PRODUCT-FREE SETS IN THE FREE SEMIGROUP

IMRE LEADER, SHOHAM LETZTER, BHARGAV NARAYANAN, AND MARK WALTERS

ABSTRACT. In this paper, we study product-free subsets of the free semigroup over a finite alphabet $\mathcal{A}$. We prove that the maximum density of a product-free subset of the free semigroup over $\mathcal{A}$, with respect to the natural measure that assigns a weight of $|\mathcal{A}|^{-n}$ to each word of length n, is precisely $1/2$.

1. INTRODUCTION

A subset S of a semigroup is said to be *product-free* if there do not exist $x, y, z \in S$ (not necessarily distinct) such that $x \bullet y = z$; it is customary to call S *sum-free* when the underlying semigroup is abelian.

It is a well known fact (and an easy exercise) that any sum-free subset of the integers has upper density at most $1/2$. Sum-free subsets of the integers, and of abelian groups in general, have been studied by very many researchers over the last fifty years. For example, from the work of Green and Ruzsa [4], there is now a complete picture of how large a sum-free set we can find in any finite abelian group. We refer the reader to the surveys of Tao and Vu [7] and Kedlaya [5] for more information on these questions.

Product-free subsets of finite non-abelian groups were first investigated by Babai and Sós [1]. Following foundational work by Gowers [3] demonstrating so-called 'product-mixing' phenomena in groups with no low-dimensional representations, there has been a great deal of recent work in the non-abelian setting; for instance, in a recent breakthrough, Eberhard [2] determined how large a product-free subset of the alternating group can be.

In light of these developments, it is natural to ask what one can say about product-free sets in infinite non-abelian structures, a setting in which our knowledge is a bit more limited. Perhaps the first natural place to look among infinite non-abelian

structures is among those that are free, so here, we shall investigate how large product-free subsets of the free semigroup can be.

2. Our results

Let $\mathcal{A}$ be a finite set. We write $\mathcal{F} = \mathcal{F}_{\mathcal{A}}$ for the free semigroup over $\mathcal{A}$; in other words, $\mathcal{F}$ is the set of all finite words over the alphabet $\mathcal{A}$ equipped with the associative operation of concatenation. While we state and prove our results for finite alphabets of all possible sizes for the sake of completeness, the reader will lose nothing by supposing that $\mathcal{A}$ is a two-element set in what follows; indeed, this case captures all the difficulties inherent in the questions we study.

Recall that a set $S \subset \mathcal{F}$ is *product-free* if, writing $\bullet$ for the operation of concatenation, there do not exist words $x, y, z \in S$ (not necessarily distinct) such that $x \bullet y = z$. There is an obvious example of a 'large' subset of $\mathcal{F}$ that is product-free: when $\mathcal{A} = \{a, b\}$ for instance, the set of words which contain an odd number of occurrences of the symbol a (or b, for that matter) is easily seen to be a product-free set that contains, roughly, half the words from $\mathcal{F}$. Our aim in this paper is to prove that these sets are, in a precise sense, the largest product-free subsets of $\mathcal{F}$. We remark in passing that there are several other product-free sets that are 'equally large': for any nonempty subset $\Gamma \subset \mathcal{A}$, the *odd-occurrence set* $\mathcal{O}_{\Gamma} \subset \mathcal{F}$ generated by Γ, namely the set of words in which the total number of occurrences of symbols from Γ is odd, is easily seen to be a product-free set; in the case where $\mathcal{A} = \{a, b\}$, our earlier example corresponds to taking $\Gamma = \{a\}$, and taking $\Gamma = \{a, b\}$ gives us the set of all words of odd length, for example.

To formally state our results, we need a way to measure the size of a set $S \subset \mathcal{F}$. For an integer $n \in \mathbb{N}$, the *layer* $\mathcal{F}(n) \subset \mathcal{F}$ is the set of words of length n, and the *ball* $\mathcal{F}_{\leq}(n) \subset \mathcal{F}$ is the set of words of length at most n. As a first attempt, one might define the density of a set $S \subset \mathcal{F}$ via its densities in balls, namely as the quantity

$$\limsup_{n \to \infty} \frac{|S \cap \mathcal{F}_{\leq}(n)|}{|\mathcal{F}_{\leq}(n)|}.$$

However, a little thought should convince the reader that the counting measure is somewhat ill-suited for our purposes. Indeed, when $|\mathcal{A}| > 1$, almost all the words in $\mathcal{F}_{\leq}(n)$ are long since $|\mathcal{F}(n)| \geq |\mathcal{F}_{\leq}(n)|/2$. Consequently, we may find product-free sets that are intuitively small, and yet have density arbitrarily close to 1 in the

above sense; for example, for any sufficiently large $c \in \mathbb{N}$, the set

$$\bigcup_{n \geq c} (\mathcal{F}_{\leq}(3^n + c) \setminus \mathcal{F}_{\leq}(3^n))$$

is product-free and has density at least $1 - 1/c$ in the above sense, provided $|\mathcal{A}| > 1$.

A more natural approach is to assign a weight of $|\mathcal{A}|^{-n}$ to each word of $\mathcal{F}(n)$, thereby ensuring that the layers $\mathcal{F}(n)$ have the same total weight for all $n \in \mathbb{N}$. To this end, for a subset $S \subset \mathcal{F}$ and an integer $n \in \mathbb{N}$, we define *the density of S in the layer* $\mathcal{F}(n)$ by $d_S(n) = |S \cap \mathcal{F}(n)|/|\mathcal{F}(n)|$. With this definition in place, most standard notions of density may now be carried over: we define the *upper asymptotic density* of S by

$$\bar{d}(S) = \limsup_{n \to \infty} \frac{\sum_{i=1}^{n} d_S(i)}{n},$$

and the *upper Banach density* of S by

$$d^*(S) = \limsup_{n-m \to \infty} \frac{\sum_{i=m}^{n} d_S(i)}{n - m + 1}.$$

Of course, the latter is a weaker notion of density than the former; indeed, it is clear that $\bar{d}(S) \leq d^*(S)$ for any $S \subset \mathcal{F}$.

It is easy to see that any odd-occurrence set has both an upper asymptotic density and an upper Banach density of $1/2$. Our aim in this note is to show that product-free sets cannot be any larger; our main result is as follows.

Theorem 2.1. *Let $\mathcal{A}$ be a finite set. If $S \subset \mathcal{F}_\mathcal{A}$ is product-free, then $d^*(S) \leq 1/2$.*

Let us mention that product-free sets in cancellative semigroups have been studied by Łuczak and Schoen [6]; while their results are sharp for such semigroups in general, these results do not give us any effective bounds on the size of a product-free subset of $\mathcal{F}$.

Before we turn to the proof of Theorem 2.1, it is worth pointing out that there is a simple argument that allows us to bound the upper Banach density of a product-free subset of $\mathcal{F}$ away from 1. Indeed, suppose that $S \subset \mathcal{F}$ is product-free. We then have

$$d_S(m)d_S(n) + d_S(m + n) \leq 1$$

for any $m, n \in \mathbb{N}$ since the sets $S \cap \mathcal{F}(m + n)$ and $(S \cap \mathcal{F}(m)) \bullet (S \cap \mathcal{F}(n))$ must be disjoint. By averaging this inequality, with m fixed and with n ranging over suitably long intervals, we learn that $(d_S(m) + 1)d^*(S) \leq 1$ for each $m \in \mathbb{N}$. It

is now immediate that $d^*(S) \le \phi$, where $\phi = (\sqrt{5} - 1)/2 \approx 0.618$ is the unique positive solution to the equation $x^2 + x = 1$.

We shall have to work somewhat harder to prove Theorem 2.1, which improves this bound of ϕ for the upper Banach density to the optimal bound of $1/2$. The proof of Theorem 2.1 is given in Section 3. We conclude this note with a discussion of some open problems in Section 4.

3. Proof of the main result

We begin by fixing our finite alphabet $\mathcal{A}$. In the sequel, $\mathcal{F}$ will always mean $\mathcal{F}_{\mathcal{A}}$, the free semigroup over this fixed alphabet $\mathcal{A}$.

It will be helpful to establish some notation. For a pair of words $x, w \in \mathcal{F}$, we say that x is a *prefix* of w if $w = x \bullet y$ for some $y \in \mathcal{F}$, and that x is a *suffix* of w if $w = y \bullet x$ for some $y \in \mathcal{F}$. For a pair of sets $S_1, S_2 \subset \mathcal{F}$, we write $S_1 \bullet S_2$ for their (Minkowski) product; in other words,

$$S_1 \bullet S_2 = \{w_1 \bullet w_2 : w_1 \in S_1, w_2 \in S_2\}.$$

For a set $S \subset \mathcal{F}$ and an integer $n \in \mathbb{N}$, we set $S(n) = S \cap \mathcal{F}(n)$. One of the key ideas in the proof of Theorem 2.1 is the following definition. For any sequence of positive integers $\ell_1 < \ell_2 < \cdots < \ell_k < n$, we define

$$S(n; \ell_1, \ell_2, \ldots, \ell_k) = \{w \in S(n) : w \text{ has no prefix in } S(\ell_1) \cup S(\ell_2) \cup \cdots \cup S(\ell_k)\};$$

in other words,

$$S(n; \ell_1, \ell_2, \ldots, \ell_k) = S(n) \setminus \left(\bigcup_{i=1}^{k} S(\ell_i) \bullet \mathcal{F}(n - \ell_i) \right).$$

Let us note, for any $S \subset \mathcal{F}$, that the sets $S(n; m)$ and $S(m) \bullet \mathcal{F}(n - m)$ are disjoint for any pair of positive integers $m < n$. Recall that $d_S(n) = |S(n)||\mathcal{F}(n)|^{-1}$; we analogously define

$$d_S(n; \ell_1, \ell_2, \ldots, \ell_k) = \frac{|S(n; \ell_1, \ell_2, \ldots, \ell_k)|}{|\mathcal{F}(n)|}.$$

When the set S in question is clear, we write $d(n)$ and $d(n; \ell_1, \ell_2, \ldots, \ell_k)$ for $d_S(n)$ and $d_S(n; \ell_1, \ell_2, \ldots, \ell_k)$, respectively. Recall that for any product-free set $S \subset \mathcal{F}$ and any $m, n \in \mathbb{N}$, we have

$$d(m)d(n) + d(m + n) \le 1.$$

4

We start by proving a generalisation of this fact.

Proposition 3.1. *If $S \subset \mathcal{F}$ is product-free, then for any sequence of positive integers $\ell_1 < \ell_2 < \cdots < \ell_k < n$, we have*

$$d(\ell_1)d(n - \ell_1) + d(\ell_2; \ell_1)d(n - \ell_2) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1})d(n - \ell_k) + d(n)$$

$$\leq d(\ell_1) + d(\ell_2; \ell_1) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1}) + d(n; \ell_1, \ell_2, \ldots, \ell_k) \leq 1.$$

Proof. First, consider the products

$$S(\ell_1) \bullet S(n - \ell_1), S(\ell_2; \ell_1) \bullet S(n - \ell_2), \ldots, S(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1}) \bullet S(n - \ell_k).$$

These subsets of $\mathcal{F}(n)$ are by definition pairwise disjoint. Let L' be the union of these k sets. Since S is product-free, L' and $S(n)$ are disjoint as well. Let $L = L' \cup S(n)$; clearly, the density of L in $\mathcal{F}(n)$ is

$$d(\ell_1)d(n - \ell_1) + d(\ell_2; \ell_1)d(n - \ell_2) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1})d(n - \ell_k) + d(n).$$

Next, consider the Minkowski products

$$S(\ell_1) \bullet \mathcal{F}(n - \ell_1), \ S(\ell_2; \ell_1) \bullet \mathcal{F}(n - \ell_2), \ \ldots, \ S(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1}) \bullet \mathcal{F}(n - \ell_k).$$

These subsets of $\mathcal{F}(n)$ are again pairwise disjoint by definition; let R' denote their union. Note that R' and $S(n; \ell_1, \ell_2, \ldots, \ell_k)$ are disjoint. Let $R = R' \cup S(n; \ell_1, \ell_2, \ldots, \ell_k)$; it is easy to see that the density of R in $\mathcal{F}(n)$ is

$$d(\ell_1) + d(\ell_2; \ell_1) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1}) + d(n; \ell_1, \ell_2, \ldots, \ell_k)$$

and that this quantity is therefore at most 1.

To finish the proof, it suffices to show that $L \subset R$. It is easy to see that $L' \subset R'$. It remains to show that $S(n)$ is a subset of $R' \cup S(n; \ell_1, \ell_2, \ldots, \ell_k)$. To see this, note that any word from $S(n)$ which has a prefix in $S(\ell_1) \cup S(\ell_2) \cup \cdots \cup S(\ell_k)$ belongs to R'. In other words, $S(n) \setminus S(n; \ell_1, \ell_2, \ldots, \ell_k) \subset R'$; the result follows. $\square$

With the above observation in hand, we are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. We prove by contradiction that the upper Banach density of a product-free set is at most $1/2$.

Suppose that $S \subset \mathcal{F}$ is product-free and that $d^*(S) > 1/2 + \varepsilon$ for some $\varepsilon > 0$. We then claim that we may find an increasing sequence of positive integers $(\ell_k)_{k \in \mathbb{N}}$ such that

$$d(\ell_1) + d(\ell_2; \ell_1) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1}) \geq \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^k} = 1 - \frac{1}{2^k}$$

for each $k \in \mathbb{N}$.

We construct this sequence inductively. Since $d^*(S) > 1/2$, it is clear that we may find $\ell_1 \in \mathbb{N}$ such that $d(\ell_1) \geq 1/2$. Having found $\ell_1 < \ell_2 < \cdots < \ell_k$ as required, we choose ℓ_{k+1} as follows. Since $d^*(S) > 1/2 + \varepsilon$, there exist arbitrarily long intervals $I \subset \mathbb{N}$ that satisfy

$$\frac{\sum_{n \in I} d(n)}{|I|} > \frac{1}{2} + \varepsilon.$$

Choose such an interval I whose length is sufficiently larger than ℓ_k; we may assume (by passing to a sub-interval if necessary) that $\min I > \ell_k$, and that $|I| > (\ell_k + 1)/\varepsilon$. We claim that it is possible to choose ℓ_{k+1} from I; in other words, we claim that there exists an $n \in I$ such that

$$d(\ell_1) + d(\ell_2; \ell_1) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1}) + d(n; \ell_1, \ell_2, \ldots, \ell_k) \geq 1 - \frac{1}{2^{k+1}}.$$

We prove this claim by contradiction. Suppose that there is no such $n \in I$. Then, by Proposition 3.1, we have

$$d(\ell_1)d(n - \ell_1) + d(\ell_2; \ell_1)d(n - \ell_2) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1})d(n - \ell_k) + d(n)$$

$$\leq d(\ell_1) + d(\ell_2; \ell_1) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1}) + d(n; \ell_1, \ell_2, \ldots, \ell_k) < 1 - \frac{1}{2^{k+1}}$$

for each $n \in I$. By summing the above bound over all $n \in I$, we get

$$\sum_{n \in I'} d(n)(1 + d(\ell_1) + d(\ell_2; \ell_1) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1})) < |I| \left(1 - \frac{1}{2^{k+1}}\right),$$

where $I' \subset I$ is the set of $n \in I$ with $n + \ell_k < \max I$; indeed, to see this, note that for every $1 \leq i \leq k$ and every $n' \in I'$, there is an $n \in I$ such that $d(n') = d(n - \ell_i)$. This implies, by the inductive hypothesis, that

$$\sum_{n \in I'} d(n)\left(2 - \frac{1}{2^k}\right) < |I|\left(1 - \frac{1}{2^{k+1}}\right),$$

or equivalently, $\sum_{n \in I'} d(n) < |I|/2$. Therefore, we have

$$\sum_{n \in I} d(n) \leq \sum_{n \in I'} d(n) + \ell_k + 1 < \frac{|I|}{2} + \ell_k + 1,$$

which contradicts the fact that $\sum_{n \in I} d(n) > |I|/2 + \varepsilon|I|$, provided $|I| > (\ell_k + 1)/\varepsilon$.

We now finish the proof of the proposition by showing that the existence of this sequence $(\ell_k)_{k \in \mathbb{N}}$ contradicts our initial assumption that $d^*(S) > 1/2 + \varepsilon$. Fix a $k \in \mathbb{N}$ large enough to ensure that

$$\frac{2^k}{2^{k+1} - 1} < \frac{1 + \varepsilon}{2}$$

and consider any interval $I \subset \mathbb{N}$ with $|I| > 4(\ell_k + 1)/\varepsilon$. We know from Proposition 3.1 that

$$d(\ell_1)d(n - \ell_1) + d(\ell_2; \ell_1)d(n - \ell_2) + \cdots + d(\ell_k; \ell_1, \ell_2, \ldots, \ell_{k-1})d(n - \ell_k) + d(n) \leq 1$$

for each $n \in \mathbb{N}$ with $n > \ell_k$; as earlier, we sum the above bound over such $n \in I$ to get

$$\sum_{n \in I'} d(n)\left(2 - \frac{1}{2^k}\right) \leq |I|,$$

where I' is the set of $n \in I$ with $n > \ell_k$ and $n + \ell_k < \max I$. Therefore,

$$\frac{\sum_{n \in I} d(n)}{|I|} \leq \frac{2^k}{(2^{k+1} - 1)} + \frac{2(\ell_k + 1)}{|I|} < \frac{1}{2} + \varepsilon,$$

which is a contradiction; this proves the claimed upper bound in Theorem 2.1. $\square$

4. CONCLUSION

A common line of enquiry in the study of product-free sets is to ask for 'asymmetric' versions of results bounding the upper density of product-free sets. In this spirit, it is natural to ask whether an analogue of Theorem 2.1 continues to hold when one wishes to solve the equation $x \bullet y = z$ with x, y and z in specified subsets of $\mathcal{F}$. More precisely, if $X, Y, Z \subset \mathcal{F}$ are such that there are no solutions to $x \bullet y = z$ with $x \in X$, $y \in Y$ and $z \in Z$, one might ask if one of X, Y or Z has an upper asymptotic density of at most $1/2$. However, it is possible to construct for any $\varepsilon > 0$, three sets $X, Y, Z \subset \mathcal{F}$, each of upper asymptotic density at least $\phi - \varepsilon$, where $\phi = (\sqrt{5} - 1)/2$, such that there are no solutions to $x \bullet y = z$ with $x \in X$, $y \in Y$ and $z \in Z$; of course, the simple argument given in the introduction shows that one cannot do any better than this. Pick a suitably large $n \in \mathbb{N}$ and choose any set $W \subset \mathcal{F}(n)$ such that $||W|/|\mathcal{F}(n)| - \phi| < \varepsilon/3$. Now take X to be the set of all words with a prefix in W, Y to be the set of all words with a suffix in W, and Z to be the set $\mathcal{F} \setminus (X \bullet Y)$. Clearly, there are no solutions to $x \bullet y = z$ with $x \in X$, $y \in Y$ and $z \in Z$; it is also not hard to check that each of X, Y and Z has an upper asymptotic density at least $\phi - \varepsilon$.

Next, it would be interesting to understand what product-free sets of maximal density look like. As we saw earlier, several non-isomorphic extremal constructions are furnished by the family of odd-occurrence sets. We suspect that these might be the only constructions of maximal density, and make the following conjecture.

Conjecture 4.1. *Let $\mathcal{A}$ be a finite set. If $S \subset \mathcal{F}_\mathcal{A}$ is product-free and $d^*(S) = 1/2$, then $S \subset \mathcal{O}_\Gamma$ for some nonempty subset $\Gamma \subset \mathcal{A}$.*

Finally, another natural direction is to study product-free subsets of the *free group* $\mathbf{F}_\mathcal{A}$ over a finite alphabet $\mathcal{A}$. Similarly to the situation in this paper, the most natural measure to consider in the case of the free group $\mathbf{F}_\mathcal{A}$ would be the one that assigns a weight of $|\mathcal{A}|(|\mathcal{A}| - 1)^{-(n-1)}$ to each irreducible word of length n. The different notions of density defined here for the free semigroup then have analogous definitions in the free group, and we believe that an analogue of Theorem 2.1 should hold in the free group as well; concretely, we conjecture the following.

Conjecture 4.2. *For any finite alphabet $\mathcal{A}$, no product-free subset of the free group $\mathbf{F}_\mathcal{A}$ has upper Banach density exceeding $1/2$.*

Note that, in the proof of Theorem 2.1, we rely crucially on the fact that there is exactly one way to write a word of length $m + n$ as the concatenation of a word of length m with a word of length n; of course, we lose this property when working with free groups, so we believe that some new ideas will be required to understand product-free sets in free groups.

Acknowledgements

References

1. L. Babai and V. T. Sós, *Sidon sets in groups and induced subgraphs of Cayley graphs*, European J. Combin. **6** (1985), 101–114.
2. S. Eberhard, *Product mixing in the alternating group*, Discrete Analysis (2016:2), 19 pp.
3. W. T. Gowers, *Quasirandom groups*, Combin. Probab. Comput. **17** (2008), 363–387.

4. B. Green and I. Z. Ruzsa, *Sum-free sets in abelian groups*, Israel J. Math. **147** (2005), 157–188.

5. K. S. Kedlaya, *Product-free subsets of groups, then and now*, Communicating mathematics, Contemp. Math., vol. 479, Amer. Math. Soc., Providence, RI, 2009, pp. 169–177.

6. T. Łuczak and T. Schoen, *Sum-free subsets of right cancellative semigroups*, European J. Combin. **22** (2001), 999–1002.

7. T. Tao and V. Vu, *Sumfree sets in groups: a survey*, J. Comb. **8** (2017), 541–552.

Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK

Email address: `i.leader@dpmms.cam.ac.uk`

ETH Institute for Theoretical Studies, 8092 Zurich, Switzerland

Email address: `shoham.letzter@math.ethz.ch`

Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA

Email address: `narayanan@math.rutgers.edu`

School of Mathematical Sciences, Queen Mary, University of London, London E1 4NS, UK

Email address: `m.walters@qmul.ac.uk`