Combinatorics, Probability and Computing (2015) 24, 486–489. © Cambridge University Press 2014
doi:10.1017/S0963548314000522

A Binary Deletion Channel With a Fixed Number of Deletions

BENJAMIN GRAHAM

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK
(e-mail: b.graham@warwick.ac.uk)

Received 24 October 2013; revised 12 June 2014; first published online 7 October 2014

Suppose a binary string $\mathbf{x} = x_1 \cdots x_n$ is being broadcast repeatedly over a faulty communication channel. Each time, the channel delivers a fixed number m of the digits ($m < n$) with the lost digits chosen uniformly at random and the order of the surviving digits preserved. How large does m have to be to reconstruct the message?

2010 *Mathematics subject classification*: Primary 68R05
Secondary 68P30

1. Introduction

A *binary deletion channel* is a communication device that accepts a sequence of n binary digits. Each digit is lost in transmission with probability p. The order of the surviving digits is preserved, but the output does not indicate the original location of those digits. The number of digits in the output binary string thus follows the Binomial$(n, 1-p)$ distribution.

There are two main questions associated with binary deletion channels; see [3] for a survey. First, can deletion channels be used to transmit information efficiently using some encoding scheme? Unlike binary symmetric channels and binary erasure channels, the exact information carrying capacity of the binary deletion channel is unknown. A lower bound on the information carrying capacity of the channel is $(1-p)/9$ [1].

The other question concerns the reconstructability of the original message when it is transmitted across the deletion channel a number of times. This question is motivated in part by the task of sequencing DNA strands. Let $\mathbf{x} \in \{0,1\}^n$ denote the message being transmitted. If $\mathbf{x}$ is chosen uniformly at random, and if p is sufficiently small, then $\mathbf{x}$ can be identified with high probability by looking at a polynomial number of samples from the deletion channel [2]. When p is large, $\exp(\mathrm{O}(\sqrt{n}\log n))$ samples are sufficient for reconstructing any $\mathbf{x}$.

To study the situation when p tends to 1, we will consider an alternative definition for the binary deletion channel. Rather than varying p, we will condition on the number of digits m in the output. This is equivalent to choosing m digits uniformly at random from the input digits; the value of p no longer matters. Our alternative definition is inspired by the difference between the two formulations $G(n,m)$ and $G(n,p)$ of the Erdős–Rényi random graph: $G(n,m)$ has a fixed number of edges while the number of edges under $G(n,p)$ has the Binomial$(\binom{n}{2},p)$ distribution.

By fixing the number of deletions, we remove the option of sending $\mathbf{x}$ through the deletion channel again and again until eventually none of the digits are deleted. It is therefore no longer clear that $\mathbf{x}$ can be reconstructed by studying the deletion channel output.

2. The (m,n)-deletion channel

Let $P(m,\mathbf{x})$ denote the probability distribution on $\{0,1\}^m$ generated by picking $1 \leqslant i_1 < \cdots < i_m \leqslant n$ uniformly at random and returning the sequence $\mathbf{y} := x_{i_1} \cdots x_{i_m}$.

Definition. Let $R(m,n)$ denote the statement

$$\text{for } \mathbf{x} \in \{0,1\}^n\text{: the map } \mathbf{x} \to P(m,\mathbf{x}) \text{ is one-to-one.}$$

If $R(m,n)$ holds then $\mathbf{x}$ can be determined by sampling repeatedly from $P(m,\mathbf{x})$. If not, there is a pair of length-n binary strings that cannot be distinguished over an (m,n)-deletion channel, no matter how many times you sample.

Definition. Let $N_m := \sup\{n : R(m,n)\}$ denote the upper bound on the length of messages an $(m,\cdot\,)$-deletion channel can convey.

Remark. The first few terms of the sequence (N_m) are

$$N_1 = 1, \quad N_2 = 3, \quad N_3 = 6, \quad N_4 = 11, \quad N_5 = 15, \quad N_6 = 29, \quad \ldots.$$

The sequence appears to be growing exponentially. Equivalently, it seems that the elements of $\{0,1\}^n$ can be distinguished using a $(\mathrm{O}(\log n),n)$-deletion channel. This is perhaps unsurprising, given the high-dimensional nature of the $P(m,\mathbf{x})$ probability distributions.

Checking that $R(m,N_m)$ holds for $m \leqslant 6$ was achieved by direct calculation. The difficulty of checking $R(m,n)$ grows very rapidly with m and n. For each $\mathbf{x} \in \{0,1\}^n$, a 2^m-dimensional vector representing $P(m,\mathbf{x})$ has to be calculated. The set of 2^n vectors then has to be searched for duplicates.

To demonstrate that $R(m,N_m+1)$ is false, we must provide a pair of binary sequences of length N_m+1 that produce a $P(m,\cdot)$ collision. Let b^a denote a copies of b, *i.e.*, $0^2 \equiv 00$ and $1^3 \equiv 111$. For $m \in \{1,2,\ldots,6\}$, examples of pairs of strings of length N_m+1 that

https://doi.org/10.1017/S0963548314000522 Published online by Cambridge University Press

produce a collision are:

$$\begin{array}{rcl} 01 \equiv 0^1 1^1 & \text{and} & 1^1 0^1 \equiv 10, \\ 0110 \equiv 0^1 1^2 0^1 & \text{and} & 1^1 0^2 1^1 \equiv 1001, \\ 0^1 1^2 0^3 1^1 & \text{and} & 1^1 0^3 1^2 0^1, \\ 0^1 1^2 0^3 1^1 0^2 1^2 0^1 & \text{and} & 1^1 0^3 1^1 0^1 1^2 0^3 1^1, \\ 0^1 1^2 0^5 1^3 0^4 1^1 & \text{and} & 1^1 0^4 1^3 0^5 1^2 0^1, \\ 0^1 1^2 0^5 1^3 0^4 1^1 0^3 1^3 0^5 1^2 0^1 & \text{and} & 1^1 0^4 1^3 0^5 1^1 0^1 1^2 0^5 1^3 0^4 1^1. \end{array}$$

We have not managed to find N_7 and N_8; without some theoretical advance they are computationally intractable. However, we have established the upper bounds $N_7 < 54$ and $N_8 < 106$. The bounds follow by checking that

$$\begin{aligned} &P(7{,}0^1 1^2 0^5 1^3 0^4 1^1 0^3 1^3 0^5 1^2 0^1 1^2 0^5 1^2 0^1 1^1 0^5 1^3 0^4 1^1) = \\ &P(7{,}1^1 0^4 1^3 0^5 1^1 0^1 1^2 0^5 1^2 0^1 1^2 0^5 1^3 0^3 1^1 0^4 1^3 0^5 1^2 0^1) \end{aligned}$$

and

$$\begin{aligned} &P(8{,}0^1 1^2 0^5 1^3 0^4 1^1 0^3 1^3 0^5 1^2 0^1 1^2 0^5 1^2 0^1 1^1 0^5 1^3 0^4 \hookleftarrow \\ &\qquad \hookrightarrow 1^1 0^3 1^3 0^5 1^1 0^1 1^2 0^5 1^2 0^1 1^2 0^5 1^3 0^3 1^1 0^4 1^3 0^5 1^2 0^1) = \\ &P(8{,}1^1 0^4 1^3 0^5 1^1 0^1 1^2 0^5 1^2 0^1 1^2 0^5 1^3 0^3 1^1 0^4 1^3 0^5 1^1 \hookleftarrow \\ &\qquad \hookrightarrow 0^1 1^2 0^5 1^3 0^4 1^1 0^3 1^3 0^5 1^2 0^1 1^2 0^5 1^2 0^1 1^1 0^5 1^3 0^4 1^1). \end{aligned}$$

These binary strings were found by experimentally concatenating long substrings of the strings that form $P(6,30)$-collisions. Extrapolating from a dangerously small amount of data, these bounds appear to be the right order of magnitude.

We also have an upper bound on the whole sequence (N_m). It is growing no more quickly than exponentially.

Theorem 2.1. *For some constant C, $N_m \leqslant C^m$.*

Proof. If $\mathbf{x} \in \{0,1\}^n$, the probability distributions $P(m,\mathbf{x})$ is characterized by the probability of seeing each of the 2^m elements of $\{0,1\}^m$. The probability of seeing any particular element of $\{0,1\}^m$ is a multiple of $1/\binom{n}{m}$. The pigeonhole principle implies that $R(m,n)$ can only hold if

$$\left(\binom{n}{m}+1\right)^{2^m} \geqslant 2^n. \tag{2.1}$$

Set $n = C^m$, take logs and use the inequality $\binom{n}{m}+1 \leqslant 2n^m$. Inequality (2.1) holds only if

$$2^m(1+m^2\log_2 C) \geqslant C^m.$$

Clearly $R(m,C^m)$ cannot hold if C is sufficiently large. □

Finding good lower bounds for (N_m) seems much more difficult. We will only prove a linear bound.

https://doi.org/10.1017/S0963548314000522 Published online by Cambridge University Press

Theorem 2.2. $N_m \geqslant 2m-1$ *for* $m \geqslant 3$.

Proof of Theorem 2.2. We will show that $\mathbf{x} \in \{0,1\}^{2m-1}$ can be identified using $P(m,\mathbf{x})$. Note that for $j \leqslant m$, we can deduce $P(j,\mathbf{x})$ from $P(m,\mathbf{x})$ using a second deletion channel which discards $m-j$ of its input digits.

Let k denote the number of ones in $\mathbf{x}$: k is simply $2m-1$ times the probability of 1 under $P(1,\mathbf{x})$. By symmetry we can assume $k < m$. The string $\mathbf{x}$ can be written as $k+1$ runs of zeros separated by the k ones. Let $i(0), i(1), \ldots, i(k) \geqslant 0$ denote the length of the runs of zeros, *i.e.*,

$$\mathbf{x} = 0^{i(0)}1^1 0^{i(1)}1^1 \cdots 1^1 0^{i(k)}.$$

The $i(j)$ are determined by the probabilities under $P(k+1,\mathbf{x})$ of the $k+1$ strings containing a single zero and k ones:

$$\text{the probability of } 1^j 0^1 1^{(k-j)} \text{ under } P(k+1,\mathbf{x}) \text{ is } \frac{i(j)}{\binom{2m-1}{k+1}}.$$

□

3. Conclusions

We have introduced an alternative model for deletion channels; it has a non-trivial reconstructability problem. We have explored the space of 'hardest-to-transmit' binary sequences to find N_m for m small. We have also found bounds on N_m for general m.

We conjecture that (N_m) grows exponentially.

References

[1] Drinea, E. and Mitzenmacher, M. (2007) Improved lower bounds for the capacity of i.i.d. deletion and duplication channels. *IEEE Trans. Inform. Theory* **53** 2693–2714.

[2] Holenstein, T., Mitzenmacher, M., Panigrahy, R. and Wieder, U. (2008) Trace reconstruction with constant deletion probability and related results. In *Proc. 19th Annual ACM–SIAM Symposium on Discrete Algorithms: SODA '08*, pp. 389–398.

[3] Mitzenmacher, M. (2009) A survey of results for deletion channels and related synchronization channels. *Probab. Surv.* **6** 1–33.

https://doi.org/10.1017/S0963548314000522 Published online by Cambridge University Press