A Remark on Plotkin's Bound

Warwick de Launey and Daniel M. Gordon

Abstract— **Let $A(n,d)$ denote the greatest number of codewords possible in a binary block code of length n and distance d. Plotkin gave a simple counting argument which leads to an upper bound $B(n,d)$ for $A(n,d)$ when $d \geq n/2$. Levenshtein proved that if Hadamard's conjecture is true then Plotkin's bound is sharp. Though Hadamard's conjecture is probably true, its resolution remains a difficult open question. So it is natural to ask what one can prove about the ratio $R(n,d) = A(n,d)/B(n,d)$. This note presents an efficient heuristic for constructing for any $d \geq n/2$, a binary code which has at least $0.495B(n,d)$ codewords. A computer calculation confirms that $R(n,d) > 0.495$ for d up to one trillion.**

Keywords— **Plotkin bound, Hadamard matrix, Paley matrix, Goldbach conjecture, high distance binary block codes**

I. Preliminaries and Overview

For $n > d > 0$, let $A(n,d)$ denote the maximum number of codewords possible in a binary block code of length n and minimum (Hamming) distance d. Notice that, if d is odd, then $\mathcal{C}$ is an (n,M,d) code if and only if the code $\mathcal{C}'$ obtained by adding a parity check bit to each codeword in $\mathcal{C}$ is an $(n+1,M,d+1)$ code. Therefore, if d is even, then $A(n,d) = A(n-1,d-1)$. So in order to understand the behaviour of $A(n,d)$ it is sufficient to understand its behaviour for d even.

An elementary counting argument gives Plotkin's bound. This states that for d even,

$$A(n,d) \leq B(n,d) = \begin{cases} 2\left\lfloor \frac{d}{2d-n} \right\rfloor & \text{if } 2d > n \geq d, \\ 4d & \text{if } n = 2d. \end{cases}$$

An $r \times n$ *partial Hadamard matrix* is a $(1,-1)$-matrix H such that

$$HH^\top = nI_r.$$

It is easy to show that $r \leq n$, and that if $r > 2$, then $n = 4t$ is divisible by four. If $r = n$, then H is a Hadamard matrix. In this case, the matrix is said to be *complete*.

The connection between partial Hadamard matrices and Plotkin's bound is provided by the following lemma which was proved in [4] by using partial Hadamard matrices in place of the Hadamard matrices in Levenshtein's well known construction for high distance binary block codes.

Lemma 1: If there is a $c2t \times 2t$ partial Hadamard matrix for all even $t \geq N$, then for all $2d \geq n \geq (2-1/N)d$,

$$cB(n,d) \leq A(n,d) \leq B(n,d).$$

In 1893, Hadamard conjectured that a complete matrix exists for any order divisible by four. So (as Levenshtein proved in [8]) if Hadamard's conjecture is true then Plotkin's bound is sharp. Even though there is an extensive literature on Hadamard matrices, the conjecture remains unproven. It is therefore natural to ask what we can prove about the ratio $R(n,d) = A(n,d)/B(n,d)$.

A recent paper [4] contains the following asymptotic result.

Theorem 2: For any $\epsilon > 0$, there exists an integer N, such that for all integers n and d satisfying $2d \geq n \geq d(2 - \frac{1}{N})$, we have $R(n,d) > \frac{1}{3}(1-\epsilon)$.

In other words, provided that the distance d is rather close to $n/2$ or $B(n,d)$ is large, for n sufficiently large Plotkin's bound is at worst off by a factor close to three.

Theorem 2 is ultimately a consequence of a recent result in analytic number theory which states that any sufficiently large odd number may be written as the sum of three primes which are all close to each other. The idea is to paste together three truncated Paley Hadamard matrices with nearly equal orders.

However, we think that there are actually enough pairs of known Hadamard matrices to allow us to replace the factor one third in Theorem 2 by one half.

Firstly, as noted at the end of [4], pairs of Paley matrices might be used to give Theorem 2 with the factor equal to one half. However a proof along these lines seems beyond our reach, since it would imply an asymptotic form of the long standing Goldbach conjecture which states that any even integer greater than two may be written as the sum of two primes. We will examine this approach in more detail later in this paper.

Secondly, if we use one Paley Hadamard and one Hadamard matrix whose existence is given by Craigen's recent asymptotic results [1] [1], then the problem reduces to finding a prime in a short arithmetic sequence. This suggests assuming the Extended Riemann Hypothesis (ERH), and seeing what can be proved. Indeed, if we let $\boldsymbol{r}(n)$ denote the great-

The authors are with the IDA Center for Communications Research, 4320 Westerra Court, San Diego, CA 92121 USA (email: {warwick,gordon}@ccrwest.org).

[1] It is interesting to note that Craigen's improvements over Seberry's earlier asymptotic existence result are essential.

est number of rows in a partial Hadamard matrix with n columns, and let $\epsilon > 0$, then in [5] the authors proved that if the ERH is true then for every sufficiently large $n \equiv 0 \pmod 4$

$$\boldsymbol{r}(n) \geq \frac{n}{2} - n^{\frac{17}{22}+\epsilon}. \tag{1}$$

This equation implies that for any $c < 1/2$, there is an integer N such that for every $n > N$ congruent to zero modulo four we have $\boldsymbol{r}(n) \geq cn$. Lemma 1 with $t = n/2$ then shows that (assuming the ERH is true) we can take the factor in Theorem 2 equal to one half.

Regardless of how high we can make the factor in Theorem 2, the theorem has two major deficiencies. Firstly, it seems to be difficult to estimate how large N needs to be, and secondly, if N needs to be large, then d will be forced to be very close to $n/2$.

In this correspondence, we study an efficient heuristic, employing only Paley Hadamard matrices and Hadamard matrices of small orders, for constructing, for any block size n and distance $d \geq n/2$, codes with at least $0.495B(n,d)$ codewords.

The heuristic relies on the following observation.

Lemma 3: If $2(t-1) = p_1 + p_2$ where p_1 and $p_2 > p_1$ are primes then there is a $2(p_1+1) \times 4t$ partial Hadamard matrix.

Proof: Paley showed that there is a Hadamard matrix of order $2(q+1)$ for any prime power. So the desired partial Hadamard matrix can be obtained by concatenating the matrices obtained by taking the first $2(p_1+1)$ rows of the Hadamard matrices obtained by taking q equal to p_1 and p_2. ■

So if we use pairs of Paley matrices, we are left with the problem of expressing $2(t-1)$ as the sum of two primes which are close together. Of course it seems to be difficult to prove that the primes p_1 and p_2 of Lemma 3 exist for every $t > 1$. Nevertheless, we can describe a simple probabilistic model which predicts with some accuracy how far apart the two primes are likely to be.

As t grows, Lemma 3 gives partial Hadamard matrices which are very close to half complete. However, for small t it is useful to use all the known Hadamard matrices of order up to, say, twelve thousand. This seems to allow us to form for any t a $c4t \times 4t$ partial Hadamard matrix where $c > 0.495$.

To test our model (and to obtain a supply of nearly half complete partial Hadamard matrices) we used a computer to find the optimal pair of primes for each even number up to 10^{12}. The computer calculation implies the following result.

Theorem 4: For all integers $d \leq 10^{12}$ and n satisfying $2d \geq n \geq d$ we have $R(n,d) > 0.495$.

The calculation also confirms the probabilistic model for d up to 10^{12}, suggesting that Theorem 4 is true for all d. In any case, for $d \leq 10^{12}$ the rate of the largest code of length $n \in [d, 2d]$ is very close to the theoretical limit $\frac{1}{n}\log_2 B(n,d)$. We note that proving $R(n,d) > 0.5$ would require a completely different approach to that taken in this paper.

II. Modeling the Offset

Definition 5: Suppose an even integer $2n$ may be represented as the sum of two primes. For any representation $2n = p_1 + p_2$ with $p_1 \leq p_2$, call $p_2 - n = n - p_1$ the *separation.* An *optimal representation* has minimal separation, which we will call the *offset* for n.

Goldbach's Conjecture simply states that the offset is well defined for every even integer larger than two.

Cramér's model (Granville, in [7], gives an interesting description of his model and modifications of it) treats primality as a random event, allowing one to use standard probabilistic methods to model the behavior of primes. For a given n and small k, the prime number theorem implies that $n-k$ will be prime with probability asymptotically equal to $1/\log n$. If $n-k$ is prime, it is likely (but unproven) that $n+k$ will be prime with probability about equal to $2/\log n$ (since $n+k$ must be odd). So if we try $(\log^2 n)/2$ values for k starting at zero and increasing by one, we would expect that about one of the values for k will have $n+k$ and $n-k$ both prime.

Now if P is the probability of success in each of a sequence of Bernoulli trials, then the expected number of trials needed to obtain a single success (on the last attempt) is $1/P$. Therefore under the current model, the expected value of the smallest k for which $n+k$ and $n-k$ are both prime would be $(\log^2 n)/2$.

This model can be made more accurate using ideas in [2]. If some prime ℓ divides n, then the probability that ℓ does not divide $n-k$ and ℓ does not divide $n+k$ is $(\ell-1)/\ell$ instead of $(\ell-1)^2/\ell^2$. On the other hand, if ℓ does not divide n, then the probability that ℓ does not divide $n-k$ and ℓ does not divide $n+k$ is $(\ell-2)/\ell$ instead of $(\ell-1)^2/\ell^2$. Thus, the expected size of the offset will be

$$\log^2 n/2 \cdot \left(\prod_{\ell \mid n} \frac{\ell-1}{\ell}\right)\left(\prod_{\ell \nmid n} \frac{(\ell-1)^2}{\ell(\ell-2)}\right),$$

where each prime ℓ appears in one of the products. See [9] for extensive computations on the accuracy of this formula.

For a random n, the probability of being divisible by ℓ is just $1/\ell$, so assuming independence of the scaling factors for each prime ℓ, the expected offset becomes

$$\begin{aligned} & \log^2 n/2 \cdot \prod_{l \text{ prime}} \left(\frac{\ell-1}{\ell} \cdot \frac{1}{\ell} + \frac{(\ell-1)^2}{\ell(\ell-2)} \cdot \frac{\ell-1}{\ell} \right) \\ = \; & \log^2 n/2 \cdot \prod_{\ell \text{ prime}} \left(1 + \frac{1}{\ell^2(\ell-2)} \right) \\ \approx \; & 0.5665 \log^2 n. \end{aligned}$$

We are of course most interested in the extremal behaviour of the offset. In the worst case, where n is prime or relatively prime to the smallest primes, the probability of a given $p, n-p$ pair both being prime is

$$\frac{2}{\log^2 n} \prod_l \frac{l(l-2)}{(l-1)^2} \approx 1.32/\log^2 n.$$

The probability of $c \log^3 n$ consecutive pairs all failing is

$$\left(1 - \frac{1.32}{\log^2 n}\right)^{c \log^3 n} \leq e^{-c \log n/0.757} = n^{-c/0.757}.$$

Notice that for $c > 0.757$, this probability is less than n^{-1}; so for any $c > 0.757$ the expected number of n for which the gap is greater than $c \log^3 n$ is finite but arbitrarily large as c approaches 0.757. For $c > 0.757$ we obtain the following estimate for the number of integers $n > N$ with offset greater than $c \log^3 n$:

$$\int_N^\infty n^{-c/0.757} dn = \frac{N^{1-c/0.757}}{c/0.757 - 1}$$

So combined with Lemma 3 our model suggests that for any $c > 0.757$ and

$$n > (c/0.757 - 1)^{-(c/0.757-1)^{-1}} \tag{2}$$

divisible by four we have

$$\boldsymbol{r}(n) \geq \frac{1}{2}(n - c \log^3 n) \tag{3}$$

which is much better than (1). Note that when $c = 2 \times 0.757$, the lower bound (2) on n is one.

III. Computational Experiments

It is easy to find good partial Hadamard matrices for a wide range of n. For large n, one simply looks for optimal pairs. These may not give the best possible bounds, but asymptotically they do very well. For smaller n, one may use tables of known Hadamard matrices, such as Table 24.33 of [1], concatenating pairs of Hadamard matrices with orders close together.

TABLE I

n with partial Hadamard matrices with $< 0.498n$ rows

n	a	b	a/n
428	212	216	0.495327
668	332	336	0.497006
716	356	360	0.497207
764	380	384	0.497382
892	444	448	0.497758
956	476	480	0.497908
1436	712	724	0.495822
1912	952	960	0.497908
1916	952	964	0.496868
3832	1908	1924	0.497912

TABLE II

Moving average of the offsets for n up to 10^{14}

n	average offset/ $\log^2 n$
10^6	.4746
10^7	.4858
10^8	.4985
10^9	.5081
10^{10}	.5149
10^{11}	.5199
10^{12}	.5252
10^{13}	.5285
10^{14}	.5315

Table III shows all $n < 10^{12}$ for which the best partial Hadamard matrix had fewer than $0.498n$ rows. They are all very small, and cannot be improved without a different construction for partial Hadamard matrices or new results about Hadamard matrices. For example, to improve $n = 428$ we would either need to find a Hadamard matrix of order 428 or construct a partial $r \times 428$ partial Hadamard matrix by some means other than concatenating two Hadamard matrices. In any case, our computer calculation confirms that there are enough good partial Hadamard matrices to prove Theorem 4.

We also tested our heuristic model for the offsets. We first computed a moving average for the offsets as n increased up to 10^{12}. For larger values, we computed the average offset for a range of 10^9 integers. Table III shows the results.

We would expect this average to have an asymptote at about 0.5665, but in order to test for the presence of the asymptote we would have to examine very large numbers. This slow convergence is not surprising; similar behavior has been observed in the size of the gaps between successive prime numbers, where computations up to 10^{14} were not sufficient to give evidence for or against Cramér's conjecture [7]. Nevertheless, Table III supports our model in the sense that the expected value of the offset is roughly

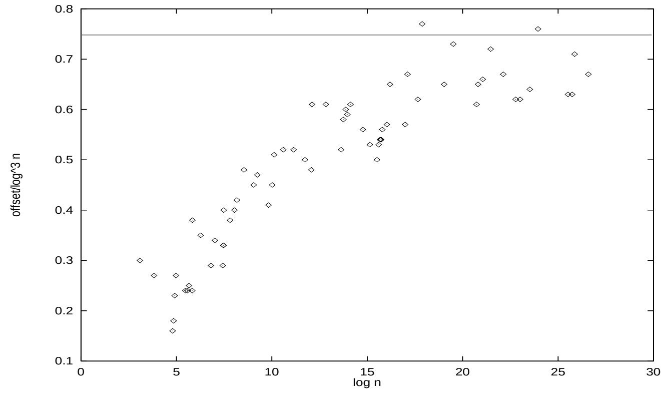

Fig. 1. Maximal raw offsets for n up to 10^{12}

proportional to $\log^2 n$.

Next we tested the distribution of maximal raw offsets as n increases. These computations were the most expensive, and were done in a weekend run on 128 nodes of a CRAY T3D. We expected that the largest offsets would grow to about $0.757 \log^3 n$ and then level off. Figure 1 shows behavior consistent with this, suggesting that the inequality (3) holds, for c close to 0.757, for all n, and hence that Theorem 4 is true for all n.

References

[1] R. Craigen, Hadamard Matrices and Designs, in *The CRC Handbook of Combinatorial Designs,* C. J. Colbourn and J. H. Dinitz, eds., CRC Press, 1996, 370–377.

[2] J-M. Deshouillers and H. J. J. te Riele, On the probabilistic complexity of numerically checking the binary Goldbach conjecture in certain intervals, *CWI Report MAS-R9820* , October, 1998.

[3] J-M. Deshouillers and H. J. J. te Riele and Y. Saouter, New experimental results concerning the Goldbach conjecture, in *Algorithmic Number Theory (ANTS-III)* J. P. Buhler, ed., Springer LNCS 1423, 1998, 204–215.

[4] Warwick de Launey, On the Asymptotic Existence of Partial Complex Hadamard Matrices and Related Combinatorial Objects, *Discrete Applied Mathematics*, **102** (2000), 37–45.

[5] Warwick de Launey and Daniel M. Gordon, A comment on the Hadamard conjecture, preprint.

[6] Andrew Granville, Unexpected Irregularities in the Distribution of Prime Numbers, in *Proceedings of the International Congress of Mathematicians* Birkhäuser, 1995, Volume 1, 388–399.

[7] Andrew Granville, Harald Cramér and the distribution of prime numbers, *Scand. Actuarial J.*, **2** (1995), 12–28.

[8] V. I. Levenshtein, The Application of Hadamard matrices to a problem in coding, *Problemy Kibernetiki*, **5** (1961) 123-136. English translation in *Problems of Cybernetics*, **5** (1964) 166-184.

[9] Jörg Richstein, Computing the number of Goldbach partitions up to $5 \cdot 10^8$, in *Algorithmic Number Theory (ANTS-IV)* W. Bosma, ed., Springer LNCS 1838, 2000, 475–490.

Daniel M. Gordon received the B.S. degree in mathematics and literature from the California Institute of Technology, Pasadena, in 1981, and the M.A. and Ph.D. degrees in mathematics from the University of California at San Diego, in 1984 and 1986, respectively. He was an Assistant Professor in the Departments of Mathematics and Computer Science at the University of Georgia, Athens, from 1986-1990, a Member of the Technical Staff at Sandia National Laboratories from 1990-1991, and an Associate Professor in the Department of Computer Science at the University of Georgia from 1991-1992. Since 1992 he has worked as a research scientist at the IDA Center for Communications Research in La Jolla, California. His research interests include computational number theory and covering designs.

Warwick de Launey received a Ph.D. in Applied Mathematics from the University of Sydney, Australia, in 1987. As a graduate student he taught Combinatorical Mathematics at that University. From 1986 to 1989 he held technical and management positions at SIROMATH a Mathematical and Statistical Consulting firm. From 1990 to 1996 he was a Senior Research Scientist at the Defence Science and Technology Organisation, Australia, and since 1996 he has worked as a research scientist at the Center for Communications Research. He has published about thirty articles on combinatorial design theory.