Sárközy's theorem

Group 3C: Dawid Lipinski, Delyan Kirov, Trunks Lei,
Satpal Hunjan, Joshua Heyworth

February 12, 2021

Abstract

Sárközy's theorem [3] states that in any subset of natural numbers of positive upper density there exist 2 distinct elements whose difference is a perfect square number, as well as 2 elements which differ by 1 less than a prime number. In this article we estimate the maximal size of a subset lacking such configurations, and investigate its behaviour as the length of intervals tend to infinity.

1 Introduction

For an interval $[1, n]$ we are interested in the cardinality of the biggest set S that contains no elements with square differences and the ratio S/n. This can be expressed in the following definition:

Definition 1.
For the interval $[1, n]$ where $n \in \mathbb{N}_+$ define $sark$: $[1, n] \longrightarrow \mathbb{Q}_+$ for $sark(n) = S/n$ where S is the biggest subset, such that for all $s \in S$ and $k \in \mathbb{N}_+$ we have $s_1 - s_2 \neq k^2$.

In this article we investigate the behaviour of this quantity for large values of n, in particular the upper and lower bounds to approximate its values. The theorem was conjectured by László Lovász before being proved independently by Hillel Furstenberg and András Sárközy in the late 70s, after whom the theorem is named by. Both Furstenberg and Sárközy used different methods, one used ergodic theoretic and the other a more harmonic analytic. Since then there have been many publications simplifying the previous proofs as well as creating better bounds. The upper bound of Sárközy's theorem can be generalized from sets that avoid square differences to sets that avoid differences in $p(\mathbb{N})$ (the values at integers of a polynomial p with integer coefficients) [3].

2 Data testing

Let us look at an interval of natural numbers and try to find the biggest subsets that does not contain elements, which have a square difference. We shall investigate the interval $[1, 10]$.

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

One way to start would be to look at the differences of 1 with the other elements and exclude numbers which produce square differences. This greedy algorithm then produces the following subset:

$$1, 3, 6, 8$$

Naturally, we may be sceptical to believe that this is the biggest subset, however it turns out that for small intervals, the greedy algorithm is accurate. A proof of this is simple: the biggest subgroups must be no bigger than size 5, since if they are any bigger subsets then they must contain at least 2 consecutive numbers, but that would mean that their difference would be a square. Thus, for the interval $[1, 10]$ the groups with five non-consecutive elements are:

$$1, 3, 5, 7, 9 \text{ and } 2, 4, 6, 8, 10$$

These subsets contain square differences, therefore the biggest subsets are of at most 4 elements.

However, the greedy algorithm becomes increasingly less precise with bigger intervals. In that case it may be more beneficial to look for an algorithm that checks every subset individually. This is, however, not computationally efficient. It may be helpful to keep a few things in mind when constructing such an algorithm:

1. The greedy algorithm is a reasonable lower bound, and the size of the interval divided by 2 is a reasonable upper bound. But finding better bounds would make computation even more efficient.
2. There is no point in checking subsets with consecutive numbers.
3. If a subset has no square differences in the interval $[1, n]$, then there is no point in checking for the same subset again in the interval $[1, n + m]$, where n and $m \in \mathbb{N}_+$.

Let us consider an example of how the true algorithm[4] would work for the interval [1, 20]. Let us say that we have decided that the reasonable bounds for this interval are $[7, 8]$. We start to look at all distinct subsets of size 7 and store them as the column vectors of a matrix. We then check that the columns do not contain square differences. The first vector that does not contain any square differences will stop this loop, if there is no vector that satisfies this condition then the program stops and the output would be: (lower bound−1)/(the size of the original interval) = 7/20 = 0.35. In this case we do detect a column that has this property, so we move on to the next value, which in this case is 8 (our upper bound) and we detect a column that has this property so our true output is 8/20 = 0.4.

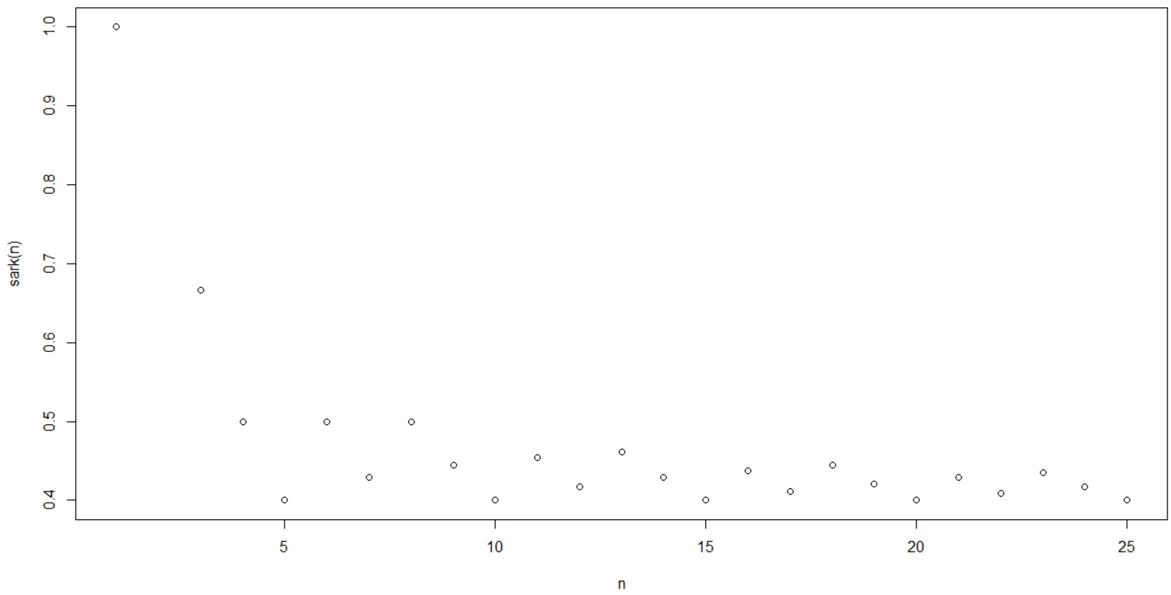

Figure 1: sark function

As can be seen from figure 1, this ratio tends to become smaller as n increases. It would not be unreasonable to assume that the $Sark(n) \longrightarrow 0$ as $n \longrightarrow \infty$.

3 prime-1 sequence

So far we have looked at subsets of intervals that do not contain elements with square differences. We showed, through Sárközy's theorem, that the ratio of the size of these subsets and the size of the interval $[1, n]$ tends to zero as $n \longrightarrow \infty$. In this final section we will see that this is also true if we consider subsets that do not contain any prime−1 differences. Let us begin with the following definitions.

Definition 3.
For the interval $[1, n]$ where $n \in \mathbb{N}_+$ define $primesark$: $[1, n] \longrightarrow \mathbb{Q}_+$ for $primesark(n) = S/n$ where S is the biggest subset, such that for all $s \in S$ and $k \in \mathbb{N}_+$ we have $s_1 - s_2 \neq k - 1$, where k is prime.

As before, we can calculate these subsets using a greedy algorithm. For the interval $[1, 10]$ we have the following set of primes-1: $\{1, 2, 4, 6, \}$. So one possible subset would be: $\{1, 4, 9\}$ which turns out to be the biggest for this interval. We can construct an algorithm that finds all values of $primesark(n)$ the same way we did for $sark(n)$. From figure 2 we can see that $primesark(n)$ seems to converge to zero as $n \longrightarrow \infty$. However, unlike the function $sark(n)$, $primesark(n)$ checks for squares and while squares are not as common as primes for small intervals, when the interval $[1, n]$ gets quite big, the number of squares greatly outnumbers the number of primes. This can be seen visually by comparing figure 1 and figure 2, notice that the values of $sark(n)$ are smaller than the values of $primesark(n)$ for all n in this range. It is thus surprising that these two functions are actually related and that they both converge to zero as we conjectured. This can be proved by Sárközy's theorem.

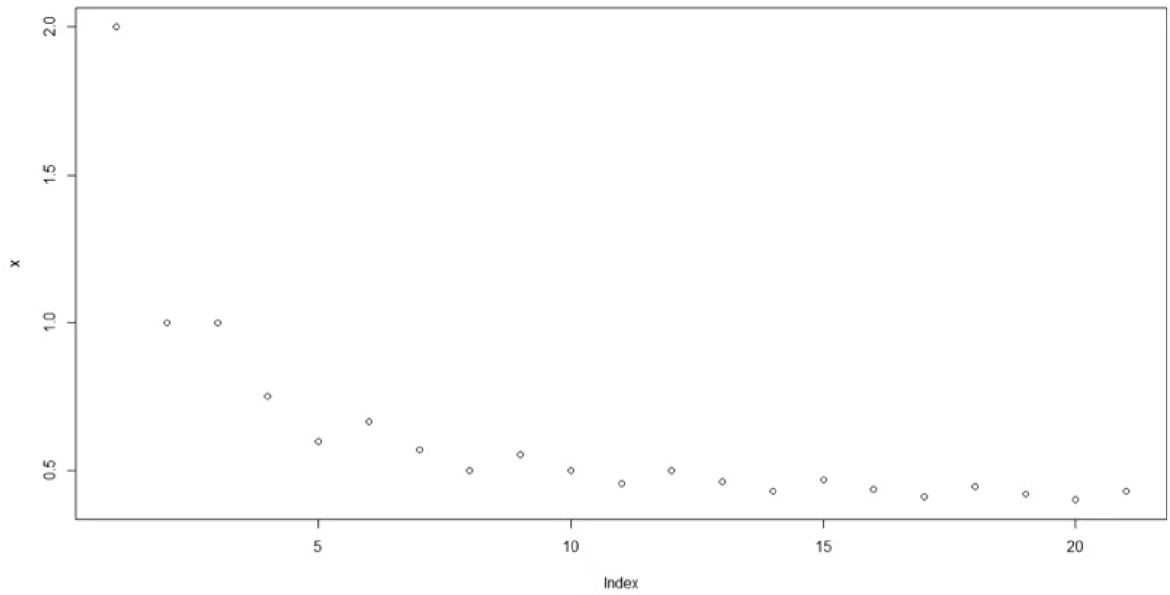

Figure 2: Graph of function p1.

4 Notation

First we need to introduce some notation, taken from [2]. Let f be a real or complex valued function and g a real valued function. Let both functions be defined on some unbounded subset of the positive real numbers, and $g(n)$ be strictly positive for all large enough values of n. Then we write:

1. **The big O notation:**

$$f(x) = O(g(x)) \iff \exists k > 0 \exists n_0 \forall n > n_0 : |f(n)| \leq kg(n) \qquad [4.1]$$

2. **The big Omega notation:**

$$f(n) = \Omega(g(n)) \iff \exists k > 0 \forall n_0 \exists n > n_0 : |f(n)| \geq kg(n) \qquad [4.2]$$

When we write Ω_ε or O_ε, this implies that the constant k is dependent on ε.

5 Lower bound

In this section we look at the lower bound for Sárközy's theorem. Here is the definition.

Definition 2. For any set $A = \{1, 2, 3, ..., n\}$ (where $n \in \mathbb{N}$) we can find a set of size zero (empty set), such that it is a subset of A. We define the lower bound as the smallest square-difference-free subset of A, then the answer is zero.

In this case what we consider to be the lower bound is the smallest number m, dependent on the size of A, such that for any A, the largest square-difference-free subset of A is at least m.

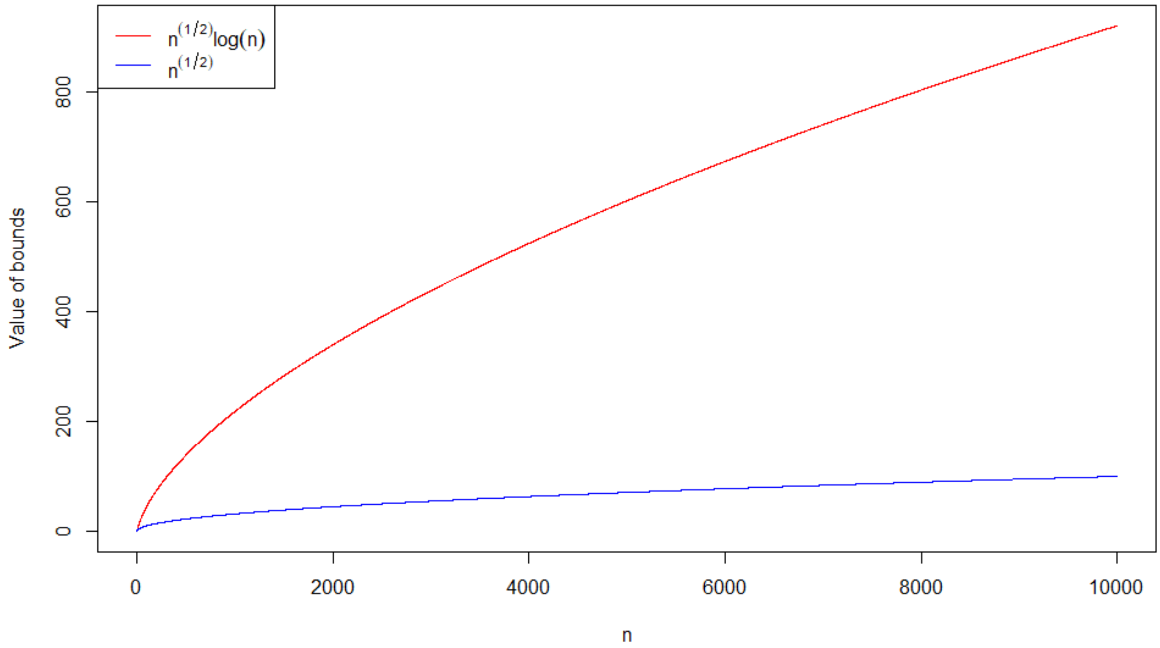

Figure 3: Plot for n from 1 to 10000 of two bounds. (Equation 5.1 and Equation 5.2)

The first conjectured lower bound was:

$$O(n^{1/2} \log^k n), \tag{5.1}$$

where k is some constant and n is the largest number in the set of natural numbers up to n [3]. This was later disproved by Sárközy, who himself conjectured that for every $\varepsilon > 0$ the bound is

$$O_\varepsilon(n^{1/2+\varepsilon}) \tag{5.2}$$

[3]. While the change may seem very insignificant if we look at the graph comparing these two functions (for $k = 1$ and $\varepsilon = 0$), we can see that the difference gets larger as n increases (Figure 3). This bound was later disproved by by Imre Z. Ruzsa, who found square-difference-free sets with up to

$$\Omega(n^{(1+\log_{65} 7)/2}) \approx n^{0.733077} \tag{5.3}$$

elements [3]. We can also plot this using R (Figure 4). In his proof Ruzsa constructed the set R which is free of square-differences modulo 65, which he was then able to use to construct a set of integers free of square differences of size $n^{(1+\log_{65} 7)/2}$ [3] [6] [7]. Then by improving the modular construction (to 205), the bound can be improved to

$$\Omega(n^{(1+\log_{205} 12)/2}) \approx n^{0.733412} \tag{5.4}$$

[3]. If we plot those two bounds side by side we can see that there is almost no difference for n up to 1000 (Figure 5).

Based on this type of construction of lower bounds, there are currently two conjectures: [3]:

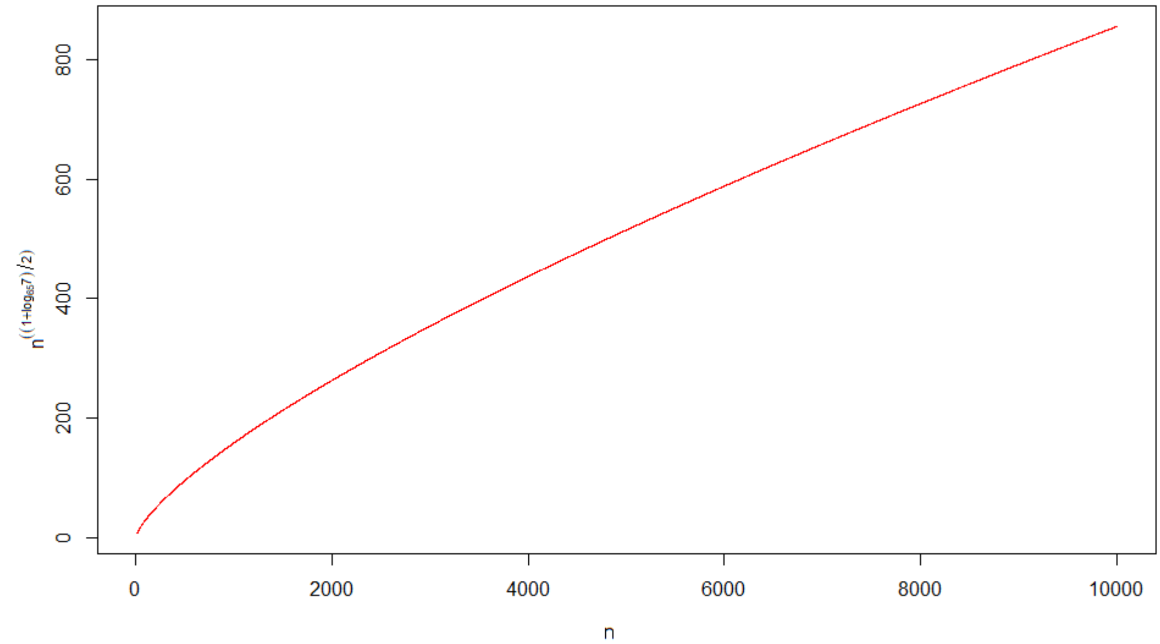

Figure 4: Plot for n from 1 to 10000 of $n^{(1+\log_{65} 7)/2}$.

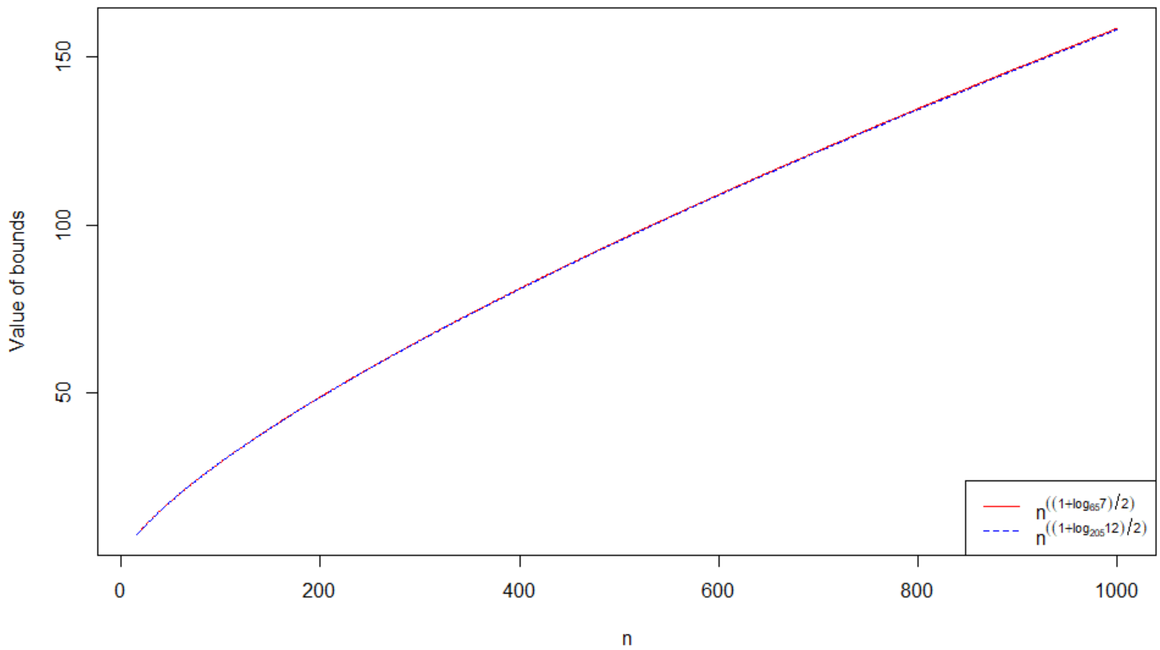

Figure 5: Comparison of two bounds. (Equations 5.3 and 5.4)

1. For every $\varepsilon > 0$ and sufficiently large n, there exist square-difference-free subsets of the numbers from 0 to n with the number of elements equal to

$$\Omega_{\varepsilon}(n^{1-\varepsilon}). \qquad [5.5]$$

What is also interesting is that if this conjecture is true, the exponent of one in the upper bounds for the Sárközy's theorem cannot be lowered.

2. The exponent $\frac{3}{4} = 0.75$ has been identified as another candidate for the true maximum growth rate of these sets. That is, the lower bound can be at most $n^{0.75}$.

6 Upper bound

The upper bound can be calculated using

$$O(n/\log(n)^{\frac{1}{4}\log(\log(\log(\log(n))))}) \qquad [6.1]$$

[3], A majority of the proofs for this use Fourier analysis or Ergodic theory however this is not required to prove a basic form of the theorem, that all square-difference-free sets have zero density [5].
We can compare both upper and lower bounds by looking at the graph (Figure 6).

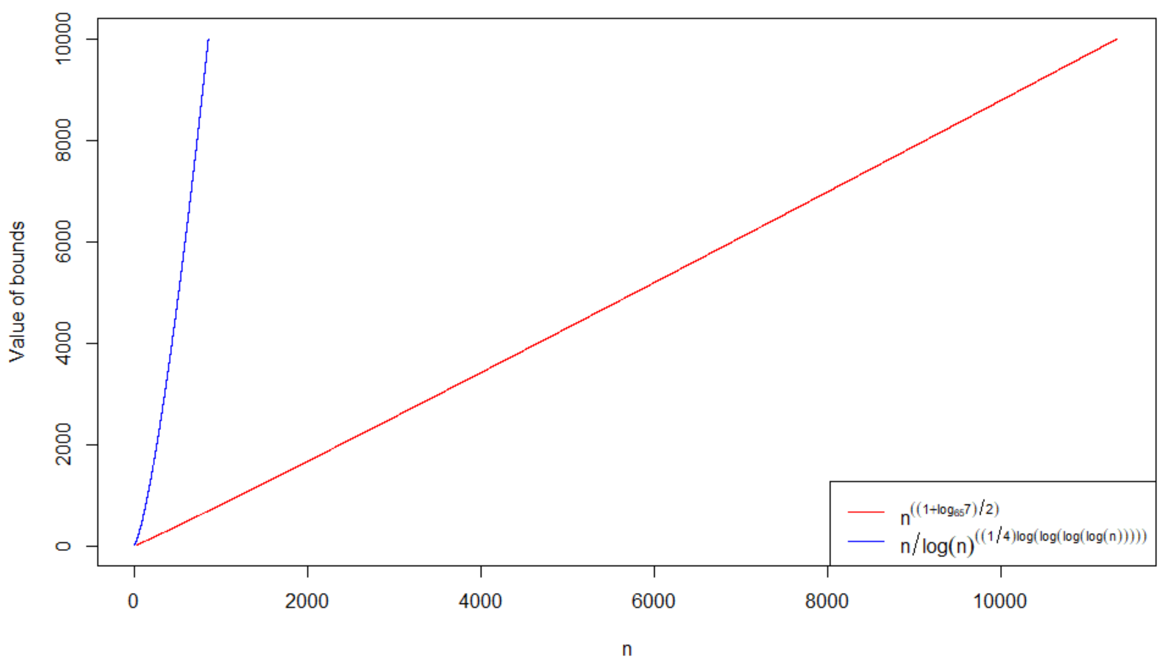

Figure 6: Comparison of upper and lower bound for n between 16 and 1000. (Our first $n = 16$, because $\log(0)$ is not defined, so we have to choose n sufficiently large to be able to graph this functions.)

As can be seen, there is a a large difference between upper and lower bound.

References

[1] Eisner, Tanja; Farkas, Bálint; Haase, Markus; Nagel, Rainer (2015), "20.5 The Furstenberg–Sárközy Theorem", Operator Theoretic Aspects of Ergodic Theory, Graduate Texts in Mathematics, 272, Cham, Switzerland: Springer, pp. 455–457

[2] https://en.wikipedia.org/wiki/Big_O_notation

[3] https://en.wikipedia.org/wiki/Furstenberg%E2%80%93S%C3%A1rk%C3%B6zy_theorem

[4] https://github.com/delyan-kirov/Sarkozy-s-function.git

[5] https://terrytao.wordpress.com/2013/02/28/a-fourier-free-proof-of-the-furstenberg-sarkozy-theorem/

[6] https://arxiv.org/pdf/0804.4892.pdf

[7] https://www.combinatorics.org/ojs/index.php/eljc/article/view/v22i1p32/pdf

[8] Nathanson, Melvyn B. (2000). Elementary Methods in Number Theory. Graduate Texts in Mathematics. 195. Springer-Verlag. ISBN 978-0387989129. Zbl 0953.11002.

All of the references have been last accessed at 12/02/2021.