JOURNAL OF NUMBER THEORY 9, 420–425 (1977)

Bases for Sets of Integers

P. ERDÖS AND D. J. NEWMAN

Department of Mathematics, Belfer Graduate School of Science, Yeshiva University, New York, New York 10033

Communicated by H. Zassenhaus

Received October 13, 1976

We are interested in expressing each of a given set of non-negative integers as the sum of two members of a second set, the second set to be chosen as economically as possible.

So let us call B a basis for A if to every $a \in A$ there exist $b, b' \in B$ such that $a = b + b'$. We concern ourselves primarily with finite sets, A, since the results for infinite sets generally follow from these by the familiar process of condensation.

TRIVIA

If then, we introduce the notation

n_A = number of elements of A,

N_A = largest element of A, and

m_A = minimum number of elements in a basis, B, of A,

we may make the following simple observations.

1. $m \leqslant n + 1$, this since the set $B = \{0\} \cup A$ is clearly a basis for A.
2. $m \leqslant (4N + 1)^{1/2}$.

We obtain this bound by choosing for B the integers $0, 1, 2, \ldots, k - 1$ together with the integers $k, 2k, \ldots, [N/k] \cdot k$. This is a basis for the whole interval $[0, N]$ and so surely for A itself. Also the number of elements in B is $k + [N/k]$ and since $\min_k(k + [N/k]) = [(4N + 1)^{1/2}]$ our result follows by choosing k appropriately.

3. $m \geqslant n^{1/2}$ (indeed $m \geqslant (2n + \frac{1}{4})^{1/2} - \frac{1}{2}$), for if B is a basis for A, having m elements, then the number of integers of the form $b + b'$, $b, b' \in B$, would have to be at least n. Since the number of couples (b, b') is at most m^2 (indeed $\binom{m+1}{2}$)) our results follow.

In summary, then, we have

THEOREM 1. $(n_A)^{1/2} \leqslant m_A \leqslant \min(n_A + 1, (4N_A + 1)^{1/2})$.

ISSN 0022-314X

Our main message is that the truth is "usually" nearer this upper bound than the lower one. As an example consider $A = \{3, 9, 27, \ldots, 3^n\}$, for B to be a basis we must have $b + b' = 3^k$, $k \leqslant n$, so that either b or b' lies in $[\frac{1}{2} \cdot 3^k, 3^k]$. Also $b + b' = 3$ implies that we must have an element of B in $[0, 1]$. These $n + 1$ intervals are disjoint, however, and so B has at least $n + 1$ elements. Hence $m = n + 1$.

"Most" Sets

In order to describe the situation for "most" sets we reverse our outlook by fixing numbers n and N and considering all those sets A for which $n_A = n$, $N_A = N$. We denote such sets as being of *type* (n, N) and we observe that the number of such is precisely $\binom{N}{n-1}$.

Next fix a number m and consider all those sets B for which $n_B = m$ and $N_B \leqslant N$. For each such B we form $B + B$, the set of all sums $b + b'$, b, $b' \in B$, and obtain thereby a set of at most m^2 distinct integers. Thus those A of type (n, N) for which $A \subseteq B + B$, i.e., for which B is a basis, number at most $\binom{m^2-1}{n-1}$. The number of such B, furthermore, is exactly $\binom{N+1}{m}$ and if we disallow those wasteful B which contain the number N but not the number 0 then this count diminishes to $\binom{N}{m} + \binom{N-1}{m-1} \leqslant 2\binom{N}{m}$.

Combining these results we obtain

4. Of all sets, A, of type (n, N) the fraction having $m_A \leqslant m$ is at most $\lambda = 2\binom{m^2-1}{n-1}\binom{N}{m}/\binom{N}{n-1}$.

As for this quantity λ we have

$$\lambda = 2 \,\lfloor\underline{m^2 - 1}\, \lfloor\underline{N - n + 1}/(\lfloor\underline{m^2 - n}\, \lfloor\underline{m}\, \lfloor\underline{N - m})$$

$$\leqslant (2m^{2\nu}/\lfloor\underline{m})(1/X^{\nu-m}),$$

where $\nu = n - 1$, $X = N - n + 1$. By the inequality $\lfloor\underline{m} \geqslant 2(m/e)^m$ we have, furthermore,

$$\lambda \leqslant m^{2\nu-m}((Xe)^m/X^\nu)$$

so that

5. $\log \lambda \leqslant \nu(2 + \log X) - (2\nu - m)(1 + \log X - \log m)$.

Now any choice of m which makes the right-hand side of 5 negative guarantees the existence of an A of type (n, N) with $m_A > m$. Also if the choice of m makes this right-hand side *large* negative then we are justified in saying that *most* sets of type (n, N) have $m_A > m$.

For example, consider the case $N = n^3$, and choose $m = [n/2]$. The right-hand side becomes essentially equal to $2n + 3n \log n - (\frac{3}{2})(1 + \log 2) n - 3n \log n = ((1 - 3 \log 2)/2)n$, and this is large negative. *Conclusion*

6. Most A of type (n, n^3) have $m_A > n/2$.

A similar calculation holds if we assume that $N \geqslant n^{2+\epsilon}$, $\epsilon > 0$. We then choose $m \approx (\epsilon/(1 + \epsilon))n$ and note, by monotonicity in N, that our expression is bounded by

$$n \log(n^{2+\epsilon}) - (2 - (\epsilon/(1 + \epsilon))n (\log(n^{2+\epsilon}) - \log(\epsilon n/(1 + \epsilon))$$
$$= -n((2 + \epsilon)/(1 + \epsilon)) \log((1 + \epsilon)/\epsilon).$$

Hence we have

7. If $N \geqslant n^{2+\epsilon}$, most A of type (n, N) satisfy $m_A > (\epsilon/(1 + \epsilon))n$.

For the general case we point out that the choice of $m = \min(n/\log N, N^{1/2}/2)$ always proves successful. Substituting this into 5, we obtain, namely, the bound

$$n \log N - (2n - (n/\log N)) (\log N - \log(N^{1/2}/2))$$
$$= (n/2) - n \cdot 2 \log 2 + n/\log N \leqslant (1 - 2 \log 2)n.$$

From this result and 7, we obtain

THEOREM 2. *Most sets A, of type (n, N) satisfy* $m_A > \min(n/\log N, N^{1/2}/2)$. *If furthermore, we have $N \geqslant n^{2+\epsilon}$, $\epsilon > 0$, then the $\log N$ may be replaced by* $(1 + \epsilon)/\epsilon$.

Certain observations present themselves. Note that when ϵ becomes very large this bound for m_A becomes very close to n (or $n + 1$) itself. In short:

8. If N grows faster than every power of n then most sets, A, of type (n, N) satisfy $m_A \sim n$.

Also observe that the only time that the lower bound in Theorem 2 is of a different order of magnitude than the upper bound in Theorem 1 is when N is of the order of n^2. Only sets with growth like the squares seem to present any real difficulty! It behooves us, therefore, to study the squares themselves.

THE SET OF THE SQUARES

We consider the set $A_0 = \{1^2, 2^2, ..., n^2\}$. Since we do not know that this set is in any way typical, Theorem 2 is not applicable and all we can use is Theorem 1 to conclude that $n^{1/2} \leqslant m_{A_0} \leqslant n + 1$.

Our purpose here is to narrow the gap between this upper and lower bound. Although we are far from closing this gap we derive the nontrivial bounds,

9. $n^{2/3-\epsilon} \leqslant m_{A_0} \leqslant n/\underset{\log n}{M}$, ϵ arbitrarily small, M arbitrarily large.

This upper bound definitely shows that the set of squares is not typical, for most sets of type (n, n^2) satisfy $m_A > n/2 \log n$, by Theorem 2 (and in fact this can be improved to $m_A > c\, n(\log\log n/\log n)$ while $m_{A_0} < n/\log^2 n$ (for example).

To derive our upper bound recall that, for each odd prime, p, the squares fall into precisely $(p + 1)/2$ residue classes (mod p). Hence if $p, q, r,\ldots$ are distinct odd primes and $P = p \cdot q \cdot r \cdots$ the Chinese remainder theorem tells us that the squares fall into precisely $(p + 1)/2 \cdot (q + 1)/2 \cdot (r + 1)/2 \cdots$ residue classes (mod P). A basis for the squares is obtained, then, by choosing these reduced residues (i.e., in $[0, P)$) together with all the multiples of P. Hence we have

$$m_{A_0} \leqslant ((p + 1)/2)((q + 1)/2) \cdots + (n^2/p \cdot q \cdot r \cdots) + 1,$$

for any distinct odd primes, $p, q, r,\ldots$.

If $p_1 < p_2 < \cdots$ denote all the odd primes below $2 \log n$ then we know, from prime number theory, that for any fixed M, $p_1 \cdot p_2 \cdots > n \log^{M+3} n$. Thus we may pick k so that

$$2n \log^{M+2} n > p_1 p_2 \cdots p_k > n \log^{M+1} n,$$

and we automatically have $(2 \log n)^k > n \log^M n$, so that $k > \log n/\log\log n$. Using these primes as our $p, q, r,\ldots$ and observing that $(p_i + 1)/2p_i \leqslant \frac{2}{3}$ we obtain

$$m_{A_0} \leqslant 2n \log^{M+2} n(\tfrac{2}{3})^{\log n/\log\log n} + (n^2/n \log^{M+1} n) + 1$$
$$\leqslant (n/\log^M n) \qquad \text{for large } n.$$

This trick can be used with some success for other sequences which, like the squares, fall into a limited number of residue classes (mod p); thus for example if A is the set of primes below x then we produce thereby a basis of size $O(x/\log\log x)^{1/2}$. Compare this to the lower bound (Theorem 1) which is $(x/\log x)^{1/2}$.

We obtain our lower bound as an immediate corollary to the following theorem (since the number of solutions to $x^2 - y^2 = k$ is known to be $O(k^\epsilon)$ for every ϵ).

DEFINITION. D_A is the maximum number of ways in which a positive integer can be written as the difference of two elements of A.

THEOREM 3. $m_A > n_A^{2/3}(D_A + 1)^{-1/3}$.

Proof. Let B be a minimum size basis for A and order the elements of B as follows: b_1 is the element involved in the least number, V_1, of representations for A, b_2 is then chosen as the element involved in the least number, V_2, of *new* representations for A (i.e., ones not involving b_1); b_3 is then chosen as the one involved in the least number, V_3, of representations not involving b_1 and b_2, etc.

Now fix i and consider the ordered couples (j, k), $j \geqslant i$, $k > i$ such that $b_i + b_j \in A$, $b_j + b_k \in A$. First of all, for fixed j, there are at least $V_i - 1$ such k and since there are exactly V_i of these j the couples number at least $V_i(V_i - 1)$. On the other hand for fixed k each j leads to the representation $(b_i + b_j) - (b_j + b_k)$ of the nonzero number $b_i - b_k$ as a difference of two members of A. Thus for each fixed k there can be at most D couples and since the number of k is less than m there are less than mD couples.

Hence $V_i(V_i - 1) < mD$, but we also know that $\sum_{i=1}^{m} V_i \geqslant n$ (since all of A is represented) and combining these inequalities shows that $(n/m)((n/m) - 1) < mD$. Thus $D > (n^2/m^3) - (n/m^2)$ and since this is $\geqslant (n^2/m^3) - 1$, by 3, our theorem follows.

It is interesting to note that Theorem 3 is, in a very strong sense, best possible. Indeed by Theorem 1 the inequality is trivial when $D \geqslant n^{1/2}$ and so we consider only numbers D and n such that $D < n^{1/2}$. For any such pair of numbers we construct an example of an A for which $D_A \leqslant D$, $n_A \geqslant n$, and $m_A \leqslant 7n^{2/3}D^{-1/3}$.

We proceed as follows: Denote $I = \{1, 2, \ldots, k\}$, $J = \{k + 1, k + 2, \ldots, 2k\}$, and to each $i \in I$ choose, at random (each element independently and with probability α), a subset $J_i \subseteq J$. The expected number of elements in J_i is $k\alpha$ and in $J_i \cap J_{i'}$ is $k\alpha^2$. A slight calculation shows in fact that, with positive probability,

(a) each J_i has at least $k\alpha/2$ elements,

(b) each $J_i \cap J_{i'}$, $i \neq i'$, has at most $2k\alpha^2$ elements,

(c) each pair $j, j'(j \neq j')$ lies in at most $2k\alpha^2$ sets J_i.

We pick such an arrangement. Next we choose numbers $b_1, b_2, \ldots, b_{2k}$ such that

(d) The sums taken 4 at a time, $b_i + b_j + b_k + b_l$, are all distinct up to permutations (for example we can pick $b_i = 4^i$).

So B is chosen (with $2k$ elements) and we pick A as the set of all $b_i + b_j$, $i \leqslant k$, $j \in J_i$ and note that $n_A \geqslant k^2\alpha/2$ (by (a)). As B is clearly a basis for A we have $m_A \leqslant 2k$. Finally we estimate D_A. Namely, for two numbers of the form $b_i + b_j - (b_{i'} + b_{j'})$ to be equal (d) ensures that they must have either the same j and j' and $i = i'$ or the same i and i' and $j = j'$. By (b) and (c) above, then, there can only be at most $2k\alpha^2$ such coincidences, and in short we have $D_A \leqslant 2k\alpha^2$.

It is a simple matter, for given n and D with $D \leqslant n^{1/2}$, to make $k^2\alpha/2 \geqslant n$ and $2k\alpha^2 \leqslant D$. Choose $\alpha = D^{2/3}/3n^{1/3}$.

Noting that the interval $[\frac{5}{2}(n^{2/3}/D^{1/3}), \frac{7}{2}(n^{2/3}/D^{1/3})]$ has length at least 1 we can choose a k lying in it. This choice of k and α then works and indeed it gives $m_A \leqslant 2k \leqslant 7n^{2/3}D^{-1/3}$ as required.

Discontinuity

Finally we wish to point out that the size of m_A depends rather delicately on the arithmetical structure of the sequence A and not just on the coarse aspects of its "rate of growth." The fact is that to every set, A, there is a fairly nearby set, A', which has a relatively small basis. This perturbed set is produced by choosing a large K and then replacing the *larger* members of A by their closest multiples of K, while leaving the smaller ones fixed. Thus A' has changed the elements of A by a relatively negligible amount and yet A' has for a basis the following (small) set: 0, the unmoved elements of A, and a basis for the set of all multiples of K up to N_A. (Indeed by 2 the multiples of K up to N_A have a basis of size only $((4N_A/K)) + 1)^{1/2}$.

To give an example of such a phenomenon consider a randomly chosen set of type (n, n^2). An elementary probability computation shows that *usually* with at most $n^{3/4}$ exceptions the gap between elements is at least $n^{2/3}$. We take as A such a set which at the same time is typical according to Theorem 2. Thus $m_A \geqslant n/2 \log n$. For A' we take the aforementioned $n^{3/4}$ exceptions together with the nearest multiples of $K = [n^{1/2}]$ to the other members. Thus A' is very near to A and yet, as previously indicated, $m_{A'} \leqslant 1 + n^{3/4} + (4n^{3/2} + 1)^{1/2} \leqslant 5n^{3/4}$.

In view of this discontinuous behavior of m as a function of A it seems difficult to even *guess* the size of m for a specific A. For example, what is the size of m for the cubes, $\{1^3, 2^3, ..., n^3\}$? If they were typical the answer would be cn: the squares are atypical, however, and so perhaps the cubes are also. We are unable to decide.

Another question which seems interesting and difficult is whether *any* set of type (n, n^2) needs cn elements in its basis. In short let $M_n = \max_A m_A$, taken over all A of type (n, n^2), is $M_n = o(n)$?