the conditional distribution of A_i given A_1, \dots, A_{i-1} is uniform on $\Sigma - \{A_{i-1}\}$. Next, we write

$$\begin{aligned} & \Pr(L_1, \dots, L_{B(n)+1}) \\ & \leq \Pr(L_1, \dots, L_{B(n)}, |X'''(n)|) \\ & \leq \Pr(L_1, \dots, L_{B(n)}). \end{aligned}$$

By Lemma 3, we have $B(n)/n \rightarrow 1/\mathbf{E}[L]$ (and $(B(n) + 1)/n \rightarrow 1/\mathbf{E}[L]$) almost surely, and therefore Lemma 5 implies that

$$\begin{aligned} & \lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr(L_1, \dots, L_{B(n)+1}) \\ &= \lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr(L_1, \dots, L_{B(n)}) \\ &= \frac{H(L)}{\mathbf{E}[L]}. \end{aligned}$$

It follows that

$$-\frac{1}{n} \log \Pr(X(n)) \rightarrow \frac{H(L) + \log(|\Sigma| - 1)}{\mathbf{E}[L]}$$

almost surely as $n \rightarrow \infty$, completing the proof. \blacksquare

The analysis of the second term in (1) is essentially the same as for the first term, but slightly more technical.

Lemma 7:

$$\lim_{n \rightarrow \infty} \frac{1}{n} H(X(n), K(n) | Y(n)) = \frac{H(G | \mathcal{L})}{\mathbf{E}[|G|]}.$$

Proof: We write

$$\begin{aligned} & H(X(n), K(n) | Y(n)) \\ &= \mathbf{E}[-\log \Pr(X(n), K(n) | Y(n))]. \end{aligned}$$

Note that $X(n) \in \Sigma^n$ and that $K(n)$ can be represented as a (possibly empty) binary string of the form

$$10^{|X_1|-1} 10^{|X_2|-1} \dots 10^{|X_{M(n)}|-1},$$

which has length $|X'(n)| \leq n - 1$. Thus, $|\text{Supp}(X(n), K(n))| \leq (2|\Sigma|)^n$, and so Lemma 4 implies that the $-\frac{1}{n} \log \Pr(X(n), K(n) | Y(n))$'s are uniformly integrable. By Theorem 2 (in Appendix), it now suffices to show that, almost surely,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr(X(n), K(n) | Y(n)) = \frac{H(G | \mathcal{L})}{\mathbf{E}[|G|]}.$$

We do this by showing that, almost surely

$$\begin{aligned} & \lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr(X(n), K(n), Y(n)) \\ &= \frac{H(G, \mathcal{L}) + \log(|\Sigma| - 1)}{\mathbf{E}[|G|]} \end{aligned} \quad (4)$$

$$\text{and } \lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr(Y(n)) = \frac{H(\mathcal{L}) + \log(|\Sigma| - 1)}{\mathbf{E}[|G|]}. \quad (5)$$

We start with (4), writing

$$\begin{aligned} & \Pr(X(n), K(n), Y(n)) \\ &= \Pr((G_1, \mathcal{L}_1, \mathcal{A}_1) \dots, (G_{M(n)}, \mathcal{L}_{M(n)}, \mathcal{A}_{M(n)}), \\ & \quad (X''(n), Y''(n), \mathcal{A}_{M(n)+1})) \\ &= \Pr((G_1, \mathcal{L}_1, \mathcal{A}'_1) \dots, (G_{M(n)}, \mathcal{L}_{M(n)}, \mathcal{A}'_{M(n)}), \\ & \quad (X''(n), Y''(n), \mathcal{A}'_{M(n)+1})). \end{aligned}$$

Therefore,

$$\begin{aligned} & \Pr((G_1, \mathcal{L}_1, \mathcal{A}'_1), \dots, (G_{M(n)+1}, \mathcal{L}_{M(n)+1}, \mathcal{A}'_{M(n)+1})) \\ & \leq \Pr(X(n), K(n), Y(n)) \\ & \leq \Pr((G_1, \mathcal{L}_1, \mathcal{A}'_1), \dots, (G_{M(n)}, \mathcal{L}_{M(n)}, \mathcal{A}'_{M(n)})). \end{aligned}$$

By Lemma 1, the $(G_i, \mathcal{L}_i, \mathcal{A}'_i)$'s are independent and, for $i \geq 2$, they have common distribution $(G, \mathcal{L}, \mathcal{A}')$ with

$$H(G, \mathcal{L}, \mathcal{A}') = H(G, \mathcal{L}) + \log(|\Sigma| - 1) < \infty.$$

Since $M(n)/n \rightarrow 1/\mathbf{E}[|G|]$ (and $(M(n)+1)/n \rightarrow 1/\mathbf{E}[|G|]$) almost surely as $n \rightarrow \infty$ (by Lemma 3), Lemma 5 tells us that, almost surely,

$$\begin{aligned} & \lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr((G_1, \mathcal{L}_1, \mathcal{A}'_1), \dots, \\ & \quad (G_{M(n)+1}, \mathcal{L}_{M(n)+1}, \mathcal{A}'_{M(n)+1})) \\ &= \lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pr((G_1, \mathcal{L}_1, \mathcal{A}'_1), \dots, \\ & \quad (G_{M(n)}, \mathcal{L}_{M(n)}, \mathcal{A}'_{M(n)})) \\ &= \frac{H(G, \mathcal{L}) + \log(|\Sigma| - 1)}{\mathbf{E}[|G|]}. \end{aligned}$$

Thus, we have established (4).

The proof of (5) is similar. We write

$$\Pr(Y(n)) = \Pr(\mathcal{L}_1, \mathcal{A}'_1, \dots, \mathcal{L}_{M(n)}, \mathcal{A}'_{M(n)}, Y''(n)),$$

so that

$$\begin{aligned} & \Pr(\mathcal{L}_1, \mathcal{A}'_1, \dots, \mathcal{L}_{M(n)+1}, \mathcal{A}'_{M(n)+1}) \\ & \leq \Pr(Y(n)) \\ & \leq \Pr(\mathcal{L}_1, \mathcal{A}'_1, \dots, \mathcal{L}_{M(n)}, \mathcal{A}'_{M(n)}). \end{aligned}$$

Lemmas 1, 3, and 5 now yield (5), completing the proof. \blacksquare

The formal analysis of the third term in (1) is fairly technical. The added difficulty is due primarily to the fact that we cannot use the same sort of almost-sure convergence trick as for the first two terms, because the relevant $\Pr(\cdot)$ random variables do not factor into independent identically distributed random variables. However, we can still write $\frac{1}{n} H(K(n) | X(n), Y(n)) = \mathbf{E}[Z_n]$ for $Z_n = -\frac{1}{n} \log \Pr(K(n) | X(n), Y(n))$. The Z_n 's are easily shown to be uniformly integrable using Lemma 4. Recalling that $m(n) \triangleq n/\mathbf{E}[|G|]$ and $\Pr(|M(n) - m(n)| > \epsilon n) \rightarrow 0$ as $n \rightarrow \infty$ for any $\epsilon > 0$, we have (heuristically)

$$\mathbf{E}[Z_n] \approx \mathbf{E}[Z_n | M(n) \approx m(n)] \approx \mathbf{E}[Z_n | M(n) = m(n)].$$

This argument is essentially the justification for (2), and from there, our previous analysis is sufficient to complete the proof of Theorem 1.

In reality, we do not prove (2) directly, but instead perform a direct analysis of the third term in (1) using the previous arguments as intuition. We now give the formal proof as a sequence of lemmas.

Lemma 8:

$$\begin{aligned} & |H(K(n) \mid X(n), Y(n)) - H(K(n) \mid X'(n), Y'(n))| \\ & \leq H(X''(n)) \end{aligned}$$

Proof: For the upper bound, we write

$$\begin{aligned} & H(K(n) \mid X(n), Y(n)) \\ & \leq H(K(n), X''(n) \mid X(n), Y(n)) \\ & = H(K(n) \mid X''(n), X(n), Y(n)) \\ & \quad + H(X''(n) \mid X(n), Y(n)) \\ & \leq H(K(n) \mid X''(n), X(n), Y(n)) + H(X''(n)) \\ & = H(K(n) \mid X'(n), X''(n), Y'(n), Y''(n)) + H(X''(n)) \\ & \leq H(K(n) \mid X'(n), Y'(n)) + H(X''(n)). \end{aligned}$$

For the lower bound, we have

$$\begin{aligned} & H(K(n) \mid X(n), Y(n)) \\ & \geq H(K(n) \mid X(n), Y(n), X''(n)) \\ & = H(K(n) \mid X'(n), X''(n), Y'(n), Y''(n)) \\ & = H(K(n) \mid X'(n), Y'(n)) \\ & \quad - I(K(n); (X''(n), Y''(n)) \mid X'(n), Y'(n)) \\ & = H(K(n) \mid X'(n), Y'(n)) \end{aligned}$$

where the fourth step is the only nontrivial one, and it follows from the fact that

$$K(n) \rightarrow (X'(n), Y'(n)) \rightarrow (X''(n), Y''(n))$$

is a Markov chain; this is evident from Figure 1. ■

Lemma 9: Consider some random variable $W \in \Sigma^*$ and any other discrete random variable Z defined on the same probability space. Then

$$H(W \mid Z) \leq H(|W| \mid Z) + \mathbf{E}[|W|] \log |\Sigma|.$$

Proof: First we write

$$\begin{aligned} H(W \mid Z) & = H(W, |W| \mid Z) \\ & = H(|W| \mid Z) + H(W \mid |W|, Z) \\ & \leq H(|W| \mid Z) + H(W \mid |W|) \end{aligned}$$

Next, we observe that

$$\begin{aligned} & H(W \mid |W|) \\ & = \mathbf{E}[-\log \Pr(W \mid |W|)] \\ & = \mathbf{E}[\mathbf{E}[-\log \Pr(W \mid |W|) \mid |W|]]. \end{aligned}$$

Finally, we have

$$\mathbf{E}[-\log \Pr(W \mid |W|) \mid |W|] \leq |W|,$$

since for any $\ell \geq 0$

$$\begin{aligned} & \mathbf{E}[-\log \Pr(W \mid |W|) \mid |W| = \ell] \\ & = \mathbf{E}[-\log \Pr(W \mid |W| = \ell) \mid |W| = \ell], \end{aligned}$$

which is the entropy of the conditional distribution of W given that $|W| = \ell$, and the entropy of any distribution with support contained in Σ^ℓ is at most $\ell \log |\Sigma|$. ■

Lemma 10: $\lim_{n \rightarrow \infty} \frac{1}{n} H(X''(n)) = 0$

Proof: We write

$$\begin{aligned} H(X''(n)) & \leq H(|X''(n)|) + \mathbf{E}[|X''(n)|] \\ & \leq \log n + \mathbf{E}[|X''(n)|] \log |\Sigma|, \end{aligned}$$

where the first step follows from Lemma 9, and the second step follows from the fact that the support of $|X''(n)|$ is contained in $\{1, \dots, n\}$. Since $(\log n)/n \rightarrow 0$ as $n \rightarrow \infty$, it suffices to show that $\mathbf{E}[|X''(n)|]/n \rightarrow 0$ as $n \rightarrow \infty$. To this end, we note that $|X''(n)|/n \leq 1$, and therefore it is enough to show that $|X''(n)|/n \rightarrow 0$ as $n \rightarrow \infty$ almost surely, by the dominated convergence theorem. Indeed, $|X''(n)| \leq |X_{M(n)+1}|$, since $X''(n)$ is a prefix of $X_{M(n)+1}$. Therefore, almost surely,

$$\begin{aligned} \frac{|X''(n)|}{n} & \leq \frac{|X_{M(n)+1}|}{n} \\ & = \frac{1}{n} \sum_{j=1}^{M(n)+1} |X_j| - \frac{1}{n} \sum_{i=1}^{M(n)} |X_j| \\ & = \frac{M(n)+1}{n} \cdot \frac{1}{M(n)+1} \sum_{j=1}^{M(n)+1} |X_j| \\ & \quad - \frac{M(n)}{n} \cdot \frac{1}{M(n)} \sum_{i=1}^{M(n)} |X_j| \\ & \rightarrow \frac{1}{\mathbf{E}[|G|]} \cdot \mathbf{E}[|G|] - \frac{1}{\mathbf{E}[|G|]} \cdot \mathbf{E}[|G|] \\ & = 0, \end{aligned}$$

by Lemma 3 and the strong law of large numbers. ■

Lemma 11: Let Z be a nonnegative random variable, and let \mathcal{V} be an event with $\Pr(\mathcal{V}) > 0$. Then $\mathbf{E}[Z \mid \mathcal{V}] \leq \mathbf{E}[Z]/\Pr(\mathcal{V})$.

Proof: Since Z is nonnegative

$$\begin{aligned} \mathbf{E}[Z] & = \Pr(\neg \mathcal{V}) \mathbf{E}[Z \mid \neg \mathcal{V}] + \Pr(\mathcal{V}) \mathbf{E}[Z \mid \mathcal{V}] \\ & \geq \Pr(\mathcal{V}) \mathbf{E}[Z \mid \mathcal{V}]. \end{aligned}$$

Lemma 12:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} H(K(n) \mid X'(n), Y'(n)) \leq \frac{h}{\mathbf{E}[|G|]}.$$

Proof: First, we note that

$$\begin{aligned} & \frac{H(K(n) \mid X'(n), Y'(n))}{n} \\ & = \mathbf{E} \left[\frac{-\log \Pr(K(n) \mid X'(n), Y'(n))}{n} \right]. \end{aligned} \quad (6)$$

As we saw in the proof of Lemma 7, $K(n)$ can be represented as a binary string of the form

$$10^{|X_1|-1} 10^{|X_2|-1} \dots 10^{|X_{M(n)}|-1},$$

which has length $|X'(n)| \leq n - 1$. Thus, $|\text{Supp}(K(n))| \leq 2^n$, and so Lemma 4 tells us that the $-\frac{1}{n} \log \Pr(K(n) | X'(n), Y'(n))$'s are uniformly integrable.

Now, for any $n \geq 1$ and $\epsilon > 0$, define the event

$$\mathcal{V}_n(\epsilon) = \left\{ \left| \frac{M(n)}{n} - \frac{1}{\mathbf{E}[|G|]} \right| \leq \epsilon \right\}.$$

Fix any $\epsilon_1, \epsilon_2 > 0$, and let $\mathbf{1}(\cdot)$ denote the indicator function. By Theorem 3 (in Appendix), there exists some $\delta > 0$ such that for all $n \geq 1$ and any event \mathcal{V} with $\Pr(\mathcal{V}) < \delta$,

$$\mathbf{E} \left[\frac{-\log \Pr(K(n) | X'(n), Y'(n))}{n} \cdot \mathbf{1}(\mathcal{V}) \right] \leq \epsilon_1.$$

By Lemma 3, $\lim_{n \rightarrow \infty} \Pr(\neg \mathcal{V}_n(\epsilon_2)) = 0$, and so for sufficiently large n , we have

$$\mathbf{E} \left[\frac{-\log \Pr(K(n) | X'(n), Y'(n))}{n} \cdot \mathbf{1}(\neg \mathcal{V}_n(\epsilon_2)) \right] \leq \epsilon_1.$$

It follows that for large enough n , we have

$$\begin{aligned} & \mathbf{E} \left[\frac{-\log \Pr(K(n) | X'(n), Y'(n))}{n} \right] \\ & \leq \epsilon_1 + \mathbf{E} \left[\frac{-\log \Pr(K(n) | X'(n), Y'(n))}{n} \mid \mathcal{V}_n(\epsilon_2) \right]. \end{aligned} \quad (7)$$

For convenience, define $K_j = (|X_1|, \dots, |X_j|)$ for $j \geq 1$. Also define

$$\begin{aligned} a(n, \epsilon_2) &= \max(3, \lceil n(-\epsilon_2 + 1/\mathbf{E}[|G|]) \rceil) \\ b(n, \epsilon_2) &= \lfloor n(\epsilon_2 + 1/\mathbf{E}[|G|]) \rfloor. \end{aligned}$$

For brevity, let $\mathcal{V} = \mathcal{V}_n(\epsilon_2)$. Then

$$\begin{aligned} & \mathbf{E} \left[\frac{-\log \Pr(K(n) | X'(n), Y'(n))}{n} \mid \mathcal{V} \right] \\ &= \mathbf{E} \left[\frac{1}{n} \sum_{j=1}^{M(n)} -\log \Pr(|X_j| | X'(n), Y'(n), K_{j-1}) \mid \mathcal{V} \right] \\ &\leq \mathbf{E} \left[\frac{1}{n} \sum_{j=1}^{b(n, \epsilon_2)} -\log \Pr(|X_j| | X'(n), Y'(n), K_{j-1}) \mid \mathcal{V} \right] \\ &= \frac{1}{n} \sum_{j=1}^{b(n, \epsilon_2)} \mathbf{E} \left[-\log \Pr(|X_j| | X'(n), Y'(n), K_{j-1}) \mid \mathcal{V} \right]. \end{aligned} \quad (8)$$

We now upper bound the terms of the sum in (8). We start with the first term. By Lemma 11,

$$\begin{aligned} & \mathbf{E} \left[-\log \Pr(|X_1| | X'(n), Y'(n)) \mid \mathcal{V} \right] \\ & \leq \frac{\mathbf{E}[-\log \Pr(|X_1| | X'(n), Y'(n))]}{\Pr(\mathcal{V})} \\ & = \frac{H(|X_1| | X'(n), Y'(n))}{\Pr(\mathcal{V})} \\ & \leq \frac{H(|X_1|)}{\Pr(\mathcal{V})}. \end{aligned}$$

We can bound the last few terms of the sum in (8) similarly:

$$\begin{aligned} & \sum_{j=a(n, \epsilon_2)+1}^{b(n, \epsilon_2)} \mathbf{E} \left[-\log \Pr(|X_j| | X'(n), Y'(n), K_{j-1}) \mid \mathcal{V} \right] \\ & \leq \sum_{j=a(n, \epsilon_2)+1}^{b(n, \epsilon_2)} \frac{\mathbf{E}[-\log \Pr(|X_j| | X'(n), Y'(n), K_{j-1})]}{\Pr(\mathcal{V})} \\ & = \sum_{j=a(n, \epsilon_2)+1}^{b(n, \epsilon_2)} \frac{H(|X_j| | X'(n), Y'(n), K_{j-1})}{\Pr(\mathcal{V})} \\ & \leq \sum_{j=a(n, \epsilon_2)+1}^{b(n, \epsilon_2)} \frac{H(|G|)}{\Pr(\mathcal{V})} \\ & \leq \frac{2n\epsilon_2 H(|G|)}{\Pr(\mathcal{V})}. \end{aligned} \quad (10)$$

Bounding the other terms in (8) requires more care. The main idea is that, conditioned on \mathcal{V} , we can modify the $\Pr(\cdot)$ random variable inside the expectation in a way that would not be valid in the original probability space. We can then remove the conditioning from the expectation through an application Lemma 11. Once the conditioning is removed from the expectation, the expectation once again becomes a conditional entropy, but a different conditional entropy than at the beginning of the proof. The new conditional entropy is easily bounded by h .

Fix some $2 \leq j \leq a(n, \epsilon_2)$. To shorten our equations, define

$$\begin{aligned} Z_{n,j} &= (X'(n), Y'(n), K_{j-1}) \\ T_i &= (X_i, Y_i) \text{ for } i \geq 1. \end{aligned}$$

Now, we have that X_j and $\{T_i : i > M(n)\}$ are conditionally independent given $Z_{n,j}$ and $j \leq M(n)$, because for any $k \geq 1$, given $j \leq M(n)$,

$$\begin{aligned} X_j &\rightarrow (X_1, \dots, X_{j-1}, X_j \cdots X_{M(n)}, Y_1, \dots, Y_{M(n)}) \\ &\rightarrow (T_{M(n)+1}, \dots, T_{M(n)+k}) \end{aligned}$$

forms a Markov chain; this is easily seen from Figure 1.

It follows that if $j \leq M(n)$, then for any $k \geq 1$,

$$\begin{aligned} & \Pr(|X_j| | X'(n), Y'(n), K_{j-1}) \\ &= \Pr(|X_j| | X'(n), Y'(n), K_{j-1}, T_{M(n)+1}, \dots, T_{M(n)+k}) \\ &= \Pr(|X_j| | Z_{n,j}, T_{M(n)+1}, \dots, T_{M(n)+k}). \end{aligned}$$

Of course, $j \leq M(n) \leq b(n, \epsilon_2)$ given \mathcal{V} , and therefore, letting $S_m = |X_1 \cdots X_m|$ for any $m \geq 1$ and $T = (T_{M(n)+1}, \dots, T_{b(n, \epsilon_2)})$, we have

$$\begin{aligned} & \mathbf{E} \left[-\log \Pr(|X_j| | X'(n), Y'(n), K_{j-1}) \mid \mathcal{V} \right] \\ &= \mathbf{E} \left[-\log \Pr(|X_j| | Z_{n,j}, T) \mid \mathcal{V} \right] \\ &= \mathbf{E} \left[-\log \Pr(|X_j| | Z_{n,j}, T, S_{b(n, \epsilon_2)}) \mid \mathcal{V} \right] \\ &\leq \frac{\mathbf{E}[-\log \Pr(|X_j| | Z_{n,j}, T, S_{b(n, \epsilon_2)})]}{\Pr(\mathcal{V})}, \end{aligned} \quad (11)$$

(9) where the third step follows from the fact that, when \mathcal{V} occurs, $S_{b(n, \epsilon_2)}$ is determined by the other random variables

on which we are conditioning, and the fourth step follows from Lemma 11.

Now,

$$\begin{aligned}
& \mathbf{E} \left[-\log \Pr(|X_j| \mid Z_{n,j}, T, S_{b(n,\epsilon_2)}) \right] \\
&= H(|X_j| \mid Z_{n,j}, T_{M(n)+1}, \dots, T_{b(n,\epsilon_2)}, S_{b(n,\epsilon_2)}) \\
&\leq H(|X_j| \mid X_j \cdots X_{b(n,\epsilon_2)}, Y_j, \dots, Y_{b(n,\epsilon_2)}) \\
&= h_{b(n,\epsilon_2)-(j-2)} \\
&\leq h.
\end{aligned} \tag{12}$$

The first step is obvious. The second step follows from the fact that

$$(Z_{n,j}, T_{M(n)+1}, \dots, T_{b(n,\epsilon_2)}, S_{b(n,\epsilon_2)})$$

determines

$$(X_j \cdots X_{b(n,\epsilon_2)}, Y_j, \dots, Y_{b(n,\epsilon_2)}).$$

The third step follows from the fact that the $(G_i, \mathcal{L}_i, \mathcal{A}'_i)$'s are identically distributed for $j \geq 2$. Finally, the fourth step follows from Lemma 2.

Combining (6)-(12) and taking limits as $n \rightarrow \infty$ now gives

$$\begin{aligned}
& \limsup_{n \rightarrow \infty} \frac{1}{n} H(K(n) \mid X'(n), Y'(n)) \\
&\leq \epsilon_1 + 2\epsilon_2 H(|G|) + \frac{h}{\mathbf{E}[|G|]},
\end{aligned}$$

and since $\epsilon_1, \epsilon_2 > 0$ are arbitrary, we can take limits as $\epsilon_1, \epsilon_2 \rightarrow 0$ to obtain the desired result. ■

For convenience, we define $K_m = (|X_1|, \dots, |X_m|)$ as in the proof of Lemma 12. Also, for $m \geq 1$, we define $X^{(m)} = X_1 \cdots X_m$ and $Y^{(m)} = (Y_1, \dots, Y_m)$ and $W_m = (X^{(m)}, Y^{(m)})$.

Lemma 13: The family

$$\left\{ -\frac{1}{m} \log \Pr(K_m \mid W_m) : m \geq 1 \right\}$$

is uniformly integrable.

Proof: We proceed using a generalization of the technique used in the proof of Lemma 4. Specifically, we show that

$$\sup_m \mathbf{E} \left[\left(-\frac{1}{m} \log \Pr(K_m \mid W_m) \right)^2 \right] < \infty,$$

by showing that as $m \rightarrow \infty$,

$$\mathbf{E} \left[(-\log \Pr(K_m \mid W_m))^2 \right] = O(m^2).$$

For $j \in \mathbb{Z}^+$, let N_j be the number of blocks of X contained in X_j . Let $N(m) = \sum_{j=1}^m N_j$ denote the total number of blocks in $X_1 \cdots X_m$. Then, given $X_1 \cdots X_m$, we can encode $(|X_1|, \dots, |X_m|)$ as a binary string of the form

$$10^{N_1-1} 10^{N_2-1} \dots 10^{N_m-1}.$$

(This string uniquely specifies $(|X_1|, \dots, |X_m|)$ when $X_1 \cdots X_m$ is known since the lengths of the blocks of X contained in $X_1 \cdots X_m$ are easily seen.) The length of the string

is clearly $N(m)$, and therefore $|\text{Supp}(K_m \mid W_m)| \leq 2^{N(m)}$. Now, as $m \rightarrow \infty$,

$$\begin{aligned}
& \mathbf{E} \left[(-\log \Pr(K_m \mid W_m))^2 \right] \\
&\leq \mathbf{E} \left[\left(\log \left(\frac{1}{\Pr(K_m \mid W_m)} + e \right) \right)^2 \right] \\
&= \mathbf{E} \left[\mathbf{E} \left[\left(\log \left(\frac{1}{\Pr(K_m \mid W_m)} + e \right) \right)^2 \mid W_m \right] \right] \\
&\leq \mathbf{E} \left[\left(\log \left(\mathbf{E} \left[\frac{1}{\Pr(K_m \mid W_m)} \mid W_m \right] + e \right) \right)^2 \right] \\
&= \mathbf{E} \left[\left(\log (\mathbf{E}[|\text{Supp}(K_m \mid W_m)|] + e) \right)^2 \right] \\
&\leq \mathbf{E} \left[\left(\log (2^{N(m)} + e) \right)^2 \right] \\
&= O(\mathbf{E}[N(m)^2]) \\
&= O \left(m \sum_{j=1}^m \mathbf{E}[N_j^2] \right) \\
&= O(m^2).
\end{aligned}$$

The first two steps are obvious. The third step follows from Jensen's inequality and the concavity of $f(x) = (\log(x+e))^2$ on $[0, \infty)$ (which is easy to check by examining the second derivative of f). The fourth, fifth, and sixth steps are obvious, and the seventh step follows from the Cauchy-Schwarz inequality. For the eighth step, looking back to the procedure in Figure 1, it is easy to see the N_j 's have a common distribution N for $j \geq 2$. Furthermore, it is also easy to see that N_1 is stochastically dominated by the sum of two independent samples from N . Thus, it suffices to show that $\mathbf{E}[N^2] < \infty$, which we do by a direct calculation.

Define random variables P_0, P_1, \dots and Z_1, Z_2, \dots and M_1, M_2, \dots and R_1, R_2, \dots and N', S' , and J such that

- $\Pr(P_0 = j) = \frac{\Pr(L=j)(1-d_0^j)}{1-\mathbf{E}[d_0^L]}$,
- P_1, P_2, \dots are i.i.d. with common distribution L ,
- $\Pr(Z_i = j) = \frac{\Pr(L=j)d_0^j}{\mathbf{E}[d_0^L]}$ for $i \geq 1$,
- M_1, M_2, \dots are i.i.d. with common distribution $\text{Geom}(1/(|\Sigma| - 1))$,
- N' has distribution $\text{Geom}(1 - \mathbf{E}[d_0^{L''}]) - 1$, where L'' is as defined in the proof of Lemma 1,
- R_1 is uniformly distributed on $\{1, \dots, |\Sigma| - 1\}$,
- S' has the distribution S from the proof of Lemma 1,
- J is the number of blocks in S' ,
- the P_i 's, Z_i 's, M_i 's, N' , S' , and R_1 are independent, and
- for $j \geq 2$, the conditional distribution of R_j given the P_i 's, Z_i 's, M_i 's, N' , S , and R_1, \dots, R_{j-1} is uniform on $\Sigma - \{0, R_{j-1}\}$ if $j \neq M_i + 1$ for any i , and uniform on $\Sigma - \{0\}$ otherwise.

It is easy to see that G has the same distribution as

$$\begin{aligned}
& 0^{P_0} R_1^{Z_1} \dots R_{M_1}^{Z_{M_1}} 0^{P_1} R_{M_1+1}^{Z_{M_1+1}} \dots R_{M_1+M_2}^{Z_{M_1+M_2}} 0^{P_2} \\
& \dots 0^{P_{N'-1}} R_{1+\sum_{i=1}^{N'-1} M_i}^{Z_{1+\sum_{i=1}^{N'-1} M_i}} \dots R_{\sum_{i=1}^{N'} M_i}^{Z_{\sum_{i=1}^{N'} M_i}} 0^{P_{N'}} S'.
\end{aligned}$$

In other (less precise) words, for any $j \geq 2$, the first block of G_j is a block of 0's whose length is distributed according to the conditional distribution of L given that at least one symbol in a block of length L is transmitted through the channel. The next blocks of G_j correspond to the blocks that are deleted between the first block and the next block of A_j 's in X_j . The number of such blocks clearly has distribution $\text{Geom}(1/(|\Sigma| - 1))$, which is the common distribution of the M_i 's. Furthermore, the symbol for each of these blocks, conditioned on all previous blocks, is uniform over all of Σ , with the exception of 0 and the symbol used for the previous block. Now, once we reach the next block of 0's in G_j , the length of that block simply has distribution L , and then the distribution of the portion of G_j between this block of 0's and the next block of 0's is determined as before. This process continues until the last block of 0's in G_j , corresponding to the last block of A_j 's in X_j . The remainder of the G_j has distribution S (by definition of S), corresponding to all of the symbols in X_j deleted after the last occurrence of A_j .

It now follows that N has the same distribution as

$$1 + N' + J + \sum_{i=1}^{N'} M_i.$$

Recall from the proof of Lemma 1 that J is stochastically dominated by a $\text{Geom}(1-p)$ random variable for some $p \in [0, 1)$, so $\mathbf{E}[J], \mathbf{E}[J^2] < \infty$. A simple calculation now yields

$$\begin{aligned} \mathbf{E}[N^2] &= \mathbf{E} \left[\left(1 + N' + J + \sum_{i=1}^{N'} M_i \right)^2 \right] \\ &\leq \mathbf{E} \left[\left(J + \sum_{i=1}^{N'+1} (M_i + 1) \right)^2 \right] \\ &= \mathbf{E}[J^2] + 2 \mathbf{E}[J] \mathbf{E} \left[\sum_{i=1}^{N'+1} (M_i + 1) \right] \\ &\quad + \mathbf{E} \left[\left(\sum_{i=1}^{N'+1} (M_i + 1) \right)^2 \right] \\ &= \mathbf{E}[J^2] + 2 \mathbf{E}[J] \mathbf{E}[N' + 1] \mathbf{E}[M_1 + 1] \\ &\quad + \mathbf{E} \left[\sum_{i=1}^{N'+1} (M_i + 1)^2 \right] \\ &\quad + \mathbf{E} \left[\sum_{1 \leq i \neq j \leq N'+1} (M_i + 1)(M_j + 1) \right] \\ &= \mathbf{E}[J^2] + 2 \mathbf{E}[J] \mathbf{E}[N' + 1] \mathbf{E}[M_1 + 1] \\ &\quad + \mathbf{E}[N' + 1] \mathbf{E}[(M_1 + 1)^2] \\ &\quad + \mathbf{E}[N'(N' + 1)] \mathbf{E}[M_1 + 1]^2 \\ &< \infty, \end{aligned}$$

completing the proof. \blacksquare

Lemma 14:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} H(K(n) | X'(n), Y'(n)) \geq \frac{h}{\mathbf{E}[|G|]}.$$

Proof: Fix any $\epsilon_1 \in (0, 1)$ and $\epsilon_2 \in (0, 1/\mathbf{E}[|G|])$. Define $\mathcal{V} = \mathcal{V}_n(\epsilon_2)$ and $a(n, \epsilon_2)$ as in the proof of Lemma 12, and define

$$c(n, \epsilon_1, \epsilon_2) = \max(2, \lfloor (1 - \epsilon_1)a(n, \epsilon_2) \rfloor).$$

For brevity, let $Q(n) = (X'(n), Y'(n), X^{(a(n, \epsilon_2))})$. Then

$$\begin{aligned} &\frac{H(K(n) | X'(n), Y'(n))}{n} \\ &\geq \frac{H(K(n) | Q(n))}{n} \\ &= \mathbf{E} \left[\frac{-\log \Pr(K(n) | Q(n))}{n} \right] \\ &\geq \Pr(\mathcal{V}) \mathbf{E} \left[\frac{-\log \Pr(K(n) | Q(n))}{n} \mid \mathcal{V} \right] \end{aligned} \quad (13)$$

and

$$\begin{aligned} &\mathbf{E} \left[\frac{-\log \Pr(K(n) | Q(n))}{n} \mid \mathcal{V} \right] \\ &= \mathbf{E} \left[\frac{1}{n} \sum_{j=1}^{M(n)} -\log \Pr(|X_j| | Q(n), K_{j-1}) \mid \mathcal{V} \right] \\ &\geq \mathbf{E} \left[\frac{1}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} -\log \Pr(|X_j| | Q(n), K_{j-1}) \mid \mathcal{V} \right] \\ &= \frac{1}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} \mathbf{E} \left[-\log \Pr(|X_j| | Q(n), K_{j-1}) \mid \mathcal{V} \right]. \end{aligned} \quad (14)$$

Now, for any $j \leq a(n, \epsilon_2)$, we claim that $|X_j|$ is conditionally independent of $X'(n) = X_1 \cdots X_{M(n)}$ and $(Y_{a(n, \epsilon_2)+1}, \dots, Y_{M(n)})$ given $j \leq M(n)$ and $X_1 \cdots X_{a(n, \epsilon_2)}$ and $Y_1, \dots, Y_{a(n, \epsilon_2)}$ and K_{j-1} . Indeed, given $j \leq M(n)$,

$$\begin{aligned} |X_j| &\rightarrow (X_1, \dots, X_{j-1}, X_j \cdots X_{a(n, \epsilon_2)}, Y_1, \dots, Y_{M(n)}) \\ &\rightarrow (X_1 \cdots X_{M(n)}, Y_{a(n, \epsilon_2)+1}, \dots, Y_{M(n)}) \end{aligned}$$

forms a Markov chain; this is evident from Figure 1.

Thus, letting $W = W_{a(n, \epsilon_2)}$, we have that for $2 \leq j \leq c(n, \epsilon_1, \epsilon_2)$, given \mathcal{V} ,

$$\begin{aligned} &\Pr(|X_j| | Q(n), K_{j-1}) \\ &= \Pr(|X_j| | X'(n), Y'(n), X_1 \cdots X_{a(n, \epsilon_2)}, K_{j-1}) \\ &= \Pr(|X_j| | X_1 \cdots X_{a(n, \epsilon_2)}, Y_1, \dots, Y_{a(n, \epsilon_2)}, K_{j-1}) \\ &= \Pr(|X_j| | W, K_{j-1}). \end{aligned} \quad (15)$$

Combining (13)-(15) now gives

$$\begin{aligned} &\frac{H(K(n) | X'(n), Y'(n))}{n} \\ &\geq \frac{\Pr(\mathcal{V})}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} \mathbf{E} \left[-\log \Pr(|X_j| | W, K_{j-1}) \mid \mathcal{V} \right] \\ &= \frac{1}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} \mathbf{E} \left[-\log \Pr(|X_j| | W, K_{j-1}) \right] \\ &\quad - \frac{1}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} \mathbf{E} \left[-1(-\mathcal{V}) \log \Pr(|X_j| | W, K_{j-1}) \right]. \end{aligned} \quad (16)$$

$$- \frac{1}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} \mathbf{E} \left[-1(-\mathcal{V}) \log \Pr(|X_j| | W, K_{j-1}) \right]. \quad (17)$$

We lower bound (16) and (17) separately, starting with (16). For $2 \leq j \leq c(n, \epsilon_1, \epsilon_2)$, Lemma 2 yields

$$\begin{aligned}
& \mathbf{E}[-\log \Pr(|X_j| \mid W, K_{j-1})] \\
&= H(|X_j| \mid X_1 \cdots X_{a(n, \epsilon_2)}, Y_1, \dots, Y_{a(n, \epsilon_2)}, K_{j-1}) \\
&= H(|X_j| \mid X_j \cdots X_{a(n, \epsilon_2)}, Y_j, \dots, Y_{a(n, \epsilon_2)}) \\
&= H(|X_2| \mid X_2 \cdots X_{a(n, \epsilon_2)-(j-2)}, Y_2, \dots, Y_{a(n, \epsilon_2)-(j-2)}) \\
&= h_{a(n, \epsilon_2)-(j-2)} \\
&\geq h_{a(n, \epsilon_2)-c(n, \epsilon_2)+2}. \tag{18}
\end{aligned}$$

Substituting (18) into (16) now gives

$$\begin{aligned}
& \frac{1}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} \mathbf{E}[-\log \Pr(|X_j| \mid W, K_{j-1})] \\
&\geq \frac{c(n, \epsilon_1, \epsilon_2) - 1}{n} h_{a(n, \epsilon_2)-c(n, \epsilon_2)+2}. \tag{19}
\end{aligned}$$

Now we lower bound (17). We start by writing

$$\begin{aligned}
& \frac{1}{n} \sum_{j=2}^{c(n, \epsilon_1, \epsilon_2)} \mathbf{E}[-1(\neg \mathcal{V}) \log \Pr(|X_j| \mid W, K_{j-1})] \\
&\leq \frac{a(n, \epsilon_2)}{n} \mathbf{E} \left[\frac{\mathbf{1}(\neg \mathcal{V})}{a(n, \epsilon_2)} \sum_{j=1}^{a(n, \epsilon_2)} -\log \Pr(|X_j| \mid W, K_{j-1}) \right] \\
&= \frac{a(n, \epsilon_2)}{n} \mathbf{E} \left[\frac{\mathbf{1}(\neg \mathcal{V})}{a(n, \epsilon_2)} (-\log \Pr(K_{a(n, \epsilon_2)} \mid W)) \right]. \tag{20}
\end{aligned}$$

We now use Lemma 13 to bound (20). Fix some $\epsilon_3 > 0$. By Lemma 13 and Theorem 3 (in Appendix), there exists some $\delta > 0$ such that for all $m \geq 1$ and any event \mathcal{V} with $\Pr(\mathcal{V}) < \delta$, we have

$$\mathbf{E} \left[\frac{\mathbf{1}(\mathcal{V})}{m} (-\log \Pr(K_m \mid X_1 \cdots X_m, Y_1, \dots, Y_m)) \right] \leq \epsilon_3.$$

Since $\lim_{n \rightarrow \infty} \Pr(\mathcal{V}) = 1$, we have $\Pr(\neg \mathcal{V}) < \delta$ for sufficiently large n . Therefore, for sufficiently large n ,

$$\mathbf{E} \left[\frac{\mathbf{1}(\neg \mathcal{V})}{a(n, \epsilon_2)} (-\log \Pr(K_{a(n, \epsilon_2)} \mid W)) \right] \leq \epsilon_3. \tag{21}$$

Substituting (19) into (16) and substituting (21) into (20) into (17) and taking limits as $n \rightarrow \infty$ now gives

$$\begin{aligned}
& \liminf_{n \rightarrow \infty} \frac{1}{n} H(K(n) \mid X'(n), Y'(n)) \\
&\geq (1 - \epsilon_1)(-\epsilon_2 + 1/\mathbf{E}[|G|])h - \epsilon_3/\mathbf{E}[|G|]. \tag{22}
\end{aligned}$$

Since (22) holds for sufficiently small $\epsilon_1, \epsilon_2, \epsilon_3 > 0$, we may take the limit in (22) as $\epsilon_1, \epsilon_2, \epsilon_3 \rightarrow 0$ to obtain the desired result. ■

Combining Lemmas 8, 10, 12, and 14 now gives the following result, which completes the proof of Theorem 1.

Lemma 15:

$$\lim_{n \rightarrow \infty} \frac{1}{n} H(K(n) \mid X(n), Y(n)) = \frac{h}{\mathbf{E}[|G|]}.$$

IV. SIMULATION RESULTS FOR THE DELETION CHANNEL

We now apply Theorem 1 to the binary deletion channel to gain a more concrete sense of our improvement on previous work. Our objective here is merely to demonstrate that the third term in Theorem 1 is significant enough to warrant further study. We leave the issue of analyzing this term more precisely for future work.

Similarly to [4], we can use Theorem 1 to lower bound the capacity of the binary deletion channel by choosing L to be some geometric distribution. Theorem 1 then tells us that the capacity of the channel is at least

$$\begin{aligned}
& \frac{H(L)}{\mathbf{E}[L]} - \frac{H(G \mid \mathcal{L})}{\mathbf{E}[|G|]} + \frac{h}{\mathbf{E}[|G|]} \\
&= \frac{H(L)}{\mathbf{E}[L]} + \frac{H(L) - H(G, \mathcal{L}) + h}{\mathbf{E}[|G|]}. \tag{23}
\end{aligned}$$

All of the quantities in (23) except for $H(G, \mathcal{L})$ and h have closed forms. $H(G, \mathcal{L})$ can be written as an infinite sum that can be numerically approximated (see Lemma 1 in [4]). These calculations become more time-consuming as the deletion probability d_0 increases because this causes the average length $\mathbf{E}[|G|]$ of a group to increase. Indeed, if L has distribution $\text{Geom}(1-p)$, we obtain $\mathbf{E}[|G|] = \frac{1+d_0-2pd_0}{1-d_0} \cdot \frac{1}{1-p}$ from the proof of Lemma 1, which increases with d_0 .

Estimating h is much more challenging, even if all we seek is a lower bound (which, as discussed in Section III, is sufficient to improve the lower bounds on the capacity shown in [4]). We write

$$\begin{aligned}
h &\geq H(|X_2| \mid X_2 X_3, \mathcal{L}_2, \mathcal{L}_3) \\
&= H(X_2, X_3, \mathcal{L}_2, \mathcal{L}_3) - H(X_2 X_3, \mathcal{L}_2, \mathcal{L}_3) \\
&= 2H(G, \mathcal{L}) - H(X_2 X_3, \mathcal{L}_2, \mathcal{L}_3), \tag{24}
\end{aligned}$$

where the first step follows from Lemma 2 and the last step follows from Lemma 1. Since we can numerically approximate $H(G, \mathcal{L})$, it suffices to estimate (or even just lower bound) $h' \triangleq H(X_2 X_3, \mathcal{L}_2, \mathcal{L}_3)$. Similarly to $H(G, \mathcal{L})$, we can write h' as an infinite sum. However, unlike for $H(G, \mathcal{L})$, numerically approximating the sum for h' seems to require an enormous amount of time.

The basic underlying reason for the increase in the complexity of the computations is the same as the reason for the increase in the difficulty of analytically deriving the third term in the expression of Theorem 1. In essence, our approach attempts *block decoding* in the standard information theoretic sense, which was previously suggested in [4]. More specifically, [4] considers a received block of zeros (or ones) as the output symbol; for better decoding, more consecutive received blocks could be considered as the new output symbol. However, besides the increased computational effort, the combinatorial nature of the approach in [4] yields very complicated formulas, even when considering output symbols consisting of only two received blocks, making it difficult to quantify the improvement due to block decoding. Here, we succinctly identify this improvement for output symbols consisting of any number of consecutive received blocks. Indeed, adding $h_m/\mathbf{E}[|G|]$ to the first two terms of Theorem 1 essentially mimics block decoding for $m-1$ blocks.

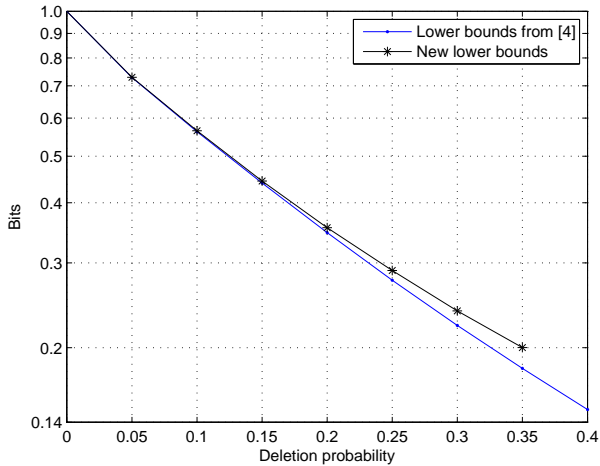


Fig. 2. Lower bounds on the capacity of the binary deletion channel.

Of course, the main drawback for block decoding is that it is computationally challenging; in our case, considering multiple blocks of the received sequence instead of just one introduces some very complicated dependencies. The details are fairly technical, but in brief, these dependencies present themselves as a large number of nested summations for h' , making it much more difficult to compute than $H(G, \mathcal{L})$, where this nesting can be reduced. Thus, we resort to estimating h' through simulation.

We estimate h' by taking $t = 5 \times 10^7$ independent samples from the distribution of $(X_2 X_3, \mathcal{L}_2, \mathcal{L}_3)$ and computing the entropy \hat{h}' of the resulting empirical distribution. Our simulations suggest that \hat{h}' converges from below as t increases, but that the chosen value of t gives a sufficiently good approximation of h' for the resulting lower bound on the capacity of the channel to be reasonably accurate (for the values of d_0 that we consider).

We give the results of our simulations are shown in Figure 2. Since the computation time increases quickly as d_0 grows, we only have results for $d_0 \leq 0.35$. However, it is evident that our theoretical results allow for improved lower bounds on the capacity. Indeed, at $d_0 = 0.35$, the new lower bound exceeds that of [4] by more than 10%. Furthermore, we expect this improvement to increase with d_0 .

V. CONCLUSIONS

We give a new approach to lower bounding the capacity of channels with i.i.d. deletions and duplications that is simple, intuitive, and provides stronger theoretical results than currently obtainable using the more standard combinatorial approach. Specifically, we consider channels governed by any deletion/duplication distribution D with $H(D) < \infty$ and we consider sources with i.i.d. block lengths given by any distribution L with $\mathbf{E}[L], H(L) < \infty$, as opposed to prior work [3], [4], which requires that D have geometrically decreasing tails and that L is geometric or has finite support. Furthermore, our results essentially reveal the limits of using i.i.d. block lengths to lower bound the channel capacity, which is currently the

standard technique for deriving such lower bounds. Finally, we show that our techniques allow for improved lower bounds on the capacity of the binary deletion channel.

ACKNOWLEDGMENT

The authors are very grateful to Michael Mitzenmacher for his comments and many useful discussions.

APPENDIX UNIFORM INTEGRABILITY

This appendix gives an overview of the concept of uniform integrability and a preview of its importance in our analysis. We include this appendix so that readers who are unfamiliar with uniform integrability can gain some additional intuition before reading our proofs that rely on it. Of course, we give only a brief overview; for more information see, for example, [5, Chapter 7.10]. We begin with the standard definition.

Definition 1: A family of random variables $\{X_n : n \geq 1\}$ is *uniformly integrable* if

$$\lim_{a \rightarrow \infty} \sup_n \mathbf{E}[|X_n| \mathbf{1}(|X_n| \geq a)] = 0,$$

where $\mathbf{1}(\cdot)$ denotes the indicator function.

At first glance, Definition 1 is not particularly enlightening. However, if we write

$$\mathbf{E}[|X_n| \mathbf{1}(|X_n| \geq a)] = \Pr(|X_n| \geq a) \mathbf{E}[|X_n| \mid |X_n| \geq a],$$

then we can see that Definition 1 formalizes the idea that as a grows very large, $\Pr(|X_n| \geq a)$ shrinks to 0, and $\mathbf{E}[|X_n| \mid |X_n| \geq a]$ does not grow fast enough to prevent the product from converging to 0. Furthermore, this effect happens *uniformly* over all n . Indeed, this uniformity is what makes uniform integrability stronger than just saying that each X_n is integrable (that is, $\mathbf{E}[|X_n|] < \infty$ for every n).

There are two important characterizations of uniform integrability that we use in our analysis. We begin with the first that we use.

Theorem 2: [5, Theorem 7.10.3] Suppose that $X_n \rightarrow X$ in probability as $n \rightarrow \infty$. Then the following are equivalent:

- 1) $\{X_n : n \geq 1\}$ is uniformly integrable.
- 2) $\mathbf{E}[|X_n|] < \infty$ for all n , $\mathbf{E}[|X|] < \infty$, and $X_n \rightarrow X$ in mean as $n \rightarrow \infty$.
- 3) $\mathbf{E}[|X_n|] < \infty$ for all n , and $\mathbf{E}[|X_n|] \rightarrow \mathbf{E}[|X|]$ as $n \rightarrow \infty$.

In our applications of Theorem 2, the family $\{X_n : n \geq 1\}$ is always nonnegative, and we know that $X_n \rightarrow X$ almost surely as $n \rightarrow \infty$. In this case, the equivalence of the first and third items in the theorem essentially tells us that uniform integrability is a characterization of when one may interchange limit and expectation operations. In particular, the theorem allows for generalizations of the dominated convergence theorem, which is the result that is usually used to justify such an interchange.

The second characterization of uniform integrability that we present is slightly more technical in appearance.

Theorem 3: [5, Lemma 7.10.6] The family $\{X_n : n \geq 1\}$ is uniformly integrable if and only if both of the following conditions hold:

- 1) $\sup_n \mathbf{E}[|X_n|] < \infty$.
- 2) For any $\epsilon > 0$, there exists some $\delta > 0$ such that for all n and any event \mathcal{V} with $\Pr(\mathcal{V}) < \delta$, we have $\mathbf{E}[|X_n| \mathbf{1}(\mathcal{V})] < \epsilon$.

In our proofs, we always know that $\{X_n : n \geq 1\}$ is uniformly integrable and nonnegative when we apply Theorem 3, and we are interested in obtaining the second item. The second item is significant to us because if we have some sequence of “good” events $\{\mathcal{V}_n\}_{n \geq 1}$ such that $\lim_{n \rightarrow \infty} \Pr(\mathcal{V}_n) = 1$, then we may write

$$\mathbf{E}[X_n] = \Pr(\mathcal{V}_n) \mathbf{E}[X_n | \mathcal{V}_n] + \mathbf{E}[X_n \mathbf{1}(\neg \mathcal{V}_n)].$$

In our proofs, we are interested in computing $\lim_{n \rightarrow \infty} \mathbf{E}[X_n]$, and we essentially do this by computing $\lim_{n \rightarrow \infty} \mathbf{E}[X_n | \mathcal{V}_n]$ (that is, the limit of the conditional expectation of X_n given the corresponding “good” event). We use this technique because conditioning on the “good” event \mathcal{V}_n provides us with a lot of structure that we can use to analyze $\mathbf{E}[X_n | \mathcal{V}_n]$. However, to argue that

$$\lim_{n \rightarrow \infty} \mathbf{E}[X_n] = \lim_{n \rightarrow \infty} \mathbf{E}[X_n | \mathcal{V}_n],$$

we need to show that

$$\lim_{n \rightarrow \infty} \mathbf{E}[X_n \mathbf{1}(\neg \mathcal{V}_n)] = 0. \quad (25)$$

Theorem 3 gives us the mechanism to prove (25). We fix some arbitrary $\epsilon > 0$ and apply Theorem 3 to guarantee the existence of some $\delta > 0$ such that for all n any event \mathcal{V} with $\Pr(\mathcal{V}) < \delta$, we have $\mathbf{E}[X_n \mathbf{1}(\mathcal{V})] < \epsilon$. Now, since $\lim_{n \rightarrow \infty} \Pr(\mathcal{V}_n) = 1$, we know that $\Pr(\neg \mathcal{V}_n) < \delta$ for sufficiently large n , and therefore

$$\limsup_{n \rightarrow \infty} \mathbf{E}[X_n \mathbf{1}(\neg \mathcal{V}_n)] \leq \epsilon.$$

Since the above equation holds for any $\epsilon > 0$, we may take the limit as $\epsilon \rightarrow 0$ to obtain (25).

REFERENCES

- [1] S. Diggavi and M. Grossglauser, “On Information Transmission over a Finite Buffer Channel,” *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1226-1237, 2006.
- [2] 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.
- [3] E. Drinea and M. Mitzenmacher, “On Lower Bounds for the Capacity of Deletion Channels,” *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4648-4657, 2006.
- [4] E. Drinea and M. Mitzenmacher, “Improved Lower Bounds for I.I.D. Deletion and Insertion Channels,” *IEEE Trans. Inf. Theory*, vol. 53, no. 8, pp. 2693-2714, 2007.
- [5] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Third Edition, Oxford University Press, 2003.
- [6] M. Mitzenmacher and E. Drinea, “A Simple Lower Bound for the Capacity of the Deletion Channel,” *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4657-4660, 2006.
- [7] S. Ross, *Applied Probability Models with Optimization Applications*. Holden-Day, 1970.