

Bounds on the Capacity of Channels with Insertions, Deletions and Substitutions

Dario Fertonani

Advisor: Prof. Tolga M. Duman
Department of Electrical Engineering
Fulton School of Engineering
Arizona State University

School of Information Theory
Northwestern University
August 11, 2009

Abstract

- Some systems affected by **synchronization errors** can be modeled as binary channels with insertions, deletions, and substitutions.
- In the Sixties, the relevant capacity was defined and the coding theorem was proved, but **the capacity is currently unknown**.
- **Capacity bounds** are available in the literature, but the gap between the upper and lower bounds is large in most scenarios.
- **We derive upper and lower bounds**, exploiting an auxiliary genie-aided system and suitable information-theoretic inequalities.
- In most scenarios, the proposed bounds improve the existing ones, significantly **narrowing the possible capacity region**.

Channel Model

- We consider the channel model proposed by Gallager in 1961, with i.i.d. insertion, deletion, and substitution errors.
- The channel input is a sequence of N bits $\mathbf{X} = \{X_n\}_{n=1}^N$.
- In the basic insertion-deletion model, each input bit gets deleted (with probability d), or experiences an insertion error (with probability i), or is correctly received (with probability $1 - d - i$).
- In the more general case, the output of the insertion-deletion channel is observed through a binary-symmetric channel with substitution probability s .
- The channel output is a sequence of M bits $\mathbf{Y} = \{Y_n\}_{n=1}^M$, M being a random variable depending on the number of insertions/deletions.
- The positions of insertions, deletions, and substitutions are random and unknown to either transmitter and receiver.

Transition Probabilities

	$X_n = 0$	$X_n = 1$
$Z_n = \emptyset$	d	d
$Z_n = 0$	$(1 - d - i)(1 - s)$	$(1 - d - i)s$
$Z_n = 1$	$(1 - d - i)s$	$(1 - d - i)(1 - s)$
$Z_n = 00$	$i/4$	$i/4$
$Z_n = 01$	$i/4$	$i/4$
$Z_n = 10$	$i/4$	$i/4$
$Z_n = 11$	$i/4$	$i/4$

Table: $P(Z_n|X_n)$

- The auxiliary non-binary output sequence $\mathbf{Z} = \{Z_n\}_{n=1}^N$ allows a memoryless description of the channel, unlike \mathbf{Y} .
- A bit that experiences an **insertion error** is replaced by two random bits (Gallager model).

Channel Capacity

- The **capacity per input bit** is defined as

$$C = \lim_{N \rightarrow \infty} \frac{1}{N} \max_{P(\mathbf{X})} I(\mathbf{X}; \mathbf{Y})$$

where $P(\mathbf{X})$ is the distribution of the input sequence, and $I(\cdot; \cdot)$ is the average mutual information between two random sequences.

- The relevant **coding theorem** was proved by Dobrushin in 1967.
- **The capacity has been unknown since the problem was formulated.**
- Only **upper bounds** and **lower bounds** on C are available.

Existing Capacity Bounds

General Case

- The benchmark **lower bound** is the one proposed by Gallager in 1961:

$$C \geq 1 + d \log_2 d + i \log_2 i + P_s \log_2 P_s + P_t \log_2 P_t ,$$

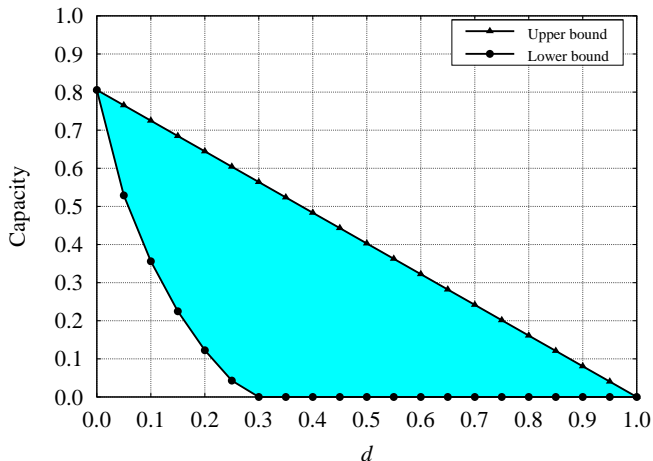
where $P_s = (1 - d - i)s$ and $P_t = (1 - d - i)(1 - s)$.

- The benchmark **upper bound** is the trivial one obtained by revealing the positions of all insertions/deletions to the receiver:

$$C \leq (1 - d - i) (1 + s \log_2 s + (1 - s) \log_2(1 - s)) .$$

Deletion Channel

- Only deletions are possible ($i = s = 0$).
- The benchmark **lower bound** was proposed by Drinea *at al.* in 2007.
- The benchmark **upper bound** was proposed by Diggavi *at al.* in 2007.

Numerical Example ($i = 0, s = 0.03$)

Large gap between the existing upper and lower bounds!

Rationale of Our Approach

- We exploit an auxiliary system identical to the considered one, with additional **genie-aided information** on the insertion/deletion process revealed to the receiver.
- The revealed information allows us to simplify

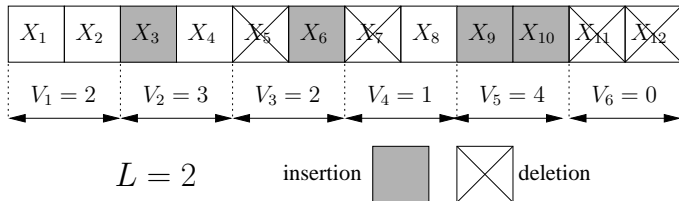
$$\lim_{N \rightarrow \infty} \frac{1}{N} \max_{P(\mathbf{X})} I(\mathbf{X}; \cdot)$$

such that only **finite-length sequences** are to be considered.

- For finite-length sequences, we can maximize $I(\mathbf{X}; \cdot)$ over the distribution $P(\mathbf{X})$ by means of the **Blahut-Arimoto algorithm**.

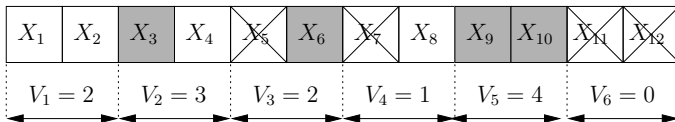
A Useful Auxiliary Process

- Let L be a positive integer parameter.
- We partition the input sequence \mathbf{X} into $Q = N/L$ subsequences $\{\mathbf{X}_q\}_{q=1}^Q$ of L consecutive bits. For example, when $L = 2$, we have $\mathbf{X}_1 = (X_1, X_2)$, $\mathbf{X}_2 = (X_3, X_4)$, $\mathbf{X}_3 = (X_5, X_6)$, and so on.
- We partition the output sequence \mathbf{Y} into Q subsequences $\{\mathbf{Y}_q\}_{q=1}^Q$ such that \mathbf{Y}_q includes the received bits related to the input subsequence \mathbf{X}_q .
- We define the random process $\mathbf{V} = \{V_q\}_{q=1}^Q$ such that V_q denotes the number of bits in the subsequence \mathbf{Y}_q .



Example of Auxiliary Process

The process \mathbf{V} is i.i.d. and does not depend on the substitution probability, since the substitutions do not alter the number of received bits.



When $L = 2$, as in the example, the probability distribution of V_q is

$$P(V_q) = \begin{cases} d^2 & \text{if } V_q = 0 \\ 2d(1 - d - i) & \text{if } V_q = 1 \\ (1 - d - i)^2 + 2di & \text{if } V_q = 2 \\ 2i(1 - d - i) & \text{if } V_q = 3 \\ i^2 & \text{if } V_q = 4 \\ 0 & \text{else} \end{cases},$$

which allows us to compute the entropy $H(V_q)$, required for the bounds.

Auxiliary System and Capacity Bounds

- We consider a system identical to the system of interest, with an **additional “parallel” channel** that provides the sequence \mathbf{V} to the receiver. Its capacity per input bit is

$$C_A = \lim_{N \rightarrow \infty} \max_{P(\mathbf{X})} \frac{1}{N} I(\mathbf{X}; \mathbf{Y}, \mathbf{V}) .$$

- Since $I(\mathbf{X}; \mathbf{Y}) = I(\mathbf{X}; \mathbf{Y}, \mathbf{V}) - I(\mathbf{X}; \mathbf{V} | \mathbf{Y})$, basic **information-theoretic inequalities** assure that

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &\leq I(\mathbf{X}; \mathbf{Y}, \mathbf{V}) \\ I(\mathbf{X}; \mathbf{Y}) &\geq I(\mathbf{X}; \mathbf{Y}, \mathbf{V}) - H(\mathbf{V}) . \end{aligned}$$

- Hence, the following **bounds on the capacity of interest** result

$$\begin{aligned} C &\leq C_A \\ C &\geq C_A - \lim_{N \rightarrow \infty} \frac{1}{N} H(\mathbf{V}) = C_A - \frac{1}{L} H(V_q) . \end{aligned}$$

Remarks

- For the bounds to be computed, **we need to evaluate C_A and $H(V_q)$** .
- $H(V_q)$ is the entropy of a simple memoryless process, which can be evaluated by means of **combinatorial analyses**.
- The capacity C_A of the genie-aided system can be written as

$$C_A = \lim_{N \rightarrow \infty} \max_{P(\mathbf{X})} \frac{1}{N} I(\mathbf{X}; \mathbf{Y}, \mathbf{V}) = \frac{1}{L} \max_{P(\mathbf{X}_q)} I(\mathbf{X}_q; \mathbf{Y}_q, V_q) = \frac{1}{L} \max_{P(\mathbf{X}_q)} I(\mathbf{X}_q; \mathbf{Y}_q)$$

since, revealed \mathbf{V} to the receiver, different subsequences $\{\mathbf{X}_q\}$ do not interfere with each other.

- Since we have now a memoryless channel with finite input/output alphabets, the **Blahut-Arimoto algorithm** allows us to evaluate C_A .

Application of the Blahut-Arimoto Algorithm

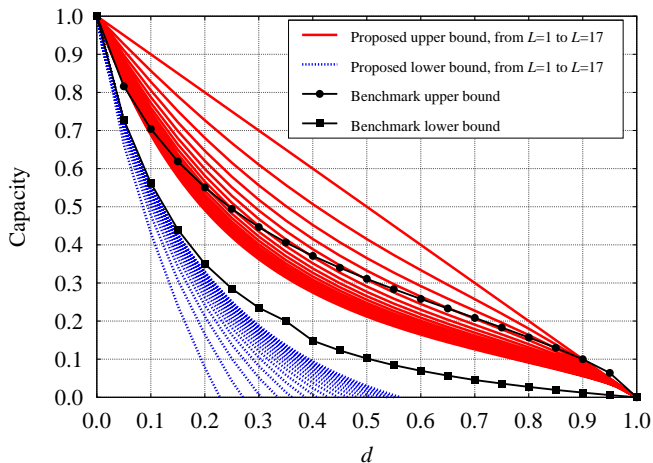
Based on tables including the **transition probabilities of the auxiliary memoryless channel**, we can evaluate the relevant capacity.

\mathbf{X}_q	\mathbf{Y}_q						
	\emptyset	0	1	00	01	10	11
00	d^2	$2rd$	0	r^2	0	0	0
01	d^2	rd	rd	0	r^2	0	0
10	d^2	rd	rd	0	0	r^2	0
11	d^2	0	$2rd$	0	0	0	r^2

Example: transition probability $P(\mathbf{Y}_q|\mathbf{X}_q)$ for $L = 2$ and $i = s = 0$, $r = 1 - d$.

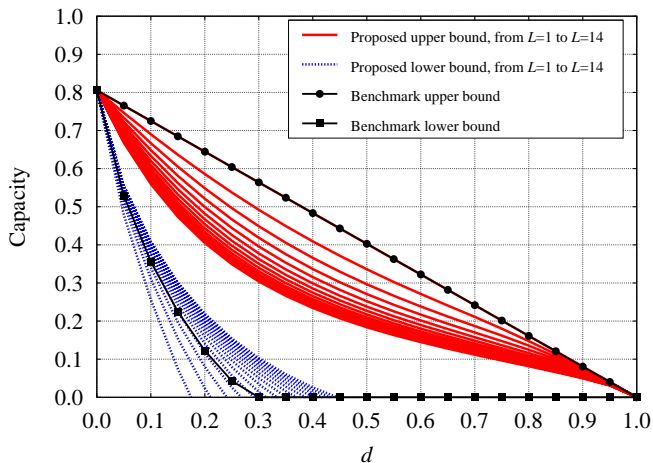
Implementation issues: the algorithm becomes prohibitively memory-consuming as L increases. The maximum values that we could manage are $L = 17$ when $i = s = 0$, and $L = 8$ in the most general case.

Comparisons ($i = s = 0$)



The upper bound is improved for most values of d .

Comparisons ($i = 0, s = 0.03$)



Both bounds are improved for all values of d .

Comparisons

d	i	s	Current LB	Novel LB	Current UB	Novel UB
0.01	0.01	0.01	0.759	0.766	0.901	0.863
0.01	0.03	0.01	0.647	0.661	0.883	0.808
0.01	0.10	0.01	0.379	0.412	0.819	0.642
0.03	0.01	0.01	0.647	0.662	0.883	0.808
0.03	0.03	0.01	0.536	0.564	0.865	0.750
0.03	0.10	0.01	0.271	0.329	0.800	0.583
0.10	0.01	0.01	0.379	0.419	0.819	0.649
0.10	0.03	0.01	0.271	0.335	0.800	0.589
0.10	0.10	0.01	0.013	0.139	0.736	0.438

Both bounds are improved!

The improvement increases as insertions and deletions become more likely.