GILBERT-VARSHAMOV AND HAMMING BOUNDS IN FINITE METRIC SPACES WITH APPLICATIONS TO DNA CODES.

BRUNO JESUS AND CHRISTIAN LOMP

Abstract. We explain some formulae that appeared in bounds on DNA codes using elementary methods in metric spaces.

1. Finite metric spaces

A metric on a non-empty set X is a map $d : X \times X \to \mathbb{R}$ satisfying the following conditions for all $x, y, z \in X$:

(1) $(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$ for all $x, y \in X$.
(2) $d(x, y) = d(y, x)$.
(3) $d(x, z) \leq d(x, y) + d(y, z)$.

1.1. For any $0 \leq r \leq n$ and x we define the *sphere of radius r around x* as

$$S_r(x) = \{y \in X \mid d(x, y) = r\}.$$

For any $0 \leq r \leq n$ and x we define the *ball of radius r around x* as

$$B_r(x) = \{y \in X \mid d(x, y) \leq r\}.$$

Obviously one has

$$B_r(x) = \dot{\bigcup_{0\leq k\leq r}} S_k(x)$$

1.2. We say that the finite metric space (X, d) has *constant volume of balls of fixed radius* if for all r there exists a number $V(r)$ such that $|B_r(x)| = V(r)$ for all $x \in X$. We say that X has *constant area of spheres of fixed radius* if for all r there exists a number $W(r)$ such that $|S_r(x)| = W(r)$ for all $x \in X$.

Lemma. *Let (X, d) be a finite metric space, then X has constant volume of balls of fixed radius if and only if X has constant areas of spheres of fixed radius.*

Proof. Suppose that X has constant areas of spheres, then for all $x \in X$:

$$|B_r(x)| = \sum_{k=0}^{r} |S_k(x)| = \sum_{k=0}^{r} W(r) =: V(r)$$

is constant. On the other hand suppose that X has constant volume of balls of fixed radius. We use induction to show that X has constant area of spheres. For $r = 0$ and $x \in X$ we have $S_0(x) = B_0(x) = \{x\}$. Hence $W(0) = V(0) = 1$ is constant. Suppose that we showed that for $r \geq 0$ there exists a number $W(r)$ such that $|S_r(x)| = W(r)$ for all $x \in X$. For any $x \in X$ we have

$$V(r+1) = |B_{r+1}(x)| = |S_{r+1}(x)| + |B_r(x)| = |S_{r+1}(x)| + V(r).$$

Hence $|S_{r+1}(x)| = V(r+1) - V(r)$ is constant. $\square$

1.3. For any subset $C \subseteq X$ with $|C| \geq 2$ define its *minimal distance* as

$$\mathrm{dist}(C) = \min\{d(x,y) \mid x, y \in X \wedge x \neq y\}.$$

Lemma (Hamming bound)**.** *Let (X,d) be a finite metric space with constant volume of balls of fixed radius. Then for any subset $C \subseteq X$ with $|C| > 1$ and minimum distance* $\mathrm{dist}(C) = \delta$ *we have*

$$|C| \leq \frac{|X|}{V(e)},$$

where $e = \frac{\delta}{2} - 1$ if δ is even and $e = \frac{\delta-1}{2}$ if e is odd.

Proof. If δ is even, then we might assume $\delta = 2(e+1)$ for $e \geq 0$, if δ is odd, then we might assume $\delta = 2e+1$ for $e \geq 0$. In both cases for $x \neq y \in X : B_e(x) \cap B_e(y) = \emptyset$, because if $z \in B_e(x) \cap B_e(y)$ then

$$d(x,y) \leq d(x,z) + d(z,y) \leq 2 * e < \delta,$$

what is not possible if $x \neq y$. Thus

$$|C|V(e) = \left| \dot{\bigcup_{x \in C}} B_e(x) \right| \leq |X|.$$

$\square$

1.4. Let (X,d) be a finite metric space with constant volume of balls and denote by $W(k)$ the area of a sphere of radius k as in 1.2, then the Hamming bound can also be expressed as

$$|C| \leq \frac{|X|}{\sum_{k=0}^{e} W(k)}$$

for e as above.

1.5. Let (X,d) be a finite metric space. Define

$$A(X,\delta) = \max\{|C| \mid C \subseteq X \text{ with } \mathrm{dist}(C) \geq \delta \text{ and } |C| > 1\}$$

Lemma (Gilbert-Varshamov bound)**.** *Let (X,d) be a finite metric space with constant volume of balls of fixed radius and $\delta \geq 1$.*

$$A(X,\delta) \geq \frac{|X|}{V(\delta-1)}$$

Proof. Let $C \subseteq X$ with $|C| = A(X, \delta)$. Let $z \in X$ and assume that $\forall x \in C : d(x, z) \geq \delta$, then $C \cup \{z\}$. By maximality of C, $z \in C$. Hence for all $z \in X \setminus C : \exists x \in C$ such that $d(x, z) < \delta$, i.e. $z \in B_{\delta-1}(x)$. Thus

$$X = \bigcup_{x \in C} B_{\delta-1}(x)$$

and hence $|X| \leq |C| V(\delta - 1)$. □

1.6. Let (X, d) be a finite metric space with constant volume of balls and denote by $W(k)$ the area of a sphere of radius k as in 1.2, then the Gilbert Varshamov bound can also be expressed as

$$A(X, \delta) \geq \frac{|X|}{\sum_{k=0}^{\delta-1} W(k)}.$$

Combining the two bounds we have:

$$\frac{|X|}{\sum_{k=0}^{\delta-1} W(k)} \leq A(X, \delta) \leq \frac{|X|}{\sum_{k=0}^{e} W(k)}$$

2. Applications to Coding Theory

Let A be a non-empty finite set. We will refer to A as an alphabet and to its elements as letters. The cartesian product A^n is a metric space with the *Hamming metric* d, i.e.

$$d(x, y) = |\{i \in \{1, \ldots, n\} \mid x_i \neq y_i\}$$

for all $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in A^n$. Usually one refers to "vectors" $x \in A^n$ as "words" in A of length n.

2.1. The finite metric space (A^n, d) has constant volume on balls of fixed radius. To see this it is enough to show that the spheres of radius r have all the same area. Let $q = |A$, $0 \leq r \leq n$ and $x \in A^n$, then

$$W(r) = |S_r(x)| = \binom{n}{r} (q-1)^r$$

since a word $y \in A^n$ differs precisely in r positions from x and at each position there are $q - 1$ possibilites. If we want to specify the area of a sphere of radius r in the metric space A^n, we will write

$$W(A, n, r) = \binom{n}{r} |A| - 1^r.$$

Hence

$$V(r) = |B_r(x)| = \sum_{k=0}^{r} \binom{n}{k} (q-1)^k$$

is the volume of a ball of radius r.

2.2. A code in A of length n is any non-empty subset $C \subseteq A^n$.

We have recovered the orignial Hamming bound for codes C of length n and minimal distance δ over a q-ary alphabet as

$$|C| \leq \frac{q^n}{\sum_{k=0}^{e} \binom{n}{k}(q-1)^k}$$

where $e = \frac{\delta}{2} - 1$ if δ is even and $e = \frac{\delta-1}{2}$ if e is odd.

The Gilbert-Varshamov bound states

$$A(X,\delta) \geq \frac{q^n}{\sum_{k=0}^{\delta-1} \binom{n}{k}(q-1)^k}.$$

2.3. Let $B \subseteq A$ and denote its complement by $\overline{B} = A \setminus B$. For any $x \in A^n$ define the B-support as

$$B - \sup(x) = \{i \in \{1, \ldots, n\} \mid x_i \in B\}.$$

As we will see later, in some aplications not the whole metric space A^n is used, but some subspace of words of costant B-support for a given subset B. Hence for any $0 \leq w \leq n$ define a metric subspace of A^n as

$$X = X(B,n,w) = \{x \in A^n \mid |B - \sup(x)| = w\}.$$

We have $|X| = \binom{n}{w}|B|^w|\overline{B}|^{n-w}$.

Lemma. *Given a finite alphabet A with non-empty subset B such that $|B| = q$ and $|\overline{B}| = p$ and numbers $0 \leq w \leq n$ any sphere of radius δ in the finite metric space $X(B,n,w)$ has constant area*

$$W(\delta) = \sum_{k=0}^{M(r)} \binom{w}{k}\binom{n-w}{k}(qp)^k \sum_{i=0}^{N(r,k)} W(B, w-k, i) W(\overline{B}, n-w-k, r-2k-i)$$

with $M(r) = \min(\frac{r}{2}, w, n-w)$ and $N(r,k) = \min(r-2k, w-k)$

3. Aplications to codes of constant weight

Let A be an alphabet of q elements with a distinguished element 0. In most cases A will be a group, a ring or a finite field and 0 will be its neutral element or zero element. The weight $w(x)$ of an element $x \in A^n$ is defined to be the number of indices i where $x_i \neq 0$. In other words $w(x) = B - \sup(x)$. for $B = A \setminus \{0\}$. The words of length n in A of constant weight w are precisely the words in $X = X(A \setminus \{0\}, n, w)$.

3.1. Note that $\overline{B} = \{0\}$ consists only of 1 element. The metric space $\overline{B}^n$ is degenerated and consists also just of one element. Hence $W(\overline{B}, n, r) \neq 0$ if and only if $r = 0$.

3.2. Acording to 2.3 we have

Lemma. *The spheres of fixed radius r in $X = X(A \setminus \{0\}, n, w)$ have area*

$$W(r) = \sum_{k=r-w}^{\min(\frac{\delta}{2},w,n-w)} \binom{w}{k}\binom{n-w}{k}\binom{n-w-k}{r-2k}(q-1)^k(q-2)^{r-2k}$$

where $q = |A|$.

Proof. Looking at the formula in 2.3 and as mentioned above $W(\overline{B}, n-w, r-2k-i) \neq 0$ if and only if $r-2k = i$. Hence if we want that the terms contribute something to the area of a sphere, i must "reach" $r-2k$, ie. $r-2k \leq w-k \Leftrightarrow r-w \leq k$. Applying 2.3 we have:

$$\begin{aligned} W(r) &= \sum_{k=r-w}^{\min(\frac{\delta}{2},w,n-w)} \binom{w}{k}\binom{n-w}{k}(q-1)^k W(B, w-k, r-2k) \\ &= \sum_{k=r-w}^{\min(\frac{\delta}{2},w,n-w)} \binom{w}{k}\binom{n-w}{k}\binom{w-k}{r-2k}(q-1)^k(q-2)^{r-2k} \end{aligned}$$

□

3.3. The Hamming bound for q-ary codes C of length n with constant weight w and minimu distance δ says now that

$$|C| \leq \frac{\binom{n}{w}(q-1)^w}{\sum_{r=0}^{e}\sum_{k=r-w}^{\min(\frac{\delta}{2},w,n-w)} \binom{w}{k}\binom{n-w}{k}\binom{n-w-k}{r-2k}(q-1)^k(q-2)^{r-2k}}$$

where e is as above.

Denote by $A^{cnst}(q, n, \delta, w)$ the cardinality of a maximal q-ary code of length n of constant weight w and minimum distance δ, then the Gibert-Varshamov bound says:

$$A^{cnst}(q, n, \delta, w) \geq \frac{\binom{n}{w}(q-1)^w}{\sum_{r=0}^{\delta-1}\sum_{k=r-w}^{\min(\frac{\delta}{2},w,n-w)} \binom{w}{k}\binom{n-w}{k}\binom{n-w-k}{r-2k}(q-1)^k(q-2)^{r-2k}}.$$

4. Applications to DNA codes

Let A be an alphabet and $B \subseteq A$ a subset with $2|B| = |A|$ and $|B| = p$. Then we have for the volume of balls for the code $X = X(B, n, w)$:

Lemma. *The volume of a ball of radius δ in X is:*

$$V(\delta) = \sum_{r=0}^{\delta} \sum_{k=0}^{\min(\frac{r}{2},w,n-w)} \binom{w}{k}\binom{n-w}{k}\binom{n-2k}{r-2k}p^{2k}(p-1)^{r-2k}.$$

Proof.

$$\begin{aligned} V(\delta) &= \sum_{r=0}^{\delta}\sum_{k=0}^{M(r)}\binom{w}{k}\binom{n-w}{k}p^{2k}\sum_{i=0}^{N(r,k)}\binom{w-k}{i}\binom{n-w-k}{r-2k-i}(p-1)^i(p-1)^{r-2k-i} \\ &= \sum_{r=0}^{\delta}\sum_{k=0}^{M(r)}\binom{w}{k}\binom{n-w}{k}p^{2k}(p-1)^{r-2k}\sum_{i=0}^{N(r,k)}\binom{w-k}{i}\binom{n-w-k}{r-2k-i} \\ &= \sum_{r=0}^{\delta}\sum_{k=0}^{\min(\frac{r}{2},w,n-w)}\binom{w}{k}\binom{n-w}{k}\binom{n-2k}{r-2k}p^{2k}(p-1)^{r-2k} \end{aligned}$$

where we use the Vandermonde convolution

$$\sum_k \binom{r}{k}\binom{s}{d-k} = \binom{r+s}{d}.$$

□

4.1. DNA codes are codes over the 4-letter alphabet $A = \{A, T, G, C\}$. In [1] the GC-content of a word x was defined as $B - \sup(x)$ where $B = \{G, C\}$ and limits for DNA codes with constant GC-content w had been established. Let $X = X(B, n, w)$. Then $|X| = \binom{n}{w}2^n$. Denote by $A^{GC}(n, \delta, w)$ the number of words in maximal DNA code of length n, minimum distance δ and GC-content w. Here we find those limits by the above general observations:

Theorem. *The volume of a ball of fixed radius δ in the space of DNA-codes of words of length n and constant GC-content w is*

$$V(\delta) = \sum_{r=0}^{\delta}\sum_{k=0}^{\min(\frac{r}{2},w,n-w)}\binom{w}{k}\binom{n-w}{k}\binom{n-2k}{r-2k}4^k$$

Hence:

$$A^{GC}(n, \delta, w) \leq \frac{\binom{n}{w}2^n}{\sum_{r=0}^{\frac{\delta}{2}-1}\sum_{k=0}^{\min(\frac{r}{2},w,n-w)}\binom{w}{k}\binom{n-w}{k}\binom{n-2k}{r-2k}4^k}.$$

$$A^{GC}(n, \delta, w) \geq \frac{\binom{n}{w}2^n}{\sum_{r=0}^{\delta-1}\sum_{k=0}^{\min(\frac{r}{2},w,n-w)}\binom{w}{k}\binom{n-w}{k}\binom{n-2k}{r-2k}4^k}.$$

References

[1] O.King, *Bounds for DNA codes with constant GC-content.*, Electron. J. Comb. 10, Research paper R33, 13 p. (2003)