Acta Mathematica Academiae Scientiarum Hungaricae
Tomus 20 (1—2), (1969), pp. 89—104.

ON SETS OF INTEGERS CONTAINING NO FOUR ELEMENTS IN ARITHMETIC PROGRESSION

By
E. SZEMERÉDI (Budapest)

In what follows we use capital letters to denote sequences of integers, $A+B$ to denote the sum of two sets of integers formed elementwise, and $A \neg B$ to denote the complement of the set B with respect to the set A.

Let us for convenience call an arithmetic progression of k (distinct) terms a *k-progression.*

If a set A contains no k-progression we say that A is k-free.

The maximal number of elements a k-free set $A \subseteq [0, n)$ can have is denoted by $\tau_k(n)$. Furthermore we set

$$\gamma_k = \overline{\lim_{n\to\infty}} \frac{\tau_k(n)}{n}.$$

Actually we can replace $\overline{\lim}$ on the right hand side by lim. For, given $\varepsilon > 0$ and n, we can find arbitrarily large m so that $\tau_k(m) \geqq (\gamma_k - \varepsilon)m$; in particular we may assume that $qn < m \leqq (q+1)n$ holds for a positive integer q. In other words there is a k-free set $A \subseteq [0, m)$ with cardinality $|A| \geqq (\gamma_k - \varepsilon)m$. Now $[0, m)$ can be split into $(q+1)$ subintervals of length at most n. One of these must contain at least $\left(\frac{1}{q+1}\right)|A|$ elements of A which clearly form a k-free set.

Hence

$$\tau_k(n) \geqq \left(\frac{1}{q+1}\right)|A| \geqq (\gamma_k - \varepsilon)\frac{m}{q+1} \geqq (\gamma_k - \varepsilon)\frac{q}{q+1} n.$$

Since ε can be taken arbitrarily small and q arbitrarily large, we have

$$\tau_k(n) \geqq \gamma_k n,$$

whence

$$\gamma_k = \lim \frac{\tau_k(n)}{n}.$$

Clearly $\gamma_k \leqq 1 - \frac{1}{k}$, and $\gamma_3 \leqq \gamma_4 \leqq \dots$. It has been proved by F. BEHREND* that either all γ_k are zero, or $\gamma_k \to 1$ as $k \to \infty$.

* On sequences of integers containing no arithmetic progression, *Časopis Mat. Fis. Praha*, **67** (1938), pp. 235—239.

In 1953 ROTH* proved that $\gamma_3=0$. In fact he proved more than that, namely

$$\tau_3(n) \ll \frac{n}{\log\log n}\,.$$

Roth's proof uses estimates of exponential sums.

In this paper we shall prove the following

THEOREM.

$$\gamma_4=0, \quad i.e. \quad r_4(n)=o(n).$$

The proof is elementary. The problem of $\gamma_5, \gamma_6, \ldots$ is left open.

The proof is indirect, so from now on we assume that

$$\gamma_4>0.$$

For convenience we write

$$\gamma=\gamma_4.$$

We shall formulate in this section the two main lemmas and deduce the theorem from them.

We write $Q(b, c, d, e)$ for the system

$$b-2c+d = c-2d+e = 0,$$

which means that either b, c, d, e form an arithmetic progression, or they are identical.

Throughout the paper $n_4(\varepsilon)$ shall mean a number (for example the smallest one) with the property that for $n \geqq n_4(\varepsilon)$ a 4-free set $A \subseteqq [0, n)$ cannot contain more than $(\gamma+\varepsilon)n$ elements. Occasionally we use the analogue meaning for $n_3(\varepsilon)$ as well.

Let $B, C, D \subseteqq [0, q)$. We regard B and C as fixed while D varies. We then define

$$D^* = \{e;\ e\in[0, q) \text{ and there are } b\in B,\ c\in C,\ d\in D \text{ such that } Q(b, c, d, e)\}.$$

With this notation we shall prove

LEMMA $(H_0, \ldots, H_k)$.** *There are absolute constants* $\varepsilon_0>0$, $\gamma'>0$, k_0 *and* q_0 *with the following property: If*

$$q \geqq q_0, \quad 3|q,$$

and if B, C *are* 4*-free sets contained in* $[0, q)$, $|B| \geqq (\gamma-\varepsilon_0)q$, $|C| \geqq (\gamma-\varepsilon_0)q$, *then there are disjoint sets*

$$H_0, \ldots, H_k, \qquad k \leqq k_0,$$

such that

$$\bigcup_{K=0}^{k} H_K = \left[\frac{1}{3}q, \frac{2}{3}q\right),$$

$$|H_0| \leqq \frac{1}{12}\gamma q; \qquad |H_K^*| \geqq \gamma' q \quad \textit{for} \quad K = 1, 2, \ldots, k,$$

* On certain sets of integers. I; II, *J. Lond. Math. Soc.*, **28** (1953), pp. 104—109; **29** (1953), pp. 20—26.

** The full force of the hypothesis that (say) C is 4-free is not needed for the proof of this lemma: see the footnote on page 95.

and such that if for some $K \neq 0$

$$G \subseteqq H_K, \qquad |G| \geqq \frac{1}{2}\gamma |H_K|,$$

then

$$|G^*| \geqq \left(1 - \frac{1}{2}\gamma\right) |H_K^*|.$$

The other main lemma is

LEMMA *BCDE. Let* $\varepsilon_1 \in (0, \gamma)$ u *and* q_0 *be given. Then there is a* $q \geqq q_0$ *and there are sets*

$$B_0, C_0, D_1, \ldots, D_u, E_1, \ldots, E_u \subseteqq [0, q),$$

all 4*-free, all with at least* $(\gamma - \varepsilon_1)q$ *elements, such that* $Q(b, c, d, e)$ *with* $b \in B_0$, $c \in C_0$, $d \in D_i$, $e \in E_i$ *is insolvable for all* $i = 1, \ldots, u$, *and such that for each* $x \in [0, q)$ *the set of all i's for which* $x \in E_i$ *holds is* 4*-free.*

We now prove the theorem using these two lemmas.

Let ε_0, γ and k_0 have the meaning of lemma $(H_0, \ldots, H_k)$. Put

$$\varepsilon_1 = \min\left(\varepsilon_0, \frac{\gamma}{20}, \frac{\gamma\gamma'}{6}\right)$$

and

$$t = n_4(\varepsilon_1).$$

Van der Waerden's Theorem* gives a number

$$u = N(k_0, t)$$

such that in any partition of $[0, u)$ into at most k_0 classes there is at least one class which contains a t-progresssion.

We apply lemma *BCDE* with this ε_1, and u, and with

$$q_0 = 3n_4(\varepsilon_1).$$

From $|D_i| \geqq (\gamma - \varepsilon_1)q, \frac{1}{3}q \geqq n_4(\varepsilon_1)$ we see that

$$\left|D_i \cap \left[\frac{1}{3}q, \frac{2}{3}q\right)\right| = |D_i| - \left|D_i \cap \left[0, \frac{1}{3}\right)\right| - \left|D_i \cap \left[\frac{2}{3}q, q\right)\right| \geqq$$

$$\geqq (\gamma - \varepsilon_1)q - 2(\gamma + \varepsilon_1)\frac{1}{3}q \geqq (\gamma - 5\varepsilon_1)\frac{1}{3}q.$$

We now define the sets H_K by lemma $(H_0, \ldots, H_k)$, using B_0, C_0 for B, C respectively.

For each $i \in (0, u]$ there is a $j = j(i) \in (0, k]$ such that

$$|D_i \cap H_j| \geqq \frac{1}{2}\gamma |H_j|.$$

* Beweis einer Baudetschen Vermutung, *Nienn. Arch. Wiskunde*, **15** (1927), pp. 212—216.

For otherwise we should get the contradiction

$$(\gamma-5\varepsilon_1)\frac{1}{3}q \leqq \left|D_i\cap\left[\frac{1}{3}q,\ \frac{2}{3}q\right)\right| = \sum_{j=0}^{k}|D_i\cap H_j| <$$
$$< |H_0|+\frac{1}{2}\gamma\sum_{j=1}^{k}|H_j| \leqq \left(\frac{1}{4}\gamma+\frac{1}{2}\gamma\right)\frac{1}{3}q \leqq (\gamma-5\varepsilon_1)\frac{1}{3}q$$

since $\varepsilon_1 \leqq \frac{1}{20}\gamma$.

Attaching such a $j(i)$ to each i, it gives a partition of the i's into k classes. Since $u=N(k_0, t)$ and $k\leqq k_0$ one of these classes contains a t-progression. In other words, there is a j_0 and an arithmetic progression $i_1, \ldots, i_t$ such that

$$|D_i\cap H_{j_0}| \geqq \frac{1}{2}g|H_{j_0}| \quad \text{for} \quad i=i_1, \ldots, i_t.$$

From lemma $(H_0, \ldots, H_k)$ we then have that

$$|(D_i\cap H_{j_0})^*| \geqq \left(1-\frac{1}{2}\gamma\right)|H_{j_0}^*|$$

where the $*$ is taken with respect to B_0 and C_0. With the trivial relation $(U\cap V)^*\subseteqq U^*\cap V^*$ this implies that

$$|D_i^*\cap H_{j_0}^*| \geqq \left(1-\frac{1}{2}\gamma\right)|H_{j_0}^*|.$$

Now $D_i^*\cap E_i=\emptyset$, for this is merely a restatement of the fact that the relations $Q(b, c, d, e)$ with $b\in B_0$, $c\in C_0$, $d\in D_i$, $e\in E_i$ are impossible.

Hence

$$|E_i\cap H_{j_0}^*|+|D_i^*\cap H_{j_0}^*| \leqq |H_{j_0}^*|,$$

so that

$$|E_i\cap H_{j_0}^*| \leqq \frac{1}{2}\gamma\,|H_{j_0}^*|$$

for $i=i_1, \ldots, i_t$.

Put

$$|H_{j_0}^*|=\alpha\cdot q, \quad [0, q)-H_{j_0}^*=M.$$

We notice that M is not empty, since otherwise the last inequality would imply that $|E_i|\leqq\frac{1}{2}\gamma q$, in contradiction with the fact that

$$|E_i| \geqq (\gamma-\varepsilon_1)q \geqq \left(\gamma-\frac{1}{20}\gamma\right)q.$$

Furthermore, lemma $(H_0, \ldots, H_k)$ shows that $\alpha\geqq\gamma'$. Therefore

$$\frac{|E_i\cap M|}{|M|} = \frac{|E_i|-|E_i\cap H_{j_0}^*|}{q-|H_{j_0}^*|} \geqq \frac{\gamma-\varepsilon_1-\frac{1}{2}\gamma\alpha}{1-\alpha} = \gamma+\frac{\frac{1}{2}\gamma\alpha-\varepsilon_1}{1-\alpha} \geqq$$
$$\geqq \gamma+\frac{1}{2}\gamma\alpha-\varepsilon_1 \geqq \gamma+\frac{1}{2}\gamma\gamma'-\varepsilon_1 \geqq \gamma+2\varepsilon_1$$

for $i=i_1, \ldots, i_t$. Summing over these i's we see that

$$\sum_{\tau=1}^{t} |E_{i_\tau} \cap M| \geqq (\gamma+2\varepsilon_1)t|M|.$$

We conclude that there is at least one $x \in M$ which occurs in not less than $(\gamma+2\varepsilon_1)t$ of the sets E_{i_τ}. By lemma *BCDE* those i_τ's for which $x \in E_{i_\tau}$ form a 4-free set. They are contained in an arithmetic progression of t terms and by the choice of $t=n_4(\varepsilon_1)$, there cannot be more than $(\gamma+\varepsilon_1)t$ numbers i_τ for which $x \in E_{i_\tau}$. Thus we have reached a contradiction and the theorem is proved.

In this section we shall prove lemma $(H_0, \ldots, H_k)$. For this we need three other lemmas. The first is almost obvious. We call it therefore

THE SIMPLE LEMMA. *Let $A \subseteqq [0, n)$ be 4-free and $|A| \geqq (\gamma-\varepsilon)n$. Let $M \subseteqq [0, n)$ have a complement that is the union of disjoint arithmetic progressions P_ϱ, $\varrho=1, \ldots, r$ each of length $|P_\varrho| \geqq n_4(\varepsilon')$. Then we have*

$$|A \cap M| \geqq \gamma|M|-(\varepsilon+\varepsilon')n.$$

PROOF. Each $A \cap P_\varrho$ as a 4-free subset of a progression fulfils

$$|A \cap P_\varrho| \leqq (\gamma+\varepsilon')|P_\varrho|.$$

Hence we have the following inequalities:

$$|A \cap M| = |A|-\sum_\varrho |A \cap P_\varrho| \geqq (\gamma-\varepsilon)n-(\gamma+\varepsilon')\sum_\varrho |P_\varrho| =$$

$$= (\gamma-\varepsilon)n-(\gamma+\varepsilon')(n-|M|) = (\gamma+\varepsilon')|M|-(\varepsilon+\varepsilon')n \geqq \gamma|M|-(\varepsilon+\varepsilon')n.$$

LEMMA $p(\delta, l)$. *For any real $\delta \in (0, 1)$ and any natural number l there exists a number $p(\delta, l)$ with the following property: If*

$$u \geqq p(\delta, l), \quad G \subseteqq [0, u), \quad |G| \geqq \delta u,$$

then G contains a set S_l of the form

$$S_l=\{y\}+\{0, x_1\}+\ldots+\{0, x_l\}$$

with natural numbers $x_1, \ldots, x_l$.

PROOF. The proof goes by complete induction and uses the box principle. The case $l=1$ is trivial, since it states only that there is a pair of elements of G. A suitable choice of $p(\delta, 1)$ is $\left[1+\frac{1}{\delta}\right]$ since this exceeds $\frac{1}{\delta}$ so that the hypothesis concerning G shows that

$$|G| \geqq \delta u > 1.$$

Now take $l \geqq 2$ and assume the case $l-1$ has been already proved. We set

$$q = p\left(\frac{\delta}{2}, l-1\right).$$

Any number u can be represented as

$$u = kq+r, \quad 0 \leqq r < q.$$

We choose $p(\delta, l)$ so that $u \geqq p(\delta, l)$ implies that

$$k > \frac{4}{\delta^2}, \quad \frac{\delta}{2} k > (q-1)^{l-1}.$$

A possible choice is, for example

$$p(\delta, l) = \max\left(\left[1+\frac{4}{\delta^2}\right] q, \ \left[1+\frac{2}{\delta}\right] q^l\right).$$

Let R be the number of those sets

$$G_K = G \cap [(K-1)q, Kq], \qquad K = 1, \dots, k$$

for which $|G_K| \geqq \frac{\delta}{2} q$. Then $R \geqq \frac{\delta}{2} k$, otherwise

$$\delta kq \leqq \delta u \leqq |G| \leqq q + \sum_{K=1}^{k} |G_K| \leqq (1+R)q + (k-R)\frac{\delta}{2} q =$$

$$= \left(1-\frac{\delta}{2}\right) Rq + \left(1+\frac{k\delta}{2}\right) q < \left(1-\frac{\delta}{2}\right)\frac{\delta}{2} kq + \left(1+\frac{k\delta}{2}\right) q =$$

$$= \delta kq - \left(\frac{\delta^2 k}{4} - 1\right) q < \delta kq.$$

By the introduction hypothesis, in each of the sets G_K a set of the type S_{l-1} can be found. In each S_{l-1} we have $1 \leqq x_1, \dots, x_{l-1} \leqq q-1$. Thus there are not more than $(q-1)^{l-1}$ different choices of $x_1, \dots, x_l$. Since $R \geqq \frac{\delta}{2} k > (q-1)^{l-1}$ there are two sets G_K containing S_{l-1} and S'_{l-1} formed with the same numbers $x_1, \dots, x_l$ but different y, y', say with $y' > y$. Then with $x_l = y' - y$ we have

$$G \supseteqq S_{l-1} \cup S'_{l-1} = S_{l-1} \cup (S_{l-1} + x_l) = S_l.$$

LEMMA $|G^*|$. *There are absolute constants* $\varepsilon_0 > 0$ *and* $\gamma' > 0$ *and a function* $g_0(\delta)$ *for* $0 < \delta < 1$ *with the following property:*
If $q \geqq q_0(\delta)$, $8|q$, $B, C \subseteqq [0, q)$ *are both 4-free,*

$$|B| \geqq (\gamma - \varepsilon_0) q, \quad |C| \geqq (\gamma - \varepsilon_0) q, \quad G \subseteqq \left[\frac{1}{3} q, \frac{2}{3} q\right) \quad |G| \geqq \frac{\delta q}{3},$$

then

$$|G^*| \geqq \gamma' q.$$

REMARK. An analogous lemma can be similarly proved with $\gamma = \gamma_3$ (instead of $\gamma = \gamma_4$) on the assumption that $\gamma_3 > 0$. We then easily arrive at a contradiction, which proves Roth's theorem $\gamma_3 = 0$. For this purpose choose a $q \geqq 3n_3(\varepsilon)$. Next choose a 3-free set $A \subseteqq [0, 3q)$ with $|A| \geqq 3\gamma q$ and represent it as

$$A = B \cup (C+q) \cup (D+2q)$$

with $B, C, D \subseteqq [0, q)$; and finally set

$$G = D \cap \left[\frac{1}{3} q, \frac{2}{3} q\right).$$

One easily obtains the inequalities $|B| \geqq (\gamma - 2\varepsilon)g$, $|C| \geqq (\gamma - 2\varepsilon)q$, $|G| \geqq (\gamma - 8\varepsilon)\frac{q}{3}$. If we take $\varepsilon \leqq \frac{1}{2}\varepsilon_0$, $\varepsilon \leqq \frac{1}{16}\gamma$ and q large enough, we can apply the lemma with $\delta = \frac{1}{2}\gamma$ and get

$$|G^*| \geqq \gamma' q > 0$$

which means that there is a triplet (b, c, d) with

$$b - 2c + d = 0.$$

But $(b, c+q, d+2q)$ is then a 3-progression in A, a set that was supposed to be 3-free.

PROOF OF LEMMA $|G^*|$. Set

$$\varepsilon_0 = \frac{1}{100}\gamma^2, \quad m = n_4(\varepsilon_0),$$

and fix an l such that $l \geqq 24\frac{m}{\gamma}$, say

$$l = \left[\frac{25m}{\gamma}\right].$$

We shall prove the lemma with

$$q_0(\delta) = 3p(\delta, l) + 3m, \quad \gamma = \frac{\gamma^2}{50 \cdot 2^l}.$$

With these choices we have $\frac{q}{3} \geqq p(\delta, l)$ and can therefore find a set of type S_l in G. We consider

$$S_i = \{y\} + \{0, x_1\} + \ldots + \{0, x_i\}$$

for all $i = 0, 1, \ldots, l$; where we take $S_0 = \{y\}$. For each i we define

$$L_i = \left\{2c - s;\quad c \in C \cap \left[\frac{1}{3} q, \frac{2}{3} q\right),\ s \in S_i\right\}.$$

Since $S_i \subseteqq \left[\frac{1}{3} q, \frac{2}{3} q\right)$ one has $L_i \subseteqq [0, q)$.

With $|C| \geqq (\gamma - \varepsilon_0)q$ and $\frac{1}{3} q > m = n_4(\varepsilon_0)$ we obtain

$$|L_0| = \left|C \cap \left[\frac{1}{3} q, \frac{2}{3} q\right)\right| \geqq (\gamma - 5\varepsilon_0)\frac{q}{3} \geqq \frac{1}{4}\gamma q,^*$$

since $5\varepsilon_0 < \frac{1}{4}\gamma$.

* The derivation of this inequality is the only extent to which we use the hypothesis that C is 4-free.

From the fact that $|L_l| \leqq q$ and $L_0 \subseteqq L_1 \subseteqq \dots$ we infer that there is some $i \leqq l$ such that

$$|L_i| - |L_{i-1}| \leqq \frac{q}{l}.$$

We decompose this L_{i-1} into maximal progression (mod x_i). We shall denote by $\bar{L}$ the union of those of these progressions which have $3m$ or more elements, and by $\bar{\bar{L}}$ the union of the remaining ones. From

$$S_i = S_{i-1} \cup (S_{i-1} + x_i)$$

one sees that

$$L_i = L_{i-1} \cup (L_{i-1} - x_i).$$

Each maximal progression (mod x_i) in L_{i-1} produces therefore one and only one new element in L_i. Hence

$$|\bar{\bar{L}}| \leqq 3m(|L_i| - |L_{i-1}|) \leqq 3m\frac{q}{l},$$

and

$$|\bar{L}| = |L_{i-1}| - |\bar{\bar{L}}| \geqq |L_0| - |\bar{\bar{L}}| \geqq \left(\frac{\gamma}{4} - \frac{3m}{l}\right) q \geqq \frac{1}{8}\gamma q$$

since by our choice of l we have $l \geqq \dfrac{24m}{\gamma}$.

Now let us drop m elements from each end of each of the progressions (mod x_i) composing $\bar{L}$, and denote the remaining set by M. Since every progression in $\bar{L}$ has a length of at least $3m$ we have

$$|M| \geqq \frac{1}{3}|\bar{L}| \geqq \frac{\gamma}{24} q.$$

By construction $[0, q) - M$ can be represented as the union of disjoint progressions (mod x_i) each of length at least m. Thus we can apply the Simple Lemma with $\varepsilon = \varepsilon' = \varepsilon_0$ and obtain

$$|L_l \cap B| \geqq |\bar{L} \cap B| \geqq |M \cap B| \geqq \gamma |M| - 2\varepsilon_0 q \geqq \frac{\gamma^2}{24} q - 2\varepsilon_0 q \geqq \frac{\gamma^2}{50} q,$$

since ε_0 has been chosen suitably.

By definition, $L_l \cap B$ is the set of those b in B which have a representation

$$b = 2c - s, \quad s \in S_l \quad c \in C \cap \left[\frac{1}{3}q, \frac{2}{3}q\right).$$

In S_l there are at most 2^l elements. Therefore at least one y contained in S_l has the property that the equation

$$b - 2c + y = 0$$

has at least $\dfrac{\gamma^2 q}{50 \cdot 2^l}$ solutions (b, c). In another notation this means that

$$|\{y\}^*| \geqq \gamma' q,$$

where we have put

$$\gamma' = \frac{\gamma^2}{50 \cdot 2^l}.$$

The statement of lemma $|G^*|$ is now immediate. From $y \in S_l \subseteqq G$ we see that

$$|G^*| \geqq |\{y\}^*| \geqq \gamma' q.$$

PROOF OF LEMMA $(H_0, \ldots, H_k)$. We first fix some number h such that $\left(1-\frac{\gamma}{2}\right)^h < \gamma'$, for example

$$h = \left[1 + \frac{\log \gamma'}{\log\left(1-\frac{\gamma}{2}\right)}\right].$$

We now start from some $G_0 \subseteqq \left[\frac{1}{3}q, \frac{2}{3}q\right)$ with $|G_0| \geqq \frac{1}{12}\gamma q$ and put $g_0 = |G^*|$. Next we define by recursion for $i = 1, \ldots, h$

$$\Gamma_i = \left\{G, G \subseteqq G_{i-1}, |G| \geqq \frac{\gamma}{2}|G_{i-1}|\right\}, \qquad g_i = \min_{G \in \Gamma_i} |G^*|$$

and fixe one G_i in Γ_i for which $|G_i^*| = g_i$.

From $G_i \in \Gamma_i$ we see that

$$|G_i| \geqq \frac{\gamma}{2}|G_{i-1}| \geqq \ldots \geqq \left(\frac{\gamma}{2}\right)^i |G_0| \geqq \left(\frac{\gamma}{2}\right)^i \frac{\gamma}{12} q,$$

$$|G_i| \geqq \frac{1}{6}\left(\frac{\gamma}{2}\right)^{h+1} q.$$

Thus, if we take $\delta = \frac{1}{2}\left(\frac{\gamma}{2}\right)^{h+1}$ and $q_0 = q_0(\delta)$ we can apply lemma $|G^*|$ for all $q \geqq q_0$ and obtain

$$g_i = |G_i^*| \geqq \gamma' q, \quad \text{for} \quad i = 1, 2, \ldots, h.$$

Since clearly $g_0 \leqq q$ there is a $j \leqq h$ such that

$$g_j \geqq \left(1-\frac{\gamma}{2}\right) g_{j-1},$$

otherwise we should have the contradiction

$$\gamma' q \leqq g_h < \left(1-\frac{\gamma}{2}\right)^h g_0 \leqq \left(1-\frac{\gamma}{2}\right)^h q < \gamma' q.$$

Set with this j $H=G_{j-1}$. From the meaning of g_j and g_{j-1} it follows that if $G\subseteqq H$, and $|G|\geqq\frac{\gamma}{2}|H|$, then $G\in\Gamma_j$ and therefore

$$|G^*|\geqq g_j\geqq\left(1-\frac{\gamma}{2}\right)g_{j-1}=\left(1-\frac{\gamma}{2}\right)|H^*|.$$

Moreover we have

$$|H|=|G_{j-1}|\geqq\frac{1}{6}\left(\frac{\gamma}{2}\right)^{h+1}q.$$

At first we apply this process to $G_0=\left[\frac{1}{3}q,\frac{2}{3}q\right)$ and call the set H obtained H_1. Then we take $G_0=\left[\frac{1}{3}q,\frac{2}{3}q\right)\urcorner H$, and if this set contains at least $\frac{1}{12}\gamma q$ elements we obtain a set H_2 from it. Next we take $G_0=\left[\frac{1}{3}q,\frac{2}{3}q\right)\urcorner(H_1\cup H_2)$ to get a set H_3, and so on. As soon as we are left with

$$\left|\left[\frac{1}{3}q,\frac{2}{3}q\right)\urcorner(H_1\cup H_2\cup\ldots\cup H_k)\right|<\frac{\gamma}{12}q$$

we stop the procedure and call this remaining set H_0.

Since the sets H_K are obviously disjoint and

$$|H_K|\geqq\frac{1}{6}\left(\frac{\gamma}{2}\right)^{h+1}q\quad\text{for}\quad K=1,2,\ldots,k$$

this occurs certainly after a finite number of steps. To be precise, we see that

$$k\leqq\frac{1}{3}q\left(\frac{1}{6}\left(\frac{\gamma}{2}\right)^{h+1}q\right)^{-1}=2\left(\frac{2}{\gamma}\right)^{h+1}.$$

By construction $H_0\cup H_1\cup\ldots\cup H_k=\left[\frac{1}{3}q,\frac{2}{3}q\right)$ and if $G\subseteqq H_K$, $|G|\geqq\frac{\gamma}{2}|H_K|$ then $|G^*|\geqq\left(1-\frac{\gamma}{2}\right)|H_K^*|$ for all $K=1,2,\ldots,k$. This is precisely the statement of lemma $(H_0,\ldots,H_k)$.

Proof of Lemma *BCDE*. Let us take n and q to be integers so that $nq\geqq 6n_4\left(\frac{\varepsilon}{3}\right)$ and let A be a 4-free set contained in $[0,4nq)$ which satisfies

$$|A|\geqq\gamma 4nq.$$

Then we can decompose A into

$$A=B\cup(C+nq)\cup(D+2nq)\cup(E+3nq)$$

with $B,C,D,E\subseteqq[0,nq)$ and (in an obvious notation)

$$B=\bigcup_{x<n}(B+xq)\quad\text{with}\quad B_x\subseteqq[0,q),$$

similarly for C, D, E. For their respective cardinalities we get easily the estimates

$$|B|,\ |C|,\ |D|,\ |E| \geqq (\gamma-\varepsilon)nq.$$

That A is 4-free is reflected in the fact that $Q(b, c, d, e)$ has no solutions with $b \in B$, $c \in C$, $d \in D$, $e \in E$. More precisely: If $Q(x, y, z, w)$ holds, then $Q(b, c, d, e)$ is insolvable with $b \in B_x$, $c \in C_y$, $d \in D_z$, $e \in E_w$. Moreover all of the sets B_x, C_y, D_z, E_w are 4-free.

Let us call a set B etc. $\subseteqq [0, q)$ *full* if $|B| \geqq (\gamma-\varepsilon_1)q$, and *poor* otherwise.

Clearly lemma *BCDE* will be proved if we can show that there are u quadruples (b, c, d, e) such that all B_b's are equal, all C_c's are equal, all B_b's, C_c's, D_d's, E_e's are full, and the e's form an arithmetic progression.

We shall use all the ideas from the proof of lemma $|G^*|$ but not only these, moreover the technique will be more involved.

We can easily provide a set $\mathfrak{B}$ with positive density (about 2^{-q}) such that all B_b for $b \in \mathfrak{B}$ are equal and full. Similarly we find a dense set $\mathfrak{C}$ with all C_c for $c \in \mathfrak{C}$ equal and full. We have then a set of type S_e in $\mathfrak{C}$ through which we 'project' $\mathfrak{B}$ onto the levels of D and E. The points e defined by $Q(b, s, *, e)$ are plentiful and are arranged into long progressions. Hence it can be shown that almost all E_e with these e's are full. The same could be done for the sets D_d with d from $Q(b, s, d, *)$ but unfortunately not in the necessary simultaneous way, since the relation between the e's and the d's is not unique and this relationship weakens the larger l is taken.

The idea which overcomes this difficulty is to use not only one set $\mathfrak{C}$, but a large number of them, $\mathfrak{C}_0, \mathfrak{C}_1, \ldots, \mathfrak{C}_{r-1}$ generated from one of them by shifting $\mathfrak{C}_\varrho = \mathfrak{C}_0 + \varrho$, such that $C_c = C_{c'}$, if c and c' belong to the same set $\mathfrak{C}_\varrho$. This again introduces long progressions on the levels of D and E, which can be exploited independently of the former ones. As a result we get u quadruples of the required type for at least one ϱ with $b \in \mathfrak{B}$ and all $C \in \mathfrak{C}_\varrho$, and so all B_b as well as C_c coincide.

We shall use the following simple counting argument a couple of times: If $\sum_{x=1}^{n} a_x \geqq (\gamma-\varepsilon_3)n$ and $a_x \leqq (\gamma+\varepsilon_2)$ for all x, then the number R of terms a_x which satisfy $a_x \leqq (\gamma-\varepsilon_1)$ is

$$R \leqq \frac{\varepsilon_2+\varepsilon_3}{\varepsilon_1} n.$$

PROOF.

$$(\gamma-\varepsilon_3)n \leqq (\gamma-\varepsilon_1)R + (\gamma+\varepsilon_2)(n-R), \quad (\varepsilon_1+\varepsilon_2)R \leqq (\varepsilon_2+\varepsilon_3)n.$$

We list now the parameters used in the proof, in the order of their dependence. The reader may check them as they occur.

ε, u and q_0 are supposed to be given,

$$\varepsilon_2 = \frac{\varepsilon_1}{16u}, \qquad l = 75m \cdot 2^q,$$

$$q = \max(q_0, n_4(\varepsilon_2)), \qquad \varepsilon_4 = \frac{\varepsilon_2^2}{600 \cdot 2^{q+2l}},$$

$$\varepsilon_3 = \frac{\varepsilon_2}{150 \cdot 2^q}, \qquad r = n_4(\varepsilon_4),$$

$$\varepsilon = \text{sufficiently small}$$

$$m = \max(2u, n_4(\varepsilon_3)), \qquad n = \text{sufficiently large},\ 6r|n.$$

We can safely dispense with specifying ε and n since there is no feedback to the other parameters. A small ε only demands a large n.

By an already repeatedly used argument we get

$$\sum_{x<\frac{n}{6}} |B_x| = \left|B \cap \left[0, \frac{1}{6} nq\right)\right| \geqq (\gamma-\varepsilon)\frac{nq}{6}$$

$$\sum_{\frac{n}{6}\leqq y<\frac{n}{3}} |C_y| = \left|C \cap \left[\frac{nq}{6}, \frac{nq}{3}\right)\right| \geqq (\gamma-\varepsilon)\frac{nq}{6}$$

provided only that n is large enough. We set $\varepsilon_2 = \frac{\varepsilon_1}{16u}$, and take $q \geqq n_4(\varepsilon_2)$, so that we then have for all x, y, z, w,

$$|B_x|,\ |C_y|,\ |D_z|,\ |E_w| \leqq (\gamma+\varepsilon_2)q.$$

By the above counting argument the number of poor B_x, $0 \leqq x < \frac{n}{6}$ and the number of poor C_y, $\frac{n}{6} \leqq y < \frac{n}{3}$ is each at most $\left(\frac{1}{16u}+\frac{\varepsilon}{\varepsilon_1}\right)\frac{n}{6} \leqq \frac{1}{8}\cdot\frac{n}{6}$ if ε is small enough. Consequently more than half of the B_x are full.

There are only 2^q subsets of $(0, q)$, so there is a full $B_{(0)} \subseteqq (0, q)$ such that

$$B_b = B_{(0)} \quad \text{for} \quad b \in \mathfrak{B} \subseteqq \left[0, \frac{n}{6}\right), \quad \text{with} \quad |\mathfrak{B}| \geqq \frac{n}{12\cdot 2^q}.$$

We next look at C, and assuming that $r|n$ we consider the r-tuples

$$(C_{mr}, C_{mr+1}, \ldots, C_{mr+r-1}), \quad \frac{n}{6r} \leqq m < \frac{n}{3r}.$$

Since not more than $\frac{1}{8}$ of the C_j are poor, not more than $\frac{1}{2}$ of the r-tuples contain more than $\frac{1}{4}$ poor sets. There are only 2^{qr} different r-tuples, so we find

$$C_{(0)}, \ldots, C_{(r-1)},$$

not more than $\frac{1}{4}$ of them being poor, and $\mathfrak{B} \subseteqq \left[\frac{n}{6}, \frac{n}{3}\right)$ so that

$$C_{c+\varrho} = C_{(\varrho)} \quad \text{for} \quad c \in \mathfrak{C} \quad \text{and} \quad \varrho \in [0, r), \qquad |\mathfrak{C}| \geqq \frac{n}{12r\cdot 2^{qr}}.$$

By lemma $p(\delta, l)$ we see that $\mathfrak{C}$ contains a subset of type

$$S_l = \{y\} + \{0, x_1\} + \ldots + \{0, x_l\}.$$

With the sets

$$S_i = \{y\} + \{0, x_1\} + \ldots + \{0, x_i\}$$

we form

$$L_i = \{3s - 2b;\ s \in S_i,\ b \in \mathfrak{B}\}.$$

Then we have

$$L_i \subseteqq \left[\frac{n}{6}, n\right), \quad |L_0| = |\mathfrak{B}| \geqq \frac{n}{12\cdot 2^q},$$

$$L_i = L_{i-1} \cup (L_{i-1} + 3x_i).$$

For a suitable $i \leqq l$ we have

$$|L_i| - |L_{i-1}| \leqq \frac{n}{l}.$$

We decompose L_{i-1} into maximal progression (mod $3x_i$), collect those progressions which are longer than $3m$ into $\bar{L}$, and the remaining ones into $\bar{\bar{L}}$; as in the proof of lemma $|G^*|$ we get

$$|\bar{\bar{L}}| \leqq 3m(|L_i| - |L_{i-1}|) \leqq \frac{3mn}{l},$$

$$|\bar{L}| \geqq |L_0| - |\bar{\bar{L}}| \geqq \left(\frac{1}{12\cdot 2^q} - \frac{3m}{l}\right) n \geqq \frac{n}{25\cdot 2^q}.$$

(Here we have taken $l \geqq 72m\cdot 2^q$). Dropping the first m and the last m elements of each of the progressions collected into $\bar{L}$, we obtain a set we shall call $\mathscr{E}$. Then

$$|\mathscr{E}| \geqq \frac{1}{3}|\bar{L}| \geqq \frac{n}{75\cdot 2^q}$$

and $[0, n) \urcorner \mathscr{E}$ is the union of disjoint progressions (mod $3x_i$), none of which contains fewer than m elements.

If we start from $S_l + \varrho \subseteqq \mathfrak{C} + \varrho$ instead of S_l, $0 \leqq \varrho < r$ we get $\mathscr{E} + 3\varrho$ instead of $\mathscr{E}$. Thus the complement of $\mathscr{E} + 3\varrho$ too is composed of disjoint progressions, each of length not less than m.

We now show that if m is large enough then almost all E_e with $e \in \mathscr{E}$ (or $\mathscr{E} + 3\varrho$) are full. In particular we show that the following conditions are sufficient:

$$m \geqq n_4(\varepsilon_3), \quad \text{where} \quad \varepsilon_3 = \frac{\varepsilon_2}{150\cdot 2^q}.$$

The set

$$M = \bigcup_{e\in\mathscr{E}} [eq, (e+1)q)$$

has the property of the set M in the Simple Lemma. (The progressions have the modulus $3qx_i$ and are each of length at least m; $\varepsilon' = \varepsilon_3$). Therefore

$$\sum_{e\in\mathscr{E}} |E_e| = |E \cap M| \geqq \gamma |M| - (\varepsilon + \varepsilon_3)qn = \gamma q|\mathscr{E}| - (\varepsilon+\varepsilon_3)qn \geqq$$

$$\geqq \gamma q|\mathscr{E}| - 2\varepsilon_3 qn \geqq (\gamma - 150\cdot 2^q \varepsilon_3)q|\mathscr{E}| = (\gamma - \varepsilon_2)q|\mathscr{E}|.$$

Since $|E_e| \leqq (\gamma + \varepsilon_2)q$ for all e, the 'counting argument' applies, showing that the number of poor E_e, $e \in \mathscr{E}$ is at most

$$\frac{2\varepsilon_2}{\varepsilon_1}|\mathscr{E}| = \frac{1}{8u}|\mathscr{E}|.$$

More generally, for each $\varrho=0, \dots, r-1$ there are at most $\frac{1}{8u}|\mathcal{E}|$ poor sets $E_{e+3\varrho}$, $e\in\mathcal{E}$.

Each $e\in\mathcal{E}$ by construction occurs in at least one quadruple (b, s, d, e) with $b\in\mathfrak{B}$ and $s\in S_l$. To each $e\in\mathcal{E}$ we attach one such quadruple making the d, as well as the b and the s, a function of e, $d=\varphi(e)$. Let

$$\mathcal{D}=\{\varphi(e);\ e\in\mathcal{E}\}.$$

Since S_l has at most 2^l elements any particular d in D can arise as a value $\varphi(e)$ at most 2^l times.

We consider the quadruples

$$(b, s+\varrho, \varphi(e)+2\varrho, e+3\varrho),\ e\in\mathcal{E},\ \varrho\in[0, r).$$

We want now to show that for at least one ϱ

$C_{s+\varrho}$ is full (independent of e since $C_{s+\varrho}=C_{(\varrho)}$),

and

almost all $D_{\varphi}(e)+2\varrho$ are full (counted with multiplicity).

We do this by considering all the ϱ together. The basic tool is again the Simple Lemma. Before applying it, however, we have to remove the multiplicities with which the $C\varphi(e)+\varrho$ occur. There are two sources of multiplicity: the mapping $\varphi(e)=d$, and the forming of the sum $d+\varrho$. We deal first with the case when φ is one to one, where only one of these sources is present.

Set

$$\mathcal{D}'=\left\{d;\ d\in\mathcal{D},\ \sum_{\varrho=0}^{r-1}|D_{d+2\varrho}|\leqq(\gamma-\varepsilon_2)qr\right\}.$$

We construct a subset $\mathcal{D}''\subseteqq\mathcal{D}'$ with the property that consecutive elements have a difference of at least $4r$, but

$$|\mathcal{D}''|\geqq\frac{1}{4r}|\mathcal{D}'|.$$

For this purpose we may go from left to right retaining for our set $\mathcal{D}''$ the first element not ruled out by the restriction upon the differences. Since we exclude at most $4r-1$ elements for each one which we keep we obtain the stated inequality.

Now, each element in

$$\mathcal{D}'''=\mathcal{D}''+\{0, 2, 4, \dots, 2(r-1)\}$$

is uniquely represented. Therefore we have

$$\Big|\bigcup_{x\in\mathcal{D}'''}D_x\Big|=\sum_{d\in\mathcal{D}''}\sum_{\varrho=0}^{r-1}|D_{d+2\varrho}|\leqq(\gamma-\varepsilon_2)rq\,|\mathcal{D}''|.$$

By construction the complement of $\mathcal{D}'''$ consists of progressions (mod 2), each of length at least r. (No difficulty arises when considering elements to the left of the first and to the right of the last elements in $\mathcal{D}'''$, respectively, since $\mathcal{D}+2\varrho\subseteqq\subseteqq\left[\frac{1}{6}n, \frac{2}{3}n\right)\Big)$. Therefore the left hand side can be estimated by the Simple Lemma.

We take

$$M = \bigcup_{x\in\mathscr{D}'''} [xq, (x+1)q); \quad r = n_4(\varepsilon_4), \quad \varepsilon_4 \leqq \frac{\varepsilon_2^2}{600\cdot 2^q}$$

and obtain

$$|\bigcup_{x\in\mathscr{D}'''} D_x| = |D\cap M| \geqq \gamma q|\mathscr{D}'''| - (\varepsilon+\varepsilon_4)qn =$$

$$= \gamma qr|\mathscr{D}''| - (\varepsilon+\varepsilon_4)qn \geqq \gamma qr|\mathscr{D}''| - z\varepsilon_4 qn.$$

Putting these estimates together gives

$$\varepsilon_2 r|\mathscr{D}''| \leqq 2\varepsilon_4 n, \quad |\mathscr{D}'| \leqq 4r|\mathscr{D}''| \leqq 8\frac{\varepsilon_4}{\varepsilon_2} n.$$

Next we have the estimate

$$(*) \qquad \sum_{d\in\mathscr{D}} \sum_{\varrho=0}^{r-1} |D_{d+2\varrho}| \geqq \sum_{d\in\mathscr{D}\urcorner\mathscr{D}'} \sum_{\varrho=0}^{r-1} |D_{d+2\varrho}| \geqq$$

$$\geqq (|\mathscr{D}|-|\mathscr{D}'|)(\gamma-\varepsilon_2)rq \geqq (\gamma-\varepsilon_2)\left(|\mathscr{D}| - 8\frac{\varepsilon_4}{\varepsilon_2}n\right) rq.$$

In the present special case we have $|\mathscr{D}| = |\mathscr{E}| \geqq \frac{n}{75\cdot 2^q}$. We therefore get the further inequality

$$\sum_{d\in\mathscr{D}} \sum_{\varrho=0}^{r-1} |D_{d+2\varrho}| \geqq (\gamma-\varepsilon_2)\left(1 - 8\cdot 75\cdot 2^q\frac{\varepsilon_4}{\varepsilon_2}\right) rq|\mathscr{D}| \geqq$$

$$\geqq (\gamma-\varepsilon_2)(1-\varepsilon_2)rq|\mathscr{D}| \geqq (\gamma-2\varepsilon_2)rq|\mathscr{D}|.$$

By the 'counting argument' we infer that not more than $3\frac{\varepsilon_2}{\varepsilon_1} r|D| = \frac{3}{16u} r|\mathscr{E}|$ sets $D_{d+2\varrho}$, taken with their multiplicity, are poor. For at most one half of the ϱ's can we have more than $\frac{3}{8u}|\mathscr{E}|$ poor sets $D_{d+2\varrho}$.

If we drop these numbers ϱ, of which there at most $\frac{1}{2}r$, and also those ϱ for which $C_{(\varrho)}$ is poor, there being no more than $\frac{1}{4}r$ of them, some of the numbers ϱ remain. So far we have proved:

There is a number $o\in[0, r)$ such that $C_{s+o} = C_{(o)}$ is full, at most $\frac{3}{8u}|\mathscr{E}|$ of the sets $D_{\varphi(e)+2o}$, $e\in\mathscr{E}$ are poor, and at most $\frac{1}{8u}|\mathscr{E}|$ of the sets E_{e+3o} are poor.

Hence for at most $\frac{1}{2u}|\mathscr{E}|$ elements $e\in\mathscr{E}$ we have either E_{e+3o} or $D_{\varphi(e)+2o}$ poor. We call these $e\in\mathscr{E}$ 'bad'. The density of the bad elements in $\mathscr{E}$ is at most $\frac{1}{2u}$. Now recall that $\mathscr{E}$ is composed of disjoint arithmetic progressions of length at least m.

We can take $m \geqq 2u$. If one of every u consecutive elements of such a progression were a bad one, the density of bad elements in any particular progression in $\mathscr{E}$ would be at least

$$\frac{2}{3u-1} > \frac{2}{3u}$$

and so therefore would be the density of bad elements in the whole of $\mathscr{E}$. Since we have disproved this there exists an arithmetic progression of at least u good elements in $\mathscr{E}$, q.e.d.

Rather little has to be changed in the general case when the elements $d \in \mathscr{D}$ are taken with the multiplicities of $d = \varphi(e)$ not necessarily all equal to one.

Set

$$\mathscr{D}^i = \{d;\ d = \varphi(e) \quad \text{for exactly } i \text{ elements} \quad e \in \mathscr{E}\}.$$

Each $\mathscr{D}^i$ can be treated in exactly the same way that $\mathscr{D}$ was until we reach the formula $(\ast)$. However, in order to make the formula useful this time we must take a smaller ε_4 (and therefore a larger r):

$$\varepsilon_4 = \frac{\varepsilon^2}{600 \cdot 2^{q+2l}}, \quad r = n_4(\varepsilon_4).$$

We have then

$$\sum_{d \in \mathscr{D}^i} \sum_{\varrho=0}^{r-1} |D_{d+2\varrho}| \geqq (\gamma - \varepsilon_2)\left(|\mathscr{D}^i| - 8\frac{\varepsilon_4}{\varepsilon_2} n\right) rq.$$

Multiplying by i and summing gives

$$\sum_{e \in \mathscr{E}} \sum_{\varrho=0}^{r-1} |D_{\varphi(e)+2\varrho}| \geqq (\gamma - \varepsilon_2)\left(|\mathscr{E}| - 8\frac{\varepsilon_4}{\varepsilon_2} n \sum_{i=1}^{2^l} i\right) rq \geqq$$

$$\geqq (\gamma - \varepsilon_2)\left(1 - 8\frac{\varepsilon_4}{\varepsilon_2} \cdot 2^{2l} \cdot 75 \cdot 2^q\right) rq\,|\mathscr{E}| \geqq (\gamma - 2\varepsilon_2) rq\,|\mathscr{E}|.$$

The counting argument again shows that there is an $o \in [0, r)$ such that for at most $\frac{3}{8u}|\mathscr{E}|$ elements $e \in \mathscr{E}$ the sets $D_{\varphi(e)+2o}$ are full, and the proof is finished as above.

We have now completed the proof of lemma *BCDE* and with it the proof of the theorem.

The author wishes to express his thanks to E. WIRSING and P. D. T. A. ELLIOTT, who helped considerably in the final formulation of the proof.

(Received 20 October 1967)

MTA MATEMATIKAI KUTATÓ INTÉZETE,
BUDAPEST, V., REÁLTANODA U. 13–15