

# Some Remarks on the Additive Structure of the Set of Primes

Christian Elsholtz

## 1 Introduction

In this paper we consider problems related to the additive decomposition of the set of primes. Ostmann [7] asked whether the set of primes  $\mathcal{P}$  has an asymptotic additive decomposition into two sets, that is, whether there exist two sets  $\mathcal{A}$  and  $\mathcal{B}$  of positive integers with at least two elements such that the sum set  $\mathcal{A} + \mathcal{B} = \{a + b : a \in \mathcal{A}, b \in \mathcal{B}\}$  is, up to finitely many exceptions, equal to the set of primes, i.e.,  $\mathcal{A} + \mathcal{B} = \mathcal{P}'$ , where  $\mathcal{P}' \cap [x_0, \infty] = \mathcal{P} \cap [x_0, \infty]$  for some  $x_0$ . For simplicity we will write such an asymptotic equality as  $\mathcal{A} + \mathcal{B} \simeq \mathcal{P}$ . It is conjectured that no such decomposition can exist, but the problem is still open. This problem is also known as the inverse Goldbach problem.

At the Journées Arithmétiques in Rome (1999), the author provided new bounds for the counting functions  $\mathcal{A}(N)$  and  $\mathcal{B}(N)$ , if an asymptotic additive decomposition of the primes exists (see (1.2) below, and [2]). These bounds imply (see Theorem 2) that the set of primes has no asymptotic additive decomposition into three sets, which settles the ternary case of the inverse Goldbach problem.

In our talk at the Millennial Conference we sketched a more direct approach to the ternary case based on new upper bounds for the number of long prime tuples. In this paper we will describe this approach in detail, and we make some remarks on various generalizations of the inverse Goldbach problem.

We begin with the following result, which gives a new upper bound for the number of long prime  $k$ -tuples below  $N$ , where  $k$  is of size  $(\log N)^r$  for some  $r > 0$ .

**Theorem 1.** *Let  $\varepsilon > 0$ . Let  $r > 1$  and let  $N \geq N_{r,\varepsilon}$  be sufficiently large. Let  $\mathcal{A} = \{a_1, \dots, a_k\} \subset [1, N]$ ,  $0 < a_1 < a_2 < \dots < a_k$ , be a set of integers with  $k \geq (\log N)^r$ . Then the number  $E_{\mathcal{A}}(N)$  of those  $n \leq N$  for which all  $n + a_i$  are simultaneously prime is bounded by*

$$E_{\mathcal{A}}(N) \ll N^{\frac{1}{2} + \frac{1}{r+1} + \varepsilon}.$$

For a more detailed investigation of results of this type and some refinements we refer the reader to forthcoming publications of the author. For example, we expect to prove a bound in which the term  $+\varepsilon$  can be replaced by a function  $-\varepsilon(N)$ . Here we will prove only the special case  $r = 6$  of this result, which will be sufficient for our application to the ternary case of the inverse Goldbach problem.

First suppose that  $\mathcal{A} + \mathcal{B} \simeq \mathcal{P}$  is an asymptotic additive decomposition of the set of primes. As the proof of Theorem 2 shows, our new bound on  $E_{\mathcal{A}}(N)$  implies that

$$N^{0.35} \ll \mathcal{A}(N) \ll N^{0.65}. \quad (1.1)$$

In [2] the author refined this estimate and obtained the bounds

$$N^{\frac{1}{2}}(\log N)^{-5} \ll \mathcal{A}(N) \ll N^{\frac{1}{2}}(\log N)^4. \quad (1.2)$$

The same bounds hold for  $\mathcal{B}(N)$ .

The idea of the proof is as follows. If  $a + b = q > x_0$  is prime, then  $a + b$  avoids the class 0 modulo primes  $p$  with  $x_0 < p < q$ . This means that the residue class  $a$  modulo  $p$  that occurs in  $\mathcal{A}$  forbids the residue class  $-a$  for  $\mathcal{B}$ . It can be shown that many residue classes modulo many primes contain elements of the sets  $\mathcal{A}$  or  $\mathcal{B}$ , since otherwise one would obtain an upper bound on the counting functions  $\mathcal{A}(N)$  and  $\mathcal{B}(N)$  that contradicts existing lower bounds. On the other hand, these residue classes in  $\mathcal{A}$  forbid many residue classes in  $\mathcal{B}$ . This gives good upper bounds on  $\mathcal{B}(N)$  and good lower bounds on  $\mathcal{A}(N)$ , etc.

The conclusions we draw from this argument go in two different directions.

Suppose that a ternary asymptotic decomposition exists, i.e., that  $\mathcal{A} + \mathcal{B} + \mathcal{C} \simeq \mathcal{P}$ . Then equation (1.1) implies that  $\mathcal{A}(N), \mathcal{B}(N), \mathcal{C}(N) \gg N^{0.35}$ . But this implies that there must be a prime  $p$  such that some element  $a + b + c \in \mathcal{A} + \mathcal{B} + \mathcal{C}$  is divisible by  $p$ . Making sure that  $a + b + c > p$  we arrive at a contradiction to  $\mathcal{A} + \mathcal{B} + \mathcal{C} \simeq \mathcal{P}$ .

Therefore, Theorem 1 will enable us to give a short proof of the following result, which settles the ternary inverse Goldbach problem.

**Theorem 2.** *There do not exist sets of integers  $\mathcal{A}, \mathcal{B}$ , and  $\mathcal{C}$  with  $|\mathcal{A}|, |\mathcal{B}|, |\mathcal{C}| \geq 2$  such that  $\mathcal{A} + \mathcal{B} + \mathcal{C} \simeq \mathcal{P}$  holds.*

In another direction, we examine how many residue classes modulo  $p$  are actually used by the sequences  $\mathcal{A}$  and  $\mathcal{B}$  in a binary decomposition  $\mathcal{A} + \mathcal{B} \simeq \mathcal{P}$ . It is known that for a sequence  $\mathcal{S}$  that avoids half of the residue classes modulo primes the counting function is (roughly) bounded by  $N^{1/2}$ . If the classes that are avoided are randomly distributed, then one would expect the counting function of  $\mathcal{S}$  to be much smaller. In view

of equation (1.2) it is an interesting question if, modulo many primes, the sets  $\mathcal{A}$  and  $\mathcal{B}$  avoid half of the residue classes. We shall make this more precise as follows.

For sufficiently large  $N$ ,  $x_0$  and any set  $\mathcal{S}$  of positive integers define

$$\nu_{\mathcal{S}}(p) = |\{s \bmod p : s \in \mathcal{S} \cap [x_0, N]\}|.$$

We will prove:

**Theorem 3.** *Let  $\mathcal{A} + \mathcal{B} \simeq \mathcal{P}$  be an asymptotic decomposition of the set of primes. Then there exist constants  $K_1, K_2, K_3, K_4 > 0$  such that for  $y = K_4 N^{1/2} (\log N)^{-K_3}$*

$$2 \log y - K_1 \log \log y \leq \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} \leq 2 \log y + K_2 \log \log y.$$

The same bounds hold for  $\nu_{\mathcal{B}}$ .

In view of the inequality  $\nu_{\mathcal{A}}(p) + \nu_{\mathcal{B}}(p) \leq p$  (see Corollary 1) this shows that, modulo many primes, the size of  $\nu_{\mathcal{A}}(p)$  and  $\nu_{\mathcal{B}}(p)$  is about  $\frac{p}{2}$ .

One can think of several generalizations of the original problem.

**Question 1:** What kind of results can be proved if  $\mathcal{A} + \mathcal{B} \simeq \mathcal{P}_1$ , where  $\mathcal{P}_1$  is a thin subset of the primes?

**Question 2:** What happens in the case when  $\mathcal{A}_1 + \mathcal{A}_2 + \dots + \mathcal{A}_k \subseteq \mathcal{P}$ ?

**Question 3:** Another possible generalization is to consider decompositions in which more than finitely many composite exceptions are allowed. Let  $\mathcal{A} + \mathcal{B} = \mathcal{P}_1 \cup \mathcal{C}$  with  $\mathcal{P}_1 \subseteq \mathcal{P}$  and  $\mathcal{C} \cap \mathcal{P} = \emptyset$ , where  $\mathcal{P}_1$  is a not too small set of primes and  $\mathcal{C}$  is a not too large set of composite integers. Can one obtain similar results for this situation?

**Acknowledgements.** The author would like to thank J. Brüdern, L.G. Lucht, and A. Sárközy for discussions and comments.

## 2 A Large Sieve Argument

We start with the proof of Theorem 2. The proof of the part of Theorem 1 which is needed for Theorem 2 follows below.

We make use of the following result, which is a special case of a theorem of Pomerance, Sárközy, and Stewart (see Theorem 3 of [8]). Note that this result gives some insight into Question 2.

**Lemma 1.** *Let  $\varepsilon > 0$ , let  $N$  be a positive integer, and let  $\mathcal{A}, \mathcal{B}, \mathcal{C}$  denote sets of positive integers. If  $N$  is sufficiently large and if*

$$\min(\mathcal{A}(N), \mathcal{B}(N), \mathcal{C}(N)) > N^{1/3+2\varepsilon},$$

then there is a prime  $p$  with  $p < N^{1/3+\varepsilon}$  such that  $\mathcal{A} + \mathcal{B} + \mathcal{C}$  contains an element which is divisible by  $p$ .

Before we turn to the ternary problem we make some observations on the binary case. Suppose that  $\mathcal{A} + \mathcal{D} \simeq \mathcal{P}$ , where  $|\mathcal{A}|, |\mathcal{D}| \geq 2$ . For sufficiently large  $N > x_0$  we give an upper bound on the number of those  $n \leq N$  for which all  $n + a_i$  are prime. (We do not need to worry about the finitely many exceptions below  $x_0$ .) By an application of a 2-dimensional sieve,  $|\mathcal{A}| \geq 2$  implies that  $\mathcal{D}(N) \ll N/(\log N)^2$ . It then follows from  $\mathcal{A}(N)\mathcal{D}(N) \gg \pi(N) \gg N/\log N$  that  $\mathcal{A}(N) \gg \log N$  and in particular  $|\mathcal{A}| \geq 8$ . Similarly,  $|\mathcal{A}| \geq 8$  implies that  $\mathcal{D}(N) \ll N/(\log N)^8$  and hence  $\mathcal{A}(N) \gg (\log N)^7 \geq (\log N)^6$ , for sufficiently large  $N \geq N_0 \geq x_0$ . We next apply Theorem 1 with  $r = 6$ . It follows that for arbitrary positive  $\varepsilon$  we have  $\mathcal{D}(N) \ll N^{\frac{1}{2} + \frac{1}{7} + \varepsilon}$  and therefore  $\mathcal{A}(N) \gg N^{\frac{1}{2} - \frac{1}{7} - 2\varepsilon}$ .

Returning to the ternary problem, we put  $\mathcal{D} = \mathcal{B} + \mathcal{C}$ . The above discussion then shows that  $\mathcal{A}(N) \gg N^{\frac{1}{2} - \frac{1}{7} - 2\varepsilon} \gg N^{0.35}$ , for sufficiently small  $\varepsilon$ . By symmetry, the same bound holds for  $\mathcal{B}(N)$  and  $\mathcal{C}(N)$ .

For  $N > (N_0)^3$  let  $\mathcal{A}_1 = \mathcal{A} \cap [N^{0.34}, \infty]$ . Then  $\mathcal{A}_1(N) \gg N^{0.35}$  still holds. Lemma 1 implies that  $\mathcal{A}_1 + \mathcal{B} + \mathcal{C}$  contains an element  $a_1 + b + c \geq N^{0.34}$  which is divisible by some prime  $p \leq N^{1/3+\varepsilon}$ . For sufficiently small  $\varepsilon$  this implies that  $a_1 + b + c$  is not a prime. This proves Theorem 2.

With regard to Question 1 one might be able to state similar results. However, in order to start the sieve iteration at the beginning of the proof all sets involved have to be sufficiently large. For example, one might be able to prove the following statements: If there exists a subset of the primes  $\mathcal{P}_1$  such that  $\mathcal{P}_1 \simeq \mathcal{A} + \mathcal{B}$  with  $\mathcal{P}_1(N) \gg N^\alpha$ , where  $\alpha > \frac{1}{2}$ , then  $\mathcal{A}(N), \mathcal{B}(N) \gg (\log N)^r$  with  $\frac{1}{2} + \frac{1}{r+1} < \alpha$  implies that

$$N^{\alpha - \frac{1}{2} - \varepsilon} \ll \mathcal{A}(N) \ll N^{\frac{1}{2} + \varepsilon}.$$

Moreover, by the methods of [2] one can replace the factor  $N^\varepsilon$  by a power of  $\log N$ .

We now turn to the proof of Theorem 1 for the case  $r = 6$ . We have  $\mathcal{A}(N) \geq (\log N)^6$ . From the Cauchy-Schwarz inequality and the bound

$$\sum_{p \leq y} \left( \frac{\log p}{p} \right)^{1/2} \gg \frac{y^{1/2}}{(\log y)^{1/2}}$$

we conclude

$$\left( \sum_{N_0 < p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} \right) \left( \sum_{N_0 < p \leq y} \frac{\nu_{\mathcal{A}}(p)}{p} \right) \gg \frac{y}{\log y}.$$

We take an integer  $m_0 = \lfloor (1/14)(\log N)/(\log \log N) \rfloor$  and set  $y = N^{1/(2m_0)}$ . This implies that  $y \sim (\log N)^7$ . If now

$$\sum_{N_0 < p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} > 3 \log N,$$

then we arrive at a contradiction by applying Gallagher’s larger sieve (see [3]) as follows:

$$\mathcal{A}(N) \leq \frac{-\log N + \sum_{N_0 < p \leq y} \log p}{-\log N + \sum_{N_0 < p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)}} < \frac{y}{2 \log N} < (\log N)^6.$$

We therefore may assume that

$$\sum_{N_0 < p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} \leq 3 \log N.$$

This implies that

$$\sum_{N_0 < p \leq y} \frac{\nu_{\mathcal{A}}(p)}{p} \gg \frac{y}{(\log y)(3 \log N)} \gg \frac{(\log N)^6}{\log \log N}.$$

It then follows by Montgomery’s large sieve (see [5]) and a result by Vaughan (see [9]) that  $\mathcal{D}(N) \leq \frac{2N}{L}$ , where

$$\begin{aligned} L &= \sum_{q \leq N^{1/2}} \mu^2(q) \prod_{p|q} \frac{\omega(p)}{p - \omega(p)} \\ &\geq \max_{m \in \mathbb{N}} \exp \left( m \log \left( \frac{1}{m} \sum_{p \leq N^{1/(2m)}} \frac{\omega(p)}{p} \right) \right). \end{aligned}$$

We may choose

$$\omega(p) = \begin{cases} \nu_{\mathcal{A}}(p) & \text{for } N_0 < p \leq y, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, taking  $m = m_0 = \lfloor (1/14)(\log N)/(\log \log N) \rfloor$ , we find that

$$\begin{aligned} \log L &\geq \frac{\log N}{14 \log \log N} \log \left( \frac{14 \log \log N}{\log N} \frac{c(\log N)^6}{\log \log N} \right) \\ &\geq \left( \frac{5}{14} - \varepsilon \right) \log N = \left( \frac{1}{2} - \frac{1}{7} - \varepsilon \right) \log N. \end{aligned}$$

Hence  $E_{\mathcal{A}}(N) \leq 2N/L \ll N^{\frac{1}{2} + \frac{1}{7} + \varepsilon}$ . This completes the proof of the case  $r = 6$  of Theorem 1.

### 3 A Combinatorial Approach to the Inverse Goldbach Problem

Let us suppose that  $\mathcal{A} + \mathcal{B} \simeq \mathcal{P}$  is an asymptotic decomposition of the set of primes. Our knowledge of the number  $\nu(p)$  of residue classes modulo  $p$  that can occur is limited. We expect that  $\nu_{\mathcal{A}}(p) \approx \nu_{\mathcal{B}}(p) \approx p/2$ . However, it could also be the case that  $\nu_{\mathcal{A}}(p)$  is very large (i.e., close to  $p$ ) for half of the primes, and very small for the other half of the primes. We shall investigate this issue in more detail in this section.

Recall that the definition of  $\nu_{\mathcal{S}}(p)$  depends on  $N$ . In the following we assume that  $N$  is sufficiently large in terms of  $p$ . (Perhaps it would be more convenient to redefine  $\nu$  as  $\nu_{\mathcal{S}}(p) = |\{s \bmod p : s \in \mathcal{S}\}|$ .)

**Lemma 2.** *Let  $p$  be a sufficiently large prime. Then we have, in the binary case,*

$$\nu_{\mathcal{A}+\mathcal{B}}(p) = p - 1.$$

This follows from the definition of  $\mathcal{A} + \mathcal{B} \simeq \mathcal{P}$ .

**Lemma 3.** *Let  $p$  be an arbitrary prime. Then we have:*

$$(a) \text{ In the binary case, } \nu_{\mathcal{A}}(p) + \nu_{\mathcal{B}}(p) - 1 \leq \nu_{\mathcal{A}+\mathcal{B}}(p).$$

$$(b) \text{ In the ternary case, } \nu_{\mathcal{A}}(p) + \nu_{\mathcal{B}}(p) + \nu_{\mathcal{C}}(p) - 2 \leq \nu_{\mathcal{A}+\mathcal{B}+\mathcal{C}}(p).$$

This follows from Lemma 2 and the Cauchy-Davenport theorem and its generalization (see Nathanson [6], Theorems 2.2 and 2.3). This result is also mentioned in Pomerance et. al. [8, p. 367].

The following corollary, of course, was also known before, but we state it for completeness.

**Corollary 1 (Binary case).** *If  $p$  is sufficiently large, then*

$$\nu_{\mathcal{A}}(p) + \nu_{\mathcal{B}}(p) \leq p.$$

This obvious remark is strong enough to imply (by means of the large sieve inequality) the nontrivial bound  $\mathcal{A}(N)\mathcal{B}(N) = O(N)$ ; see [10], [8] and [4]. A more elementary proof of  $\mathcal{A}(N)\mathcal{B}(N) = O(N)$  can be found in [1].

**Theorem 4.** *For some constant  $c$  and all sufficiently large  $y$  the following bound holds:*

$$\sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} + \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{B}}(p)} \geq 4 \log y + c.$$

*Proof.* We have

$$\begin{aligned} \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} + \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{B}}(p)} &= \sum_{p \leq y} \log p \left( \frac{1}{\nu_{\mathcal{A}}(p)} + \frac{1}{\nu_{\mathcal{B}}(p)} \right) \\ &\geq \sum_{p \leq y} \log p \left( \frac{1}{\nu_{\mathcal{A}}(p)} + \frac{1}{p - \nu_{\mathcal{A}}(p)} \right) \\ &\geq \sum_{p \leq y} \log p \left( \frac{1}{(p-1)/2} + \frac{1}{(p+1)/2} \right) + O(1) \\ &\geq \sum_{p \leq y} \frac{4 \log p}{p} \left( 1 + \frac{1}{p^2} + \frac{1}{p^4} + \dots \right) + O(1) \\ &\geq 4 \log y + c. \quad \square \end{aligned}$$

In particular, it follows that with at least one of  $\nu = \nu_{\mathcal{A}}$  or  $\nu = \nu_{\mathcal{B}}$  we have

$$\sum_{p \leq y} \frac{\log p}{\nu(p)} \geq 2 \log y + \frac{c}{2}.$$

Of course, which of the two sets  $\mathcal{A}$  and  $\mathcal{B}$  achieves this inequality may depend on  $y$ .

We now turn to the proof of Theorem 3. Recall that Theorem 3 states that there exist constants  $K_1, K_2 > 0$  such that

$$2 \log y - K_1 \log \log y \leq \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} \leq 2 \log y + K_2 \log \log y.$$

This means that it is not possible that  $\nu(p)$  is small modulo half of the primes  $p \leq y$ , and large modulo the other half of the primes. In fact, the result shows that for most primes we have  $\nu(p) = p/2 + O(p/\log p)$ .

From Theorem 3 and Theorem 4 we immediately obtain the following corollary:

**Corollary 2.**

$$4 \log y + c \leq \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} + \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{B}}(p)} \leq 4 \log y + 2K_2 \log \log y.$$

*Proof of Theorem 3.* Suppose, to get a contradiction, that

$$\sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} \geq 2 \log y + K_2 \log \log y.$$

Let  $y = K_4 N^{1/2} / (\log N)^{K_3}$  with some positive constants  $K_3$  and  $K_4$  that will be chosen below. It follows that

$$\begin{aligned} 2 \log y + K_2 \log \log y &= 2 \log K_4 + \log N - 2K_3 \log \log N + \\ &\quad K_2 \left( \log \log N + \log \frac{1}{2} + o(1) \right) \\ &= \log N + (-2K_3 + K_2) \log \log N + O(1). \end{aligned}$$

Hence, by Gallagher's larger sieve,

$$\begin{aligned} \mathcal{A}(N) &\leq \frac{y}{-\log N + \sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)}} \\ &\leq \frac{K_4 N^{1/2}}{(\log N)^{K_3} ((-2K_3 + K_2) \log \log N + O(1))}. \end{aligned}$$

Taking  $K_3 = 5$  and  $K_2 > 10$ , this yields a contradiction to the lower bound  $\mathcal{A}(N) \gg N^{1/2} / (\log N)^5$  in (1.2). Hence we have the upper bound

$$\sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{A}}(p)} \leq 2 \log y + K_2 \log \log y.$$

It follows from Theorem 4 that

$$\sum_{p \leq y} \frac{\log p}{\nu_{\mathcal{B}}(p)} \geq 2 \log y + c - K_2 \log \log y \geq 2 \log y - K_1 \log \log y.$$

Since the problem is symmetric in  $\nu_{\mathcal{A}}(p)$  and  $\nu_{\mathcal{B}}(p)$ , we have proven Theorem 3.  $\square$

## References

- [1] D. Bshouty and N.H. Bshouty, *A note on prime  $n$ -tuples*, Rocky Mountain J. Math. **27** (1997), 775–778.
- [2] C. Elsholtz, *The inverse Goldbach problem*, Mathematika, to appear.
- [3] P.X. Gallagher, *A larger sieve*, Acta Arith. **18** (1971), 77–81.
- [4] A. Hofmann and D. Wolke, *On additive decompositions of the set of primes*, Arch. Math. **67** (1996), 379–382.
- [5] H.L. Montgomery, *The analytic principle of the large sieve*, Bull. Amer. Math. Soc. **84** (1978), 547–567.

- [6] M.B. Nathanson, *Additive number theory. Inverse problems and the geometry of sumsets*, Springer-Verlag, New York, 1996.
- [7] H.-H. Ostmann, *Additive Zahlentheorie. 1. Teil: Allgemeine Untersuchungen*, (1968).
- [8] C. Pomerance, A. Sárközy, and C.L. Stewart, *On divisors of sums of integers. III*, Pacific J. Math. **133** (1988), 363–381.
- [9] R.C. Vaughan, *Some applications of Montgomery's sieve*, J. Number Theory **5** (1973), 64–79.
- [10] E. Wirsing, *Über additive Zerlegungen der Primzahlmenge*, unpublished manuscript; abstract in Tagungsbericht 28/1972, Oberwolfach.