ON THE EXPECTATION AND VARIANCE OF HAMMING DISTANCE BETWEEN TWO I.I.D RANDOM VECTORS*

FU FANGWEI (符方伟) SHEN SHIYI (沈世镒)

(*Department of Mathematics, Nankai University, Tianjin* 300071, *China*)

Abstract

By using the generalized MacWilliams theorem, we give new representations for expectation and variance of Hamming distance between two i.i.d random vectors. By using the new representations, we derive a lower bound for the variance, and present a simple and direct proof of the inequality of [1].

Key words. Hamming distance, random vector, expectation, variance, generalized MacWilliams theorem

1. Introduction

Let $F_2^n = \{0,1\}^n$ be an n-dimensional vector space over the binary field $F_2 = \{0,1\}$. The Hamming distance between two vectors $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ is the number of coordinates where they differ, and is denoted by $d_H(x,y)$,

$$d_H(x,y) = \sum_{i=1}^{n} |x_i - y_i|.$$

The Hamming weight of x is the number of non-zero coordinates, and is denoted by $w_H(x)$. Obviously $w_H(x) = d_H(x, \mathbf{0})$, where $\mathbf{0}$ is the zero vector.

The scalar product of x and y is

$$\langle x, y \rangle = x_1y_1 + \cdots + x_ny_n \qquad \text{in } F_2.$$

For a set $A \subseteq F_2^n$, $|A|$ denotes the cardinality of A. The average distance in A is defined by

$$\operatorname{dist}(A) = \frac{1}{|A|^2} \sum_{x\in A} \sum_{y \in A} d_H(x,y). \tag{1.1}$$

Received July 31, 1993.

* This research is supported by Young Teacher Foundation of Chinese Educational Ministry and Natural Science Foundation of China.

The variance of dist (A) is defined by

$$\operatorname{var}(A) = \frac{1}{|A|^2} \sum_{x\in A} \sum_{y\in A} \left[d_H(x,y) - \operatorname{dist}(A)\right]^2. \tag{1.2}$$

Althöfer and Sillke[1] proved

Theorem 1.1. Every non-empty set $A \subseteq F_2^n$ satisfies the inequality

$$\operatorname{dist}(A) \geq \frac{n+1}{2} - \frac{2^{n-1}}{|A|}, \tag{1.3}$$

where equality is possible only for $|A| = 2^n$ and for $|A| = 2^{n-1}$ with A being a subcube.

This inequality yields only negative values as lower bounds for $|A| < \frac{2^n}{n+1}$. Therefore it is only meaningful for large subsets.

In this paper, we derive the following inequality for var (A).

Theorem 1.2. Every non-empty set $A \subseteq F_2^n$ satisfies the inequality

$$\operatorname{var}(A) \geq \frac{n-1}{4} + \frac{2^{n-1}}{|A|} - \frac{2^{2n-2}}{|A|^2}, \tag{1.4}$$

where equality holds for $|A| = 2^n$ and for $|A| = 2^{n-1}$ with A being a subcube.

If $A = F_2^n$, we have

$$\frac{n+1}{2} - \frac{2^{n-1}}{|F_2^n|} = \frac{n+1}{2} - \frac{1}{2} = \frac{n}{2},$$
$$\frac{n-1}{4} + \frac{2^{n-1}}{|F_2^n|} - \frac{2^{2n-2}}{|F_2^n|^2} = \frac{n-1}{4} + \frac{1}{2} - \frac{1}{4} = \frac{n}{4}.$$

For a given $x \in F_2^n$,

$$\begin{aligned}\sum_{y\in F_2^n} d_H(x,y) &= \sum_{i=0}^{n} \sum_{y\in F_2^n : d_H(x,y)=i} d_H(x,y)\\ &= \sum_{i=0}^{n} i \left|\{y \in F_2^n : d_H(x,y) = i\}\right| = \sum_{i=0}^{n} i\binom{n}{i} = n2^{n-1},\\ \sum_{y\in F_2^n} \left[d_H(x,y)\right]^2 &= \sum_{i=0}^{n} i^2\binom{n}{i} = n(n+1)2^{n-2}.\end{aligned}$$

Hence

$$\begin{aligned}\operatorname{dist}(F_2^n) &= \frac{1}{2^{2n}} \sum_{x\in F_2^n} \sum_{y\in F_2^n} d_H(x,y) = \frac{n2^{n-1}2^n}{2^{2n}} = \frac{n}{2},\\ \operatorname{var}(F_2^n) &= \frac{1}{2^{2n}} \sum_{x\in F_2^n} \sum_{y\in F_2^n} \left[d_H(x,y)\right]^2 - \left[\operatorname{dist}(F_2^n)\right]^2 = \frac{2^n n(n+1)2^{n-2}}{2^{2n}} - \frac{n^2}{4} = \frac{n}{4}.\end{aligned}$$

Therefore the lower bounds in Theorem 1.1 and Theorem 1.2 are tight for $A = F_2^n$.

If A is a subcube, $A = F_2^{n-1} \times \{0\} = \left\{(x,0) \mid x \in F_2^{n-1}\right\}$, we have

$$\frac{n+1}{2} - \frac{2^{n-1}}{|F_2^{n-1} \times \{0\}|} = \frac{n-1}{2},$$

$$\frac{n-1}{4}+\frac{2^{n-1}}{|F_2^{n-1}\times\{0\}|}-\frac{2^{2n-2}}{|F_2^{n-1}\times\{0\}|^2}=\frac{n-1}{4}.$$

In the same way, we have

$$\operatorname{dist}\left(F_2^{n-1}\times\{0\}\right)=\frac{n-1}{2},\qquad \operatorname{var}\left(F_2^{n-1}\times\{0\}\right)=\frac{n-1}{4}.$$

Therefore the lower bounds in Theorem 1.1 and Theorem 1.2 are tight for $|A|=2^{n-1}$ with A being a subcube.

Let X,Y be two independent identical distributed (i.i.d) random vectors. The common probability distribution is $P=\{P(x)\,|\,x\in F_2^n\}$. The expectation of $d_H(X,Y)$ is

$$E\,d_H(X,Y)=\sum_{x\in F_2^n}\sum_{y\in F_2^n}P(x)P(y)\,d_H(x,y). \tag{1.5}$$

The variance of $d_H(X,Y)$ is

$$\begin{aligned} D\,d_H(X,Y)=&E\big[d_H(X,Y)\big]^2-\big[Ed_H(X,Y)\big]^2 &(1.6)\\ =&\sum_{x\in F_2^n}\sum_{y\in F_2^n}P(x)P(y)\left[d_H(x,y)-Ed_H(X,Y)\right]^2. &(1.7)\end{aligned}$$

Denote

$$L(P)=2^{n-1}\sum_{x\in F_2^n}\left[P(x)-\frac{1}{2^n}\right]^2. \tag{1.8}$$

$L(P)$ measures how unequally P is distributed.

Althöfer and Sillke[1] proved

Theorem 1.3. $\quad Ed_H(X,Y)\geq\frac{n}{2}-L(P).\qquad$ (1.9)

We derive a lower bound for $Dd_H(X,Y)$.

Theorem 1.3. $\quad Dd_H(X,Y)\geq\frac{n}{4}-\left[L(P)\right]^2.\qquad$ (1.10)

For a set $A\subseteq F_2^n$, let X_A,Y_A be two i.i.d random vectors with common distribution

$$P_A(x)=\begin{cases}\frac{1}{|A|}, & \text{if } x\in A,\\ 0, & \text{otherwise.}\end{cases}$$

It is easy to see that

$$Ed_H(X_A,Y_A)=\operatorname{dist}(A),\qquad Dd_H(X_A,Y_A)=\operatorname{var}(A),$$
$$L(P_A)=2^{n-1}\left[|A|\left(\frac{1}{|A|}-\frac{1}{2^n}\right)^2+(2^n-|A|)\left(\frac{1}{2^n}\right)^2\right]=\frac{2^{n-1}}{|A|}-\frac{1}{2}.$$

Then

$$\frac{n}{2}-L(P_A)=\frac{n+1}{2}-\frac{2^{n-1}}{|A|},\qquad \frac{n}{4}-\left[L(P_A)\right]^2=\frac{n-1}{4}+\frac{2^{n-1}}{|A|}-\frac{2^{2n-2}}{|A|^2}.$$

Therefore Theorem 1.1 and Theorem 1.2 could be derived from Theorem 1.3 and Theorem 1.4 respectively in a straightforward way.

Althöfer and Sillke proved Theorem 1.3 by using induction method. In this paper we first introduce the generalized MacWilliams theorem, then we give new representations for $Ed_H(X,Y)$ and $Dd_H(X,Y)$. Finally we derive Theorem 1.4, and present a simple and direct proof of Theorem 1.3 by using the new representations.

2. Generalized MacWilliams Theorem

$\mathcal{R}$ is the field of real numbers. $f: F_2^n \longrightarrow \mathcal{R}$ is a function. Denote

$$M = \sum_{u \in F_2^n} f(u) \neq 0, \tag{2.1}$$

$$B_i = \sum_{u \in F_2^n : w_H(u) = i} f(u), \qquad i = 0, 1, \cdots, n. \tag{2.2}$$

$\{B_0, B_1, \cdots, B_n\}$ is called the weight distribution of f. The weight enumerator of f is defined by

$$W_f(z) = \sum_{u \in F_2^n} f(u) z^{w_H(u)} = \sum_{i=0}^{n} B_i z^i. \tag{2.3}$$

The Hadamard transform of f is

$$\overline{f}(u) = \frac{1}{M} \sum_{v \in F_2^n} (-1)^{\langle u,v \rangle} f(v), \qquad u \in F_2^n. \tag{2.4}$$

$\overline{f}$ is also a function from F_2^n to $\mathcal{R}$. Denote

$$\overline{B}_i = \sum_{u \in F_2^n : w_H(u) = i} \overline{f}(u), \qquad i = 0, 1, \cdots, n. \tag{2.5}$$

$\{\overline{B}_0, \overline{B}_1, \cdots, \overline{B}_n\}$ is the weight distribution of $\overline{f}$. The weight enumerator of $\overline{f}$ is

$$W_{\overline{f}}(z) = \sum_{i=0}^{n} \overline{B}_i z^i. \tag{2.6}$$

Theorem 2.1. (Generalized MacWilliams Theorem)

$$W_{\overline{f}}(z) = \frac{1}{M}(1+z)^n W_f\left(\frac{1-z}{1+z}\right), \tag{2.7}$$

$$W_f(z) = \frac{M}{2^n}(1+z)^n W_{\overline{f}}\left(\frac{1-z}{1+z}\right). \tag{2.8}$$

The proof of the generalized MacWilliams theorem could be found in [2] (pp. 132–137).

3. Several Lemmas

$P = \{P(u) \mid u \in F_2^n\}$ is a probability distribution on F_2^n. Function $f_P : F_2^n \longrightarrow \mathcal{R}$ is defined by

$$f_P(u) = \sum_{a,b \in F_2^n :\, a+b=u} P(a)P(b). \tag{3.1}$$

Obviously,

$$M_P = \sum_{u \in F_2^n} f_P(u) = 1.$$

The weight distribution of f_P is $\{B_0(P), B_1(P), \cdots, B_n(P)\}$.

$$B_i(P) = \sum_{u \in F_2^n:\, w_H(u)=i} \; \sum_{a,b \in F_2^n:\, a+b=u} P(a)P(b). \tag{3.2}$$

The weight enumerator of f_P is $W_{f_P}(z)$.

$$W_{f_P}(z) = \sum_{i=0}^{n} B_i(P) z^i. \tag{3.3}$$

The Hadamard transform of f_P is $\overline{f}_P$.

$$\overline{f}_P(u) = \sum_{v \in F_2^n} (-1)^{\langle u,v\rangle} f_P(v) \tag{3.4}$$

$$= \sum_{v \in F_2^n} (-1)^{\langle u,v\rangle} \sum_{a,b \in F_2^n:\, a+b=v} P(a)P(b) \tag{3.5}$$

$$= \sum_{a,b \in F_2^n} (-1)^{\langle u,a+b\rangle} P(a)P(b) \tag{3.6}$$

$$= \Big[\sum_{a \in F_2^n} (-1)^{\langle u,a\rangle} P(a) \Big]^2 \geq 0. \tag{3.7}$$

The weight distribution of $\overline{f}_P$ is $\{\overline{B}_0(P), \overline{B}_1(P), \cdots, \overline{B}_n(P)\}$.

$$\overline{B}_i(P) = \sum_{u \in F_2^n:\, w_H(u)=i} \overline{f}_P(u) \tag{3.8}$$

$$= \sum_{u \in F_2^n:\, w_H(u)=i} \Big[\sum_{a \in F_2^n} (-1)^{\langle u,a\rangle} P(a) \Big]^2 \geq 0. \tag{3.9}$$

The weight enumerator of $\overline{f}_P$ is

$$W_{\overline{f}_P}(z) = \sum_{i=0}^{n} \overline{B}_i(P) z^i. \tag{3.10}$$

From the generalized MacWilliams theorem, we have

$$W_{f_P}(z) = \frac{1}{2^n}(1+z)^n W_{\overline{f}_P}\left(\frac{1-z}{1+z}\right). \tag{3.11}$$

Lemma 3.1. ([2], p.134, Problem (13))

$$\sum_{v \in F_2^n} (-1)^{\langle u,v\rangle} = \begin{cases} 2^n, & \text{if } u = \mathbf{0}, \\ 0, & \text{if } u \neq \mathbf{0}. \end{cases} \tag{3.12}$$

Lemma 3.2.
$$Ed_H(X,Y)=\sum_{i=0}^{n} iB_i(P), \tag{3.13}$$
$$E\big[d_H(X,Y)\big]^2=\sum_{i=0}^{n} i^2B_i(P). \tag{3.14}$$

Proof.
$$\begin{aligned}Ed_H(X,Y)&=\sum_{a\in F_2^n}\sum_{b\in F_2^n}P(a)P(b)d_H(a,b)=\sum_{a\in F_2^n}\sum_{b\in F_2^n}P(a)P(b)w_H(a+b)\\&=\sum_{u\in F_2^n}w_H(u)\sum_{a,b\in F_2^n:\,a+b=u}P(a)P(b)=\sum_{u\in F_2^n}w_H(u)f_P(u)\\&=\sum_{i=0}^{n}\sum_{u\in F_2^n:\,w_H(u)=i}w_H(u)f_P(u)=\sum_{i=0}^{n}iB_i(P).\end{aligned}$$

(3.14) could be proved in the same way.

4. A Direct and Simple Proof of Theorem 1.3

Lemma 4.1.
$$Ed_H(X,Y)=\frac{n}{2}-\frac{\overline{B}_1(P)}{2}. \tag{4.1}$$

Proof. From (3.3), we know that the differentiation of $W_{f_P}(z)$ is

$$W'_{f_P}(z)=\sum_{i=0}^{n} i\,B_i(P)z^{i-1}. \tag{4.2}$$

From (4.2) and Lemma 3.2, we have

$$W'_{f_P}(1)=\sum_{i=0}^{n} i\,B_i(P)=Ed_H(X,Y). \tag{4.3}$$

From (3.11), we have

$$W'_{f_P}(z)=\frac{1}{2^n}\left[n(1+z)^{n-1}W_{\overline{f}_P}\left(\frac{1-z}{1+z}\right)-2(1+z)^{n-2}W'_{\overline{f}_P}\left(\frac{1-z}{1+z}\right)\right]. \tag{4.4}$$

Then
$$W'_{f_P}(1)=\frac{1}{2^n}\left[n2^{n-1}W_{\overline{f}_P}(0)-2^{n-1}W'_{\overline{f}_P}(0)\right]. \tag{4.5}$$

From (3,9), (3,10), we have

$$W_{\overline{f}_P}(0)=\overline{B}_0(P)=1,\qquad W'_{\overline{f}_P}(0)=\overline{B}_1(P). \tag{4.6}$$

From (4.5), (4.6), we have
$$W'_{f_P}(1)=\frac{n}{2}-\frac{\overline{B}_1(P)}{2}. \tag{4.7}$$

Therefore from (4.3) and (4.7), we have

$$Ed_H(X,Y)=\frac{n}{2}-\frac{\overline{B}_1(P)}{2}.$$

For a given $u \in F_2^n$, $w_H(u) \geq 2$, we denote

$$G_u^0 = \{a \in F_2^n \mid \langle u, a\rangle = 0\}, \qquad G_u^1 = \{a \in F_2^n \mid \langle u, a\rangle = 1\}.$$

Lemma 4.2.
$$\overline{B}_1(P) \leq 2L(P); \tag{4.8}$$
equality holds only when

$$P(G_u^0) = P(G_u^1) \qquad \text{for every } u \in F_2^n,\ w_H(u) \geq 2. \tag{4.9}$$

Proof. From (3.9), we have

$$\begin{aligned}
\overline{B}_1(P) &= \sum_{u\in F_2^n:\, w_H(u)=1} \Big[\sum_{a\in F_2^n} (-1)^{\langle u,a\rangle} P(a)\Big]^2 \\
&= \sum_{u\in F_2^n} \Big[\sum_{a\in F_2^n} (-1)^{\langle u,a\rangle} P(a)\Big]^2 - 1 - \sum_{u\in F_2^n:\, w_H(u)\geq 2} \Big[\sum_{a\in F_2^n} (-1)^{\langle u,a\rangle} P(a)\Big]^2 \\
&\leq -1 + \sum_{u\in F_2^n} \Big[\sum_{a\in F_2^n} (-1)^{\langle u,a\rangle} P(a)\Big]^2 = -1 + \sum_{u\in F_2^n} \sum_{a,b\in F_2^n} (-1)^{\langle u,a+b\rangle} P(a)P(b) \\
&= -1 + \sum_{a,b\in F_2^n} P(a)P(b) \sum_{u\in F_2^n} (-1)^{\langle u,a+b\rangle}.
\end{aligned}$$

From Lemma 3.1, we have

$$\sum_{u\in F_2^n} (-1)^{\langle u,a+b\rangle} = \begin{cases} 2^n, & \text{if } a = b, \\ 0, & \text{if } a \neq b. \end{cases}$$

Then

$$\overline{B}_1(P) \leq -1 + 2^n \sum_{a\in F_2^n} P^2(a) = 2^n \sum_{a\in F_2^n} \Big[P(a) - \frac{1}{2^n}\Big]^2 = 2L(P);$$

equality holds only when $\forall\, u \in F_2^n$, $w_H(u) \geq 2$

$$\sum_{a\in F_2^n} (-1)^{\langle u,a\rangle} P(a) = P(G_u^0) - P(G_u^1) = 0.$$

Proof of Theorem 1.3 from Lemma 4.1 and Lemma 4.2.
From Lemma 4.1 and Lemma 4.2, we have

$$Ed_H(X,Y) = \frac{n}{2} - \frac{\overline{B}_1(P)}{2} \geq \frac{n}{2} - L(P);$$

equality holds only when (4.9) is true.

5. Proof of Theorem 1.4

Lemma 5.1.
$$E\big[d_H(X,Y)\big]^2 = \frac{n(n+1)}{4} - \frac{n}{2}\overline{B}_1(P) + \frac{\overline{B}_2(P)}{2}.$$

Proof. From (4.2), we have

$$W''_{f_P}(z) = \sum_{i=0}^{n} i\,(i-1)B_i(P)z^{i-2}.$$

Then from Lemma 3.2, we have

$$\begin{aligned} W''_{f_P}(1) =& \sum_{i=0}^{n} i\,(i-1)B_i(P) = \sum_{i=0}^{n} i^2 B_i(P) - \sum_{i=0}^{n} iB_i(P) \\ =& E\big[d_H(X,Y)\big]^2 - Ed_H(X,Y). \end{aligned} \tag{5.1}$$

From (4.4), we have

$$\begin{aligned} W''_{f_P}(z) = \frac{1}{2^n}\Big[& n(n-1)(1+z)^{n-2}W_{\overline{f}_P}\Big(\frac{1-z}{1+z}\Big) - 4(n-1)(1+z)^{n-3}W'_{\overline{f}_P}\Big(\frac{1-z}{1+z}\Big) \\ & + 4(1+z)^{n-4}W''_{\overline{f}_P}\Big(\frac{1-z}{1+z}\Big)\Big]. \end{aligned}$$

Hence

$$\begin{aligned} W''_{f_P}(1) =& \frac{1}{2^n}\big[n(n-1)2^{n-2}W_{\overline{f}_P}(0) - (n-1)2^{n-1}W'_{\overline{f}_P}(0) + 2^{n-2}W''_{\overline{f}_P}(0)\big] \\ =& \frac{n(n-1)}{4}\overline{B}_0(P) - \frac{n-1}{2}\overline{B}_1(P) + \frac{1}{2}\overline{B}_2(P). \end{aligned} \tag{5.2}$$

From (5.1), (5.2), (4.1), we have

$$\begin{aligned} E\big[d_H(X,Y)\big]^2 =& \frac{n(n-1)}{4} - \frac{n-1}{2}\overline{B}_1(P) + \frac{1}{2}\overline{B}_2(P) + \frac{n}{2} - \frac{1}{2}\overline{B}_1(P) \\ =& \frac{n(n+1)}{4} - \frac{n}{2}\overline{B}_1(P) + \frac{1}{2}\overline{B}_2(P). \end{aligned}$$

Proof of Theorem 1.4 from Lemma 5.1 and Lemma 4.2.

$$\begin{aligned} Dd_H(X,Y) =& E\big[d_H(X,Y)\big]^2 - \big[Ed_H(X,Y)\big]^2 \\ =& \frac{n(n+1)}{4} - \frac{n}{2}\overline{B}_1(P) + \frac{1}{2}\overline{B}_2(P) - \Big[\frac{n}{2} - \frac{1}{2}\overline{B}_1(P)\Big]^2 \\ =& \frac{n}{4} - \frac{1}{4}\big[\overline{B}_1(P)\big]^2 + \frac{1}{2}\overline{B}_2(P). \end{aligned}$$

From (3.9), (4.8), we have

$$\overline{B}_i(P) \ge 0, \qquad i = 0,1,\cdots,n; \qquad \overline{B}_1(P) \le 2L(P).$$

Therefore

$$Dd_H(X,Y) \ge \frac{n}{4} - \big[L(P)\big]^2.$$

References

[1] I. Althöfer and T. Sillke. An "Average Distance" Inequality for Large Subsets of the Cube. *Journal of Combinatorical Theory* (Series B), 1992, 56: 296–301.

[2] F.J. MacWilliams and N.J.A. Sloane. The Theory of Error-Correcting Codes. North-Holland, Amsterdam, 1981 (Third printing).