Almost disjoint families of 3-term arithmetic progressions

Hayri Ardal[*], Tom C. Brown[†], Peter A.B. Pleasants[‡]

Communicated by R.L. Graham.

Citation data: Hayri Ardal, Tom Brown, and Peter A.B. Pleasants, *Almost disjoint families of 3-term arithmetic progressions*, J. Combin. Theory Ser. A **109** (2005), 75–90.

Abstract

We obtain upper and lower bounds for the size of a largest family of 3-term arithmetic progressions contained in $[0,n-1]$, no two of which intersect in more than one point. Such a family consists of just under a half of all the 3-term arithmetic progressions contained in $[0,n-1]$.

MSC: 05D99

Keywords: Arithmetic progression, Almost disjoint

1 Introduction

This paper studies the problem of the maximum size of a family of 3-term arithmetic progressions of integers from the interval $[0,n-1]$ such that no two have more than one integer in common. This seemingly simple problem turns out to have a lot of interesting structure. To give it some context, it can be viewed as an extremal set system problem with an arithmetic constraint. A partial t-(n,k)-design is a family of k-element subsets of $[0,n-1]$ such that every two have less than t elements in common (see [2]). A special case of a more general result of Deza et al. [1] (also [2, Theorem 6.3]) is that for large n the number of subsets in a partial t-(n,k)-design is at most

$$\binom{n}{t}/\binom{k}{t} \tag{1}$$

and Rödl [3] (also [2, Theorem 6.4]) has shown that the maximum size of a partial t-(n,k)-design is in fact asymptotic to $\binom{n}{t}/\binom{k}{t}$ as $n\to\infty$. These are pure set-theoretic results and we should like to investigate the effect of introducing some arithmetic constraints. There is little scope for arithmetic structure in a 2-element set, but a 3-element set of integers can be given arithmetic structure by insisting it is an arithmetic progression, that is, it has the form $\{a,a+d,a+2d\}$, $a,d\in\mathbb{N}$.

We shall use the abbreviation "AP" for "arithmetic progression" and we denote a particular 3-term AP $\{a,a+d,a+2d\}$ more briefly by $\langle a;d\rangle$. The problem of the maximum number of disjoint 3-term

[*]Department of Mathematics, Boğaziçi university, Bebek 80815, Istanbul, Turkey
[†]Department of Mathematics, Simon Fraser University, Burnaby, BC Canada V5A 1S6
[‡]Department of Mathematics, University of Queensland, QLD 4072, Australia

APs that can be packed into an interval $[0,n-1]$ is trivial: obviously the maximum number is $\leq \lfloor n/3 \rfloor$ and this number is achieved by the family $\langle 0;1\rangle, \langle 3;1\rangle, \ldots, \langle 3\lfloor n/3\rfloor - 3;1\rangle$. This can be described as the problem of finding a large partial 1-$(n,3)$-design consisting of APs. The next level of complexity is to look for large partial 2-$(n,3)$-designs consisting of APs, which is our problem of finding large families of 3-term APs in $[0,n-1]$ such that no two APs have more than one number in common. We shall call such a family of 3-term APs *almost disjoint*. We shall show that, as for unrestricted partial 2-$(n,3)$-designs, the maximum size of a family of almost disjoint 3-term APs is asymptotic to a constant multiple of n^2, but that the constant is smaller than the value 1/6 given by (1) for the unrestricted case.

The total number of 3-term APs in $[0,n-1]$ is

$$\binom{\lfloor n/2\rfloor}{2} + \binom{\lceil n/2\rceil}{2} = \frac{n^2}{4} + O(n) \tag{2}$$

(the number of pairs of integers in $[0,n-1]$ whose difference is even) so we shall express the sizes of families of 3-term APs in terms of multiples of $n^2/4$, so that the multiplier tells us what asymptotic proportion of the total our family is.

Being almost disjoint puts little constraint on a family of 3-term APs because any AP $\langle a;d\rangle$ has two numbers in common with at most six others, namely

$$\langle a;d/2\rangle, \langle a+d;d/2\rangle, \langle a-d;d\rangle, \langle a+d;d\rangle, \langle a-2d;2d\rangle, \langle a;2d\rangle \tag{3}$$

(Of these the first two occur only when d is even and some of the others do not occur when $a < 2d$ or $a \geq n-4d$.) This immediately shows that any maximal almost disjoint family in $[0,n-1]$ contains at least 1/7 of all 3-term APs in $[0,n-1]$, so has size $\geq \frac{1}{7}(n^2/4) - O(n)$, and this bound can be increased to $\frac{3}{14}(n^2/4) - O(n)$ if we also take account of the fact that a proportion of about 1/12 of the 3-term APs $\langle a;d\rangle$ in $[0,n-1]$ (those with d odd, $a < d$ and $a \geq n-3d$) are completely disjoint from all others. Similar considerations give an upper bound $\frac{73}{84}(n^2/4) + O(n)$ for the size of a maximal almost disjoint family of 3-term APs in $[0,n-1]$, though this is weaker than the upper bound $\frac{2}{3}(n^2/4)$ obtained from the more widely applicable estimate (1).

Our main result is to show that the maximum size of an almost disjoint family of 3-term APs in $[0,n-1]$ is asymptotic to $C(n^2/4)$, for some C in the range $0.476 < C < 0.485$. We also obtain a lower bound of the form $\exp(cn^2)$ for the number of families that achieve this maximum size.

2 Dyadic APs and the key relation

A feature that makes our problem tractable (in addition to the fact mentioned above that a 3-term AP can have two numbers in common with at most six others) is that (as can be seen from (3)) two 3-term APs with two numbers in common either have the same common difference or else the common difference of one is exactly twice that of the other. This means that if we partition a set of 3-term APs into subsets according to the largest odd factor of the common difference then APs from different subsets never have two members in common, so that our problem is equivalent to finding the maximum size of an almost disjoint family within each subset. With this in mind, we call a 3-term AP whose common difference

is a power of 2 *dyadic* and call a collection of dyadic 3-term APs such that each two have at most one number in common an *almost disjoint dyadic family*.

Definition 1. *Let $A(n)$ be the set of all 3-term APs contained in $[0,n-1]$ and $A_2(n)$ the set of all dyadic 3-term APs contained in $[0,n-1]$. We denote by $F(n)$ the maximum size of any almost disjoint family in $A(n)$ and by $f(n)$ the maximum size of any almost disjoint family in $A_2(n)$. We also extend f to non-negative real values of its argument by linear interpolation between its values at integers.*

The size of $A(n)$ is given by (2), and for the total number of dyadic 3-term APs in $[0,n-1]$ we have

$$|A_2(n)| = n\log_2 n + O(n)$$

since this is the number of pairs of integers in $[0,n-1]$ whose difference is a positive power of 2. So we shall express our results for the size of $f(n)$ in terms of multiples of $n\log_2 n$.

The following proposition gives a key relationship between F and f (and is the reason why we interpolate f between integers).

Proposition 2. *For every positive integer n we have*

$$F(n) = f(n) + 3f(n/3) + 5f(n/5) + \cdots. \tag{4}$$

Proof. We note that the sum on the right is finite, since $f(x) = 0$ for $x \le 2$, and is an integer, since $f(n/m)$, as a linear interpolation between integer values at integers, has denominator a divisor of m.

We partition $A(n)$ as

$$A(n) = \bigcup_{m \text{ odd}} \bigcup_{a=0}^{m-1} A_{a,m}(n) \tag{5}$$

where $A_{a,m}(n)$ consists of those APs in $[0,n-1]$ whose common difference is a power of 2 times m and whose first term is congruent to a mod m. Then $A_{0,1}(n) = A_2(n)$ and subtracting a and dividing by m gives a one-one order-preserving affine map from $A_{a,m}(n)$ to $A_2(v(n,m,a))$, where $v(n,m,a)$ is the number of integers in $[0,n-1]$ congruent to a mod m. (So $v(n,m,a)$ is $\lceil n/m \rceil$ when $a < m\{n/m\}$ and $\lfloor n/m \rfloor$ when $a \ge m\{n/m\}$, where $\{x\}$ means the fractional part of x.) No two APs from different subsets $A_{a,m}$ intersect in two points because APs from sets with different m's have incompatible lengths and APs from sets with the same m but different a's lie entirely in separate residue classes mod m so do not intersect at all. Hence, any maximum-sized almost disjoint family in $A(n)$ is a union of maximum-sized almost disjoint families in the $A_{a,m}(n)$'s and

$$F(n) = \sum_{m \text{ odd}} \sum_{a=0}^{m-1} f(v(n,m,a))$$

Since

$$\sum_{a=0}^{m-1} v(n,m,a) = n \tag{6}$$

and f is defined between $\lfloor n/m \rfloor$ and $\lceil n/m \rceil$ by linear interpolation, the inner sum is $mf(n/m)$, giving (4). □

We end this section with some simply obtained bounds for $f(n)$ and $F(n)$ which (for $F(n)$) improve the bounds sketched in the introduction but which we shall further improve to asymptotic formulae later.

Proposition 3.

(i) $\frac{1}{2}n - 1 \leq f(n) < \frac{1}{3}(n\log_2 n)$.

(ii) $\frac{1}{3}(n^2/4) - O(n) < F(n) < \frac{2}{3}(n^2/4)$.

Proof. The lower bound in (i) is immediate, since the $\lceil n/2 \rceil - 1$ 3-term APs $\langle a; 1\rangle$ with a an even number in $[0, n-3]$ are almost disjoint. In fact, more generally, we clearly have $f(n+2) \geq f(n)+1$ for $n \geq 1$.

For the upper bound in (i) we count the number of pairs of numbers contained in all the 3-term APs of a family. A dyadic almost disjoint family of maximum size in $[0, n-1]$ contains $3f(n)$ distinct pairs of numbers whose difference is a power of 2. Since the total number of pairs in $[0, n-1]$ differing by a power of 2 is

$$\sum_{e=0}^{\lfloor \log_2 n \rfloor} (n - 2^e) < n(\log_2 n + 1) - n,$$

this gives (i).

The proof of the upper bound in (ii) is the same, but we drop the restriction that the pairs differ by a power of 2 and note that the total number of pairs in $[0, n-1]$ is $\binom{n}{2} < n^2/2$.

For the lower bound in (ii), let S be an almost disjoint family in $[0, k-1]$ of maximum size and consider what 3-term APs from $[0, k]$ can be added to it. If $k/2 \leq i < 2k/3$ then $\{2i-k, i, k\}$ can be added unless $\{2i-k, (3i-k)/2, i\} \in S$. But in that case $k-i$ is even and $\{(3i-k)/2, i, (k+i)/2\} \notin S$, so $\{i, (k+i)/2, k\}$ can be added. Hence for every $i \in [k/2, 2k/3)$ a 3-term AP containing $\{i, k\}$ can be added, and any two of these AP's meet only in k. So $f(k+1) - f(k) > k/6 - 1$ and summing from $k = 2$ to $n-1$ gives the lower bound in (ii). □

We note that the argument that gives the upper bound in (ii) makes no use of the fact that the triples are APs and, as a result, gives a bound coinciding with (1), valid for unrestricted partial 2-$(n, 3)$-designs.

For $f(n)$, the lower bound given by this proposition is not of the right order of magnitude — we shall see later (Theorem 9) that $f(n)$ is in fact asymptotic to $\frac{1}{3}n\log_2 n$. We have already mentioned that $F(n)$ is asymptotic to $C(n^2/4)$ with $C > 0.476$. It is at first sight paradoxical, in view of the direct relationship (4), that $f(n)$ and $F(n)$ should be asymptotic to different proportions of the sizes of $A_2(n)$ and $A(n)$: if, for each n, the maximum number of almost disjoint dyadic 3-term APs in $[0, n-1]$ is approximately 1/3 of the total number of dyadic APs and the set of all 3-term APs in $[0, n-1]$ is a disjoint union of sets in one-one correspondence with the set of dyadic APs in $[0, l-1]$ for some $l \leq n$, then how can significantly more than 1/3 of the 3-term APs in $[0, n-1]$ be almost disjoint? The answer is that a large and non-decreasing proportion of the numbers $l \approx n/m$ are small enough that $f(l)$ is not yet close to its limiting order of magnitude; put more concisely, small values of n/m make a major contribution to sum (4). A simpler example of this phenomenon is the set $P(n)$ of pairs of numbers in $[0, n-1]$, which decomposes into disjoint subsets each in one-one correspondence with a set of dyadic pairs (that is, pairs whose difference is a power of 2) in such a way that the parity of differences is preserved. As n tends to

infinity the proportion of dyadic pairs in $[0,n-1]$ with odd difference tends to 0. (There are dyadic pairs with odd difference, but only those with difference 1.) But for every n more than half the general pairs in $[0,n-1]$ have odd difference.

3 The asymptotic formula

The following lemma enables us to show that $F(n)$ is asymptotically a constant multiple of n^2.

Lemma 4. *Let $\phi(x)$ be any function that is defined for $x>0$, is linear between consecutive integer values of x, is 0 for $0<x\le 1$ and satisfies $\phi(x)=O(x\ln x)$ as $x\to\infty$. Then the function $\Phi(n)$, defined by*

$$\Phi(n)=\sum_{\substack{m=1\\ m\ odd}}^{\infty} m\phi(n/m) \tag{7}$$

satisfies

$$\Phi(n)=B(n^2/4)+O(n^{5/3}\ln n), \tag{8}$$

where

$$B=\sum_{k=2}^{\infty}\frac{2\phi(k)}{k(k^2-1)}. \tag{9}$$

If, further, $\phi(x+1)-\phi(x)=O(\ln x)$ as $x\to\infty$ then the error term in (8) can be decreased to $O(n^{3/2}\ln n)$.

Proof. Split sum (7) into ranges of the form

$$\frac{n}{k+1}<m\le\frac{n}{k}\qquad(k=1,2,3,\ldots), \tag{10}$$

in each of which $\phi(n/m)$ is linear. By the linearity,

$$\begin{aligned} m\phi(n/m) &= ((k+1)m-n)\phi(k)+(n-km)\phi(k+1)\\ &= ((k+1)\phi(k)-k\phi(k+1))m+(\phi(k+1)-\phi(k))n \end{aligned}$$

for m in range (10). Now summing over all odd m in a given range gives

$$\begin{aligned} &\frac{n^2}{4}((k+1)\phi(k)-k\phi(k+1))\left(\frac{1}{k^2}-\frac{1}{(k+1)^2}\right)\\ &\qquad+O\left(\frac{n}{k}((k+1)\phi(k)-k\phi(k+1))\right)\\ &\qquad+\frac{n^2}{2}(\phi(k+1)-\phi(k))\left(\frac{1}{k}-\frac{1}{k+1}\right)+O(n(\phi(k+1)-\phi(k)))\\ &=\frac{n^2}{4}\left(\frac{\phi(k)}{k^2(k+1)}+\frac{\phi(k+1)}{k(k+1)^2}\right)+O\left(n\left(\phi(k+1)-\phi(k)+\frac{\phi(k)}{k}\right)\right), \end{aligned} \tag{11}$$

where we have used the fact that the sum of an arithmetic progression is equal to the number of terms times the average of the end terms.

Finally, summing from $k=1$ to K gives

$$\Phi(n) = B(n^2/4) + O\left(\frac{n^2 \ln K}{K}\right) + O(nK^2 \ln K), \tag{12}$$

where the main term is the sum from 1 to infinity of the main term in (11), the first error term comes from the tail of the series from $K+1$ to infinity, and the second error term is the sum from 1 to K of the error term in (11). Now taking $K = \lfloor n^{1/3} \rfloor$ gives (8), where the fact that $\phi(1) = 0$ has been used to put B in form (9). When $\phi(x+1) - \phi(x) = O(\ln x)$ the second error term in (12) decreases to $O(nK \ln K)$ and taking $K = \lfloor n^{1/2} \rfloor$ gives an error term $O(n^{3/2} \ln n)$ in (8). □

Corollary 5. *$F(n) = C(n^2/4) + O(n^{3/2} \ln n)$, where*

$$C = \sum_{k=3}^{\infty} \frac{2f(k)}{k(k^2-1)} \tag{13}$$

Proof. Only the estimate $f(k+1) - f(k) = O(\ln k)$ remains to be checked, since this implies that $f(k) = O(k \ln k)$. This is immediate from the fact that there are only $\lfloor \log_2 k \rfloor$ dyadic 3-term APs in $[0,k]$ whose last term is k. □

For later use, we note that two 3-term APs with last term k whose common differences are consecutive powers of 2 are not almost disjoint, and hence that

$$f(k+1) - f(k) \le \lceil \lfloor \log_2 k \rfloor / 2 \rceil. \tag{14}$$

We mentioned earlier, in discussing relation (4), that $F(n)$ for large n, depends heavily on $f(k)$ for small values of k. Expression (13) for the constant C in the asymptotic formula for $F(n)$ shows this dependence very explicitly.

The value of C can be estimated from Proposition 3(i). For the lower bound, sum (13) with $f(k)$ replaced by $\lceil k/2 \rceil - 1$ can be explicitly summed to $1 - \ln 2 = 0.306\ldots$, weaker than the lower bound $C \ge \frac{1}{3}$ implied by the left hand inequality of Proposition 3(ii). For the upper bound, we define a function $\phi(k)$ for $k \ge 3$ recursively by

$$\phi(3) = 1, \qquad \phi(k+1) = \min(\lfloor ((e+1)(k+1) - 2^{e+1} + 1)/2 \rfloor, \phi(k) + \lceil e/2 \rceil),$$

where 2^e is the largest power of 2 that is $\le k$, the first term in the minimum comes from the proof of the upper bound in Proposition 3(i), and the second term in the minimum comes from (14). Now replacing $f(k)$ in (13) by $\phi(k)$ for $3 \le k \le 4097$ and using the upper bound $(12 + 1/\ln 2)/6144$ for the tail of the series from $k = 4098$ onwards, derived from comparison with

$$\int_{4096}^{\infty} \frac{2 \log_2 x dx}{3x^2}$$

gives $C < 0.506$.

We next improve these estimates for C by computing $f(k)$ exactly for some small values of k.

4 Particular values of $f(n)$

There are of the order of n^n families of dyadic 3-term APs in $[0,n-1]$, so it soon becomes infeasible to look at them all and select the maximum-sized disjoint families. The number of families we need to look at is vastly reduced if we know in advance the possible numbers $f_e(n)$ $(e=0,1,\ldots,\lceil\log_2 n\rceil-2)$ of APs with common difference 2^e in a maximum-sized almost disjoint family. These numbers satisfy

$$f_e(n)\le 2^e(\{n/2^e\}\lceil\lceil n/2^e\rceil l/2\rceil+(1-\{n/2^e\})\lceil\lfloor n/2^e\rfloor/2\rceil-1) \tag{15}$$

for $0\le e\le\lceil\log_2 n\rceil-2$, and

$$f_e(n)\le n-2^{e+1}-2f_{e+1}(n), \tag{16}$$

for $0\le e\le\lceil\log_2 n\rceil-3$. Though they look complicated in this notational form, these inequalities are based on two simple observations. The first is that the leading terms of the $n-2^{e+1}$ 3-term APs in $[0,n-1]$ with common difference 2^e fall into 2^e residue classes mod 2^e and that consecutive members of the same residue class cannot occur within an almost disjoint family. (This leads to (15).) The second observation is that each AP with common difference 2^{e+1} in an almost disjoint family excludes two APs with common difference 2^e and that the APs excluded by any two almost disjoint APs with common difference $2e+1$ are different. (This gives (16).) An upper bound for $f(n)$ is given by the maximum, over all vectors $(f_0(n),f_1(n),\ldots,f_l(n))$ satisfying (15) and (16), of the sum $f_0(n)+f_1(n)+\cdots+f_l(n)$. (Here $l=\lceil\log_2 n\rceil-2$.) Table 1 lists the vectors $(f_0(n),f_1(n),\ldots,f_l(n))$ with maximum sum for $n=3,\ldots,15$ and the number of almost disjoint families corresponding to each vector, computed (for the larger values of n) by a tree-searching algorithm. When there is no such family we list the vectors of successively smaller sums until we come to one for which there is an almost disjoint family. The first instance of this is $n=10$ and the first instance where one of the maximum sum vectors has no corresponding family is $n=9$. Beyond 15 the number of maximum sum vectors starts to get large, so for $n=16$–22, instead of listing the individual vectors, Table 1 simply lists the number of vectors with the maximum sum and with each successively smaller sum down to the largest for which there exists a corresponding almost disjoint family. The first instance of there being no family corresponding either to the maximum sum or to one less than the maximum sum is $n=17$.

For all values of n covered by the table $f(n)=g(n)$, where $g(n)$, defined in Definition 6 in the next section, is an easily computed function. This continues to hold at least up to $n=29$, as we have verified by using the methods of this section to check that there are no almost disjoint dyadic families of size $g(n)+1$. (By Theorem 7, $f(n)\ge g(n)$.) The extra values of f are

$$f(23)=26, f(24)=27, f(25)=29, f(26)=30$$

$$f(27)=32, f(28)=33, f(29)=34.$$

By using the exact values of f for $n\le 29$ in (13), we obtain the improved estimate

$$0.419<C<0.485$$

where for $n>29$ in the lower bound we have used the estimate $f(n)\ge f(n-2)+1$ (mentioned in the

Vectors with large sums satisfying (15) and (16) and the number of almost disjoint dyadic families corresponding to them

n	3	4	5	6	7
	(1) 1	(1) 2	(1,1) 1 (2,0) 1	(2,1) 2	(3,1) 1
n	8	9	10	11	12
	(2,2) 2 (3,1) 6	(3,2,1) 0 (4,1,1) 2	(4,2,2) 0 (3,2,2) 0 (4,1,2) 4 (4,1,2) 4 (4,2,1) 2	(5,1,3) 1 (5,2,2) 1	(5,2,3) 6
n	13	14	15	16	17
	(5,3,3) 3 (6,1,4) 0 (6,2,3) 4	(6,2,4) 3 (6,3,3) 6	(7,3,4) 0 (6,3,4) 0 (7,2,4) 0 (7,3,3) 4	14;2 0 13;5 222	17;1 0 16;5 0 15;13 60
n	18	19	20	21	22
	19;1 0 18;5 0 17;14 24	21;1 0 20;6 0 19;17 8	22;1 0 21;7 0 20;21 124	23;5 0 22;18 50	25;1 0 24;7 24

Table 1: For $n \geq 16$, $s;v\ f$ indicates a set of v vectors of sum s with a total of f corresponding families. For $n = 3,\ldots,15$, the vector listed last is the vector for the standard dyadic family, described in Section 5.

proof of Proposition 3) and in the upper bound we have estimated the tail of the series as in the previous section.

5 The standard families

To get a better lower bound for C we need a good lower bound for $f(k)$, which we find by identifying, for each k, a large "standard" almost disjoint family of 3-term APs. We shall see later that these "standard" families are those that are produced by certain greedy algorithms. They are maximum-sized for n up to 29 at least, and they enable us to improve the lower bound for $f(n)$ in Proposition 3(i).

Definition 6. *The* standard dyadic family $S_2(n)$ *in* $A_2(n)$ *is the family of 3-term dyadic APs* $\langle a;2^{e-1}\rangle$ *in* $[0,n-1]$ *such that the final* e *binary digits of* a *begin with a string of 0's of odd length. We write* $g(n) = |S_2(n)|$. *Again we extend the function* g *to* $\mathbb{R}^+$ *by linear interpolation between integers.*

Theorem 7. *The standard dyadic family is almost disjoint. Hence* $g(n) \leq f(n)$.

Proof. Let $\langle a;2^{e-1}\rangle \in S_2(n)$. The six 3-term APs that, according to (3), are candidates for having two numbers in common with this one are

$$\langle a;2^{e-2}\rangle, \langle a+2^{e-1};2^{e-2}\rangle, \langle a-2^{e-1};2^{e-1}\rangle, \langle a+2^{e-1};2^{e-1}\rangle, \langle a-2^e;2^e\rangle, \langle a;2^e\rangle$$

(The first two do not exist when $e = 1$, and some of the last four may be out of range, depending on the values of a, e and n.) It is easily checked that none of these is in $S_2(n)$: for the first two, a and $a+2^{e-1}$

agree with a in their last $e-1$ digits, which therefore begin with a string of 0's of even length (possibly empty); for the next two, the last e digits of $a \pm 2^{e-1}$ begin with a 1; and for the last two, $a-2^e$ and a agree with a in their last e digits so their last $e+1$ digits either begin with 1 or a string of 0's of even length. □

Since the dyadic family $S_2(n)$ is almost disjoint for all n so is the family

$$S(n) = \bigcup_{m \text{ odd}} \bigcup_{a=0}^{m-1} (mS_2(v(n,m,a)) + a) \subset \bigcup_{m \text{ odd}} \bigcup_{a=0}^{m-1} A_{a,m}(n) = A(n)$$

We call this simply the *standard family* in $[0, n-1]$ and denote its size by $|S(n)| = G(n)$. An alternative description of it is that it is the family of all 3-term APs $\langle a; 2^{e-1}m \rangle \in A(n)$ with m odd and the final e binary digits of $\lfloor a/m \rfloor$ beginning with a string of 0's of odd length. Clearly $G(n) \le F(n)$ and, by the same argument as in the proof of Proposition 2, we have

$$G(n) = g(n) + 3g(n/3) + 5g(n/5) + \cdots \tag{17}$$

In view of (17) and the fact, mentioned in the proof of Corollary 5, that there are only $\lfloor \log_2 k \rfloor$ dyadic 3-term APs in $[0,k]$ with last term k, the strong form of Lemma 4 is applicable with g and G in place of ϕ and Φ, and we have:

Theorem 8. *The function $G(n)$ satisfies*

$$G(n) = D(n^2/4) + O(n^{3/2} \ln n) \tag{18}$$

where

$$D = \sum_{k=3}^{\infty} \frac{2g(k)}{k(k^2-1)} \tag{19}$$

Since $G(n) \le F(n)$ for all n, $D \le C$. In the next section, we obtain a remarkable explicit formula for D that enables us to significantly improve our lower bound for C.

The standard dyadic families are large enough to give the correct asymptotic size of $f(n)$.

Theorem 9. *The function $f(n)$ satisfies*

$$f(n) = \frac{1}{3} n \log_2 n + O(n)$$

Proof. The upper bound has already been established in Proposition 3 and the lower bound comes from bounding below the size of $S_2(n)$. For a given e, the numbers a whose last e binary digits begin with a string of 0's of odd length have period 2^e and there are $(2^e - (-1)^e)/3$ of them per period, since this is the number of e-digit binary numbers with an odd number of leading 0's. All such a's in the range $0 \le a < n - 2^e$ give APs $\langle a; 2^e - 1 \rangle$ in $S_2(n)$, so the number of these is

$$\ge \frac{1}{3}(2^e - 1)(\lfloor n/2^e \rfloor - 1) > \frac{1}{3} n (1 - 2^{-e}) - \frac{1}{3} 2^{e+1}.$$

Now summing from $e = 1$ to $\lfloor \log_2 n \rfloor$ gives $f(n) > \frac{1}{3} n \log_2 n - O(n)$. □

6 The value of D

We now show how to obtain a remarkably simple expression for D as a sum involving values of the Riemann ζ-function. This makes it easy to approximate D extremely accurately and hence get a good numerical lower bound for C.

Let $g_e(k)$[1] be the number of APs in the standard dyadic family $S_2(k)$ with common difference 2^{e-1}. We shall calculate sum (19) by replacing $g(k)$ by $g_e(k)$ and then summing over e. Let χ_e, for $e \geq 1$, be the characteristic function of the set of non-negative integers whose final e binary digits begin with a string of 0's of odd length. Clearly χ_e has period 2^e, and in the range $[0, 2^e]$ it changes value only at powers of 2. By the definition of the standard dyadic family, $g_e(k)$ is the sum of $\chi_e(j)$ over $j \in [0, k-2^e)$, which by the periodicity of χ_e is the same as the sum over $j \in [2^e, k)$. Thus

$$\sum_{k=3}^{\infty} \frac{2g_e(k)}{k(k^2-1)} = \sum_{k=2^e}^{\infty} \frac{2}{k(k^2-1)} \sum_{j=2^e}^{k-1} \chi_e(j). \tag{20}$$

Since $1/k(k^2-1)$ is a second difference of the sequence $1/k$ and χ_e is constant over long ranges, the sum on the right can be greatly simplified by partial summation in the form

$$\sum_{k=K}^{\infty} (a_{k-1} - a_k) b_k = a_{K-1} b_K + \sum_{k=K}^{\infty} a_k (b_{k+1} - b_k).$$

With $a_k = 1/k(k+1)$ and $K = 2^e$ the right hand side of (20) becomes

$$\sum_{k=2^e}^{\infty} \frac{\chi_e}{k(k+1)}. \tag{21}$$

To carry out a second partial summation we write $k = 2^e q + r$, with $0 \leq r < 2^e$, and note that

$$\chi_e(k) - \chi_e(k-1) = \begin{cases} \bar{e} & \text{if } r = 0, \\ (-1)^{e-i} & \text{if } r = 2^i \text{ with } 0 \leq i < e, \\ 0 & \text{otherwise,} \end{cases}$$

where $\bar{e} \in \{0, 1\}$ is the residue of e modulo 2. Now partial summation with $a_k = \chi_e(k)$, $b_k = 1/k$ and $K = 2^e$ transforms (21) into

$$\sum_{q=1}^{\infty} \left(\frac{\bar{e}}{2^e q} + \sum_{i=0}^{e-1} \frac{(-1)^{e-i}}{2^e q + 2^i} \right) = \sum_{q=1}^{\infty} \sum_{i=0}^{e-1} \frac{(-1)^{e-i}}{2^i} \left(\frac{1}{2^{e-i} q + 1} - \frac{1}{2^{e-i} q} \right).$$

Summing over e from 1 to infinity gives an absolutely convergent double sum for D, which can be evaluated by making the substitution $e' = e - i$ and inverting the order of summation to get

$$D = \sum_{q=1}^{\infty} \sum_{i=0}^{\infty} \frac{1}{2^i} \sum_{e'=1}^{\infty} \frac{(-1)^{e'+1}}{2^{e'} q (2^{e'} q + 1)} = 2 \sum_{q=1}^{\infty} \sum_{e=1}^{\infty} \frac{(-1)^{e'+1}}{2^e q (2^e q + 1)}$$

[1] This differs from the analogous notation f_e in Section 4 in that the common difference was 2^e but here, in order to keep formulae short, we take it to be 2^{e-1}.

where at the final step we have renamed the variable e' as e.

In view of the somewhat artificial nature of our problem, this expression for D is remarkably simple, but the sum over q converges too slowly to make it easy to approximate D to any great accuracy. To remedy this we put the expression in an even more concise and amenable form by expanding each term as a power series in $1/2^e q$, evaluating the sum over e, and reversing the order of summation in the remaining variables. Explicitly

$$\begin{aligned} D &= 2\sum_{q=1}^{\infty}\sum_{e=1}^{\infty}(-1)^{e+1}\sum_{m=2}^{\infty}\left(\frac{-1}{2^e q}\right)^m \\ &= 2\sum_{q=1}^{\infty}\sum_{m=2}^{\infty}\frac{(-1)^m}{(2^m+1)q^m} = 2\sum_{m=2}^{\infty}\frac{(-1)^m\zeta(m)}{2^m+1} \end{aligned}$$

where $\zeta(m)=\sum_{m=1}^{\infty}1/q^m$ is the Riemann ζ-function. For even m, $\zeta(m)$ can be given explicitly in terms of π and the Bernoulli numbers, and for odd m there are ways of efficiently approximating it. The sum over m then converges geometrically, and is even alternating and decreasing in size, allowing simple error bounds and enabling D to be calculated to almost unlimited accuracy.

As a result of these last three sections we have

Theorem 10.

(i) $D = 2\sum_{m=2}^{\infty}\frac{(-1)^m\zeta(m)}{2^m+1} = 0.47621693\ldots$

(ii) $D \le C < 0.485$

It is of interest to relate the sum in (i) over values of the ζ-function to binary representations more directly than via the function g that measures the standard dyadic families. Keeping k fixed and summing $\chi_e(k)$ over values of e with $1 < 2^e \le k$ counts the number of 0's in the binary representation of k that are an odd number of places from the end of a string of 0's. This number is $\frac{1}{2}z(k)+\frac{1}{2}s(k)$, where $z(k)$ is the total number of 0's in the binary representation of K and $s(k)$ is the number of strings of 0's of odd length. So summing (21) over e, inverting the order of summation on the left and using our previous evaluation on the right, gives

$$\sum_{k=1}^{\infty}\frac{z(k)}{2k(k+1)}+\sum_{k=1}^{\infty}\frac{s(k)}{2k(k+1)} = D.$$

The first of the sums on the left straightforwardly evaluates to $1-\ln 2$, by partial summation, giving the value of the second sum on the left as $D+\ln 2-1$.

7 The number of maximum-sized families

We see from Table 1 that maximum-sized almost disjoint families of dyadic 3-term APs are far from unique in general, and as a consequence the same applies to unrestricted 3-term APs. In this section, we use Lemma 4 to obtain a reasonably sharp estimate for the number of maximum-sized almost disjoint families.

Definition 11. *Let $H(n)$ be the number of maximum-sized almost disjoint families in $A(n)$ and $h(n)$ the number of maximum-sized almost disjoint families in $A_2(n)$. We extend the function h to $\mathbb{R}^+$ by requiring that $h(x) = 1$ for $0 \le x \le 2$ (when the only almost disjoint family is the empty family!) and that $\ln h(x)$ is linear between integer values of x for $x \ge 2$.*

Because an almost disjoint family in $A(n)$ is maximum-sized if and only if its intersection with each of the parts $A_{a,m}(n)$ in partition (5) is maximum-sized we have

$$H(n) = \prod_{m \text{ odd}} \prod_{a=0}^{m-1} h(v(n,m,a)) = \prod_{m \text{ odd}} h(n/m)^m \tag{22}$$

where the last step follows from (6) and the linearity of $\ln h$ between integers. Since

$$|h(n)| \le 2^{|A_2(n)|},$$

Lemma 4 with $\phi(x) = \log_2 h(x)$ and $\Phi(n) = \log_2 H(n)$ gives

$$\log_2 H(n) = E(n^2/4) + O(n^{5/2} \ln n),$$

where

$$E = \sum_{k=3}^{\infty} \frac{2 \log_2 h(k)}{k(k^2-1)}. \tag{23}$$

Bounds for E can be calculated from the values of $h(k)$ for small k, which are implicit in Table 1. In Table 2, we list these values explicitly as far as we have calculated them. We have also collected in Table 2 the corresponding values of f (implicit in Table 1 too), of F (calculated from (4)) and of H (calculated from (22)).

We have

$$E \ge \sum_{k=3}^{22} \frac{2 \log_2 h(k)}{k(k^2-1)} = 0.102555\ldots$$

and

$$E \le \sum_{k=3}^{22} \frac{2 \log_2 h(k)}{k(k^2-1)} + \sum_{k=23}^{513} \frac{2 \log_2 \binom{a(k)}{b(k)}}{k(k^2-1)} + \frac{9 + 1/\ln 2}{256} - \frac{2}{257} < 0.447,$$

where

$$a(k) = \sum_{e=1}^{\lfloor \log_2 k \rfloor} (k - 2^e) \text{ and } b(k) = \left\lfloor \frac{1}{3} \sum_{e=0}^{\lfloor \log_2 k \rfloor} (k - 2^e) \right\rfloor$$

are upper bounds for $|A_2(k)|$ and $f(k)$ (derived from the proof of Proposition 3(i)), and for the tail of series (23) beyond 513 we have used the cruder estimate $\log_2 h(k) < |A_2(k)| \le k \log_2 k - 2k + 2$. A more intuitive way of describing these bounds is to say that the number of maximum-sized almost disjoint families of 3-term APs in $[0, n-1]$ is asymptotically greater than the 10th root of the total number of families of 3-term APs and asymptotically less than the square root of the total number of families.

We note that there is a unique maximum-sized almost disjoint family only for $n = 3$ and $n = 7$, since for $n \ge 10$ there is always an odd $m \ge 3$ with $3 < n/m < 7$ and then $h(n/m) > 1$.

n	$f(n)$	$F(n)$	$h(n)$	$H(n)$
3	1	1	1	1
4	1	1	2	2
5	2	2	2	2
6	3	3	2	2
7	4	5	1	1
8	4	6	8	8
9	6	9	2	2
10	7	10	6	12
11	9	13	2	8
12	10	15	6	48
13	11	18	7	56
14	12	21	9	72
15	13	25	4	32
16	13	27	222	3552
17	15	31	60	1920
18	17	35	24	1536
19	19	40	8	512
20	20	44	124	7936
21	22	50	50	1600
22	24	54	24	6144

Table 2: Values of $f(n), F(n), h(n)$ and $H(n)$

8 Greedy Algorithms

By a *greedy algorithm* for constructing an almost disjoint family we mean choosing an ordering of $A(n)$ (or $A_2(n)$) then inspecting the members of $A(n)$ or $A_2(n)$ in order, discarding only those that overlap in at least two places some member already inspected and not discarded. The following lemma enables us to see why several different greedy algorithms produce the same standard families.

Lemma 12. *If $\langle a; 2^{e-1}\rangle$ is a member of $A_2(n)$ that is not in $S_2(n)$ then at least one of $\langle a; 2^{e-2}\rangle$ and $\langle a-2^{e-1}; 2^{e-1}\rangle$ is in $S_2(n)$.*

Proof. Since $\langle a; 2^{e-1}\rangle \notin S_2(n)$ the last e binary digits of a begin with a string of 0's of even length. If this string of 0's is not empty then $e > 1$ and the last $e-1$ binary digits of a begin with a string of 0's of odd length, so $\langle a; 2^e - 2\rangle \in S_2(n)$. On the other hand, if the digit e from the end of a is 1 then $a \geq 2^{e-1}$ and the last e digits of $a - 2^{e-1}$ begin with a non-empty string of 0's. If this string has odd length then $\langle a-2^{e-1}; 2^{e-1}\rangle \in S_2(n)$ and if it has even length then $\langle a; 2^{e-2}\rangle \in S_2(n)$, since a and $a-2^{e-1}$ have the same last $e-1$ digits. □

Lemma 12 shows that any greedy algorithm for $A_2(n)$ that always lists $\langle a; 2^{e-1}\rangle$ before either of $\langle a+2^{e-1}; 2^{e-1}\rangle$ or $\langle a; 2^e\rangle$ produces precisely the family $S_2(n)$. Since $S_2(n)$ is almost disjoint no member of $S_2(n)$ will be rejected until at least one AP not in $S_2(n)$ has been retained. Suppose $\langle a; 2^{e-1}\rangle$ is the first AP not in $S_2(n)$ to be retained. By Lemma 12, at least one of $\langle a-2^{e-1}; 2^{e-1}\rangle$ and $\langle a; 2^{e-2}\rangle$ is in $S_2(n)$ so has already been retained. But both these APs have 2-point intersection with $\langle a; 2^{e-1}\rangle$, contrary to the supposition that $\langle a; 2^{e-1}\rangle$ can be retained.

Now expression (5) for $A(n)$ as a union of families, each the image of some $A_2(v)$ under an order-preserving affine map and with no APs of different families having 2-point intersection, shows that

the standard family $S(n)$ is produced by any greedy algorithm for $A(n)$ that lists $\langle a_1;d\rangle$ before $\langle a_2;d\rangle$ whenever $a_1 < a_2$ and $\langle a;d_1\rangle$ before $\langle a;d_2\rangle$ whenever $d_1 < d_2$. The following four greedy algorithms are all of this type:

1. Order the $\langle a;d\rangle$'s first on increasing a and then on increasing d.

2. Order the $\langle a;d\rangle$'s first on increasing d and then on increasing a.

3. Order the $\langle a;d\rangle$'s first on increasing $a+2d$ (the last term of $\langle a;d\rangle$), then on increasing a, then on increasing d.

4. Order the $\langle a;d\rangle$'s first on increasing $a+2d$, then on increasing d, then on decreasing a.

It was noticing that (1) and (2) always give the same family that led us to the description of the standard families. All greedy algorithms are equally efficient for a single value of n, but (3) and (4) are better when $S(n)$ is wanted for a range of values of n, since they find $S(3), S(4), S(5), \ldots$ successively on the way to finding $S(n)$. The proof of the lower bound in Proposition 3(ii) can be regarded as a partial analysis of the operation of algorithms (3) and (4).

Although these different algorithms produce the same standard families, we know that maximum-sized almost disjoint families are not unique for any $n > 7$. In particular, the standard family has mirror symmetry in the point $(n-1)/2$ only for $n = 3, 5$ or 7, as can be seen from the pattern of AP's with $d = 4$ (noting that the APs with $d = 3$ exclude the case $n = 11$). Any greedy algorithm that ordered on increasing d but decreasing a would find the mirror images of the standard families.

9 Summary

In this paper, we have established that the maximum size of an almost disjoint family of 3-term APs in $[0, n-1]$ is asymptotic to $C(n^2/4)$ for some constant C which we have estimated to within 1% (roughly, $0.476 < C < 0.485$).

We have also shown that the number of families achieving the maximum size is asymptotically larger than the 10th root of the total number of 3-term APs in $[0, n-1]$.

In the course of doing this we have stumbled across the following remarkable identity, where $s(k)$ is the number of strings of 0's in the binary representation of k that have odd length and ζ is the Riemann ζ-function:

$$\sum_{k=1}^{\infty} \frac{s(k)}{2k(k+1)} = \ln 2 - 1 + 2 \sum_{m=2}^{\infty} \frac{(-1)^m \zeta(m)}{2^m + 1}$$

References

[1] M. Deza, Paul Erdős, and P. Frankl, *Intersection properties of systems of finite sets*, Proc. London Math. Soc. **36** (1978), no. 3, 368–384.

[2] P. Frankl, *Extremal set systems*, Handbook of Combinatorics (R.L. Graham, L. Grötschel, and L. Lovász, eds.), Elsevier Science B.V., Amsterdam, MIT Press, Cambridge, MA, 1995, pp. 1293—-1329.

[3] Vojtek Rödl, *On a packing and covering problem*, European J. Combin. **6** (1985), 69–78.