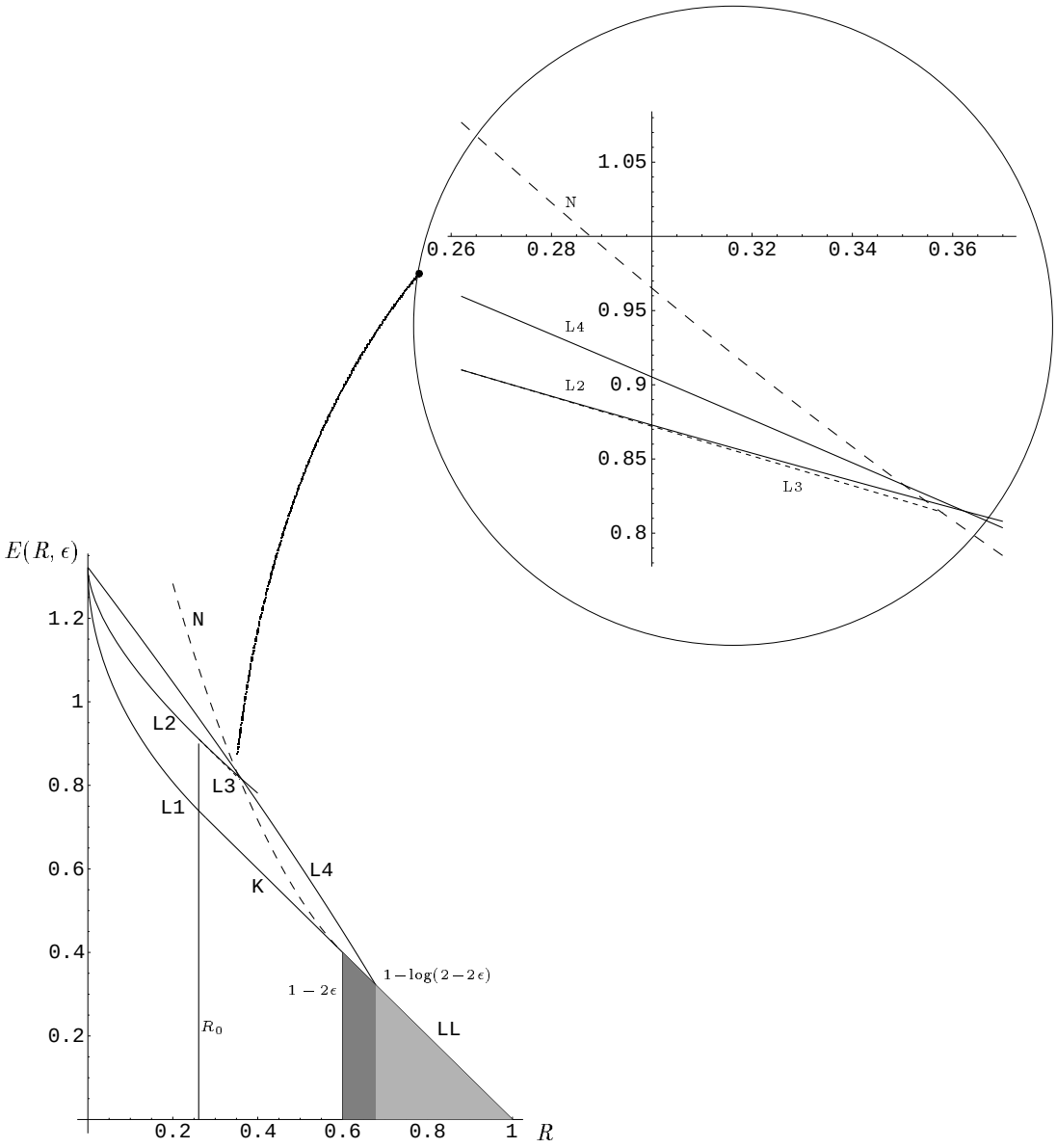


Figure 1: Bounds on $E(R,\epsilon)$, $\epsilon = 0.2$. The shaded areas show the segment on which $E(R,\epsilon) = 1 - R$. Straight-line bound L3 (dotted) is still slightly better than the other bounds. Bound L4 is improved everywhere.

[7] T. Helleseth, T. Kløve, and J. Mykkeltveit, "The weight distribution of irreducible cyclic codes with block lengths $n_1((q^l-1)/N)$," *Discrete Math.*, **18** (1977), 179-211.

[8] G. L. Katsman and M. A. Tsfasman, "Spectra of algebraic geometric codes," *Problemy Peredachi Informatsii*, **23** (4) (1988), 19–34.

[9] G. L. Katsman, M. A. Tsfasman, and S. G. Vladuts, "Spectra of linear codes and error probability of decoding," in H. Stichtenoth and M. A. Tsfasman, Eds., *Coding Theory and Algebraic Geometry*, Lect. Notes Math., vol. **1518**, Berlin: Springer (1992), pp. 82–98.

[10] T. Kløve and V. I. Korzhik, *Error Detecting Codes*, Kluwer (1995).

[11] T. Kløve, "Bounds on the weight distribution of cosets," *IEEE Trans. Inform. Theory*, **42** (6) (1996), 2257–2260.

[12] V. I. Korzhik, "Bounds on the probability of undetected error and optimum group codes in a channel with feedback," *Radiotechnika*, **20** (1) (1965), 27–33. English translation in *Telecommun. Radio Eng.*, **20** (1, pt.2) (1965), 87–92.

[13] V. K. Leontiev, "Error-detecting encoding," *Problemy Peredachi Informatsii*, **8** (3) (1972), 6–14

[14] V. I. Levenshtein, "Bounds on the probability of undetected error," *Problemy Peredachi Informatsii*, **13** (1) (1977), 3–18.

[15] V. I. Levenshtein, "Straight-line bound for the undetected error exponent," *Problemy Peredachi Informatsii*, **25** (1) (1989), 33–37.

[16] F. J. MacWilliams, "A theorem on the distribution of weights in a systematic codes," *Bell Syst. Techn. Journal*, **42** (1963), 79–94.

[17] J. Simonis, "The effective length of subcodes," *Applicable Algebra in Engineering, Communication and Computing*, **5** (1994), 371–377.

[18] D. J. A. Welsh, *Complexity: knots, colourings and counting*, London Math. Society Lecture Note Series, **186**, Cambridge Univ. Press, Cambridge (1993).

[19] J. K. Wolf, A. M. Michelson, and A. H. Levesque, " On the probability of undetected error for linear block codes," *IEEE Trans. Commun*, **30** (1982), 317–324.

Bound (K) is due to Korzhik [12]. Bounds (L1)-(L4) were proved by Levenshtein [15], [14]. Bound (LL) is due to Leontiev [13] and Levenshtein [15].

Some remarks may be in order. First, by (K) and (LL) the function $E(R,\epsilon)$ is known exactly for $\log(2-2\epsilon) \le R \le 1$. Next, the "minimum-distance" bound (L2) is an analog of the minimum-distance bound of Shannon, Gallager and Berlekamp for the probability of decoding error. The straight-line bound (L3) [15] is a segment with slope -1 that can be drawn from any point of the minimum-distance bound. The point R_0 is chosen for this segment to be tangent to (L2) and therefore is optimal.

These bounds should be compared to the bound of Theorem 12. First, this theorem shows that the interval on which the function $E(R,\epsilon)$ is known exactly (and equals $1-R$) is extended to $1-2\epsilon \le R \le 1$. Next, bound (N) from this theorem is better than the bound (LL) for a certain interval of rates.

Bounds from both theorems are drawn in Fig. 1, where for $\rho(R)$ we have chosen the upper bound

$$\rho(R) = \frac{1}{2} - \sqrt{H^{-1}(R)(1-H^{-1}(R))}, \tag{19}$$

which is the best presently known for small code rates. The straight-line bound (L3) still is the best in a certain interval of rates $R > R_0(\epsilon)$. Finally, the best upper bound currently known is given in the following proposition.

Proposition 14 *Let $\rho(R)$ be defined by (19) and let R_0 be the root of $\rho'(R) = (\log(\epsilon/(1-\epsilon))^{-1}$.*

$$E(R,\epsilon) \le \begin{cases} D(\rho(R),\epsilon), & 0 \le R \le R_0, \\ \min\{D(\rho(R_0),\epsilon) + R_0 - R, & \\ \qquad - H(R) - (1-R)\log\epsilon - R\log(1-2\epsilon)\}, & R_0 \le R \le 1-2\epsilon, \end{cases}$$

$$E(R,\epsilon) = 1-R, \qquad 1-2\epsilon \le R \le 1.$$

We conclude with the following quotation from [15]: "Surprisingly, the question whether the bound $E(R,\epsilon) \le 1-R$ is tight for all rates R, $1-H(\epsilon) \le R \le 1$, is still open."

Acknowledgment. We are grateful to Vladimir Levenshtein and H.F. Mattson, Jr., for the careful reading of an earlier draft of this paper.

References

[1] K.A.S. Abdel-Ghaffar, "A lower bound on the undetected error probability and strictly optimal codes," *IEEE Trans. Inform. Theory*, **43** (5) (1997), 1489–1502.

[2] A. Barg, "The matroid of supports of a linear code," *Applicable Algebra in Engineering, Communication and Computing*, **8** (1997), 165–172.

[3] T. Brylawski and J. G. Oxley, "The Tutte polynomial and its applications," in N. White, Ed., in *Matroid Applications*, Encyclopedia of Mathematics and Its Applications, **40**, Cambridge Univ. Press (1992).

[4] N. J. Fine, *Hypergeometric Series and Applications*, AMS (1988).

[5] C. Greene, "Weight enumeration and the geometry of linear codes," *Stud. Appl. Math.*, **55** (1976), 119–128.

[6] T. Helleseth, T. Kløve, and V.I. Levenshtein, "On the information function of an error-correcting code," *IEEE Trans. Inform. Theory*, **43** (2) (1997), 549–557.

for the probability of undetected error. Let $n \to \infty$, $M = 2^{Rn}$, $0 \le R \le 1$. All logarithms below are base 2. The exponent of the probability of undetected error is defined as

$$E(n, R, \epsilon) = \Big(-\frac{1}{n} \log P(2, n, \lceil 2^{Rn} \rceil, \epsilon) \Big),$$
$$E(R, \epsilon) = \lim_{n \to \infty} E(n, R, \epsilon)$$

provided that this limit exists. We are interested in lower bounds on $\liminf_{n\to\infty} E(n, R, \epsilon)$ and upper bounds on $\limsup_{n\to\infty} E(n, R, \epsilon)$. By abuse of notation, in both cases below we write $E(R, \epsilon)$.

The following theorem is an easy corollary of the bound (18).

Theorem 12

$$E(R, \epsilon) \le \begin{cases} -H(R) - (1-R)\log\epsilon - R\log(1-2\epsilon), & 0 \le R \le 1-2\epsilon, \quad (N) \\ 1 - R, & 1-2\epsilon \le R \le 1. \end{cases}$$

Proof: Let us write out the generic term in the sum in (18) in the exponential form, putting $w = n\omega$. We get in the exponent

$$-n[-H(\omega) - (\omega - 1 + R) - \omega\log\epsilon - (1-\omega)\log(1-2\epsilon)].$$

The minimum of this is attained for $\omega = 2\epsilon$. Hence this expression gives an upper bound on $E(R, \epsilon)$ if the corresponding term is within the summation range, or if $2\epsilon > 1 - R$. Substituting $\omega = 2\epsilon$, we get $1 - R$ in the exponent. If $2\epsilon \le 1 - R$, the maximum in the sum is attained for $\omega = 1 - R$. ∎

Let us quote the known results to which this result should be compared (see [15] for a survey). Let

$$D(u, v) = -u\log v - (1-u)\log(1-v),$$
$$H(u) = -u\log u - (1-u)\log(1-u).$$

Theorem 13 *Let* $\delta(R) = H^{-1}(1-R)$. *Then*

$$E(R, \epsilon) \ge \begin{cases} D(\delta(R), \epsilon), & 0 \le R \le 1 - H(\epsilon) \quad (L1) \\ 1 - R, & 1 - H(\epsilon) \le R \le 1. \quad (K) \end{cases}$$

Let $\rho(R)$ *be any upper bound on the distance of codes of rate* R *such that* $\rho''(R) \ge 0$ *and let* $R_0 = R_0(\epsilon)$ *be defined by* $\rho'(R_0) = (\log(\epsilon/(1-\epsilon)))^{-1}$. *Let* $\tau(R) = -R/\log(2^{1-R} - 1)$. *Then*

$$E(R, \epsilon) \le \begin{cases} D(\rho(R), \epsilon), & 0 \le R \le R_0, \quad (L2) \\ D(\rho(R_0), \epsilon) + R_0 - R, & R_0 \le R \le 1, \quad (L3) \\ D(\tau(R), \epsilon) - R, & 0 \le R \le \log(2-2\epsilon), \quad (L4) \\ 1 - R, & \log(2-2\epsilon) \le R \le 1. \quad (LL) \end{cases}$$

It is given by the following sum:

$$\sum_{j=1}^{w} \sum_{c \in C} \sum_{E \in \binom{N}{w}} \sum_{\substack{c' \in C \\ \mathsf{supp}(c,c') \subseteq E \\ \mathsf{dist}(c,c')=j}} 1.$$

This sum is equal to $|C|$ times the left-hand side of the claimed equality. Interchanging the first and the third sums, we observe that it is also equal to the right-hand side. ∎

To use this lemma in the expression (8), we need to estimate the sum on the right-hand side. We get the following analog of Lemma 2.

Lemma 10 *Suppose that* $\mathsf{supp}\,(C) = n$. *Let* $E \in \binom{N}{w}$. *Then*

$$\sum_{i=1}^{s(E)} |\mathcal{C}_i(E)|(|\mathcal{C}_i(E)| - 1) \geq \max\{0, |C|(q^{w-n}|C| - 1)\}.$$

Proof: As remarked above, $\sum_i |\mathcal{C}_i(E)| = |C|$. If $|C| \leq q^{n-w}$, estimate each $|\mathcal{C}_i(E)|$ by 1. For $|C| > q^{n-w}$, we have

$$\sum_i |\mathcal{C}_i(E)|(|\mathcal{C}_i(E)| - 1) = \sum_i |\mathcal{C}_i(E)|^2 - |C|.$$

Using Lagrange multipliers, we see that the sum $\sum |\mathcal{C}_i(E)|^2$ attains the minimum when all numbers $|\mathcal{C}_i(E)|$ are the same (and equal $\frac{|C|}{s(E)} \geq |C|q^{w-n}$). Hence, we get

$$\sum_i |\mathcal{C}_i(E)|(|\mathcal{C}_i(E)| - 1) \geq |C|^2 q^{w-n} - |C|.$$

∎

These two lemmas again lead to the bound of the form (4). We quote this result for future use.

Proposition 11 *Let* $P(q, n, M, \epsilon)$ *be defined by* (3). *Then*

$$P(q, n, M, \epsilon) \geq \sum_{w=n-k+1}^{n} \binom{n}{w} (Mq^{w-n} - 1) \left(\frac{\epsilon}{q-1}\right)^w \left(1 - \frac{q\epsilon}{q-1}\right)^{n-w}. \tag{18}$$

The bound in [1] is a little sharper than this because the estimate in Lemma 10 can be improved by using integer parts.

Though it may not be clear at the first glance, linearity is irrelevant to Proposition 7 and the ensuing discussion as well. Its general analog is proved upon defining the average rank function of any code C. We believe that the details can be left to an interested reader.

4 Asymptotics

In this section we confine ourselves to binary codes. Let us analyze the asymptotic case. Our source for previous results is [15], which gives a comprehensive overview of the asymptotic bounds

Thus, by (15) and (10), we have expressed $P(C,\epsilon)$ in terms of support weight distributions of C (and by (17), also of $C^\perp$).

3 The general case

Let us consider unrestricted codes. Our proofs of properties of binomial moments (Lemma 2 and Lemma 3) depend on linearity. However, binomial moments of the distance distribution of a nonlinear code still retain the properties claimed in these lemmas, and the linearity turns out to be irrelevant!

We need two definitions.

Definition. Let $\mathcal{A} \subseteq F_q^n$ be a set of vectors. Define the support of $\mathcal{A}$ as follows:

$$\mathsf{supp}\,\mathcal{A} = \{e \in N : \exists_{c,c' \in \mathcal{A}}(c_e \neq c'_e)\}.$$

Definition. Let $C \subseteq F_q^n$ be a code and $E \subseteq N$. By $\mathcal{C}(E)$ we denote a subcode of C such that (a), $\mathsf{supp}\,\mathcal{C}(E) \subseteq E$, and (b), $\forall c \in C, c \notin \mathcal{C}(E) : \mathsf{supp}\,(\mathcal{C}(E) \cup c) \not\subseteq E$.

Thus, $\mathcal{C}(E)$ are maximal subcodes of C with support in E.

Example. Let C be a code of 6 words:

$$c_1 = 000$$
$$c_2 = 101$$
$$c_3 = 111$$
$$c_4 = 010$$
$$c_5 = 100$$
$$c_6 = 110,$$

and let $E = \{1\}$. Then there are 4 subcodes $\mathcal{C}(E) = \mathcal{C}_i(E)$, namely $(c_1, c_5), (c_2), (c_3), (c_4, c_6)$.

Generally for a given subset E, $|E| = w$, there are at most q^{n-w} subcodes $\mathcal{C}_i(E)$, and $\sum_i \mathcal{C}_i(E) = |C|$. If C is linear, we get back to our original definition of $\mathcal{C}(E)$.

For the rest of this section we assume that $\mathsf{supp}\,C = N$. Denote by $s(E)$ the number of subcodes $\mathcal{C}_i(E)$.

Let us prove an analog of Lemma 3, which is essentially Lemma 1 in [1], in different notation.

Lemma 9 *Let A_w be as defined in* (1). *Then*

$$B_w := \sum_{j=1}^{w} \binom{n-j}{n-w} A_j = (1/|C|) \sum_{E \in \binom{N}{w}} \sum_{i=1}^{s(E)} |\mathcal{C}_i(E)|(|\mathcal{C}_i(E)| - 1).$$

Proof: Count in two ways the size of the set

$$\left\{(E,(c,c')) \mid E \in \binom{N}{w};\ c,c' \in C,\ \mathsf{supp}\,(c,c') \subseteq E, 1 \leq \mathsf{dist}\,(c,c') \leq w\right\}.$$

with it). These functions play an important role in combinatorics of incidence geometries; see [3],[18]. More on this link in the linear case is found in [5],[2],[6][3]. In particular, [6] puts forward the problem of estimating the numbers $U^C_{w\alpha}$. By Proposition 7, a solution would yield new estimates of $P(C,\epsilon)$.

The *Whitney function*, by definition, is a polynomial $R_C(x,y) = \sum_{u=0}^{n}\sum_{v=0}^{k} R^v_u x^u y^v$, *where* $R^v_u := \left|\left\{ E \subseteq N \ : \ k - \mathsf{rank}\, G(E) = u, |E| - \mathsf{rank}\, G(E) = v \right\}\right|$. The *Tutte polynomial* $T_C(x,y)$ can be formally defined as $T_C(x,y) := R_C(x-1, y-1)$. The following proposition is a variation of Greene's result [5].

Proposition 8

$$P(C,\epsilon) = \frac{\epsilon^{n-k}(q-1-q\epsilon)^k}{(q-1)^n} T_C\Big(\frac{q-1}{q-1-q\epsilon}, \frac{(q-1)(1-\epsilon)}{\epsilon}\Big) - (1-\epsilon)^n.$$

Proof: Let $A(z) := A(1,z) = \sum_{i=0}^{n} A_i z^i$ be another form of the weight enumerator of a code C. The result of [5] states that

$$A(z) = (1-z)^k z^{n-k} T_C\Big(\frac{1+(q-1)z}{1-z}, \frac{1}{z}\Big).$$

Substitution in (2) completes the proof. ∎

The final topic of this section is a relation of $P(C,\epsilon)$ and *support weight distributions* of C. The set of support weight distributions is an extension of the notion of the weight distribution of a code, which is related to the functions studied in this section.

Let $\mathcal{C} \subset C$ be a subcode of C. The subset $E = \cup_{c\in\mathcal{C}} \mathsf{supp}\,(c)$ is called the support of $\mathcal{C}$, denoted $\mathsf{supp}\,\mathcal{C}$. The mth support weight distribution is the set of numbers $A^m_0, A^m_1, \ldots, A^m_n$, where

$$A^m_u = |\{\mathcal{C} \subseteq C \ : \ |\mathsf{supp}\,\mathcal{C}| = u, \mathsf{dim}\,\mathcal{C} = m\}|.$$

Support weight distributions have received much attention in the literature; see [7],[17],[2],[6] and references therein. J. Simonis [17, Lemma 1] proved the following identity, which is a natural extension of Lemma 3:

$$\sum_{u=0}^{n-w} \binom{n-u}{w} A^{k-m}_u = \sum_{\alpha=0}^{m} \begin{bmatrix} k-\alpha \\ k-m \end{bmatrix} U^C_{w\alpha}, \quad 0 \le w \le n, 0 \le m \le k,$$

where $\begin{bmatrix} u \\ v \end{bmatrix}$ is a q-binomial coefficient (Gaussian number). In order to express the numbers B_w (hence, $P(C,\epsilon)$) via the support weight distributions of C, we need to solve this set of equations with respect to the numbers $U_{w\alpha}$. For this we use the orthogonality relation of q-binomial coefficients [4]

$$\sum_{u=i}^{j} (-1)^{u-i} q^{\binom{u-i}{2}} \begin{bmatrix} u \\ i \end{bmatrix} \begin{bmatrix} j \\ u \end{bmatrix} = \delta_{ij}.$$

One obtains

$$U^C_{w\alpha} = \sum_{j=0}^{\alpha} (-1)^j q^{j(j-1)/2} \begin{bmatrix} k-\alpha+j \\ j \end{bmatrix} \sum_{i=0}^{n} A^{k-\alpha+j}_i \binom{n-i}{w}, \quad 0 \le w \le n, 0 \le m \le k.$$

[3] For reader's convenience we note that in [6] the numbers $U^C_{w\alpha}$ and $U^C_{w,w-\alpha}$ are denoted by $e_{w\alpha}(C)$ and $m_{w\alpha}(C^\perp)$, respectively. The notation for $U^{C^\perp}_{w,w-\alpha}$ in [17] is N^α_w.

Hence for $0 \le \alpha \le w-n+k-1$ we have $U^C_{w\alpha} = 0$. Thus, the summation interval of α in (13) can be extended to $0 \le \alpha \le w$. This completes the proof. ∎

The probability of undetected error for C is expressed via these functions as follows.

Proposition 7 *We have*

$$P(C,\epsilon) = \Big(1-\frac{\epsilon q}{q-1}\Big)^n U_{C^\perp}\Big(\frac{\epsilon q}{q-1-\epsilon q}, \frac{1}{q}\Big) - (1-\epsilon)^n \tag{14}$$

$$= \frac{q^k\epsilon^n}{(q-1)^n} U_C\Big(\frac{q-1-\epsilon q}{\epsilon}, \frac{1}{q}\Big) - (1-\epsilon)^n \tag{15}$$

$$= q^k(1-\epsilon)^n P\Big(C^\perp, \frac{q-1-q\epsilon}{q-q\epsilon}\Big) + q^{k-n} - (1-\epsilon)^n. \tag{16}$$

Proof: By (6) and Lemma 3,

$$B_w = \sum_{\alpha=0}^{n-k}(q^{w-\alpha}-1)U^{C^\perp}_{w\alpha} = \sum_{\alpha=0}^{n-k} q^{w-\alpha}U^{C^\perp}_{w\alpha} - \binom{n}{w}. \tag{17}$$

Therefore by (8),

$$\begin{aligned} P(C,\epsilon) &= \sum_{w=0}^{n}\sum_{\alpha=0}^{n-k}(q^{w-\alpha}-1)U^{C^\perp}_{w\alpha}\Big(1-\frac{\epsilon q}{q-1}\Big)^{n-w}\Big(\frac{\epsilon}{q-1}\Big)^w \\ &= \Big(1-\frac{\epsilon q}{q-1}\Big)^n\Big[\sum_{w=0}^{n}\sum_{\alpha=0}^{n-k}U^{C^\perp}_{w\alpha}\Big(\frac{\epsilon q}{q-1-\epsilon q}\Big)^w\Big(\frac{1}{q}\Big)^\alpha - \sum_{w=0}^{n}\binom{n}{w}\Big(\frac{\epsilon}{q-1-\epsilon q}\Big)^w\Big] \\ &= \Big(1-\frac{\epsilon q}{q-1}\Big)^n U_{C^\perp}\Big(\frac{\epsilon q}{q-1-\epsilon q}, \frac{1}{q}\Big) - (1-\epsilon)^n. \end{aligned}$$

This proves (14). The second equality is immediate from Lemma 6. To prove (16) we use (14) for $C^\perp$.

$$P\Big(C^\perp, \frac{q-1-q\epsilon}{q-q\epsilon}\Big) = \Big(1-\frac{q}{q-1}\frac{q-1-q\epsilon}{q-q\epsilon}\Big)^n U_C\Big(\frac{q-1-q\epsilon}{q(1-\epsilon)}\frac{q}{q-1-q\frac{q-1-q\epsilon}{q(1-\epsilon)}}, \frac{1}{q}\Big)$$
$$- \Big(1-\frac{q-1-q\epsilon}{q(1-\epsilon)}\Big)^n = \frac{\epsilon^n}{(q-1)^n(1-\epsilon)^n}U_C\Big(\frac{q-1-q\epsilon}{\epsilon}, \frac{1}{q}\Big) - \frac{1}{q^n(1-\epsilon)^n}.$$

Further,

$$q^k(1-\epsilon)^n P\Big(C^\perp, \frac{q-1-q\epsilon}{q-q\epsilon}\Big) + q^{k-n} - (1-\epsilon)^n = \frac{q^k\epsilon^n}{(q-1)^n}U_C\Big(\frac{q-1-\epsilon q}{\epsilon}, \frac{1}{q}\Big) - (1-\epsilon)^n,$$

as claimed. ∎

Note that relation (16) between $P(C,\epsilon)$ and $P(C^\perp,\cdot)$ has appeared in the literature (see Theorem 3.4.1 in [10]). The other parts of this proposition are new.

While the function $U_C(x,y)$ is a natural object from the coding-theoretic point of view, in combinatorics one studies different, though closely related polynomials, namely the Whitney corank-nullity function and the Tutte polynomial of the code (or of the matroid $\mathcal{M}(C)$ associated

Since our approach in the linear case links $P(C,\epsilon)$ and dependence relations among the columns of the parity-check matrix of C, we may also directly express $P(C,\epsilon)$ via the generating function of linearly (in)dependent subsets of columns in the dual code $C^\perp$. Moreover, since these generating functions for C and $C^\perp$ are related, we can also express $P(C,\epsilon)$ via the rank distribution in C itself. Namely, let G be the generator matrix of C. Let

$$U_C(x,y) = \sum_{w=0}^{n}\sum_{\alpha=0}^{k} U_{w\alpha}^C x^w y^\alpha, \tag{10}$$

where

$$U_{w\alpha}^C \quad = \quad |\{E \subset N : |E| = w, \mathsf{rank}\, G(E) = \alpha\}|,$$

By definition, let $U_{0\alpha}^C = \delta_{0,\alpha}$.

Let $E \subseteq N$, $|E| = w$. The following relation is a version of the fundamental duality for the rank function in matroid theory [18]:

$$k - \mathsf{rank}\, G(N-E) = w - \mathsf{rank}\, H(E). \tag{11}$$

In the linear case it follows from an easy linear-algebraic argument [2]: Suppose E occupies the first part of N. Represent G in the form $\left[\begin{array}{c|c} A & 0 \\ \hline B & D \end{array}\right]$, where the subset E is formed by all the columns on the left of the vertical divider. Further, suppose that A has the maximal possible number of rows, i.e., the rows of D are a basis of the row-space of the part of G to the right of the divider. As in (5), we have $\mathsf{dim}\, \mathcal{C}(E) = |E| - \mathsf{rank}\, H(E)$. Also $\mathsf{rank}\, (B|D) = \mathsf{rank}\, D$, and therefore $\mathsf{dim}\, \mathcal{C}(E) = k - \mathsf{rank}\, G(N-E)$.

This implies the following version of the MacWilliams identities:

$$U_{w\alpha}^{C^\perp} = U_{n-w,k-w+\alpha}^C. \tag{12}$$

This gives rise to the following lemma.

Lemma 6

$$U_{C^\perp}(x,y) = x^n y^{n-k} U_C(\frac{1}{xy}, y).$$

Proof: We have

$$\begin{aligned} U_{C^\perp}(x,y) &= \sum_{w=0}^{n}\sum_{\alpha=0}^{n-k} U_{w\alpha}^{C^\perp} x^w y^\alpha = \sum_{w=0}^{n}\sum_{\alpha=0}^{n-k} U_{n-w,k-w+\alpha}^C x^w y^\alpha \\ &= x^n \sum_{w=0}^{n}\sum_{\alpha=0}^{n-k} U_{w,k-n+w+\alpha}^C x^{-w} y^\alpha = x^n y^{n-k} \sum_{w=0}^{n}\sum_{\alpha=0}^{n-k} U_{w,w-\alpha}^C (xy)^{-w} y^{w-\alpha} \\ &= x^n y^{n-k} \sum_{w=0}^{n} \sum_{\alpha=w-n+k}^{w} U_{w\alpha}^C (xy)^{-w} y^\alpha. \end{aligned} \tag{13}$$

Now let $|N-E| = w, \mathsf{rank}\, G(N-E) = \alpha$. Then $|E| = n - w$ and by (11),

$$0 \le \mathsf{rank}\, H(E) = |E| - k + \alpha = \alpha - (w - n + k).$$

∎

B. An improvement for the binary case. Our proof shows that the linear case treated by Theorem 1 is essentially nothing else than a reformulation of Lemma 2. It is much simpler than the general case in [1]. However, below we argue that our approach yields the result of [1] in full generality.

Let us make some further remarks on the linear case. From Lemma 2 and the proof of Theorem 1 we see that the bound can be improved if we can get better estimates on the dimension of subcodes. As remarked above, the equality in (4) is attained on MDS codes. In the binary case the only existing linear MDS codes are codes with parameters $[n,k]$, where $k \in \mathcal{K} = \{n-1, 1, n, 0\}$. Each of these codes is unique, and therefore, the value of $P(2,n,2^k,\epsilon)$ for $k \in \mathcal{K}$ is known exactly. Otherwise, the estimate on the numbers B_w is related to the difficult problem of estimating from above the number of subsets E, $|E| = n-k$, of linearly independent columns in H, i.e., the number of information sets in the dual code $C^\perp$. However, we can observe that in the binary case the bound on $P(2,n,2^k,\epsilon)$ for $k \notin \mathcal{K}$ is strictly greater than the right-hand side of (4). Namely, Let C be a binary linear $[n,k,d]$ code, $k \neq 0, 1, n, n-1$, and A_d the number of words of weight d in it. Then

$$P(2,n,2^k,\epsilon) \geq A_d \epsilon^d (1-2\epsilon)^{n-d} + \sum_{w=n-k+1}^{n} \binom{n}{w} (2^{w-n+k}-1)\epsilon^w (1-2\epsilon)^{n-w},$$

since $B_d = A_d$ and $d \leq n-k$. Elaborating on this argument, we get the following.

Proposition 5 *Let $k, n, 1 \leq k \leq n$ be any integers. Suppose $d^* = d^*[n,k]$ is an upper bound on the distance of any $[n,k]$ code C.*

$$P(2,n,2^k,\epsilon) \geq \sum_{w=d^*}^{n-k} \frac{\binom{n-d^*}{n-w}}{\binom{w-1}{d^*-1}} \epsilon^w (1-2\epsilon)^{n-w} + \sum_{w=n-k+1}^{n} \binom{n}{w} (2^{w-n+k}-1)\epsilon^w (1-2\epsilon)^{n-w}.$$

Proof: Apply Lemma 4 to estimate B_w for $d^* \leq w \leq n-k$ and Lemma 2 to estimate them for $n-k+1 \leq w \leq n$. ∎

In particular, let $k \geq 2$ and let C be an $[n,k]$ linear code. By the Griesmer bound,

$$n \geq \sum_{i=0}^{k-1} \left\lceil \frac{d}{2^i} \right\rceil \geq d + \frac{d}{2} + (k-2) = \frac{3}{2}d + k - 2,$$

where d is the distance of C. This implies $d \leq d^* := \lfloor \frac{2}{3}(n-k+2) \rfloor$. Hence $B_{d^*} \geq 1$ and we can apply this proposition to get a bound better than (4). The comparison performed in [1] shows that this bound is in many cases the best known for binary linear codes. Any further information about a class of codes (estimates of the minimum distance, automorphisms, etc.) can further improve the bound for this class.

C. Combinatorial connections. According to (7), the numbers B_w can be called *binomial moments* of the weight spectrum of the code. They appear already in [16], where it is observed that they also count the cumulative size of subcodes in the dual code of C. Lemma 3 found in [8] gives a nice combinatorial interpretation to these numbers. Later they played a role in [9], [17], [2] and conceivably in many other works.

If the bound in this lemma holds for every E with equality, then C is an MDS code. It will follow from the proof below that bound (4) is equivalent to this lemma. Therefore, this bound holds with equality if and only if C is an MDS code.

Proof of Theorem 1: Following [8], let us define numbers B_w from the expansion

$$\widetilde{A}(x,y) = \sum_{w=1}^{n} A_w x^{n-w} y^w = \sum_{w=1}^{n} B_w (x-y)^{n-w} y^w. \tag{6}$$

Equating coefficients of these generating functions, we obtain relations between A_w and B_w:

$$B_w = \sum_{j=1}^{w} \binom{n-j}{n-w} A_j, \quad A_w = \sum_{j=1}^{w} (-1)^{j-w} \binom{n-j}{n-w} B_j, \quad 0 \le w \le n. \tag{7}$$

Clearly, $B_w \ge 0$. Using (6), rewrite (2) as follows:

$$P(C,\epsilon) = \sum_{w=1}^{n} B_w \left(1 - \frac{\epsilon q}{q-1}\right)^{n-w} \left(\frac{\epsilon}{q-1}\right)^{w}. \tag{8}$$

A second ingredient of the proof is the following

Lemma 3 [16],[8] *We have*

$$B_w = \sum_{E \in \binom{N}{w}} (|\mathcal{C}(E)| - 1),$$

where the summation is over all w-subsets of N.

Proof: This is a double counting argument. Let $E \subseteq N$, $|E| = w$. Let $c \in C \setminus \{0\}$, $\mathsf{wt}(c) = j$, $\mathsf{supp}(c) \subseteq E$. We can choose such a subset E in $\binom{n-j}{n-w}$ different ways, or

$$\sum_{E \in \binom{N}{w}} (|\mathcal{C}(E)| - 1) = \sum_{j=1}^{w} \binom{n-j}{n-w} A_j. \tag{9}$$

The lemma now follows from (7). ∎

Now let us use (8) to complete the proof of the theorem. Namely, by Lemma 2 we have $B_w \ge \binom{n}{w}(q^{w-(n-k)} - 1)$ for $n-k+1 \le w \le n$. For $1 \le w \le n-k$ just replace B_w by 0, ∎

First of the equations (7) implies the following general property of the numbers B_w.

Lemma 4 *Let $1 \le s < w \le n$. Then*

$$B_w \ge B_s \binom{n-s}{n-w} \Big/ \binom{w-1}{s-1}.$$

Proof:

$$\begin{aligned} B_w &= \sum_{i=1}^{w} \binom{n-i}{n-w} A_i \ge \sum_{i=1}^{w-1} \binom{n-i}{n-(w-1)} \frac{n-(w-1)}{w-1-(i-1)} A_i \\ &\ge \frac{n-w+1}{w-1} B_{w-1} \ge \cdots \ge \frac{(n-w+1)(n-w+2)\ldots(n-s)}{(w-1)(w-2)\ldots s} B_s \\ &= \frac{(n-s)(n-s-1)\ldots(n-s-(w-s)+1)}{(w-1)\ldots(w-1-(w-s)+1)} B_s = B_s \binom{n-s}{n-w} \Big/ \binom{w-1}{s-1}. \end{aligned}$$

where $\widetilde{A}(x,y) = \sum_{w=1}^{n} A_w x^{n-w} y^w$ is the distance function of C[1]. The problem that we address is to derive lower bounds on the function

$$P(q,n,M,\epsilon) = \min_{q\text{-ary codes of size } M} P(C,\epsilon). \tag{3}$$

This problem has a long history in coding theory; see Ch.3 in [10].

In [1] K. A. S. Abdel-Ghaffar derived a new lower bound on this quantity. We observe that his proof relies on binomial moments of the distance distribution of codes. This enables us to give a very simple proof of Abdel-Ghaffar's theorem. Though the approach discussed below applies to general codes, we first focus on the linear case because it is especially intuitive and reveals useful combinatorial insights, see Sect. 2. The general case is treated in Section 3. Section 4 deals with the asymptotics of this bound. It turns out that it gives an improvement of known upper bounds on the exponent of the probability of undetected error.

2 The linear case

Let C be an $[n,k]$ linear q-ary code[2]. The following result is proved in [1] (Theorem 2 and Eq. (12)).

Theorem 1 *Let $0 \le \epsilon \le (q-1)/q$. Let C be a linear code of size q^k with no all-zero coordinates. Then*

$$P(q,n,q^k,\epsilon) \ge \sum_{w=n-k+1}^{n} \binom{n}{w} (q^{k+w-n}-1) \left(\frac{\epsilon}{q-1}\right)^w \left(1-\frac{q\epsilon}{q-1}\right)^{n-w}. \tag{4}$$

This theorem improves some previously known bounds on the probability of undetected error and, in particular, a bound for linear codes from [19], [11].

A. Proof of Theorem 1. Our observation is that in the linear case this theorem is equivalent to the obvious fact that any submatrix of the parity-check matrix of an $[n,k]$ linear code C has rank at most $n-k$. We begin with reformulating this fact. Let $N = \{1,2,\ldots,n\}$. For any matrix A with n columns and any $E \subseteq N$ let $A(E)$ be the submatrix formed by columns with numbers in E.

Let H be a parity-check matrix of C with columns $(h_e,\ e \in N)$. Let $\mathcal{C}(E)$ be the linear subcode of C with zeros outside E, i.e., $\mathcal{C}(E) = \mathsf{ker}\,(H(E))$.

Lemma 2 *Let $E \subseteq N$, $|E| = w$. Then $|\mathcal{C}(E)| \ge \max(1, q^{w-(n-k)})$.*

Proof: Indeed,

$$\mathsf{dim}\,\mathcal{C}(E) = |E| - \mathsf{rank}\,H(E). \tag{5}$$

∎

[1] we write $\widetilde{A}$ because the sum in the standard definition of the distance function A includes the term $w = 0$.
[2] in this case F_q is the field of q elements.

Binomial moments of the distance distribution and the probability of undetected error

A. Barg * A.Ashikhmin †

Abstract

In [1] K. A. S. Abdel-Ghaffar derives a lower bound on the probability of undetected error for unrestricted codes. The proof relies implicitly on the binomial moments of the distance distribution of the code. We use the fact that these moments count the size of subcodes of the code to give a very simple proof of the bound in [1] by showing that it is essentially equivalent to the Singleton bound. This proof reveals connections of the probability of undetected error to the rank generating function of a code and to related polynomials (Whitney function, Tutte polynomial, and higher weight enumerators). We also discuss some improvements of this bound.

Finally, we analyze asymptotics. We show that an upper bound on the undetected error exponent that corresponds to the bound of [1] improves known bounds on this function.

Keywords: Distance distribution, binomial moments, rank generating function, undetected error.

1 Introduction

Let C be a code of size M over a q-ary alphabet F_q. Suppose C is used for transmission over a q-ary symmetric channel with crossover probability $\epsilon < \frac{q-1}{q}$. Let $P(C,\epsilon)$ be the probability that a transmitted codeword $c \in C$ is corrupted in the channel by an error vector e so that $c+e$ is also a codeword in C. Such an error cannot be detected by the decoder; therefore, the quantity $P(C,\epsilon)$ is called the *probability of undetected error*. The set of $n+1$ numbers

$$A_w = \frac{1}{M}\sum_{c\in C} |\{c' \in C : \mathsf{dist}\,(c,c') = w\}| \tag{1}$$

is called the (average) distance distribution of C. Then clearly,

$$\begin{aligned} P(C,\epsilon) &= \sum_{w=1}^{n} A_w \left(\frac{\epsilon}{q-1}\right)^w (1-\epsilon)^{n-w} \\ &= \widetilde{A}\Big(1-\epsilon, \frac{\epsilon}{q-1}\Big), \end{aligned} \tag{2}$$

*Bell Laboratories, Lucent Technologies, 600 Mountain Avenue 2C-375, Murray Hill, NJ 07974, e-mail abarg@research.bell-labs.com

†Los Alamos National Laboratory, Group CIC-3, Mail Stop P990, Los Alamos, NM 87545, e-mail alexei@c3serve.c3.lanl.gov.