Optimal Coding for the Binary Deletion Channel With Small Deletion Probability

Yashodhan Kanoria, *Student Member, IEEE*, and Andrea Montanari, *Senior Member, IEEE*

***Abstract*—The binary deletion channel is the simplest point-to-point communication channel that models lack of synchronization. Input bits are deleted independently with probability d, and when they are not deleted, they are not affected by the channel. Despite significant effort, little is known about the capacity of this channel and even less about optimal coding schemes. In this paper, we develop a new systematic approach to this problem, by demonstrating that capacity can be computed in a series expansion for small deletion probability. We compute three leading terms of this expansion, and find an input distribution that achieves capacity up to this order. This constitutes the first optimal random coding result for the deletion channel. The key idea employed is the following: We understand perfectly the deletion channel with deletion probability $d = 0$. It has capacity 1 and the optimal input distribution is iid Bernoulli$(1/2)$. It is natural to expect that the channel with small deletion probabilities has a capacity that varies smoothly with d, and that the optimal input distribution is obtained by smoothly perturbing the iid Bernoulli$(1/2)$ process. Our results show that this is indeed the case.**

***Index Terms*—Capacity achieving code, channel capacity, deletion channel, series expansion.**

I. Introduction

The binary deletion channel accepts bits as inputs, and deletes each transmitted bit independently with probability d. Computing or providing systematic approximations to its capacity is one of the outstanding problems in information theory [1]. An important motivation comes from the need to understand synchronization errors and optimal ways to cope with them.

In this paper, we suggest a new approach. We demonstrate that capacity can be computed in a series expansion for small deletion probability, by computing the first two orders of such an expansion. Our main result is the following.

Theorem I.1: Let $C(d)$ be the capacity of the deletion channel with deletion probability d. Then, for small d and any $\epsilon > 0$,

$$C(d) = 1 + d \log d - A_1\, d + A_2\, d^2 + O(d^{3-\epsilon}), \qquad (1)$$

where

$$\begin{aligned}
A_1 &\equiv \log(2e) - \sum_{l=1}^{\infty} 2^{-l-1} l \log l \;\approx 1.15416377\\
A_2 &\equiv c_3 + c_4 + \\
&\quad \frac{1}{4\ln 2}\left(2 + \frac{3}{2}c_2^2 + \sum_{l=1}^{\infty} 2^{-l}\,(l \ln l)^2 - c_2 \sum_{l=1}^{\infty} 2^{-l} l^2 \ln l\right)\\
&\approx\; 1.67814594\\
c_2 &\equiv \sum_{l=1}^{\infty} 2^{-l} l \ln l \approx 1.78628364\\
c_3 &\equiv \frac{1}{2}\left(-1 + \sum_{l=3}^{\infty} 2^{-l}\left\{\binom{l}{2}\log\binom{l}{2} - l^2 \log l + \right.\right.\\
&\qquad \left.\left.(l-1)(l-3)\log(l-1) + (l-2)\log(l-2)\right\}\right)\\
&\approx\; -0.88636960\\
c_4 &\equiv \sum_{j=4}^{\infty} 2^{-(2+j)}\,(j-1)(j-3)\,h\!\left(\frac{1}{j-1}\right)\\
&\quad + \sum_{i=2}^{\infty}\sum_{j=4}^{\infty} 2^{-(i+j+1)}\,(i+j-1)(j-3)\,h\!\left(\frac{i+1}{i+j-1}\right)\\
&\approx 0.69001321.
\end{aligned}$$

Here, $h(\,\cdot\,)$ is the binary entropy function, i.e., $h(p) \equiv -p\log p - (1-p)\log(1-p)$.

Further, the binary stationary source defined by the property that the times at which it switches from 0 to 1 or vice versa form a renewal process with holding time distribution $p_L(l) = 2^{-l}(1 + d(l \ln l - c_2 l/2))$, achieves rate within $O(d^{3-\epsilon})$ of capacity.

Given a binary sequence, we will call "runs" its maximal blocks of contiguous 0s or 1s. We shall refer to binary sources such that the switch times form a renewal process as sources (or processes) with iid runs. The "rate" of a given binary source is the maximum rate at which information can be transmitted through the deletion channel using input sequences distributed as the source. A formal definition is provided later (see Definition II.3). Logarithms denoted by log here (and in the rest of the paper) are understood to be in base 2.

Manuscript received April 28, 2011; revised February 19, 2013; accepted April 21, 2013. Date of publication May 16, 2013; date of current version September 11, 2013. This work was supported in part by the NSF under Grants CCF-0743978 and CCF-0915145 and in part by a Terman fellowship. Y. Kanoria was supported by a 3Com Corporation Stanford Graduate Fellowship. This paper was presented in part at the 2010 IEEE International Symposium on Information Theory.

Y. Kanoria was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with the Decision, Risk, and Operations Division, Columbia Business School, New York, NY 10027 USA (e-mail: ykanoria@columbia.edu)

A. Montanari is with the Departments of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: montanari@stanford.edu).

Communicated by S. Diggavi, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2262020

A few remarks on Theorem I.1 are in order.

Bounds versus asymptotic expansions: The proof of Theorem I.1 consists in establishing upper and lower bounds on capacity that match up to quadratic order in d. However, we explicitly evaluating the constants in the error terms, and hence, (1) does not provide either an upper or a lower bound at $d > 0$. It would be very interesting to obtain explicit expressions for these constants. Although technically daunting, we do not see any conceptual obstacle to such a calculation.

While (1) is only asymptotically exact as $d \to 0$, it provide useful guidance in designing concrete coding schemes. If a coding scheme aims at achieving capacity for small d, its rate should match (1) up to higher order terms. This test can be very stringent. In particular, our proof of Theorem I.1 implies the following.

Remark I.2: There exists $d_* > 0$ such that for any $d \in (0, d_*)$ no coding scheme such that the empirical distribution of codewords is given by a Markov process [2] or a hidden Markov process with state space of bounded cardinality can achieve capacity.

Indeed Markov processes or hidden Markov processes have run-length distribution that is exponential or sum of exponentials, thus not matching the distribution $p_L(l)$. Our proof, in fact, establishes that the rate achieved by Markov processes is $\Omega(d^2)$ below capacity (Theorem VI.1 states this for first-order Markov processes). Notice that the best previous bounds could not rule out the hypothesis the Markov sources are capacity achieving.

Optimal coding schemes: Theorem I.1 shows that the stationary process consisting of iid runs with the specified run-length distribution, achieves a rate to within $O(d^{3-\epsilon})$ of capacity. In particular, a random codebook that achieves this rate is given as follows. For blocklength n, and rate R, generate 2^{nR} codewords independently. Each codeword $x \in \{0, 1\}^n$, has iid run lengths $\{l_i\}_{i \in \mathbb{N}}$, with $l_i \sim p_L(\,\cdot\,)$. (We refer to Section IV for further details.) Decoding can be performed by maximum likelihood.

Notice that this is not a *practical* coding scheme in terms encoding and decoding complexity. However, as often in information theory, it can provide useful intuition toward the construction of a practical scheme.

Why $d \to 0$?: The regime $d \to 0$ appears to be particularly appealing for the methods developed here. On one hand, the case $d = 0$ is trivial, and hence, one can hope to accurately approximate the capacity in a neighborhood of this limit case. On the other, synchronization errors are infrequent in many applications, in natural correspondence with the regime under consideration. For instance, the deletion channel in the $d \to 0$ regime has bearing on the problem of file synchronization. This connection has been explored in recent work [3], building on the conference version of this paper [14].

Higher order terms: Finally, asymptotic expansions as the one studied here allow to isolate different sources of uncertainty, and order them by their impact for small d. As clarified by the proof of Theorem I.1, the $(d \log d - A_1 d)$ term in (1) is due to the occurrence of a single deletion in a run or a small sequence of runs, and hence, to the uncertainty about its location. An optimal scheme has to cope with this uncertainty optimally.

Computing further terms in the capacity expansion (1) reveals additional structure. For instance, at the moment we cannot disprove the hypothesis that a source with iid runs achieves capacity over an interval $[0, d_0)$. However, we suspect that computing the next, $\Theta(d^3)$ and $\Theta(d^4)$, terms in the expansion will solve in negative sense this question.

A related open question is whether the small d series is absolutely convergent up to some radius $d_0 > 0$. If this was the case, the small d expansion would provide a systematic way to address the capacity problem for all $d \in [0, d_0)$. See Section VI for further comments.

The underlying philosophy of this study is that whenever capacity of a channel is known for a specific value of the channel parameter, and the corresponding optimal input distribution is unique and well characterized, it should be possible to compute an asymptotic expansion around that value. In the present context, the special channel is the perfect channel, i.e., the deletion channel with deletion probability $d = 0$. The corresponding input distribution is the iid Bernoulli$(1/2)$ process. Similar approaches have been successful in other contexts, e.g., hidden Markov chains and related channels [4].

A. Related Work

Dobrushin [5] proved a coding theorem for the deletion channel, and other channels with synchronization errors. He showed that the maximum rate of reliable communication is given by the maximal mutual information per bit, and proved that this can be achieved through a random coding scheme. This characterization has so far found limited use in proving concrete estimates. An important exception is provided by the work of Kirsch and Drinea [6] who use the Dobrushin coding theorem to prove lower bounds on the capacity of channels with deletions and duplications. We will also use the Dobrushin theorem in a crucial way, although most of our effort will be devoted to proving upper bounds on the capacity.

Several capacity bounds have been developed, starting with achievability results by Gallager [7], which have been significantly improved in recent years [2], [8]–[10]. Diggavi and Grossglauser [8], [9] suggested codebooks with memory for the deletion channel, in particular Markovian codebooks. Drinea, Kirsch, and Mitzenmacher [2], [6] improved lower bounds using better decoders, and also considered codebooks with iid run lengths. However, numerical results were again restricted to the special case of first-order Markov inputs, with the best first-order Markov process being estimated numerically. An upper bound on capacity is proved in [13] by optimizing communication rate over an augmented channel over input distributions with iid runs. The augmented channel essentially sends a synchronization symbol at the end of each run in the input. The optimal input for the augmented channel is quite different from the optimal input for the deletion channel, since sending short runs does not cause synchronization difficulties in the augmented channel.

A trivial upper bound to the capacity of the deletion channel is $1-d$, the capacity of the corresponding erasure channel. It has been proved that, in fact, $C(d) = \Theta(1-d)$ as $d \to 1$ [10]. The papers [11]–[13] improve the upper bound in this limit obtaining

$\limsup_{d\to 1} C(d)/(1-d) \le 0.413$. However, determining the asymptotic behavior in this limit [i.e., finding a constant B_1 such that $C(d) = B_1(1-d) + o(1-d)$] is an open problem. The authors in [11] and [13] obtained upper bounds for general deletion probabilities, using various augmented channels. When applied to the small d regime, none of the known upper bounds actually captures the correct behavior as stated in (1). A simple calculation shows that the first upper bound in [13] has asymptotics of $1 + (3/4)d \log d$. Another work [11] shows that $C \le 1 - 4.19\,d$ as $d \to 0$. The recent survey by Mitzenmacher [1] provides a useful entry point to this literature.

Against this backdrop, our study proves that random codebooks with iid runs are optimal for small deletion probability up to corrections of order $O(d^{3-\epsilon})$. We thus provide the first rigorous justification for the use of iid run lengths. We further determine analytically the optimal distribution of the runs for small d. As a byproduct of our analysis, we are able to characterize the performance of first-order Markov inputs analytically, and find that such inputs are suboptimal by $\Theta(d^2)$ terms. In asymptotic sense (for small d), Markovian inputs are no better than an iid Bernoulli$(1/2)$ input (cf., Section VI).

An earlier version of this paper was presented at the IEEE International Symposium on Information Theory 2010 [14]. That paper determined the $\Theta(d)$ and $\Theta(d \log d)$ terms in the expansion, namely $C(d) = 1+d\log d - A_1\,d + O(d^{3/2-\epsilon})$, and proved that this rate is achievable by iid Bernoulli$(1/2)$ input. Concurrent work by Kalai, Mitzenmacher, and Sudan [15], presented at the same conference, established that $C(d) = 1+d\log d + O(d)$ using a very different counting argument. As should be clear from the proof in this paper, proving Theorem I.1 is significantly more challenging than proving the results in [14] and[15]. We undertook this challenge because computing the $\Theta(d^2)$ term leads to new insights in the capacity achieving codebook.

1) On one hand, [14] and [15] provided limited coding insights, for two reasons. First of all, it is unsurprising (and follows from earlier bounds) that, as $d \to 0$, Bernoulli$(1/2)$ achieves capacity, a continuity argument being sufficient. Second, Markovian codebooks (hence, a fortiori Bernoulli$(1/2)$ codebooks) were already well studied before these works.
2) On the other, this paper presents a codebook (iid runs with explicitly given run-length distribution p_L) that was not known before, and achieves capacity to the desired order.

B. Numerical Illustration of Results

We can numerically evaluate the expression in (1) (dropping the error term) to obtain estimates of capacity for small deletion probabilities.

$$C_{\text{est}} = 1 + d \log d - A_1\, d + A_2\, d^2\,.$$

The values of C_{est} are presented in Table I and Fig. 1. We compare with the best known numerical lower bounds [2] and upper bounds [11], [13].

We stress here that C_{est} is neither an upper nor a lower bound on capacity. It is an estimate based on taking the leading terms of the asymptotic expansion of capacity for small d, and is expected to be accurate for small values of d. Indeed, we see that for d larger than 0.4, our estimate C_{est} *exceeds* the upper bound. This simply indicates that we should not use C_{est} as an estimate for such large d.

TABLE I
TABLE SHOWING BEST KNOWN NUMERICAL BOUNDS ON CAPACITY (FROM [2], [11], AND [13]) COMPARED WITH OUR ESTIMATE BASED ON THE SMALL d EXPANSION

d	Best lower bound	C_{est}	Best upper bound
0.05	0.7283	0.7304	0.8160
0.10	0.5620	0.5692	0.6890
0.15	0.4392	0.4541	0.5790
0.20	0.3467	0.3719	0.4910
0.25	0.2759	0.3163	0.4200
0.30	0.2224	0.2837	0.3620
0.35	0.1810	0.2715	0.3150
0.40	0.1484	0.2781	0.2750
0.45	0.1229	0.3020	0.2410
0.50	0.1019	0.3425	0.2120

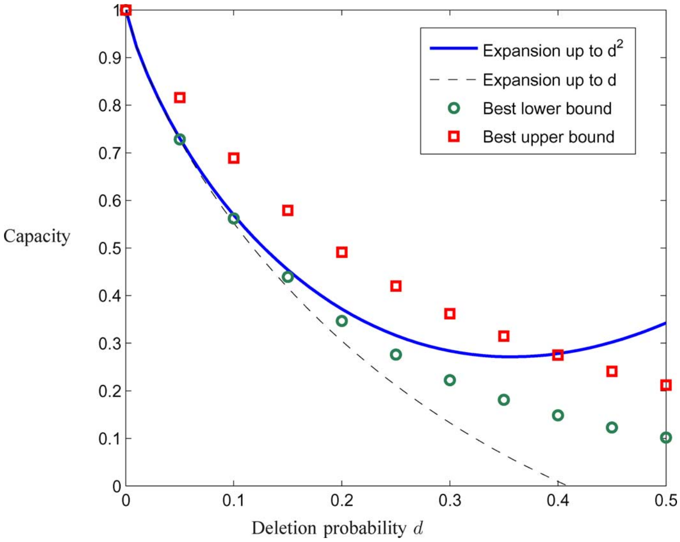

Fig. 1. Plot showing best known numerical bounds on capacity (from [2], [11], and [13]) compared with our estimate based on the small d expansion.

C. Notation

We borrow $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ notations from the computer science literature. We define these as follows to fit our needs. Let $f : [0,1] \to \mathbb{R}$ and $g : [0,1] \to \mathbb{R}_+$. We say:

1) $f = O(g)$, if there is a constant $c < \infty$ such that $|f(x)| \le cg(x)$ for all $x \in [0,1]$.
2) $f = \Omega(g)$, if there is a constant $c > 0$ such that $f(x) \ge cg(x)$ for all $x \in [0,1]$.
3) $f = \Theta(g)$, if there are constants $c < \infty$, $c' > 0$ such that $cg(x) \ge f(x) \ge c'g(x)$ for all $x \in [0,1]$.

Throughout this paper, we adhere to the convention that the aforementioned constants c, c' should not depend on the processes $\mathbb{X}, \mathbb{Y}, \ldots$ etc., under consideration, if there are such processes.

D. Outline of the Paper

Section II contains the basic definitions and results necessary for our approach to estimating the capacity of the deletion channel. We show that it is sufficient to consider stationary ergodic input sources, and define their corresponding rate (mutual information per bit). Capacity is obtained by maximizing this quantity over stationary processes. In Section III, we present an informal argument that contains the basic intuition leading to our main result (see Theorem I.1), and allows us to correctly guess the optimal input distribution. Section IV states a small number of core lemmas, and shows that they imply Theorem I.1. Finally, Section V states several technical results (proved in Appendix) and uses them to prove the core lemmas. We conclude with a short discussion, including open problems, in Section VI.

II. Preliminaries

For the reader's convenience, we restate here some known results that we will use extensively, along with some definitions and auxiliary lemmas.

Consider a sequence of channels $\{W_n\}_{n\geq 1}$, where W_n allows exactly n inputs bits, and deletes each bit independently with probability d. The output of W_n for input X^n is a binary vector denoted by $Y(X^n)$. The length of $Y(X^n)$ is a binomial random variable. We want to find maximum rate at which we can send information over this sequence of channels with vanishingly small error probability.

The following characterization follows from [5].

Theorem II.1: Let

$$C_n \equiv \frac{1}{n} \max_{p_{X^n}} I(X^n; Y(X^n)) \,. \tag{2}$$

Then, the following limit exists

$$C \equiv \lim_{n\to\infty} C_n = \inf_{n\geq 1} C_n, \tag{3}$$

and is equal to the capacity of the deletion channel.

Note that in (2), we know that $\sup_{p_{X^n}} I(X^n; Y(X^n))$ is achieved since $I(X^n; Y(X^n))$ is a continuous function on a compact (the set of possible input distributions p_{X^n}).

A further useful remark [5, Th. 5] is that, in computing capacity, we can assume $(X_1, \ldots, X_n)$ to be n consecutive coordinates of a stationary ergodic process. We denote by $\mathcal{S}$ the class of stationary and ergodic processes that take binary values. This result of Dobrushin is restated formally below.

Lemma II.2: Let $\mathbb{X} = \{X_i\}_{i\in\mathbb{Z}}$ be a stationary and ergodic process, with X_i taking values in $\{0,1\}$. Then, the limit $I(\mathbb{X}) \equiv \lim_{n\to\infty} \frac{1}{n} I(X^n; Y(X^n))$ exists and

$$C = \sup_{\mathbb{X}\in\mathcal{S}} I(\mathbb{X}) \,.$$

We use the following natural definition of the *rate* achieved by a stationary ergodic process.

Definition II.3: For stationary and ergodic $\mathbb{X}$, we call $I(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} I(X^n; Y(X^n))$ the rate achieved by $\mathbb{X}$.

Proofs of Theorem II.1 and Lemma II.2 are provided in Appendix for the convenience of the reader.

Given a stationary process $\mathbb{X}$, it is convenient to consider it from the point of view of a "uniformly random" block/run. Intuitively, this corresponds to choosing a large integer n and selecting as reference point the beginning of a uniformly random block in $X_1, \ldots, X_n$. Notice that this approach naturally discounts longer blocks for finite n. While such a procedure can be made rigorous by taking the limit $n \to \infty$, it is more convenient to make use of the notion of *Palm measure* from the theory of point processes [17], [18], which is, in this case, particularly easy to define. To a binary source $\mathbb{X}$, we can associate in a bijective way a subset of times $\mathbb{S} \subseteq \mathbb{Z}$, by letting $t \in \mathbb{S}$ if and only if X_t is the first bit of a run. The Palm measure $\mathbb{P}_1$ is then the distribution of $\mathbb{X}$ conditional on the event $1 \in \mathbb{S}$. We refer to Appendix for further details.

We denote by L the length of the block starting at 1 under the Palm measure, and denote by p_L its distribution. As an example, if $\mathbb{X}$ is the iid Bernoulli$(1/2)$ process, we have $p_L = p_L^*$ where $p_L^*(l) \equiv 2^{-l}$. We will also call p_L the block-perspective run-length distribution or simply the run-length distribution, and let

$$\mu(\mathbb{X}) \equiv \mathbb{E}[L] = \sum_{l=1}^{\infty} p_L(l)\, l,$$

be its average. Let L_0 be the length of the block containing bit X_0 in the stationary process $\mathbb{X}$. A standard calculation [17], [18] yields $\mathbb{P}(L_0 = l) = l p_L(l)/\mu(\mathbb{X})$. Since L_0 is a well defined and almost surely finite (by ergodicity), we necessarily have $\mu(\mathbb{X}) < \infty$.

In our main result, Theorem I.1, a special role is played by processes $\mathbb{X}$ such that the associated switch times form a stationary renewal process. We will refer to such an $\mathbb{X}$ as a process with iid runs.

III. Intuition Behind the Main Theorem

In this section, we provide a heuristic/nonrigorous explanation for our main result. The aim is to build intuition and motivate our approach, without getting bogged down with the numerous technical difficulties that arise. In fact, we focus here on heuristically deriving the optimal input process $\mathbb{X}^\dagger$, and do not actually obtain the quadratic term of the capacity expansion. We find $\mathbb{X}^\dagger$ by computing various quantities to leading order and using the following observation (cf., Remark IV.2).

A. Key Observation

The process that achieves capacity for small d should be "close" to the Bernoulli$(1/2)$ process, since $H(\mathbb{X})$ must be close to 1.

We have

$$I(X^n; Y(X^n)) = H(Y) - H(Y|X^n) \,. \tag{4}$$

Let D^n be a binary vector containing a one at position i if and only if X_i is deleted from the input vector. We can write

$$H(Y|X^n) = H(Y, D^n|X^n) - H(D^n|X^n, Y) \,.$$

But Y is a function of (X^n, D^n), leading to $H(Y, D^n|X^n) = H(D^n|X^n) = H(D^n) = nh(d)$, where we used the fact that D^n is iid Bernoulli(d), independent of X^n. It follows that

$$H(Y|X^n) = nh(d) - H(D^n|X^n, Y) \,. \tag{5}$$

The term $H(D^n|X^n, Y)$ represents ambiguity in the location of deletions, given the input and output strings. Now, since d is small, we expect that most deletions occur in "isolation," i.e., far away from other deletions. Make the (incorrect) assumption that all deletions occur such that no three consecutive runs have more than one deletion in total. In this case, we can unambiguously associate runs in $\mathbb{Y}$ with runs in $\mathbb{X}$. Ambiguity in the location of a deletion occurs if and only if a deletion occurs in a run of length $l > 1$. In this case, each of l locations is equally likely for the deletion, leading to a contribution of $\log l$ to $H(D^n|X^n, Y)$. Now, a run of length l should suffer a deletion with probability $\approx ld$. Thus, we expect

$$\frac{1}{n}H(D^n|X^n,Y) \approx \frac{d}{\mu(\mathbb{X})}\sum_{l=1}^{\infty} p_L(l) l \log l\,.$$

We know that $H(\mathbb{X})$ is close to 1, implying $\mu(\mathbb{X})$ is close to 2 and p_L is close to $p_L^*(l) \equiv 2^{-l}$. This leads to

$$\begin{aligned}
&\frac{1}{n}H(D^n|X^n,Y)\\
&\approx \frac{d}{2}\sum_{l=1}^{\infty} p_L(l) l \log l - \frac{d(\mu(\mathbb{X})-2)}{4}\sum_{l=1}^{\infty} p_L^*(l) l \log l\\
&= \frac{d}{2}\left[\frac{c_2}{\ln 2} + \sum_{l=1}^{\infty} p_L(l) l \left(\log l - \frac{c_2}{2\ln 2}\right)\right]\,. \qquad (6)
\end{aligned}$$

Consider $H(Y)$. Now, if the input X^n is drawn from a stationary process $\mathbb{X}$, we expect the output $Y(X^n)$ to also be a segment of some stationary process $\mathbb{Y}$. (It turns out that this is the case.) Moreover, we expect that the channel output has $n(1-d) + o(n)$ bits, leading to $H(Y) \approx n(1-d)H(\mathbb{Y})$. Denote the run-length distribution in $\mathbb{Y}$ by $q_L(\cdot)$. Define $\mu(\mathbb{Y}) \equiv \sum_{l=1}^{\infty} q_L(l) l$. Let $L_{\mathbb{Y}}$ denote the length of a random run drawn according to $q_L(\cdot)$. It is not hard to see that

$$H(\mathbb{Y}) \leq H(L_{\mathbb{Y}})/\mu(\mathbb{Y}),$$

with equality if and only if $\mathbb{Y}$ consists of iid runs, which occurs if and only if $\mathbb{X}$ consists of iid runs. Define $q_L^*(l) \equiv 2^{-l}$. An explicit calculation yields $H(L_{\mathbb{Y}}) = \mu(\mathbb{Y}) - D(q_L\|q_L^*)$. We know that $H(\mathbb{Y})$ is close to 1, implying $\mu(\mathbb{Y})$ is close to 2 and $D(q_L\|q_L^*)$ is small. Thus,

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}H(Y)\\
&= (1-d)H(\mathbb{Y})\\
&\leq (1-d)(1 - D(q_L\|q_L^*)/\mu(\mathbb{Y}))\\
&\approx 1 - d - D(q_L\|q_L^*)/2\,.
\end{aligned}$$

Notice that an iid Bernoulli$(1/2)$ input results in an iid Bernoulli$(1/2)$ output from the deletion channel. The following is made precise in Lemma V.9: Let Δ be the "distance" between p_L and p_L^*. Then, a short calculation tells us that the distance between p_L and q_L should be $O(d^{1-\epsilon}\Delta)$. In other words p_L and q_L are very nearly equal to each other.

So we obtain, to leading order,

$$\lim_{n\to\infty}\frac{1}{n}H(Y) \lesssim 1 - d - D(p_L\|p_L^*)/2, \qquad (7)$$

with (approximate) equality if and only if $\mathbb{X}$ consists of iid runs.

Putting (4)–(7) together, we have

$$\begin{aligned}
I(\mathbb{X}) &= \lim_{n\to\infty}\frac{1}{n}I(X^n;Y)\\
&\lesssim 1 - d - D(p_L\|p_L^*)/2 - h(d) +\\
&\quad \frac{d}{2}\left[\frac{c_2}{\ln 2} + \sum_{l=1}^{\infty} p_L(l) l \left(\log l - \frac{c_2}{2\ln 2}\right)\right]\\
&\approx 1 - d\log(1/d) - A_1 d - \frac{1}{2}D(p_L\|p_L^*) +\\
&\quad \frac{d}{2}\left[\sum_{l=1}^{\infty} p_L(l) l \left(\log l - \frac{c_2}{2\ln 2}\right)\right]\,.
\end{aligned}$$

Since this (approximate) upper bound on $I(\mathbb{X})$ depends on input $\mathbb{X}$ only through p_L, we choose $\mathbb{X}$ consisting of iid runs so that (approximate) equality holds.

We expect p_L to be close to $p_L^*(l)$. A Taylor expansion gives

$$\begin{aligned}
D(p_L\|p_L^*) &= \sum_{l=1}^{\infty} p_L(l)(l + \log p_L(l))\\
&\approx \frac{1}{\ln 2}\sum_{l=1}^{\infty}\Big(\left(p_L(l) - 2^{-l}\right) + 2^{l-1}\left(p_L(l) - 2^{-l}\right)^2\Big)\\
&= \frac{1}{\ln 2}\sum_{l=1}^{\infty} 2^{l-1}\left(p_L(l) - 2^{-l}\right)^2\,. \qquad (8)
\end{aligned}$$

Thus, we want to maximize

$$\begin{aligned}
&\frac{1}{2\ln 2}\sum_{l=1}^{\infty} 2^{l-1}\left(p_L(l) - 2^{-l}\right)^2 +\\
&\quad \frac{d}{2}\left[\sum_{l=1}^{\infty} p_L(l) l \left(\log l - \frac{c_2}{2\ln 2}\right)\right],
\end{aligned}$$

subject to $\sum_{l=1}^{\infty} p_L(l) = 1$, in order to achieve the largest possible $I(\mathbb{X})$. A simple calculation tells us that the maximizing distribution is $p_L^\dagger(l) = 2^{-l}(1 + d(l\ln l - c_2 l/2))$.

IV. Proof of the Main Theorem: Outline

In this section, we provide the proof of Theorem I.1 after stating the key lemmas involved. We defer the proof of the lemmas to the next section. Sections V-A–E develop the technical machinery we use, and the proofs of the lemmas are in Section V-F.

Given a (possibly infinite) binary sequence, a *run* of 0s (of 1s) is a maximal subsequence of consecutive 0s (1s), i.e., a subsequence of 0s bordered by 1s (respectively, of 1s bordered by 0s). The first step consists of proving achievability by estimating $I(\mathbb{X})$ for a process having iid runs with appropriately chosen distribution.

Lemma IV.1: Let $\mathbb{X}^\dagger$ be the process consisting of iid runs with distribution $p_L^\dagger(l) = 2^{-l}(1 + d(l\log l - c_2 l/2))$. Then, for any $\epsilon > 0$, we have

$$I(\mathbb{X}^\dagger) = 1 + d\log d - A_1\, d + A_2\, d^2 + O(d^{3-\epsilon})\,.$$

Lemma IV.1 is proved in Section V-F.

TABLE II
EXAMPLE SHOWING HOW $\mathbb{X}$ IS DIVIDED INTO SUPER RUNS

$\dots$	b_{-4}	b_{-3}	b_{-2}	b_{-1}	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	$\dots$
$\dots$	1	0	0	1	0	0	0	1	1	0	1	0	0	$\dots$

Lemma II.2 allows us to restrict our attention to stationary ergodic processes in proving the converse. For a process $\mathbb{X}$, we denote by $H(\mathbb{X})$ its *entropy rate*. Define

$$H(Y_{\mathbb{X}}) \equiv \lim_{n\to\infty} \frac{H(Y(X^n))}{n(1-d)}\,. \tag{9}$$

A simple argument shows that this limit exists and is bounded above by 1 for any stationary process $\mathbb{X}$ and any d, with $H(Y_{\mathbb{X}}) = 1$ if and only if $\mathbb{X}$ is the iid Bernoulli(1/2) process.

In light of Lemma IV.1, we can restrict consideration to processes $\mathbb{X}$ satisfying $I(\mathbb{X}) > 1 - d^{1-\epsilon}$ whence $H(\mathbb{X}) > 1 - d^{1-\epsilon}, H(Y_{\mathbb{X}}) > 1 - d^{1-\epsilon}$:

Remark IV.2: There exists $d_0(\epsilon) > 0$ such that for all $d < d_0(\epsilon)$, if $I(\mathbb{X}) > C - d$, we have $I(\mathbb{X}) > 1 - d^{1-\epsilon}$ and hence also $H(\mathbb{X}) > 1 - d^{1-\epsilon},\ \ H(Y_{\mathbb{X}}) > 1 - d^{1-\epsilon}$.

We define a "super run" next.

Definition IV.3: A super run consists of a maximal contiguous sequence of runs such that all runs in the sequence after the first one (on the left) have length one. In other words, each super run is in one-to-one correspondence with a run of length 2 or larger. The super run includes that run plus (eventually) one or more contiguous runs of length one.

We divide a realization of $\mathbb{X}$ into super runs $\dots, S_{-2}, S_{-1}, S_0, S_1, S_2, \dots$. Here, S_1 is the super run including the bit at position 1.

See Table II for an example showing division into super runs.

Denote by $\mathcal{S}$ the set of all stationary ergodic processes and by $\mathcal{S}_{L^*}$ the set of stationary ergodic processes such that, with probability one, no super run has length larger than L^*.

Our next lemma tightens the constraint given by Remark IV.2 further for processes in $\mathcal{S}_{\lfloor 1/d \rfloor}$.

Lemma IV.4: Consider any $\epsilon > 0$ and constant κ. There exists $d_0(\epsilon, \kappa) > 0$ such that the following happens for any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$. For any $d < d_0$, if

$$I(\mathbb{X}) \geq C - \kappa d^{2-(\epsilon/2)},$$

then

$$H(Y_{\mathbb{X}}) \geq 1 - d^{2-\epsilon}\,.$$

We show an upper bound for the restricted class of processes $\mathcal{S}_{L^*}$.

Lemma IV.5: For any $\epsilon > 0$, there exists $d_0 = d_0(\epsilon) > 0$ and $\kappa < \infty$ such that the following happens. If $d < d_0(\epsilon)$, for any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$,

$$I(\mathbb{X}) \leq 1 + d\log d - A_1 d + A_2 d^2 + \kappa d^{3-\epsilon}\,.$$

Finally, we show a suitable reduction from the class $\mathcal{S}$ to the class $\mathcal{S}_{L^*}$.

Lemma IV.6: For any $\epsilon > 0$, there exists $d_0 = d_0(\epsilon) > 0$ such that the following happens for all $d < d_0$, and all $\gamma > 0$. For any $\mathbb{X} \in \mathcal{S}$ such that $H(Y_{\mathbb{X}}) > 1 - d^{\gamma}$ and for any $L^* > 2\gamma\log(1/d)$, there exists $\mathbb{X}_{L^*} \in \mathcal{S}_{L^*}$ such that

$$I(\mathbb{X}) \leq I(\mathbb{X}_{L^*}) + d^{\gamma-\epsilon}(L^*)^{-1}\log L^*, \tag{10}$$
$$H(Y_{\mathbb{X}}) \geq H(Y_{\mathbb{X}_{L^*}}) - d^{\gamma-\epsilon}(L^*)^{-1}\log L^*\,. \tag{11}$$

Lemmas IV.4, IV.5, and IV.6 are proved in Section V-F.

The proof of Theorem I.1 follows from these lemmas with Lemma IV.6 being used twice.

Proof of Theorem I.1: For the converse, we start with a process $\mathbb{X} \in \mathcal{S}$ such that $I(\mathbb{X}) > C - d^3$. By Remark IV.2, $H(Y_{\mathbb{X}}) > 1 - d^{1-\delta}$ for any $\delta > 0$ and $d < d_0(\delta)$. Use Lemma IV.6, with $\gamma = 1 - \delta$, $L^* = \lfloor 1/d \rfloor$ and $\epsilon = \delta/2$. It follows that for $d < d_0(\delta/2)$,

$$I(\mathbb{X}_{L^*}) > C - d^{2-2\delta},$$
$$H(Y_{\mathbb{X}}) \geq H(Y_{\mathbb{X}_{L^*}}) - d^{2-2\delta}\,.$$

We now use Lemma IV.4 on $\mathbb{X}_{L^*}$ which yields $H(Y_{\mathbb{X}_{L^*}}) \geq 1 - d^{2-2\delta}$, and hence, by (11), $H(Y_{\mathbb{X}}) \geq 1 - 2d^{2-2\delta} \geq 1 - d^{2-3\delta}$ for small d. Now, we can use Lemma IV.6 again with $\gamma = 2 - 3\delta$, $L^* = \lfloor 1/d \rfloor$, $\epsilon = \delta/2$. We obtain

$$I(\mathbb{X}_{L^*}) \geq C - d^{3-4\delta}\,.$$

Finally, using Lemma IV.5, we get the required upper bound on C. This completes the proof of the converse.

Constructing a codebook: As part of the proof of achievability in the channel coding theorem [5, Th. 1], Dobrushin in fact establishes (also using previous work [16]) that given a sequence of input distributions $(p_{X^n})_{n\geq 1}$ such that

$$\lim_{n\to\infty, X^n \sim p_{X^n}} I(X^n; Y(X^n)) = I_*,$$

the following is true. Given any $R < I_*$ and $\epsilon > 0$ for all n large enough, there exists a codebook $\mathcal{C} \subseteq \{0,1\}^n$ with $|\mathcal{C}| = 2^{nR}$, achieving error probability smaller than ϵ for every codeword under maximum likelihood decoding. Moreover, this codebook is constructed (see [16]) simply by letting $\mathcal{C} = \{x^{(1)}, \dots, x^{(M)}\}$, $M = 2^{nR}$, $x^{(\ell)} \in \{0,1\}^n$ where $x^{(\ell)} \sim p_{X^n}$ are independent. Now, we have constructed $\mathbb{X}^{\dagger}$ with the property

$$\lim_{n\to\infty} I((X^{\dagger})^n; Y((X^{\dagger})^n)) = I(\mathbb{X}^{\dagger}) > C - O(d^{3-\delta})\,.$$

Moreover, sampling $x \sim p_{(X^{\dagger})^n}$ is easy, and can be achieved as follows. First define

$$p^{\dagger}_{L_1^+}(l) = \frac{1}{Z}\sum_{l'=l}^{\infty} p^{\dagger}_L(l'), \tag{12}$$

for Z a normalization constant. Then, sample $l_1 \sim p^{\dagger}_{L_1^+}$ and set the first l_1 bits of x all to 0, or all to 1 with equal probability. Assume, to be definite, that these first bits were set to

0. Successively sample $l_2, l_3, \ldots, \sim p_L^{\dagger}$ and set $x_{l_1+1}^{l_1+l_2} = 1$, $x_{l_1+l_2+2}^{l_1+l_2+l_3} = 0$, and so on until n bits have been fixed.

The random codebook $\mathcal{C}$, thus, constructed achieves capacity up to $O(d^{3-\epsilon})$ under maximum likelihood decoding. ■

V. Proofs of the Lemmas

In Section V-A, we show that, for any stationary ergodic $\mathbb{X}$ that achieves a rate close to capacity, the run-length distribution must be close to the distributions obtained for the iid Bernoulli$(1/2)$ process. In Section V-B, we suitably rewrite the rate $I(\mathbb{X})$ achieved by stationary ergodic process $\mathbb{X}$ as the sum of three terms. In Section V-C, we construct a modified deletion process that allows accurate estimation of $H(Y|X^n)$ in the small d limit. Section V-D proves a key bound on $H(Y_{\mathbb{X}})$ that leads directly to Lemma IV.4. Finally, in Section V-F, we present proofs of the Lemmas quoted in Section IV using the tools developed.

We will often write X_a^b for the random vector $(X_a, X_{a+1}, \ldots, X_b)$ where the X_i's are distributed according to the process $\mathbb{X}$.

A. Characterization in Terms of Runs

Let the r.v. m_n be the number of runs in X^n. Let $L_1^+, L_2, \ldots, L_{m_n}$ be the run lengths (L_1^+ being the length of the intersection of the run containing X_1 with X^n). It is clear that $H(X^n) \leq 1 + H(m_n, L_1^+, L_2, \ldots, L_{m_n})$ (where one bit is needed to remove the $0,1$ ambiguity). By ergodicity, $m_n/n \to 1/\mathbb{E}[L]$ almost surely as $n \to \infty$. Also $m_n \leq n$ implies $H(m_n)/n \leq \log n/n \to 0$. Further, $\limsup_{n\to\infty} H(L_1^+, L_2, \ldots, L_{m_n})/n \leq \lim_{n\to\infty} H(L) m_n/n = H(L)/\mathbb{E}[L]$. If $H(\mathbb{X})$ is the entropy rate of the process $\mathbb{X}$, by taking the $n \to \infty$ limit, it is easy to deduce that

$$H(\mathbb{X}) \leq \frac{H(L)}{\mathbb{E}[L]}, \tag{13}$$

with equality if and only if $\mathbb{X}$ is a process with iid runs with common distribution p_L.

We know that given $\mathbb{E}[L] = \mu$, the probability distribution with largest possible entropy $H(L)$ is geometric with mean μ, i.e., $p_L(l) = (1 - 1/\mu)^{l-1} 1/\mu$ for all $l \geq 1$, leading to

$$\frac{H(L)}{\mathbb{E}[L]} \leq -\big(1 - \frac{1}{\mu}\big) \log\big(1 - \frac{1}{\mu}\big) - \frac{1}{\mu} \log \frac{1}{\mu} \equiv h(1/\mu)\,. \tag{14}$$

Here, we introduced the notation $h(p) = -p \log p - (1 - p)\log(1 - p)$ for the binary entropy function.

Using this, we are able to obtain sharp bounds on p_L and $\mu(\mathbb{X})$.

Lemma V.1: There exists $d_0 > 0$ such that the following occurs. For any $\beta > 1/2$ and $d < d_0$, if $\mathbb{X} \in \mathcal{S}$ is such that $H(\mathbb{X}) > 1 - d^{\beta}$, we have

$$|\mu(\mathbb{X}) - 2| \leq 3\, d^{\beta/2}\,. \tag{15}$$

Proof: By (13) and (14), we have $h(1/\mu) \geq 1 - d^{\beta}$. By Pinsker's inequality $h(p) \leq 1 - (1 - 2p)^2/(2 \ln 2)$, and therefore,

$$|1 - (2/\mu)|^2 \leq (2 \ln 2)\, d^{\beta}\,. \tag{16}$$

We deduce that for sufficiently small $d_0 > 0$, we have $|1 - (2/\mu)| < 0.1$ for all $d < d_0$. Here, we need $\beta > 1/2$ to obtain d_0 that does not depend on β, with $1/2$ being an arbitrarily chosen real number in the interval $(0, 1)$. It follows that $\mu < 2.3$ for $d < d_0$. Plugging back in (16), we have

$$|\mu - 2| \leq \mu\sqrt{2 \ln 2}\, d^{\beta/2} < 3\, d^{\beta/2} \tag{17}$$

for all $d < d_0$. ■

Lemma V.2: There exists $d_0 > 0$ and $\kappa' < \infty$ such that the following occurs for any $\beta > 1/2$ and $d < d_0$. For any $\mathbb{X} \in \mathcal{S}$ such that $H(\mathbb{X}) > 1 - d^{\beta}$, we have

$$\sum_{l=1}^{\infty} \left| p_L(l) - \frac{1}{2^l} \right| \leq \kappa' d^{\beta/2}\,. \tag{18}$$

Proof: Let $p_L^*(l) = 1/2^l,\ l \geq 1$ and recall that $\mu(\mathbb{X}) = \mathbb{E}[L] = \sum_{l\geq 1} p_L(l) l$. An explicit calculation yields

$$H(L) = \mu(\mathbb{X}) - D(p_L \| p_L^*)\,. \tag{19}$$

Now, by Pinsker's inequality,

$$D(p_L \| p_L^*) \geq \frac{2}{\ln 2} \| p_L - p_L^* \|_{\mathrm{TV}}^2\,. \tag{20}$$

Combining Lemma V.1, (13), (19), and (20), we get the desired result. ■

We now state a tighter bound on probabilities of large run lengths. We will find this useful, for instance, to control the number of bit flips in going from general $\mathbb{X}$ to $\mathbb{X}_{L^*}$ having bounded run lengths.

Lemma V.3: There exists $d_0 > 0$ such that the following occurs: Consider any $\beta > 1/2$, and define $\ell \equiv \lfloor 2\beta \log(1/d) \rfloor$. For all $d < d_0$, if $\mathbb{X} \in \mathcal{S}$ is such that $H(\mathbb{X}) > 1 - d^{\beta}$, we have

$$\sum_{l=\ell}^{\infty} l p_L(l) \ \leq 20 d^{\beta}\,. \tag{21}$$

Proof of Lemma V.3: Combining (13), Lemma V.1, and (19), it follows that for small enough d, we must have

$$D(p_L \| p_L^*) \leq 3 d^{\beta} \tag{22}$$

to achieve $H(\mathbb{X}) \geq 1 - d^{\beta}$. Now define $\Delta \equiv \sum_{l=\ell}^{\infty} l p_L(l)$. Take $\alpha = e^{3/5}$. We have

$$\sum_{l=\ell}^{\infty} \frac{l}{\alpha^l} = \frac{\ell \alpha^{-\ell}}{(1 - \alpha)^2} < d^{\beta}$$

for sufficiently small d, since $\alpha^{-\ell} \approx \exp\left\{\frac{6}{5}\beta \log d\right\}$. Thus,

$$\sum_{l=\ell}^{\infty} l(p_L(l) - \alpha^{-l}) \geq \Delta - d^\beta$$

$$\Rightarrow \qquad \sum_{l\in\mathcal{I}} l p_L(l) \geq \Delta - d^\beta$$

where $\mathcal{I} = \{l : l \geq \ell, p_L(l) \geq \alpha^{-l}\}$.

This yields,

$$\sum_{l\in\mathcal{I}} p_L(l) \log \frac{p_L(l)}{p_L^*(l)} \geq \sum_{l\in\mathcal{I}} l p_L(l) \log\frac{2}{\alpha} \geq \log(2/\alpha)(\Delta - d^\beta). \tag{23}$$

It remains to show that the sum of terms from outside $\mathcal{I}$ is not too small. It is easy to see that

$$\sum_{l\in\mathcal{I}} p_L(l) \leq \Delta/\ell \qquad \Rightarrow \qquad \sum_{l\notin\mathcal{I}} p_L(l) \geq 1 - \Delta/\ell. \tag{24}$$

With a fixed sum constraint on $(p_L(l), l \notin \mathcal{I})$, the smallest value of $\sum_{l\notin\mathcal{I}} p_L(l) \log \frac{p_L(l)}{p_L^*(l)}$ is achieved when

$$\frac{p_L(l)}{p_L^*(l)} = \kappa = \frac{\sum_{l\notin\mathcal{I}} p_L(l)}{\sum_{l\notin\mathcal{I}} 2^{-l}} \geq \sum_{l\notin\mathcal{I}} p_L(l) \qquad \forall l \notin \mathcal{I}. \tag{25}$$

Note that $\kappa \leq 1$. It follows from (25) that

$$\begin{aligned}\sum_{l\notin\mathcal{I}} p_L(l) \log \frac{p_L(l)}{p_L^*(l)} &\geq \sum_{l\notin\mathcal{I}} p_L(l) \log \kappa \\ &\geq \log \kappa \\ &\geq \log\Big(\sum_{l\notin\mathcal{I}} p_L(l)\Big). \end{aligned} \tag{26}$$

Further, from (24), for small d, we have

$$\log\Big(\sum_{l\notin\mathcal{I}} p_L(l)\Big) \geq -2\Delta/\ell \tag{27}$$

since we know that $\Delta \leq \mu(\mathbb{X}) \leq 3$, and hence, $\Delta/\ell \leq 1/10$. Combining (26) and (27), we have

$$\sum_{l\notin\mathcal{I}} p_L(l) \log \frac{p_L(l)}{p_L^*(l)} \geq -2\Delta/\ell. \tag{28}$$

The lemma follows by combining (23), (28), and $D(p_L \| p_L^*) \leq 3d^\beta$. ■

We use $L(k)$ to denote the vector of lengths $(L_1, L_2, \ldots, L_k)$ of a randomly selected block of k consecutive runs (a "k-block"). Formally, $(L_1, L_2, \ldots, L_k)$ is the vector of lengths of the first k runs starting from bit X_1, under the Palm measure $\mathbb{P}_1$ introduced in Section II.

Corollary V.4: There exists $d_0 > 0$ such that the following occurs: Consider any positive integer k and any $\beta > 1/2$, and define $\ell \equiv \lfloor 2\beta \log(1/d) \rfloor$. For all $d < d_0$, if $\mathbb{X} \in \mathcal{S}$ is such that $H(\mathbb{X}) > 1 - d^\beta$, we have

$$\sum_{l_1+\cdots+l_k \geq k\ell} (l_1 + \ldots + l_k) p_{L(k)}(l_1, \ldots, l_k) \leq 20k^2 d^\beta. \tag{29}$$

Proof of Corollary V.4: Clearly $L_1 + \cdots + L_k \geq k\ell$ occurs only if at least one of the L_i's is at least ℓ. Also, the distribution $p_{L(k)}$ has a marginal p_L for each individual L_i. We have

$$\begin{aligned} &\sum_{l_1+\cdots+l_k \geq k\ell} (l_1 + \cdots + l_k) p_{L(k)}(l_1, \ldots, l_k) \\ &\leq \sum_{i=1}^{k} \sum_{l_1+\cdots+l_k\geq k\ell} \mathbb{I}[l_i \text{ is the largest}]\, k l_i\, p_{L(k)}(l_1, \ldots, l_k) \\ &\leq \sum_{i=1}^{k} \sum_{l_i=\ell}^{\infty} k l_i\, p_L(l_i) \\ &= k^2 \sum_{l=\ell}^{\infty} l p_L(l). \end{aligned}$$

The result now follows from the first inequality in Lemma V.3. ■

Clearly, $\mathbb{E}[L_1 + \cdots + L_k] = k\mu(\mathbb{X})$. We have

$$H(\mathbb{X}) \leq \frac{H(L_1, L_2, \ldots, L_k)}{k\mu(\mathbb{X})}.$$

A stronger form of Lemma V.2 follows.

Lemma V.5: Let $p^*_{L(k)}(l_1, \ldots, l_k) \equiv 2^{-\sum_{i=1}^{k} l_i}$. For the same κ' and $d_0 > 0$ as in Lemma V.2, the following occurs. Consider any positive integer k and any $\beta > 1/2$. For all $d < d_0$, if $\mathbb{X} \in \mathcal{S}$ is such that $H(\mathbb{X}) > 1 - d^\beta$, we have

$$\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty}\cdots\sum_{l_k=1}^{\infty} \left| p_{L(k)}(l_1, \ldots, l_k) - p^*_{L(k)}(l_1, \ldots, l_k) \right| \leq \kappa' \sqrt{k}\, d^{\beta/2}.$$

Proof of Lemma V.5: Repeat proof of Lemma V.2. ■

We now relate the run-length distribution in $\mathbb{X}$ and in $Y(X^n)$ (as $n \to \infty$). For this, we first need a characterization of Y in terms of a stationary ergodic process. Let $\mathbb{D} = (\ldots, D_{-1}, D_0, D_1, D_2, \ldots)$ be an iid Bernoulli(d) process, independent of $\mathbb{X}$. Construct $\mathbb{Y}$ as follows. Look at $X_1, X_2, \ldots$. Delete bits X_i corresponding to $D_i = 1$ for $i = 1, 2, 3, \ldots$, the bits remaining are $Y_1, Y_2, \ldots$ in order. Similarly, in $X_0, X_{-1}, X_{-2}, \ldots$ delete bits X_i corresponding to $D_i = 1$ for $i = 0, -1, -2, \ldots$, the bits remaining are $Y_0, Y_{-1}, \ldots$ in order.

Proposition V.6: The process $\mathbb{Y}$ is stationary and ergodic for any stationary ergodic $\mathbb{X}$.

Proof of Proposition V.6: A time shift by a constant in $\mathbb{Y}$ corresponds to a time shift by a random amount in $\mathbb{X}$. The random shift in $\mathbb{X}$ depends only on the $\mathbb{D}$ and is, hence, independent of $\mathbb{X}$. Also, $\mathbb{D}$ is independent identically distributed. Thus, stationarity of $\mathbb{X}$ implies stationarity of $\mathbb{Y}$. ■

Notice on the other hand that $(\mathbb{X}, \mathbb{Y})$ are *not* jointly stationary.

The channel output $Y(X^n)$ is then $(\mathbb{Y})_1^M$ where $M \sim \text{Binomial}(n, 1-d)$. It is easy to check that

$$H(\mathbb{Y}) = H(Y_{\mathbb{X}})$$

[cf., (9)]. We will, henceforth, use $H(\mathbb{Y})$ instead of the more cumbersome notation $H(Y_{\mathbb{X}})$.

Let q_L denote the block perspective run-length distribution for $\mathbb{Y}$. Denote by $q_{L(k)}$ the block perspective distribution for k-blocks in $\mathbb{Y}$. Lemmas V.1–V.5 and Corollary V.4 hold for any stationary ergodic process, hence, they hold true if we replace $(\mathbb{X}, p)$ with $(\mathbb{Y}, q)$.

In proving the upper bound, it turns out that we are able to establish a bound of $H(\mathbb{Y}) > 1 - d^{2-\epsilon}$ for $\epsilon > 0$ and small d, but no corresponding bound for $H(\mathbb{X})$. Next, we establish that if $H(\mathbb{Y})$ is close to 1, this leads to tight control over the tail for $p_L(\,\cdot\,)$. This is a corollary of Lemma V.3.

Lemma V.7: There exists $d_0 > 0$ such that the following occurs: Consider any $\gamma > 1/2$, and define $\ell \equiv \lfloor 2\gamma \log(1/d) \rfloor$. For all $d < d_0$, if $H(\mathbb{Y}) \geq 1 - d^\gamma$, we have

$$\sum_{l=2\ell}^{\infty} l p_L(l) \ \leq 80 d^\gamma\,.$$

Note that p_L refers to the block length distribution of $\mathbb{X}$, not $\mathbb{Y}$.

Proof of Lemma V.7: Consider a run R of length $l \geq 2\ell$ in $\mathbb{X}$. With probability at least $(1-d)^2$, the runs bordering R do not disappear due to deletions. Independently, with probability $\mathbb{P}[\mathrm{Binomial}(l, 1-d) \geq l/2]$ at least half the bits of R survive deletion. Thus, for small d, with probability at least $1/2$, R leads to a run of length at least $l/2$ in $\mathbb{Y}$. Moreover, runs can only disappear in going from $\mathbb{X}$ to $\mathbb{Y}$. It follows that

$$\sum_{l=\ell}^{\infty} l q_L(l) \ \geq \sum_{l=2\ell}^{\infty} \left(\frac{l}{2}\right)\left(\frac{p_L(l)}{2}\right)\,.$$

From Lemma V.3 applied to $\mathbb{Y}$, we know that

$$\sum_{l=\ell}^{\infty} l q_L(l) \ \leq 20 d^\beta\,.$$

The result follows. ■

Corollary V.8: There exists $d_0 > 0$ such that the following occurs: Consider any positive integer k and $\gamma > 1/2$, and define $\ell \equiv \lfloor 2\gamma \log(1/d) \rfloor$. For all $d < d_0$, if $H(\mathbb{Y}) \geq 1 - d^\gamma$, we have

$$\sum_{l=2k\ell}^{\infty} (l_1 + \cdots + l_k) p_{L(k)}(l_1, \ldots, l_k) \ \leq 80 k^2 d^\gamma\,.$$

Proof of Corollary V.8: Analogous to proof of Corollary V.4. ■

Consider $\mathbb{X}$ being iid Bernoulli$(1/2)$. Clearly, this corresponds to $\mathbb{Y}$ also iid Bernoulli$(1/2)$. Hence, each has the same run-length distribution $p_L^*(l) = q_L^*(l) = 2^{-l}$. This happens irrespective of the deletion probability d. Now suppose $\mathbb{X}$ is not iid Bernoulli$(1/2)$ but approximately so, in the sense that $H(\mathbb{X})$ close to 1. The next lemma establishes, that in this case also, the run-length distribution of $\mathbb{Y}$ is very close to that of $\mathbb{X}$, for small run lengths and small d.

Lemma V.9: There exists a function $(\kappa, \epsilon) \mapsto d_0(\kappa, \epsilon) > 0$ and constants $\kappa_1 < \infty$, $\kappa_2 < \infty$ such that the following happens, for any $\beta \in (1/2, 2)$, $\epsilon > 0$ and $\kappa < \infty$.

i) For all $d < d_0$, for all $\mathbb{X}$ such that $H(\mathbb{X}) > 1 - d^\beta$, and all $l < \kappa \log(1/d)$, we have

$$|p_L(l) - q_L(l)| \leq \kappa_1 d^{1+\beta/2-\epsilon}\,.$$

ii) For all $d < d_0$ and all $\mathbb{X}$ such that $H(\mathbb{X}) > 1 - d^\beta$, we have

$$|\mu(\mathbb{X}) - \mu(\mathbb{Y})| \leq \kappa_2 d^{1+\beta/2}\,. \tag{30}$$

Proof of Lemma V.9: We adopt two conventions. First, when we use the $O(\cdot)$ or the $\Omega(\cdot)$ notation, the constant involved does not depend on the particular $\mathbb{X}, \mathbb{Y}$ under consideration. Second, we use "typical" in this proof to refer to events having a probability $\Omega(d^{2-\delta})$, for some $\delta > 0$. Thus, an event with probability $2d^2$ is not typical, but an event with probability $d^{1.5}$ is typical.

We ignore boundary effects due to runs at the beginning and end.

First, we estimate the factor due to disappearance of runs in moving from $\mathbb{X}$ in $\mathbb{Y}$. Define

$$r(\mathbb{X}) \equiv \lim_{n \to \infty} \frac{\text{Number of runs in } Y(X^n)}{\text{Number of runs in } X^n}.$$

We have almost sure convergence of this ratio to a constant value due to ergodicity.

Runs disappear typically due to runs of length 1 being deleted, and the runs at each end being fused with each other (i.e., neither of them is deleted). Such an event reduces the number of runs by 2. Nontypical run deletions lead to a correction factor that is $O(d^2)$. Hence, the expected number of runs in Y per run in X^n is $1 - 2p_L(1)d + O(d^2)$. It follows from a limiting argument that

$$r = 1 - 2p_L(1)d + O(d^2). \tag{31}$$

In this proof, we make use of the following implication of Lemma V.5.

$$\left| p_{L(k)}(l_1, \ldots, l_k) - 2^{-\sum_{i=1}^{k} l_i} \right| \leq \kappa' \sqrt{k} d^{\beta/2}. \tag{32}$$

We immediately have $p_L(1) = 1/2 + O(d^{\beta/2})$, and hence, $r = 1 - d + O(d^{1+\beta/2})$.

Consider $q_L(1)$. Blocks of length 1 in Y typically arise due to blocks in $\mathbb{X}$ of length 1 or 2. In case of a block of length 1, we require that it is not deleted, and also that bordering blocks are not deleted. Consider a randomly selected run in $\mathbb{X}$ (Formally, we pick a run uniformly at random in X^n, and then, take the limit $n \to \infty$). The run has length $L = 1$ with probability $p_L(1)$. Define

1) $\mathsf{E}_1 \equiv$ no bordering block of length 1. We have $\mathbb{P}[\mathsf{E}_1, L = 1] = (1/8) + O(d^{\beta/2})$;
2) $\mathsf{E}_2 \equiv$ one bordering block of length 1. We have $\mathbb{P}[\mathsf{E}_2, L = 1] = (1/4) + O(d^{\beta/2})$;
3) $\mathsf{E}_3 \equiv$ two bordering blocks of length 1. We have $\mathbb{P}[\mathsf{E}_3, L = 1] = (1/8) + O(d^{\beta/2})$.

Probabilities were estimated using $p_L(1) = 1/2 + O(d^{\beta/2})$, $p_{L(2)}(1,1) = 1/4 + O(d^{\beta/2})$ and $p_{L(3)}(1,1,1) = 1/8 + O(d^{\beta/2})$, and their immediate consequences $p_{L(3)}(1,1,>1) = 1/8 + O(d^{\beta/2})$, $p_{L(3)}(>1,1,1) = 1/8 + O(d^{\beta/2})$ and $p_{L(3)}(>1,1,>1) = 1/8 + O(d^{\beta/2})$. We made use of (32).

Probability of arising from block of length 1 is

$$\begin{aligned}(1-d)\Big\{&\mathbb{P}[\mathsf{E}_1, L=1](1-O(d^2))+\\ &\mathbb{P}[\mathsf{E}_2, L=1](1-d)(1-O(d^2))+\\ &\mathbb{P}[\mathsf{E}_3, L=1](1-d)^2\Big\}\\ =p_L(1)&(1-2d)+O(d^{1+\beta/2}).\end{aligned}$$

Probability of arising from a block of length 2 is $p_L(2)2d + O(d^2) = d/2 + O(d^{1+\beta/2})$, using (32). It follows that

$$\begin{aligned}q_L(1) &= \frac{p_L(1)(1-2d)+d/2+O(d^{1+\beta/2})}{r}\\ &= p_L(1)+O(d^{1+\beta/2})\end{aligned}$$

as required.

Now consider $q_L(l)$ for $1 < l < \kappa\log(1/d)$. Typical modes of creation of such a run in $\mathbb{Y}$ are:

1) run of length l in $\mathbb{X}$ that goes through unchanged;
2) two runs in $\mathbb{X}$ being fused due to the length 1 run between them being deleted. Fused runs have no deletions. They have l bits in total;
3) run of length $l+1$ in $\mathbb{X}$ that suffers exactly one deletion. Bordering runs do not disappear.

For mode 1, we define events $\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3$ as aforementioned. Probability estimates are:

1) $\mathbb{P}[\mathsf{E}_1, L=l] = 2^{-l-2} + O(d^{\beta/2})$,
2) $\mathbb{P}[\mathsf{E}_2, L=l] = 2^{-l-1} + O(d^{\beta/2})$,
3) $\mathbb{P}[\mathsf{E}_3, L=l] = 2^{-l-2} + O(d^{\beta/2})$,

using (32) as we did for $L=1$. Thus, probability of creation from randomly selected run via mode 1 is

$$\begin{aligned}(1-d)^l\big\{&\mathbb{P}[\mathsf{E}_1, L=l](1-O(d^2))+\\ &\mathbb{P}[\mathsf{E}_2, L=l](1-d)(1-O(d^2))+\\ &\mathbb{P}[\mathsf{E}_3, L=l](1-d)^2\big\}\\ =p_L(l)&-2^{-l}(l+1)d+O(d^{1+\beta/2-\epsilon})\end{aligned}$$

for any $\epsilon > 0$, since $l < \kappa\log(1/d)$.

The probability of a random set of three consecutive runs being such that the middle run has length 1 and bordering runs have total length l is $(l-1)2^{-l-1} + O(d^{\beta/2-\epsilon})$, using (32) and $l < \kappa\log(1/d) < d^{-\epsilon}$ for small enough d. Probability of the middle run being deleted and the other two runs being left intact, along with bordering runs of this set of three runs not being deleted, is $d + O(ld^2)$. Thus, probability of creation via mode 2 is $(l-1)2^{-l-1}d + O(d^{1+\beta/2-\epsilon})$.

It is easy to check that the probability of mode 3 working on a randomly selected run is $(l+1)\,2^{-l-1}d + O(d^{1+\beta/2})$.

Combining, we have

$$\begin{aligned}q_L(l) = r^{-1}\Big\{&p_L(l) - 2^{-l}(l+1)d + (l-1)2^{-l-1}d +\\ &(l+1)\,2^{-l-1}d + O(d^{1+\beta/2-\epsilon})\Big\}\\ = p_L(l) &+ O(d^{1+\beta/2-\epsilon}).\end{aligned}$$

This completes the proof of (i).

For (ii), simply note that

$$\frac{\mu(\mathbb{X})}{\mu(\mathbb{Y})} = r(\mathbb{X})\left[\lim_{n\to\infty}\frac{n}{\text{Length of } Y(X^n)}\right] = \frac{r(\mathbb{X})}{1-d}.$$

It follows from (31) that

$$|\mu(\mathbb{X}) - \mu(\mathbb{Y})| \le 4\big|p_L(1) - 1/2\big|d + \kappa_3 d^2 \tag{33}$$

for some $\kappa_3 < \infty$. Equation (30) follows using Lemma V.2 to bound $p_L(1)$. ■

Let us emphasize that κ_1 and κ_2 do not depend at all on β, ϵ, κ, whereas d_0 does not depend on β in the aforementioned lemma. Analogous comments apply to the remaining lemmas in this section.

As before, we are able to generalize this result to blocks of k consecutive runs.

Lemma V.10: There exist a function $(\kappa, \epsilon) \mapsto d_0(\kappa, \epsilon) > 0$ and a constant $\kappa < \infty$ such that the following happens, for any $\beta \in (1/2, 2)$, $\epsilon > 0$ and $\kappa < \infty$.

For all $d < d_0$, for all integers $k > 0$ and $(l_1, l_2, \ldots, l_k)$ such that $\sum_{i=1}^k l_i < \kappa\log(1/d)$, and all $\mathbb{X}$ such that $H(\mathbb{X}) > 1 - d^\beta$, we have

$$|p_{L(k)}(l_1, \ldots, l_k) - q_{L(k)}(l_1, \ldots, l_k)| \le \kappa'\, d^{1+\beta/2-\epsilon}\,.$$

Proof of Lemma V.10: Similar to proof of Lemma V.9(i). We use (32) again, and make use of $k \le \sum_{i=1}^k l_i \le \kappa\log(1/d)$ to deduce that $\sqrt{k+2} \le d^{-\epsilon/2}$ for small enough d. ■

In proving the lower bound, we have $H(\mathbb{X}^\dagger) = 1 - O(d^2)$, but no corresponding bound for $H(\mathbb{Y})$. The next lemma allows us to get tight control over the tail of $q_L^\dagger(\cdot)$.

Lemma V.11: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$ and $\kappa < \infty$ such that the following occurs: Consider any $\beta \in (1/2, 2]$, and define $\ell \equiv \lfloor 4\log(1/d)\rfloor$. For all $d < d_0$, if $H(\mathbb{X}) \ge 1 - d^\beta$, we have

$$\sum_{l=\ell}^{\infty} lq_L(l) \;\le \kappa d^{\beta-\epsilon}\,.$$

Proof of Lemma V.11: From Lemma V.9(ii), we know that

$$\left|\sum_{l=1}^{\infty} lp_L(l) - \sum_{l=1}^{\infty} lq_L(l)\right| \le \kappa_1 d^{1+\beta/2}. \tag{34}$$

Recall $\ell \equiv \lfloor 4\log(1/d)\rfloor$. Using Lemma V.9(i), we deduce

$$\left|\sum_{l=1}^{\ell-1} lp_L(l) - \sum_{l=1}^{\ell-1} lq_L(l)\right| \le \kappa_2 d^{1+\beta/2-\epsilon/2}. \tag{35}$$

From Lemma V.3, we know that

$$\sum_{l=\ell}^{\infty} lp_L(l) \le \kappa_3 d^\beta. \tag{36}$$

Note that $\kappa_1, \kappa_2,$ and κ_3 do not depend on β.

Combining (34)–(36), and using $\beta \le 2$, we arrive at the desired result. ■

Define $p^*_{L(k)}(l_1,\dots,l_k) \equiv 2^{-\sum_{i=1}^k l_i}$. We show, using Lemma V.10, that if $H(\mathbb{Y})$ is close to 1, then one can bound the distance between $p_{L(k)}(\,\cdot\,)$ and $p^*_{L(k)}(\,\cdot\,)$.

Lemma V.12: There exists a function $(\kappa,\epsilon) \mapsto d_0(\kappa,\epsilon) > 0$ and constants $\kappa_1 < \infty$, $\kappa_2 < \infty$ such that the following happens, for any $\epsilon > 0$, $\kappa < \infty$ and $\gamma \in (1/2, 2)$.

i) For all $d < d_0$, all sources $\mathbb{X}$ such that $H(\mathbb{X}) > 1 - d^{0.6}$ and $H(\mathbb{Y}) > 1 - d^{\gamma}$, and all integers $k > 0$ and $(l_1, l_2, \dots, l_k)$ such that $\sum_{i=1}^k l_i < \kappa \log(1/d)$, we have

$$|p_{L(k)}(l_1,\dots,l_k) - p^*_{L(k)}(l_1,\dots,l_k)| \le d^{\gamma/2-\epsilon}, \tag{37}$$

$$|p_{L(k)}(l_1,\dots,l_k) - q_{L(k)}(l_1,\dots,l_k)| \le d^{1+\gamma/2-\epsilon}. \tag{38}$$

ii) For all $d < d_0$, all sources $\mathbb{X}$ such that $H(\mathbb{X}) > 1 - d^{0.6}$ and $H(\mathbb{Y}) > 1 - d^{\gamma}$, we have

$$|\mu(\mathbb{X}) - 2| \le \kappa_1 d^{\gamma/2}, \tag{39}$$

$$|\mu(\mathbb{X}) - \mu(\mathbb{Y})| \le \kappa_2 d^{1+\gamma/2}. \tag{40}$$

Proof of Lemma V.12: By Lemma V.5 applied to $\mathbb{Y}$, we know that

$$\sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty}\cdots\sum_{l_k=1}^{\infty}\Big|q_{L(k)}(l_1,l_2,\dots,l_k) - p^*_{L(k)}(l_1,\dots,l_k)\Big| \le \kappa_5\sqrt{k}\,d^{\gamma/2}.$$

Using Lemma V.10, we have for $d < d_0(\kappa,\gamma)$, for any integer k and $(l_1,\dots,l_k)$ such that $\sum_{i=1}^k l_i < \kappa\log(1/d)$.

$$\big|p_{L(k)}(l_1,l_2,\dots,l_k) - q_{L(k)}(l_1,l_2,\dots,l_k)\big| \le \kappa_6\, d.$$

Thus, we obtain (37), using $k < \kappa\log(1/d) < d^{-\epsilon}$ for small d. Also, note that we can deduce

$$|p_L(1) - p^*_L(1)| \le 2\kappa_5 d^{\gamma/2} \tag{41}$$

for small enough d. We repeat the proof of Lemma V.9(i) (or Lemma V.10), using (37) instead of (32) to obtain (38). This completes the proof of (i).

For (ii), we proceed as follows to prove (39) and (40). In the proof of Lemma V.9(ii), we deduced that $|\mu(\mathbb{X}) - \mu(\mathbb{Y})| \le 4\big|p_L(1) - 1/2\big|d + \kappa_7 d^2$ (this is (33) with the constant renamed). Using (41) to bound $p_L(1)$, we obtain (40). From Lemma V.1 applied to $H(\mathbb{Y})$, we know that $|\mu(\mathbb{Y}) - 2| \le 7d^{\gamma/2}$. Equation (39) follows. ■

The next lemma assures us that if $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d\rfloor}$, then very few runs in $\mathbb{Y}$ are much longer than $\lfloor 1/d\rfloor$. In fact, we show that $q_L(\lambda\lfloor 1/d\rfloor)$ decays exponentially in λ.

Lemma V.13: There exists $d_0 > 0$ such that, for all $d < d_0$, the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d\rfloor}$ such that $H(\mathbb{X}) > 1 - d^{2/3}$. Then, for all $\lambda > 2$ such that the product $\lambda\lfloor 1/d\rfloor$ is an integer, we have

$$q_L(\lambda\lfloor 1/d\rfloor) \le d^{\lambda-2}.$$

Proof of Lemma V.13: Associate each run in $\mathbb{Y}$ with the run in $\mathbb{X}$ from which its first bit came. Consider any run R_P in $\mathbb{X}$. If it gives rise to a run in $\mathbb{Y}$ of length $\lambda\lfloor 1/d\rfloor$, then we know that the runs $R_{P+1}, R_{P+3}, \dots, R_{P+2\lfloor\lambda-0.1\rfloor-1}$ were all deleted (since $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d\rfloor}$). This occurs with probability at most $d^{\lfloor\lambda-0.1\rfloor}$. Further, for each run in $\mathbb{X}$, there are $\mu(\mathbb{X})(1-d)/\mu(\mathbb{Y})$. This implies

$$q_L(\lambda\lfloor 1/d\rfloor) \le \frac{\mu(\mathbb{Y})}{\mu(\mathbb{X})(1-d)}d^{\lfloor\lambda-0.1\rfloor}.$$

From Lemmas V.1 and V.9(ii), we know that $|\mu(\mathbb{X}) - 2| < 0.1$ and $|\mu(\mathbb{Y}) - 2| < 0.1$ for small enough d. Plugging into the aforementioned equation yields the desired result. ■

Next, we prove some analogous results for super runs, cf., Definition IV.3, that we also need.

We denote by $\widetilde{L}^{\rm rep}$ the length of the first run in a random super run and by $\widetilde{L}^{\rm alt}$ the total length of the remaining runs of the same super run. More precisely, we repeat here the construction of Section II, and define a new Palm measure, $\mathbb{P}_{s1}$, which is the measure of $\mathbb{X}$ conditional on X_1 being the first bit of a super run. Then, $\widetilde{L}^{\rm rep}$ the length of the first run of this super run, and $\widetilde{L}^{\rm alt}$ is the residual length of the same super run, always under the Palm measure $\mathbb{P}_{s1}$. Here, "rep" indicates "repeated" with $\widetilde{L}^{\rm rep}$ being the number of repeated bits and "alt" indicates "alternating" with $\widetilde{L}^{\rm alt}$ being the number of alternating bits. We denote the type of a random super run by $\widetilde{T} \equiv (\widetilde{L}^{\rm rep}, \widetilde{L}^{\rm alt})$ and the length by $\widetilde{L} \equiv \widetilde{L}^{\rm alt} + \widetilde{L}^{\rm rep}$. We need versions of Lemmas V.3 and V.7 for super runs.

Define $\widetilde{\mu}(\mathbb{X}) \equiv \mathbb{E}[\widetilde{L}]$. It is easy to see that

$$H(\mathbb{X}) \le \frac{H(\widetilde{T})}{\widetilde{\mu}(\mathbb{X})}. \tag{42}$$

We denote by $p_{\widetilde{T}}$ the distribution of $\widetilde{T}$. Define $p^*_{\widetilde{T}}(l_1,l_2) \equiv 2^{-l_1-l_2}$, this being the distribution for the iid Bernoulli(1/2) process $\mathbb{X}^*$. We denote by $p_{\widetilde{L}}$ the distribution of $\widetilde{L}$ in $\mathbb{X}$. Clearly,

$$p_{\widetilde{L}}(l) = \sum_{l^{\rm rep}=2}^{l} p_{\widetilde{T}}(l^{\rm rep}, l - l^{\rm rep}).$$

Lemma V.14: There exists $d_0 > 0$ such that the following occurs. For any $\beta > 1/2$ and $d < d_0$, if $\mathbb{X} \in \mathcal{S}$ is such that $H(\mathbb{X}) > 1 - d^{\beta}$, we have

$$|\widetilde{\mu}(\mathbb{X}) - 4| \le 4\,d^{\beta/2}.$$

Proof of Lemma V.14: We make use of (42). Maximizing $H(\widetilde{T})$ for fixed $\widetilde{\mu}$, it is not hard to deduce that

$$\frac{H(\widetilde{T})}{\widetilde{\mu}} \le f(\widetilde{\mu}) \tag{43}$$

$$\text{where } f(x) \equiv -\frac{2}{x} - \left(1 - \frac{2}{x}\right)\log(x-2) + \log x$$

with equality if and only if $\mathbb{X}$ consists of iid super runs with $p_{\widetilde{T}}(l^{\rm rep}, l - l^{\rm rep}) = (\lambda-1)^2\lambda^{-l}$ where $\lambda = \widetilde{\mu}/(\widetilde{\mu}-2)$. Now, using (42), $H(\mathbb{X}) \le H(\widetilde{T})/\widetilde{\mu}$, and (43), we know that we must have $f(\widetilde{\mu}) \ge 1 - d^{\beta}$. Now, we have $f(4) = 1$. Further, it is easy to check that $f(\cdot)$ achieves its unique global and local maximum at 4, increasing monotonically before that and decreasing monotonically after that. It follows that for any fixed $\epsilon > 0$, for small enough d, we must have $|\widetilde{\mu} - 4| \le \epsilon$. It then follows from

Taylor's theorem that $f(\widetilde{\mu}) \le 1-(\widetilde{\mu}-4)^2/15$, so that we must have $|\widetilde{\mu}-4| \le 4d^{\beta/2}$ for $d \le d_0$, where $d_0 > 0$. ■

Lemma V.15: There exists $d_0 > 0$ such that the following occurs: Consider any $\beta > 1/2$, and define $\ell \equiv \lfloor 2\beta \log(1/d) \rfloor$. For all $d < d_0$, if $\mathbb{X} \in \mathcal{S}$ is such that $H(\mathbb{X}) > 1 - d^{\beta}$, we have

$$\sum_{l=\ell}^{\infty} l p_{\widetilde{L}}(l) \le 40 d^{\beta} .$$

Proof of Lemma V.15: An explicit calculation yields

$$H(\widetilde{T}) = \widetilde{\mu}(\mathbb{X}) - D(p_{\widetilde{T}} \| p^*_{\widetilde{T}}).$$

The proof now mirrors the proof of Lemma V.3, making use of Lemma V.14 in place of Lemma V.1. ■

Let $q_{\widetilde{L}}(\cdot)$ the distribution of super-run lengths in $\mathbb{Y}$, and $\widetilde{\mu}(\mathbb{Y})$ denote the mean length of a super run in $\mathbb{Y}$.

Lemma V.16: There exists $d_0 > 0$ such that the following occurs: Consider any $\gamma > 1/2$, and define $\ell \equiv \lfloor 2\gamma \log(1/d) \rfloor$. For all $d < d_0$, if $H(\mathbb{X}) \ge 1 - d^{0.6}$ and $H(\mathbb{Y}) \ge 1 - d^{\gamma}$, we have

$$\sum_{l=\ell}^{\infty} l p_{\widetilde{L}}(l) \le 80 d^{\gamma} .$$

Note that $p_{\widetilde{L}}$ refers to the super-run-length distribution of $\mathbb{X}$, not $\mathbb{Y}$.

Proof of Lemma V.16: It is easy to see that $f_{\mathbb{X}} = \sum_{l=\ell}^{\infty} l p_{\widetilde{L}}(l)/\widetilde{\mu}(\mathbb{X})$ is the asymptotic fraction of bits in $\mathbb{X}$ that are part of super runs of length at least ℓ. Similarly, $f_{\mathbb{Y}} = \sum_{l=\ell}^{\infty} l q_{\widetilde{L}}(l)/\widetilde{\mu}(\mathbb{Y})$ is the asymptotic fraction of bits in $\mathbb{Y}$ that are part of super runs of length at least ℓ.

We argue that $f_{\mathbb{Y}} \ge 0.9 f_{\mathbb{X}}$. Consider any bit b_P at position P in $\mathbb{X}$ that is part of a super run S_i with length $\widetilde{L}_i \ge \ell$. Consider a contiguous substring of S_i that includes b_P of length exactly ℓ. Clearly such a substring exists. The probability that it does not undergo any deletion is at least $1 - \ell d \ge 0.9$ for small enough d. Further, if this substring does not undergo any deletion, then all bits in this substring are part of the same super run in $\mathbb{Y}$, which must, therefore, have length at least ℓ. It follows that bit b_P is part of a super run of length at least ℓ in $\mathbb{Y}$ with probability at least 0.9. Thus, we have proved $f_{\mathbb{Y}} \ge 0.9 f_{\mathbb{X}}$. From Lemma V.14, it follows that $\widetilde{\mu}(\mathbb{X}) \le 5$ and $\widetilde{\mu}(\mathbb{Y}) \ge 3$ for small enough d. Putting these facts together leads to the result.

$$\sum_{l=\ell}^{\infty} l p_{\widetilde{L}}(l) \le 5 f_{\mathbb{X}} \le 5 f_{\mathbb{Y}}/0.9 \le \frac{5}{0.9 \cdot 3} \sum_{l=\ell}^{\infty} l q_{\widetilde{L}}(l) \le 80 d^{\gamma},$$

where we have made use of Lemma V.15 applied to $\mathbb{Y}$. ■

Corollary V.17: There exists $d_0 > 0$ such that the following occurs: Consider any positive integer k, any $\gamma > 1/2$, and define $\ell \equiv \lfloor 2\gamma \log(1/d) \rfloor$. For all $d < d_0$, if $H(\mathbb{X}) \ge 1 - d^{0.6}$ and $H(\mathbb{Y}) \ge 1 - d^{\gamma}$, we have

$$\sum_{l_1+\cdots+l_k \ge k\ell}^{\infty} (l_1 + \ldots + l_k) p_{\widetilde{L}(k)}(l_1, \ldots, l_k) \le 80 k^2 d^{\gamma} .$$

Proof of Corollary V.17: Analogous to proof of Corollary V.4. ■

TABLE III
PROCEDURE FOR GENERATING $Y(1), Y(2), \ldots, Y(M)$ AND $X(1), X(2), \ldots, X(M)$ GIVEN X^n AND $Y(X^n)$
(ADAPTED FROM [6, Fig. 1])

```
Set $X(1) = Y(1) =$the empty string.
$j \leftarrow 1$
For $i = 1$ to $m$ do
    $\sigma \leftarrow b(i)^{L_i}$
    $\omega \leftarrow$ the bits in $Y$ that arise from $i$th run in $X$
        % $\sigma$ is a (possibly empty) string of all $b(i)$'s.
        % $Y(j)$ is a (possibly empty) string of all $b(j)$'s.
    If $b(i) = b(j)$ or $|\omega| = 0$ then
        % $\omega$ is contained in the current block $Y(j)$ of $Y$
        $Y(j) \leftarrow Y(j)\omega$
        $X(j) \leftarrow X(j)\sigma$
    Else % $\omega$ is a prefix of $Y(j+1)$
        $j \leftarrow j+1$
        $Y(j) \leftarrow Y(j)\omega$
        $X(j) \leftarrow X(j)\sigma$
    End If
End For
```

B. Rate Achieved by a Process

We make use of an approach similar to that of Kirsch and Drinea [6] to evaluate $I(\mathbb{X})$ for a stationary ergodic process $\mathbb{X}$ that may be used to generate an input for the deletion channel. A fundamental difference is that [6] only considers processes with iid runs. Our analysis is instead general. This enables us to obtain tight upper and lower bounds (up to $O(d^{3-\epsilon})$), hence, leading to an estimate for the channel capacity.

We depart from the notation of Kirsch and Drinea, retaining X_i for the ith bit of X, and using $Y(j)$ to denote the jth run in $Y(X^n)$. Denote by $L_1, L_2, \ldots, L_m$ the lengths of runs in X_1^n (where m is a nondecreasing function of n for any fixed X_1^{∞}). Let the ith run consist of $b(i)$'s, where $b(i) \in \{0,1\}$. For instance, if the first run consists of 0s, then $b(i) = i+1 \pmod 2$.

We use $X(j)$ to denote the concatenation of runs in X that led to $Y(j)$, with the first run in $X(j)$ contributing at least one bit [if the run is completely deleted, then it is part of $X(j-1)$]. $X(1)$ is an exception. This is made precise in Table III, which is essentially the same as [6, Fig. 1], barring changes in notation. We call runs in $X(j)$ the *parent runs* of the run $Y(j)$.

We define $K(X^n)$ as the vector of $|X(j)|$, where $|\cdot|$ denotes the number of bits. Let the total number of runs in $Y(X^n)$ be M. Thus,

$$\begin{aligned} Y(X^n) &= Y(1) \ldots Y(M-1) Y(M), \\ X^n &= X(1) \ldots X(M-1) X(M), \\ K(X^n) &= (|X(1)|, \ldots, |X(M-1)|, |X(M)|) . \end{aligned}$$

Note that $X(j)$ consists of an odd number of runs for $1 < j < M$.

We write

$$I(X^n; Y(X^n)) = H(Y) - H(Y, K|X^n) + H(K|X^n, Y), \tag{44}$$

which is analogous to the identity $I(X^n; Y(X^n)) = H(X^n) - H(X^n, K|Y) + H(K|X^n, Y)$ used in [6], but more convenient for our proof.

Let $L_{\mathbb{Y}}$ be an integer random variable having the distribution q_L, i.e., the distribution of run length in $\mathbb{Y}$. It is easy to see that

$$\lim_{n\to\infty} \frac{H(Y(X^n))}{n(1-d)} = H(\mathbb{Y}) \le \frac{H(L_{\mathbb{Y}})}{\mu(\mathbb{Y})}$$

holds, similar to (13). It turns out that this suffices for our upper bound (cf., Lemma IV.4).

Consider the second term in (44). Let D^n denote the n-bit binary vector that indicates which bit locations in X^n have suffered deletions. We have

$$\begin{aligned} H(Y,K|X^n) &= H(D^n|X^n) - H(D^n|X^n,Y,K) \\ &= nh(d) - H(D^n|X^n,Y,K)\,. \end{aligned} \tag{45}$$

We study $H(D^n|X^n,Y,K)$ by constructing an appropriate modified deletion process in Section V-C

Consider the third term in (44). From [6], we know that

$$\begin{aligned} &\lim_{n\to\infty} \frac{H(K|X^n,Y)}{n} \\ &= \frac{\lim_{n\to\infty} H(\,|X(2)|\,|X(2)\dots X(M),Y(2)\dots Y(M))}{\mathbb{E}[|X(2)|]}\,. \end{aligned}$$

Here, $X(2)\dots X(M)$ denotes the string obtained by concatenating $X(2),\dots,X(M)$, without separation marks, and analogously for $Y(2)\dots Y(M)$. Roughly, single deletions do not lead to ambiguity in $|X(2)|$ if $X(2)\dots$ and $Y(2)\dots$ are known. Thus, this term is $O(d^2)$. It turns out, we can get a good estimate for this term by computing it for the iid Bernoulli$(1/2)$ case.

Lemma V.18: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$, and $\kappa < \infty$ such that for all $d < d_0$ the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$ such that $H(\mathbb{X}) > 1 - d^{1-\epsilon}$ and $\max\{H(\mathbb{X}), H(\mathbb{Y})\} > 1 - d^{\gamma}$ for some $\gamma \in (1/2, 2)$. Then

$$\left| \lim_{n\to\infty} \frac{1}{n} H(K(X^n)|X^n, Y(X^n)) - d^2 c_4 \right| \le \kappa d^{1+\gamma-\epsilon/2}, \tag{46}$$

where

$$\begin{aligned} c_4 \equiv &\sum_{j=4}^{\infty} 2^{-(2+j)}\,(j-1)(j-3)\,h\!\left(\frac{1}{j-1}\right) \\ &+ \sum_{i=2}^{\infty}\sum_{j=4}^{\infty} 2^{-(i+j+1)}\,(i+j-1)(j-3)\,h\!\left(\frac{i+1}{i+j-1}\right). \end{aligned}$$

Note that with $\gamma = 2 - \epsilon/2$, we obtain $|\delta| \le \kappa d^{3-\epsilon}$.

The proof of Lemma V.18 is quite technical and uses a so-called "perturbed" deletion process[1] (cf., Section V-C). We defer it to Appendix.

Lemma V.19: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$ such that if $H(\mathbb{X}) > 1 - d^{0.6}$ and $H(\mathbb{Y}) \ge 1 - d^{2-\epsilon/2}$, then

$$H(\mathbb{Y}) \le 1 - \frac{1}{2}\sum_{l=1}^{\infty} q_L(l)\big(\log q_L(l) + l\big) + d^{3-\epsilon},$$

for all $d < d_0$.

The proof of this lemma is fairly straightforward.

Proof of Lemma V.19: An explicit calculation yields $H(q_L) = \mu(\mathbb{Y}) - D(q_L\|q_L^*)$ where q_L^* is the run-length distribution corresponding to the iid Bernoulli$(1/2)$ half process (cf., proof of Lemma V.2). We know $H(\mathbb{Y}) \le H(q_L)/\mu(\mathbb{Y})$. It follows that

$$H(\mathbb{Y}) \le 1 - D(q_L\|q_L^*)/\mu(\mathbb{Y})\,. \tag{47}$$

Using Lemma V.12(ii), we have $|\mu(\mathbb{Y}) - 2| < \kappa_1 d^{1-\epsilon/4} < 0.1$, leading to

$$\left|\frac{1}{\mu(\mathbb{Y})} - \frac{1}{2}\right| \le |\mu(\mathbb{Y}) - 2| \le \frac{1}{3}\, d^{1-\epsilon/2},$$

and, in particular, $\mu(\mathbb{Y}) < 3$ for small d. Hence, substituting into (47) and using the lower bound $H(\mathbb{Y}) \ge 1 - d^{2-\epsilon/2}$, we have $D(q_L\|q_L^*) < 3d^{2-\epsilon/2}$. Explicit calculation gives $D(q_L\|q_L^*) = \sum_{l=1}^{\infty} q_L(l)\big(\log q_L(l) + l\big)$. The result follows by plugging into (47). ■

C. Modified Deletion Process

We want to get a handle on the term $H(D^n|X^n,Y,K)$. The main difficulty in achieving this is that a fixed run in Y can arise in many ways from parent runs, via a countable infinity of different deletion "patterns." For example, consider that a run in Y may have *any* odd number of parent runs. Moreover, a countable infinity of these deletion patterns "contribute" to $H(D^n|X^n,Y,K)$.

However, we expect that deletions are typically well separated at small deletion probabilities, and as a result, there are only a few dominant "types" of deletion patterns that influence the leading order terms in $H(D^n|X^n,Y,K)$. Deletions that "act" in isolation from other deletions should contribute an order d term: for instance, a positive fraction of runs in X^n should have a length 4, and with probability of order d, they should shrink to runs of length 3 in Y due to one deletion. Each time this occurs, there are four (equally likely) candidate positions at which the one deletion occurred, contributing $\log(4)$ to $H(D^n|X^n,Y,K)$. Similarly, pairs of "nearby" deletions (for instance, in the same run of X^n) should contribute a term of order d^2. We should be able to ignore instances of more than two deletions occurring in close proximity, since (intuitively) they should have a contribution of $O(d^3)$ on $H(D^n|X^n,Y,K)$.

We formalize this intuition by constructing a suitable modified deletion process that allows us to focus on the dominant deletion patterns in our estimate of this term. We bound the error in our estimate due to our modification of the deletion process, leading to an estimate of $H(D^n|X^n,Y,K)$ that is exact up to order d^2.

We restrict attention to $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$. Denote by R_j the jth run in $\mathbb{X}$ (where the run including bit 1 is labeled R_1). R_j has length L_j. Recall that the deletion process $\mathbb{D}$ is an iid Bernoulli(d) process, independent of $\mathbb{X}$, with D_1^n being the n-bit vector that contains a 1 if and only if the corresponding bit in X^n is deleted by the channel W_n. We define an auxiliary sequence of channels $\widehat{W}_n$ whose output -denoted by $\widehat{Y}(X^n)$– is obtained by modifying the deletion channel output: $\widehat{Y}(X^n)$ contains all bits

[1] The perturbed deletion process is constructed using a different modification to the deletion process.

TABLE IV
EXAMPLE SHOWING HOW $\widehat{\mathbb{D}}(\mathbb{X}, \mathbb{D})$ IS CONSTRUCTED

			R_1		R_2	R_3			R_4						
$\mathbb{X}$:	...	1	0	0	1	0	0	0	1	1	1	1	0	0	...
$\mathbb{D}$:	...	0	0	1	1	1	0	1	1	0	1	1	0	0	...
$\widehat{\mathbb{D}}$:	...	0	0	1	1	1	0	1	0	0	0	0	0	0	...

present in $Y(X^n)$ and some of the deleted bits in addition. Specifically, whenever there are *three* or more deletions in a single run R_i under $\mathbb{D}$, the run R_i suffers no deletions in $\widehat{Y}(X^n)$.

Formally, we construct this sequence of channels when the input is a stationary process $\mathbb{X}$ as follows. For all integers i, define:

$\mathbb{Z}^i \equiv$ Binary process that is zero throughout except if R_i contains 3 or more deletions, in which case $Z_l^i = 1$ if and only if $X_l \in R_i$ and $D_l = 1$.

Define $\mathbb{Z}$ by

$$Z_l = \begin{cases} 1 & \text{if } \exists i \text{ s.t. } Z_l^i = 1\,, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, define $\widehat{\mathbb{D}}(\mathbb{D}, \mathbb{X}) \equiv \mathbb{D} \oplus \mathbb{Z}$ (where $\oplus$ is componentwise sum modulo 2). The output of the channel $\widehat{W}_n$ is simply defined by deleting from X^n those bits whose positions correspond to 1s in $\widehat{\mathbb{D}}$. We define $\widehat{K}(X^n)$ for the modified deletion process in the same way as $K(X^n)$. The sequence of channels W_n are defined by $\mathbb{D}$, and the coupled sequence of channels $\widehat{W}_n$ are defined by $\widehat{\mathbb{D}}$. We emphasize that $\widehat{\mathbb{D}}$ is a function of $(\mathbb{X}, \mathbb{D})$.

Note that if $D_l = 0$, then $Z_l = 0$, and hence, $\widehat{D}_l = 0$. Thus, $\widehat{\mathbb{D}}$ is obtained by flipping the 1s in $\mathbb{D}$ that also correspond to 1s in $\mathbb{Z}$. If $Z_i = 1$, i.e., $D_i = 1, \widehat{D}_i = 0$, we will say that a deletion is *reversed* at position i. See Table IV for an example: The deletions in run R_4 are reversed because there are three of them, whereas the deletions in runs R_1, R_2, and R_3 are not affected.

It is not hard to see that the process $\mathbb{Z}$ is stationary. (In fact $(\mathbb{X}, \mathbb{D}, \mathbb{Z}, \widehat{\mathbb{D}})$ are jointly stationary.) Define $z \equiv \mathbb{P}(Z_i = 1)$, where i is arbitrary.

The expected number of deletions reversed due to a run with length ℓ is bounded above by

$$\ell d - \ell d(1-d)^{\ell-1} - 2\binom{\ell}{2} d^2 (1-d)^{\ell-2} \\ \leq \ell(\ell-1)(\ell-2)d^3 \leq \ell^3 d^3, \tag{48}$$

using $(1-d)^{\ell-1} \geq 1-(\ell-1)d$ and $(1-d)^{\ell-2} \geq 1-(\ell-2)d$.

We know that each run has length at least 1. Thus, we have the following.

Fact V.20: For arbitrary stationary process $\mathbb{X}$, the probability z of a reversed deletion at an arbitrary position i is bounded as $z \leq d^3 \mathbb{E}[L^3]$.

Now $\mathbb{E}[L^3] \leq d^{-2}\mathbb{E}[L]$ for $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$. If $\max\{H(\mathbb{X}), H(\mathbb{Y})\} \geq 1 - d^\gamma$, Lemmas V.3 and V.7 yield $\mathbb{E}[L] < \kappa d^\gamma$. Combining, we deduce:

Fact V.21: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$ and $\kappa < \infty$ such that for any $d < d_0$ the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$ such that $\max\{H(\mathbb{X}), H(\mathbb{Y})\} \geq 1 - d^\gamma$. Then, we have $\mathbb{E}[L^3] < \kappa d^{\gamma-2}$.

Note that $\max\{H(\mathbb{X}), H(\mathbb{Y})\} \geq 1 - d^{2-\epsilon/2}$ holds for relevant processes $\mathbb{X}$ (see Lemma IV.4), justifying our aforementioned assumption.

The next proposition follows immediately from Facts V.20 and V.21.

Proposition V.22: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$ and $\kappa < \infty$ such that for any $d < d_0$ the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$ such that $\max\{H(\mathbb{X}), H(\mathbb{Y})\} \geq 1 - d^\gamma$. Then, we have $z < \kappa d^{1+\gamma}$.

We now analyze the modified deletion process with the aim of estimating $H(\widehat{D}^n | X^n, \widehat{Y}, \widehat{K})$. Notice that for any run R_i, either all deletions in R_i are reversed (in which case we say that R_i suffers deletion reversal), or none of the deletions are reversed (in which case we say that R_i is unaffected by reversal). It follows that

$$H(\widehat{D}^n | X^n, \widehat{Y}, \widehat{K}) = \sum_{j=1}^{M} H(\widehat{D}(j) | \widehat{X}(j), \widehat{Y}(j)), \tag{49}$$

where $\widehat{D}(j)$ consists of the substring of $\widehat{D}^n$ corresponding to $\widehat{X}(j)$. As before, when we study $H(\widehat{D}^n | X^n, \widehat{Y}, \widehat{K})/n$ in the limit $n \to \infty$, the terms corresponding to $j = 1$ and $j = M$ can be neglected, and we can perform the calculation by considering the stationary processes $\mathbb{X}$, $\mathbb{Y}$, and $\mathbb{D}$.

Recall the definition of the parent runs $\widehat{X}(j)$ of a run $\widehat{Y}(j)$ for $j > 1$ from Section V-B. Consider the possibilities for how many runs $\widehat{X}(j)$ contains, and the resultant ambiguity (or not) in the position of deletions (under $\widehat{\mathbb{D}}$) in the parent run(s):

Single parent run: Let the parent run be R_P. The parent run should not disappear;[2] by definition it should contribute at least one bit to $\widehat{Y}(j)$. The run R_{P+1} should not disappear (else it is also a parent). R_P can suffer $0, 1$, or 2 deletions (else we have a deletion pattern not allowed under $\widehat{\mathbb{D}}$). The cases of 1 or 2 deletions lead to ambiguity in the location of deletions.

Note that if R_{P-1} disappears, then R_{P-2} also disappears (else R_{P-2}, R_{P-1} are also parents of $\widehat{Y}(j)$), and so on.

Combination of Three Parent Runs: Let the parent runs be R_P, R_{P+1}, and R_{P+2}. We know that R_P and R_{P+3} did not disappear and R_{P+1} has disappeared, by definition of $X(j)$ (cf., Table III). If R_P and R_{P+2} suffer no deletions, this leads to no ambiguity in the location of deletions. Ambiguity can arise in case R_P and R_{P+2} suffer between one and four deletions in total.

Note that if R_{P-1} disappears, then R_{P-2} also disappears, and so on.

Combination of $2k+1$ Parent Runs, for $k = 2, 3, \ldots$: Let the parent runs be $R_P, R_{P+1}, \ldots, R_{P+2k}$. The runs $R_{P+1}, R_{P+3}, \ldots, R_{P+2k-1}$ must disappear and R_P does not disappear. The runs $R_P, R_{P+2}, \ldots, R_{P+2k}$ must suffer between one and $2(k+1)$ deletions in total for ambiguity to arise in the location of deletions.

[2] We emphasize that we are referring here to deletions under $\widehat{\mathbb{D}}$.

Define

$$p_{L(3)}(>1, l_2, l_3) \equiv \sum_{l_1=2}^{\infty} p_{L(3)}(l_1, l_2, l_3),$$
$$p_{L(3)}(>1, l_2, >1) \equiv \sum_{l_1=2}^{\infty} \sum_{l_3=2}^{\infty} p_{L(3)}(l_1, l_2, l_3),$$

and so on.

The following lemma shows the utility of the modified deletion process. We obtain this result by adding the contributions of the cases enumerated earlier.

Lemma V.23: There exists $d_0 > 0$ such that for any $d < d_0$ the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$. Then,

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} H(\widehat{D}^n | X^n, \widehat{Y}, \widehat{K}) = \\
&\frac{d}{\mu(\mathbb{X})} \sum_{l=2}^{\infty} p_L(l)\, l \log l \\
&+ \frac{d^2}{\mu(\mathbb{X})} \sum_{l=2}^{\infty} p_L(l) \left\{ \binom{l}{2} \log \binom{l}{2} - l^2 \log l \right\} \\
&+ \frac{d^2}{\mu(\mathbb{X})} \sum_{l=2}^{\infty} \left\{ p_{L(3)}(>1, l, >1)\, l \log l - p_{L(3)}(1, l, 1)\, l \log l \right\} \\
&+ \frac{d^2}{\mu(\mathbb{X})} \Bigg(\sum_{l_0 > 1, l_2} \left\{ p_{L(3)}(l_0, 1, l_2)\, (l_0 + l_2) \log(l_0 + l_2) \right\} \\
&\qquad + \sum_{1,1,l_2} \left\{ p_{L(3)}(1, 1, l_2)\, l_2 \log l_2 \right\} \Bigg) + \delta, \qquad (50)
\end{aligned}$$

where

$$-11 d^3 \log(1/d) \mathbb{E}[L^3] \le \delta \le 140 d^3 \log(1/d) \mathbb{E}[L^3]\,. \qquad (51)$$

The proof of Lemma V.23 is quite technical.

Proof of Lemma V.23: We make use of (49) and the fact that $\mathbb{X}$ is stationary and ergodic. Consider a randomly chosen run R_P in $\mathbb{X}$. We associate $H(\widehat{D}(j)|\widehat{X}(j), \widehat{Y}(j))$ with R_P if R_P is the first run in $\widehat{X}(j)$. Denote by L_{P+i} the length of R_{P+i} for any integer i. We add contributions from the three possibilities of how $\widehat{Y}(j)$ arose under $\widehat{D}(j)$:

1) From a single parent run: Define

$$B_1 \equiv \{R_P \text{ suffers one or two deletions under } \mathbb{D} \text{ and } \exists j \text{ s.t. } \widehat{X}(j) = R_P\}.$$

Clearly, B_1 is exactly the event we are interested in here. We will restrict attention to a subset of B_1 and the prove that we are missing a very small contribution. Define

$$E_1 \equiv B_1 \cap \{R_{P-1} \text{ and } R_{P+1} \text{ do not disappear under } \mathbb{D}.\}.$$

Consider $B_1 \backslash E_1$. For this event, one of the following must occur.

1) Run R_{P-1} disappears under $\mathbb{D}$ but not under $\widehat{\mathbb{D}}$. For this, we need at least three deletions in run R_{P-1}. A simple calculation shows that this occurs with probability less than $d^3 L_{P-1}^3$.
2) Run R_{P-1} disappears under $\widehat{\mathbb{D}}$ as well. In this case, R_{P-2} also disappears under $\widehat{\mathbb{D}}$. Thus, we need R_{P-1} and R_{P-2} both to disappear under $\mathbb{D}$ which occurs with probability at most d^2. Moreover, we require at least one deletion in R_P (probability less than $L_P d$). Thus, the overall probability is bounded above by $d^3 L_P$.
3) Run R_{P+1} disappears under $\mathbb{D}$ but not under $\widehat{\mathbb{D}}$. For this, we need at least three deletions in run R_{P+1}. This occurs with probability less than $d^3 L_{P+1}^3$.

Thus, $0 \le \mathbb{P}(B_1 \backslash E_1) < d^3(L_{P-1}^3 + L_P + L_{P+1}^3)$. The largest possible value of $H(\widehat{D}(j)|\widehat{X}(j), \widehat{Y}(j))$ for a particular occurrence of $B_1 \backslash E_1$ is $\max_{i=1,2} \log \binom{L_P}{i} \le 2 \log L_P$. Thus, the additive error introduced by restricting to E_1 in our estimate of $\lim_{n\to\infty} \frac{1}{n} H(\widehat{D}^n|X^n, \widehat{Y}, \widehat{K})$ is

$$\begin{aligned}
0 \le \delta_{1E}(d, \mathbb{X}) &\le d^3 \mathbb{E}[2(L_{P-1}^3 + L_P + L_{P+1}^3) \log L_P] \\
&\le 6 d^3 \mathbb{E}[L^3 \log L] \qquad (52)
\end{aligned}$$

where we have made use of Proposition A.1.

Partition E_1 into two events:

$$B_{11} \equiv E_1 \cap \{R_P \text{ undergoes one deletion under } \mathbb{D}\} \qquad (53)$$
$$B_{12} \equiv E_1 \cap \{R_P \text{ undergoes two deletions under } \mathbb{D}\}. \qquad (54)$$

Let T_1 be the contribution of B_1 to $H(\widehat{D}(j)|\widehat{X}(j), \widehat{Y}(j))$, T_{11} be the contribution of B_{11} and T_{12} be the contribution of B_{12}. Then, we have

$$T_1 = T_{11} + T_{12} + \delta_{1E} \qquad (55)$$

1) One deletion in R_P:
Consider B_{11}. The contribution of a particular occurrence is $\log L_P$. Now

$$\begin{aligned}
&\mathbb{P}(B_{11}, L_P = l, L_{P-1} = l_{P-1}, L_{P+1} = l_{P+1}) \\
&= p_{L(3)}(l_{P-1}, l, l_{P+1})\, (1 - d^{l_{P-1}})\, (1 - d^{l_{P+1}}) \\
&\quad \cdot\ l_P d (1-d)^{l_P - 1}. \qquad (56)
\end{aligned}$$

We have, for $l > 1$,

$$\begin{aligned}
&p_{L(3)}(>1, l, >1)\, l d (1-d)^{l-1}\, (1 - 2d^2) \\
&\le \mathbb{P}(B_{11}, L_P = l, L_{P-1} > 1, L_{P+1} > 1) \\
&\le p_{L(3)}(>1, l, >1)\, l d (1-d)^{l-1}
\end{aligned}$$

since probability that R_{P-1} of length greater than 1 disappears is bounded above by d^2 and similarly for R_{P+1}. It follows that

$$\begin{aligned}
&\mathbb{P}(B_{11}, L_P = l, L_{P-1} > 1, L_{P+1} > 1) \\
&= p_{L(3)}(>1, l, >1)\, l d (1 - (l-1)d) + \eta_{1,1}(l), \\
&-2d^3 p_{L(3)}(>1, l, >1)\, l \le \eta_{1,1}(l) \\
&\qquad \le d^3 p_{L(3)}(>1, l, >1)\, l \binom{l-1}{2}.
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
&\mathbb{P}(B_{11}, L_P = l, L_{P-1} = 1, L_{P+1} = 1) \\
&= p_{L(3)}(1, l, 1)\, l d (1 - (l+1)d) + \eta_{1,4}(l), \\
&0 \le \eta_{1,4}(l) \le d^3 p_{L(3)}(1, l, 1)\, l \binom{l+1}{2}
\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(B_{11}, L_P = l, L_{P-1} > 1, L_{P+1} = 1)\\ &\quad = p_{L(3)}(>1, l, 1)\, ld(1-ld) + \eta_{1,3}(l),\\ &-d^3 p_{L(3)}\,(>1, l, 1)\, l \le \eta_{1,3}(l) \le d^3 p_{L(3)}(>1, l, 1)\, l\binom{l}{2}\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(B_{11}, L_P = l, L_{P-1} = 1, L_{P+1} > 1)\\ &\quad = p_{L(3)}(1, l, >1)\, ld(1-ld) + \eta_{1,2}(l),\\ &-d^3 p_{L(3)}(1, l, >1)\, l \le \eta_{1,2}(l) \le d^3 p_{L(3)}(1, l, >1)\, l\binom{l}{2}.\end{aligned}$$

Combining, we arrive at the following contribution of B_{11} to $\lim_{n\to\infty} H(\widehat{D}^n | X^n, \widehat{Y}, \widehat{K})/n$:

$$\begin{aligned}T_{11} &= \frac{1}{\mu(\mathbb{X})}\sum_{l=2}^{\infty}\mathbb{P}(B_{11}, L_P = l)\log l\\ &= \frac{d}{\mu(\mathbb{X})}\sum_{l=2}^{\infty}\Big\{p_{L(3)}(>1, l, >1)\, l\big(1-(l-1)d\big)\log l\\ &\qquad + p_{L(3)}(1, l, 1)\, l\big(1-(l+1)d\big)\log l\\ &\qquad + \big(p_{L(3)}(1, l, >1) + p_{L(3)}(>1, l, 1)\big)\\ &\qquad \cdot\, l\big(1-ld\big)\log l\Big\} + \delta_{11},\end{aligned}\tag{57}$$

with

$$\begin{aligned}-\frac{2d^3}{\mu(\mathbb{X})}\sum_{l=2}^{\infty}p_L(l)\, l\log l \le\ &\delta_{11} = \delta_{11}(d, \mathbb{X})\\ &\le \frac{d^3}{\mu(\mathbb{X})}\sum_{l=2}^{\infty}p_L(l)\, l\binom{l+1}{2}\log l.\end{aligned}\tag{58}$$

We have normalized by $\mu(\mathbb{X})$ to move from a per run contribution to a per bit contribution.

It is easy to infer

$$-d^3\mathbb{E}[L^3\log L] \le \delta_{11} \le d^3\mathbb{E}[L^3\log L]\tag{59}$$

from (58).

2) Two deletions in R_P:
Consider B_{12}. If $L_P = l > 2$, then entropy contribution is $\log\binom{l}{2}$. We have, for $l > 2$,

$$\begin{aligned}\mathbb{P}(B_2, L_P = l) &= p_L(l)\binom{l}{2}d^2(1-d)^{l-2}\\ &\cdot\ \mathbb{P}(R_{P-1} \text{ and } R_{P+1} \text{ do not disappear under } \mathbb{D}).\end{aligned}$$

It follows that

$$\begin{aligned}p_L(l)\binom{l}{2}d^2(1-d)^l &\le \mathbb{P}(B_2, L_P = l)\\ &\le p_L(l)\binom{l}{2}d^2(1-d)^{l-2}\end{aligned}$$

leading to

$$\begin{aligned}\mathbb{P}(B_2, L_P = l) &= p_L(l)\binom{l}{2}d^2 + \eta_2,\\ -d^3 p_L(l)\, l\binom{l}{2} &\le \eta_2 \le 0.\end{aligned}$$

Combining, we arrive at the following contribution to $\lim_{n\to\infty} H(\widehat{D}^n | X^n, \widehat{Y}, \widehat{K})/n$:

$$\begin{aligned}T_{12} &= \frac{1}{\mu(\mathbb{X})}\sum_{l=3}^{\infty}\mathbb{P}(B_2, L_P = l)\log\binom{l}{2}\\ &= \frac{d^2}{\mu(\mathbb{X})}\sum_{l=3}^{\infty}p_L(l)\binom{l}{2}\log\binom{l}{2} + \delta_{12}\end{aligned}\tag{60}$$

with

$$\begin{aligned}-d^3\mathbb{E}[L^3\log L] &\le -\frac{d^3}{\mu(\mathbb{X})}\sum_{l=3}^{\infty}p_L(l)\, l\binom{l}{2}\log\binom{l}{2}\\ &\le \delta_{12} = \delta_{12}(d, \mathbb{X}) \le 0.\end{aligned}\tag{61}$$

Plugging (57) and (60) into (55), we obtain our desired estimate on the contribution T_1 of the event B_1,

$$\begin{aligned}&T_1\\ &= \frac{d}{\mu(\mathbb{X})}\sum_{l=2}^{\infty}\Big\{p_{L(3)}(>1, l, >1)\, l\log l\big(1-(l-1)d\big)\\ &\qquad + p_{L(3)}(1, l, 1)\, l\log l\big(1-(l+1)d\big)\\ &\qquad + \big(p_{L(3)}(1, l, >1)\\ &\qquad + p_{L(3)}(>1, l, 1)\big)\, l\log l\big(1-ld\big)\Big\}\\ &\quad + \frac{d^2}{\mu(\mathbb{X})}\sum_{l=3}^{\infty}p_L(l)\binom{l}{2}\log\binom{l}{2} + \delta_1,\end{aligned}$$

where $\delta_1 = \delta_{1E} + \delta_{11} + \delta_{12}$ is bounded using (52), (59), and (61) as

$$-2d^3\mathbb{E}[L^3\log L] \le \delta_1 \le 7d^3\mathbb{E}[L^3\log L]\,.\tag{62}$$

2) From a combination of three parent runs: Define

$$\begin{aligned}B_3 \equiv \{&R_P \text{ and } R_{P+2} \text{ suffer at least one deletion in total}\\ &\text{under } \widehat{\mathbb{D}} \text{ and } \exists j \text{ s.t. } \widehat{X}(j) = (R_P\, R_{P+1}\, R_{P+2})\}.\end{aligned}$$

We are interested in the contribution due to occurrence of event B_3.

Again, we will restrict attention to a subset of B_3 and the prove that we are missing a very small contribution. Define

$$E_3 \equiv B_3 \cap \{R_{P-1} \text{ and } R_{P+3} \text{ do not disappear under } \mathbb{D}\}.$$

Similar to our analysis for Case 1, we can show that

$$0 \le \mathbb{P}(B_3 \backslash E_3) < d^3(L_{P-1}^3 + L_P + L_{P+2} + L_{P+1}^3).$$

The largest possible value of $H(\widehat{D}(j)|\widehat{X}(j),\widehat{Y}(j))$ for a particular occurrence of $B_3 \backslash E_3$ is

$$\max_{i=1,2,3,4} \log \binom{L_P + L_{P+2}}{i} \le 4\log(L_P + L_{P+2})$$

since R_P and R_{P+2} can suffer at most four deletions in total under $\widehat{\mathbb{D}}$. Thus, the additive error introduced by restricting to E_3 in our estimate of $\lim_{n\to\infty} \frac{1}{n} H(\widehat{D}^n|X^n,\widehat{Y},\widehat{K})$ is

$$\begin{aligned}&0 \le \delta_{3E}(d,\mathbb{X})\\ &\le d^3\mathbb{E}[4(L^3_{P-1} + L_P + L_{P+2} + L^3_{P+1})\log(L_P + L_{P+2})].\end{aligned} \tag{63}$$

Now, $\log(L_P + L_{P+2}) \le \log(2L_P L_{P+2}) = 1 + \log L_P + \log L_{P+2}$. From Proposition A.1, $\mathbb{E}[L^3_{P-1}\log L_P] \le E[L^3 \log L]$, also $\mathbb{E}[L_P \log L_{P+2}] \le \mathbb{E}[L\log L]$, and so on. Plugging into (63), we arrive at

$$0 \le \delta_{3E}(d,\mathbb{X}) \le d^3\mathbb{E}[16L^3 + 32L^3\log L]. \tag{64}$$

Now, we further restrict to a subset of E_3. Define

$$\begin{aligned}B_{31} = E_3 \cap \{&\text{One deletion in total in } R_P, R_{P+2}\}\\ &\cap\{L_{P+1} = 1\}.\end{aligned}$$

Consider the event $E_3\backslash B_{31}$. This can occur due to one of the following.

1) *More than one deletion in* R_P, R_{P+2}*:* This occurs with probability at most $\binom{L_P+L_{P+2}}{2}d^3$ (since we also need R_{P+1} to disappear).
2) $L_{P+1} > 1$*:* Now the probability that R_{P+1} disappears is at most d^2. Thus, the probability of $\mathbb{P}(E_3 \cap \{L_{P+1} > 1\}) \le (L_P + L_{P+2})d^3$.

It follows from union bound that $\mathbb{P}(E_3\backslash B_{31}) \le d^3(L_P + L_{P+2})^2$. As before, the largest possible value of $H(\widehat{D}(j)|\widehat{X}(j),\widehat{Y}(j))$ for a particular occurrence of $E_3\backslash B_{31}$ is $4\log(L_P + L_{P+2})$. Thus, the additive error introduced by restricting to B_{31} in estimating the contribution of E_3 is

$$0 \le \delta_{32} \le 4d^3(L_P + L_{P+2})^2 \log(L_P + L_{P+2}).$$

Now, we use $\log(L_P + L_{P+2}) \le 1 + \log L_P + \log L_{P+2}$ and Proposition A.1 to obtain

$$0 \le \delta_{32} \le d^3\mathbb{E}[16L^2 + 32L^2\log L]. \tag{65}$$

Denoting by T_{31} the contribution of B_{31}, and T_3 the contribution of B_3, we have

$$T_3 = T_{31} + \delta_{3E} + \delta_{32}. \tag{66}$$

We consider two cases in estimating T_{31}:

1) $L_P > 1$.
The value of $H(\widehat{D}(j)|\widehat{X}(j),\widehat{Y}(j))$ for a particular occurrence is $\log(L_P + L_{P+2})$. We have

$$\begin{aligned}&\mathbb{P}(B_{31}, L_P = l_0, |R_{P+2}| = l_2)\\ &= d^2 p_{L(3)}(l_0,1,l_2)(l_0+l_2) + \eta_{3,1},\\ &-d^3 p_{L(3)}(l_0,1,l_2)(l_0+l_2)^2 \le \eta_{3,1} \le 0.\end{aligned}$$

2) $L_P = 1$.
The value of $H(\widehat{D}(j)|\widehat{X}(j),\widehat{Y}(j))$ for a particular occurrence is $\log L_{P+2}$ since R_P should not disappear. We have

$$\begin{aligned}\mathbb{P}(B_3, L_P = 1, L_{P+2} = l_2) &= d^2 p_{L(3)}(1,1,l_2)l_2 + \eta_{3,2}\\ &-d^3 p_{L(3)}(1,1,l_2)l_2^2 \le \eta_{3,2} \le 0.\end{aligned}$$

Combining the two cases, we arrive at the following estimate:

$$\begin{aligned}T_{31} &= \frac{1}{\mu(\mathbb{X})}\sum_{l_0,l_2}\mathbb{P}(B_3, L_P = l_0, L_{P+2} = l_2)\\ &\qquad\cdot\log\left(l_2 + l_0\mathbb{I}(l_0 > 1)\right)\\ &= \frac{d^2}{\mu(\mathbb{X})}\Bigg(\sum_{l_0>1,l_2} p_{L(3)}(l_0,1,l_2)\,(l_0+l_2)\log(l_0+l_2)\\ &\qquad + \sum_{l_2} p_{L(3)}(1,1,l_2)\,l_2\log l_2\Bigg) + \delta_{31}\end{aligned} \tag{67}$$

where

$$\begin{aligned}-\frac{d^3}{\mu(\mathbb{X})}\sum_{l_0,l_2} p_{L(3)}(l_0,1,l_2)\,(l_0+l_2)^2\log(l_0+l_2)&\\ \le \delta_{31} = \delta_{31}(d,\mathbb{X}) \le 0.&\end{aligned}$$

Again, we use $\log(L_P + L_{P+2}) \le 1 + \log L_P + \log L_{P+2}$ and Proposition A.1 to obtain

$$-d^3\mathbb{E}[4L^2 + 8L^2\log L] \le \delta_{31} = \delta_{31}(d,\mathbb{X}) \le 0. \tag{68}$$

Finally, we plug (67) into (66) to obtain

$$\begin{aligned}T_3 = \frac{d^2}{\mu(\mathbb{X})}\Bigg(&\sum_{l_0>1,l_2} p_{L(3)}(l_0,1,l_2)\,(l_0+l_2)\log(l_0+l_2)\\ &+ \sum_{l_2} p_{L(3)}(1,1,l_2)\,l_2\log l_2\Bigg) + \delta_3\end{aligned}$$

where $\delta_3 = \delta_{3E} + \delta_{32} + \delta_{31}$. Using (64), (65), and (68), we obtain

$$-d^3\mathbb{E}[4L^2 + 8L^2\log L] \le \delta_3 \le d^3\mathbb{E}[32L^3 + 64L^3\log L]. \tag{69}$$

3) *From a combination of five parent runs:* Define

$$\begin{aligned}B_5 \equiv \{&R_P, R_{P+2}, R_{P+4} \text{ suffer at least one deletion in total}\\ &\text{under } \widehat{\mathbb{D}} \text{ and } \exists j \text{ s.t. } \widehat{X}(j) = (R_P R_{P+1} R_{P+2} R_{P+3} R_{P+4})\}.\end{aligned}$$

We have $\mathbb{P}(B_5) \le d^3(L_P + L_{P+2} + L_{P+4})$ since R_{P+1} and R_{P+3} must disappear. Also, the largest possible value of $H(\widehat{D}(j)|\widehat{X}(j), \widehat{Y}(j))$ for a particular occurrence is

$$\max_{i=1,2,\dots,6} \log \binom{L_P + L_{P+2} + L_{P+4}}{i} \\ \le 6\log(L_P + L_{P+2} + L_{P+4})$$

since each run can suffer at most two deletions under $\widehat{\mathbb{D}}$. Thus, the contribution of B_5 is δ_5, where

$$\begin{aligned} 0 &\le \delta_5 \\ &\le 6d^3\mathbb{E}[(L_P + L_{P+2} + L_{P+4})\log(L_P + L_{P+2} + L_{P+4})] \\ &\le d^3\mathbb{E}[36L + 54L\log L] \end{aligned} \tag{70}$$

where we have used $\log(L_P + L_{P+2} + L_{P+4}) \le 2 + \log L_P + \log L_{P+2} + \log L_{P+4}$ and Proposition A.1.

From a combination of $2k+1$ *parent runs for* $k \ge 3$*:* Define

$$B_{2k+1} \equiv \exists j \text{ s.t. } \widehat{X}(j) = (R_P R_{P+1} \dots R_{P+2k}).$$

We need k runs to disappear, and this occurs with probability at most d^k. The largest possible value of $H(\widehat{D}(j)|\widehat{X}(j), \widehat{Y}(j))$ for a particular occurrence is $2(k+1)\log(L_P + L_{P+2} + \dots + L_{P+2k}) \le 2(k+1)\log((k+1)/d)$ since no run has length exceeding $1/d$. Thus, the contribution of B_{2k+1} is bounded above by $d^k 2(k+1)\log((k+1)/d)$. Summing, we find that the overall contribution $T_{\mathrm{gt}5}$ of $B_7, B_9, \dots$ is bounded as

$$0 \le T_{\mathrm{gt}5} \le \sum_{k=3}^{\infty} d^k 2(k+1)\log((k+1)/d) \le 10d^3\log(1/d) \tag{71}$$

for small enough d.

Finally, we obtain

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n}H(\widehat{D}^n|X^n, \widehat{Y}, \widehat{K}) = T_1 + T_3 + T_5 + T_{\mathrm{gt}5} \\ &= \frac{d}{\mu(\mathbb{X})}\sum_{l=2}^{\infty}\Big\{p_{L(3)}(>1, l, >1)\, l\log l\big(1-(l-1)d\big) \\ &\qquad + \big(p_{L(3)}(1, l, >1) \\ &\qquad + p_{L(3)}(>1, l, 1)\big)\, l\log l\big(1-ld\big) \\ &\qquad + p_{L(3)}(1, l, 1)\, l\log l\big(1-(l+1)d\big)\Big\} \\ &+ \frac{d^2}{\mu(\mathbb{X})}\sum_{l=3}^{\infty} p_L(l)\binom{l}{2}\log\binom{l}{2} \\ &+ \frac{d^2}{\mu(\mathbb{X})}\Big(\sum_{l_0>1, l_2} p_{L(3)}(l_0, 1, l_2)\,(l_0 + l_2)\log(l_0 + l_2) \\ &\qquad + \sum_{1,1,l_2} p_{L(3)}(1, 1, l_2)\, l_2\log l_2\Big) + \delta \end{aligned}$$

where $\delta = \delta_1 + \delta_3 + \delta_5 + T_{\mathrm{gt}5}$. Rearranging gives (50), whereas (51) follows for small enough d from (62), (69), (70), and (71) and the fact that no run has length exceeding $1/d$. ■

Making use of the estimates of $p_{L(k)}(\cdot)$ derived in Section V-A, we obtain the following corollary of Lemma V.23.

Corollary V.24: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$ and $\kappa < \infty$ such that for any $d < d_0$ the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$ such that $H(\mathbb{X}) \ge 1 - d^{1-\epsilon}$ and $\max\{H(\mathbb{X}), H(\mathbb{Y})\} \ge 1 - d^{\gamma}$ for some $\gamma \in (0, 2)$. Then

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n}H(\widehat{D}^n|X^n, \widehat{Y}, \widehat{K}) \\ &\qquad = \frac{d}{\mu(\mathbb{X})}\Big\{\sum_{l=2}^{\infty} p_L(l)\, l\log l\Big\} + d^2 c_3 + \xi, \end{aligned}$$

where $|\xi| \le \kappa d^{1+\gamma-\epsilon/2}$. Recall that

$$\begin{aligned} c_3 \equiv \frac{1}{2}\Big(-1 + \sum_{l=3}^{\infty} 2^{-l}\Big\{&\binom{l}{2}\log\binom{l}{2} - l^2\log l \\ &+ (l-1)(l-3)\log(l-1) + (l-2)\log(l-2)\Big\}\Big). \end{aligned}$$

Note that with $\gamma = 2 - \epsilon/2$, we obtain $|\xi| \le \kappa d^{3-\epsilon}$ from Corollary V.24.

Proof of Corollary V.24: We prove the corollary assuming $H(\mathbb{Y}) > 1 - d^{\gamma}$. The proof assuming $H(\mathbb{X}) > 1 - d^{\gamma}$ is analogous.

It follows from Fact V.21 that if $H(\mathbb{Y}) \ge 1 - d^{\gamma}$, then δ [cf., (51)] is bounded as $|\delta| < \kappa_1 d^{1+\gamma}\log(1/d) \le d^{1+\gamma-\epsilon/2}$ for small enough d, for some $\kappa_1 < \infty$.

Consider $\sum_{l=2}^{\infty} p_L(l) l^2\log l$. We separately analyze the first $l_0 = \lfloor 4\log(1/d) \rfloor$ terms of the sum. We use Lemma V.12(i) [see (37)] to deduce that

$$\sum_{l=2}^{l_0} p_L(l) l^2\log l = \sum_{l=2}^{\infty} p_L^*(l) l^2\log l + \xi_1, \tag{72}$$
$$\text{with}\quad |\xi_1| \le \kappa_4 d^{\gamma/2-\epsilon/4}(l_0)^3 \le \kappa_5 d^{\gamma/2-\epsilon/2},$$

for small enough d. Next, we use Lemma V.7 to deduce that

$$\begin{aligned} \sum_{l=l_0+1}^{\infty} p_L(l) l^2\log l &= \sum_{l=l_0+1}^{\lfloor 1/d \rfloor} p_L(l) l^2\log l \\ &\le \kappa_6 d^{\gamma}(1/d)\log(1/d) \\ &\le \kappa_7 d^{\gamma-\epsilon/2} \end{aligned} \tag{73}$$

for small enough d. Finally, Lemma V.12(ii) tells us that

$$|\mu(\mathbb{X}) - 2| \le \kappa_3 d^{\gamma/2}.$$

Combining with (72) and (73), it follows that

$$\frac{d^2}{\mu(\mathbb{X})}\sum_{l=2}^{\infty} p_L(l) l^2\log l = \frac{d^2}{2}\Big\{\sum_{l=2}^{\infty} p_L^*(l) l^2\log l\Big\} + \eta_2$$

where $|\eta_2| \le \kappa_8 d^{2+\gamma/2-\epsilon/2} \le \kappa_8 d^{1+\gamma-\epsilon/2}$, for small enough d.

Other terms in (50) can be similarly analyzed. The result follows. ■

We need to show that our estimate for the modified deletion process is also a good estimate for original deletion process. The following simple fact helps us do this:

Fact V.25: Suppose $U, \widehat{U}$, and V are random variables with the property that U is a deterministic function of $\widehat{U}$ and V, and also $\widehat{U}$ is a deterministic function of U and V. (Denote this property by $U \overset{V}{\longleftrightarrow} \widehat{U}$.) Then

$$|H(U) - H(\widehat{U})| \leq H(V)\,.$$

Proof: We have $H(U) \leq H(\widehat{U}, V) \leq H(\widehat{U}) + H(V)$. Similarly, $H(\widehat{U}) \leq H(U) + H(V)$. ■

It is not hard to see that

$$(X^n, Y, K, D^n) \overset{Z^n}{\longleftrightarrow} (X^n, \widehat{Y}, \widehat{K}, \widehat{D}^n) \quad \text{and}$$
$$(X^n, Y, K) \overset{Z^n}{\longleftrightarrow} (X^n, \widehat{Y}, \widehat{K})\,.$$

Using Fact V.25, we obtain

$$\begin{aligned}|H(\widehat{D}^n|X^n, \widehat{Y}, \widehat{K}) - H(D^n|X^n, Y, K)| &\leq 2H(Z^n)\\ &\leq 2nh(z)\,. \quad (74)\end{aligned}$$

Combining (74) with Corollary V.24, we obtain an estimate for the second term in (44). For future convenience, we form an estimate in terms of $q_L(\cdot)$ instead of $p_L(\cdot)$, using Lemma V.12 to make the switch.

Corollary V.26: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$ and $\kappa < \infty$ such that for any $d < d_0$ the following occurs: Define $\ell \equiv \lfloor 4\log(1/d) \rfloor$. Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$ such that $H(\mathbb{X}) \geq 1 - d^{1-\epsilon}$ and $\max\{H(\mathbb{X}), H(\mathbb{Y})\} \geq 1 - d^{2-\epsilon/2}$. Then

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n} H(Y(X^n), K(X^n)|X^n)\\ &= -\frac{d}{2}\sum_{l=2}^{\ell} q_L(l)\, l\log l + \frac{dc_2}{4\ln 2}\sum_{l=1}^{\ell} q_L(l) l + d\log(1/d)\\ &\quad + \frac{d}{\ln 2}\left(1 - \frac{c_2}{2}\right) + d^2\left(- c_3 - \frac{1}{2\ln 2}\right) + \delta,\end{aligned}$$

where $|\delta| \leq \kappa d^{3-\epsilon}$. Recall $c_2 \equiv \sum_{l=1}^{\infty} 2^{-l} l \ln l$.

Proof of Corollary V.26: We prove the corollary assuming $H(\mathbb{Y}) > 1 - d^{\gamma}$. The proof assuming $H(\mathbb{X}) > 1 - d^{\gamma}$ is analogous.

By definition, D^n is independent of X^n, so $H(D^n) = H(D^n|X^n) = nh(d)$, where $h(\cdot)$ is the binary entropy function. We have, for $Y = Y(X^n)$,

$$\begin{aligned}H(Y, K|X^n) &= H(D^n|X^n) - H(D^n|X^n, Y, K)\\ &= nh(d) - H(\widehat{D}^n|X^n, \widehat{Y}, \widehat{K}) + n\delta_1\end{aligned}$$

with $|\delta_1(d, \mathbb{X})| \leq 2H(Z^n)/n \to 2h(z)$. It follows from Corollary V.24, with $\gamma = 2 - \epsilon/2$, that

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n} H(Y(X^n), K(X^n)|X^n)\\ &= h(d) - \frac{d}{\mu(\mathbb{X})}\sum_{l=2}^{\infty} p_L(l)\, l\log l - d^2 c_3 + \delta_2 \qquad (75)\end{aligned}$$

with $|\delta_2| \leq 2h(z) + \kappa_1 d^{3-\epsilon}$. From Proposition V.22, we know that $z < \kappa_1 d^{3-\epsilon/2}$. It follows that $h(z) \leq \kappa_2 d^{3-\epsilon}$, and hence, $|\delta_2| \leq \kappa_3 d^{3-\epsilon}$. Simple calculus gives

$$h(d) = d\log(1/d) + (d - d^2/2)/\ln 2 + \delta_3 \qquad (76)$$

$|\delta_3| \leq \kappa_4 d^3$. Using Lemma V.12(i) [see (38)] and Lemma V.7, we obtain

$$\sum_{l=2}^{\infty} p_L(l)\, l\log l = \sum_{l=2}^{\ell} q_L(l)\, l\log l + \delta_4 \qquad (77)$$

where $|\delta_4| \leq \kappa_5 d^{2-\epsilon}$ for small enough d. Using Lemma V.12(ii)[see(40)] and $\mu(\mathbb{X}) > 1$ (from Lemma V.1), we obtain

$$\left|\frac{1}{\mu(\mathbb{X})} - \frac{1}{\mu(\mathbb{Y})}\right| \leq \kappa_6 d^{2-\epsilon}. \qquad (78)$$

Also, it follows from $|\mu(\mathbb{Y}) - 2| \leq 7d^{1-\epsilon/4}$ (Lemma V.1 applied to $\mathbb{Y}$) and elementary calculus that

$$\begin{aligned}\{\mu(\mathbb{Y})\}^{-1} &= 1 - \frac{1}{4}\mu(\mathbb{Y}) + \delta_5\\ &= 1 - \frac{1}{4}\sum_{l=1}^{\ell} q_L(l) l + \delta_6 \qquad (79)\end{aligned}$$

where $|\delta_6| \leq \kappa_7 d^{2-\epsilon}$. Here, we have used Lemma V.3 (applied to $\mathbb{Y}$) to bound $\sum_{l=\ell+1}^{\infty} q_L(l) l$.

Plugging (76)–(79) into (75), we obtain the result. ■

D. Self-Improving Bound on $H(\mathbb{Y})$

Our next lemma constitutes a "self-improving" bound on the closeness of $H(\mathbb{Y})$ to 1 and leads directly to Lemma IV.4.

Lemma V.27: There exists a function $(\kappa, \epsilon) \mapsto d_0(\kappa, \epsilon) > 0$ such that the following happens for any $\epsilon > 0$, and constants $\kappa > 0$ and $\gamma \in (1/2, 2)$. For any $d < d_0$ and any $\mathbb{X} \in S_{\lfloor 1/d \rfloor}$ such that

$$I(\mathbb{X}) \geq 1 - d\log(1/d) - A_1 d - \kappa d^{2-(\epsilon/4)}$$

and $H(\mathbb{Y}) \geq 1 - d^{\gamma}$, we have

$$H(\mathbb{Y}) \geq 1 - d^{1+\gamma/2-\epsilon/2}\,.$$

Proof: From (44) and (45), we have

$$\begin{aligned} I(\mathbb{X}) &= \lim_{n\to\infty} \frac{1}{n}\big\{H(Y) - H(D^n) + H(D^n|X^n,Y,K) \\ &\qquad + H(K|X^n,Y)\big\} \\ &= (1-d)H(\mathbb{Y}) - h(d) \\ &\quad + \lim_{n\to\infty} \frac{1}{n}\left\{H(D^n|X^n,Y,K) + H(K|X^n,Y)\right\}. \end{aligned} \tag{80}$$

Using (74) and Proposition V.22, we have

$$\begin{aligned} \frac{1}{n}\big|H(D^n|X^n,Y,K) - H(\widehat{D}^n|X^n,\widehat{Y},\widehat{K})\big| \\ \le \kappa_1 d^{1+\gamma}\log(1/d). \end{aligned}$$

It follows from $H(\mathbb{X}) > I(\mathbb{X})$ and our assumed lower bound on $I(\mathbb{X})$, that $H(\mathbb{X}) > 1-d^{1-\epsilon}$ for some $\epsilon > 0$. Using Corollary V.24, $|\mu(\mathbb{X})-2| \le \kappa_2 d^{\gamma/2}$ from Lemma V.12(ii), and Lemmas V.12(i) and V.7 to control $p_L(\cdot)$, we have

$$\lim_{n\to\infty} \frac{1}{n} H(\widehat{D}^n|X^n,\widehat{Y},\widehat{K}) = \frac{d}{2}\Big\{\sum_{l=2}^{\infty} 2^{-l}\, l\log l\Big\} + \delta_1,$$

where $|\delta_1| \le \kappa_3 d^{1+\gamma/2-\epsilon/4}$.

Lemma V.18 gives

$$\lim_{n\to\infty} H(K|X^n,Y) \le \kappa_4 d^{1+\gamma/2-\epsilon/4}.$$

We used here $\gamma < 2$.

Plugging back into (80), we obtain

$$I(\mathbb{X}) \le H(\mathbb{Y}) - d\log(1/d) - A_1 d + \kappa_5 d^{1+\gamma/2-\epsilon/4}.$$

The result follows from the assumption on $I(\mathbb{X})$. ■

E. Auxiliary Lemmas for the Lower Bound

Lemma V.28: Recall $\mathbb{X}^\dagger$ is the process consisting of iid runs with distribution $p_L^\dagger(l) = 2^{-l}(1 + d(l\log l - c_2 l/2))$ (cf., Lemma IV.1). There exists $d_0 > 0$ such that, for any $d < d_0$, we have the following: For any integer i and any $x_{-\infty}^{i-1}$, we have

$$\big|\mathbb{P}\{X_i^\dagger = 1 \big| (X^\dagger)_{-\infty}^{i-1} = x_{-\infty}^{i-1}\} - 1/2\big| \le 0.05.$$

Proof: Without loss of generality, suppose $x_{i-1} = 1$. Also, suppose that it is the lth consecutive 1 to occur. Now, since the runs' starting points form a renewal process under $\mathbb{X}^\dagger$, we have

$$\frac{\mathbb{P}\{X_i^\dagger = 0 \big| (X^\dagger)_{-\infty}^{i-1} = x_{-\infty}^{i-1}\}}{\mathbb{P}\{X_i^\dagger = 1 \big| (X^\dagger)_{-\infty}^{i-1} = x_{-\infty}^{i-1}\}} = \frac{p_L^\dagger(l)}{\sum_{l'>l} p_L^\dagger(l')}.$$

Routine calculus yields

$$\sum_{l'>l} p_L^\dagger(l') = 2^{-l}\left(1 + d\{l\log l + \eta_{1,l}\}\right),$$

where $|\eta_{1,l}| \le \kappa_1 l$ for some $\kappa_1 < \infty$. In comparison, $p_L^\dagger(l) = 2^{-l}\left(1 + d\{l\log l - c_2 l/2\}\right)$.

Case I ($l < 1/\sqrt{d}$): In this case, we have $p_L^\dagger(l) = 2^{-l}(1+\eta_{2,l})$ with $|\eta_{2,l}| \le d^{0.4}$ and $\sum_{l'>l} p_L^\dagger(l') = 2^{-l}(1+\eta_{3,l})$ with $|\eta_{3,l}| \le d^{0.4}$, for sufficiently small d. The result follows.

Case II ($l \ge 1/\sqrt{d}$): In this case, $\{l\log l + \eta_{1,l}\} = \{l\log l - c_2 l/2\}(1+\eta_{4,l})$, where $|\eta_{4,l}| \le 0.01$ provided d is small enough. It follows that

$$\left|\frac{p_L^\dagger(l)}{\sum_{l'>l} p_L^\dagger(l')} - 1\right| \le 0.02.$$

The result follows. ■

Lemma V.29: Let $q_L^\dagger(\cdot)$ be the run-length distribution of $\mathbb{Y}^\dagger$ corresponding to input $\mathbb{X}^\dagger$. Then, there exists d_0 (same as in Lemma V.28) such that, for any $d < d_0$, we have $q_L^\dagger(l) \le (0.55)^l$ for all l.

Proof: Consider a realization s of the deletion process, i.e., $\mathbb{D} = s$ for some fixed $s \in \{0,1\}_{-\infty}^{\infty}$ (recall that a 1 indicates that a deletion occurred at that location). Suppose that bit $Y_i^\dagger$ comes from bit $X_j^\dagger$ in the input. (In particular, this means $Y_i^\dagger = X_j^\dagger$ and $s_j = 0$. The location j is uniquely determined by (i,s).) From Lemma V.28, we know that for any $x_{-\infty}^{j-1}$, we have

$$\big|\mathbb{P}\{Y_i^\dagger = X_j^\dagger = 1 \big| (X^\dagger)_{-\infty}^{j-1} = x_{-\infty}^{j-1}\} - 1/2\big| \le 0.05,$$

for $d < d_0$. Summing over the bits in $(X^\dagger)_{-\infty}^{j-1}$ where s takes the value 1 (indicating that those bits are deleted), we obtain

$$\big|\mathbb{P}\{Y_i^\dagger = 1 \big| (Y^\dagger)_{-\infty}^{i-1} = y_{-\infty}^{i-1}, D_{-\infty}^{j} = s_{\infty}^{j}\} - 1/2\big| \le 0.05,$$

where s is the realized deletion process (with $s_j = 0$). Finally, summing over possible realizations s of $\mathbb{D}$, we obtain

$$\big|\mathbb{P}\{Y_i^\dagger = 1 \big| (Y^\dagger)_{-\infty}^{i-1} = y_{-\infty}^{i-1}\} - 1/2\big| \le 0.05.$$

Now think of sampling a realization y of $\mathbb{Y}$, one bit at a time. Every time a bit is sampled, it ends the current run with probability at least 0.45, using the aforementioned inequality. This gives $q_L^\dagger(l) \le (0.55)^l$, implying the result. ■

F. Proofs of Lemmas IV.1, IV.4, IV.5, and IV.6

We first prove Lemma IV.6, followed by Lemmas IV.1, IV.4, and IV.5.

Proof of Lemma IV.6: We construct $\acute{\mathbb{X}} \in \mathcal{S}_{L^*}$ from $\mathbb{X}$ as follows: Suppose a super run starts at X_j and continues until X_{j+L^*}. We flip one or both of X_{j+L^*+1} and X_{j+L^*+2} such that the super run ends at X_{j+L^*}. (It is easy to verify that this can always be done. If multiple different choices work, then pick an arbitrary one.) The density of flipped bits in $\mathbb{X}$ is upper bounded by

$$\alpha = 2\mathbb{E}[\widetilde{L}\mathbb{I}(\widetilde{L} \ge L^*)]/L^*. \tag{81}$$

The expected fraction of bits in the channel output $\acute{Y} = Y(\acute{X}^n)$ that have been flipped relative to $Y = Y(X^n)$ (output of the same channel realization with different input) is also at most α. Let $F = F(\mathbb{X}, \mathbb{D})$ be the binary vector having the same length as Y, with a 1 wherever the corresponding bit in $\acute{Y}$ is flipped relative to Y, and 0s elsewhere. The expected fraction of 1s in F is at most α. Therefore,

$$H(F) \le n(1-d)h(\alpha) + \log(n+1)\,. \qquad (82)$$

Recall Fact V.25. Notice that $Y \overset{F}{\longleftrightarrow} \acute{Y}$, whence

$$|H(Y) - H(\acute{Y})| \le H(F)\,. \qquad (83)$$

Further, $\mathbb{X} - \acute{\mathbb{X}} - \acute{X}^n - \acute{Y}$ form a Markov chain, and $\acute{\mathbb{X}}$, $\acute{X}^n$ are deterministic functions of $\mathbb{X}$. Hence, $H(\acute{Y}|\acute{X}^n) = H(\acute{Y}|\acute{\mathbb{X}})$. Similarly, $H(Y|X^n) = H(Y|\mathbb{X})$. Therefore, [the second step is analogous to (83)]

$$|H(\acute{Y}|\acute{X}^n) - H(Y|X^n)| = |H(\acute{Y}|\acute{\mathbb{X}}) - H(Y|\mathbb{X})| \le H(F)\,. \qquad (84)$$

It follows from Lemma V.16 and $L^* > 2\gamma\log(1/d)$ that $\alpha \le 80d^{\gamma}/L^*$ for sufficiently small d. Hence, $h(\alpha) \le d^{\gamma-\epsilon}\log L^*/L^*$ for $d < d_0(\epsilon)$, for some $d_0(\epsilon) > 0$. Now (82) and (83) gives (10), whereas (11) follows by combining (82)–(84) to bound $|I(\mathbb{X}) - I(\acute{\mathbb{X}})|$. ■

Proof of Lemma IV.1: We first make some preliminary observations. Direct calculation using (19) and (8)[3] leads to $H(\mathbb{X}^\dagger) = H(p_L^\dagger)/\mu(\mathbb{X}^\dagger) = 1 - O(d^2)$, and $|\mu(\mathbb{X}^\dagger) - 2| = O(d)$. From Lemma V.9(ii), we deduce $|\mu(\mathbb{Y}^\dagger) - 2| = O(d)$.

Since $\mathbb{X}^\dagger$ consists of independent runs, the same is true for $\mathbb{Y}^\dagger$. Hence, recalling the notation $q_L^*(l) = 2^{-l}$, we have

$$\begin{aligned} H(\mathbb{Y}^\dagger) &= H(q_L^\dagger)/\mu(\mathbb{Y}^\dagger) = 1 - D(q_L^\dagger\|\{2^{-l}\})/\mu(\mathbb{Y}^\dagger) \\ &= 1 - \frac{1}{\mu(\mathbb{Y}^\dagger)}\sum_{l=1}^{\infty} q_L^\dagger(l)\big(\log q_L^\dagger(l) + l\big)\,. \end{aligned}$$

Define $\ell \equiv \lfloor 4\log(1/d)\rfloor$. It follows from Lemma V.29 that $\sum_{l=\ell+1}^{\infty} q_L^\dagger(l) l = O(d^3)$, leading to

$$H(\mathbb{Y}^\dagger) \ge 1 - \frac{1}{\mu(\mathbb{Y}^\dagger)}\sum_{l=1}^{\ell} q_L^\dagger(l)\big(\log q_L^\dagger(l) + l\big) + O(d^3)\,. \qquad (85)$$

Now, from Lemma V.9(i), we know that

$$|q_L^\dagger(l) - p_L^\dagger(l)| \le \kappa_2 d^{2-\epsilon/2} \qquad (86)$$

for $l < \ell$.

Taylor's theorem yields

$$\begin{aligned} &\sum_{l=1}^{\ell} q_L^\dagger(l)\big(\log q_L^\dagger(l) + l\big) \\ &= \frac{1}{\ln 2}\sum_{l=1}^{\ell}\left(\left(q_L^\dagger(l) - 2^{-l}\right) + 2^{l-1}\left(q_L^\dagger(l) - 2^{-l}\right)^2\right) \\ &\quad + O(d^{3-\epsilon}). \end{aligned}$$

Now, $\sum_{l=0}^{\infty} q_L^\dagger(l) = \sum_{l=0}^{\infty} 2^{-l} = 1$, so $\sum_{l=1}^{\ell}\left(q_L^\dagger(l) - 2^{-l}\right) = -\sum_{l=\ell+1}^{\infty}\left(q_L^\dagger(l) - 2^{-l}\right) = O(d^3)$, using Lemma V.29 and our choice of ℓ. Thus, the first term in the aforementioned Taylor expansion can be ignored. Using (86) for the second term, we obtain

$$\begin{aligned} &\sum_{l=1}^{\ell} q_L^\dagger(l)\big(\log q_L^\dagger(l) + l\big) \\ &= \frac{d^2}{2\ln 2}\sum_{l=1}^{\ell} 2^{-l}\left(-c_2 l/2 + l\ln l\right)^2 + O(d^{3-\epsilon}) \\ &= \frac{d^2}{2\ln 2}\sum_{l=1}^{\infty} 2^{-l}\left(-c_2 l/2 + l\ln l\right)^2 + O(d^{3-\epsilon}) \\ &= \frac{d^2}{2\ln 2}\left(\frac{3}{2}c_2^2 + \sum_{l=1}^{\infty} 2^{-l}\left((l\ln l)^2 - c_2 l^2\ln l\right)\right) \\ &\quad + O(d^{3-\epsilon})\,. \end{aligned} \qquad (87)$$

Plugging back into (85) and using $|\mu(\mathbb{Y}^\dagger) - 2| = O(d)$, we obtain

$$\begin{aligned} H(\mathbb{Y}^\dagger) \ge 1 - \frac{d^2}{4\ln 2}&\left(\frac{3}{2}c_2^2 + \sum_{l=1}^{\infty} 2^{-l}\left((l\ln l)^2 - c_2 l^2\ln l\right)\right) \\ &+ O(d^{3-\epsilon})\,. \end{aligned} \qquad (88)$$

We construct $\acute{\mathbb{X}}^\dagger \in \mathcal{S}_{\lfloor 1/d\rfloor}$ from $\mathbb{X}^\dagger$ by flipping a few bits as in the proof of Lemma IV.6. Using (81), the fraction of flipped bits, both in $\mathbb{X}^\dagger$ and in $\mathbb{Y}^\dagger$, is at most $\alpha = 2\mathbb{E}[\widetilde{L}\mathbb{I}(\widetilde{L} \ge \lfloor 1/d\rfloor)]/\lfloor 1/d\rfloor = O(2^{-1/(2d)}) = O(d^4)$. Proceeding as in the proof of Lemma IV.6, cf., (82) and (84), we have

$$\left|H(\acute{Y}^\dagger|(\acute{X}^\dagger)^n) - H(Y^\dagger|(X^\dagger)^n)\right| \le nh(\alpha) = nO(d^3)\,. \qquad (89)$$

For each bit that is flipped, the number of runs in Y can change by at most 2, and the number of runs of a particular length can change by at most 3. It follows that

$$\left|\frac{1}{\mu(\mathbb{Y}^\dagger)} - \frac{1}{\mu(\acute{\mathbb{Y}}^\dagger)}\right| \le 2\alpha = O(d^4),$$

and, for any positive integer l,

$$\left|\frac{q_L^\dagger(l)}{\mu(\mathbb{Y}^\dagger)} - \frac{\acute{q}_L^\dagger(l)}{\mu(\acute{\mathbb{Y}}^\dagger)}\right| \le 3\alpha = O(d^4)\,.$$

We then deduce from the above that

$$\left|\mu(\mathbb{Y}^\dagger) - \mu(\acute{\mathbb{Y}}^\dagger)\right| = O(d^4),$$

and for any $l > 0$,

$$\left|q_L^\dagger(l) - \acute{q}_L^\dagger(l)\right| \le \kappa_1 d^4,$$

where $\acute{q}_L^\dagger(\cdot)$ is the distribution of runs under $\acute{Y}$. From (86), it follows that for $l < \ell$,

$$\left|\acute{q}_L^\dagger(l) - p_L^\dagger(l)\right| \le 2\kappa_2 d^{2-\epsilon/2}\,. \qquad (90)$$

[3] The approximation error in (8) can be easily controlled for $\mathbb{X}^\dagger$.

We have $H(\acute{Y}^\dagger|(\acute{X}^\dagger)^n) = H(\acute{Y}^\dagger, K^\dagger|(\acute{X}^\dagger)^n) - H(K^\dagger|(\acute{X}^\dagger)^n, \acute{Y}^\dagger)$ where $K^\dagger \equiv K((\acute{X}^\dagger)^n)$. We use Corollary V.26 and Lemma V.18 to arrive at

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}H(\acute{Y}^\dagger|(\acute{X}^\dagger)^n)\\&= d\log(1/d) - \frac{d}{2}\sum_{l=2}^{\ell}\acute{q}_L^\dagger(l)\, l\log l + \frac{dc_2}{4\ln 2}\sum_{l=1}^{\ell}\acute{q}_L^\dagger(l)l\\&\quad+\left(1-\frac{c_2}{2}\right)\frac{d}{\ln 2} - \left(c_3+c_4+\frac{1}{2\ln 2}\right)d^2 + O(d^{3-\epsilon})\,.\end{aligned} \tag{91}$$

Combining (89)–(91), we obtain,

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}H(Y^\dagger|(X^\dagger)^n)\\&= d\log(1/d) - \frac{d}{2}\sum_{l=2}^{\ell}p_L^\dagger(l)\, l\log l + \frac{dc_2}{4\ln 2}\sum_{l=1}^{\ell}p_L^\dagger(l)l\\&\quad+\left(1-\frac{c_2}{2}\right)\frac{d}{\ln 2} - \left(c_3+c_4+\frac{1}{2\ln 2}\right)d^2 + O(d^{3-\epsilon})\,.\end{aligned}$$

A calculation yields

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}H(Y^\dagger|(X^\dagger)^n)\\&= d\log(1/d) + \left(1-\frac{c_2}{2}\right)\frac{d}{\ln 2}\\&\quad- d^2\Bigg(\frac{1}{4\ln 2}\Big[2+3c_2^2+2\sum_{l=1}^{\infty}2^{-l}\big((l\ln l)^2 - c_2 l^2\ln l\big)\Big]\\&\qquad+c_3+c_4\Bigg) + O(d^{3-\epsilon})\,.\end{aligned} \tag{92}$$

Finally,

$$I(\mathbb{X}^\dagger) = (1-d)H(\mathbb{Y}^\dagger) + \lim_{n\to\infty}\frac{1}{n}H(Y^\dagger|(X^\dagger)^n)\,.$$

The result now follows by using the estimates in (88) and (92).

We obtain

$$I(\mathbb{X}^\dagger) \geq 1 - d\log(1/d) - A_1 d + A_2 d^2 + O(d^{3-\epsilon}),$$

where

$$\begin{aligned}A_1 &= \log(2e) - \frac{c_2}{2\ln 2},\\A_2 &= -\frac{1}{4\ln 2}\left(\frac{3}{2}c_2^2 + \sum_{l=1}^{\infty}2^{-l}\left((l\ln l)^2 - c_2 l^2\ln l\right)\right)\\&\quad+c_3+c_4\\&\quad+\frac{1}{4\ln 2}\left(2+3c_2^2+2\sum_{l=1}^{\infty}2^{-l}\left((l\ln l)^2 - c_2 l^2\ln l\right)\right)\\&= c_3+c_4+\\&\quad\frac{1}{4\ln 2}\left(2+\frac{3}{2}c_2^2+\sum_{l=1}^{\infty}2^{-l}(l\ln l)^2 - c_2\sum_{l=1}^{\infty}2^{-l}l^2\ln l\right).\end{aligned}$$

■

Proof of Lemma IV.4: Let $\gamma_* = \sup\{\gamma : H(\mathbb{Y}) \geq 1 - d^\gamma\}$. Then $\gamma_* \geq 1 + \gamma_*/2 - \epsilon/2$ must hold, else Lemma V.27 leads to a contradiction. It follows that $\gamma_* \geq 2-\epsilon$, hence the result.

We use here the fact that d_0 in Lemma V.27 does not depend on γ. ■

Proof of Lemma IV.5: Fix $\epsilon > 0$. Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d\rfloor}$. Assume

$$I(\mathbb{X}) \geq 1 - d\log(1/d) - A_1 d - d^{2-(\epsilon/8)}\,.$$

(If not, we are done, for small enough d.)

By Lemma IV.4, we know that $H(\mathbb{Y}) > 1 - d^{2-(\epsilon/2)}$. Now, we use Lemma V.19, Corollary V.26 and Lemma V.18 for the three terms in (44), to arrive at

$$\begin{aligned}I(\mathbb{X}) &\leq 1 - d\log(1/d) - \frac{1}{2}\sum_{l=1}^{\infty}q_L(l)\big(\log q_L(l) + l\big)\\&\quad+\frac{d}{2}\sum_{l=2}^{4\log(1/d)}q_L(l)\, l\log l - \frac{dc_2}{4\ln 2}\sum_{l=1}^{4\log(1/d)}q_L(l)l\\&\quad+\tilde{c}_1 d + \tilde{c}_2 d^2 + \kappa_1 d^{3-\epsilon},\end{aligned} \tag{93}$$

where $\tilde{c}_1$ and $\tilde{c}_2$ can be explicitly computed in terms of aforementioned constants, and $\kappa_1 < \infty$ is independent of q_L. The precise value of these constants is irrelevant for the argument below.

Since we know that $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d\rfloor}$, Lemma V.13 tells us that the tail of q_L is small. Define $\ell \equiv \lfloor 8/d\rfloor$. We deduce that

$$\sum_{l=\ell+1}^{\infty}q_L(l) \leq d^4, \qquad \sum_{l=\ell+1}^{\infty}lq_L(l) \leq d^4,$$

for small enough d. From elementary calculus, we obtain

$$\begin{aligned}&\sum_{l=\ell+1}^{\infty}q_L(l)\big(\log q_L(l) + l\big)\\&\geq \sum_{l=\ell+1}^{\infty}q_L(l)\log\left(\frac{\sum_{l=\ell+1}^{\infty}q_L(l)}{2^{-\ell}}\right)\\&\geq \ell d^4 + d^4\log d^4\\&\geq d^{3-\epsilon/2}\,.\end{aligned} \tag{94}$$

From Lemma V.3, we deduce

$$\sum_{l=4\log(1/d)}^{\ell}q_L(l)l \leq d^{2-\epsilon}\,. \tag{95}$$

Plugging the bounds in (94) and (95) into (93), we obtain

$$\begin{aligned}I(\mathbb{X}) &\leq 1 - d\log(1/d) - \frac{1}{2}\sum_{l=1}^{\ell}q_L(l)\big(\log q_L(l) + l\big)\\&\quad+\frac{d}{2}\sum_{l=2}^{\ell}q_L(l)\, l\log l - \frac{dc_2}{4\ln 2}\sum_{l=1}^{\ell}q_L(l)l\\&\quad+\tilde{c}_1 d + \tilde{c}_2 d^2 + \kappa_2 d^{3-\epsilon},\end{aligned}$$

where $\kappa_2 < \infty$ is independent of q_L.

Now, we simply maximize the bound over “distributions" $\{q_L(l)\}_{l=1}^{\ell}$ satisfying $\sum_{l\leq\ell} q_L(l) \leq 1$, to arrive at an optimal distribution

$$q_L^*(l) = B(d)2^{-l}2^{d(-Sl/2+l\log l)}$$

for $l \leq \ell$, where $B(d)$ is such that $\sum_{l\leq\ell} q_L^*(l) = 1$, and $S = c_2/\ln 2$. Note that $q_L^*(l)$ has no dependence on the process $\mathbb{X}$ we started with.

It is easy to verify that

$$B(d) = 1 + O(d^{2-\epsilon/2})\,.$$

This leads to

$$q_L^*(l) = \begin{cases} 2^{-l}\big[1 + d(-c_2 l/2 + l\ln l) \\ \qquad + O(d^{2-\epsilon/2})\big] & \text{for } l \leq \ell, \\ 2^{-l/2}O(1) & \text{otherwise.} \end{cases}$$

We now have

$$\begin{aligned} I(\mathbb{X}) \leq{} & 1 - d\log(1/d) - \frac{1}{2}\sum_{l=1}^{\ell} q_L^*(l)\big(\log q_L^*(l) + l\big) \\ & + \frac{d}{2}\sum_{l=2}^{\ell} q_L^*(l)\; l\log l - \frac{dc_2}{4\ln 2}\sum_{l=1}^{\ell} q_L^*(l) l \\ & + \tilde{c}_1 d + \tilde{c}_2 d^2 + \kappa_4 d^{3-\epsilon}, \end{aligned} \tag{96}$$

for some $\kappa_4 < \infty$. Again, calculus yields

$$\begin{aligned} &\sum_{l=1}^{\lfloor 6\log(1/d)\rfloor} q_L^*(l)\big(\log q_L^*(l) + l\big) = \\ &\frac{d^2}{2\ln 2}\left(\frac{3}{2}c_2^2 + \sum_{l=1}^{\infty} 2^{-l}\left((l\ln l)^2 - c_2 l^2 \ln l\right)\right) + O(d^{3-\epsilon})\,. \end{aligned}$$

We substitute into (96) to get the result. ■

VI. Discussion

The previous best lower bounds on the capacity of the deletion channel were derived using first-order Markov sources. In contrast, we found that the optimal coding scheme for small d consists of independent runs with run-length distribution $p_L^{\dagger}(l) = 2^{-l}(1 + d(l\log l - c_2 l/2))$. This leads to the natural question “How much “loss" do we incur if we are only allowed to use an input distribution that is a first-order Markov source?"

The following theorem is fairly straightforward to prove using the results we have derived. It provides an upper bound on the rate achievable with a Markov source, and also a precise analytical characterization of the optimal Markov source for small d.

Theorem VI.1: Fix any $\epsilon > 0$. Consider the class of first-order Markov sources. There exists $\kappa < \infty$ and $d_0 \equiv d_0(\epsilon) > 0$, such that for and any $\mathbb{X}$ in this class,

$$I(\mathbb{X}) \leq 1 - d\log(1/d) - A_1 d + A_2' d^2 + \kappa d^{3-\epsilon}$$

holds for any $d < d_0$, where

$$\begin{aligned} A_2' &\equiv 2c_5^2/\ln 2 + c_3 + c_4 + 1/(2\ln 2), \\ c_5 &\equiv \frac{\ln 2}{4}\sum_{l=1}^{\infty}\left\{l(l-3)2^{-l}\log l\right\}\,. \end{aligned}$$

Denote the symmetric first-order Markov source with $\mathring{p}(d) \equiv \mathbb{P}(X_i = b|X_{i-1} = b) = 1/2 + c_5 d$ for $b \in \{0,1\}$, by $\mathring{\mathbb{X}}$. We have

$$I(\mathring{\mathbb{X}}) \geq 1 - d\log(1/d) - A_1 d + A_2' d^2 + \kappa d^{3-\epsilon}\,.$$

Numerical evaluation yields $A_2' \approx 1.57796256$ and $c_5 \approx 0.60409609$. We have $A_2 - A_2' \approx 0.10018339$, implying that the restriction to Markov sources leads to a rate loss of $0.10018339\,d^2$ bits per channel use, with respect to the optimal coding scheme.

Remark VI.2: Lower bounds are derived in [2] using Markov sources and “jigsaw" decoding. In this case, we can show (using [6] and Lemma V.18) that the best achievable rate is

$$1 - d\log(1/d) - A_1 d + (A_2' - c_4)d^2 + O(d^{3-\epsilon}),$$

and that $\mathring{\mathbb{X}}$ achieves this rate to within $O(d^{3-\epsilon})$. Thus, the lower bounds in [2] are off by $A_2 - A_2' - c_4 \approx 0.7902d^2$, to leading order.

Remark VI.3: The utility of our asymptotic analysis is confirmed by considering the prescription for the optimal Markov source $\mathring{\mathbb{X}}$ provided by Theorem VI.1. Drinea and Mitzenmacher [2] optimized numerically over Markov sources obtaining, for instance, $p = 0.53$ for $d = 0.05$. Our analytical prediction yields $\mathring{p}(0.05) \approx 0.530204804$.

In comparison, we have shown that $I(\mathbb{X}^{\dagger}) = C - O(d^{3-\epsilon})$. In fact, we conjecture that an even stronger bound holds.

Conjecture VI.4: $I(\mathbb{X}^{\dagger}) = C - \Theta(d^4)$

The reasoning behind this conjecture is as follows: We expect the next order correction to the optimal input distribution to be quadratic in d. If $I(\mathbb{X})$ is a “smooth" function of the input distribution, a change of order d^2 in the input distribution should imply that $I(\mathbb{X})$ decreases by an amount $\Theta((d^2)^2) = \Theta(d^4)$ below capacity.

Our work leaves several open questions.

1) Can the capacity be expanded as

$$C = 1 - d\log(1/d) - A_1 d + \sum_{i=2}^{\infty} A_i d^i$$

for small d? If yes, is this series convergent? In other words, is there a $d_0 > 0$ such that for all $d < d_0$, the infinite sum on the right has terms that decay exponentially in magnitude? We expect that the answer to both these questions is in the affirmative. We provide a very coarse reasoning for this below.

The analysis carried out in this paper suggests that the optimal input distribution for $d < d_0$ does not have “long range dependence." In particular, we expect correlations to decay exponentially in the distance between bits. Suppose we are computing contribution to capacity due to “clusters" of k nearby deletions. These “clusters" should correspond to k deletions occurring within $2k+1$ consecutive runs. This should give us a term $A_k d^k$ with the error being bounded by the probability of seeing $(k+1)$ deletions in $2k+1$ consecutive runs. This error should decay exponentially in k for $d < d_0$, assuming our hypothesis on correlation decay.

2) What is the next order correction to the optimal input distribution? It appears that this correction should be of order d^2 and should involve nontrivial dependence between the run-length distribution of consecutive runs. It would be illuminating to shed light on the type of dependence that would be most beneficial in terms of maximizing rate $I(\mathbb{X})$ achieved. Moreover, it appears that computing this correction heuristically may, in fact, be tractable, using some of the estimates derived in this study.
3) Can the results here be generalized to nonbinary alphabet, and to other channel models of insertions/deletions?
4) What about the deletion channel in the large deletion probability regime, i.e., $d \to 1$? What is the best coding scheme in this limit? It seems this limit may be harder to analyze than the $d \to 0$ limit studied in the present work: for $d = 1$, the channel capacity is 0 and there is no specific coding scheme that we can modify continuously in order to achieve good performance for d close to 1. This is in contrast to the case $d = 0$, where we know that the iid Bernoulli$(1/2)$ input achieves capacity.
5) We did not compute explicitly the constants in the error terms of our upper and lower bounds. As mentioned in Section I, it would be interesting to compute them. This would lead to improvements over existing upper and lower bounds on capacity.

APPENDIX

Proof of Theorem II.1: This is just a reformulation of Theorem 1 in [5], to which we add the remark $C = \inf_{n\geq 1} C_n$, which is of independent interest. In order to prove this fact, consider the channel W_{m+n}, and let $X^{m+n} = (X_1^m, X_{m+1}^{m+n})$ be its input. The channel W_{m+n} can be realized as follows. First the input is passed through a channel $\widetilde{W}_{m+n}$ that introduces deletions independently in the two strings X_1^m and X_{m+1}^{m+n} and outputs $\widetilde{Y}(X_1^{m+n}) \equiv (Y(X_1^m), |, Y(X_{m+1}^{m+n}))$ where $|$ is a marker. Then, the marker is removed.

This construction proves that W_{m+n} is physically degraded with respect to $\widetilde{W}_{m+n}$, whence

$$
\begin{aligned}
(m+n)C_{m+n} &\leq \max_{p_{X^{m+n}}} I(X^{m+n}; \widetilde{Y}(X_1^{m+n})) \\
&\leq mC_m + nC_n\,.
\end{aligned}
$$

Here, the last inequality follows from the fact that $\widetilde{W}_{m+n}$ is the product of two independent channels, and hence, the mutual information is maximized by a product input distribution.

Therefore, the sequence $\{nC_n\}_{n\geq 1}$ is subadditive, and the claim follows from Fekete's lemma. ■

Proof of Lemma II.2: This is essentially [5, Th. 5]. The proof is provided for the convenience of the reader. Take any stationary $\mathbb{X}$, and let $I_n = I(X^n; Y(X^n))$. Notice that $Y(X_1^m) - X_1^m - X_{m+1}^{n+m} - Y(X_{m+1}^{n+m})$ form a Markov chain. Define $\widetilde{Y}(X^{n+m})$ as in the proof of Theorem II.1. We, therefore, have

$$
\begin{aligned}
I_{n+m} &\leq I(X^{n+m}; \widetilde{Y}(X^{n+m})) \\
&\leq I(X_1^m; Y(X_1^m)) + I(X_{m+1}^{m+n}; Y(X_{m+1}^{m+n})) = I_m + I_n
\end{aligned}
$$

(the last identity follows by stationarity of $\mathbb{X}$). Thus, $I_{m+n} \leq I_n + I_m$ and the limit $\lim_{n\to\infty} I_n/n$ exists by Fekete's lemma, and is equal to $\inf_{n\geq 1} I_n/n$.

Clearly, $I_n \leq C_n$ for all n. Fix any $\varepsilon > 0$. We will construct a process $\mathbb{X}$ such that

$$
I_N/N \geq C - \varepsilon \qquad \forall\, N > N_0(\varepsilon), \tag{97}
$$

thus proving our claim.

Fix n such that $C_n \geq C - \varepsilon/2$. Construct $\mathbb{X}$ with iid blocks of length n with common distribution $p^*(n)$ that achieves the supremum in the definition of C_n. In order to make this process stationary, we make the first complete block to the right of the position 0 start at position s uniformly random in $\{1, 2, \ldots, n\}$. We call the position s the offset. The resulting process is clearly stationary and ergodic.

Now consider $N = kn + r$ for some $k \in \mathbb{N}$ and $r \in \{0, 1, \ldots, n-1\}$. The vector X_1^N contains at least $k-1$ complete blocks of size n, call them $x(1), x(2), \ldots, x(k-1)$ with $x(i) \sim p^*(n)$. The block $x(1)$ starts at position s. There will be further $r + n - s + 1$ bits at the end, so that $X_1^N = (X_1^{s-1}, x(1), x(2), \ldots, x(k-1), X_{s+kn}^N)$. We write $y(i)$ for $Y(x(i))$. Given the output Y, we define $\widetilde{Y} = (Y(X_1^{s-1})\text{¦}y(1)\text{¦}y(2)\text{¦}\ldots\text{¦}y(k-1)\text{¦}Y(X_{s+(k-1)n}^N))$, by introducing k synchronization symbols ¦. There are at most $(n+1)^k$ possibilities for $\widetilde{Y}$ given Y (corresponding to potential placements of synchronization symbols). Therefore, we have

$$
\begin{aligned}
H(Y) &= H(\widetilde{Y}) - H(\widetilde{Y}|Y) \\
&\geq H(\widetilde{Y}) - \log((n+1)^k) \\
&\geq (k-1)H(y(1)) - k\log(n+1),
\end{aligned}
$$

where we used the fact that the $(x(i), y(i))$'s are iid. Further

$$
H(Y|X^N) \leq H(\widetilde{Y}|X^N) \leq (k-1)H(y(1)|x(1)) + 2n,
$$

where the last term accounts for bits outside the blocks. We conclude that

$$
\begin{aligned}
I(X^N; Y(X^N)) &= H(Y) - H(Y|X^N) \\
&\geq (k-1)nC_n - k\log(n+1) - 2n \\
&\geq N(C_n - \varepsilon/2)
\end{aligned}
$$

provided $\log(n+1)/n < \varepsilon/10$ and $N > N_0 \equiv 10n/\varepsilon$. Since $C_n \geq C - \varepsilon/2$, this in turn implies (97). ■

In this short appendix, we recall a few basic facts about Palm measures. We refer to [17] and [18] for more substantial background.

For the sake of simplicity, we shall focus on the case of interest to us, namely the one of point processes on the integer line $\mathbb{Z}$. The key intuition is that there is two important ways to study such a process. The first one is to look at it from a "uniformly random point" on the line: this is the stationary view. The second is to look at it from a "uniformly random point" in the process. Of course, these intuitions must be formulated differently in order to be rigorous.

Formally, we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a random variable $\mathbb{S}$ that takes values in subsets of the integer line, i.e., $\mathbb{S} : \Omega \to \{0,1\}^{\mathbb{Z}}$, $\mathbb{S} : \omega \mapsto \mathbb{S}(\omega) \subseteq \mathbb{Z}$. The process $\mathbb{S}$ is measurable when we endow $\{0,1\}^{\mathbb{Z}}$ with the product sigma-algebra. Given $A \subseteq \mathbb{Z}$, its shift by ℓ units to the right is denoted by $R_\ell A$. We assume that $\mathbb{S}$ is stationary, i.e., that $R_\ell \mathbb{S}$ is distributed as $\mathbb{S}$ for each integer ℓ. In order to avoid trivial cases, we further assume that $\mathbb{S}$ is ergodic and nonempty.

Notice that, by stationarity, the quantity

$$\frac{1}{\mu} \equiv \mathbb{P}(i \in \mathbb{S}) \tag{98}$$

is well defined and independent of $i \in \mathbb{Z}$. Further, $\mu \in (0, \infty)$ because $\mathbb{S}$ is nonempty. The *Palm measure* $\mathbb{P}_1(\,\cdot\,)$ is defined as the conditional probability measure

$$\mathbb{P}_1(A) = \mathbb{P}(A | 1 \in \mathbb{S}) \,. \tag{99}$$

This corresponds to the idea of looking at $\mathbb{S}$ from one of the points in $\mathbb{S}$. Obviously, under $\mathbb{P}_1$, $1 \in \mathbb{S}$ almost surely and in particular under $\mathbb{P}_1$, $\mathbb{S}$ is not stationary.

The above defines $\mathbb{P}_1$ on the basis of $\mathbb{P}$. It is often useful to consider the reverse direction, i.e., construct $\mathbb{P}$ from $\mathbb{P}_1$. To this end, consider the random variable

$$L(\mathbb{S}) = \min\big\{ j \geq 2 \text{ such that } j \in \mathbb{S} \,\big\} - 1 \,. \tag{100}$$

In terms of the associated binary process $\mathbb{X}$, cf., Section II, $L(\mathbb{S})$ is the length of the run starting at 1. The distribution of L is $p_L(l)$. We then define a new probability measure $\mathbb{P}^*$ through its Radon–Nikodym derivative with respect to $\mathbb{P}_1$:

$$\frac{\mathrm{d}\mathbb{P}^*}{\mathrm{d}\mathbb{P}_1}(\omega) = \frac{L(\mathbb{S}(\omega))}{\mathbb{E}_1\{L\}}, \tag{101}$$

where $\mathbb{E}_1$ denotes expectation with respect to $\mathbb{P}_1$. The measure $\mathbb{P}^*$ corresponds intuitively to the following procedure: choose a “uniformly random" point i on the line, and set the origin at the first point on the left of i. The aforementioned Radon–Nikodym derivative corresponds to the fact that a uniformly random point is more likely to fall in a large interval between consecutive points of $\mathbb{S}$.

Finally, $\mathbb{P}$ is constructed from $\mathbb{P}^*$ by shifting the origin to a uniformly random point between 0 and $L(\mathbb{S}) - 1$. In formulae, for any measurable f,

$$\begin{aligned} &\int f(\mathbb{S}(\omega))\, \mathrm{d}\mathbb{P}(\omega) \\ &= \int \mathbb{E}_U f(R_{-U}\mathbb{S}(\omega))\, \mathrm{d}\mathbb{P}_*(\omega), \end{aligned} \tag{102}$$

with U a uniformly random variable in $\{0, 1, \dots, L(\mathbb{S}) - 1\}$. Hence

$$\begin{aligned} &\int f(\mathbb{S}(\omega))\, \mathrm{d}\mathbb{P}(\omega) \\ &= \frac{1}{\mathbb{E}_1\{L\}} \int \mathbb{E}_U f(R_{-U}\mathbb{S}(\omega))\, L(\mathbb{S}(\omega))\, \mathrm{d}\mathbb{P}_1(\omega) \,. \end{aligned} \tag{103}$$

As a sanity check, let us compute μ as defined by (98). To this end, we let $f(\mathbb{S}) = \mathbb{I}(1 \in \mathbb{S})$, and obtain, from (103),

$$\begin{aligned} \frac{1}{\mu} &= \mathbb{P}(1 \in \mathbb{S}) \\ &= \frac{1}{\mathbb{E}_1\{L\}} \int \mathbb{E}_U \mathbb{I}(1 + U \in \mathbb{S}(\omega))\, L(\mathbb{S}(\omega))\, \mathrm{d}\mathbb{P}_1(\omega) \\ &= \frac{1}{\mathbb{E}_1\{L\}} \int \mathbb{P}_U(U = 0)\, L(\mathbb{S}(\omega))\, \mathrm{d}\mathbb{P}_1(\omega) \\ &= \frac{1}{\mathbb{E}_1\{L\}} \int 1 \ \mathrm{d}\mathbb{P}_1(\omega) \\ &= \frac{1}{\mathbb{E}_1\{L\}} \,. \end{aligned}$$

In other words, $\mu = \mathbb{E}_1(L)$ as expected.

The proof of Lemma V.18 is quite intricate and requires us to define a new modification to the deletion process in terms of super runs.

Now, we define a new modification to the deletion process, and we call the resulting process $\breve{\mathbb{D}}$ the *perturbed* deletion process to avoid confusion with the modified deletion process $\widehat{\mathbb{D}}$.

The input process $\mathbb{X}$ is divided into super runs as $\dots, S_{-1}, S_0, S_1, \dots$ (cf., Definition IV.3). For all integers i, define:

$\breve{\mathbb{Z}}^i \equiv$ Binary process that is zero throughout except if (S_i, S_{i+1}, S_{i+2}) have three or more deletions in total, in which case $\breve{Z}^i_l = 1$ if and only if $X_l \in S_i$ and $D_l = 1$.

Define $\breve{\mathbb{Z}}$ by

$$\breve{Z}_l = \begin{cases} 1 & \text{if } \exists i \text{ s.t. } \breve{Z}^i_l = 1 \,, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, define $\breve{\mathbb{D}}(\mathbb{D}, \mathbb{X}) \equiv \mathbb{D} \oplus \breve{\mathbb{Z}}$ (where $\oplus$ is componentwise sum modulo 2). The output of the channel is simply defined by deleting from X^n those bits whose positions correspond to 1s in $\breve{\mathbb{D}}$. We define $\breve{K}$ for the perturbed deletion process similarly to K.

We make use of the following fact:

Proposition A.1: Consider any integer $m > 0$. Let $U_1, U_2, \dots, U_m$ be random variables, taking values in $\mathbb{N}$, that have the same marginal distribution, i.e., $U_i \sim U$ for $i = 1, 2, \dots, m$, and arbitrary joint distribution. Let $f_1, f_2, \dots, f_m : \mathbb{N} \to \mathbb{R}_+$ be nondecreasing functions. Then, we have

$$\mathbb{E}\Big[\prod_{i=1}^m f_i(U_i)\Big] \leq \mathbb{E}\Big[\prod_{i=1}^m f_i(U)\Big].$$

Proof of Proposition A.1: We prove the result for $m = 2$. The proof can easily be extended to arbitrary $m \in \mathbb{N}$.

We want to show that for random variables U and V, with $U \sim V$, and nondecreasing, nonnegative valued functions f, g, we have

$$\mathbb{E}[f(U)g(V)] \leq \mathbb{E}[f(U)g(U)].$$

Part I: Define $\mathcal{H} = \{f : \mathbb{E}[f(U)\mathbb{I}(V \geq b)] \leq \mathbb{E}[f(U)\mathbb{I}(U \geq b)]\ \forall b \in \mathbb{R}\}$.

Claim: The class $\mathcal{H}$ contains all nonnegative, nondecreasing functions f.

Proof of Claim:

(i) We have $\mathbb{I}_{[a,\infty)} \in \mathcal{H}\ \forall a \in \mathbb{R}$.

$$\begin{aligned}&\mathbb{E}[\mathbb{I}(U \geq a)\mathbb{I}(V \geq b)]\\ &\leq \min\left\{\mathbb{P}(U \geq b), \mathbb{P}(U \geq a)\right\}\\ &= \mathbb{P}(U \geq \max(a,b)) = \mathbb{E}[\mathbb{I}(U \geq a)\mathbb{I}(U \geq b)].\end{aligned}$$

(ii) If $f_1, f_2 \in \mathcal{H}$, then $c_1 f_1 + c_2 f_2 \in \mathcal{H}$ for any $c_1 > 0, c_2 > 0$.

This follows from linearity of expectation.

Define the class of "simple increasing functions"

$$\mathcal{I} \equiv \{f : \exists k \in \mathbb{N} \text{ s.t. } f = \sum_{i=1}^{k} c_i \mathbb{I}_{[a_i,\infty)} \text{ for some } c_i > 0,\ a_i \in \mathbb{R} \text{ for } i = 1, 2, \ldots, k\}.$$

(iii) It follows from (i) and (ii) that $\mathcal{I} \subseteq \mathcal{H}$.

Now, it is not hard to see that for any nonnegative nondecreasing f, we can find a monotone nondecreasing sequence of functions $(f_n)_{n=1}^{\infty} \in \mathcal{I}$ such that $f_n \uparrow f$. By the monotone convergence theorem, we have

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[f_n(U)\mathbb{I}(V \geq b)] &= \mathbb{E}[f(U)\mathbb{I}(V \geq b)],\\ \lim_{n\to\infty} \mathbb{E}[f_n(U)\mathbb{I}(U \geq b)] &= \mathbb{E}[f(U)\mathbb{I}(U \geq b)].\end{aligned}$$

Combining with (iii), we infer that $f \in \mathcal{H}$, proving our claim.

Part II: Define $\widehat{\mathcal{H}}_f = \{g : \mathbb{E}[f(U)g(V)] \leq \mathbb{E}[f(U)g(U)]\}$.

From Part I, we infer that $\mathbb{I}(V \geq b) \in \widehat{\mathcal{H}}_f$ for all $b \in \mathbb{R}$. We now repeat the steps in the proof of the Claim in Part I, to obtain the result "The class $\widehat{\mathcal{H}}_f$ contains all nonnegative, nondecreasing functions g." This completes our proof of the proposition. ■

Lemma A.2: There exists $d_0 > 0$ such that for any $d < d_0$ the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$. Then

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} H(\breve{K}(X^n)|X^n, \breve{Y}(X^n)) = \frac{d^2}{\mu(\mathbb{X})}\Bigg\{ &\sum_{k=2}^{\infty}\sum_{l_{k+1}=2}^{\infty} p_{L(k+2)}\big(1, 1, \ldots (k+1 \text{ ones}), l_{k+1}\big)\\ &\cdot \big(k - 1 + l_{k+1}\big)\, h\bigg(\frac{1}{k-1+l_{k+1}}\bigg)\\ &+ \sum_{l_0=2}^{\infty}\sum_{k=2}^{\infty}\sum_{l_{k+1}=2}^{\infty} p_{L(k+2)}\big(l_0, 1, 1, \ldots (k \text{ ones}), l_{k+1}\big)\\ &\cdot \big(l_0 + k - 1 + l_{k+1}\big)\, h\bigg(\frac{l_0+1}{l_0+k-1+l_{k+1}}\bigg)\Bigg\} + \delta \qquad (104)\end{aligned}$$

for some δ such that $|\delta| \leq 18 d^3 \mathbb{E}[\widetilde{L}^2]$.

Proof of Lemma A.2: Using the chain rule, we obtain

$$\begin{aligned}&H(\breve{K}(X^n)|X^n, \breve{Y}(X^n))\\ &= \sum_{j=1}^{M} H(|\breve{X}(j)|\,|\breve{X}(j)\ldots\breve{X}(M), \breve{Y}(j)\ldots\breve{Y}(M)).\end{aligned}$$

Consider the term

$$t_j \equiv H(|\breve{X}(j)||\breve{X}(j)\ldots\breve{X}(M), \breve{Y}(j)\ldots\breve{Y}(M)).$$

Suppose the first bit in $\breve{X}(j)\ldots$ is part of super run S_i. Let the first run in $\breve{X}(j)$ be R_P, with length L_P. By the construction of the perturbed deletion process, we know that S_i, S_{i+1} and S_{i+2} cannot have more than two deletions in total.

Different cases may arise:

1) $L_P < |\breve{Y}(j)|$.
If $L_P + L_{P+2} \geq |\breve{Y}(j)|$, then we know that $\breve{X}(j) = (R_P, R_{P+1}, R_{P+2})$. If not, then we know that $\breve{X}(j) = (R_P, R_{P+1}, R_{P+2}, R_{P+3}, R_{P+4})$. In either case, $t_j = 0$.

2) $L_P > |\breve{Y}(j)|$.
It must be that $\breve{X}(j) = R_P$. Again, $t_j = 0$

3) $L_P = |\breve{Y}(j)|$.
In this case, if $L_{P+1} > 1$ or $L_{P+2} > 1$, then we know that $\breve{X}(j) = R_P$ and $t_j = 0$. Suppose $L_{P+1} = L_{P+2} = 1$. Now consider the possibility that $\breve{X}(j) = (R_P, R_{P+1}, R_{P+2})$ (this is the only alternative to $\breve{X}(j) = R_P$). For this possibility to exist, the following condition must hold

$$\begin{aligned}\mathcal{C} \equiv \Big\{&\breve{Y}(j)\breve{Y}(j+1)\breve{Y}(j+2)\ldots \text{ must match exactly}\\ &R_P R_{P+3} R_{P+4}\ldots \text{ until the end of } S_{i+2}\Big\}\\ &\cap \{L_{P+1} = L_{P+2} = 1\}.\end{aligned}$$

(Else, we would need more than two deletions in (S_i, S_{i+1}, S_{i+2}), a contradiction.)

Note that in any case, there are at most two possibilities for $\breve{X}(j)$, so we have $t_j \leq 1$.

Let us understand $\mathcal{C}$ better. Let S_i include k runs to the right of R_P, i.e., $L_{P+1} = L_{P+2} = \cdots = L_{P+k} = 1$ and $L_{P+k+1} > 1$. Condition $\mathcal{C}$ can arise, along with $\breve{X}(j)$ starting at R_P if and only if:

1) run R_{P-1} does not disappear under $\breve{\mathbb{D}}$;
2) super runs (S_i, S_{i+1}, S_{i+2}) undergo no more than two deletions in total. Event E;
3) one of the following deletion patterns occur.
   - (Only if $L_P > 1$) The bit R_{P+1} is deleted and one deletion in R_P. Event E_1.
   - The bits R_{P+1} and R_{P+2} are deleted. Event E_2.
   - The bits R_{P+2} and R_{P+3} are deleted. Event E_3.

   $\vdots$

   - The bits R_{P+k-1} and R_{P+k} are deleted. Event E_k.
   - The bit R_{P+k} is deleted and one deletion in R_{P+k+1}. Event E_{k+1}.

Define $p_0 \equiv (1-d)^{\widetilde{L}_i + \widetilde{L}_{i+1} + \widetilde{L}_{i+2} - 2}$. It is easy to see that $\mathbb{P}(\mathsf{E}_1 \cap \mathsf{E}) = p_0 d^2 L_P$, $\mathbb{P}(\mathsf{E}_l \cap \mathsf{E}) = p_0 d^2$ for $l = 2, 3, \ldots, k$, and $\mathbb{P}(\mathsf{E}_{k+1} \cap \mathsf{E}) = p_0 d^2 L_{P+k+1}$. We know that exactly one of these has occurred. $(\mathsf{E}_1 \cap \mathsf{E}) \cup (\mathsf{E}_2 \cap \mathsf{E})$ leads to

$\check{X}(j) = (R_P, R_{P+1}, R_{P+2})$, whereas all other possibilities lead to $\check{X}(j) = R_P$. It follows that if $\mathcal{C}$ holds, $L_P = l_P$ and $L_{P+k+1} = l_{P+k+1}$,

$$t_j = h\left(\frac{l_P\mathbb{I}(l_P > 1) + 1}{l_P\mathbb{I}(l_P > 1) + k - 1 + l_{P+k+1}}\right).$$

Let R_P be a uniformly random run (cf., Section II). The probability of seeing $L_P = l_P$, exactly k runs of unit length after R_P, $L_{P+k+1} = l_{P+k+1}$ and $(\mathsf{E}_1 \cup \mathsf{E}_2 \cup \ldots \cup \mathsf{E}_k) \cap \mathsf{E}$ is

$$p_{L(k+2)}(l_P, 1, 1, \ldots (k \text{ ones}), l_{P+k+1})\; p_0 d^2 \cdot (l_P\mathbb{I}(l_P > 1) + k - 1 + l_{P+k+1}).$$

It is easy to see that $p_0 \in (1 - d(\widetilde{L}_i + \widetilde{L}_{i+1} + \widetilde{L}_{i+2}), 1)$. Also, the conditional probability of R_{P-1} not disappearing is in $(1-d, 1)$. Thus, the expected contribution of R_P to the sum is

$$d^2\left\{\sum_{l_P=1}^{\infty}\sum_{k=2}^{\infty}\sum_{l_{P+k+1}=2}^{\infty} p_{L(k+2)}(l_P, 1, 1, \ldots (k \text{ ones}), l_{P+k+1}) \cdot \left(l_P\mathbb{I}(l_P > 1) + k - 1 + l_{P+k+1}\right) \cdot h\left(\frac{l_P\mathbb{I}(l_P > 1) + 1}{l_P\mathbb{I}(l_P > 1) + k - 1 + l_{P+k+1}}\right)\right\} + \delta$$

where

$$0 \geq \delta \geq -d^3 E[(1 + \widetilde{L}_i + \widetilde{L}_{i+1} + \widetilde{L}_{i+2})(\widetilde{L}_i + \widetilde{L}_{i+1})] \geq -2d^3 E[(\widetilde{L}_i + \widetilde{L}_{i+1} + \widetilde{L}_{i+2})^2].$$

We have used $l_P + k - 1 + l_{P+k+1} \leq \widetilde{L}_i + \widetilde{L}_{i+1}$, and $p_0 \in (1 - d(\widetilde{L}_i + \widetilde{L}_{i+1} + \widetilde{L}_{i+2}), 1)$, and that the conditional probability of R_{P-1} not disappearing is in $(1-d, 1)$. Now using Fact A.1, we have $E[(\widetilde{L}_i + \widetilde{L}_{i+1} + \widetilde{L}_{i+2})^2] \leq 9\mathbb{E}[\widetilde{L}^2]$, yielding $|\delta| \leq 18d^3\mathbb{E}[\widetilde{L}^2]$. The result follows. ■

Corollary A.3: For any $\epsilon > 0$, there exists $d_0 \equiv d_0(\epsilon) > 0$, and $\kappa < \infty$ such that for any $d < d_0$ the following occurs: Consider any $\mathbb{X} \in \mathcal{S}_{\lfloor 1/d \rfloor}$ such that $H(\mathbb{X}) > 1 - d^{1-\epsilon}$ and $\max\{H(\mathbb{X}), H(\mathbb{Y})\} > 1 - d^{\gamma}$ for some $\gamma \in (1/2, 2)$. Then,

$$\lim_{n\to\infty}\frac{1}{n}H(\check{K}(X^n)|X^n, \check{Y}(X^n)) = \frac{d^2}{2}\Bigg\{ \sum_{k=2}^{\infty}\sum_{l_{k+1}=2}^{\infty} 2^{-(1+k+l_{k+1})}\left(k - 1 + l_{k+1}\right) h\left(\frac{1}{k - 1 + l_{k+1}}\right) + \sum_{l_0=2}^{\infty}\sum_{k=2}^{\infty}\sum_{l_{k+1}=2}^{\infty} 2^{-(l_0+k+l_{k+1})}\left(l_0 + k - 1 + l_{k+1}\right) \cdot h\left(\frac{l_0 + 1}{l_0 + k - 1 + l_{k+1}}\right)\Bigg\} + \eta \qquad (105)$$

for some η such that $|\eta| \leq \kappa d^{2+\gamma/2-\epsilon/2}$.

Proof of Corollary A.3: We prove the corollary assuming $H(\mathbb{Y}) > 1 - d^{\gamma}$. The proof assuming $H(\mathbb{X}) > 1 - d^{\gamma}$ is analogous.

Consider the second summation in (104). Define $\ell \equiv \lfloor 4\log(1/d)\rfloor$. Consider any term with $l_0 \leq \ell$, $k \leq \ell$, $l_{k+1} \leq \ell$. Using Lemma V.12(i) ((37)), we have

$$\left|p_{L(k+2)}\left(l_0, 1, 1, \ldots (k \text{ ones}), l_{k+1}\right) - 2^{-(l_0+k+l_{k+1})}\right| \leq d^{\gamma/2-\epsilon/4}$$

for $d < d_0(\epsilon)$. Note that d_0 does not depend on l_0, k, l_{k+1}. It follows that

$$\sum_{l_0=2}^{\ell}\sum_{k=2}^{\ell}\sum_{l_{k+1}=2}^{\ell} p_{L(k+2)}\left(l_0, 1, 1, \ldots (k \text{ ones}), l_{k+1}\right) \cdot \left(l_0 + k - 1 + l_{k+1}\right) h\left(\frac{l_0 + 1}{l_0 + k - 1 + l_{k+1}}\right) = \sum_{l_0=2}^{\infty}\sum_{k=2}^{\infty}\sum_{l_{k+1}=2}^{\infty} 2^{-(l_0+k+l_{k+1})}\left(l_0 + k - 1 + l_{k+1}\right) \cdot h\left(\frac{l_0 + 1}{l_0 + k - 1 + l_{k+1}}\right) + \delta_{21},$$

where $|\delta_{21}| \leq d^{\gamma/2-\epsilon/2}$.

We make use of Lemma V.16 to bound the error due to the missed terms. Let $\widetilde{l}_0$ be the length of the super run containing the initial run of length l_0. Clearly, $\widetilde{l}_0 \geq l_0 + k$. Let $\widetilde{l}_1$ be the length of the next super run to the right. Clearly, $\widetilde{l}_1 \geq l_{k+1}$. Now

$$\begin{aligned}&\{l_0 > \ell\} \text{ OR } \{k > \ell\} \text{ OR } \{l_{k+1} > \ell\}\\ &\Rightarrow \{l_0 + k + l_{k+1} > \ell\}\\ &\Rightarrow \{\widetilde{l}_0 + \widetilde{l}_1 > \ell\}.\end{aligned}$$

Also, $\left(l_0 + k - 1 + l_{k+1}\right) \leq \widetilde{l}_0 + \widetilde{l}_1$ and $h(p) \leq 1$ for any p. It follows that the missed terms contribute

$$\delta_{22} \leq \sum_{\widetilde{l}_0+\widetilde{l}_1 \geq 4\ell} p_{\widetilde{L}(2)}(\widetilde{l}_0, \widetilde{l}_1)\left(\widetilde{l}_0 + \widetilde{l}_1\right) \leq d^{\gamma/2-\epsilon/2}$$

to the sum, where we have used Lemma V.16 in the second inequality.

Thus, we have established

$$\sum_{l_0=2}^{\infty}\sum_{k=2}^{\infty}\sum_{l_{k+1}=2}^{\infty} p_{L(k+2)}\left(l_0, 1, 1, \ldots (k \text{ ones}), l_{k+1}\right) \cdot\left(l_0 + k - 1 + l_{k+1}\right) h\left(\frac{l_0 + 1}{l_0 + k - 1 + l_{k+1}}\right) = \sum_{l_0=2}^{\infty}\sum_{k=2}^{\infty}\sum_{l_{k+1}=2}^{\infty}\left\{2^{-(l_0+k+l_{k+1})}\left(l_0 + k - 1 + l_{k+1}\right) \cdot h\left(\frac{l_0 + 1}{l_0 + k - 1 + l_{k+1}}\right)\right\} + \delta_2$$

with $|\delta_2| \leq 2d^{\gamma/2-\epsilon/2}$ for $d < d_0(\epsilon)$. The first summation in (104) can be similarly handled. Finally, Lemma V.12(ii) tells us that $|\mu(\mathbb{X}) - 2| \leq d^{\gamma/2}$ for small enough d. Putting the estimates together yields the result. ■

Proof of Lemma V.18: We prove the lemma assuming $H(\mathbb{Y}) > 1 - d^{\gamma}$. The proof assuming $H(\mathbb{X}) > 1 - d^{\gamma}$ is analogous.

It is easy to verify that the right hand side of (105) is, in fact, $d^2 c_4 + \eta$. We show that

$$\lim_{n\to\infty}\frac{1}{n}\left|H(\check{K}(X^n)|X^n, \check{Y}(X^n)) - H(K(X^n)|X^n, Y(X^n))\right| \leq 3d^{1+\gamma-\epsilon/2} \qquad (106)$$

whence (46) follows using Corollary A.3.

Consider $\breve{Z}^n$ defined in our construction of the perturbed deletion process. We define $U(X^n, D^n, Z^n) \in \{\mathtt{t}, 0, 1\}^{|\breve{Y}|}$ constructed as follows: start from the first bit in $\breve{Y}$ and consider bits sequentially

1) for each bit also present in Y, U has a $\mathtt{t}$;
2) for each bit not present in Y, U has 0 if that bit 0 and a 1 if that bit is 1.

Clearly, the corresponding stationary process $\mathbb{U}$ can also be defined.

Recall Fact V.25. It is not hard to see that

$$(X^n, Y) \overset{U}{\longleftrightarrow} (X^n, \widehat{Y}) \qquad \text{and}$$
$$(X^n, Y, K) \overset{(U,Z)}{\longleftrightarrow} (X^n, \widehat{Y}, \widehat{K})\,.$$

It follows that

$$|H(\widehat{K}(X^n)|X^n, \widehat{Y}(X^n)) - H(K(X^n)|X^n, Y(X^n))| \le 2H(U) + H(Z)\,.$$

Let $\breve{z} \equiv \mathbb{P}[\breve{Z}_j = 1]$ for arbitrary j. The number of deletions reversed in a random super run is at most $d^3 \sum_{l_0,l_1,l_2} p_{\widetilde{L}(3)}(l_0, l_1, l_2)(l_0 + l_1 + l_2)^3$ in expectation (similar to (48)). Using Proposition A.1, this is bounded above by $27d^3\mathbb{E}[\widetilde{L}^3]$. Since each super run has length at least one, it follows that $\breve{z} \le 27d^3\mathbb{E}[\widetilde{L}^3]$. Using Lemma V.16 and $\widetilde{L} \le 1/d$ w.p. 1, we find that $\mathbb{E}[\widetilde{L}^3] \le d^{\gamma-2}$ for small enough d. Hence, $\breve{z} \le 27d^{1+\gamma}$. It follows that $H(\breve{\mathbb{Z}}) \le h(\breve{z}) \le d^{1+\gamma-\epsilon/2}$ for small enough d.

Let $u \equiv \mathbb{P}(U_j \ne \mathtt{t})$ for arbitrary j. Then, $u = \breve{z}/(1-d)$. It follows that $H(\mathbb{U}) \le u + h(u) \le d^{1+\gamma-\epsilon/2}$ for small enough d. Finally, we have

$$\lim_{n\to\infty} \frac{2H(U) + H(Z)}{n} = 2(1-d)H(\mathbb{U}) + H(\mathbb{Z}) \le 3d^{1+\gamma-\epsilon/2},$$

leading to the desired bound (106). ■

Acknowledgment

We would like to thank M. Mitzenmacher for introducing the us to the deletion channel, and later directing our attention to the reference [6]. We would like to thank the three anonymous reviewers, whose careful reading and suggestions greatly improved the paper.

References

[1] M. Mitzenmacher, “A survey of results for deletion channels and related synchronization channels,” *Probab. Surv.*, vol. 6, pp. 1–33, 2009.

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

[3] N. Ma, K. Ramchandran, and D. Tse, “Efficient file synchronization: A distributed source coding approach,” *IEEE Int. Symp. Inf. Theory*, pp. 583–587, 2011.

[4] G. Han and B. H. Marcus, “Asymptotics of entropy rate in special families of hidden Markov chains,” *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1287–1295, Mar. 2010.

[5] R. L. Dobrushin, “Shannon’s theorems for channels with synchronization errors,” *Problemy Peredachi Inf.*, vol. 3, pp. 18–36, 1967.

[6] A. Kirsch and E. Drinea, “Directly lower bounding the information capacity for channels with I.I.D. deletions and duplications,” *IEEE Int. Symp. Inf. Theory*, pp. 1731–1735, 2007.

[7] R. G. Gallager, Sequential Decoding for Binary Channels with Noise and Synchronization Errors Lincoln Lab., 1961.

[8] S. Diggavi and M. Grossglauser, “On transmission over deletion channels,” in *Proc. Aller. Conf. Commun., Control, Comput.*, 2001, pp. 573–582.

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

[10] E. Drinea and M. Mitzenmacher, “A simple lower bound for the capacity of the deletion channel,” *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4657–4660, Oct. 2006.

[11] D. Fertonani and T. M. Duman, “Novel bounds on the capacity of the binary deletion channel,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2753–2765, Jun. 2010.

[12] M. Dalai, “A new bound for the capacity of the deletion channel with high deletion probabilities,” *IEEE Int. Symp. Inf. Theory*, pp. 499–502, 2011.

[13] S. Diggavi, M. Mitzenmacher, and H. Pfister, “Capacity upper bounds for deletion channels,” *IEEE Int. Symp. Inf. Theory*, pp. 1716–1720, 2007.

[14] Y. Kanoria and A. Montanari, “On the deletion channel with small deletion probability,” *IEEE Int. Symp. Inf. Theory*, pp. 1002–1006, 2010.

[15] A. Kalai, M. Mitzenmacher, and M. Sudan, “Tight asymptotic bounds for the deletion channel with small deletion probabilities,” *IEEE Int. Symp. Inf. Theory*, pp. 997–1001, 2010.

[16] R. L. Dobrushin, “A general formulation of the fundamental theorem of Shannon in the theory of information,” *Uspekhi Mat. Nauk.*, vol. 14, pp. 3–104, 1959.

[17] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*. New York, NY, USA: Springer-Verlag, 2008.

[18] F. Baccelli and P. Brémaud, *Elements of Queuing Theory*. New York, NY, USA: Springer-Verlag, 2003.

Yashodhan Kanoria (S’07) is an Assistant Professor in the Decision, Risk and Operations Division at Columbia Business School. He obtained a PhD in Electrical Engineering at Stanford University in 2012.

He was awarded a Student Paper award for the conference version of this paper at the IEEE International Symposium on Information Theory, 2010. His current research interests include matching markets, social networks, probability and game theory.

Andrea Montanari (SM’13) is an associate professor in the Departments of Electrical Engineering and of Statistics, Stanford University. He received the Laurea degree in physics in 1997, and the Ph.D. degree in theoretical physics in 2001, both from Scuola Normale Superiore, Pisa, Italy. He has been a Postdoctoral Fellow with the Laboratoire de Physique Théorique of Ecole Normale Supérieure (LPTENS), Paris, France, and the Mathematical Sciences Research Institute, Berkeley, CA. From 2002 to 2010 has been Chargé de Recherche at LPTENS. In September 2006, he joined the faculty of Stanford University. Dr. Montanari was co-awarded the ACM SIGMETRICS Best Paper Award in 2008. He received the CNRS Bronze Medal for Theoretical Physics in 2006 and the National Science Foundation CAREER award in 2008. His research focuses on algorithms on graphs, graphical models, statistical inference and estimation.