On Information Transmission over a Finite Buffer Channel

Suhas Diggavi* Matthias Grossglauser†
School of Computer and Communication Sciences (I&C)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
{suhas.diggavi,matthias.grossglauser}@epfl.ch

Abstract

We study information transmission through a finite buffer queue. We model the channel as a finite-state channel whose state is given by the buffer occupancy upon packet arrival; a loss occurs when a packet arrives to a full queue. We study this problem in two contexts; one where the state of the buffer is known at the receiver, and the other where it is unknown. In the former case, we show that the capacity of the channel depends on the long-term loss probability of the buffer. Thus, even though the channel itself has memory, the capacity depends only on the stationary loss probability of the buffer. The main focus of this paper is on the latter case.

When the receiver does not know the buffer state, this leads to the study of deletion channels, where symbols are randomly dropped and a subsequence of the transmitted symbols is received. In deletion channels, unlike erasure channels, there is no side-information about which symbols are dropped. We study the achievable rate for deletion channels, and focus our attention on simple (mismatched) decoding schemes. We show that even with simple decoding schemes, with i.i.d. input codebooks, the achievable rate in deletion channels differs from that of erasure channels by at most $H_0(p_d) - p_d \log \frac{K}{K-1}$ bits, for $p_d < 1 - K^{-1}$, where p_d is the deletion probability, K is the alphabet size and $H_0(\cdot)$ is the binary entropy function. Therefore the difference in transmission rates between the erasure and the deletion channels is not large for reasonable alphabet sizes. We also develop sharper lower bounds with the simple decoding framework for the deletion channel by analyzing it for Markovian codebooks. Here it is shown that the difference between the deletion and erasure capacities is even smaller than that with i.i.d. input codebooks and for a larger range of deletion probabilities. We also examine the noisy deletion channel where a deletion channel is cascaded with a symmetric

*Contact author, School of Computer and Communication Sciences (I&C), EPFL, Lausanne, Switzerland. Email: suhas.diggavi@epfl.ch

†School of Computer and Communication Sciences (I&C), EPFL, Lausanne, Switzerland. Email: matthias.grossglauser@epfl.ch

discrete memoryless channel (DMC). We derive a single letter expression for an achievable rate for such channels. For the binary case, we show that this result simplifies to $\max(0, 1 - [H_0(\theta) + \theta H_0(p_e)])$ where p_e is the cross-over probability for the binary symmetric channel.

1 Introduction

In a packet-switched communication network, such as the Internet, the source of a session encodes information in a set of packets, which are transported as independent units through a set of links to reach their destination. A packet reaches its destination if there exists a route to the destination, and if there is buffer space available at every node along the path followed by this packet. The context motivating our problem formulation is that of a packet-switched communication network where packet flows *share* resources, which gives rise to random packet losses due to the randomness of packet arrivals to buffers in the network, and the effects of congestion control protocols such as TCP that regulate the packet generation rate of flows. We assume that (a) a packet either reaches its destination or is lost completely, and (b) the original order of packets is conserved. We propose an abstraction for this finite buffer channel, and examine reliable transmission rates over this channel. We ignore the possibility of information transmission through timing (*i.e.,* through interarrival times), but do allow for coding of successive packets[1].

We study this problem in two scenarios, one when the locations of the packet drops are known at the receiver, and the other when they are not. In the case where the location of the packet drops is known, we formulate this problem as transmission over a finite-state channel where the transitions of the finite-state channel occur due to arrivals and departures of packets to the buffer. This is a finite-state channel with memory whose transitions are *not* Markovian in general. We show that under some regularity conditions, the capacity is determined by the long term stationary loss probability of the buffer. Thus we get the intuitively satisfying result that the capacity is the product of the capacity of the DMC and the long term probability of a packet not being dropped[2]. This is the case even when we allow for feedback.

The main limitation of the erasure channel as a model for a finite-buffer queue (or a sequence thereof) is that there is no mechanism in the network to explicitly signal a dropped packet to the destination. Rather, transport protocols such as TCP use *sequence numbers* in the packet header[3] to detect lost packets. The sequence number uses up a certain number of bits to detect lost packets

[1]Transmission of information through inter-arrival times, though elegant, is not optimal when the packet sizes (*i.e.,* alphabet size of the transmitted symbol) is large enough. This was demonstrated in Theorem 10 in [1], where it was shown that the capacity of the queue can be achieved without coding for timing information when the alphabet size is large. This is the case in current networks, where packet sizes range from a few tens of bytes to a few thousand. Furthermore, it would be difficult to characterize the timing transfer function for a flow of packets in a multihop packet-switched network shared by many other traffic flows.

[2]This is akin to the result showed for memoryless erasure networks in [2].

[3]Strictly speaking, in the TCP header.

and to request retransmission of those packets. A fundamental question therefore arises: if we do not assume *a-priori* the existence of sequence numbers, what is the capacity of the resulting channel? This question naturally leads to the *deletion channel*, which essentially differs from the erasure channel in that the destination receives no explicit symbol indicating loss of a packet. Instead, the received sequence of symbols is shorter than the original sequence, with deleted symbols simply removed.

The deletion channel is a special case of insertion/deletion/substitution channels[4] which model the effect of synchronization errors and have a long history [3, 4, 5, 6, 7, 8]. A coding theorem for such channels in terms of maximizing mutual information over input distributions was proved in [6]. However, even in the presence of i.i.d. deletions, this does not lead to a single-letter characterization for achievable rates. Gallager, in an unpublished report [4], analyzed the performance of convolutional codes over insertion/deletion/substitution channels. He proposed adding a pseudo-random sequence to convolutional codes to correct such synchronization errors and derived lower bounds for achievable rates using sequential decoding for these codes. A similar idea was studied in "watermarking" codes proposed in [9], where LDPC codes and iterative decoding were used. Zigangirov [7] studied a more general insertion/deletion channel and improved the lower bounds of [4] for the performance of convolutional codes with sequential decoding. The bounds of [4, 7] coincide for i.i.d. deletion channels and are given by

$$C_{del} \geq 1 - H_0(p_d), \quad p_d \leq 0.5 \tag{1}$$

where $H_0(p_d) = -(1-p_d)\log(1-p_d) - p_d \log(p_d)$ is the binary entropy function. Ullman [5] studied the binary insertion/deletion channel from a zero-error point of view rather than vanishing error probability. He provided combinatorial upper bounds to insertion/deletion channels when asymptotically (in the codeword block size) the number of synchronization errors is a fraction of the codeword block size. His bounds are (see (33) and (44) in [5])

$$1 - p\log_2 e^2 \left(\frac{3}{2p} + \frac{15}{16}\right)\left(\frac{3}{2p} + \frac{47}{16}\right) \leq C \leq 1 - (1+p)\log_2(1+p) + p\log_2(2p), \tag{2}$$

where p in his notation is the fraction of total insertion/deletion errors asymptotically in the codeword block size. The zero-error rate bounds are more pessimistic than the bounds when the error is allowed to vanish asymptotically and hence the lower bound (1) in [4, 7] is sharper than (2).

The insertion/deletion/substitution channel was also pioneered by Levenshtein [3], where asymptotic bounds in the number of codewords capable of correcting up to a *finite* number of synchronization errors was studied. He also provided number-theoretic constructions for such codes and motivated a large body of literature on this topic (see [8] for a recent survey of such code constructions).

Our focus in this paper is on asymptotic information-theoretic bounds rather than on code construction, for the case when the number of deletions is a non-zero fraction of the codeword block

[4]In an insertion channel, additional symbols can randomly be inserted into the codeword. Substitutions are the familiar symbol errors for noisy channels.

size. More specifically, we are interested in achievable rates when the error probability vanishes to zero asymptotically, and not in zero-error achievable rates. Our main results are the following. In Section 4.2, we provide an alternative (and simpler) proof for the Gallager-Zigangirov result given in (1). The proof is based on a random coding argument which analyzes an i.i.d. (memoryless) codebook with a simple (not maximum likelihood, *i.e.,* mismatched) decoder. The advantages of this proof technique are two-fold. The simple decoder analysis is easily extended to larger (non-binary) alphabet sizes. It is also applicable when the deletion process is only stationary and ergodic (but not necessarily i.i.d.). Moreover, the simple decoding technique is more efficient than optimal (maximum likelihood) detection. Our key insight is that the transmission capacity of deletion channels is quite close to that of the erasure channel. That is, the penalty in achievable rate for the decoder not knowing the packet losses is small, though the code design problem is much harder. We do examine the deletion channel when the optimal decoding is used, and unfortunately we have not found any single letter characterization of the achievable rate for this case. However, we extend our framework of analyzing simple decoding techniques to input codebooks with memory, using Markovian codebooks. In Section 4.3 we derive bounds that improve (1) by using such codebooks. An important component of this analysis is the study of common subsequences between independent Markov processes, which might be of independent interest. One of the main insights of this results is that codebooks with memory can significantly increase the achievable rate for deletion channels. Our result shows that even with a first order Markovian codebook, we can *provably* show significant improvement in the achievable rate. Moreover, this is the first result to show that the achievable rate of the deletion channel is non-zero even for deletion probabilities p_d close to 1. In Section 5, we analyze the noisy deletion channel where a deletion channel is cascaded with a symmetric DMC. We give a single letter expression for the achievable rate which naturally generalizes the results for i.i.d. input codebooks of Section 4.2. In summary, the main contribution of this paper is the development of analysis techiniques that allow for single letter expressions for achievable rates over (noisy) deletion channels. A preliminary version of this work had appeared in [10], and since then there have been some interesting follow-up work [11, 12] which have further developed on our framework.

The remainder of this paper is structured as follows. Section 2 formally states the problem. In Section 3, we analyze the erasure channel, which models the transmission over a finite buffer channel with receiver side-information about packet losses. In Section 4, we present the main results of the paper, which are the analysis of achievable rates over deletion channels, including some illustrative numerical results. In Section 5, we generalize some of our results to the noisy deletion channel, which is the concatenation of a deletion and a discrete memoryless channel (DMC). Section 6 concludes the paper with a discussion of several open issues.

2 Problem Formulation

In the finite buffer channel, whether a packet (symbol) gets through to its destination depends on whether the buffer in a router is full when the packet arrives, in which case the packet is dropped; otherwise, the packet is delivered, possibly subject to some corruption for other reasons, such as bit errors in optical fibers or fading effects over a wireless link.

The finite buffer queue viewed as a channel is reminiscent of the finite-state Markov channel. The state of such a channel is governed by an underlying Markov chain. Information transmission over such channels has been studied for the case where the channel state evolves independently of its input and output [13, 14].

For the finite buffer channel, the channel state is determined by the state of the buffer. However, there is in general no reason why the finite buffer channel should be driven by a Markov process, given that the evolution of the queue depends on the arrival processes of all the flows sharing the queue, which in turn depend on the other queues these flows have already traversed, as well as other effects, such as congestion control protocols regulating the rate of packet generation. Hence the channel memory could be much more complicated. Therefore, we need to rely on weaker regularity conditions on the state process under which a coding theorem exists. Note that in our case, channel memory need not be finite and hence the results of finite-state channels (with finite memory) given in [15] are not directly applicable. However, we use results from [16] which allow the state process to have longer memory. More details are given in Section 3.

In Section 3, we make the assumption that the receiver knows when the packet is dropped, *i.e.,* it receives an explicit erasure symbol. In practice, this is achieved through a sequence number associated with each packet to allow the receiver to detect missing packets. This channel is equivalent to an *erasure channel*, albeit with complicated memory.

If such side information about dropped packets is not available at the decoder then the channel is equivalent to a *deletion channel*. The K-ary deletion channel is defined as follows. Let $x^n = (x_1, \ldots, x_n)$ be a codeword, where $x_i \in \{1, \ldots, K\}$. A *deletion pattern* d^n is a binary vector $(d_1, \ldots, d_n)$, where $d_i = 1$ indicates that the i-th symbol of x is deleted, and $d_i = 0$ indicates that the i-th symbol is received at the output. Let e be the total number of 1's in d^n, *i.e.*, the number of deletions. We define $i(k)$ as the position of the k-th 0 in the sequence d; clearly $0 \leq k \leq n - e$. Then the received sequence is $\boldsymbol{Y} = (y_1, \ldots, y_{n-e})$, with $y_k = x_{i(k)}$, $1 \leq k \leq n - e$. In other words, $\boldsymbol{Y}$ is a sequence of length $n - e$ containing the non-deleted symbols in x. The difference between erasure and deletion channels is illustrated in Figure 1.

We are mainly interested in i.i.d. distributions (with $\mathbb{P}\{D_i = 1\} = p_d$) for the binary sequence D_i, but results in Section 4.2 also apply when D^n is stationary and ergodic. Note that the deletion channel has memory in that $p(\boldsymbol{Y}|x^n)$ does not become a product distribution even for an i.i.d. deletion process. Intuitively, this is because the k-th output symbol y_k depends on the entire history of the deletion process up to $i(k)$. In Section 4, we focus on the pure deletion channel, *i.e.,*

erasure channel		**deletion channel**	
X^n	**0 0 1 0 1 1 1 1 0 0 1**	X^n	**0 0 1 0 1 1 1 1 0 0 1**
Q^n	**0 0 1 0 0 1 1 0 0 0 1**	D^n	**0 0 1 0 0 1 1 0 0 0 1**
Y^n	**0 0 E 0 1 E E 1 0 0 E**	Y^M	**0 0 0 1 1 0 0**

Figure 1: The output of a binary erasure and of a binary deletion channel for identical input X^n and identical erasure/deletion pattern $Q^n = D^n$. Note that the length of the output of the erasure channel is n, while for the deletion channel it is random.

where the symbols that are not deleted are received without error. Section 5 gives some results on the noisy deletion channel where there is a DMC following the deletion channel.

We define a $(|\mathcal{C}|, n)$ code as a set of $|\mathcal{C}|$ codewords $\mathcal{C} = \{x(1), \ldots, x(|\mathcal{C}|)\}$ of length n. The encoding function results in a codeword $x(j)$ to be sent when a message $j \in \{1, \ldots, |\mathcal{C}|\}$ is drawn. The deletion process D causes the random sequence $Y = y$ of length at most n to be received. We also define a decoding function $\hat{W} : \mathcal{S} \to \{1, \ldots, |\mathcal{C}|\}$, where $\mathcal{S}$ is the set of all K-ary sequences of length at most n. The average probability of error for a given codebook $\mathcal{C}$ and decoding function $\hat{W}$ is defined as

$$P_e(\mathcal{C}, \hat{W}) = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \mathbb{P}\left\{\hat{W}(Y) \neq j | X = x(j)\right\}. \tag{3}$$

We define a rate R to be achievable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes and a decoding rule $\hat{W}$ such that the average probability of error $P_e(\mathcal{C}, \hat{W})$ tends to 0 as $n \to \infty$.

We use the following notation in the remainder of the paper:

i	sample index of point process
j	index of a codeword in the codebook $\mathcal{C}$
Q_i	state of erasure channel upon arrival of packet i (1: packet erased; 0: packet gets through)
D_i	state of deletion channel upon arrival of packet i (1: packet deleted; 0: packet gets through)
p_d	probability of packet drop
$\theta = 1 - p_d$	probability that packet gets through
n	block size
X_i	information packet
Y_i	received information packet, or erasure symbol E for dropped packet
Z_i	received information packet from concatenation of deletion channel and DMC
$\boldsymbol{Y}, \boldsymbol{Z}$	sequence of received packets of *random* length
X^n, D^n	sequence of length n of information packet and deletion states
K	alphabet size of X_i
$H_0(.)$	binary entropy function, $H_0(x) = -x \log x - (1-x)\log(1-x)$
$\|x\|$	the length of a sequence x
weight of x	the number of 1's in x
$y = d \circ x$	y is the result of applying the deletion pattern d to codeword x
Δx	the derivative of x, defined as a sequence of length $\|x\| - 1$, whose l^{th} component is 1 if $x_{l+1} \neq x_l$, and 0 otherwise

3 The erasure channel with memory and feedback

In this section we study the finite-state model described in the previous section, where we assume that the decoder knows which symbols (packets) have been erased, because it receives an explicit erasure symbol E for each such symbol. Given the finite-state model, we calculate the capacity as the maximum mutual information between the input and the output. In order for this to make sense, there has to be some regularity conditions imposed on the state process. In this context, there is a result in [16] (see also [17]) which states that if the input process $\{X_i\}$ is stationary and ergodic, and the state process $\{Q_i\}$ is weakly mixing and stationary, then the output process $\{Y_i\}$ is jointly stationary and ergodic with the input process $\{X_i\}$. In this case, there is information stability, which ensures that the mutual information is the operational rate [18].

We use this result to compute the capacity of the finite buffer channel, where we also allow for receiver feedback. We make the following definition. Let θ denote the long-term fraction of packets that are *not* dropped, i.e.,

$$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\{Q_i = 0\}. \tag{4}$$

We consider causal feedback, *i.e.,* when the received sequence $\{y_i\}$ is fed back noiselessly to the transmitter after it has been received. Therefore the transmitter knows the buffer state in the

previous transmission but not the current state of the buffer. Given that the capacity is determined by the long term loss probability of the buffer and the capacity of the DMC, it seems unlikely that feedback would help to improve this. The following result formalizes this intuition.

Proposition 3.1 *If the state process $\{Q_i\}$ is stationary and weakly mixing, then the capacity of the finite buffer channel with feedback is given by*

$$C = \theta C_0, \tag{5}$$

where C_0 denotes the capacity of the DMC. The capacity is achieved for an i.i.d. input process $\{X_i\}$ that achieves the capacity of the DMC.

This result can be generalized if the probability distribution maximizing the mutual information for the DMC for each of the states is identical. Under this condition the capacity is given by

$$C = \sum_{q=0}^{Q-1} \pi_q C^{(q)}$$

where π_q is the long term probability of the state to be q and $C^{(q)}$ is the capacity of that state. This does not hold if this compatibility condition is not met.

Finally, we note that (47) is the capacity per arrival of the input. This can be converted into capacity per unit time using arguments similar to Theorem 5 in Appendix 6.2 in [19].

4 The deletion channel

In this section, we develop lower bounds for the "noiseless" deletion channel formulated in Section 2. In Section 5 we analyze the noisy deletion channel for some special cases. We start in Section 4.1 with achievable rates for optimal detection. The rest of the section is focused on the analysis of a simpler decoding scheme which uses subsequence matching. This framework not only allows us to assess the performance of low complexity decoders, it also gives a single letter characterization of its achievable rate (which lower bounds the capacity of the deletion channel). In Section 4.2, we analyze the simpler decoder for i.i.d. input codebooks. This bound is valid even when there is memory in the deletion process, i.e., when the deletion process is just assumed to be stationary and ergodic. Given that input codebooks with memory would do better, in Section 4.3 we analyze the Markovian input codebooks in the simple decoding framework. This improved bound is valid only for i.i.d. deletion patterns. However, a bound using this approach can also be worked out when the deletion patterns are Markovian.

4.1 Mutual information formulation for deletion channels

If the decoder does not have access to the packet loss pattern, it receives a subsequence of the transmitted sequence. There is a subtle assumption made in the model in that we assume that the decoder knows when to start decoding, *i.e.,* it does not expect more received symbols. In practice this can be done through a time-out mechanism. In [6], it was shown that if we employ optimal decoding, the capacity can be written in terms of the mutual information. It was proved that this is the case when there are rare synchronization symbols available which allow the transmitter and receiver to synchronize block boundaries. Given a transmission block of size n, the capacity [6] is written in terms of maximizing the mutual information.

Theorem 4.1 (Dobrushin, [6]) *For the deletion channel, a constant C is defined for which reliable transmission is possible if and only if $R < C$, where*

$$C = \lim_{n\to\infty} \frac{1}{n} C_n, \quad \text{where} \quad C_n = \sup_{p(x^n)} I(X^n; \mathbf{Y}), \tag{6}$$

where $\mathbf{Y}$ is the received sequence.

Note that the received sequence $\boldsymbol{Y}$ is of random length $M = |\boldsymbol{Y}|$.

Dobrushin proved this result for more general insertion/deletion channels. However, there has not been any single letter characterization of the mutual information.

The mutual information can be written as,

$$I(X^n; \boldsymbol{Y}) = H(X^n) - H(X^n|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|X^n) \tag{7}$$

For a given input codebook it is not difficult to calculate $H(X^n)$, but the main difficulty is in characterizing or bounding $H(X^n|\boldsymbol{Y})$. The other expressions give rise to similar difficulties. We can write $H(X^n|\boldsymbol{Y})$ as

$$H(X^n|\boldsymbol{Y}) = H(D^n|\boldsymbol{Y}) + H(X^n|D^n, \boldsymbol{Y}) - H(D^n|\mathbf{X}, \boldsymbol{Y}), \tag{8}$$

where $D^n = \{D_1, D_2, \ldots, D_n\}$ is the deletion pattern. In order to see the relationship between the erasure and the deletion channel notice that if the side information about the deletion pattern was known at the receiver, the mutual information of interest would be $I(X^n; \boldsymbol{Y}, D^n)$ and this can be written as

$$I(X^n; \boldsymbol{Y}, D^n) = I(X^n; \boldsymbol{Y}|D^n) = H(X^n) - H(X^n|D^n, \boldsymbol{Y}), \tag{9}$$

since the deletion process is independent of the input. Comparing (8) and (9) we see that the difference comes from $H(D^n|\boldsymbol{Y}) - H(D^n|X^n, \boldsymbol{Y}) = I(X^n; D^n|\boldsymbol{Y})$, that is, the dependence between the deletion process and the input, given the received sequence. This argument of course ignores the fact that the optimal input distribution for the i.i.d. erasure channel and the i.i.d. deletion channels are different. But it gives a flavor of where the difference might lie. Unfortunately we have not been able to give a single letter characterization of the mutual information of the deletion channel.

Optimal decoder: Achieving the rate defined in (6) requires optimal decoding. In the case of the deletion channel, if we know the block boundaries, we receive a sequence $\boldsymbol{Y}$ of length M where $n - M$ is the number of deletions. We need to find the most likely codeword that would have produced $\boldsymbol{Y}$. For the i.i.d. deletion channel this corresponds to finding the codeword $X^n(j) \in \mathcal{C}$ such that $\boldsymbol{Y}$ occurs most number of times as a subsequence of $X^n(j)$. This would mean finding for all codewords in $\mathcal{C}$ the number of times $\boldsymbol{Y}$ occurs as a subsequence. In the rest of the section we analyze a simple decoder that finds the codeword that contains $\boldsymbol{Y}$ as a subsequence, and if it is not unique, declares a decoding error. The advantages of this suboptimal decoder are its simplicity and the fact that we are able to characterize its achievable rate. However, we expect that more sophisticated decoders can achieve higher transmission rates.

4.2 Simple decoding framework

For the rest of the section we analyze a simple decoder defined as follows. We check the number of codewords in the codebook that could have produced the received sequence under *any* deletion pattern, *i.e.,* the codewords that contain the received sequence as a subsequence. If there is more than one possible codeword, then we declare a *collision error*. If there is no collision, we are certain that the unique candidate codeword is the correct one (as the channel is not noisy), and the transmission is successful.

In this section, we assume a stationary and ergodic model for the deletion process D. Therefore, for large n, the fraction of deleted packets is close to $(1-\theta) \stackrel{def}{=} p_d$ with high probability. Note that, as mentioned in Section 1, if the deletion patterns were known (through sequence numbers for example) then the channel would be equivalent to an erasure channel, whose capacity is $\theta \log(K)$. Clearly, conveying the deletion pattern to the receiver constitutes side-information and therefore this rate is an upper bound to the deletion channel capacity.

To study this problem we use the following lemma proved in [20] for common subsequences of random sequences.

Lemma 4.1 *[20] For a given K-ary sequence y of length $|y|$, the number $F(n, y, K)$ of K-ary sequences of length n which contain sequence y as a subsequence is given by:*

$$F(n, |y|, K) = \sum_{j=|y|}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) (K-1)^{n-j} \tag{10}$$

Note that the function $F(\cdot)$ depends on y *only* through its length $|y|$. Lemma 4.1 implies that if X^n is an i.i.d. sequence with uniform distribution over its K-ary alphabet then,

$$\mathbb{P}\{\, \boldsymbol{Y} \text{ subsequence of } X\} = \frac{F(n, |\, \boldsymbol{Y}|, K)}{K^n} \tag{11}$$

as all sequences are equally likely. Given this result, we prove the following lower bound on the capacity of the deletion channel.

Theorem 4.2 *Given a stationary and ergodic deletion channel with stationary deletion probability $p_d = 1-\theta$ (with $p_d < 1-1/K$), and an input alphabet size K, the capacity of this channel is lower bounded as*

$$C_{del} \geq \log\left(\frac{K}{K-1}\right) + \theta\log(K-1) - H_0(\theta). \tag{12}$$

Proof: Generate a random codebook of 2^{nR} i.i.d. codewords chosen uniformly from a K-ary alphabet. As the channel randomly deletes symbols from a codeword, the length of the received sequence $M = |\boldsymbol{Y}|$ is a random variable. Assuming that the deletion process is stationary and ergodic, it follows that

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{M}{n} - \theta\right| > \epsilon\right\} \to 0. \tag{13}$$

We use the following decoding rule. If the received sequence $\boldsymbol{Y}$ has fewer than $m = (\theta-\epsilon)n$ symbols, we declare an error. Because of (13), the probability of this error goes to zero. If the received sequence $\boldsymbol{Y}$ has at least m symbols, we check the number of codewords in the codebook that could have produced $\boldsymbol{Y}$ under any deletion pattern, i.e., the codewords that contain $\boldsymbol{Y}$ as a subsequence. If there is more than one possible codeword, then we declare a *collision error*. If there is no collision, we are certain that the unique candidate codeword is the correct one (as the subsequent channel after deletions is not noisy), and the transmission is successful.

We now compute the asymptotic probability of collision error. We first consider the *pairwise* probability of collision error between two codewords X_1 and X_2 from the codebook, averaged over all the codebooks. A collision error occurs between two codewords[5] X_1 and X_2 if the received sequence $\boldsymbol{Y} = D_1^n \circ X_1$ generated by a random deletion pattern D_1^n is a subsequence of X_2 (because this implies that there exists a deletion pattern D_2^n such that $D_2^n \circ X_2 = \boldsymbol{Y}$.)

Consider the error probability conditional on the number of received symbols M. This probability obviously decreases with M, because the probability of a common subsequence decreases with its length. Therefore, an upper bound for the collision probability can be obtained by setting $m = \lfloor(\theta-\epsilon)n\rfloor$ and assuming $M = m$. For computational reasons, we set $m = (\theta-\epsilon)n - 1$, which is conservative. Therefore, we can write the pairwise error probability that X_2 collides with the

[5]By slight abuse of notation we use X_1, X_2 for the codeword sequences of length n.

transmitted codeword X_1 averaged over random codebooks as,

$$\begin{aligned}
\bar{P}_2 &= \mathbb{E}_{\mathcal{C}}[\mathbb{P}\{\text{error}|X_1, M\}] = \sum_{\boldsymbol{Y},|\boldsymbol{Y}|=m} \mathbb{P}\{y \text{ is a subsequence of } X_2|X_1\}\,\mathbb{P}\{\boldsymbol{Y} = D^n \circ X_1\} \quad (14)\\
&\overset{(a)}{=} \frac{F(n,m,K)}{K^n} \sum_{\boldsymbol{Y},|\boldsymbol{Y}|=m} \mathbb{P}\{\boldsymbol{Y} = D^n \circ X_1\} \overset{(b)}{=} \frac{F(n,m,K)}{K^n}\\
&\overset{(c)}{\leq} \frac{1}{K^n} n \left(\begin{array}{c} n \\ m \end{array} \right) (K-1)^{n-m} \overset{(d)}{\leq} \frac{1}{K^n} n 2^{nH_0(\frac{m}{n})} (K-1)^{n-m}
\end{aligned}$$

where (a) follows from (11) and the independence of X_1 and X_2; (b) follows because the probability summed over all deletion patterns (conditioned on the weight of the deletion pattern) is unity; (c) follows from the inequality that $F(n,m,K) \leq n \left(\begin{array}{c} n \\ m \end{array} \right) (K-1)^{n-m}$ (which can be easily verified [20]); and (d) follows from the inequality $\left(\begin{array}{c} n \\ m \end{array} \right) \leq 2^{nH_0(\frac{m}{n})}$ (see [21], Chapter 12, pp 284).

A union bound over all codewords X_2 bounds the error probability $\bar{P}_e$ (averaged over codebooks) as,

$$\bar{P}_e \leq 2^{nR} \mathbb{E}_{\mathcal{C}}[\mathbb{P}\{\text{error}|X_1, M > m\}] + \mathbb{P}\{M \leq m\} \leq n \left[\frac{(K-1)2^R 2^{H_0(\frac{m}{n})}}{K(K-1)^{\frac{m}{n}}} \right]^n + \delta_n. \quad (15)$$

Therefore, if the first term decreases exponentially with n, the probability of error goes to zero asymptotically in n as $\delta_n \to 0$ from (13). This happens when

$$R < \log\left(\frac{K}{K-1}\right) + \theta \log(K-1) - H_0(\theta) \quad (16)$$

Therefore, by using the well-known random coding argument [21], there exists a deterministic codebook which has an achievable rate given by R. Note that this is in the regime where θ is such that $\log(\frac{K}{K-1}) + \theta \log(K-1) - H_0(\theta) > 0$, which occurs for $p_d < 1 - 1/K$. ■

Therefore, as the capacity of the deletion channel is upper bounded by that of the erasure channel, we obtain the following double inequality,

$$\log\left(\frac{K}{K-1}\right) + \theta \log(K-1) - H_0(\theta) \leq C_{del} \leq \theta \log(K) \quad (17)$$

For the binary case ($K = 2$), Theorem 4.2 coincides with the achievable rate proved in [4, 7]. However, the decoding technique is not the sequential decoding rule used in [4, 7], but is a common subsequence (mismatched) detection rule.

Corollary 4.1 *Given a stationary and ergodic deletion channel with long term deletion probability given by $1-\theta$ (with $\theta > 1/2$), and a binary input alphabet, the capacity of this channel is lower bounded as*

$$C_{del} \geq 1 - H_0(\theta), \tag{18}$$

where $H_0(.)$ is the binary entropy function.

4.3 Markov lower bound

In Section 4.2, the codewords were i.i.d. However, we believe that the optimal codebook construction has memory, because the deletion channel has memory. In this section, we sharpen the result in Section 4.2 by using input codebooks which are generated from a Markov process. For simplicity, we consider only first-order Markov chains. We continue with the simple decoding framework introduced in Section 4.2.

In analogy with the analysis for i.i.d. codewords in the previous section, we need to find the probability that two independent Markov chains of length n have a common subsequence of length greater than m. In order to simplify the analysis, we assume that the deletion sequences are i.i.d. processes[6].

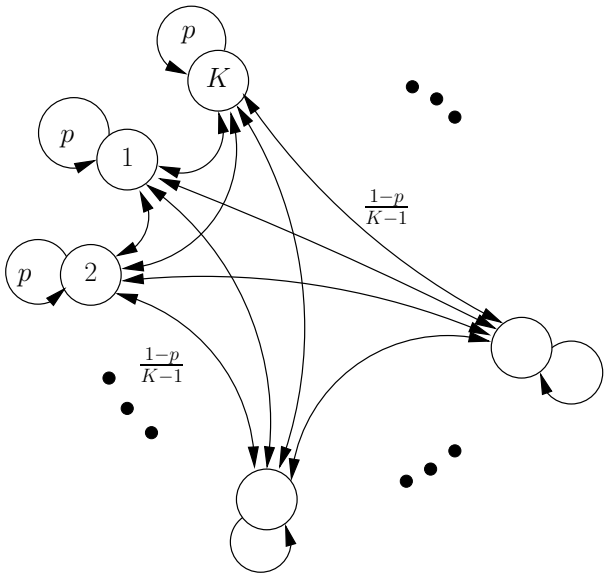

Figure 2: Markov chain generating codebook.

[6]The technique developed here can be easily extended to the case where the deletion sequences are finite-state Markov processes as well.

Theorem 4.3 *Given an i.i.d. deletion channel with deletion probability given by $p_d = 1-\theta$, and a K-ary input alphabet, the capacity of this channel is lower bounded as*

$$C_{del} \geq \sup_{\substack{\gamma>0 \\ 0<p<1}} \left[-\theta \log\{(1-q)A + qB\} - \gamma\right] \text{ nats} \tag{19}$$

where $A = A(p,\gamma) = e^{-\gamma}\frac{(1-p)}{(K-1)\left[1-e^{-\gamma}\left\{1-\frac{1-p}{K-1}\right\}\right]}$, $B = B(p,\gamma) = e^{-2\gamma}\frac{(1-p)^2}{(K-1)\left[1-e^{-\gamma}\left\{1-\frac{1-p}{K-1}\right\}\right]} + pe^{-\gamma}$ *and* $q = q(p) = \frac{1}{K}\left[1 + \frac{\theta(K-1)(pK-1)}{(K-1)-(1-\theta)(pK-1)}\right]$.

Proof: We generate a random codebook $\mathcal{C}$ of e^{nR} codewords, with each codeword an independent sequence generated using the first-order Markov chain illustrated in Figure 2. We use the same decoding framework as the one used in the proof of Theorem 4.2, *i.e.*, we declare an error either when the received sequence $\boldsymbol{Y}$ has fewer than $m = (\theta-\epsilon)n$ symbols, or when there is more than one codeword in $\mathcal{C}$ that contains $\boldsymbol{Y}$ as a subsequence. In order to compute the asymptotic collision error probability, we again consider the *pairwise probability of error* $\bar{P}_2$ between two codewords X_1 and X_2 averaged over random codebooks. As in (14) this can be written as,

$$\bar{P}_2 = \mathbb{E}_{\mathcal{C}}[\mathbb{P}\{\text{error}|X_1, M\}] = \sum_{\boldsymbol{Y}, |\boldsymbol{Y}|=m} \mathbb{P}\{\boldsymbol{Y} \text{ is a subsequence of } X_2|X_1\}\,\mathbb{P}\{\boldsymbol{Y} = D^n \circ X_1\} \tag{20}$$

The central difference between the calculation done in (14) and here is that when the input codebook has memory, the term $\mathbb{P}\{y \text{ is a subsequence of } X_2|X_1\}$ depends explicitly on the subsequence $\boldsymbol{Y}$ not just through its length M as was the case in (14). Moreover, this implies that we need to also explicitly calculate the term $\mathbb{P}\{\boldsymbol{Y} = D^n \circ X_1\}$ when the deletion process is i.i.d. and X_1 is a Markov process. For both these calculations, the weight of the derivative[7] $\Delta\boldsymbol{Y}$ of $\boldsymbol{Y}$ becomes important. Note that the number N_i of K-ary sequences y of length $|y| = m$ with Δy having weight i is given by

$$N_i = K(K-1)^i \begin{pmatrix} m-1 \\ i \end{pmatrix}. \tag{21}$$

Therefore, to calculate the pairwise error probability expression in (20) we use the following results, which are proved in the Appendix. Once having calculated the pairwise error probability we can prove the claim in (19).

Result 1: The probability of a given subsequence y occurring through an i.i.d. deletion process D in X, averaged over the input codebook, is given by

$$\mathbb{P}\{\boldsymbol{Y} = D^n \circ X\} = \frac{1}{K}\left(\frac{1-q}{K-1}\right)^i q^{m-1-i}, \tag{22}$$

[7] $\Delta\boldsymbol{Y}$ is defined as a sequence of length $|\boldsymbol{Y}|-1$, whose l^{th} component is 1 if $Y_{l+1} \neq Y_l$, and 0 otherwise. See also the table of definitions at the end of Section 2.

where i is the weight of $\Delta \boldsymbol{Y}$, $|\boldsymbol{Y}| = m$, and $q = \frac{1}{K}\left[1 + \frac{\theta(K-1)(pK-1)}{(K-1)-(1-\theta)(pK-1)}\right]$.

Result 2: The probability of a given subsequence y occurring in a Markov sequence X_2 generated by the transition probability illustrated in Figure 2 is bounded by,

$$\mathbb{P}\{\boldsymbol{Y} \text{ is a subsequence of } X_2\} \leq \inf_{\gamma>0} F e^{\gamma n} A^i B^{m-1-i}, \tag{23}$$

where

$$A = e^{-\gamma}\frac{(1-p)}{(K-1)\left[1 - e^{-\gamma}\left\{1 - \frac{1-p}{K-1}\right\}\right]}, \tag{24}$$

$$B = e^{-2\gamma}\frac{(1-p)^2}{(K-1)\left[1 - e^{-\gamma}\left\{1 - \frac{1-p}{K-1}\right\}\right]} + pe^{-\gamma}, \tag{25}$$

and

$$F = \frac{1}{K}e^{-\gamma}\left[1 + \frac{(1-p)e^{-\gamma}}{1 - e^{-\gamma}\left\{1 - \frac{1-p}{K-1}\right\}}\right]. \tag{26}$$

Again, in (23) i denotes the weight of $\Delta \boldsymbol{Y}$, and $|\boldsymbol{Y}| = m$.

Using (21), (22) and (23) in (20) we obtain,

$$\begin{aligned}
\bar{P}_2 &\leq \sum_{i=0}^{m-1} N_i \inf_{\gamma>0}\left[F e^{\gamma n} A^i B^{m-1-i}\frac{1}{K}\left(\frac{1-q}{K-1}\right)^i q^{m-1-i}\right] \qquad (27)\\
&\leq \inf_{\gamma>0}\left[F e^{\gamma n}\sum_{i=0}^{m-1}\binom{m-1}{i}\{(1-q)A\}^i\{Bq\}^{m-1-i}\right]\\
&\overset{(a)}{\leq} F \inf_{\gamma>0}\left[e^{\gamma n}\{(1-q)A + qB\}^{m-1}\right],
\end{aligned}$$

where (a) follows by using the binomial expansion of $(a+b)^{m-1}$.

As in the proof of Theorem 4.2, we use a union bound over all codewords X_2, if X_1 was the transmitted codeword to obtain,

$$\begin{aligned}
\bar{P}_e &\leq e^{nR}\mathbb{E}_{\mathcal{C}}[\mathbb{P}\{\text{error}|x_1, M > m\}] + \mathbb{P}\{M \leq m\} \qquad (28)\\
&\leq e^{nR} F \inf_{\gamma>0}\left[e^{\gamma n}\{(1-q)A + qB\}^{m-1}\right] \qquad (29)\\
&\leq F \inf_{\gamma>0}\left[\frac{1}{(1-q)A + qB}\left\{e^R e^{\gamma}[(1-q)A + qB]^{\frac{m}{n}}\right\}^n\right] + \delta_n.
\end{aligned}$$

Therefore the probability of error goes to zero asymptotically in n if,

$$R < \sup_{\substack{\gamma>0\\ 0<p<1}} \left[-\theta \log\{(1-q)A + qB\} - \gamma\right] \text{ nats} \tag{30}$$

giving us the desired result.

■

Remarks:

- Note that for $p = \frac{1}{K}$, it can be shown that the result in Theorem 4.3 reduces to that in Theorem 4.2 specialized to the i.i.d. deletion channel.
- Also note that the optimization of (19) for a given p can be accomplished in closed form as it results in a simple quadratic equation in γ.

We can specialize the result in Theorem 4.3 to the binary $K = 2$ case, as follows.

Corollary 4.2 *Given an i.i.d. deletion channel with deletion probability given by $p_d = 1 - \theta$, and a binary input alphabet, the capacity of this channel is lower bounded as*

$$C_{del} \geq \sup_{\substack{\gamma>0 \\ 0<p<1}} \left[-\theta \log\{(1-q)A + qB\} - \gamma\right] \text{ nats} \tag{31}$$

where $A = A(p,\gamma) = \frac{(1-p)e^{-\gamma}}{(1-pe^{-\gamma})}$, $B = B(p,\gamma) = \frac{(1-p)^2 e^{-2\gamma}}{(1-pe^{-\gamma})} + pe^{-\gamma}$ and $q = q(p) = 1 - \frac{1-p}{1+(1-\theta)(1-2p)}$.

It is interesting to note that with a Markovian codebook, non-zero capacity is achievable even for $p_d > 1 - 1/K$, the cutoff deletion probability for i.i.d. codebooks in our decoding framework. This is further examined in the next subsection, where we give some numerical results. This leads us to conjecture that a promising approach to achieve higher rates over the deletion channel is through codebooks generated by higher-order Markov processes, as it is unclear whether i.i.d. geometrically distributed runlengths are optimal.

4.4 Numerical results

In order to gain insight into the behavior of deletion channels, we give some numerical examples. In Figure 3, we plot the achievable rates derived in Sections 4.2 and 4.3. We also plot the "upper" bound derived by Ullman [5] given in (2). Note that this bound is only given for reference, as the underlying assumptions differ from ours: it relates to the combinatorial zero-error rate, while we are interested in vanishing error probability. Therefore, the achievable rate for the deletion channel is *not* necessarily upper bounded by the Ullman *zero-error upper bound*. Note that the lower bound derived in 4.3 (labeled as the Markov bound) is non-zero up to $p_d = 0.96$ and in particular improves the previously known lower bound given in (1) [4, 7]. In Figure 4, we illustrate that for non-binary alphabet sizes, the difference between the deletion channel and the erasure channel rates can be quite small. In particular, we have compared the lower bound (12) derived in Theorem 4.2 to the erasure channel rate. As can be seen from (12), the difference is at most $H_0(p_d) - p_d \log \frac{K}{K-1}$ bits, which is at most 1 bit. This is true with i.i.d. codebooks, and the bound becomes sharper with Markovian codebooks. This suggests that the use of sequence numbers to detect deletions is quite inefficient[8]. For example, in TCP, 32 bits per packet are sacrificed for the sequence number, while our result shows that at most one bit of redundancy per packet is necessary to convert a deletion channel into an erasure channel.

[8]Strictly speaking, sequence numbers are not a feasible coding scheme, as they require $\log n$ bits per packet.

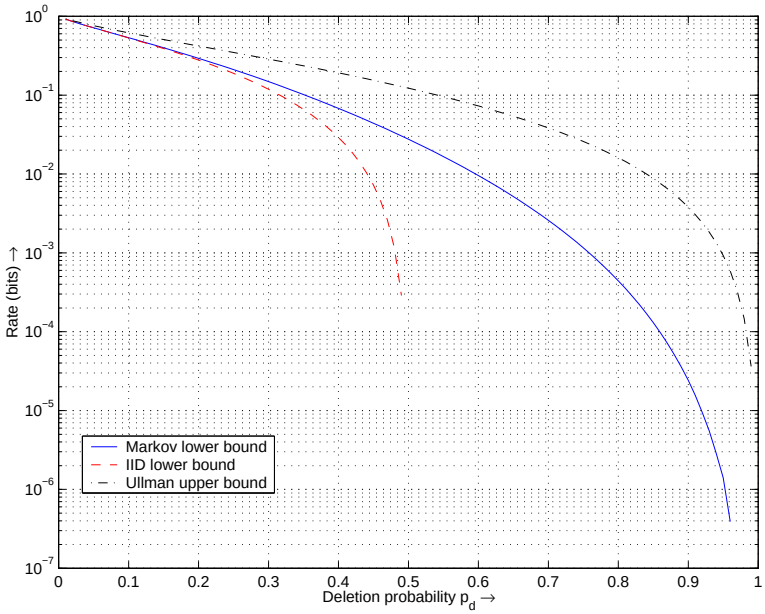

Figure 3: Achievable rates versus deletion probability for binary alphabet size.

5 Noisy deletion channel

In this section we briefly examine the effect of cascading a deletion channel with a discrete memoryless channel (DMC) as shown in Figure 5. We will mainly develop the lower bounds for i.i.d. input codebooks as done in Section 4.2. Also for simplicity we consider symmetric DMCs for which it is known that the optimal input distribution is uniform over the input alphabet (see [21], pp 190). We first start with a description of a decoder suitable for the noisy deletion channel. Using this decoder we provide a single letter achievable rate bound for the noisy deletion channel.

Consider a received sequence $\boldsymbol{Z}$ of random length M which is the output of the cascade of the deletion channel and the DMC. Informally, the decoder looks at the set $A_\epsilon(\tilde{Y}^M|\boldsymbol{Z})$ sequences of length m which are "typical" with respect to the *particular* noisy received $\boldsymbol{Z}$, where typicality is defined with respect to the discrete memoryless channel. Now, the decoder constructs a list $\mathcal{L}$ of all codewords for which *any* sequence in $A_\epsilon(\tilde{Y}^M|\boldsymbol{Z})$ is a subsequence of the codeword. Unless $|\mathcal{L}| = 1$, the decoder declares an error. If $|\mathcal{L}| = 1$, then the decoder declares that j is the index of the transmitted codeword, where $j \in \mathcal{L}$. Therefore, an error occurs either if $|\mathcal{L}| \neq 1$ or if $j \in \mathcal{L}$ is not the transmitted message. We bound the rate for which this error probability diminishes to zero asyptotically in n and this provides an achievable rate for the noisy deletion channel.

Consider a discrete memoryless channel (DMC) defined by the transition probability $p(z|y)$, where we take Y as the input and Z as the output of the DMC. As mentioned earlier, we restrict our attention to symmetric DMCs. The m^{th} extension of the DMC is given by the product distribution, $p(z^m|y^M) = \prod_{i=1}^{M} p(z_i|y_i)$. We define the set $A_\epsilon(\tilde{Y}^M|\boldsymbol{Z})$ to be the set of input sequences $\tilde{y}^M$, that

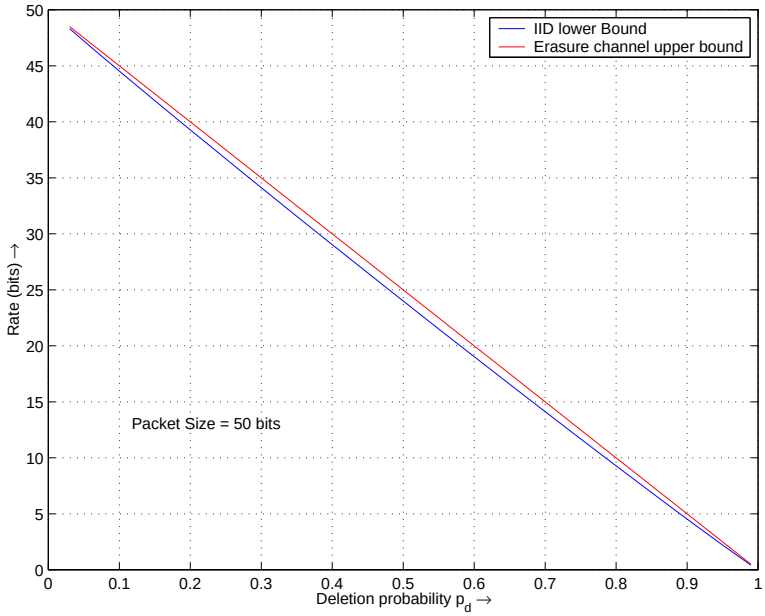

Figure 4: Erasure and deletion channels with packet size of $K = 50$ bits.

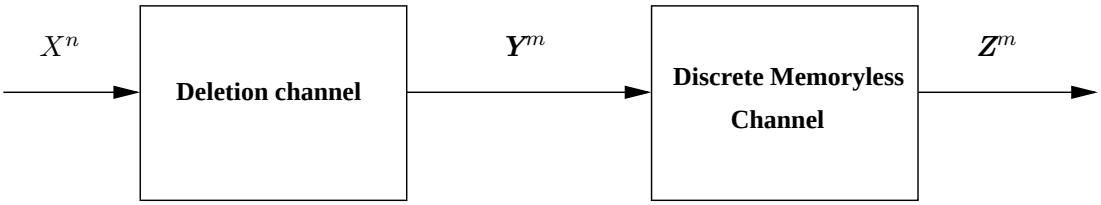

Figure 5: A deletion channel followed by a discrete memoryless channel.

are jointly $\epsilon - typical$ with a particular output sequence $\boldsymbol{Z}$ [21]. Therefore, we immediately have the following result (see [21], Theorem 14.2.2) that for sufficiently large m,

$$|A_\epsilon(\tilde{Y}^M|\boldsymbol{Z})| \leq 2^{m(H(Y|Z)+2\epsilon)}. \tag{32}$$

Theorem 5.1 *Consider a stationary and ergodic deletion channel with stationary deletion probability $p_d = 1 - \theta$ (with $p_d < 1 - 1/K$), and an input alphabet size K. If the deletion channel is cascaded with a symmetric DMC (as shown in Figure 5) then capacity of this channel is lower bounded as*

$$C_{del} \geq \max\left\{\log\left(\frac{K}{K-1}\right) + \theta\log(K-1) - [H_0(\theta) + \theta H(Y|Z)]\right\}. \tag{33}$$

Proof: Generate a random codebook $\{x^n(1), \ldots, x^n(2^{nR})\}$ of 2^{nR} n-length i.i.d. codewords chosen uniformly from a K-ary alphabet. As in the proof of Theorem 4.2, if the received sequence $\boldsymbol{Z}$ has fewer than $m = (\theta - \epsilon)n$ symbols, we declare an error. As before due to (13), the probability of

this error goes to zero asymptotically in n. Without loss of generality, let us assume that $i = 1$ was the transmitted message, *i.e.,* $x^n(1)$ was the transmitted sequence. If we denote the output of the deletion channel as $\boldsymbol{Y}$, then clearly it is a subsequence of $x^n(1)$. However note that the noisy sequence $\boldsymbol{Z}$ observed after $\boldsymbol{Y}$ may not be a subsequence of $x^n(1)$.

The decoder generates a list $\mathcal{L}$ of all messages for which there exists a sequence in $A_\epsilon(\tilde{Y}^M|\boldsymbol{Z})$ which is a subsequence of the codeword. That is,

$$\mathcal{L} = \left\{ j : \exists \tilde{y}^M \in A_\epsilon(\tilde{Y}^M|\boldsymbol{Z}) \text{ such that } \tilde{y}^M \text{ is a subsequence of } x^n(j) \right\} \tag{34}$$

Let us define the event

$$E_j = \{ j \in \mathcal{L} \}, \;\; j = 1, \ldots, 2^{nR}. \tag{35}$$

Clearly, we have from the union bound that

$$\mathbb{P}\left\{ \text{error} | x^n(1) \right\} \leq \mathbb{P}\left\{ E_1^c \right\} + \sum_{j \neq 1} \mathbb{P}\left\{ E_j \right\}. \tag{36}$$

Since $\boldsymbol{Y}$ the output of the deletion channel is a subsequence of $x^n(1)$, the event E_1^c will occur only if $\boldsymbol{Y}$ is not jointly typical with the DMC output $\boldsymbol{Z}$. Therefore by the definition of the typical set, we see that $\mathbb{P}\left\{ E_1^c \right\} \leq \epsilon$. Now for the event $E_j, j \neq 1$, it is easy to see that

$$\mathbb{P}\left\{ E_j \right\} \leq \sum_{\tilde{y}^M \in A_\epsilon(\tilde{Y}^M|\boldsymbol{Z})} \mathbb{P}\left\{ \tilde{y}^M \text{ is a subsequence of } x^n(j) \right\} \overset{(a)}{\leq} |A_\epsilon(\tilde{Y}^M|\boldsymbol{Z})| \frac{F(n,m,K)}{K^n} \tag{37}$$

In the above (a) is because we are considering uniform i.i.d. codebooks, and hence (a) is a consequence of (11). Using (37) and (32), and following steps similar to (15) we see that

$$\mathbb{P}\left\{ \text{error} | x^n(1) \right\} \leq \delta_n + \epsilon_n + \left[\frac{(K-1) 2^R 2^{H_0(\frac{m}{n})} 2^{\frac{m}{n}[H(Y|Z)+2\epsilon_n]}}{K(K-1)^{\frac{m}{n}}} \right]^n \tag{38}$$

Hence, as $\delta_n, \epsilon_n \to 0$, we see that the probability of error goes to zero asymptotically in n if,

$$R < \log\left(\frac{K}{K-1} \right) + \theta \log(K-1) - \left\{ H_0(\theta) + \theta H(Y|Z) \right\}. \tag{39}$$

This gives the desired achievable rate using standard random coding arguments. ■

Corollary 5.1 *Consider a stationary and ergodic deletion channel with long term deletion probability given by $1-\theta$ (with $\theta > 1/2$), and a binary input alphabet. If we cascade this with a binary symmetric channel with cross-over probability p_e, then the capacity of this channel is lower bounded as*

$$C_{del} \geq \max\left\{ 0, 1 - \left[H_0(\theta) + \theta H_0(p_e) \right] \right\}, \tag{40}$$

where $H_0(.)$ is the binary entropy function.

6 Discussion

In this paper, we have studied information transmission over finite buffer channels. We considered two cases. First, if the packet loss pattern is known at the receiver, the channel is equivalent to an erasure channel. For this case, we focus on the impact of non-Markovian channel memory and on feedback. Second, if the receiver does not have this side information, it is equivalent to a deletion channel. For this case, we need to make stronger assumptions about the deletion process. We developed bounds for the achievable rate of deletion channels when we use a simple (but mismatched) decoder.

In Section 3, we studied the case where the packet loss is known at the receiver, *i.e.,* the deletion pattern is known. This led to a model of the channel as a finite-state channel (with memory) that sends a special erasure symbol when the buffer is full. For this model we showed that even when the state-process has complicated memory an i.i.d. channel input process achieves capacity, which is not true for general finite-state channels. Moreover, for this model we showed that feedback does not change the capacity as long as the sender does not change the arrival rate of the packets.

In Section 4, we studied the case where the packet loss is not known at the receiver. Under this assumption, the finite buffer channel is a deletion channel, whose study is more complicated than the erasure channel, and whose precise capacity remains elusive. We were able to significantly improve upon previously known capacity bounds for the deletion channel. This was done using Markovian input codebooks and a stronger assumption about the deletion process than for the erasure channel. Specifically, we assumed an i.i.d. deletion process for those results, although our techniques could be extended to finite-state Markovian deletion processes in a straightforward manner. Also, we showed that a simple subsequence matching decoder can perform quite well for larger alphabet sizes K, although its performance remains far below the erasure channel in the binary case even with Markovian codebooks when the deletion probability approaches 1. We also examined the noisy deletion channel where the deletion channel is cascaded with a symmetric DMC. Here we gave a single letter expression for an achievable rate that naturally extended from the bound for deletion channels with i.i.d. inputs.

Several interesting questions remain open about deletion channels. Of course, the central question is the single-letter characterization of its capacity. Even in the absence of such a characterization, it would be important to develop tighter upper and lower bounds for achievable rates. In particular, good upper bounds for small alphabet sizes will be useful in gaining insight into the behavior of deletion channels.

The problem of code construction has a long history and still has a vast number of unresolved problems (see [8] for example). A key insight from our work is that for small deletion probabilities, random i.i.d. codebooks perform well, while for higher deletion probabilities one needs to introduce memory into the codebook design. We believe that this stems from the need of having "runs" of identical symbols in codewords, such that most of these runs survive the deletion process. This

leads us to conjecture that coding in run-lengths is a useful design technique for deletion channel codes.

Acknowledgments

We would like to thank Yiannis Kontoyiannis, Michael Mitzenmacher, Alon Orlitsky, Neil Sloane, Emre Telatar, and Vinay Vaishampayan for stimulating discussions on the topic of this paper. We also thank the reviewers for detailed and constructive inputs. Their comments also encouraged us to obtain additional results on the noisy deletion channel.

A Proof of Theorem 3.1

Proof: We first prove the result for the case without feedback. Since under the regularity conditions for $\{Q_i\}$, mutual information has an operational interpretation, the capacity C is given by

$$C = \lim_{n\to\infty} \frac{1}{n} \max_{p(X^n)} I(X^n; Y^n).$$

Thus we can write,

$$C_n = I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) = \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X_i, X^{i-1}) \qquad (41)$$

The channel output Y_i in turn is a function of the current channel state Q_i and the current input symbol X_i. We can identify the following Markov chain.

$$X^{i-1} \leftrightarrow (X_i, Q_i) \leftrightarrow Y_i$$

Furthermore, conditional on Q^{i-1} or Y^{i-1} (or both), Q_i is independent of X^{i-1}, X_i, i.e.,

$$(X^{i-1}, X_i) \leftrightarrow Y^{i-1} \leftrightarrow Q_i \qquad (42)$$

This is because the channel state is independent of the channel input.

We can now compute the two entropies in (41). Recall that for a dropped packet ($Q_i = 1$), the decoder receives the special erasure symbol $Y_i = E$, which embodies the assumption that the receiver has side information about dropped packets.

$$
\begin{aligned}
H(Y_i|Y^{i-1}) &= \mathbb{E}\left[-\log\left\{p\left(Y_i|Q_i=0,Y^{i-1}\right)p\left(Q_i=0|Y^{i-1}\right)+p\left(Y_i|Q_i=1,Y^{i-1}\right)p\left(Q_i=1|Y^{i-1}\right)\right\}\right] \\
&= \mathbb{E}\left[-\log\left\{\mathbf{1}_{\{Y_i\neq E\}}p\left(Y_i|Q_i=0,Y^{i-1}\right)p\left(Q_i=0|Y^{i-1}\right)+\mathbf{1}_{\{Y_i=E\}}p\left(Q_i=1|Y^{i-1}\right)\right\}\right] \\
&= \mathbb{E}\left[-\mathbf{1}_{\{Y_i\neq E\}}\log\left\{p\left(Y_i|Q_i=0,Y^{i-1}\right)p\left(Q_i=0|Y^{i-1}\right)\right\}\right]+ \\
&\quad \mathbb{E}\left[-\mathbf{1}_{\{Y_i=E\}}\log\left\{p\left(Q_i=1|Y^{i-1}\right)\right\}\right]
\end{aligned} \tag{43}
$$

$$
\begin{aligned}
H(Y_i|Y^{i-1},X_i,X^{i-1}) &= \mathbb{E}\left[-\log\left\{p\left(Y_i|Q_i=0,X_i\right)p\left(Q_i=0|Y^{i-1},X_i,X^{i-1}\right)\right.\right. \\
&\quad \left.\left.+p\left(Y_i|Q_i=1,Y^{i-1},X_i,X^{i-1}\right)p\left(Q_i=1|Y^{i-1},X_i,X^{i-1}\right)\right\}\right] \\
&= \mathbb{E}\left[-\log\left\{\mathbf{1}_{\{Y_i\neq E\}}p(Y_i|Q_i=0,X_i)p\left(Q_i=0|Y^{i-1}\right)+\mathbf{1}_{\{Y_i=E\}}p\left(Q_i=1|Y^{i-1}\right)\right\}\right] \\
&= \mathbb{E}\left[-\mathbf{1}_{\{Y_i\neq E\}}\log\left\{p(Y_i|Q_i=0,X_i)p\left(Q_i=0|Y^{i-1}\right)\right\}\right]+ \\
&\quad \mathbb{E}\left[-\mathbf{1}_{\{Y_i=E\}}\log\left\{p\left(Q_i=1|Y^{i-1}\right)\right\}\right]
\end{aligned} \tag{44}
$$

Putting together (41), (43) and (44), we get

$$
C_n = \sum_{i=1}^{n}\mathbb{P}\left\{Q_i=0\right\}\left[H(Y_i|Q_i=0,Y^{i-1})-H(Y_i|Q_i=0,X_i)\right] \tag{45}
$$

Since for $Q_i=1$, we have $H(Y_i|Q_i=1,Y^{i-1})=H(Y_i|Q_i=1,X_i)=0$, using this in (45) we have therefore shown that

$$
C_n = I(X^n;Y^n) = \sum_{i=1}^{n}\left[H(Y_i|Q_i,Y^{i-1})-H(Y_i|Q_i,X_i)\right]. \tag{46}
$$

For a given marginal distribution $p(X_i)$, the first term of (46) is maximized for $\{X_i\}$ i.i.d., because conditioning reduces entropy. Therefore,

$$
\begin{aligned}
C &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\max_{p(X_i)} I(X_i;Y_i|Q_i) \\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\left\{Q_i=0\right\}\max_{p(X_i)} I(Y_i;X_i|Q_i=0) \\
&= C_0\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\left\{Q_i=0\right\} = \theta C_0
\end{aligned} \tag{47}
$$

where $C_0=\max_{p(X)} I(X;Y|Q=0)$ is the capacity of the DMC,

and the existence of $\theta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\{Q_i = 0\}$ is guaranteed because of the regularity conditions on $\{Q_i\}$. Therefore, the capacity is proportional to the fraction of time the channel is in the "good" state, i.e., when there is no buffer overflow.

We now show that feedback does not increase the capacity of the channel. We assume that the encoder has access up to the previous received symbol, i.e., it knows Y^{i-1}, and hence $X_i = f(W, Y^{i-1})$.

$$\begin{aligned}
I(W;Y^n) &\overset{(a)}{=} I(W;Y^n,Q^n) \qquad (48)\\
&\overset{(b)}{=} I(W;Y^n|Q^n)\\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1},Q^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1},W,Q^n)\\
&\overset{(c)}{=} \sum_{i=1}^{n} H(Y_i|Y^{i-1},Q^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1},W,Q^n,X_i)\\
&\overset{(d)}{=} \sum_{i=1}^{n} H(Y_i|Y^{i-1},Q^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1},Q^n,X_i)\\
&\overset{(e)}{\leq} \sum_{i=1}^{n} H(Y_i|Q_i) - \sum_{i=1}^{n} H(Y_i|Y^{i-1},Q^n,X_i)\\
&\overset{(f)}{=} \sum_{i=1}^{n} H(Y_i|Q_i) - \sum_{i=1}^{n} H(Y_i|Q_i,X_i)\\
&= \sum_{i=1}^{n} I(X_i;Y_i|Q_i), \qquad (49)
\end{aligned}$$

where (a) is because Q_n is a function of Y_n, (b) is because the message W is independent of Q^n (hence $I(W;Q_n) = 0$), (c) is because $X_i = f(W, Y^{i-1})$, (d) follows from Y_i conditional on X_i and Q^n is independent of W, inequality (e) is due to conditioning decreasing entropy, and (f) is because we have a DMC and therefore conditioning on Q_i, X_i makes Y_i independent of Y^{i-1}, Q^{i-1}.

The crucial thing to note in (49) is that $I(X_i;Y_i|Q_i)$ depends on i because X_i is a function of Y^{i-1}, and hence this could be a non-stationary process. However, the mutual information is a concave function in its input probability for a fixed channel. Hence by choosing an average distribution averaged over the Y^{i-1}, i.e.,

$$\bar{p}(X_i|W) = \sum_{y^{i-1}} p(X_i|y^{i-1},W)p(y^{i-1}|W), \qquad (50)$$

the mutual information would increase. Hence if we define the stationary process $\{\bar{X}_i\}$, where the X_i are i.i.d. with the above marginal distribution, and using the concavity of mutual information, we obtain

$$I(X_i(y^{i-1},W);Y_i|Q_i) \leq I(\bar{X}_i(W);Y_i|Q_i) \qquad (51)$$

Now using (51) in (48) we obtain

$$\frac{1}{n}I(W;Y^n) \leq \frac{1}{n}\sum_{i=1}^{n} I(\bar{X}_i(W);Y_i|Q_i), \tag{52}$$

and hence the feedback capacity C_{fb} is clearly lower bounded by the capacity C_{nfb} without feedback. However, from the above argument we also have

$$C_{nfb} \leq C_{fb} \leq \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\max_{p(\bar{X}_i)} I(\bar{X}_i;Y_i|Q_i) = C_{nfb}. \tag{53}$$

Therefore, $C_{fb} = C_{nfb}$. ■

B Proofs of (22) and (23)

In this section we provide the details of the proof for two results used in the proof of Theorem 4.3.

B.1 Proof of (22)

In order to prove (22), we need to look at the joint process of the i.i.d. binary deletion process D and the first order Markov chain X generating the random codebook. This can be done by extending the state space of the Markov process of Figure 2 to include the state of the deletion process. This yields a $2K$-state Markov process, (X, D), with each state being the concatenation of the state of the Markov process and the deletion process. The process $\boldsymbol{Y} = D \circ X$ is generated by observing only the states where the deletion process is 0, *i.e.*, the sequence generated through the deletion channel. These states constitute a subset of the extended Markov chain state space, and by the Strong Markov Property [22], this randomly sampled Markov chain is itself a Markov chain. The transition probability $\bar{\mathbf{P}}$ of the Markov chain generated by the "watched" set is given by

$$\bar{\mathbf{P}} = \theta\mathbf{P}\left[\mathbf{I} - (1-\theta)\mathbf{P}\right]^{-1}, \tag{54}$$

where $\mathbf{P}$ is the transition matrix of the Markov chain shown in Figure 2. It can be easily shown that the Markov chain corresponding to $\bar{\mathbf{P}}$ is also symmetric (just as the Markov chain in Figure 2) with parameter $q = \frac{1}{K}\left[1 + \frac{\theta(K-1)(pK-1)}{(K-1)-(1-\theta)(pK-1)}\right]$. Therefore, the subsequence $\boldsymbol{Y}$ is obtained by running this Markov chain for m steps. This yields

$$\mathbb{P}\{\,\boldsymbol{Y} \text{ with } \Delta\boldsymbol{Y} \text{ of weight } i\} = \frac{1}{K}\left(\frac{1-q}{K-1}\right)^{i} q^{m-1-i} \tag{55}$$

B.2 Proof of (23)

To obtain (23), we examine the probability of a sequence y of length m occurring in a K-state Markov chain X of length n, given in Figure 2. We will consider the number of transitions N of the Markov chain needed to sequentially match every symbol of y.

As X is symmetric, the stationary probabilities for both states are identical. Thus, we can assume $y_1 = 0$ w.l.g. We first compute N_1, the number of transitions to match y_1. Its distribution is given by

$$\mathbb{P}\left\{N_1 = i\right\} = \begin{cases} \frac{1}{K} & i = 1 \\ (1 - \frac{1}{K})\frac{1-p}{K-1}\left[1 - \frac{1-p}{K-1}\right]^{i-2} & i \geq 2 \end{cases}. \tag{56}$$

Now assume we have matched y_{l-1} in X, and let us consider how many transitions are needed to match y_l. We have to distinguish two cases: (a) $y_l \neq y_{l-1}$, and (b) $y_l = y_{l-1}$. Let N_a and N_b denote the number of transitions up to the next match for case (a) and (b), respectively. To this end, note that because of the symmetry of X, case (a) corresponds to the number of transitions to reach state 1 when we do *not* start from state 1 (cf. Fig. 2). For this we obtain

$$\mathbb{P}\left\{N_a = i\right\} = \frac{1-p}{K-1}\left[1 - \frac{1-p}{K-1}\right]^{i-1} \quad i \geq 1. \tag{57}$$

Similarly, case (b) corresponds to the number of transitions to reach state 1 starting from state 1. Therefore we obtain

$$\mathbb{P}\left\{N_b = i\right\} = \begin{cases} p & i = 1 \\ \frac{(1-p)^2}{K-1}\left[1 - \frac{1-p}{K-1}\right]^{i-2} & i \geq 2 \end{cases}. \tag{58}$$

We calculate the moment generating functions for N_a and N_b as

$$\Phi_{N_a}(\gamma) \stackrel{def}{=} \mathbb{E}e^{-\gamma N_a} = e^{-\gamma}\frac{(1-p)}{(K-1)\left[1 - e^{-\gamma}\left\{1 - \frac{1-p}{K-1}\right\}\right]}, \tag{59}$$

where it is inherently assumed that $|e^{-\gamma}\left\{1 - \frac{1-p}{K-1}\right\}| < 1$, which is always true because, $\left\{1 - \frac{1-p}{K-1}\right\} \leq 1 < e^{\gamma}$ for $\gamma > 0$. Similarly, we can calculate the following,

$$\Phi_{N_b}(\gamma) \stackrel{def}{=} \mathbb{E}e^{-\gamma N_b} = e^{-2\gamma}\frac{(1-p)^2}{(K-1)\left[1 - e^{-\gamma}\left\{1 - \frac{1-p}{K-1}\right\}\right]} + pe^{-\gamma}. \tag{60}$$

Let N_l denote the number of transitions to match y_l. Then

$$N_l = \begin{cases} N_1 & l = 1 \\ N_{a,l} & (\Delta y)_l = 1 \\ N_{b,l} & (\Delta y)_l = 0 \end{cases}, \tag{61}$$

where $N_{a,l}$ and $N_{b,l}$ are independent and distributed like N_a and N_b. The total number of iterations to match y is $N = \sum_{l=1}^{m} N_l$. Therefore we have

$$\mathbb{P}\{y \text{ with } \Delta y \text{ of weight } i \text{ occurs in } (X_1, \ldots, X_n)\} = \mathbb{P}\{N \leq n\}. \tag{62}$$

Using the Chernoff bound we can upper bound this probability as

$$\begin{aligned} \mathbb{P}\{N \leq n\} &\leq \inf_{\gamma>0} e^{\gamma n} \mathbb{E} e^{-\gamma \sum_{l=1}^{m} N_l} \\ &\overset{(a)}{=} \inf_{\gamma>0} e^{\gamma n} \mathbb{E}[e^{-\gamma N_1}] \{\mathbb{E} e^{-\gamma N_a}\}^i \{\mathbb{E} e^{-\gamma N_b}\}^{m-1-i}, \end{aligned} \tag{63}$$

where (a) is obtained by the definition of N_l in (61) and the fact that the N_l are independent of each other, giving us the desired result.

References

[1] V. Anantharam and S. Verdu, "Bits through Queues," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 4–18, 1996.

[2] D. Julian, "Erasure networks," in *Proc. IEEE International Symposium on Information Theory (ISIT), Lausanne, Switzerland*, p. 138, 2002.

[3] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics - Doklady*, vol. 10, pp. 707–710, February 1966.

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

[5] J. D. Ullman, "On the capabilities of codes to correct synchronization errors," *IEEE Transactions on Information Theory*, vol. 13, pp. 95–105, January 1967.

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

[7] K. S. Zigangirov, "Sequential decoding for a binary channel with drop-outs and insertions," *Problems Information Transmission*, vol. 5, no. 2, pp. 17–22, 1969. Translated from Problemy Peredachi Informatsii, vol. 5, no. 2, pp 23–30, 1969.

[8] N. J. A. Sloane, "On Single-Deletion-Correcting Codes." In D. Ray-Chaudhuri Festschrift, 2001. See: http://www.research.att.com/ njas/doc/dijen.ps.

[9] M. C. Davey and D. J. C. MacKay, "Reliable communication over channels with insertions, deletions and substitutions," *IEEE Transactions on Information Theory*, vol. 47, pp. 687–698, February 2001.

[10] S. Diggavi and M. Grossglauser, "On transmission over deletion channels," in *Proc. Allerton Conference, Monticello, Illinois*, October 2001.

[11] E. Drinea and M. Mitzenmacher, "On lower bounds for the capacity of deletion channels," in *Proc. IEEE International Symposium on Information Theory (ISIT), Chicago, Illinois*, p. 227, 2004.

[12] A. Kavcic and R. Motwani, "Insertion/deletion channels: Reduced-state lower bounds on channel capacities," in *Proc. IEEE International Symposium on Information Theory (ISIT), Chicago, Illinois*, p. 229, 2004.

[13] R. G. Gallager, *Information theory and reliable communications*. New York: John Wiley and Sons, 1968.

[14] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channels," *IEEE Transactions on Information Theory*, vol. 35, no. 6, pp. 1277–1290, 1989.

[15] J. Wolfowitz, *Coding Theorems of Information Theory*. Berlin: Springer-Verlag, 2nd ed., 1964.

[16] R. L. Adler, "Ergodic and mixing properties of infinite memory channels," *Proc. American Math. Society*, vol. 12, no. 6, pp. 924–930, 1961.

[17] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[18] S. Verdu and T. S. Han, "A General Formula for Channel Capacity," *IEEE Transactions on Information Theory*, vol. 40, pp. 1147–1157, July 1994.

[19] B. Prabhakar and R. G. Gallager, "Entropy and timing capacity of discrete queues," *IEEE Transactions on Information Theory*, vol. 49, pp. 357–370, February 2003.

[20] V. Chvatal and D. Sankoff, "Longest common subsequence of two random sequences," *Journal of Applied Probability*, no. 12, pp. 306–315, 1975.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, Inc., 1991.

[22] R. Durrett, *Probability: theory and examples*. Duxbery Press, 2nd ed., 1995.