

# A generalisation of a theorem of Erdos

Tanmoy Chakraborty

## Introduction

This is my project work in the summer of 2004, that followed a reading course in Probabilistic Methods under Prof. K. V. Subrahmanyam. I found a theorem of Paul Erdos, which said that any arbitrary set of  $n$  distinct integers contains a sum-free subset of size at least  $n/3$  (A set is called sum-free if no two numbers in the set (might not be distinct) add up to a third).

I generalised this result to all linear relations (sum-free corresponds to forbidding a specific linear relation, namely,  $x_1 + x_2 = x_3$ ). I proved that for every linear relation  $R$ , one can find a constant fraction of an arbitrary set of integers that are  $R$ -independent (discussed in details below). I also proved that the result holds for real numbers.

This write-up contains these results and their proofs in details.

## Terminology

We say that the ordered tuple  $(c_1, c_2 \dots c_l)$  satisfy a relation  $R$  if

$$\sum_{i=1}^w p_i c_i = \sum_{j=1}^{l-w} q_j c_{w+j}$$

where  $c_1, c_2 \dots c_l$  are real numbers,  $p_1, p_2 \dots p_w, q_1, q_2 \dots q_{l-w}$  are constant integer coefficients,  $l > w > 0$ ,  $\sum_{i=1}^w p_i = p$ ,  $\sum_{j=1}^{l-w} q_j = q$ , and  $q > p$ .

Let  $S$  be any set of real numbers.

**Definition 1.** We say that  $S$  is  $R$ -independent if  $\nexists c_1, c_2 \dots c_l \in S$  satisfying  $R$ . (It is permissible that  $c_i = c_j$  even if  $i \neq j$ .)

**Definition 2.** We say that  $S$  is strongly  $R$ -independent if  $\nexists c_1, c_2 \dots c_l \in S$  such that

$$\left| \sum_{i=1}^w p_i c_i - \sum_{j=1}^{l-w} q_j c_{w+j} \right| < q$$

(Again, it is permissible that  $c_i = c_j$  even if  $i \neq j$ .)

We define the “real circle modulo  $r$ ”, where  $r \in \mathbf{R}$  as follows: The circle has circumference of length  $r$  units. An arbitrary point on the circle is marked as zero, and  $\forall x \in \mathbf{R}$ ,  $0 < x < r$ , the point at a distance of  $x$  units along the circumference, in anticlockwise direction, is marked as  $x$ .

We consider intervals on this real circle  $[a, b]$  as the points(numbers) on the circumference that lie on the arc that is traversed while going in anti-clockwise direction to  $a(\text{mod } r)$  to  $b(\text{mod } r)$ .

## Some remarks and lemmas

**Remark 1.**  $S \subset \mathbf{R}$  is  $R$ -independent if  $\exists x \in \mathbf{N}$  such that  $S_x = \{xs | s \in S\}$  is  $R$ -independent. Also, if  $S$  is  $R$ -independent, then  $S_x$  is  $R$ -independent  $\forall x \in \mathbf{N}$ .

**Remark 2.** If  $S \subset \mathbf{R}$  is  $R$ -independent or strongly  $R$ -independent in  $\mathbf{Z}_n$  (the ring of integers modulo  $n \geq 2$ ), then  $S$  is  $R$ -independent or strongly  $R$ -independent, respectively, in  $\mathbf{Z}$ .

**Remark 3.** If  $S \subset \mathbf{R}$  is  $R$ -independent or strongly  $R$ -independent, then so is any of its subsets.

**Lemma 1.**  $S \subset \mathbf{R}$  is  $R$ -independent if the set  $S' = \{\lfloor x \rfloor | x \in S\}$  is strongly  $R$ -independent.

*Proof.* We shall prove the contrapositive.

Suppose  $S$  is not  $R$ -independent. Then,  $\exists c_1, c_2 \dots c_l \in S$ , such that

$$\sum_{i=1}^w p_i c_i = \sum_{j=1}^{l-w} q_j c_{w+j} = v \text{ (say)}$$

Then,

$$v \geq \sum_{i=1}^w p_i \lfloor c_i \rfloor > v - \sum_{i=1}^w p_i = v - p$$

and  $v \geq \sum_{j=1}^{l-w} q_j c_{w+j} > v - q$ . Now, since  $q > p$ , so

$$\left| \sum_{i=1}^w p_i \lfloor c_i \rfloor - \sum_{j=1}^{l-w} q_j \lfloor c_{w+j} \rfloor \right| < v - (v - q) = q$$

Hence,  $S'$  is not strongly  $R$ -independent. □

**Lemma 2.** Let  $a, b \in \mathbf{N}$  such that  $\gcd(a, b) = 1$ . Then,  $\forall N \in \mathbf{N}, \exists k \in \mathbf{N}, k > N$  such that  $ak + b$  is prime.

*Proof.* Direct consequence of Dirichlet's theorem. □

**Proposition 1.** Let  $R$  be any relation of the dicussed type. Let  $r$  be a prime of the form  $(q^2 - p^2)k + (q^2 - p^2 - 1)$ . ( $r$  exists,  $\because \gcd(q^2 - p^2, q^2 - p^2 - 1) = 1$ ).

1. Let  $D = \{pk + 1, pk + 2 \dots qk\}$  ( $|D| = (q - p)k$ ). Then,  $D$  is strongly  $R$ -independent in  $\mathbf{Z}_r$  and hence in  $\mathbf{Z}$ .
2. Let  $E = \{pk + p, pk + p + 1 \dots qk + q - 1\}$  ( $|E| = (q - p)(k + 1)$ ). Then,  $E$  is  $R$ -independent in  $\mathbf{Z}_r$  and hence in  $\mathbf{Z}$ .

*Proof.* By intervals, we shall mean those on the “real circle modulo  $r$ ”.

*Case 1.* If all the variables are drawn from  $D$ , then the L.H.S. of the equation for  $R$  clearly lies in  $[p^k + p, pqk]$ . Also, range of R.H.S. =  $[pqk + q, q^2k] = [pqk + q, p^2k - (q^2 - p^2 - 1)]$  (since  $q^2k \equiv p^2k - (q^2 - p^2 - 1) \pmod{r}$ ). So, the ranges do not overlap, and are infact separated by  $\min\{q, q^2 - p^2 - 1 + p\} \geq q$ . So,  $D$  is strongly  $R$ -independent in  $\mathbf{Z}_r$ , and hence, by Remark 2, in  $\mathbf{Z}$ .

*Case 2.* In this case, range of L.H.S. =  $[p^k + p^2, pqk + pq - p]$ , and range of R.H.S. =  $[pqk + pq, q^2k + q^2 - q] = [pqk + pq, p^2k + p^2 - (q + 1)]$ . Again, the ranges do not overlap, so  $E$  is  $R$ -independent in  $\mathbf{Z}_r$ , and, by Remark 2, in  $\mathbf{Z}$ . □

## The Main Results

**Theorem 1.** Let  $A = \{a_1, a_2 \dots a_n\}$  be any set of  $n$  non-zero distinct real numbers. Then,  $\exists B \subseteq A, |B| \geq \frac{n}{p+q}$ , such that  $B$  is  $R$ -independent. Moreover, if  $(p + q) \nmid n$ , we can find  $B$  such that  $|B| > \frac{n}{p+q}$ .

*Proof.* Let us fix  $\epsilon \in \mathbf{R}, 0 \leq \epsilon \leq \frac{1}{p+q}$ .

To show:  $\exists B \subseteq A, |B| > (\frac{n}{p+q} - n\epsilon)$ , such that  $B$  is  $R$ -independent.

Let  $\alpha = \min\left\{\{|a_i - a_j| \mid 1 \leq i < j \leq n\} \cup \{|a_i| \mid 1 \leq i \leq n\}\right\}$ . Choose  $N \in \mathbf{N}, N > \frac{6}{\epsilon(p+q)\alpha}$ . Let  $x_i = Na_i, 1 \leq i \leq n$ .

Then,  $|x_i| > \frac{6}{\epsilon(p+q)}, 1 \leq i \leq n$ . Also,  $\lfloor x_i \rfloor \neq \lfloor x_j \rfloor, \forall 1 \leq i < j \leq n$ . Let  $C = \{x_1, x_2 \dots x_n\}$ . Let  $\beta = \max_{1 \leq i \leq n} \{|x_i|\}$ . Let us also choose a prime  $r$  of the form  $(q^2 - p^2)k + (q^2 - p^2 - 1)$ , where  $k > \frac{3\beta}{\epsilon}$ .

Let us pick  $z \in \mathbf{N}$ ,  $1 \leq z < r$ , randomly, according to a uniform distribution on  $\{1, 2 \dots r - 1\}$ . Let  $\lfloor zx_i \rfloor \equiv d_i \pmod{r}$ ,  $1 \leq i \leq n$ . Let  $D = \{pk + 1, pk + 2 \dots qk\}$ .

**Claim.**  $Pr(d_i \in D) > \frac{1}{p+q} - \epsilon, \forall 1 \leq i \leq n$ .

*Proof.* Let  $P = \{z | \lfloor zx_i \rfloor \pmod{r} \in D; z \in \{1, 2 \dots r - 1\}\}$ . We shall find  $|P|$ .

Let  $I$  be the interval  $[pk+1, qk]$ .

Now, if we plot  $x_i, 2x_i, 3x_i \dots (r-1)x_i$ , modulo  $r$ , on the “real circle modulo  $r$ ”, in that order, it is same as moving  $x_i$  units on the circle in every “step”, starting from the point  $x_i$ , in anti-clockwise direction. We take  $(r-2)$  such steps. To estimate  $|P|$ , we need to count how many times we “step” on  $I$ .

In making a complete rotation of the circle, at least  $\lfloor \frac{(q-p)k}{x_i} \rfloor$  “steps” must fall on  $I$ , since  $I$  has length  $(q-p)k$ . Also, we make at least  $\lfloor \frac{(r-2)x_i}{r} \rfloor \geq \lfloor x_i \rfloor - 1$  complete rotations of the cycle in the process.

Now,  $zx_i \pmod{r} \in I \Rightarrow z \in P$ . Hence,

$$\begin{aligned} |P| &\geq (\lfloor x_i \rfloor - 1) \left( \lfloor \frac{(q-p)k}{x_i} \rfloor \right) \\ &> (x_i - 2) \left( \frac{(q-p)k}{x_i} - 1 \right) \\ &> \left( 1 - \frac{2}{x_i} \right) (q-p)k - x_i \end{aligned}$$

$$\begin{aligned}
\therefore Pr(d_i \in D) &= \frac{|P|}{r-1} \\
&> \frac{(1 - \frac{2}{x_i})(q-p)k}{(q^2-p^2)(k+1)} - \frac{x_i}{(q^2-p^2)k} \\
&\quad (\because (q^2-p^2)(k+1) > r-1 > (q^2-p^2)k) \\
&> \left(1 - \frac{2}{x_i}\right) \left(1 - \frac{1}{k}\right) \left(\frac{1}{q+p}\right) - \frac{x_i}{k} \\
&= \frac{1}{q+p} - \left(\frac{2}{x_i(q+p)} + \frac{1}{(q+p)k} - \frac{2}{x_i k(q+p)} + \frac{x_i}{k}\right) \\
&> \frac{1}{q+p} - \left(\frac{2}{(q+p)} \times \frac{\epsilon(q+p)}{6} + \frac{\epsilon}{(q+p)(3\beta)} + \frac{\epsilon x_i}{3\beta}\right) \\
&\quad (\because |x_i| > \frac{6}{\epsilon(q+p)}; k > \frac{3\beta}{\epsilon}; \beta > x_i) \\
&> \frac{1}{q+p} - \left(\frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}\right) = \frac{1}{q+p} - \epsilon
\end{aligned}$$

Hence the claim is proved.  $\square$

Let us introduce indicator random variables  $X_i$  which is 1 if  $d_i \in D$  and 0 otherwise,  $\forall 1 \leq i \leq n$ , and let  $X = |\{x_i | d_i \in D\}|$ . Then,  $X = \sum_{i=1}^n X_i$ .

$\therefore$  By linearity of expectation,

$$\begin{aligned}
E[X] &= \sum_{i=1}^n E[X_i] = \sum_{i=1}^n Pr(d_i \in D) \\
\Rightarrow E[X] &> \frac{n}{p+q} - n\epsilon
\end{aligned}$$

This is true  $\forall 0 < \epsilon \leq \frac{1}{q+p}$ .

$\therefore \exists z$  such that  $X > \frac{n}{p+q} - n\epsilon$ . So,  $\exists z$  such that  $F_z = \{zx_i | [zx_i] \pmod{r} \in D\}$ ,  $|F_z| \geq E[X]$ . But, by Proposition 1 (i), and Remark 3,  $F'_z = \{[x] | x \in F_z\}$  is *strongly R-independent*, and so, by Lemma 1,  $F_z$  is *R-independent*. By Remark 1, it follows that  $Q = \{x_i | zx_i \in F_z\}$  is *R-independent*, and has cardinality  $|F_z| \geq E[X] > \frac{n}{p+q} - n\epsilon$ .

Hence,  $\exists Q \subseteq C$  such that  $Q$  is *R-independent* and  $|Q| > (\frac{n}{p+q} - n\epsilon)$ , and, by Remark 1,  $\exists T \subseteq A$  such that  $T$  is *R-independent* and  $|T| > (\frac{n}{p+q} - n\epsilon)$ .

If  $(p+q)|n$ , we can choose  $\epsilon < \frac{1}{2n}$ , and find an *R-independent* subset of size  $> \frac{n}{p+q} - \frac{1}{2}$ . Since size of the subset is integral, its size must be  $\geq \frac{n}{p+q}$ .

Otherwise (if  $(p+q) \nmid n$ ), we can write  $\frac{n}{p+q} = a+f$ , where  $a \in \mathbf{N} \cup \{0\}$ ,  $f \in \mathbf{R}$ ,  $0 < f < 1$ . Choose  $\epsilon < \frac{f}{2n}$ . Then, we can find an  $R$ -independent subset of size  $i$ ,  $(a+f) - \frac{1}{2}f = a + \frac{1}{2}f$ . Again, since size of the subset is integral, its size must be  $> a + f = \frac{n}{p+q}$ .

Hence proved. □

**Theorem 2.** *Let  $A = \{a_1, a_2 \dots a_n\}$  be any set of  $n$  non-zero (distinct) rational numbers. Then,  $\exists B \subseteq A, |B| > \frac{n}{p+q}$ , such that  $B$  is  $R$ -independent.*

*Proof.* Let  $N = \prod_{i=1}^n (\text{denominator of } a_i)$ . Let  $x_i = Na_i$ ,  $1 \leq i \leq n$ . Then,  $C = \{x_1, x_2 \dots x_n\}$  is a set of non-zero distinct integers. So, by Remark 1, we only need to prove the theorem for  $C$ , a set of integers.

Let us choose a prime  $r$  of the form  $(q^2 - p^2)k + (q^2 - p^2 - 1)$ , where  $k > \max_{1 \leq i \leq n} x_i$  (this is possible, by Lemma 2). Let us pick  $z \in \mathbf{N}, 1 \leq z < r$ , randomly, according to a uniform distribution on  $\{1, 2 \dots r-1\}$ . Let  $zx_i \equiv d_i \pmod{r}$ ,  $0 \leq d_i < r, 1 \leq i \leq n$ . As  $z$  varies from 1 to  $r-1$ ,  $d_i$  takes every value in the set  $\{1, 2 \dots r-1\}$  exactly once.

Let  $E = \{pk + p.pk + p + 1 \dots qk + q - 1\}$ . Then,

$$Pr(d_i \in E) = \frac{|E|}{r-1} > \frac{(q-p)(k+1)}{(q^2-p^2)(k+1)} = \frac{1}{p+q}, \forall 1 \leq i \leq n.$$

Let us introduce indicator random variables  $X_i$  which is 1 if  $d_i \in E$  and 0 otherwise,  $\forall 1 \leq i \leq n$ , and let  $X = |\{x_i | d_i \in E\}|$ . Then,  $X = \sum_{i=1}^n X_i$ .

$$\text{So, by linearity of expectation, } E[X] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n Pr(d_i \in E).$$

$$\therefore, E[X] > \frac{n}{p+q}.$$

Hence,  $\exists z$  such that if  $F_z = \{x_i | zx_i \pmod{r} \in E\}$ , then  $|F_z| > \frac{n}{p+q}$ . Since  $E$  is  $R$ -independent (by Proposition 1 (ii)), so by Remark 1 and 3,  $F_z$  is  $R$ -independent. □

## Acknowledgement

I am indebted to Prof. K. V. Subrahmanyam for motivating me for this generalisation, and also for checking my proofs for the results obtained.