

Math 3500-101
Instructor: Dr. V. Rykov

End of Class Report

VARSHAMOV-GILBERT BOUNDS FOR PARTITION CODES

Jaime Lopez
University of Nebraska at Omaha
Undergraduate Program
E-mail: jalopez@mail.unomaha.edu

Adviser : Dr. Vyacheslav Rykov
Professor of Mathematics
University of Nebraska at Omaha
E-mail: vrykov@mail.unomaha.edu

1. Abstract

Partition codes characterize the partitioning of n elements into q partitions, with $2 \leq q < n$. The space generated by such codes is nonhomogeneous, resulting in the inability to define a closed formula for determining the volume of a sphere for various n and q . With the use of a Java program developed by the team, we calculate the Varshamov-Gilbert bounds as generated from the average sphere volume for various values of n and q .

2. Statement of the Problem

Partition codes have the potential to be applicable for many societal needs, most notably the diagnosis of mental impairments in youth. One method for diagnosing autism involves having the child divide a finite set of objects into partitions of their choosing. The psychiatrist administering the test compares the results against those from other children that have been deemed autistic. By using some type of distance measure between the child's results and the mean for autistic children, the psychiatrist can make a guess as to if the child also has autism.

Clearly, this method contains several areas that could result in a misdiagnosis. First, the method for evaluating the distance from one child's results to another's is, by enlarge, subjective, resulting in variant diagnosis among psychiatrists. Secondly,

the distance away from the autistic mean that still constitutes a positive diagnosis is not as quantifiably sound as it could be.

Developing the theory of partition codes would provide a way to quantifiably support a diagnosis that is neither haphazard nor subjective. By defining a distance metric and calculating the lower bound properties of various q-ary sequences of length n, we hope to help in the accurate and quantifiable diagnosis of such mental illnesses.

3. Definitions and Theorems

Definition 1: Let $n > q \geq 2$ be fixed integers, $A_q = \{0, 1, \dots, q-1\}$ be the standard q-ary alphabet, and $M_q = \{\mu\}$, $\mu = \mu(x)$, be the set of all $q!$ one-to-one permutations of A_q , i.e.

$$y = \mu(x), \quad x = \mu^{-1}(y), \quad x, y \in A_q = \{0, 1, \dots, q-1\}, \quad \mu, \mu^{-1} \in M_q. \quad [1]$$

Definition 2: Let $x = (x_1, x_2, \dots, x_n) \in A_{q \times n}$ be an arbitrary q-ary n-sequence which identifies an unordered q-partition $\bar{x} = \{E_0(x), E_1(x), \dots, E_{q-1}(x)\}$ of the set

$$[n] = E_0(x) + E_1(x) + \dots + E_{q-1}(x), \quad \text{where } E_x(x) = \{i: x_i = x\}, \quad x \in A_q. \quad [1]$$

Given these definitions, a distance metric between two arbitrary q-ary n-sequences, \bar{x} and \bar{y} , can be defined as follows:

Definition 3: The minimum distance between two arbitrary q-ary n-sequences \bar{x} and \bar{y} is described as:

$$d(\bar{x}, \bar{y}) = \min(H(\mu(\bar{x}), \bar{y})),$$

where $\mu(\bar{x})$ denotes any permutation of \bar{x} on A_q and $H(\bar{x}, \bar{y})$ denotes the standard Hamming distance between two q-ary numerical sequences of length n.

Example 1: Let $\bar{x} = (1, 1, 2, 1, 2, 3, 3, 2, 2, 1)$ and $\bar{y} = (1, 2, 3, 1, 1, 2, 2, 3, 1, 3)$. The distance between \bar{x} and \bar{y} would be 4 because by permuting \bar{y} , we can get $\bar{y} = (1, 2, 3, 1, 1, 2, 2, 3, 1, 3) = (1, 3, 2, 1, 1, 3, 3, 2, 1, 2)$, whose Hamming distance to \bar{x} is 4.

Definition 4: The Stirling number of the second kind, $S(n, k)$, is defined as the number of ways to partition a set of n elements into k nonempty subsets, and can be shown as:

$$S(n, q) = \frac{1}{q!} \sum_{i=0}^n (-1)^i \binom{q}{i} (q-i)^n$$

Definition 5: The Bell number, B_n , is defined as the number of partitions of a set

of n elements, and can be shown as the sum of Stirling numbers, in that:

$$B_n = \sum_{i=0}^n S(n,i)$$

The Bell number provides a method to calculate the sphere size of distance d around the origin, or in our case, the partition code with only one partition (i.e. 11111...1), via the following relationship:

$$V(n,q,d) = 1 + \sum_{k=1}^d \binom{n}{k} B_{q-1} = 1 + \sum_{k=1}^d \binom{n}{k} \sum_{i=1}^{q-1} S(k,i)$$

4. Methodology

Given the aforementioned distance metric, the task quickly became that of finding the total number of partitions of each distance, d , for all codewords in A_q , for various n and q . To accomplish this, we wrote a Java program that cycles over all representative partitions, calculates the distance between it and all other representative partitions, and maintains a count for all distance. We define a representative codeword to be one that can be represented in multiple ways. For example, 11112 and 11113 are the same partition, and, thus, should not both be used in calculations.

In an attempt to make speed a priority, an open source implementation of Hungarian Algorithm was integrated into the program for use with calculating the distance. Using this algorithm, most commonly used for solving assignment problems in an attempt to minimize “cost,” we used it to maximize the cost of permuting a partition. Subtracting the sum of the “costs” chosen by the algorithm from the length of either codeword (they should be equal) results in the minimum distance value.

Example 2: Given the same x' and y' as *Example 1*, a matrix containing the frequency of mapping each partition in x' to each partition in y' was generated.

x'/y'	1	2	3
1	2	1	1
2	2	0	2
3	0	2	0

The Hungarian Algorithm then takes this matrix and chooses the three assignments that maximize the “cost,” such that only one is chosen from each row and column, namely, $1 \rightarrow 1$, $3 \rightarrow 2$, and $2 \rightarrow 3$, resulting in the permutation of $\bar{y} = (1, 3, 2, 1, 1, 3, 3, 2, 1, 2)$. By subtracting the sum of the frequencies chosen (i.e. $2+2+2=6$) from the length of \bar{x} (or subsequently \bar{y}), we get the resulting value of $10-6=4$.

As mentioned, the distance value is calculated for every $x'-y'$ pair in A_q , and the frequency of each is collected. Upon the completion of the distance calculations, the average sphere size for each distance value is calculated as shown:

$$V_q^{avg} = \frac{f(d)}{\binom{q}{d} V(n,q,d)} \quad (1)$$

where $f(d)$ denotes the frequency of occurrence for distance d and $V(n, q, d)$ denotes the sphere size derived from the Bell number for a given n, q , and d .

Dividing by $\binom{q}{d}$ is necessary because there exist $\binom{q}{d}$ code words that are the d positions away from the origin (i.e. 111...11).

The results are shown in Table 1.

Table 1: The Average Sphere Sizes for Various n and q Values

n	4	4	4	6	6	6	8	8	9	10
q	2	3	4	2	3	4	2	4	3	2
d	Average Sphere Size									
0	0	0	0	0	0	0	0	0	0	0
1	3	6	5	13	67	95	56	671	1965	95
2		2	1		64	70		593	2127	70
3			0			38		727		38

Now that the average sphere sizes are calculated, Turan's Theorem is needed to help determine the Varshamov-Gilbert bound

Theorem 1 (Turan's Theorem): Let G be any simple graph on q^n vertices and e edges. G contains a subgraph of size M if:

$$\frac{q^{2n}}{2} \left(1 - \frac{1}{M-1}\right) < e$$

The goal is to find the lower bound for the size of $\tilde{C}_q(n, d)$. Therefore, we define an edge set, G , where an edge (x, y) exists in G iff $d(x, y) \geq d$. How many edges are contained in G can be found by taking spheres of radius $d-1$ around each vector then counting the number of vectors outside the particular sphere. However, since edges can be doubly counted, the value must be divided by two:

$$e = \frac{1}{2} \sum_{x \in F_q^n} (q^n - V_q(n, x, d-1)) = \frac{1}{2} (q^{2n} - q^n \frac{\sum_{x \in F_q^n} V_q(n, x, d-1)}{q^n}) = \frac{q^{2n}}{2} (q^n - V_q^{avg}(n, d-1))$$

Therefore, by Turan's Theorem, we know that if:

$$\frac{q^{2n}}{2} \left(1 - \frac{1}{M-1}\right) < \frac{q^{2n}}{2} (q^n - V_q^{avg}(n, d-1))$$

then a code of size at least M will exist, implying that:

$$M \leq |\tilde{C}_q(n, d)|$$

Taking $M = \left\lceil \frac{q^n}{V_q^{avg}(d-1)} \right\rceil$, it can easily be shown that

$$\frac{q^n}{V_q^{avg}(n, d-1)} \leq M < 1 + \frac{q^n}{V_q^{avg}(n, d-1)}$$

Therefore, by Turan's Theorem, we can make the following relationship:

$$\frac{q^n}{V_q^{avg}(d-1)} \leq |\tilde{C}_q(n, d)| \quad [2]$$

The significance of this relationship is such that we can use the average sphere size to calculate the Varshamov-Gilbert bound for each distance, d . The resulting calculations were performed on our average sphere sizes from Table 1, and are shown in Table 2.

Table 2: The Varshamov-Gilbert Bound for Various n and q Values

n	4	4	4	6	6	6	8	8	9	10
q	2	3	4	2	3	4	2	4	3	2
d	Varshamov-Gilbert Bound									
2	5.33	13.50	51.20	4.92	10.88	43.12	4.57	97.67	10.02	10.78
3		40.50	256.00		11.39	58.51		110.52	9.25	14.63
4						107.79		90.15		26.95

5. Conclusion

A Java program was developed that generates all codes for specified values of n and q , calculates the distance between every possible combination of such code words with the use of an implementation of the Hungarian Algorithm, calculates the average sphere size for each distance value, as well as the Varshamov-Gilbert bound values. This work will allow us to generate properly-sized codewords for various values of n and q , with the application being to more quantifiably support the diagnosis of mental illnesses.

6. Future Work

The program used for calculating the various values is fairly straight-forward. However, several improvements could be implemented to increase the performance.

- The implementation of the Hungarian Algorithm used is noted as being more applicable for smaller matrices (i.e. 2x2 and 3x3). It seems to have issues with calculating assignments given any matrix with dimension greater than 3. Approximately 1% of the distance calculations result in error from the algorithm, and are, thus, not included in the results. Obviously, this is not ideal. A new implementation of the algorithm that is capable of handling greater dimensions should be implemented.
- Obviously, the time needed to run the cases has an exponential behavior, with respects to both length and size (i.e. n and q). Time should be taken to run more cases with greater values for n and q .
-

7. References

- [1] A. D'yachkov, V.Rykov, D.Torney, S.Yekhanin, *Partition Codes*
- [2] V. Ufimtsev, *Generation of DNA Codes*, University of Nebraska at Omaha, 2005
- [3] V. Ufimtsev, *Discrete bounds on Partition Codes*