SYMMETRIC SUBSETS OF LATTICE PATHS

Oleg Verbitsky[1]
Institute of Information Systems, Vienna University of Technology, Favoritenstraße 9, 1090 Wien, Austria
oleg@dbai.tuwien.ac.at

Received: 11/10/99, Revised: 3/30/00, Accepted: 4/13/00, Published: 5/19/00

Abstract

Let $[n] = \{0, 1, \ldots, n\}$. A subset S of $[n]$ is symmetric if $S = g - S$ for a natural number g. We show that, for every $f : [n] \to [2n]$ with the restriction $1 \le f(i+1) - f(i) \le 2$ for all $i < n$, there is some $S \subset [n]$ such that $|S| \ge 2 \ln n - O(1)$, with the property that both S and $f(S)$ are symmetric. We prove this result by finding a lower bound for the length of a symmetric pattern whose abelian occurrences are encountered in all binary words of length n. We also show that if M is such that for every $f : [n] \to [2n]$ as above, there is at least one $S \subset [n]$ with $|S| \ge M$ and both S and $f(S)$ symmetric, then $M \le (7 + o(1))\sqrt{n}$. This result is based on the construction of appropriate Sidon B_2-sequences. In another interpretation, our results can be formulated as lower and upper bounds for M such that every path of length n along basis vectors of a two-dimensional lattice contains an M-point centrally symmetric set.

1. Introduction

Let $[n] = \{0, 1, 2, \ldots, n\}$, with 0 included for our convenience. We consider injective order-preserving transformations $f : [n] \to [2n]$ with restriction $f(i+1) - f(i) \le 2$ for all $i < n$. We wonder to which extent such transformations can violate the regular structure of $[n]$. Namely, suppose that $\mathcal{P}$ is a regularity property of a set of integers, say, one of being an arithmetic progression. We then wish to know the maximum $M = M(n)$ such that, for every f as above, at least one set $S \subseteq [n]$ with $|S| \ge M$ has property $\mathcal{P}$ and its image $f(S)$ still has the same property.

In the case of arithmetic progressions, it is easy to observe an equivalent reformulation of the question. Let $V = \{v_0, \ldots, v_n\}$ be a sequence of points in the grid $\mathbb{Z}^2$ with each difference $v_{i+1} - v_i$ being either $a = (1, 1)$ or $b = (1, 2)$. Now the question is what is the maximum M such that every V contains an M-term arithmetic progression of vectors. To see the equivalence of the two problems, it suffices to view a set V as the graph of a map f. Clearly, f preserves an arithmetic progression $S \subseteq [n]$ iff $\{\,(x, f(x)) : x \in S\}$ is an arithmetic progression in V. Notice

[1]On leave from the Department of Mechanics & Mathematics, Lviv University, Ukraine. Research supported in part by a Lise Meitner Fellowship of the Austrian Science Foundation (FWF) and by grant INTAS-96-0753.

that the specification of differences a and b is actually irrelevant — those could be any other pair of non-collinear vectors as well, say, $a = (1, 0)$ and $b = (0, 1)$.

As the choice of the initial point v_0 does not affect anything, a set V is characterized by the sequence of differences $v_1 - v_0, \ldots, v_n - v_{n-1}$, which can be regarded as a word $w(V)$ of length n over alphabet $\{a, b\}$. In this way we arrive at yet another reformulation of the problem under consideration. We call an arbitrary sequence of variables a *pattern*. An *abelian occurrence* of a pattern in a word is a subword obtainable from the pattern by substituting nonempty words in place of variables so that words replacing the same variable may differ only in order of letters (see Section 2 for more details). It is not hard to observe a one-to-one correspondence between $(m+1)$-term arithmetic progressions in V and abelian occurrences of the pattern x^m in $w(V)$. Thus, the value of $M(n)$ is the maximum number M such that every word of length n over the binary alphabet has an abelian occurrence of x^{M-1}.

Dekking [11] constructs an infinite word in the binary alphabet without abelian occurrences of x^4. It immediately follows [21, theorem 6.13] that $M(n) \le 4$, i.e. 5-term arithmetic progressions can all be destroyed by some transformation f.

This motivates an extension of property $\mathcal{P}$. A set $S \subseteq \mathbb{Z}^k$ such that $S = g - S$ for a lattice point $g \in \mathbb{Z}^k$ is called *symmetric* (with respect to the center at rational point $\frac{1}{2}g$). From now on the property $\mathcal{P}$ extended to being symmetric will be our main concern. Given $V \subseteq \mathbb{Z}^k$, let $MS(V)$ denote the maximum cardinality of a symmetric subset of V.

A pattern is *symmetric* if it reads the same backward as forward, like xyx. With notation introduced above, we again have a one-to-one correspondence between sets $S \subseteq [n]$ whose symmetry is preserved by f, symmetric subsets of the graph V of f, and abelian occurrences of symmetric patterns in the word $w(V)$. Correspondingly, we have the following equivalences whose proof is given in more detail in Section 2.

Lemma 1.1 *The statements below are equivalent.*

1. $M(n) = \min_{f:[n]\to[2n]} \max_{S\subseteq[n]} \{|S| : \text{both } S \text{ and } f(S) \text{ are symmetric}\}$, *where the minimum is taken over all f with*

$$1 \le f(i+1) - f(i) \le 2 \quad \text{for} \quad i < n. \tag{1}$$

2. *$M(n)$ is the minimum of $MS(V)$ over all subsets $V = \{v_0, v_1, \ldots, v_n\}$ of $\mathbb{Z}^2$ with each $v_{i+1} - v_i$ equal to either a or b, where a and b are arbitrarily fixed non-collinear vectors.*

3. *$M(n)$ is the maximum M such that every word of length n over the binary alphabet has abelian occurrence of a symmetric pattern of length at least $M - 1$.*

In contrast to the case of arithmetic progressions, $M(n)$ now grows with n, that is, no f is able to destroy symmetric subsets so well as arithmetic progressions. To show this, consider an infinite sequence of symmetric patterns

$$\begin{aligned} P_1 &= x, \\ P_2 &= xyx, \\ P_3 &= xyxzxyx, \\ P_4 &= xyxzxyxuxyxzxyx, \\ &\vdots \end{aligned} \quad (2)$$

where P_{i+1} is the result of inserting a new variable between two copies of P_i. In combinatorics of words, members of this sequence are called *sesquipowers* or *Zimin's patterns*. Coudrain and Schützenberger [10] proved that each P_i must occur in all long enough words over a finite alphabet. Here we mean literal rather than abelian occurrence, i.e. the same variable is substituted everywhere by the same word. The unavoidability of sesquipowers immediately implies that $M(n)$ goes to the infinity with n increasing. However, this argument gives a very small lower bound for $M(n)$, actually, a kind of the inverse tower function (see Lemma 2.3).

In Section 3 we prove a better lower bound $M(n) = \Omega(\ln n)$ based on estimation of how long symmetric pattern is represented by an abelian occurrence in every binary word of length n. Similarly to the O-notation, we write $\Omega(h(n))$ to refer to a function of n that everywhere exceeds $c \cdot h(n)$ for a positive constant c.

In Section 4 we prove upper bound $M(n) = O(\sqrt{n})$. As the main technical tool we use B_2-sequences introduced by Sidon and investigated by many authors (see [21, section 4.1] for survey and references). A set X of integers is called a *B_2-sequence* if for any integer g the equation $x + y = g$ has at most one solution in X with $x \leq y$. In other words, a B_2-sequence X is a highly asymmetric set characterized by $MS(X) \leq 2$. There are several constructions [24, 6, 13, 14, 9, 7, 19] of dense B_2-sequences in $[n]$. We employ a fairly simple and explicit construction of [19], making use of an additional uniformity property of it.

We conclude the paper with discussion of open problems in Section 5.

Related work. The van der Waerden theorem can be restated so that every infinite subsequence $v_0, v_1, \ldots$ of $\mathbb{N}$ with $v_{i+1} - v_i = O(1)$ contains arbitrarily long arithmetic progressions (see [8]). As Dekking's result shows, a similar statement in $\mathbb{Z}^2$ is false. However, Ramsey and Gerver [23] prove that every infinite sequence $v_0, v_1, \ldots$ in $\mathbb{Z}^2$ with bounded distances $||v_{i+1} - v_i||$ between any two successive points contains arbitrarily large subsets of collinear points. Pomerance [22] shows this holds true even under the weaker assumption that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} ||v_{i+1} - v_i|| < \infty. \quad (3)$$

These results can be viewed as two-dimensional analogs of the van der Waerden theorem and its density version of Szemerédi, with collinear subsets instead of arithmetic progressions. In this respect our result on behavior of $M(n)$, in view of item 2 of Lemma 1.1, can serve as yet another two-dimensional analog of van der Waerden's theorem, with arithmetic progressions replaced by symmetric subsets. The multi-dimensional analog of Szemerédi's theorem is also true as shown by Banakh [5], who observed that condition (3) guarantees the existence of arbitrarily long symmetric subsequences in an infinite sequence $v_0, v_1, \ldots$ of points in $\mathbb{Z}^k$, $k \geq 1$. It should be noted that in the case of $k = 2$ the latter result strengthes the claim that $M(n) \to \infty$ but provides no satisfactory lower bound for $M(n)$.

Banakh and Protasov [3, 4] prove that the minimal number of colors required for coloring the n-dimensional integer grid $\mathbb{Z}^n$ avoiding infinite symmetric monochromatic subsets is $n + 1$. Unavoidable symmetries in words are investigated by Fouché [16].

2. Preliminaries

In this section we prove Lemma 1.1 and then show that $M(n) \to \infty$ as $n \to \infty$. Recall that throughout the paper $MS(V)$ denotes the cardinality of the largest symmetric subset of V.

The proof of the equivalence of statements 1 and 2 of Lemma 1.1 in the case that

$$a = (1, 1), \quad b = (1, 2) \tag{4}$$

follows arguments outlined in the introduction for arithmetic progressions. With a function f we associate its graph $V = \{v_0, \ldots, v_n\}$, where $v_i = (i, f(i))$. The bounds (1) imply that $v_{i+1} - v_i \in \{a, b\}$. Vice versa, any set $V = \{v_0, \ldots, v_n\}$ in $\mathbb{Z}^2$ with the latter condition can be viewed as the graph of a function f of the prescribed kind. A set $S \subseteq [n]$ and its image $f(S)$ are both symmetric iff $S' = \{\, (i, f(i)) : i \in S\}$ is a symmetric subset of V. This completes the proof in the case (4).

The case of arbitrary non-collinear a and b reduces to the case (4). Really, consider two sets $V = \{v_0, \ldots, v_n\}$ and $V' = \{v'_0, \ldots, v'_n\}$ in $\mathbb{Z}^2$ with all $v_{i+1} - v_i \in \{(1, 1), (1, 2)\}$ and $v'_{i+1} - v'_i \in \{a, b\}$, where a and b are non-collinear. Let ϕ be the affine transformation of $\mathbb{Z}^2$ into itself that takes v_0 to v'_0, $(1, 1)$ to a, and $(1, 2)$ to b. Then ϕ establishes a one-to-one correspondence between V and V' that matches symmetric subsets in V and symmetric subsets in V'. It follows that $MS(V) = MS(V')$, thereby proving the equivalence of statements 1 and 2.

Before proving the equivalence of statements 2 and 3, let us recall the relevant notions of the formal language theory. A *pattern* is a word over the alphabet of variables $\{x_1, x_2, \ldots\}$. Pattern $x_{i_1} x_{i_2} \ldots x_{i_l}$ is *symmetric* if $i_j = i_{l+1-j}$ for all $j \leq l$. Let $A = \{a_1, \ldots, a_m\}$ be a finite alphabet. The number of occurrences of letter a_i in a word w over A is denoted by $|w|_{a_i}$. A *commutative index* of w is the tuple $\langle |w|_{a_1}, \ldots, |w|_{a_m} \rangle$. A subword u of a word w is an *occurrence* of a pattern $P = x_{i_1} \ldots x_{i_l}$ if u can be obtained from P by substituting nonempty words in place of each variable, where the same variable is everywhere replaced with the same word. If the same variable may be replaced by (possibly distinct) words with the same commutative index, u is called an *abelian occurrence* of P.

Example. In word $a_1 a_1 a_2 a_1 a_2 a_1 a_3$, subwords $a_1 a_1$, $a_1 a_2 a_1 a_2$, and $a_2 a_1 a_2 a_1$ are occurrences of pattern x_1x_1. In addition, $a_1a_1a_2a_1a_2a_1$ is an abelian occurrence of the same pattern.

Given a sequence of vectors $V = \{v_0, v_1, \ldots, v_n\}$ in $\mathbb{Z}^k$ with all $v_i - v_{i-1}$ in a finite set $A \subset \mathbb{Z}^k$, we associate with V the sequence $w(V)$ of differences $v_1 - v_0, v_2 - v_1, \ldots, v_k - v_{k-1}$ which will be viewed as a word of length n over alphabet A.

Lemma 2.1

1. *If $w(V)$ has an abelian occurrence of a symmetric pattern of length l, then $MS(V) \geq l+1$.*
2. *Conversely, suppose that A is a linearly independent set of vectors. Then $w(V)$ has an abelian occurrence of a symmetric pattern of length at least $MS(V) - 1$.*

Proof. 1. Recall that word $w(V)$ is a sequence of vectors $v_1 - v_0, \ldots, v_n - v_{n-1}$. Given a subword $u = v_{i+1} - v_i \ldots v_j - v_{j-1}$, $i < j$, we call v_i the initial point and v_j the terminal point of u. Let $u = u_1 \ldots u_l$ be an abelian occurrence of a symmetric pattern P of length l, where u_s is substituted in place of s-th variable of P. Let $v_{i_{s-1}}$ and v_{i_s} be the initial and terminal points of u_s. Then the set $\{v_{i_0}, \ldots, v_{i_l}\}$ is symmetric. This can be shown by easy induction. Really, assume that v_{i_1} and $v_{i_{l-1}}$ are symmetric with respect to the center $\frac{1}{2}g$, that is, $v_{i_1} + v_{i_{l-1}} = g$. As u_1 and u_l differ only in order of their letters, we have $v_{i_1} - v_{i_0} = v_{i_l} - v_{i_{l-1}}$. Consequently, v_{i_0} and v_{i_l} are symmetric with respect to $\frac{1}{2}g$ too.

2. Let $l = MS(V)$ and $v_{i_0}, \ldots, v_{i_l}$ be a symmetric subsequence of V. Denote a subword of $w(V)$ whose initial and terminal points are $v_{i_{s-1}}$ and v_{i_s} by u_s. Then $u = u_i \ldots u_l$ is an abelian occurrence of a symmetric pattern of length l. It suffices to show that commutative indices of words u_s and u_{l+1-s} are the same. Those are uniquely determined by expansions of vectors $v_{i_s} - v_{i_{s-1}}$ and $v_{i_{l+1-s}} - v_{i_{l-s}}$ in basis A. It remains to notice that the last two vectors are equal by symmetricalness of $\{v_{i_0}, \ldots, v_{i_l}\}$. □

The equivalence of statements 2 and 3 of Lemma 1.1 now follows directly from Lemma 2.1. The proof of Lemma 1.1 is complete.

Proposition 2.2 *$M(n) \to \infty$ as $n \to \infty$.*

At this point we prefer the statement 3 of Lemma 1.1. Let $L^{non-abel}(n)$ be the maximal l such that every word of length n over the binary alphabet has an occurrence of a symmetric pattern of length at least l. As $M(n) > L^{non-abel}(n)$, it suffices to show that $L^{non-abel}(n) \to \infty$ for $n \to \infty$. The latter follows from a result of Coudrain and Schützenberger [10] which we state below in a form convenient for our purposes.

Lemma 2.3 *([10]) If $L^{non-abel}(n) \geq l$, then $L^{non-abel}((n+1)(2^n+1)) \geq 2l+1$.*

Proof. Assume that every binary word of length n has occurrence of a symmetric pattern P of length l. Any binary word of length $(n+1)(2^n+1)$ contains two identical subwords of length n separated by a nonempty word. Thus, there is an occurrence of the symmetric pattern PxP, where x is a new variable absent in P. □

Notice that the above argument ensures that each pattern P_i of the sequence (2) occurs in any long enough binary word.

3. Lower Bound

The proof of Proposition 2.2 based on Lemma 2.3 gives us an extremely small lower bound for $M(n)$ that is even smaller than the inverse tower function. In this section we improve it to $M(n) \geq 2\ln n - O(1)$. We first prove an auxiliary fact. Notice that whenever below we refer to the number of subwords of a word, we distinguish all occurrences of a subword, that is, a subword is counted each time it occurs in the word.

Lemma 3.1 *Given a word w, let $\nu(w)$ denote the number of pairs $\{u_1, u_2\}$, where u_1 and u_2 are disjoint subwords of w with the same commutative index. Let $N(n)$ be the minimum of $\nu(w)$ over all binary words w of length n. Then*

$$N(n) \geq (\ln n - O(1))n^2/4.$$

Proof. Consider a binary word w of length n and estimate the value $\nu(w)$ from below. Expand $\nu(w)$ to the sum $\sum_t \nu_t(w)$, where the t-th term counts pairs of subwords with length t. Let $\sigma_t(i)$ denote the number of subwords of w with length t and commutative index $\langle i, t-i\rangle$. As the total number of subwords of length t is equal to $n+1-t$, notice the equality $\sigma_t(0) + \sigma_t(1) + \ldots + \sigma_t(t) = n+1-t$. As a subword of length t can overlap with at most $2t-1$ subwords of the same length, we have

$$\nu_t(w) \geq \frac{1}{2}\sum_{i=0}^{t} \sigma_t(i)(\sigma_t(i) - (2t-1)).$$

Taking into account that

$$\sum_{i=0}^{t} \sigma_t(i)^2 \geq (t+1)\left(\frac{\sum_{i=0}^{t}\sigma_t(i)}{t+1}\right)^2,$$

we conclude that

$$\begin{aligned} \nu_t(w) &\geq \frac{1}{2}\left((t+1)\left(\frac{n+1-t}{t+1}\right)^2 - (2t-1)(n+1-t)\right) \\ &= \frac{(n+2)^2}{2(t+1)} - (t+\frac{1}{2})(n+1) + t^2 - \frac{1}{2}. \end{aligned}$$

Let us sum these inequalities over t from 1 to s, dropping the last term $t^2 - \frac{1}{2}$ in the right hand side (anyway it would give us no gain). Summing the first term in the right hand side, we take into account that $\sum_{t=1}^{s} 1/t - \ln s$ approaches Euler's constant as s increases. Therefore,

$$\sum_{t=1}^{s} \nu_t(w) \geq \frac{1}{2}(\ln s - O(1))(n+2)^2 - \frac{s(s+2)}{2}(n+1).$$

Setting $s = \lceil \sqrt{n}\, \rceil$, we obtain the proclaimed bound for $\nu(w)$ and hence for $N(n)$. □

Theorem 3.2 $M(n) \geq 2\ln n - O(1)$.

Proof. We adhere to the statement 2 of Lemma 1.1. Let $V = \{v_0, v_1, \ldots, v_n\}$ be a set of points in $\mathbb{Z}^2$ with $v_{i+1} - v_i \in \{a, b\}$. Denote $G = \left\{ \frac{1}{2}(v_i + v_j) : 0 \leq i \leq j \leq n \right\}$, the set of all potential centers of symmetry. Let m_g denote the "multiplicity" of an element g in G, that is, the number of pairs (i, j) such that $g = \frac{1}{2}(v_i + v_j)$ and $i \leq j$. Clearly,

$$\sum_{g \in G} m_g = (n+1)(n+2)/2.$$

Furthermore, let N denote the total number of quadruples

$$(v_l, v_i, v_j, v_k) \text{ with } l < i \leq j < k \text{ and } v_i - v_l = v_k - v_j. \tag{5}$$

Clearly,

$$N \leq \sum_{g \in G} \binom{m_g}{2}$$

(actually, the linear independence of a and b implies the equality here). It follows that

$$N < \frac{1}{2} \sum_{g \in G} m_g^2 \leq \frac{1}{2} \left(\max_{g \in G} m_g \right) \sum_{g \in G} m_g = \frac{1}{4} n^2 \left(1 + O(\frac{1}{n}) \right) \max_{g \in G} m_g. \tag{6}$$

Recall that with the set V we associate a word $w(V)$ over alphabet $\{a, b\}$. It is easy to observe a one-to-one correspondence between quadruples (5) in V and pairs of disjoint subwords u_1 and u_2 with the same commutative index in $w(V)$. By Lemma 3.1 we have

$$N \geq (\ln n - O(1)) n^2/4.$$

Together with (6), this gives

$$\max_{g \in G} m_g \geq \ln n - O(1).$$

It remains to observe that, for every center $g \in G$, the set V contains a subset that is symmetric with respect to g and has at least $2m_g - 1$ elements. □

4. Upper Bound

In this section we prove an upper bound for $M(n)$.

Theorem 4.1 $M(n) \le (7 + o(1))\sqrt{n}$.

We use a two-dimensional geometric interpretation of $M(n)$ given by statement 2 of Lemma 1.1. We will construct a set $V = \{v_0, v_1, \dots, v_n\}$ of points in $\mathbb{Z}^2$ such that each difference $v_{i+1} - v_i$ is either $(1,0)$ or $(0,1)$ and $MS(V) \le (7 + o(1))\sqrt{n}$.

Our construction will be completely determined by two sets of integers $X = \{x_1, \dots, x_q\}$ and $Y = \{y_1, \dots, y_q\}$ listed in the ascending order. Given X and Y, consider a sequence of points in $\mathbb{Z}^2$

$$(x_1, y_1),\ (x_2, y_1),\ (x_2, y_2),\ (x_3, y_2),\ (x_3, y_3),\ \dots,\ (x_q, y_q) \tag{7}$$

We define V by $V = V_1 \cup V_2$, where

$$V_1 = \bigcup_{i=1}^{q} \{\, (x_i, y) : y_{i-1} < y \le y_i\} \text{ and } V_2 = \bigcup_{i=1}^{q-1} \{\, (x, y_i) : x_i < x \le x_{i+1}\}$$

(we set $y_0 = y_1 - 1$ for convenience). Thus, (7) are "corner" points of V, at which difference $v_{i+1} - v_i$ changes its value from $(1,0)$ to $(0,1)$ or vice versa. Clearly, V consists of $x_q + y_q + 1 - x_1 - y_1$ points.

Given a set $Z = \{z_1, \dots, z_q\}$ of integers listed in the ascending order, define $D(Z) = \max_{1 \le i < q}(z_{i+1} - z_i)$.

Lemma 4.2 *Suppose that V has been constructed based on q-element sets X and Y as described above. Then*

$$MS(V) < MS(X)D(Y) + MS(Y)D(X) + 2q. \tag{8}$$

Proof. Let S be the maximum subset of V symmetric with respect to center $\frac{1}{2}g$, i.e. $S = V \cap (g - V)$. Clearly,

$$S = (V_1 \cap g - V_1) \cup (V_2 \cap g - V_2) \cup (V_1 \cap g - V_2) \cup (V_2 \cap g - V_1).$$

Let us estimate the cardinality of each member of the union.

$V_1 \cap g - V_1$ is a symmetric subset of V_1. As the projection of $V_1 \cap g - V_1$ onto the first coordinate is symmetric too, the cardinality of this projection does not exceed $MS(X)$. As any cut of V_1 by vertical line (i.e. along the second coordinate) containes at most $D(Y)$ points, we have $|V_1 \cap g - V_1| \le MS(X)D(Y)$. Similarly, $|V_2 \cap g - V_2| \le MS(Y)D(X)$.

Observe now that all points of V_1 differ in the second coordinate and have only q values for the first coordinate, while all points of V_2 differ in the first coordinate and have only q values for the second coordinate. As a consequence, both $V_1 \cap g - V_2$ and $V_2 \cap g - V_1$ have less than q points. The bound (8) follows. □

We now need to choose X and Y so as to make the right hand side of (8) as small as possible. The idea is to use a B_2-sequence $X = Y$, which gives us the best possible $MS(X) = MS(Y) = 2$. It easily follows from [13] that $D(X) \geq q(1 - o(1))$ for any B_2-sequence $X = \{x_1, \ldots, x_q\}$. We use a construction of [19] that provides us with $D(X) \leq (3 + o(1))q$.

Lemma 4.3 ***(Krückeberg [19])*** *For any prime q there is a sequence of integers $X = \{x_1, \ldots, x_q\}$ with $MS(X) = 2$ and $D(X) < 3q$. Moreover, $x_1 = 0$ and $x_q = 2q^2 - 2q - 1$.*

We include the proof of this lemma given in [19], because it contains a simple explicit construction of the needed B_2-sequences, thereby making our construction of V explicit too.

Proof. Set $x_{i+1} = 2qi - (i^2 \bmod q)$ for $0 \leq i < q$, where expression $i^2 \bmod q$ stands for the least non-negative residue of i^2 modulo q. Obviously, $q < x_{i+1} - x_i < 3q$. To show that X is a B_2-sequence, assume that $x_i + x_j = x_{i'} + x_{j'}$, $i \leq j$, $i' \leq j'$. It is easy to derive from this that

$$\begin{cases} i + j = i' + j' & \pmod q \\ i^2 + j^2 = (i')^2 + (j')^2 & \pmod q \end{cases}$$

Since in the field $\mathbb{F}_q$ a system of kind

$$\begin{cases} i + j = a \\ i^2 + j^2 = b \end{cases}$$

can have only a unique solution i, j with $i \leq j$, we conclude that $i = i'$ and $j = j'$. □

Let us summarize our construction of the set $V = \{v_0, v_1, \ldots, v_n\}$. Let q be the prime next to $(\sqrt{n+3} + 1)/2$ and X be the B_2-set given by Lemma 4.3. Applying the construction described in the beginning of the section with $Y = X$, we obtain a set $V' = \{v_0, v_1, \ldots, v_n, \ldots\}$ of $4q^2 - 4q - 1 \geq n + 1$ points in $\mathbb{Z}^2$. Leaving aside some last elements of V', we get the set V. By Lemma 4.2, $MS(V) \leq MS(V') < 14q$. Since the prime next to m does not exceed $m + O(m^\alpha)$, where $0 < \alpha < 1$ ([18], see also [2, pp. 225, 256] for references on the best current values of α), we have $MS(V) \leq (7 + o(1))\sqrt{n}$. The proof of Theorem 4.1 is complete.

Remark 4.4 The choice of Krückeberg's B_2-sequence is essentially best possible, because the right hand side of (8) cannot be smaller than $\sqrt{2n}$, whatever sets X and Y are. Let us prove this fact. First, observe relation

$$MS(X) \geq q/D(X). \tag{9}$$

for a set of integers $X = \{x_1, \ldots, x_q\}$. This is a consequence of inclusion $X \subseteq \bigcup_{g=0}^{D(X)-1}(g + x_1 + x_q - X)$ which implies $|X \cap (g - X)| \geq q/D(X)$ for some g. By (9)

$$MS(X)D(Y) + MS(Y)D(X) \geq 2(MS(X)D(Y)MS(Y)D(X))^{1/2} \geq 2q$$

and therefore the right hand side of (8) is at least $4q$.

Further, observe that $MS(V) > \max\{D(X), D(Y)\}$. Using this, we have $n = |V| - 1 \leq q(D(X) + D(Y)) < 2q\,MS(V)$. Therefore, the right hand side of (8) exceeds $2n/MS(V)$. It remains to notice that one of the values $MS(V)$ and $2n/MS(V)$ is at least $\sqrt{2n}$.

Remark 4.5 Consider a random set $\mathbf{V} = \{v_0, v_1, \dots, v_n\}$ in $\mathbb{Z}^2$ with $v_{i+1} - v_i \in \{a, b\}$ for non-collinear a and b. We mean that the underlying word $w(\mathbf{V})$ is uniformly distributed on $\{a, b\}^n$. The mean value of $MS(\mathbf{V})$ could serve as an upper bound for $M(n)$. Unfortunately, this probabilistic argument cannot give anything better that the constructive bound of Theorem 4.1 by the following reason.

Just for simplicity assume that $n = 2m$ is even. Let $\mathbf{s}$ denote the cardinality of the maximum subset of $\mathbf{V}$ symmetric with respect to the center at the medium point v_m. Consider now two independent sequences $\xi_1, \dots, \xi_m$ and $\zeta_1, \dots, \zeta_m$ of unbiased Bernoulli trials, that is, all ξ_i and ζ_j are mutually independent random variables that take on equiprobable values 0 and 1. Denote the number of k such that $\sum_{i=1}^k \xi_i = \sum_{i=1}^k \zeta_i$ by $\mathbf{t}$. In coding $a = 0$ and $b = 1$, it becomes clear that $\mathbf{s} = 2\mathbf{t} + 1$. Estimate the expectation of $\mathbf{t}$ from below.

Let $p_k = \mathbb{P}\left[\sum_{i=1}^k \xi_i = \sum_{i=1}^k \zeta_i\right]$. By linearity of mathematical expectation, $\mathbb{E}[\mathbf{t}] = \sum_{k=1}^m p_k$. Using Chernoff's bound, we have

$$p_k = \sum_{l=0}^{k} \mathbb{P}\left[\sum_{i=1}^{k} \xi_i = l\right]^2 > \sum_{k/2-\sqrt{k} \le l \le k/2+\sqrt{k}} \mathbb{P}\left[\sum_{i=1}^{k} \xi_i = l\right]^2 \ge$$
$$(2\sqrt{k} - 1)\left(\frac{\mathbb{P}\left[k/2 - \sqrt{k} \le \sum_{i=1}^k \xi_i \le k/2 + \sqrt{k}\right]}{2\sqrt{k}+1}\right)^2 \ge \frac{(1 - 2\exp(-2))^2}{2\sqrt{k}+7}.$$

Therefore, $\mathbb{E}[\mathbf{t}] = \Omega\left(\sum_{k=1}^m 1/\sqrt{k}\right) = \Omega(\sqrt{m})$. As $\mathbb{E}[\mathbf{s}] = 2\mathbb{E}[\mathbf{t}] + 1$, we conclude that the mean value of $MS(\mathbf{V})$ is $\Omega(\sqrt{n})$.

In conclusion we discuss one more aspect of the upper bound proven in this section. Given n, we have constructed a set $\{v_0, v_1, \dots, v_n\}$ with

$$MS(\{v_0, v_1, \dots, v_n\}) = O(\sqrt{n}). \tag{10}$$

Question 4.6 Is it possible to construct an infinite set $\{v_0, v_1, v_2 \dots\}$ such that (10) is true for all n?

We could achieve this goal with the same construction, if we had an infinite B_2-sequence $X = \{x_1, x_2, \dots\}$ with $D(\{x_1, \dots, x_q\}) = O(q)$ for all q. However, the latter condition implies $|X \cap [m]| = \Omega(\sqrt{m})$ for all m, whereas no B_2-sequence satisfies this condition by a result of Erdős. Erdős proves that there is a constant c such that for any infinite B_2-sequence X the inequality $|X \cap [m]| \le c\sqrt{m/\ln m}$ is true for infinitely many m (see [15]). The best known construction of [1] gives $|X \cap [m]| = \Omega\left((m \ln m)^{1/3}\right)$. Nevertheless, we are able at least to approach (10) with an infinite V.

Proposition 4.7 *There is an infinite sequence $V = \{v_0, v_1, v_2, \dots\}$ with each difference $v_{i+1} - v_i$ either $(1, 0)$ or $(0, 1)$ and such that*

$$MS(\{v_0, v_1, \dots, v_n\}) = n^{1/2+O(1/\ln\ln n)} \tag{11}$$

for all n.

Proof. We apply the straightforward infinite version of the construction described in the beginning of this section with $X = Y = \{1, 4, 9, 16, \ldots\}$, the set of integer squares. By Lemma 4.2, for any integer q and $n = 2q^2 - 2$

$$MS(\{v_0, v_1, \ldots, v_n\}) < 2MS(\{1, 4, \ldots, q^2\})D(\{1, 4, \ldots, q^2\}) + 2q.$$

We obviously have $D(\{1, 4, \ldots, q^2\}) = 2q - 1$ and, by Lemma 4.8 below,

$$MS(\{1, 4, \ldots, q^2\}) = q^{O(1/\ln\ln q)}.$$

This proves (11) for all $n = 2q^2 - 2$. Equality (11) is true for any other n as well, because the next to n number of kind $2q^2 - 2$ does not exceed $n + \sqrt{(n+2)/2} + 1 = n(1 + o(1))$. □

The following lemma in other terms estimates the number of representations of an integer as a sum of two squares. Though this estimate easily follows from the well-known number-theoretic facts, we give a proof for the sake of completeness.

Lemma 4.8 $MS(\{1, 4, \ldots, q^2\}) = q^{O(1/\ln\ln q)}$.

Proof. It is easy to see that the maximum subset of $\{1, 4, \ldots, q^2\}$ symmetric with respect to $\frac{1}{2}g$ has as many elements as the number of solutions of equation $z_1 + z_2 = g$ in $\{1, 4, \ldots, q^2\}$. The Jacobi theorem (see e.g. [12, theorem 65]) says that if $g = 2^k m$ with odd m, then the total number of integer solutions of the equation $x^2 + y^2 = g$ is equal to $4E$, where E is the excess of the number of divisors $t \equiv 1 \pmod 4$ of m over the number of divisors $t \equiv 3 \pmod 4$ of m. We use the bound $E \le d(m)$, where $d(m)$ denotes the total number of positive divisors of m. It is known that $d(m) = m^{O(1/\ln\ln m)}$ ([25], see also [20] for the best currently known constant in the exponent). As $m \le g$ and it makes sense to consider only $g < 2q^2$, we have $d(m) = q^{O(1/\ln\ln q)}$. Summarizing, we obtain $MS(\{1, 4, \ldots, q^2\}) \le 4E \le 4d(m) = q^{O(1/\ln\ln q)}$. □

5. Concluding Remarks And Open Problems

5.1 The Main Problem

The main problem left open is to make closer the exponential gap between our bounds

$$2\ln n - O(1) \le M(n) \le (7 + o(1))\sqrt{n}.$$

Remark 4.4 shows that our method for upper bounding $M(n)$ cannot do it better. As for possibilities to improve our lower bound, it is a question if one can improve the intermediate lower bound of Lemma 3.1.

5.2 An Observation

It turns out that if we impose some (strong at first sight) restrictions on the structure of $V \subseteq \mathbb{Z}^2$, then the minimal possible value of $MS(V)$ will not change much. We first prove the following auxiliary fact.

Lemma 5.1 *Let V, U, and A be finite subsets of $\mathbb{Z}^k$. Suppose that for any $u \in U$ there is $v \in V$ with $u - v \in A$. Then $MS(V) \geq MS(U)/|A|^2$.*

Proof. Fix a correspondence $\phi : U \to V$ such that $u - \phi(u) \in A$ for all u. Let S be a symmetric subset of U containing $MS(U)$ elements. Among all $MS(U)$ ordered pairs (u_1, u_2) of symmetric points of S, let us consider those for which the pair $(u_1 - \phi(u_1), u_2 - \phi(u_2))$ is the same. We can pick up at least $MS(U)/|A|^2$ such pairs. The corresponding pairs $(\phi(u_1), \phi(u_2))$ are clearly pairwise distinct and, moreover, have a common center $g + \frac{1}{2}(\phi(u_1) - u_1) + \frac{1}{2}(\phi(u_2) - u_2)$, where g is the center of symmetry of S. Therefore, they form a symmetric subset of V with at least $MS(U)/|A|^2$ elements. □

As before, let a and b be arbitrary but fixed non-collinear vectors in $\mathbb{Z}^2$. Define $M'(n)$ to be the minimum of $MS(U)$ over all subsets $U = \{u_0, v_1, \ldots, u_n\}$ of $\mathbb{Z}^2$ with each difference $u_{i+1} - u_i$ in $\{a, b\}$ and such that the underlying word $w(U)$ does not contain subwords aa and bbb.

Proposition 5.2 $\frac{1}{9}M'(2n-1) \leq M(n) \leq M'(n)$.

Proof. The second inequality is trivial. To prove the first one, consider a set $V = \{v_0, v_1, \ldots, v_n\}$ of points of $\mathbb{Z}^2$ with each $v_{i+1} - v_i$ either $(1, 1)$ or $(1, 2)$. Extend V to $U = \{u_0, u_1, \ldots\}$ inserting a new point $v_i + (1, 0)$ between v_i and v_{i+1} with $v_{i+1} - v_i = (1, 1)$ and two new points $v_i + (1, 0)$ and $v_{i+1} - (0, 1)$ between v_i and v_{i+1} with $v_{i+1} - v_i = (1, 2)$. Notice that for any u_j there is v_i with $u_j - v_i \in \{(0, 0), (1, 0), (0, -1)\}$. By Lemma 5.1, $MS(V) \geq MS(U)/9$. In its turn $MS(U) \geq M'(2n-1)$, because difference between any two successive points of U is either $a = (1, 0)$ or $b = (0, 1)$, and word $w(U)$ is free of subwords a^2 and b^3 and has length at least $2n - 1$. The proposition follows. □

Similarly with item 3 of Lemma 1.1, $M'(n)$ could be alternatively defined as the maximum m such that every binary word of length n without squares of one letter and cubes of another letter has abelian occurrence of a symmetric pattern of length at least $m - 1$. Proposition 5.2 shows that the function $M(n)$ and its restricted version $M'(n)$ have the same order of magnitude.

Question 5.3 Can prohibition of a^2 and b^3 help? Is estimating $M'(n)$ somehow easier than estimating $M(n)$?

5.3 Higher Dimensions

The function $M(n)$, if considered from the geometric point of view in accordance with item 2 of Lemma 1.1, admits a natural generalization. Given $A \subset \mathbb{Z}^k$, define

$$M_{k,A}(n) = \min \Big\{ MS(\{v_0, v_1, \ldots, v_n\}) : v_i \in \mathbb{Z}^k,\ v_{i+1} - v_i \in A \Big\}.$$

Remark 5.4 We assume that $V = \{v_0, v_1, \ldots, v_n\}$ is an $(n+1)$-element set, or a sequence without self-crossing. Nevertheless, if we allow self-crossing and consider V to be *multiset*, then everything that is claimed below holds true under the suitable definition of a *symmetric multisubset*.

If $A \subset \mathbb{Z}^k$ is a linearly independent system of k vectors, then $M_{k,A}(n)$ does not depend on the particular choice of A. In this case we will drop subscript A and write simply $M_k(n)$. In this notation $M(n) = M_2(n)$.

Define also $L_k(n)$ to be the maximal l such that every word of length n over k letter alphabet has an abelian occurrence of a symmetric pattern of length at least l. The proof of Lemma 1.1 can be directly extended to show that

$$\begin{aligned} M_k(n) &= L_k(n) + 1, \\ M_{k,A}(n) &\geq L_{|A|}(n) + 1 = M_{|A|}(n). \end{aligned} \tag{12}$$

Let us focus on the case of difference set $A_c = \{ v : 0 < \|v\| \leq c\}$, where the norm $\| \cdot \|$ on $\mathbb{Z}^k$ is defined by $\|(z_1, \ldots, z_k)\| = \sum_{i=1}^k |z_i|$. As long as we are concerned with asymptotics in n and consider parameters k and c fixed, we may restrict our attention to the difference set A_1, consisting of k unit vectors of the standard basis and k more opposite vectors. We will not loose much, because

$$\frac{M_{k,A_1}(n)}{(2c+1)^{2k}} \leq M_{k,A_c}(n) \leq M_{k,A_1}(n).$$

The first inequality follows from Lemma 5.1 (given a set $V = \{v_0, v_1, \ldots, v_n\}$ with difference set A_c, we extend it to a set U with difference set A_1 inserting intermediate points between v_i and v_{i+1} whenever $\|v_{i+1} - v_i\| > 1$, and then relate $MS(V)$ with $MS(U)$). To facilitate notation, denote $M_k^{\pm}(n) = M_{k,A_1}(n)$. Notice relations

$$M_{2k}(n) \leq M_k^{\pm}(n) \leq M_k(n),$$

where the first inequality is true by (12).

If $k > 2$, all that we can say is that

$$M_k(n) \to \infty \quad \text{as} \quad n \to \infty. \tag{13}$$

This fact is a consequence of inequalities $M_k(n) > L_k(n) > L_k^{non-abel}(n)$, where $L_k^{non-abel}(n)$ generalizes the function $L^{non-abel}(n)$ introduced in Section 2 to the k-letter alphabet. Similarly with Lemma 2.3, if $L_k^{non-abel}(n) \geq l$, then $L_k^{non-abel}((n+1)(k^n+1)) \geq 2l+1$. This results in a lower bound for $M_k(n)$ that tends to the infinity but more slowly than the inverse tower function.

Question 5.5 Prove better (than the inverse tower function) lower bounds for $M_3(n)$ and $M_2^{\pm}(n)$.

Question 5.6 For $k > 2$ prove a better (than $O(\sqrt{n})$) upper bound for $M_k(n)$.

Question 5.7 Try to prove a better (than $O(\sqrt{n})$) upper bound for $M_2^{\pm}(n)$.

5.4 Odd Vs. Even Cardinality

Dekking [11] proves that in the binary alphabet there is an infinite word without abelian occurences of pattern x^4 and that in the ternary alphabet there is an infinite word without abelian occurences of pattern x^3. Keränen [17] reports a (computer aided) proof that in the 4-letter alphabet there is an infinite word free of abelian occurences of pattern x^2. The latter result implies that there is an infinite sequence of points in $\mathbb{Z}^4$ with differences between consequtive points only $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, and $(0,0,0,1)$ and without symmetric subsets of odd cardinality (excepting singletons). At the same time by (13) it must contain arbitrarily long symmetric subsets of even cardinality.

Acknowledgments.

I am grateful to Taras Banakh, whose proof of Proposition 2.2 in geometric terms inspired this work, and to Oleg Pikhurko, whose observation improved the original proof of Theorem 3.2 and led to a better lower bound for $M(n)$. I thank an anonymous referee for several corrections and useful expository suggestions.

References

[1] M. Ajtai, J. Komlós and E. Szemerédi. A dense infinite Sidon sequence. *European J. Combin.*, 2:1–11, 1981.

[2] E. Bach and J. Shallit. *Algorithmic number theory.* Vol. 1, MIT Press, Cambridge, 1996.

[3] T. Banakh. On asymmetric colorings of Abelian groups. In: *Paul Erdős and his Mathematics. Research Communications of the conference held in the memory of Paul Erdős*, July 4–11, 1999, Budapest, Janos Bolyai Mathematical Society, 30–32.

[4] T. O. Banakh and I. V. Protasov. Asymmetric partitions of Abelian groups (in Russian). *Mat. Zametki*, 66(1):10–19, 1999.

[5] T. Banakh, I. Kmit, and O. Verbitsky. On asymmetric colorings of integer grids. September 1999, accepted for publication in *Ars Combinatoria*.

[6] R. C. Bose. An affine analogue of Singer's theorem. *J. Indian Math. Soc.* (new series), 6:1–15, 1942.

[7] R. C. Bose and S. Chowla. Theorems in the additive theory of numbers. *Comment. Math. Helv.*, 37:141–147, 1962.

[8] T. C. Brown. Variations on van der Waerden's and Ramsey's theorems. *Am. Math. Mon.*, 82:993–995, 1975.

[9] S. Chowla. Solution of a problem of Erdős and Turan in additive number theory. *Proc. Natl. Acad. Sci. India*, 14:1–2, 1944.

[10] M. Coudrain and M. P. Schützenberger. Une condition de finitude des monoides finiment engendres. *C. R. Acad. Sci., Paris, Ser. A*, 262:1149–1151, 1966.

[11] F. M. Dekking. Strongly non-repetitive sequences and progression-free sets. *J. Combin. Theory*, 27:181–185, 1979.

[12] L. E. Dickson. *Introduction to the theory of numbers.* Dover Publ., New York, 1957.

[13] P. Erdős and P. Turan. On a problem of Sidon in additive number theory, and on some related problems. *J. London Math. Soc.*, 16:212–215, 1941.

[14] P. Erdős. On a problem of Sidon in additive number theory and on some related problems. Addendum. *J. London Math. Soc.*, 19:208, 1944.

[15] P. Erdős, message to A. Stöhr, published in: A. Stöhr, Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe. II. *J. Reine Angew. Math.*, 194:132–133, 1955.

[16] W. L. Fouché. Unavoidable regularities and factor permutations of words. *Proc. Royal Soc. Edinb.*, 125A(3):519–524, 1995.

[17] V. Keränen. Abelian squares are avoidable on 4 letters. In: *Proc. of ICALP'92*, LNCS vol. 623, Springer, 41–52, 1992.

[18] G. Hoheisel. Primzahlprobleme in der Analysis. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, 573:580–588, 1930.

[19] F. Krückeberg. B_2-Folgen und verwandte Zahlenfolgen. *J. Reine Angew. Math.*, 206:53–60, 1961.

[20] J.-L. Nicolas and G. Robin. Majorations explicites pour le nombre de diviseurs de N. *Canad. Math. Bull.*, 26:485–492, 1983.

[21] C. Pomerance and A. Sárközy. Combinatorial number theory. In *Handbook of Combinatorics*, chapter 20, pages 967–1018. Elsevier Publ., 1995.

[22] C. Pomerance. Collinear subsets of lattice point sequences — an analog of Szemerédi's theorem. *J. Combin. Theory A*, 28:140–149, 1980.

[23] L. T. Ramsey and J. L. Gerver. On certain sequences of lattice points. *Pacific J. Math*, 83:357–363, 1979.

[24] J. Singer. A theorem in finite projective geometry and some applications to number theory. *Trans. Am. Math. Soc.*, 43:377–385, 1938.

[25] S. Wigert. Sur l'odre de grandeur du nombre des diviseurs d'un entier. *Arkiv för Matematik, Astronomi och Fisik*, 3(18):1–9, 1906–1907.