THE UNDETECTED ERROR PROBABILITY FOR SHORTENED HAMMING CODES

Technical Report

to

NASA
Goddard Space Flight Center
Greenbelt, Maryland

Grant Number NAG 5-234

Principal Investigators

Daniel J. Costello, Jr.
Department of Electrical Engineering
Illinois Institute of Technology
Chicago, Illinois 60616

Shu Lin
Department of Electrical Engineering
University of Hawaii at Manoa
2540 Dole Street
Honolulu, Hawaii 96822

December 12, 1984

1. Introduction

Hamming or shortened Hamming codes are widely used for error detection in data communications. For example, the CCITT (International Telegraph and Telephone Consultative Committee) recommendation X.25 for packet-switched data networks adopts a distance-4 cyclic Hamming code with 16 parity-check bits for error detection [1]. The code is generated either by the polynomial,

$$\begin{aligned}\bar{g}_1(X) &= (X+1)(X^{15}+X^{14}+X^{13}+X^{12}+X^4+X^3+X^2+X+1)\\ &= X^{16}+X^{12}+X^5+1 ,\end{aligned} \tag{1}$$

or by the polynomial

$$\begin{aligned}\bar{g}_2(X) &= (X+1)(X^{15}+X^{14}+1)\\ &= X^{16}+X^{14}+X+1 ,\end{aligned} \tag{2}$$

where $X^{15}+X^{14}+X^{13}+X^{12}+X^4+X^3+X^2+X+1$ and $X^{15}+X^{14}+1$ are primitive polynomials of degree 15. The natural length of this code is $n = 2^{15}-1 = 32,767$. In practice the length of a data packet is no more than a few thousand bits which is much shorter than the natural length of the code. Consequently, a shortened version of the code is used. Often the length of a data packet varies, say from a few hundred bits to a few thousand bits, hence the code must be shortened by various degrees. Shortening affects the performance of the code. This is the subject of investigation in this paper.

For a random-error channel with bit error rate (or transition probability) ε, it was proved by Korzhik [2] that there exist (n,k) linear codes with probability P_e of an undetected error satisfying the following upper bound:

$$P_e \leq 2^{-(n-k)}[1 - (1-\varepsilon)^k] \tag{3}$$

for all n, k and ε with $0\leq\varepsilon\leq 1/2$. Korzhik's proof is an existence proof, and no general method has been found for constructing codes satisfying the bound given by (3). Only a few classes of known codes [3-6] have been proved to satisfy a weaker bound,

$$P_e \leq 2^{-(n-k)} \ . \qquad (4)$$

A code is said to be good for error detection if it satisfies the above bound, because the probability of an undetected error for the code is no greater than $2^{-(n-k)}$ even for the worst channel condition with $\varepsilon = 1/2$. In fact for small ε, the error probability P_e is much smaller than $2^{-(n-k)}$. Strict-sense Hamming codes, distance-4 Hamming codes, double-error-correcting and some triple-error-correcting primitive BCH codes of natural length are known to satisfy the bound given by (4) and their error probability P_e decreases monotonically as ε decreases [3-6]. Hence these codes are good error-detecting codes. Using a good error-detecting code with a moderate number of parity-check bits (say $n-k = 16\sim32$) in an automatic-repeat-request (ARQ) system, the probability of an undetected error can be made very small and virtually error-free data transmission can be achieved.

Even though a Hamming code of natural length satisfies the error probability bound, $P_e \leq 2^{-(n-k)}$, given by (4), a shortened Hamming code does not necessarily obey the bound [3]. Whether a shortened Hamming code satisfies the bound $2^{-(n-k)}$ depends on the degree of shortening. Because Hamming codes are normally used in shortened forms, it is important to know whether a specific shortened Hamming code satisfies the bound $2^{-(n-k)}$. In this paper we investigate the probability of an undetected error for shortened Hamming codes, particularly the shortened Hamming codes generated by the polynomials given by (1) or (2). A method for computing the probability of an undetected error is presented. We show that the codes generated by the polynomial given by (1) yield better performance than the corresponding codes generated by the polynomial given by (2).

2. Evaluation of Undetected Error Probability of Shortened Cyclic Hamming Codes

Consider a binary (n,k) linear code C. Let $P(C,\varepsilon)$ denote the probability of an undetected error when code C is used for error detection on a binary symmetric channel with transition probability ε. Let A_i and B_i be the number of codewords of weight i in C and its dual $C^{\perp}$ respectively. Then $P(C,\varepsilon)$ can be expressed in the following two forms, one is in terms of A_i and the other is in terms of B_i [7,8,9]:

$$P(C,\varepsilon) = \sum_{i=1}^{n} A_i \varepsilon^i (1-\varepsilon)^{n-i} \tag{5}$$

$$= 2^{-(n-k)} \sum_{i=0}^{n} B_i (1-2\varepsilon)^i - (1-\varepsilon)^n . \tag{6}$$

From (5) and (6), we see that, to compute the exact error probability of a linear code, one needs to know either the weight distribution $\{A_i : 0 \leq i \leq n\}$ of the code or the weight distribution $\{B_i : 0 \leq i \leq n\}$ of its dual. Theoretically, we can compute the weight distribution of an (n,k) linear code by examining its 2^k codewords or by examing the 2^{n-k} codewords of its dual. However, for large n, k and n-k, the computation becomes practically impossible. Except for some short linear codes and a few classes of linear codes [8-11], the weight distributions for most linear codes are still unknown. Consequently, it is very difficult, if not impossible, to compute the probability of an undetected error for a great many codes.

For Hamming codes, a simple formula for enumerating A_i or B_i is known [8-11], but no general formula is known for shortened Hamming codes. In general, for shortened Hamming codes, $n-k \leq k$. Hence it requires less effort in computing $\{B_i : 0 \leq i \leq n\}$ than in computing $\{A_i : 0 \leq i \leq n\}$. The weight distribution of the dual of a shortened Hamming code can be computed by generating all linear combinations of a parity-check matrix for moderate values of n-k. We call this the direct method. In the following, more effective methods for computing the weight distributions of the dual codes of shortened Hamming codes are presented. The computation is feasible for moderate values of n-k.

For any positive integer $m \geq 3$, there exists a cyclic Hamming code of length 2^m-1 and minimum distance 3. The generator polynomial of the code is a primitive polynomial $\bar{p}(X)$ of degree m. Let

$$\bar{p}(X) = \sum_{j=0}^{m} p_j X^j \qquad (7)$$

where $p_0=p_m=1$. Thus the code is a $(2^m-1, 2^m-m-1)$ code with m parity-check symbols. Let C_{2^m-1} denote this Hamming code. The dual code of C_{2^m-1}, denoted $C_{2^m-1}^{\perp}$, is a maximum-length-sequence code [8-11] which consists of the all-zero codeword and 2^m-1 maximum-length-sequences. Each maximum-length-sequence has weight 2^{m-1} and cyclicly shifting any maximum-length-sequence generates all the other maximum-length-sequences.

A distance-4 Hamming code of length 2^m-1 is simply the even weight subcode of C_{2^m-1}. It is generated by the polynomial $\bar{g}(X) = (X+1)\bar{p}(X)$ [9,11]. We denote this code by $C_{2^m-1,e}$. The dual code of $C_{2^m-1,e}$ is the first-order cyclic Reed-Muller code of length 2^m-1 which also has minimum weight 2^{m-1} [9,11].

For any positive integer n with $m<n<2^m$, let C_n be a shortened $(n,n-m)$ code of C_{2^m-1}. C_n is obtained from C_{2^m-1} by deleting the first 2^m-1-n information symbols from each codeword in C_{2^m-1} [9-11]. Let $A_{n,i}$ and $B_{n,i}$ be the number of codewords of weight i in C_n and its dual $C_n^{\perp}$ respectively. Let β be an element in the Galois field $GF(2^m)$. The trace of β, denoted $Tr(\beta)$, is defined as follows:

$$Tr(\beta) = \sum_{j=0}^{m-1} \beta^{2^j} , \qquad (8)$$

which is either 0 or 1. Let α be a root of $\bar{p}(X)$ and let

$$a_i = Tr(\alpha^i) \qquad (9)$$

for $0 \leq i < 2^m-1$. Since $\bar{p}(\alpha^{2^h})=0$ for $0 \leq h < m$, it follows from the linearity of trace $Tr(\cdot)$ that, for $0 \leq i < 2^m-1$,

$$a_{i+m2^h} = \sum_{j=0}^{m-1} p_j \, a_{i+j2^h} \, , \tag{10}$$

where the suffixes are to be taken modulo 2^m-1. It is known [8] that for a nonnegative integer u less than 2^m, $\bar{v} = (a_u, a_{u+1}, \ldots, a_{u+2^m-2})$ is a maximum-length-sequence in C_{2^m-1}. Since the weight of $\bar{v} = (a_u, a_{u+1}, \ldots, a_{u+2^m-2})$ is 2^{m-1}, it is easy to see that

$$B_{2^m-1-n,i} = B_{n,2^{m-1}-i} \; . \tag{11}$$

For $1 \leq i < 2^m$, let N_i denote the weight of $(a_u, a_{u+1}, \ldots, a_{u+i-1})$ which is a prefix of $\bar{v}$. Let $N_0=0$. Then $B_{n,i}$ is equal to the number of occurrences of integer i in the sets

$$\{N_{j+n} - N_j : \; 0 \leq j \leq 2^m-1-n\} \; , \tag{12}$$

and

$$\{2^{m-1} - N_j + N_{j-2^m+1+n} \; : \; 2^m-1-n < j < 2^m-1\} \; . \tag{13}$$

For instance, we chan choose u such that $a_{u+i}=0$ for $0 \leq i < m-1$ and $a_{u+m-1}=1$.

Now we estimate the order of computation time for finding N_j. Let p be the number of nonzero coefficients of the generator polynomial. We consider the following two methods.

Method-I

For small p, we can generate a_{u+i} with $0 \leq i < 2^m$ by using recurrence formula,

$$a_{u+i} = \sum_{j=0}^{m-1} p_j a_{u+i+j-m} \; ,$$

and obtain N_i from N_{i-1} by increasing N_i by one only if a_{u+i} is one. Then the computing time is upper bounded by $c_0 p 2^m$, where c_0 is a constant. Hereafter, c_i denotes a constant.

Method-II

If m andp are large, then we can use the following procedure for computing N_j. We assume that (i) the word length of computer is 2^h or greater where $0<h<m<2^{m-h}$, and (ii) word operations, "bit-wise Logical-AND" and "bit-wise Exclusive-OR" are available. For $0\leq i<2^{m-h}$, let

$$\bar{a}_i = (a_{i2^h+u}, a_{i2^h+u+1}, \ldots, a_{i2^h+u+2^h-1}) \ . \tag{14}$$

Then it follows from (10) that

$$\bar{a}_{i+m} = \sum_{j=0}^{m-1} p_j \bar{a}_{i+j} \ , \tag{15}$$

We first generate $(0,0,\ldots,0,1,a_{u+m},a_{u+m+1},a_{u+(m-1)2^h-1})$, i.e., $\bar{a}_0,\bar{a}_1,\ldots,\bar{a}_{m-1}$ by using

$$a_{u+i+m} = \sum_{j=0}^{m-1} p_j a_{u+i+j} \ .$$

The computing time is upper bounded by c_1pm2^h. Next we compute $\bar{a}_m,\bar{a}_{m+1},\ldots,\bar{a}_{2^{m-h}-1}$ by using (15), the computing time of which is upper bounded by $c_2p(2^{m-h}-m)$. From $\bar{a}_i$ with $0\leq i<2^{m-h}$, $N_1,N_2,\ldots,N_j,\ldots,N_{2^m-1}$ can be found sequentially as follows. Let j be $i2^h+r$, where $0\leq r<2^h$. Then N_{j+1} can be obtained from N_j by extracting the (r+1)-th bit of $\bar{a}_i$. N_j is increased by one if and only if the result of the extraction is nonzero. The computing time is c_32^m. Thus the total computing time is at most $c_1pm2^h+c_2p(2^{m-h}-m)+c_32^m$. For most cases, the first term is much smaller than the other terms.

If $c_2p2^{-h}+c_3<c_0p$, then the second method is more effective than the first one. For both methods, $\{B_{n,i}\colon\ 0\leq i<2^m\}$ can be found from (11) to (13) by c_42^m computing time. Hence, the total computing time for finding $\{B_{n,i}\colon\ 0\leq i<2^m\}$ for q different code-lengths is upper bounded as follows:

(1) For the first method,

$$(c_0p + c_4q)2^m \ . \tag{16}$$

(2) For the second method,

$$c_1pm2^h + (c_2p2^{-h}+c_3+c_4q)2^m \ . \tag{17}$$

Now we compare the above methods with a "direct method" for computing $\{B_{n,i}: \ 0\leq i<2^m\}$ which generates all linear combinations of the rows of a parity check matrix of C_n. The computing time for generating a parity check matrix is upper bounded by c_5pmn. To generate all linear combinations of the rows efficiently, we can use the Gray code in such a way that a new combination is obtained from preceding one by adding a row to it [12,13]. If we use word operations, bit-wise logical-AND and bit-wise Exclusive-OR, then the computing time is proportional to $n2^m/\ell$, where ℓ is the word length of computer. We assume that the set of code lengths n_j with $1\leq j\leq q$ for which $\{B_{n_j,i}: \ 0\leq i<2^m\}$ is to be found is given beforehand. Note that we don't need this assumption for the methods described above. If we use word operation "find the weight of a word", then the order of the total computing time for finding the distributions $\{B_{n_j,i}: \ 0\leq i<2^m\}$ for $1\leq j\leq q$ can be estimated as

$$c_5pmn_{max} + (c_6n_{max}/\ell + c_7q)2^m \tag{18}$$

where

$$n_{max} = \max\{n_j,2^m-1-n_j: \ 1\leq j\leq q\} \ . \tag{19}$$

Since $c_2\approx c_6$, $c_4\approx c_7$ and n_{max}/ℓ is much greater than p, for most cases the first or second method is more efficient than the direct method, at least if $q<n_{max}/\ell$.

Let $C_{n,e}$ denote the even weight subcode of C_n. In fact, $C_{n,e}$ is a shortened code of the distance-4 Hamming code $C_{2^m-1,e}$ generated by $\bar{g}(X) = (1+X)\bar{p}(X)$. The number of codewords of weight i in the dual code $C_{n,e}^{\perp}$ of $C_{n,e}$, denoted $B_{n,i,e}$, is

$$B_{n,i,e} = B_{n,i} + B_{n,n-i} \ . \tag{20}$$

For $15<n<2^{15}$, let $C_{n,e}^{(1)}$ and $C_{n,e}^{(2)}$ be the even weight shortened codes of length n generated by $\bar{g}_1(X)$ of (1) and $\bar{g}_2(X)$ of (2) respectively. For

$\bar{p}_1(X) = X^{15}+X^{14}+X^{13}+X^{12}+X^4+X^3+X^2+X+1$ and $\bar{p}_2(X) = X^{15}+X^{14}+1$, N_i's with $1\leq i<2^{15}$ are computed by the first method. From these N_i's, the weight distributions $\{B_{n,i,e}: 0\leq i\leq n\}$ of the dual codes of $C^{(1)}_{n,e}$ and $C^{(2)}_{n,e}$ for $16<n<2^{15}$ are obtained. From these weight distributions and (6), the error probabilities, $P(C^{(1)}_{n,e},\varepsilon)$ and $P(C^{(1)}_{n,e},\varepsilon)$, are computed and plotted in Figures 1 and 2 respectively as functions of channel bit-error-rate ε for code length $n=2^{\ell}$ with $5\leq\ell\leq 14$ and $n=2^{15}-1$.

From Figures 1 and 2, we see that if the two distance-4 Hamming codes recommended by CCITT X.25 are shortened too much, the shortened codes do not obey the bound $2^{-(n-k)}$ given by (4), i.e., their error-detection performance becomes poor as ε becomes large. Therefore, in order to maintain the data reliability, the length of a data packet should not be too short. In Tables 1 and 2, we tabulate some code lengths for which the error probabilities, $P(C^{(1)}_{n,e},\ )$ and $P(C^{(1)}_{n,e},\ )$, do not obey the bound $2^{-(n-k)}$ given by (4). We also tabulate the peak values of error probabilities and the values of channel bit-error-rate ε where the peak values occur. Note that in most cases, peak values occur for $4/n<\varepsilon<5/n$. For the longer values $n=2^{\ell}$ with $8\leq\ell\leq 14$ for $C^{(1)}_{n,e}$ and with $10\leq\ell\leq 14$ for $C^{(2)}_{n,e}$, no peak is detected within accuracy in computation. The peak values of the probability of undetected error for $C^{(1)}_{n,e}$ and $C^{(2)}_{n,e}$ are plotted in Figure 3 as functions of code length n. From Tables 1 and 2 and Figures 1-3, we see that the codes generated by the polynomial $\bar{g}_1(X)$ of (1) give better performance than the corresponding codes generated by the polynomial $\bar{g}_2(X)$ of (2).

3. Conclusion

In this paper, we have investigated the error-detection performance of shortened Hamming codes, particularly the shortened codes obtained from the two distance-4 Hamming codes adopted by CCITT recommendation X.25. First two methods for computing the weight distributions of the dual codes of shortened

Hamming codes have been presented. We have shown that these methods are in general more effective than the direct method. Using the weight distributions of the dual codes, we have evaluated the probability of undetected error for the codes obtained from shortening the two X.25 distance-4 Hamming codes. We have shown that shortening does affect the error-detection performance of the two X.25 codes. If the codes are shortened too much, the shortened codes do not obey the bound 2^{-16}. We have also shown that the codes generated by $\bar{g}_1(X) = X^{16}+X^{12}+X^5+1$ give better performance than the corresponding codes generated by $\bar{g}_2(X) = X^{16}+X^{14}+X+1$.

REFERENCES

1. CCITT: Recommendation X.25, "Interface Between Data Terminal Equipment (DTE) and Data Circuit-terminating Equipment (DCE) for Terminals Operating in Packet Mode on Public Data Networks," with Plenary Assembly, Doc. No. 7, Geneva, 1980.

2. V.I. Korzhik, "Bound on Undetected Error Probability and Optimum Group Codes in a Channel with Feedback," Radiotecknika, 20, Vol. 1, pp. 27-33, 1965. (English translation: Telecommunication and Radio Engineering, 2, pp. 87-92, January, 1965.)

3. S.K. Leung-Yan-Cheong and M.E. Hellman, "Concerning a Bound on Undetected Error Probability," IEEE Transaction on Information Theory, Vol. IT-22, No. 2, pp. 235-237, March, 1976.

4. S.K. Leung-Yan-Cheong, E.R. Barnes, and D.U. Friedman, "Some Properties of Undetected Error Probability of Linear Codes," IEEE Transaction on Information Theory, Vol. IT-25, No. 1, pp. 110-112, January, 1979.

5. T. Kasami, T. Kløve, and S. Lin, "Error Detection with Linear Block Codes," IEEE Transaction on Information Theory, Vol. IT-29, No. 1, January, 1983.

6. J.K. Wolf, A.M. Michelson, and A.H. Levesque, "On the Probability of Undetected Error for Linear Block Codes," IEEE Transaction on Communication, Vol. COM-30, No. 2, pp. 317-324, February, 1982.

7. F.J. MacWilliams, "A Theorem on the Distribution of Weights in Systematic Code," Bell System Technical Journal, Vol. 42, pp. 79-94, 1963.

8. F.J. MacWilliams and N.J.A. Sloane, Theory of Error-Correcting Codes, North Holland, Amsterdam, 1977.

9. S. Lin and D.J. Costello, Jr., Error Control Coding: Fundamentals and Applications, Prentice-Hall, New Jersey, 1983.

10. E.R. Berlekamp, Algebraic Coding Theory, McGraw-Hill, New York, 1968.

11. W.W. Peterson and E.J. Weldon, Jr., Error-Correcting Codes, 2nd Edition, MIT Press, Cambridge, MA, 1972.

12. T. Kasami, T. Fujiwara, and S. Lin, "An Approximation of the Weight Distribution of Binary Linear Codes," Proceedings of the 6th Conference on Information Theory and Its Applications, Matsuyama, Japan, November, 1983.

13. T. Fujiwara, A. Kitai, S. Yamamura, T. Kasami and S. Lin, "On the Probability of Undetected Error for Shortened Cyclic Hamming Codes," Proceedings of the 5th Conference on Information Theory and Its Applications, Hachimantai, Japan, October, 1982.

Table 1

The maximum values of $P(C_{n,e},\varepsilon)$ for $0<\varepsilon\leq 1/2$

n	ε	$P(C_{n,e},\varepsilon)$
22	1.85×10^{-1}	1.82×10^{-4}
24	1.71×10^{-1}	1.69×10^{-4}
26	1.59×10^{-1}	1.50×10^{-4}
28	1.48×10^{-1}	1.31×10^{-4}
30	1.39×10^{-1}	1.15×10^{-4}
32	1.30×10^{-1}	1.00×10^{-4}
40	1.05×10^{-1}	7.83×10^{-5}
50	8.55×10^{-2}	5.12×10^{-5}
64	7.03×10^{-2}	3.18×10^{-5}
128	4.55×10^{-2}	1.70×10^{-5}

Table 2

The maximum values of $P(C_{n,e},\varepsilon)$ for $0<\varepsilon\leq 1/2$

n	ε	$P(C_{n,e},\varepsilon)$
22	$1.98\text{x}10^{-1}$	$2.10\text{x}10^{-4}$
24	$1.84\text{x}10^{-1}$	$1.94\text{x}10^{-4}$
26	$1.70\text{x}10^{-1}$	$1.72\text{x}10^{-4}$
28	$1.58\text{x}10^{-1}$	$1.50\text{x}10^{-4}$
30	$1.46\text{x}10^{-1}$	$1.38\text{x}10^{-3}$
32	$1.36\text{x}10^{-1}$	$1.64\text{x}10^{-4}$
40	$1.08\text{x}10^{-1}$	$1.89\text{x}10^{-4}$
50	$8.63\text{x}10^{-2}$	$1.67\text{x}10^{-4}$
64	$6.73\text{x}10^{-2}$	$1.32\text{x}10^{-4}$
128	$3.47\text{x}10^{-2}$	$5.12\text{x}10^{-5}$
256	$1.93\text{x}10^{-2}$	$2.28\text{x}10^{-5}$
512	$1.19\text{x}10^{-2}$	$1.67\text{x}10^{-5}$

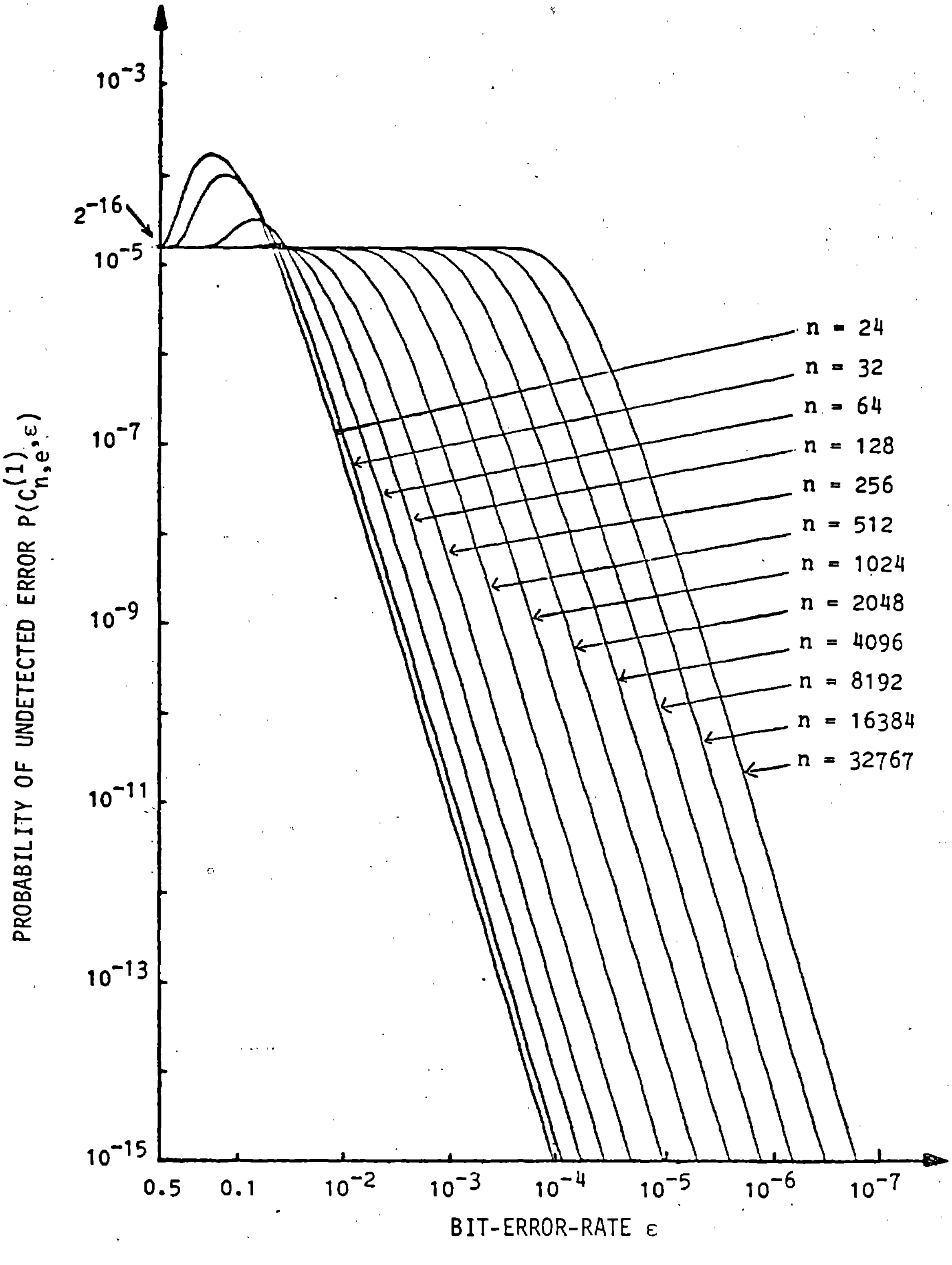

Figure 1 Actual values of probability of undetected error for the shortened cyclic Hamming code of length n generated by $\bar{g}_1(X) = 1+X^5+X^{12}+X^{16}$ as a function of channel bit error rate ε.

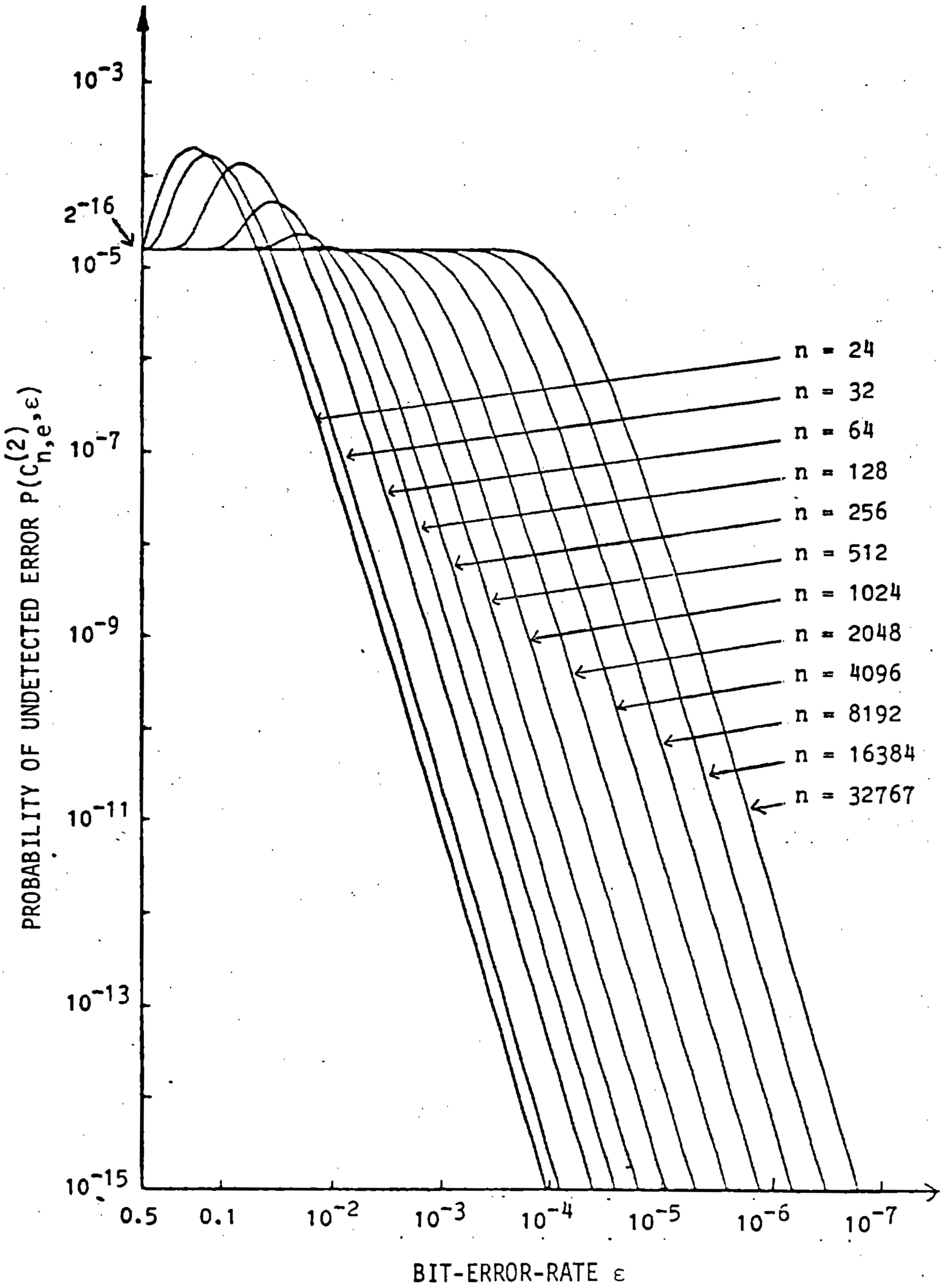

Figure 2 Actual values of probability of undetected error for the shortened cyclic Hamming code of length n generated by $\bar{g}_2(X) = 1+X+X^{14}+X^{16}$ as a function of channel bit error rate ε.

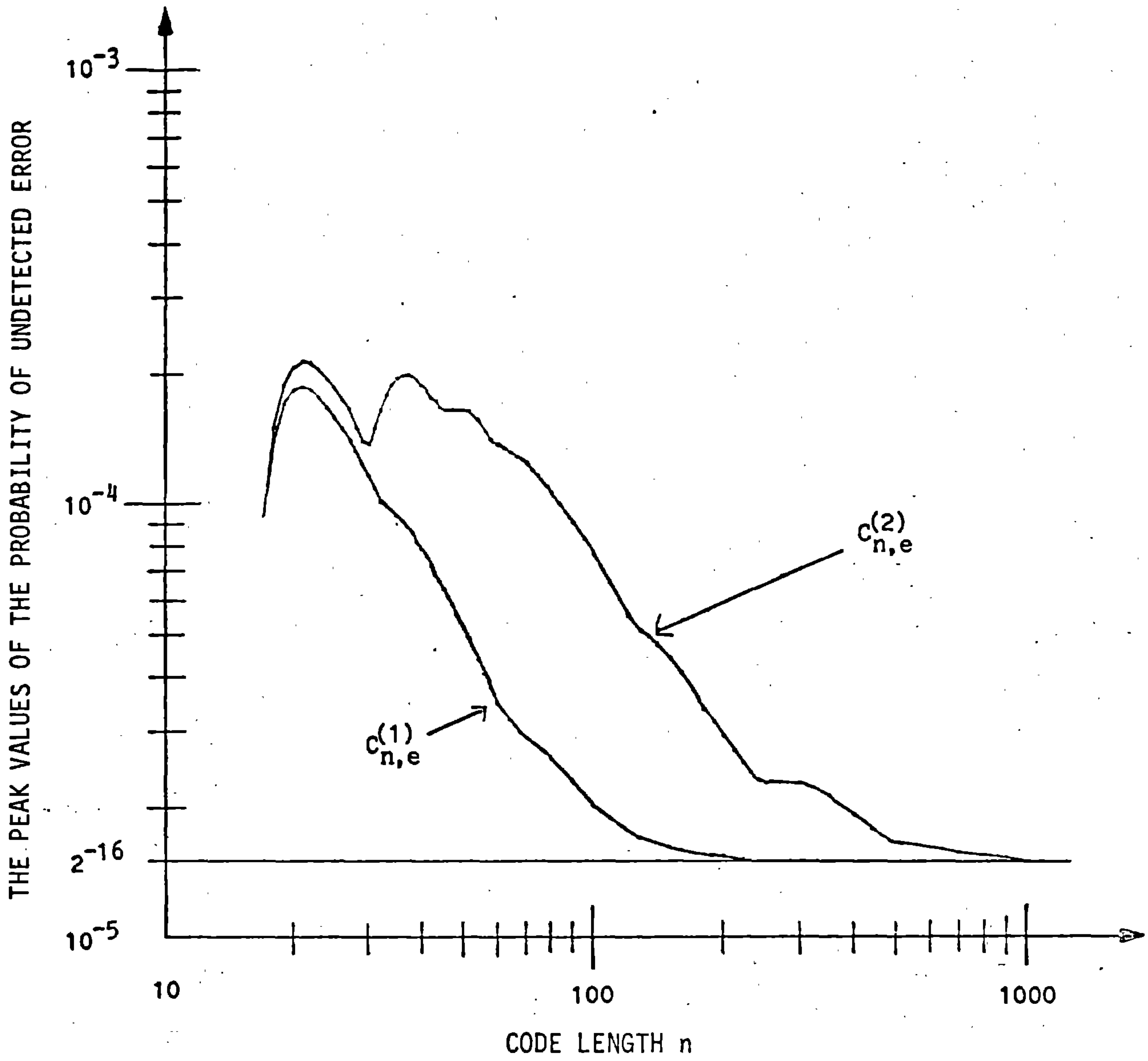

Figure 3 The peak values of the probability of undetected error for $C^{(1)}_{n,e}$ and $C^{(2)}_{n,e}$.