Upper Capacity Bounds on Binary Deletion Channels

Ray Li

Kalai, Mitzenmacher, Sudan (2010): "Tight Asymptotic Bounds for the Deletion Channel with Small Deletion Probabilities"

Dec 2014

Introduction

Outline

1. Review Deletion Codes
2. Explore existing bounds on BDC capacity
3. Examine a proof for a tight upper bound on BDC capacity for small p. $(C \leq 1 - (1 - o(1))H(p))$

Definitions and Notation

Definition

A *binary deletion channel with deletion probability p* takes a binary string and deletes each bit independently with probability p.

Definition

For transmission of a string $X \in \{0,1\}^n$ according to a binary deletion channel, the *deletion pattern* A is an increasing subsequence of $[n] = \{1, \ldots, n\}$ representing the bits that are *not* deleted.

Notation

For string $X \in \{0,1\}^n$, X_A represents the transmission of X through a deletion channel with deletion pattern A.

Definitions and Notation

Example

Suppose we send $X = 101010$ across a BDC and position 3,4,5 are deleted.

$$X = 10\textcolor{red}{101}0$$
$$A = [1, 2, 6]$$
$$X_A = 100$$

The Big Question

Determine the capacity of a binary deletion channel with deletion probability p.

The Big Question

Determine the capacity of a binary deletion channel with deletion probability p.

- BSC: Well understood
- BEC: Well understood
- BDC: Don't know capacity

Existing Bounds

- Lower bounds
  - [Mitzenmacher '06] $(1-p)/9$
  - $1-2H(p)$
  - **[Gallager '61, Zigangirov '69]** $1-H(p)$
- Upper bounds
  - $1-p$
  - [Mitzenmacher '07] Computer optimized bound, beating $1-p$

d	LB	UB
0.05	0.7283	0.816
0.10	0.5620	0.704
0.15	0.4392	0.6188
0.20	0.3467	0.5507
0.25	0.2759	0.4943
0.30	0.2224	0.4466
0.35	0.1810	0.4063
0.40	0.1484	0.3711
0.45	0.1229	0.33987
0.50	0.1019	0.31082
0.55	0.08432	0.28382
0.60	0.06956	0.25815
0.65	0.05686	0.2331
0.70	0.04532	0.2083
0.75	0.03598	0.183
0.80	0.02727	0.157
0.85	0.01938	0.1298
0.90	0.01238	0.0999*
0.95	0.00574	0.064*

TABLE I

THE LOWER BOUND FROM [6] AND THE UPPER BOUND DERIVED FROM THEOREM 2. ENTRIES DENOTED * ARE WORSE THAN THE $1-d$ BOUND.

    for $p \le .9$.
  - [Mitzenmacher '07] $.7918(1-p)$ as $p \to 1$.

Intuition for Upper Bound of $1 - H(p)$

- Binary Symmetric Channel: Each of N codewords must have approximately $\geq \binom{n}{pn} \approx 2^{H(p)n}$ length n words which map to it under a decoder, so $N2^{H(p)n} \leq 2^n \implies \frac{\log N}{n} \leq 1 - H(p)$
- Binary Deletion Channel: For "most" of the N codewords, when p is small, you can recover the deletion pattern with nontrivial probability. Using the same type of argument, each of the $2^{n(1-p)}$ recieved words should map to one of approximately $N2^{H(p)n}$ codeword-deletion pattern pairs. Then we also get $\frac{\log N}{n}$ is roughly going to be $\leq 1 - H(p)$.

Intuition for Upper Bound of $1 - H(p)$

- Binary Symmetric Channel: Each of N codewords must have approximately $\geq \binom{n}{pn} \approx 2^{H(p)n}$ length n words which map to it under a decoder, so $N2^{H(p)n} \leq 2^n \implies \frac{\log N}{n} \leq 1 - H(p)$
- Binary Deletion Channel: For "most" of the N codewords, when p is small, you can recover the deletion pattern with nontrivial probability. Using the same type of argument, each of the $2^{n(1-p)}$ recieved words should map to one of approximately $N2^{H(p)n}$ codeword-deletion pattern pairs. Then we also get $\frac{\log N}{n}$ is roughly going to be $\leq 1 - H(p)$.
- **Challenge of BDCs**: Asymmetry in channels. Compare deleting bit from 101010 vs 000000

Main Result

Theorem (KMS 2010, Abridged Version)

Suppose that we have a code C and a decoder which can successfully decode for BDC_p with probability at least δ. Then if the length of the code n is sufficiently large, the dimension of the code $\log|C|$ satisfies

$$\frac{\log|C|}{n} \leq 1 - (1 - o(1))H(p)$$

where $o(1)$ vanishes as $p \to 0$.

Main Result

Theorem (KMS 2010)

Suppose there is a code C and a decoder which can successfully decode for BDC_p with probability at least δ, and suppose $n \geq 12\log(4/\delta)/p$. Let $\gamma = 3\log(4/\delta)/np$ and $q' = (1+\gamma)np$. then the dimension of the code $\log|C|$ satisfies,

$$\log|C| \leq n - np(1-\gamma) - \log\binom{n}{np(1-\gamma)} + \log\frac{4}{\delta} + \log\beta$$

where β is given by $\beta = t'(6t'/q')^{3q'+1}$ for $t' = \lceil 3q'\log\frac{ne}{q'} + \log 4\delta\rceil$

Key Claim

Definition

A (q, n) *deletion channel* is a channel that deletes *exactly* q bits of a codeword, with the set of deleted bits chosen uniformly at random.

Claim

Suppose there exists a code C and a decoder for C that suceeds on the (q, n) deletion channel with probability at least δ, where $n \geq 12\log(2/\delta)/p$. Then the dimension of the code satisfies

$$\log|C| \leq n - q - \log(q) + \log\frac{2}{\delta} + \log\alpha$$

where α is given by $\alpha = t(6t/q)^{3q+1}$ for $t = \lceil 3q\log\frac{ne}{q} + \log\frac{2}{q}\rceil$

Proof Sketch: Key Claim $\implies$ Theorem

Goal: Find a q^* near pn such that our decoder succeeds on the (q^*, n) deletion channel with nontrivial probability.

Proof Sketch: Key Claim $\implies$ Theorem

Goal: Find a q^* near pn such that our decoder succeeds on the (q^*, n) deletion channel with nontrivial probability.
Choosing $\gamma = \sqrt{3\log(4/\gamma)/np}$ and $n \geq 12\log(4/\delta)/p$ gives $\gamma \leq \frac{1}{2}$.
Then there must be $q^* \in [(1-\gamma)pn, (1+\gamma)pn]$ such that the success probability of the (q^*, n) deletion channel is at least $\delta/2$.

Proof Sketch: Key Claim $\implies$ Theorem

Goal: Find a q^* near pn such that our decoder succeeds on the (q^*, n) deletion channel with nontrivial probability.
Choosing $\gamma = \sqrt{3\log(4/\gamma)/np}$ and $n \geq 12\log(4/\delta)/p$ gives $\gamma \leq \frac{1}{2}$.
Then there must be $q^* \in [(1-\gamma)pn, (1+\gamma)pn]$ such that the success probability of the (q^*, n) deletion channel is at least $\delta/2$.
Then we use the Key Claim to obtain

$$\log N \leq n - q^* - \log \binom{n}{q^*} + \log \frac{4}{\delta} + \log \alpha^*$$

and using $(1-\gamma)pn \leq q^* \leq (1+\gamma)pn$ we can finish.

$$\log |C| \leq n - np(1-\gamma) - \log \binom{n}{np(1-\gamma)} + \log \frac{4}{\delta} + \log \beta$$

Proof of Key Claim

Definition

The distance between two deletion patterns of equal length A, B is

$$\Delta(A, B) = |\{i | a_i \neq b_i\}|$$

Definition

A word $X \in \{0,1\}^n$ is called t-bad if there exist distinct deletion patterns A, B such that $\Delta(A, B) \geq t$ and $X_A = X_B$.

Examples

If $A = [1, 3, 4, 5], B = [1, 4, 5, 6]$, are deletion patterns for $n = 6$, then $\Delta(A, B) = 3$.
11110000 is 6-bad but not 7-bad.
10101010 is not 1-bad.

Proof of Key Claim

Lemma

For any $t \geq 1$, there are at most $\binom{n}{q}^2 2^{n-t}$ different t-bad strings $X \in \{0,1\}^n$.

Using the lemma, we can choose a large t ($t = 3q \log \frac{ne}{q} + \log \frac{2}{\delta}$) so that

$$\Pr_{Z \in C}[\mathrm{Dec}(Z_A) = Z \wedge Z \text{ is not } t\text{-bad}] \geq \delta - \binom{n}{q}^2 \frac{2^{n-t}}{N} \geq \delta/2$$

Proof of Key Claim

Lemma

Take $\alpha = t(6t/q)^{3q+1}$. For any A, α is an upper bound on the number of B such that $\Delta(A,B) \leq t-1$.
(One can first compute $(t-1)\binom{2q+t}{2q+1}\binom{q+t-1}{q} < \alpha$ as an upper boun)

Conditioned on decoding suceeding and codeword not being t-bad, each deletion pattern is equally likely, so we can recover the deletion pattern with probability at least α^{-1}.
Then the probability that we can recover the codeword *and* the deletion pattern is $\geq \delta\alpha^{-1}/2$, But the probability of recovering deletion is at most $\frac{2^{n-q}}{N\binom{n}{q}}$, so it follows that $\frac{2^{n-q}}{N\binom{n}{q}} \geq \delta\alpha^{-1}/2$ and

$$\log N \leq n - q - \log(q) + \log\frac{2}{\delta} + \log\alpha$$

Open Question

- Tighten capacity upper bounds for general p.