Distance Spectrum Formula for the Largest Minimum Hamming Distance of Finite-Length Binary Block Codes

Ling-Hua Chang*, Carol Wang†, Po-Ning Chen*, Yunghsiang S. Han‡ and Vincent Y. F. Tan†
†Dept. of Elec. and Computer Eng. & Dept. of Math., National Univ. of Singapore, Singapore
‡School of Electrical Engineering & Intelligentization, Dongguan University of Technology, Dongguan, China

Abstract—In this paper, an exact distance spectrum formula for the largest minimum Hamming distance of finite-length binary block codes is presented. The exact formula indicates that the largest minimum distance of finite-length block codes can be fully characterized by the information spectrum of the Hamming distance between two independent and identically distributed (i.i.d.) random codewords. The distance property of finite-length block codes is then connected to the distance spectrum. A side result of this work is a new lower bound to the largest minimum distance of finite-length block codes. Numerical examinations show that the new lower bound improves the finite-length Gilbert-Varshamov lower bound and can reach the minimum distance of existing finite-length block codes.

I. INTRODUCTION

In 2000, a general distance spectrum formula for the asymptotic largest minimum distance (in blocklength) of deterministic block codes under generalized distance measures (not necessarily additive, symmetric and bounded) was established [1]. A limitation of this formula is that it is asymptotic in nature, and only holds as the code block length tends to infinity. Identification of an information-spectrum formula of the largest minimum distance attainable for a block code of a given block length is still unresolved.

In this work, we revisit the same problem formulation in [1] but now tackle the largest minimum distance of finite-length block codes. For convenience in exposition, Hamming distance and binary block codes are assumed. Unlike the expurgated random-coding technique used in [1], a different approach in proving the distance spectrum formula of finite-length block codes has been taken. Specifically, the largest code size attainable, subject to a target largest minimum distance, is first derived. The largest minimum distance spectrum formula for finite-length block codes is subsequently obtained by duality. As it turns out, the resultant formula for codes with finite block length has a similar form to the asymptotic one for codes with block length growing to infinity in [1]. Yet, the cumbersome exponent in the asymptotic distance spectrum formula of finite-length block codes is considered. This is anticipated to facilitate the future characterization of the optimal code distribution that achieves the largest minimum distance of finite-length block codes.

The rest of the paper is organized as follows. The distance-spectrum formula for the largest minimum distance of finite-length block codes is derived in Section II, followed by the presentation of a new lower bound to this exact distance-spectrum formula. Also given in this section are examples of the distance-spectrum formula and the new lower bound, as well as their comparison with the finite-length Gilbert-Varshamov lower bound. Final comments appear in Section III.

II. DISTANCE SPECTRUM FORMULAS OFFINITE-LENGTH BLOCK CODES

A. Largest Code Size under a Fixed Minimum Distance

Some notations are introduced before presenting the main results. The n-tuple code alphabet is denoted by $\mathbb{X}^n = \{0, 1\}^n$. For any two elements $\hat{x}^n$ and $x^n$ in $\mathbb{X}^n$, we use $\mu(\hat{x}^n, x^n)$ to denote the Hamming distance of these two elements. An $(n, M)$-code with block length $n$ and size $M$ is understood as a codebook

$$\mathcal{C} \triangleq \{x^n_1, x^n_2, x^n_3, \ldots, x^n_M\},$$

where $x^n_m \triangleq (x^n_{m,1}, x^n_{m,2}, \ldots, x^n_{m,n}) \in \mathbb{X}^n$ [2]. As an extension, an $(n, M, d)$-code denotes a block code of length $n$ and size $M$ with the minimum pairwise distance equal to $d$, i.e.,

$$d = \min_{1 \leq i < j \leq M} \mu(x^n_i, x^n_j).$$

Definition 1: Define the cumulative distance function corresponding to $n$-fold Hamming distance measure $\mu(\cdot, \cdot)$ and $n$-fold distribution (i.e., distribution on $\mathbb{X}^n$) $P_{X^n}$ as

$$F_{X^n}(d) \triangleq \Pr[\mu(\hat{X}^n, X^n) < d],$$

where in this definition (and also throughout the paper), $\hat{X}^n$ and $X^n$ are used to denote two independent $n$-fold random variables with common distribution $P_{X^n}$.

Based on this definition, we first show in the next theorem that the maximal number of codewords subject to a specified minimum pairwise distance can be related to the cumulative distance function.

This work was supported by Minister of Sci. and Tech. of Taiwan under Grants MOST 105-2221-E-009-009-MY3 and MOST 105-2221-E-009-011-MY3. Han’s work was supported by Natl Natural Science Foundation of China (Grant No. 61671007).
Theorem 1: The largest number of codewords \( M_n^*(d) \) that can achieve a target (integer) minimum pairwise distance \( 0 < d \leq n \) for codes of block length \( n \) is given by

\[
M_n^*(d) = \sup_{X^n} \frac{1}{F_{X^n}(d)}
\]

\[
= \sup_{X^n} \frac{1}{\Pr[\|X^n, X^n\| < d]}.
\]

Proof: We defer the proof to Appendix A for better readability.

It may be interesting to point out that the reciprocal of the cumulative distribution function \( F_{X^n}(d) \) is often not an integer for general \( X^n \). However, it can be inferred from the proof of Theorem 1 that an optimal \( X^n \) that achieves \( \inf_{X^n} F_{X^n}(d) \) is uniformly distributed over a support of size \( M = M_n^*(d) \) with all pairwise distances no less than \( d \). Therefore,

\[
F_{X^n}(d) = \sum_{i=1}^{M} \left( \frac{1}{M} \right) \left( \frac{1}{M} \right) = \frac{1}{M} = \frac{1}{M_n^*(d)}
\]

and the reciprocal of \( F_{X^n}(d) \) does equal the integer-valued \( M_n^*(d) \). This indicates that determination of \( M_n^*(d) \) via the cumulative distance function is somehow equivalent to finding the optimal code. Despite the challenge of determining the optimizer, Theorem 1 indeed provides a useful tool to lower-bound \( M_n^*(d) \).

Corollary 1: The largest number of codewords \( M_n^*(d) \) that can achieve a target (integer) minimum pairwise distance \( 0 < d \leq n \) for codes of block length \( n \) satisfies

\[
M_n^*(d) \geq M_{n,d}(d) \triangleq \frac{1}{F_{n,d}(d)}
\]

for an \( n \)-fold distribution \( P_{X^n} \).

B. Largest Minimum Distance under a Fixed Code Size

Based on Theorem 1, we proceed to derive the distance spectrum formula for the largest minimum distance of finite-length \( (n, M) \) block codes by duality.

Theorem 2: The largest minimum pairwise distance \( d_1^*(M) \) for an \( (n, M) \) block code with \( 2 \leq M \leq 2^n \) must satisfy

\[
d_1^*(M) = \max \left\{ a \in \mathbb{N}_n : M_n^*(a) \geq M \right\}
\]

\[
= \max \left\{ a \in \mathbb{N}_n : \inf_{X^n} F_{X^n}(a) \leq \frac{1}{M} \right\},
\]

where \( \mathbb{N}_n \triangleq \{1, 2, ..., n\} \).

Unlike the optimization over all infinite-dimensional distributions \( X = \{X^n\} \), Theorems 1 and 2 only rely on the finite-dimensional optimization for \( \inf_{X^n} F_{X^n}(d) \). However, the optimization of \( F_{X^n}(d) \) over all possible finite-dimensional distributions is still challenging. Based on some preliminary attempts to find the optimal \( X^n \) for small \( n \) as well as the proof of Theorem 1, we have the following two observations:

1) The proof of Theorem 1 suggests that the optimizer of \( \inf_{X^n} F_{X^n}(d) \) should be a uniform distribution over certain supports.

2) Extensive numerical examination hints that the set of optimizers for \( \inf_{X^n} F_{X^n}(d) \) may always include one with all one-dimensional marginal distributions being uniform, i.e., for every \( 1 \leq i \leq n \) and every \( x_i \in X \),

\[
P_{X_i}(x_i) = \sum_{(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \in X^{n-1}} P_{X^n}(x^n)
\]

\[
= \frac{1}{2}.
\]

As an example, with \( n = d = 3 \), an optimal distribution \( P_{X^3} \) that achieves

\[
M_n^*(3) = \frac{1}{F_{X^3}(3)} = 2
\]

is given by

\[
P_{X^3}(x^3) = \frac{1}{2} \quad \text{for} \quad x^3 \in \{011, 100\}.
\]

It can be easily verified that

\[
P_{X_1}(0) = P_{X_1}(1) = \frac{1}{2} \quad \text{for} \quad i = 1, 2, 3.
\]

Similarly, with \( n = 3 \) and \( d = 2 \), an optimal distribution \( P_{X^3} \) that achieves

\[
M_n^*(2) = \frac{1}{F_{X^3}(2)} = 4
\]

satisfies

\[
P_{X^3}(x^3) = \frac{1}{4} \quad \text{for} \quad x^3 \in \{001, 010, 100, 111\},
\]

which again verifies

\[
P_{X_i}(0) = P_{X_i}(1) = \frac{1}{2} \quad \text{for} \quad i = 1, 2, 3.
\]

Again, an immediate consequence of Theorem 2 is that any distribution \( P_{X^n} \) provides a lower bound to \( d_n^*(M) \). We summarize this observation in the next corollary.

Corollary 2 (Cumulative distance function bound): For \( 2 \leq M \leq 2^n \),

\[
d_n^*(M) \geq d_X^*(M) \triangleq \sup \left\{ a \in \mathbb{R} : F_{X^n}(a) \leq \frac{1}{M} \right\}.
\]

Based on this corollary, we provide lower bounds of \( d_n^*(M) \) for \( n = 6, 7, 8, 11 \) and \( M = 2^k \) for integer \( 1 \leq k < n \), and compare them to the finite-length Gilbert-Varshamov lower bounds [3]. The \( n \)-fold distributions are chosen to be uniformly distributed over a \( k \)-dimensional subspace with \( k \) length-\( n \) vectors as its basis, which satisfies that all pairwise distances among these \( k \) basis vectors are no less than \( d \). The best known block codes listed in [4] (specifically, [6], [7], [8] and [9]) are also given as benchmarks. The results are illustrated in
Figure 1. We observe from this figure that the new lower bound improves the finite-length Gilbert-Varshamov lower bound and approaches the minimum distance of the best known block codes.

Notably, placing uniform probability masses on codewords of any existing \((n, M, d_{\text{min}})\)-code \(\mathcal{C}\) can provide a straightforward lower bound to \(M_n^*(d)\) as

\[
M_n^*(d) \geq \sum_{\hat{x}^n \in \mathcal{C}} \sum_{x^n \in \mathcal{C}} 1\{\mu(\hat{x}^n, x^n) < d\} = \frac{M^2}{M + \sum_{\hat{x}^n \in \mathcal{C}} \sum_{x^n \in \mathcal{C}, x^n \neq \hat{x}^n} 1\{\mu(\hat{x}^n, x^n) < d\}}
\]

where \(1\{\cdot\}\) is the set indicator function. However, the resulting lower bound is often loose since

\[
\sum_{\hat{x}^n \in \mathcal{C}} \sum_{x^n \in \mathcal{C}, x^n \neq \hat{x}^n} 1\{\mu(\hat{x}^n, x^n) < d\} > 0
\]

except when \(1 \leq d \leq d_{\text{min}}\). Particularly, under \(1 \leq d \leq d_{\text{min}}\), we always have \(M_n^*(d) \geq M\).

C. Asymptotic Extensions to Normalized Minimum Distances and Code Rates

In this subsection, we extend the formulas in Theorems 1 and 2 to a form that is amenable to asymptotic study. In this regard, the normalized minimum distance \(\delta = d/n\) for an \((n, M, d)\)-code under a specified code rate \(R = (1/n)\log(M)\) is considered.

**Theorem 3:** The largest normalized minimum pairwise distance \(\delta_n^*(R)\) for an \((n, M)\)-code with rate \(R = (1/n)\log(M)\) satisfies

\[
\delta_n^*(R) = \max \left\{ a \in \mathbb{R}_n : \sup_{X^n} \left( -\frac{1}{n} \log \Pr \left\{ \frac{1}{n} \mu(\hat{X}^n, X^n) < a \right\} \right) \geq R \right\}, \tag{2}
\]

where \(\mathbb{R}_n = \{ 1/n, 2/n, \ldots, (n-1)/n, 1 \}\).

**Proof:** Theorem 1 indicates that the largest code rate attainable for an \((n, M)\)-code, subject to a target normalized minimum pairwise distance \(\delta = d/n\), is equal to

\[
R_n^*(\delta) = \frac{1}{n} \log M_n^*(n\delta) = \sup_{X^n} \left( -\frac{1}{n} \log \Pr \left\{ \frac{1}{n} \mu(\hat{X}^n, X^n) < \delta \right\} \right), \tag{3}
\]

which is exactly the largest exponent of the probability that the normalized \(n\)-fold distance between independent \(\hat{X}^n\) and \(X^n\) is less than \(\delta\). The theorem is proven by duality to (3) as

\[
\delta_n^*(R) = \max \{ a \in \mathbb{R}_n : R_n^*(a) \geq R \}. \tag{4}
\]

**Remark.** The formula in (2) provides a different form to the general asymptotic distance spectrum formula of the largest minimum distance of block codes from the ones in [1], which we recapitulate here to facilitate subsequent comparison:

\[
\sup_{X} \bar{\Lambda}_X(R) \triangleq \sup_{X} \inf \left\{ a \in \mathbb{R} : \limsup_{n \to \infty} \left( \Pr \left\{ \frac{1}{n} \mu(\hat{X}^n, X^n) > a \right\} \right)^M = 0 \right\} \tag{5}
\]

and

\[
\sup_{X} \Delta_X(R) \triangleq \sup_{X} \inf \left\{ a \in \mathbb{R} : \liminf_{n \to \infty} \left( \Pr \left\{ \frac{1}{n} \mu(\hat{X}^n, X^n) > a \right\} \right)^M = 0 \right\} \tag{6}
\]

where in (5) and (6), \(M = e^{nR}\) is the exponentially growing code size, and \(\mathbb{R}\) is the set of real numbers.\(^2\) We note that the new form in (2) not only removes the cumbersome exponent of \(M\) in (5) and (6), but also performs the task of taking the supremum over all asymptotic distributions \(X = \{X^n\}_{n=1}^{\infty}\) before the functional optimization of parameter \(a\) over the real line. This reduces the set of all distributions to be examined from being of infinite dimension to being of finite dimension, and hence may help the numerical characterization of the optimal distribution that achieves the largest minimum distance of block codes, particularly when \(n\) is small. In particular, with the exact formula for every \(n\), we also exclude the necessity of sandwiching the quantity of interest in terms of (5) and (6), and the mathematical peculiarity, such as equality holding except for a countable number of points, is thus removed.

III. Conclusion

In this paper, we showed that the maximal size of finite-length block codes, satisfying a specified pairwise minimum distance constraint, can be characterized by the cumulative distance function of two independent and identically distributed codewords. Based on this result, we obtained a new formula for the asymptotic largest minimum distance of block codes as a function of code rate, which is anticipated to be useful in the determination of the optimal code distribution for each block length \(n\). A side result is a better lower bound than the finite-length Gilbert-Varshamov lower bound, as confirmed for small \(n\). A future work of theoretical significance is to construct a sequence of \(n\)-fold distributions that can lead to an asymptotic bound (in block length \(n\)) for the largest minimum distance of block codes and to compare it with the asymptotic Gilbert-Varshamov lower bound.

\(^2\)By denoting the largest minimum distance of an \((n, M)\)-code as \(d_{n,M}\), it was shown in [1, Thm. 1] that given \(R = \lim_{n \to \infty} (1/n) \log(M)\),

\[
\limsup_{n \to \infty} \frac{d_{n,M}}{n} = \sup_{X} \bar{\Lambda}_X(R) \quad \text{and} \quad \liminf_{n \to \infty} \frac{d_{n,M}}{n} = \sup_{X} \Delta_X(R)
\]

except possibly at the (countably many) points of discontinuities of \(\sup_{X} \bar{\Lambda}_X(R)\) and \(\sup_{X} \Delta_X(R)\).
Fig. 1: Cumulative distance function (CDF) bounds and finite-length Gilbert-Varshamov lower bound. The best known existing codes listed in [4] are marked in red diamonds, among which two codes of block length \( n = 11 \) have their code sizes not equal to powers of two.

REFERENCES


APPENDIX A

PROOF OF THEOREM 1

In order to simplify the notations, we use \( \overline{x} \) instead of \( x^n \) to denote the \( n \)-tuple in \( X^n \) in the proof. The same abbreviation will be applied to other notations.

The proof will be done in two steps: 1. \( M^*_n(d) \leq 1/|\inf_X F_X(d)| \) and 2. \( M^*_n(d) \geq 1/|\inf_X F_X(d)| \), based on which we can conclude that \( M^*_n(d) = 1/|\inf_X F_X(d)| \).
1. $M_n^*(d) \leq 1/\left[ \inf_X F_X(d) \right]$. 

For an $(n, M, d)$-code $C \triangleq \{x_1, x_2, \ldots, x_M\}$ with $M = M_n^*(d)$, let $P_Z$ be the $n$-fold distribution that places probability mass $1/M$ on each codeword of $C$. Then,

$$F_Z(d) = \Pr(\mu(Z, Z) < d) = \sum_{i=1}^M P_Z(x_i) P_Z(x_i) = \sum_{i=1}^M \left( \frac{1}{M} \right) \left( \frac{1}{M} \right) = \frac{1}{M} = \frac{1}{M_n^*(d)}.$$ 

This immediately implies

$$\inf_X F_X(d) \leq F_Z(d) = \frac{1}{M_n^*(d)}.$$ 

2. $M_n^*(d) \geq 1/\left[ \inf_X F_X(d) \right]$. 

Suppose that the smallest support of an $n$-fold distribution $P_Z$ consists of $\ell$ distinct elements, denoted as

$$\text{Supp}(Z) = \{x_1, x_2, \ldots, x_\ell\},$$

where $P_Z(x_i) = p_i > 0$ for every $1 \leq i \leq \ell$. We then claim that if $\ell \geq M_n^*(d)$, there exists another $n$-fold distribution $P_W$, whose support contains only $M_n^*(d)$ elements, such that

$$F_W(d) \leq F_Z(d).$$

As a result of this claim, $\inf_X F_X(d)$ can be reduced to

$$\inf_X F_X(d) = \inf_{\hat{X}:|\text{Supp}(\hat{X})| \leq M_n^*(d)} F_X(d),$$

where $|\cdot|$ denotes the size of a set. We then derive for any distribution $\hat{X}$ with $\text{Supp}(\hat{X}) = \ell \leq M_n^*(d)$,

$$F_{\hat{X}}(d) \geq \Pr[\hat{X} = \hat{X}] = \sum_{i=1}^\ell P_{\hat{X}}(x_i) P_{\hat{X}}(x_i) = \sum_{i=1}^\ell p_i^2 \geq \frac{1}{\ell} \geq \frac{1}{M_n^*(d)},$$

where (8) follows from Cauchy-Schwarz inequality, and (9) is valid due to the assumption that $\ell \leq M_n^*(d)$. We thus conclude

$$\inf_X F_X(d) = \inf_{\hat{X}:|\text{Supp}(\hat{X})| \leq M_n^*(d)} F_{\hat{X}}(d) \geq \frac{1}{M_n^*(d)}.$$ 

It remains to substantiate the claim, which can be proved through a Huffman-like coding procedure as follows. Among the support $\text{Supp}(\hat{Z}) = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_\ell\}$ of size $\ell > M_n^*(d)$, there must exist two distinct elements with Hamming distance smaller than $d$. Without loss of generality, we let this pair be $\hat{x}_1$ and $\hat{x}_2$, and denote for $1 \leq k \leq 2$,

$$q_k = \sum_{j=3}^\ell p_j \{\mu(\hat{x}_k, \hat{x}_j) < d\},$$

where $P_{\hat{Z}}(\hat{x}_i) = p_i > 0$ for $1 \leq i \leq \ell$. Assume without loss of generality that $q_1 \geq q_2$. Construct $P_W(\cdot)$ as

$$P_W(\cdot) = \hat{p}_1 = \begin{cases} 0, & i = 1; \\ p_1 + p_2, & i = 2; \\ p_i, & 3 \leq i \leq \ell. \end{cases}$$

Note that the support of $P_W(\cdot)$ consists of $(\ell - 1)$ elements, i.e.,

$$\text{Supp}(P_W(\cdot)) = \{\hat{x}_2, \hat{x}_3, \ldots, \hat{x}_\ell\}.$$

Then, we derive

$$F_W(d) - F_{W(\cdot)}(d) \leq \sum_{i=1}^\ell \sum_{j=1}^\ell p_i p_j \{\mu(\hat{x}_i, \hat{x}_j) < d\} - \sum_{i=1}^\ell \sum_{j=1}^\ell \hat{p}_i \hat{p}_j \{\mu(\hat{x}_i, \hat{x}_j) < d\}$$

$$= \sum_{i=1}^\ell \sum_{j=1}^\ell (p_i p_j - \hat{p}_i \hat{p}_j) \{\mu(\hat{x}_i, \hat{x}_j) < d\} = \sum_{i=1}^\ell \sum_{j=1}^\ell (p_i p_j + \hat{p}_i \hat{p}_j - \hat{p}_i p_j - \hat{p}_j p_i) \{\mu(\hat{x}_i, \hat{x}_j) < d\} \geq 0,$$

where (10) holds because

$$\sum_{i=1}^\ell \sum_{j=1}^\ell p_i \hat{p}_j \{\mu(\hat{x}_i, \hat{x}_j) < d\} = \sum_{i=1}^\ell \sum_{j=1}^\ell \hat{p}_i \hat{p}_j \{\mu(\hat{x}_i, \hat{x}_j) < d\},$$

and the last inequality follows from the assumption that $q_1 \geq q_2$. If $|\text{Supp}(W(\cdot))| = \ell - 1 > M_n^*(d)$, we can similarly construct $P_W(\cdot)$ with $|\text{Supp}(W(\cdot))| = \ell - 2$ and $F_W(\cdot) \leq F_{W(\cdot)}(d)$. By repeating such a construction $(\ell - M_n^*(d))$ times, we can obtain the desired $P_W = P_W(\ell - M_n^*(d))$, with $|\text{Supp}(W(\cdot))| = \ell - (\ell - M_n^*(d)) = M_n^*(d)$ and

$$F_W(d) \leq F_Z(d).$$