On lower bounds for the capacity of deletion channels

Eleni Drinea[1], Michael Mitzenmacher[2]
Division of Engineering and Applied Sciences, Harvard University
e-mail: edrinea@deas.harvard.edu, michaelm@eecs.harvard.edu

Abstract — **We consider binary deletion channels, where bits are deleted independently with probability d. We extend the framework used to analyze the capacity of binary deletion channels established by Diggavi and Grossglauser [1], improving on their lower bounds. In [1], the codewords are generated by a first order Markov chain. Our improvements arise from two considerations. First, [1] considers typical outputs, where an output is typical if an N bit input produces an output at least $N(1-d)(1-\epsilon)$ bits. We provide a stronger notion of a typical output that yields better bounds even for the cases studied in [1]. Second, we consider codewords generated by more general processes than first order Markov chains.**

I. Codebook Generation

To lower bound the capacity of binary deletion channels, Diggavi and Grossglauser consider random codes, with each codeword of length N generated by a symmetric first order Markov process with transition probability p. The decoder successfully decodes if the received sequence is a subsequence of exactly one codeword; otherwise, it fails. The transmission rate can be determined by requiring that the decoder fail with probability that goes to 0 asymptotically. The capacity of a channel with deletion probability d is lower bounded by the maximum rate over all choices of the transition probability p.

We extend this framework by allowing the codewords of length N to consist of alternating blocks of zeros and ones generated independently by the following stochastic process. Let P be a distribution on the integers $j > 0$ with suitably decreasing tails. Whether the first block consists of zeros or ones is determined by a fair coin flip. The length of each successive block is independently chosen from the distribution P, having length j with probability P_j. The last block is truncated so that the total length is N (this does not affect the resulting asymptotics).

When a codeword generated as above is transmitted through the deletion channel, the received sequence also consists of blocks of zeros and ones. We show that the lengths of the blocks in this sequence are independently and identically distributed, according to a distribution $\mathcal{P}$ (more details appear in [3]).

II. Decoder and Lower bounds for capacity

The improved results of this paper arise from two considerations. First, in [1], *typical outputs* consist of at least $N(1-d)(1-\epsilon)$ bits, for $\epsilon = o(1)$, while any output that is atypical can be considered an error (since the probability of an atypical output goes to 0). We use a stronger notion of a typical output: the number the number of blocks of each length in the received sequence must be close to its expectation. See Thm 1 in [3] for an exact definition. This technique improves the bounds even for the codebooks in [1] (see Corollary 1 in [3] for the analytical results).

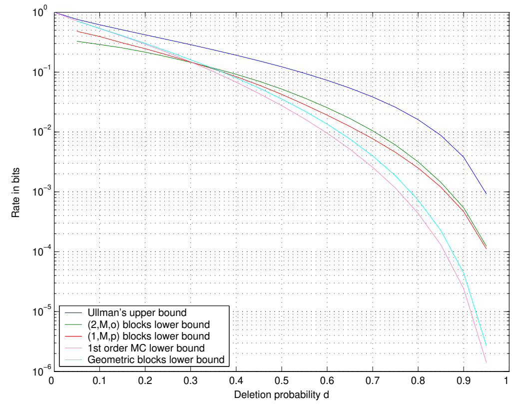

Figure 1: Improvements in rates with our framework, for geometric, $(1, M, p)$ and $(2, M, p)$ distributions.

Modifying the definition of a typical output improves the capacity bound rather mildly as shown in Figure 1. Our more important improvement comes by considering alternative block length distributions P. The first order Markov chains give block lengths that are geometrically distributed. There is no reason a priori why the geometric distribution is the right choice, either for reaching capacity or for proving lower bounds in this fashion. We suspected that for high deletion rates, greater variability in block lengths would lead to improved code rates. We considered the case where a block is assigned length 1 with probability p and integer length $M > 1$ with probability $1-p$. Denote this distribution by $(1, M, p)$. Intuitively, this coding mimics Morse code, with dots (short blocks) and dases (long blocks). Figure 1 shows the improvement such a distribution yields over the geometric distribution when $d \geq 0.4$. Distributions $(2, M, p)$, defined analogously to $(1, M, p)$, further improve the rates for the same range of d; however their rates for smaller d worsen. Finally, the rates in this figure are an underestimate of the maximum rates since the optimization of p is only over 2 decimal digits. We have found no simple closed form for the general lower bound for the capacity we provide in Thm 2 in [3]. Optimal choices for the distribution P under our framework are still not known.

References

[1] S. Diggavi and M. Grossglauser, **On Transmission over Deletion Channels**, Allerton Conference, Monticello, Illinois, October 2001.

[2] N. J. A. Sloane, **On Single-Deletion-Correcting Codes**, appeared in Ray-Chaudhuri Festschrift 2001, URL: http://www.research.att.com/~njas/docs/dijen.ps.

[3] E. Drinea, M. Mitzenmacher, **On lower bounds for the capacity of deletion channels**, Technical Report No-???, Harvard University, April 2004.

[1] Supported by NSF grant CCR-0118701.
[2] Supported by NSF grants CCR-9983832, CCR-0118701, CCR-0121154, and an Alfred P. Sloan Research Fellowship.