On the Size Distribution of Levenshtein Balls with Radius One

Geyang Wang* Qi Wang

Department of Computer Science and Engineering
Southern University of Science and Technology (SUSTech), China

March 10, 2022

Table of contents

Hamming ball

- Hamming distance:
  For two $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, $d_H(\boldsymbol{x}, \boldsymbol{y}) =$ min# substitutions needed to transform $\boldsymbol{x}$ into $\boldsymbol{y}$.

Hamming ball

- Hamming distance:

  For two $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, $d_H(\boldsymbol{x}, \boldsymbol{y}) =$ min# substitutions needed to transform $\boldsymbol{x}$ into $\boldsymbol{y}$.

- Hamming ball:

  For $\boldsymbol{x} \in \mathbb{Z}_m^n$, the Hamming ball centered at $\boldsymbol{x}$ is

$$B_t(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathbb{Z}_m^n \mid d_H(\boldsymbol{x}, \boldsymbol{y}) \le t\}.$$

Hamming ball

- Hamming distance:

  For two $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, $d_H(\boldsymbol{x}, \boldsymbol{y}) =$ min# substitutions needed to transform $\boldsymbol{x}$ into $\boldsymbol{y}$.

- Hamming ball:

  For $\boldsymbol{x} \in \mathbb{Z}_m^n$, the Hamming ball centered at $\boldsymbol{x}$ is

$$B_t(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathbb{Z}_m^n \mid d_H(\boldsymbol{x}, \boldsymbol{y}) \leq t\}.$$

The size of a Hamming t-ball can be explicitly determined:

$$|B_t(\boldsymbol{x})| = \sum_{i=0}^{t} \binom{n}{i} (m-1)^i.$$

Hamming ball

- Hamming distance:

  For two $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, $d_H(\boldsymbol{x}, \boldsymbol{y}) =$ min# substitutions needed to transform $\boldsymbol{x}$ into $\boldsymbol{y}$.

- Hamming ball:

  For $\boldsymbol{x} \in \mathbb{Z}_m^n$, the Hamming ball centered at $\boldsymbol{x}$ is

$$B_t(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathbb{Z}_m^n \mid d_H(\boldsymbol{x}, \boldsymbol{y}) \leq t\}.$$

The size of a Hamming t-ball can be explicitly determined:

$$|B_t(\boldsymbol{x})| = \sum_{i=0}^{t} \binom{n}{i} (m-1)^i.$$

$|B_t(\boldsymbol{x})|$ is independent with $\boldsymbol{x}$.

Error-correcting codes in Hamming distance

- For a code $C \subseteq \mathbb{Z}_m^n$,

  C can correct t substitution errors.

  $\Leftrightarrow$ $\min\{d_H(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x}, \boldsymbol{y} \in C\} = d \geq 2t + 1$.

  $\Leftrightarrow$ All $B_t(\boldsymbol{x})$'s are disjoint.

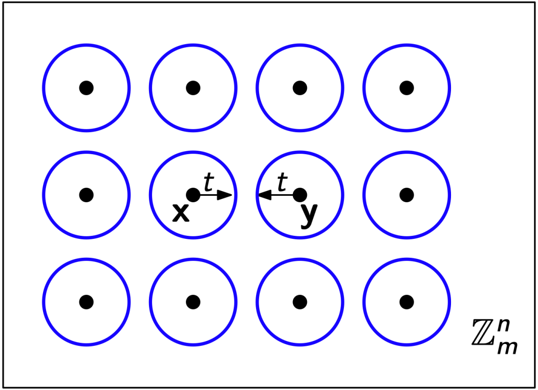
Error-correcting codes in Hamming distance

- For a code $C \subseteq \mathbb{Z}_m^n$,

  C can correct t substitution errors.

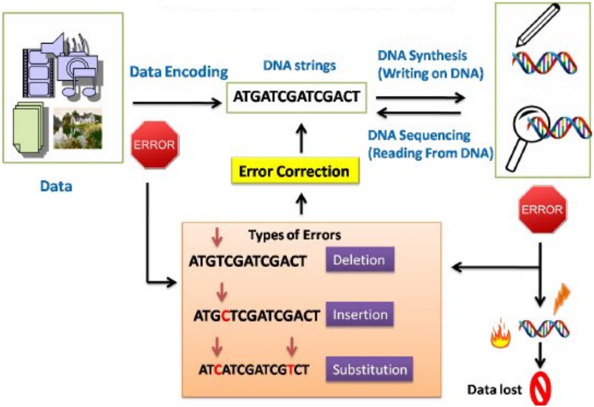
  $\Leftrightarrow$ $\min\{d_H(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x}, \boldsymbol{y} \in C\} = d \geq 2t + 1$.

  $\Leftrightarrow$ All $B_t(\boldsymbol{x})$'s are disjoint.



Insertion/deletion channels

DNA storage systems



[1]

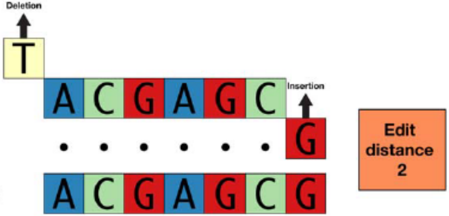
Types of errors

- substitution, insertion, deletion.

[1] picture from Limbachiya et al, Natural data storage: A review on sending information from now to then via nature.

For two words $\boldsymbol{x}, \boldsymbol{y}$.

Levenshtein (edit) distance: $d_E(\boldsymbol{x}, \boldsymbol{y}) = \min\#$ of insertions and deletions needed to transform $\boldsymbol{x}$ into $\boldsymbol{y}$.



Levenshtein balls

For two words $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$.

Fixed-length Levenshtein (FLL) distance: $d_L(\boldsymbol{x}, \boldsymbol{y})$ = the smallest t such that $\boldsymbol{x}$ can be transformed into $\boldsymbol{y}$ by t insertions and t deletions.

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = 2d_L(\boldsymbol{x}, \boldsymbol{y}).$$

Levenshtein balls

For two words $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$.

Fixed-length Levenshtein (FLL) distance: $d_L(\boldsymbol{x}, \boldsymbol{y})$ = the smallest t such that $\boldsymbol{x}$ can be transformed into $\boldsymbol{y}$ by t insertions and t deletions.

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = 2d_L(\boldsymbol{x}, \boldsymbol{y}).$$

The Levenshtein ball centered at $\boldsymbol{x}$ is

$$L_t(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathbb{Z}_m^n \mid d_L(\boldsymbol{x}, \boldsymbol{y}) \leq t\},$$

t is called radius.

What do we know about $L_t(\boldsymbol{x})$?

> *"Channels with synchronization errors, including both insertions and deletions as well as more general timing errors, are simply not adequately understood by current theory. Given the near-complete knowledge we have for channels with erasures and errors . . . our lack of understanding about channels with synchronization errors is truly* ***remarkable****."*
>
> — *Mitzenmacher, 2009.*[2]

Even the fundamental problem of counting $|L_t(\boldsymbol{x})|$ still remains elusive!

- Explicit formula of $|L_1(\boldsymbol{x})|$. Bounds of $|L_t(\boldsymbol{x})|$ for $t > 1$;
  [2013, Sala and Dolecek[3]]
- Minimum, maximum and average value of $|L_1(\boldsymbol{x})|$;
  [2021, Bar-Lev, Etzion, and Yaakobi[4]]

[2] A survey of results for deletion channels and related synchronization channels, Probab. Surv. vol. 6, pp. 1-33, 2009.

[3] Counting sequences obtained from the synchronization channel, IEEE ISIT 2013.

[4] On Levenshtein balls with radius one, IEEE ISIT 2021.

The Size of $L_1(\boldsymbol{x})$

Runs and maximal alternating segments

The size of $L_1(\boldsymbol{x})$ is related to the following two functions. For two distinct elements $a, b \in \mathbb{Z}_m$.

The Size of $L_1(\boldsymbol{x})$

Runs and maximal alternating segments

The size of $L_1(\boldsymbol{x})$ is related to the following two functions. For two distinct elements $a, b \in \mathbb{Z}_m$.

- Runs: A maximal subsequence with the form $aa\dots a$.
  $\rho(\boldsymbol{x}) = \#$ runs in $\boldsymbol{x}$.

The Size of $L_1(\boldsymbol{x})$

Runs and maximal alternating segments

The size of $L_1(\boldsymbol{x})$ is related to the following two functions. For two distinct elements $a, b \in \mathbb{Z}_m$.

- Runs: A maximal subsequence with the form $aa \ldots a$.
  $\rho(\boldsymbol{x}) = \#$ runs in $\boldsymbol{x}$.
- Maximal alternating segments: A maximal subsequence with the form $abab \ldots ab$ or $abab \ldots ba$.
  $a(\boldsymbol{x}) = \#$ maximal alternating segments in $\boldsymbol{x}$.

The Size of $L_1(\boldsymbol{x})$

Runs and maximal alternating segments

The size of $L_1(\boldsymbol{x})$ is related to the following two functions. For two distinct elements $a, b \in \mathbb{Z}_m$.

- Runs: A maximal subsequence with the form $aa\ldots a$.
  $\rho(\boldsymbol{x}) = \#$ runs in $\boldsymbol{x}$.
- Maximal alternating segments: A maximal subsequence with the form $abab\ldots ab$ or $abab\ldots ba$.
  $a(\boldsymbol{x}) = \#$ maximal alternating segments in $\boldsymbol{x}$.
- For binary $\boldsymbol{x}$, $\rho(\boldsymbol{x}) + a(\boldsymbol{x}) = n + 1$.

The Size of $L_1(\boldsymbol{x})$

Runs and maximal alternating segments

The size of $L_1(\boldsymbol{x})$ is related to the following two functions. For two distinct elements $a, b \in \mathbb{Z}_m$.

- Runs: A maximal subsequence with the form $aa\ldots a$.
  $\rho(\boldsymbol{x}) = \#$ runs in $\boldsymbol{x}$.
- Maximal alternating segments: A maximal subsequence with the form $abab\ldots ab$ or $abab\ldots ba$.
  $a(\boldsymbol{x}) = \#$ maximal alternating segments in $\boldsymbol{x}$.
- For binary $\boldsymbol{x}$, $\rho(\boldsymbol{x}) + a(\boldsymbol{x}) = n + 1$.

Example

Let $\boldsymbol{x} = 01100101$.

- Runs: 0,11,00,1,0,1; $\rho(\boldsymbol{x}) = 6$.
- Maximal alternating segments: 01,10,0101. $a(\boldsymbol{x}) = 3$.
- $\rho(\boldsymbol{x}) + a(\boldsymbol{x}) = 9$.

The Size of $L_1(\boldsymbol{x})$

For all $\boldsymbol{x} \in \mathbb{Z}_m^n$, [5]

$$|L_1(\boldsymbol{x})| = \rho(\boldsymbol{x})(mn - n - 1) + 2 - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2 + \frac{3}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i - a(\boldsymbol{x}),$$

where s_i, for $1 \leq i \leq a(\boldsymbol{x})$, is the length of the i-th maximal alternating segment of $\boldsymbol{x}$.

[5] F. Sala and L. Dolecek, Counting sequences obtained from the synchronization channel, IEEE ISIT 2013.

Minimum, maximum and average size of $L_1(\boldsymbol{x})$

Lemma (Bar-Lev et al. ISIT 2021)

Let $n > t \geq 0$. The minimum of $|L_t(\boldsymbol{x})|$ is obtained if and only if $\boldsymbol{x} = a^n$ for some $a \in \mathbb{Z}_m$ (eg. $\boldsymbol{x} = 00\ldots0$). In this case, $L_t(\boldsymbol{x}) = B_t(\boldsymbol{x})$.

Minimum, maximum and average size of $L_1(\boldsymbol{x})$

Lemma (Bar-Lev et al. ISIT 2021)

Let $n > t \geq 0$. The minimum of $|L_t(\boldsymbol{x})|$ is obtained if and only if $\boldsymbol{x} = a^n$ for some $a \in \mathbb{Z}_m$ (eg. $\boldsymbol{x} = 00\ldots0$). In this case, $L_t(\boldsymbol{x}) = B_t(\boldsymbol{x})$.

- Minimum size of $L_1(\boldsymbol{x})$:

$$\min_{\boldsymbol{x}\in\mathbb{Z}_2^n} |L_1(\boldsymbol{x})| = n(m-1)+1.$$

Minimum, maximum and average size of $L_1(\boldsymbol{x})$

Lemma (Bar-Lev et al. ISIT 2021)

Let $n > t \geq 0$. The minimum of $|L_t(\boldsymbol{x})|$ is obtained if and only if $\boldsymbol{x} = a^n$ for some $a \in \mathbb{Z}_m$ (eg. $\boldsymbol{x} = 00\ldots0$). In this case, $L_t(\boldsymbol{x}) = B_t(\boldsymbol{x})$.

- Minimum size of $L_1(\boldsymbol{x})$:

$$\min_{\boldsymbol{x}\in\mathbb{Z}_2^n} |L_1(\boldsymbol{x})| = n(m-1)+1.$$

- Average size of $L_1(\boldsymbol{x})$:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_m^n}[|L_1(\boldsymbol{x})|] = n^2(m+\frac{1}{m}-2)+2-\frac{n}{m}+\frac{m^n-1}{m^n(m-1)}.$$

Minimum, maximum and average size of $L_1(\boldsymbol{x})$

Lemma (Bar-Lev et al. ISIT 2021)

Let $n > t \geq 0$. The minimum of $|L_t(\boldsymbol{x})|$ is obtained if and only if $\boldsymbol{x} = a^n$ for some $a \in \mathbb{Z}_m$ (eg. $\boldsymbol{x} = 00\ldots0$). In this case, $L_t(\boldsymbol{x}) = B_t(\boldsymbol{x})$.

- Minimum size of $L_1(\boldsymbol{x})$:

$$\min_{\boldsymbol{x}\in\mathbb{Z}_2^n} |L_1(\boldsymbol{x})| = n(m-1)+1.$$

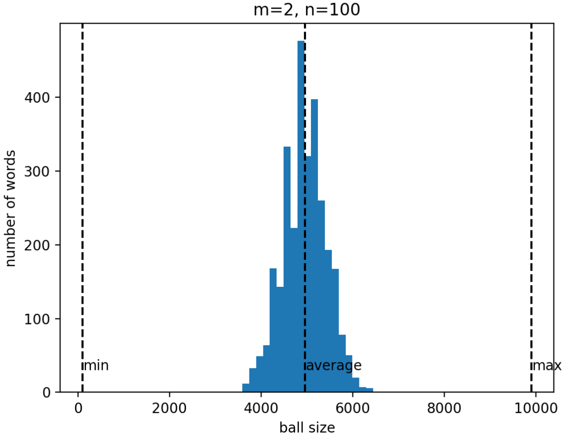
- Average size of $L_1(\boldsymbol{x})$:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_m^n}[|L_1(\boldsymbol{x})|] = n^2(m+\frac{1}{m}-2)+2-\frac{n}{m}+\frac{m^n-1}{m^n(m-1)}.$$

- Maximum size of $L_1(\boldsymbol{x})$:

$$\max_{\boldsymbol{x}\in\mathbb{Z}_m^n} |L_1(\boldsymbol{x})| = \begin{cases} n^2(m-1)-n+2, & \text{if } m>2; \\ n^2-\sqrt{2}n^{\frac{3}{2}}+O(n). & \text{if } m=2. \end{cases}$$

We study the distribution of $|L_1(\boldsymbol{x})|$ for random $\boldsymbol{x}$'s.



The distribution of $L_1(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{Z}_2^{100}$.

$\min_{\boldsymbol{x}} |L_1(\boldsymbol{x})| = 101, \mathbb{E}[|L_1(\boldsymbol{x})|] = 4953. \max_{\boldsymbol{x}} |L_1(\boldsymbol{x})| \geq 9902.$

Martingale

Definition (martingale)

A martingale is a sequence of real random variables $Z_0, \ldots, Z_n$ with finite expectation such that for each $0 \leq i < n$,

$$\mathbb{E}\left[Z_{i+1} | Z_i, Z_{i-1}, \ldots, Z_0\right] = Z_i.$$

Martingale

Definition (martingale)

A martingale is a sequence of real random variables $Z_0, \ldots, Z_n$ with finite expectation such that for each $0 \leq i < n$,

$$\mathbb{E}\left[Z_{i+1} | Z_i, Z_{i-1}, \ldots, Z_0\right] = Z_i.$$

Let $X_1, \ldots, X_n$ be the underlying random variables (not necessary independent) and f be a function over $X_1, \ldots, X_n$. The Doob martingale $Z_0, \ldots, Z_n$ is defined by

$$Z_0 = \mathbb{E}\left[f(X_1, \ldots, X_n)\right];$$
$$Z_i = \mathbb{E}\left[f(X_1, \ldots, X_n) | X_1, \ldots, X_i\right] \text{ for } i \in [n].$$

Note: $[n] = \{1, 2, \ldots, n\}$, $Z_n = f(X_1, \ldots, X_n)$.

Azuma's inequality [6]

Let $Z_0, Z_1, \ldots, Z_n$ be a martingale such that for each $i \in [n]$,

$$|Z_i - Z_{i-1}| \leq c_i.$$

Then for every $\lambda > 0$, we have

$$\Pr(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-\lambda^2}{2(c_1^2 + \cdots + c_n^2)}\right),$$

and

$$\Pr(Z_n - Z_0 \leq -\lambda) \leq \exp\left(\frac{-\lambda^2}{2(c_1^2 + \cdots + c_n^2)}\right).$$

[6] Alon, Noga, and Joel H. Spencer. *The probabilistic method.* John Wiley & Sons, 2016.

Main results (binary case)

Let $n > 3$ be an integer[7] and $x_1, \ldots, x_n$ be independent random variables such that $\Pr(x_i = 0) = \Pr(x_i = 1) = \frac{1}{2}$ for $i \in [n]$. Then for the word $\boldsymbol{x} = x_1, \ldots, x_n$, we have

$$\Pr\left(|L_1(\boldsymbol{x})| - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_2^n}[|L_1(\boldsymbol{x})|] \geq cn\sqrt{n-1}\right) \leq e^{-2c^2},$$

and

$$\Pr\left(|L_1(\boldsymbol{x})| - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_2^n}[|L_1(\boldsymbol{x})|] \leq -cn\sqrt{n-1}\right) \leq e^{-2c^2},$$

where $\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_2^n}[|L_1(\boldsymbol{x})|] = \frac{n^2}{2} - \frac{n}{2} - \frac{1}{2^n} + 3$, and c is a positive constant.

[7] The case when $n \leq 3$ is trivial.

Main results (m-ary case)

Let $m > 2, n > 3$ be integers, and $x_1, \ldots, x_n$ be independent random variables such that $\Pr(x_i = j) = \frac{1}{m}$ for $i \in [n]$, $j \in \mathbb{Z}_m$. Then for the word $\boldsymbol{x} = x_1, \ldots, x_n$, we have

$$\Pr\left(|L_1(\boldsymbol{x})| - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_m^n}[|L_1(\boldsymbol{x})|] \geq c(m + \frac{1}{m})n\sqrt{n-1}\right) \leq e^{-c^2/2},$$

and

$$\Pr\left(|L_1(\boldsymbol{x})| - \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_m^n}[|L_1(\boldsymbol{x})|] \leq -c(m + \frac{1}{m})n\sqrt{n-1}\right) \leq e^{-c^2/2},$$

where $\mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathbb{Z}_m^n}[|L_1(\boldsymbol{x})|] = n^2(m + \frac{1}{m} - 2) + 2 - \frac{n}{m} + \frac{1}{m-1} - \frac{1}{m^n(m-1)}$, and c is a positive constant.

Proof of the binary case (sketch) I

Recall that for $\boldsymbol{x} \in \mathbb{Z}_m^n$,

$$|L_1(\boldsymbol{x})| = \rho(\boldsymbol{x})(mn - n - 1) + 2 - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2 + \frac{3}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i - a(\boldsymbol{x}),$$

Putting $m = 2$, we have

$$|L_1(\boldsymbol{x})| = \rho(\boldsymbol{x})(n - 1) + 2 - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2 + \frac{3n}{2} - a(\boldsymbol{x}).$$

Recall that $a(\boldsymbol{x}) + \rho(\boldsymbol{x}) = n + 1$ for $m = 2$, we have

$$|L_1(\boldsymbol{x})| = \rho(\boldsymbol{x})n - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2 + \frac{n}{2} + 1.$$

Proof of the binary case (sketch) II

Define

$$f_n(\boldsymbol{x}) = \rho(\boldsymbol{x})n - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2.$$

Then $|L_1(\boldsymbol{x})| = f_n(\boldsymbol{x}) + 1 + \frac{n}{2}$, it suffices to consider the distribution of $f_n(\boldsymbol{x})$. Define the Doob martingale $Z_0 = \mathbb{E}\left[f_n(\boldsymbol{x})\right]$, $Z_i = \mathbb{E}\left[f_n(\boldsymbol{x}) | x_1, \ldots, x_i\right]$.

Claim

$$|Z_1 - Z_0| = 0;$$

$$|Z_i - Z_{i-1}| \leq \frac{n}{2} \text{ for } 2 \leq i \leq n.$$

Then by Azuma's inequality, we have $\Pr(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-2\lambda^2}{n^2(n-1)}\right)$.

Take $\lambda = cn\sqrt{n-1}$, the result then follows.

Proof of the claim

First, show that

$$f_n(\boldsymbol{x}) = \begin{cases} f_n(\boldsymbol{x}_{[1,i]}) + f_n(\boldsymbol{x}_{[i+1,n]}) - n & \text{if } x_i = x_{i+1}, \\ f_n(\boldsymbol{x}_{[1,i]}) + f_n(\boldsymbol{x}_{[i+1,n]}) - t(\boldsymbol{x}_{[1,i]})h(\boldsymbol{x}_{[i+1,n']}) & \text{if } x_i \neq x_{i+1}. \end{cases}$$

- $\boldsymbol{x}_{[1,i]} = (x_1, x_2, \ldots, x_i)$.
- $h(\cdot), t(\cdot)$ = length of the first/last maximal alternating segment.

Then calculate Z_i.

$$Z_i = \mathbb{E}\left[f_n(\boldsymbol{x}) \mid x_{[1,i]}\right] = \frac{1}{2}\mathbb{E}\left[f_n(\boldsymbol{x}) \mid x_{[1,i]}, x_{i+1} = x_i\right] + \frac{1}{2}\mathbb{E}\left[f_n(\boldsymbol{x}) \mid x_{[1,i]}, x_{i+1} \neq x_i\right].$$

Finally, bound $|Z_i - Z_{i-1}|$.

Proof of the m-ary case (sketch)

Define

$$f_{m,n}(\boldsymbol{x}) = \rho(\boldsymbol{x})(mn - n - 1) - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2 + \frac{3}{2}\sum_{i=1}^{a(\boldsymbol{y})} s_i - a(\boldsymbol{x}).$$

Then for $\boldsymbol{x} \in \mathbb{Z}_m^n$, $|L_1(x)| = f_{m,n}(\boldsymbol{x}) + 2$. Define Doob martingale $Z_0 = \mathbb{E}\left[f_{m,n}(\boldsymbol{x})\right]$, $Z_i = \mathbb{E}\left[f_{m,n}(\boldsymbol{x})|x_1, \dots, x_i\right]$. Then estimate Z_i by "breaking" $f_{m,n}(\boldsymbol{x})$ into two parts and bounding the differences, we have

$$|Z_i - Z_{i-1}| \leq \begin{cases} 0, & \text{if } i = 1; \\ n(m + \frac{1}{m} - 2) + 3, & \text{if } i = 2; \\ n(m + \frac{1}{m}), & \text{if } 2 < i < n - 1; \\ n(m - 1), & \text{if } i = n - 1; \\ n(m + \frac{1}{m} - 2), & \text{if } i = n; \end{cases}$$

The result then follows from Azuma's inequality.

Conclusion

- We prove that the size of the Levenshtein ball with radius one is highly concentrated around its mean.

[8] Gilbert-Varshamov-like lower bounds for deletion-correcting codes, 2014 IEEE ITW

Conclusion

- We prove that the size of the Levenshtein ball with radius one is highly concentrated around its mean.
- Explicitly determine or improve the bounds on $|L_t(\boldsymbol{x})|$ for $t > 1$.

[8] Gilbert-Varshamov-like lower bounds for deletion-correcting codes, 2014 IEEE ITW

Conclusion

- We prove that the size of the Levenshtein ball with radius one is highly concentrated around its mean.

- Explicitly determine or improve the bounds on $|L_t(\boldsymbol{x})|$ for $t > 1$.
- Study the size distribution for $|L_t(\boldsymbol{x})|$.

[8] Gilbert-Varshamov-like lower bounds for deletion-correcting codes, 2014 IEEE ITW

Conclusion

- We prove that the size of the Levenshtein ball with radius one is highly concentrated around its mean.

- Explicitly determine or improve the bounds on $|L_t(\boldsymbol{x})|$ for $t > 1$.
- Study the size distribution for $|L_t(\boldsymbol{x})|$.
- Let $C_m(n, t)$ denote the size of the largest t-insertion-deletion-correcting code over $\mathbb{Z}_m^n$. Improve the bound on $C_m(n, t)$.

  Connection[8] between codes and independent set of graphs.
  Let $V = \mathbb{Z}_m^n$ and $(\boldsymbol{x}, \boldsymbol{y}) \in E$ iff $d_L(\boldsymbol{x}, \boldsymbol{y}) \le t$.
  Gilbert-Vashamov bound, Turan's theorem, Caro-Wei bound...

[8] Gilbert-Varshamov-like lower bounds for deletion-correcting codes, 2014 IEEE ITW

On the Size Distribution of Levenshtein Balls with Radius One

Geyang Wang* Qi Wang

Department of Computer Science and Engineering
Southern University of Science and Technology (SUSTech), China

March 10, 2022

Let $G(V,E)$ be a simple graph such that

- $V = \mathbb{Z}_m^n$.
- $(\boldsymbol{x},\boldsymbol{y}) \in E$ iff $d_L(\boldsymbol{x},\boldsymbol{y}) \le t$.

Therefore, for a (t,t)-deletion-insertion correcting code $C \subseteq \mathbb{Z}_m^n$.

- for every $\boldsymbol{x},\boldsymbol{y} \in C$, $d_L(\boldsymbol{x},\boldsymbol{y}) > t$.

$\Leftrightarrow$ C forms an independent set of G.

Caro-Wei bound (Yair Caro 1979, V.K.Wei. 1981):

$$\alpha(G) \ge \sum_{\boldsymbol{x}\in\mathbb{Z}_m^n} \frac{1}{1+\deg(\boldsymbol{x})}.$$

- $\alpha(G)$: the size of the largest independent set in G.
- $\deg(\boldsymbol{x}) = |L_t(\boldsymbol{x})|$.

Therefore, a lower bound of $|C|$ could be derived from $|L_t(\boldsymbol{x})|$ [Sala, Dolecek,[9]].

[9] Gilbert-Varshamov-like lower bounds for deletion-correcting codes, 2014 IEEE ITW