.

LONG GAPS IN SIEVED SETS

KEVIN FORD, SERGEI KONYAGIN, JAMES MAYNARD, CARL POMERANCE, AND TERENCE TAO

ABSTRACT. To each prime p, let $I_p \subset \mathbb{Z}/p\mathbb{Z}$ denote a collection of at most C_0 residue classes modulo p, whose cardinality $|I_p|$ is equal to 1 on the average. We show that for sufficiently large x, the sifted set $\{n \in \mathbb{Z} : n \pmod p \notin I_p \text{ for all } p \leqslant x\}$ contains gaps of size $x(\log x)^{1/\exp(CC_0)}$ for an absolute constant $C > 0$; this improves over the "trivial" bound of $\gg x$. As a consequence, we show that for any degree d polynomial $f : \mathbb{Z} \to \mathbb{Z}$ mapping the integers to itself with positive leading coefficient, the set $\{n \leqslant X : f(n) \text{ composite}\}$ contains an interval of consecutive integers of length $\geqslant (\log X)(\log\log X)^{1/\exp(Cd)}$ for some absolute constant $C > 0$ and sufficiently large X.

CONTENTS

1. INTRODUCTION

In this paper, we will show the existence of large gaps within sets formed by sieving out from the integers a bounded number of residue classes modulo each prime. The setup will be as follows.

Definition 1 (Sieving system)**.** *Let $\mathcal{P} = \{2, 3, 5, \dots\}$ denote the primes. A* sieving system *is a collection $\mathcal{I} = (I_p)_{p\in\mathcal{P}}$ of sets $I_p \subset \mathbb{Z}/p\mathbb{Z}$ of residue classes modulo p for each prime $p \in \mathcal{P}$. We say that the sieving system is* non-degenerate *and C_0-*bounded *if the cardinalities $|I_p|$ obey the bound*

$$|I_p| \leqslant \min(C_0, p-1) \quad (p \in \mathcal{P}). \tag{1.1}$$

KF was supported by National Science Foundation grant DMS-1501982. JM was supported by a Clay Research Fellowship and a Fellowship of Magdalen College, Oxford. TT was supported by a Simons Investigator grant, the James and Carol Collins Chair, the Mathematical Analysis & Application Research Fund Endowment, and by NSF grant DMS-1266164. Part of this work was carried out at MSRI, Berkeley during the Spring semester of 2017, supported in part by NSF grant DMS-1440140.

2010 Mathematics Subject Classification: Primary 11N35, 11N32, 11B05.

Keywords and phrases: gaps, prime values of polynomials, sieves.

We say that the sieving system is one-dimensional *if we have the weighted prime number theorem*[1]

$$\sum_{p\leqslant x} |I_p| = \frac{x}{\log x} + O\left(\frac{x}{\log x(\log_2 x)^2}\right). \tag{1.2}$$

For any integer b, bounded set $J \subset \mathbb{R}$ and sieving system $\mathcal{I}$, we define the sifted set $S_J(b,\mathcal{I}) \subset \mathbb{Z}$ *by the formula*

$$S_J(b,\mathcal{I}) := \{n \in \mathbb{Z} : n - b \pmod p \notin I_p \, \forall p \in J \cap \mathcal{P}\}.$$

Remark 1. If $|I_p| = p$ for some $p \in J$ (the degenerate case), then clearly $S_J(b,\mathcal{I})$ is empty. Otherwise, $S_J(b,\mathcal{I})$ is a P_J- periodic set with density $\sigma_J(\mathcal{I})$, where P_J and $\sigma_J(\mathcal{I})$ are defined as

$$P_J := \prod_{p\in J} p, \qquad \sigma_J(\mathcal{I}) := \prod_{p\in J}\left(1 - \frac{|I_p|}{p}\right). \tag{1.3}$$

The b parameter just shifts the sifted sets $S_J(b_J,\mathcal{I})$ by translation:

$$S_J(b,\mathcal{I}) = b + S_J(0,\mathcal{I}). \tag{1.4}$$

The main result of this paper establishes somewhat large gaps inside non-degenerate C_0-bounded sifted sets.

Theorem 1 (Main theorem)**.** *Let $\mathcal{I} = (I_p)_{p\in\mathcal{P}}$ be a non-degenerate, C_0-bounded one-dimensional sieving system for some $C_0 \geqslant 1$. Then, for all sufficiently large numbers x, the sifted set $S_{[2,x]}(0,\mathcal{I})$ contains a gap of length at least $x(\log x)^{1/\exp(CC_0)}$ for some absolute constant $C > 0$.*

Remark 2. Note that by (1.4), to prove Theorem 1 it is sufficent to show there is some $b \in \mathbb{Z}/P_{[2,x]}\mathbb{Z}$ with

$$S_{[2,x]}(b,\mathcal{I}) \cap [1, x(\log x)^{1/\exp(CC_0)}] = \emptyset.$$

The conclusion of Theorem 1 should be compared with the "trivial" bound

$$S_{[2,x]}(b,\mathcal{I}) \cap [1, c'x] = \emptyset \tag{1.5}$$

for some constant $c' > 0$ and large x. To deduce this, we first use (1.1), (1.2), and partial summation, to obtain a Mertens-type product estimate

$$\sigma_{[2,z]}(\mathcal{I}) = \prod_{p\leqslant z}\left(1 - \frac{|I_p|}{p}\right) = \frac{C_1}{\log z}\left(1 + O\left(\frac{1}{\log_2 z}\right)\right), \tag{1.6}$$

where C_1 is a certain positive constant (in general, C_1 depends on the behavior of $|I_p|$ for small p, and can have great variation even over systems with a constant C_0). If x is large, then it follows that there is some b modulo $P_{[2,x/2]}$ for which $\mathcal{A} := S_{[2,x/2]}(b,\mathcal{I}) \cap [1, x/(8C_0C_1)]$ satisfies $|\mathcal{A}| \leqslant \frac{x}{4C_0\log x}$. On the other hand, by (1.2),

$$\sum_{x/2<q\leqslant x} |I_q| \sim \frac{x/2}{\log x}$$

and hence by (1.1),

$$\#\{x/2 < q \leqslant x : |I_q| \geqslant 1\} \geqslant \frac{x}{4C_0\log x}. \tag{1.7}$$

[1]The notation $\log_2 x$ means $\log\log x$, $\log_3 x$ means $\log\log\log x$, etc. The symbol p always denotes a prime.

Hence, we may pair up each element $a \in \mathcal{A}$ with a unique prime $q = q_a \in (x/2, x]$ for which $|I_q| \geqslant 1$. For each such pair a, q_a let $v_a \in I_{q_a}$ and suppose that $b \equiv a - v_a \pmod q$. It follows that $S_{[2,x]}(b, \mathcal{I}) \cap [1, x/(8C_0C_1)] = \emptyset$, proving (1.5).

Example 1. The "Eratosthenes" sieving system $\mathcal{I} = (\{0 \pmod p\})_{p\in\mathcal{P}}$ is non-degenerate, 1-bounded, and 1-dimensional. The Eratosthenes sieve asserts that

$$\mathcal{P} \cap (\sqrt{x}, x] = S_{[2,\sqrt{x}]}(0, \mathcal{I}) \cap (\sqrt{x}, x]. \tag{1.8}$$

The problem of finding large gaps between consecutive primes has a long history, and it is currently known that gaps of size

$$\gg \log x \, \frac{\log_2 x \, \log_4 x}{\log_3 x} \tag{1.9}$$

exist below x if x is large enough [7]; this improves over the "trivial" bound of $\gg \log x$ from the prime number theorem or the more elementary Chebyshev estimates. The existing methods for finding large gaps between primes, however, seem not to be adaptable to find large gaps in more general sieving systems. We will discuss the reasons for this in detail below. Our proof of Theorem 1 follows a very different route, relying on a new probabilistic method.

Example 2. Given a polynomial $f : \mathbb{Z} \to \mathbb{Z}$ of degree $d \geqslant 1$, the system

$$I_p := \{n \in \mathbb{Z}/p\mathbb{Z} : f(n) \equiv 0 \pmod p\}$$

is d-bounded, and is non-degenerate if and only if there does not exist a prime p that divides all the values $f(n), n \in \mathbb{Z}$. For irreducible f, the 1-dimensionality (1.2) follows quickly from Landau's Prime Ideal Theorem [13] (see also [3, pp. 35–36]). As a variant of (1.8), we observe that

$$\{n \in A : f(n) \in \mathcal{P}\} \subset S_{[2,x]}(0, \mathcal{I}) \tag{1.10}$$

for any $x \geqslant 1$ and any set A such that $f(n) > x$ for all $n \in A$. Now set $x := \frac{1}{2} \log X$. By Theorem 1, the set $S_{[2,x]}(0, \mathcal{I})$ contains a gap of length $\gg \log X (\log_2 X)^{1/\exp(Cd)}$. The period of this set is $P_{[2,x]}$, which by the prime number theorem is $X^{1/2+o(1)}$. Thus, this set contains such a long gap inside the interval $[X/2, X]$. Assuming that f has a positive leading coefficient and that X is large, on this interval, $f(n) > x$, and hence $f(n)$ is composite for every $n \in [X/2, X] \backslash S_{[2,x]}(0, \mathcal{I})$. We thus obtain the following.

Theorem 2. *Let $f : \mathbb{Z} \to \mathbb{Z}$ be a polynomial of degree $d \geqslant 1$ with positive leading term. Then for sufficiently large X, there is a string of consecutive natural numbers $n \in [1, X]$ of length $\geqslant \log X (\log_2 X)^{1/\exp(Cd)}$ for which $f(n)$ is composite, for some absolute constant $C > 0$.*

We note that when f has degree two or greater, it is still an open conjecture (of Bunyakovsky [1]) that there are infinitely many n for which $f(n)$ is prime; we do not address this conjecture at all in this paper. Also, Theorem 2 follows trivially in the "degenerate" cases, when either f is reducible, or there is some prime p with $|I_p| = p$.

Remark 3. While the polynomial f must be in $\mathbb{Q}[x]$, it need not have integer coefficients, e.g. $f(n) = \frac{n^7-n+7}{7}$ satisfies the hypotheses of Theorem 2.

Example 3. A simple example to keep in mind (and in fact the original example we studied) is $f(n) = n^2 + 1$. In this case, $I_2 = \{1\}$, I_p is empty for $p \equiv 3 \pmod 4$, and $I_p = \{\iota_p, -\iota_p\}$ for $p \equiv 1 \pmod 4$, where $\iota_p \in \mathbb{Z}/p\mathbb{Z}$ is one of the square roots of -1. Here one can use the prime number theorem in arithmetic progressions rather than the Prime Ideal theorem to establish one-dimensionality. For this example (and for any quadratic polynomial), our methods produce (taking $M = 4 + 10^{-15}$ in (2.1)) consecutive composite strings of length $\gg (\log X)(\log_2 X)^{K_2}$, where $K_2^{-1} \approx 1.144 \times 10^{20}$. It is certain that further numerical improvements are possible.

Theorem 1 has another application, to a problem on the coprimality of consecutive values of polynomials.

Corollary 1. *Let $f : \mathbb{Z} \to \mathbb{Z}$ be a non-constant polynomial. Then there exists an integer $G_f \geqslant 2$ such that for any integer $k \geqslant G_f$ there are infinitely many integers $n \geqslant 0$ with the property that none of $f(n+1), \dots, f(n+k)$ is coprime to all the others.*

Proof. Let $d = \deg f$. Then $d!f(x) \in \mathbb{Z}[x]$. Let $f_0(x) \in \mathbb{Z}[x]$ be a primitive irreducible factor of $d!f(x)$. If $p > d$ is a prime and $p \mid f_0(m)$ for some integer m, then $p \mid f(m)$. So it will suffice to consider the case that f is irreducible and show in this case that for all large k there are infinitely many $n \geqslant 0$ such that for each $i \in \{1, \dots, k\}$ there is some $j \in \{1, \dots, k\}$ with $j \neq i$ and $\gcd(f(n+i), f(n+j))$ divisible by some prime $> d$.

Again, we consider the system $\mathcal{I}$ defined by

$$I_p := \{n \in \mathbb{Z}/p\mathbb{Z} : f(n) \equiv 0 \pmod p\},$$

but for $p \leqslant d$, we take $I_p = \emptyset$. By Theorem 1, for all sufficiently large numbers x the set $S_{[2,x]}(0, \mathcal{I})$ contains a gap of length $\geqslant k = \lfloor 2x \rfloor$. Thus, there are infinitely many n such that each $f(n+1), \dots, f(n+k)$ has a prime factor p with $d < p \leqslant x$. For each $i \in \{1, \dots, k\}$, take a prime factor p of $f(n+i)$ with $d < p \leqslant x$. Since $k = \lfloor 2x \rfloor$, $p \leqslant x$ and $I_p \neq \emptyset$, it must be that p divides at least two terms of the sequence $f(n+1), \dots, f(n+k)$, thus proving the assertion. □

For linear polynomials the result of the corollary is well-known. For quadratic and cubic polynomials in $\mathbb{Z}[x]$ it was recently proven by Sanna and Szikszai [15].

1.1. **Discussion of methods.** In the case of the Eratosthenes sieving system $\mathcal{I} = (0 \pmod p)_{p \in \mathcal{P}}$, for any set J and shift b, the largest gap in the infinite periodic set $S_J(b, \mathcal{I})$ is equal to $j(P_J)$, where $j(n)$ is *Jacobsthal's function* of n, defined as the largest gap between the integers coprime to n. The best bounds known for $j(P_{[2,x]})$ for primorials $P_{[2,x]}$ are

$$\frac{x \log x \log_3 x}{\log_2 x} \ll j(P_{[2,x]}) \ll x^2 \tag{1.11}$$

for sufficiently large x; see [7] for the lower bound and [11] for the upper bound. Combining the lower bound in (1.11) with the sieve of Erathosthenes (1.8) gives the lower bound (1.9) for large gaps between primes obtained in [7]. The implication only works in one direction; the upper bound in (1.11) is not known to imply any upper bound on gaps between primes. We refer the reader to [7] for further discussion and for a history of the problem.

For the Eratosthenes sieving system $\mathcal{I} = (0 \pmod p)_{p \in \mathcal{P}}$, it is clear that $S_{[2,x]}(0, \mathcal{I})$ avoids the interval $[2, x]$, which already gives the "trivial" lower bound $j(P_{[2,x]}) \gg x$. Most of the improvements to this bound in previous literature (including those in [7]) rely on the following variant of this

observation: if $x \geqslant z \geqslant 2$, then the sifted set $S_{(z,x]}(0, \mathcal{I})$, when restricted to the interval $[1, y)$ with y slightly larger than x, only consists of numbers of the form a or ap, where p is a prime in $(x, y]$, and a is a *z-smooth* (or *z-friable* number), which means that all the prime factors of a do not exceed z. Using the known bounds on the frequency of smooth numbers (see e.g. [2]), the set of numbers of the first type a can be made to be negligible by choosing z appropriately, leaving one with a set which is essentially a collection of primes p (the additional factor of a in ap can be eliminated by working with numbers close to x and doing some additional sieving at very small primes). Most of the remaining effort in establishing lower bounds for $j(P_{[2,x]})$ then goes into covering this collection of primes with congruence classes as efficiently as possible. The most recent works in this direction [14, 6, 7] rely on recent results concerning tuples of linear forms that attain many prime values simultaneously, with the bounds in [7] also relying on a combinatorial hypergraph covering lemma of Pippenger-Spencer type. Again, we refer the reader to [7] for further discussion.

Unfortunately, when considering the more general sieving systems in Definition 1, in which the cardinalities $|I_p|$ are allowed to vanish for many primes p, bounds for smooth numbers cannot be used to bound $S_{(z,x]}(0)$, and without this first crucial step the existing methods only yield the trivial lower bound of $\gg x$ for the gap size. We overcome the barrier by using a very different method; rather than choosing a specific b for some interval J, we make a random choice of b; this is done via the Chinese Remainder Theorem by choosing $b \bmod p$ for primes $p \in J$ at random, but with a complex dependency structure. This will be described in full in the next section. As in [7], we utilize a hypergraph covering lemma to obtain the best bounds, though it is possible to obtain weaker improvements to the trivial bound (gaining a small power of $\log_2 x$ rather than $\log x$ in Theorem 1) without using such a lemma.

Remark 4. Unfortunately our methods only seem to give good results in the one-dimensional case. Consider for instance the set $\{n \in \mathcal{P} : n + 2 \in \mathcal{P}\}$ of (the lower) twin primes. This corresponds to a two-dimensional system in which $I_p = \{0 \pmod p, 2 \pmod p\}$ for all primes p. The "trivial" bound coming from these methods would give a bound of $\gg \log X \log_2 X$ for the largest gap between lower twin primes up to X (or between the largest such twin prime and X), and one could possibly hope to improve this bound by a small power of $\log_2 X$; however, a sieve upper bound (e.g., [9, Cor. 2.4.1]) combined with the pigeonhole principle already gives a bound of $\gg \log^2 X$ in this case.

1.2. **Notation.** We use $X \ll Y, Y \gg X$, or $X = O(Y)$ to denote the estimate $|X| \leqslant CY$ for some constant $C > 0$, and write $X \asymp Y$ for $X \ll Y \ll X$. Throughout the remainder of the paper, we adopt the convention that all implied constants in O- and related order estimates may depend on C_0, C_1 and the implied constants in the O-bounds in the estimates (1.2) and (1.6). Constant implies by O- and related symbols will not depend on any other quantity. We also assume that the quantity x is sufficiently large in terms of all of these parameters.

The notation $X = o(Y)$ as $x \to \infty$ means $\lim_{x\to\infty} X/Y = 0$ (holding other parameters fixed).

If S is a statement, we use 1_S to denote its indicator, thus $1_S = 1$ when S is true and $1_S = 0$ when S is false.

We will rely on probabilistic methods in this paper. Boldface symbols such as $\mathbf{n}$, $\mathbf{S}$, $\boldsymbol{\lambda}$ etc. denote random variables (which may be real numbers, random sets, random functions, etc.). Most of these random variables will be discrete (in fact they will only take on finitely many values), so that we may ignore any technical issues of measurability; however it will be convenient to use some continuous

random variables in the appendix. We use $\mathbb{P}(E)$ to denote the probability of a random event E, and $\mathbb{E}\mathbf{X}$ to denote the expectation of the random (real-valued) variable $\mathbf{X}$.

For a natural number d, we use $\omega(d)$ to denote the number of primes p with $p \mid d$.

Unless specified, all sums are over the natural numbers. An exception is made for sums over the variables p or q (as well as variants such as p_1, p_2, etc..), which will always denote primes.

Given two subsets A, B of an additive group G, we define the sumset and difference set

$$A+B := \{a+b : a \in A, b \in B\},$$
$$A-B := \{a-b : a \in A, b \in B\}.$$

Similarly we define the translates

$$A+h = h+A := \{a+h : a \in A\},$$
$$A-h := \{a-h : a \in A\}$$

for $h \in G$, and for any integer n, we define the dilate

$$n \cdot A := \{na : a \in A\}.$$

2. Overall strategy

In this section we describe the high-level strategy of proof, and perform two initial reductions on the problem, ultimately leaving one with the task of proving Theorem 3 below.

Let $\mathcal{I}$ be a non-degenerate, C_0-bounded one-dimensional sieving system. Let $\varepsilon > 0$ be an arbitrarily small, fixed real number, and let $M \geqslant 1$ be a fixed, positive real number which will be optimized later. Suppose x is large (think of $x \to \infty$), and define the quantities y and δ_1 as

$$y := x(\log x)^{\delta_1}, \quad \delta_1 := \frac{1-\varepsilon}{M\exp(5.6MC_0)}. \tag{2.1}$$

Our goal is to show that $S_{[2,x]}(b,\mathcal{I}) \cap [1,y] = \emptyset$ for some b. We make two trivial observations:

- A linear shift of any single set I_p (that is, replacing I_p by $c+I_p$ for some integer c) does not affect the structure of $S_J(b,\mathcal{I})$ (up to changing the b parameter). Thus, the same is true for linear shifts (depending on p) for any finite set of primes p. In particular, we may shift the sets I_p so that all nonempty sets I_p contain the zero element, without changing the structure of $S_J(b,\mathcal{I})$.
- The set $S_J(b,\mathcal{I})$ depends only on b modulo P_J. By the Chinese Remainder Theorem, $b \bmod P_J$ is uniquely determined by the set of residue classes $b \bmod p$ for primes $p \in J$.

Therefore, we may assume without loss of generality that $0 \in I_p$ whenever I_p is nonempty, and we will select b by choosing residue classes b modulo primes $p \leqslant x$. We introduce the parameter

$$z := \frac{y}{(\log x)^{(1-\varepsilon)/M}}, \tag{2.2}$$

so that for x large enough we have

$$1 \leqslant z \leqslant x/2 \leqslant x \leqslant y.$$

We will select the parameter b modulo the primes $p \leqslant x$ in three stages:

(1) (Uniform random stage) First, we choose b modulo $P_{[2,z]}$ uniformly at random; equivalently, for each prime $p \leqslant z$, we choose $b \bmod p$ randomly with uniform probability, independently for each p;

(2) (Greedy stage) Secondly, choose b modulo $P_{(z,x/2]}$ randomly, but dependent on the choice of b modulo $P_{[2,z]}$. A bit more precisely, for each prime $q \in (z, x/2]$ with $|I_q| \geqslant 1$, we will select $b \equiv b_q \pmod q$ so that $\{b_q + kq : k \in \mathbb{Z}\} \cap [1, y]$ knocks out nearly as many elements of the random set $S_{[2,z]}(b, \mathcal{I}) \cap [1, y]$ as possible. Note that we are focusing only on those residues sifted by the element $0 \in I_q$, and ignoring all other possible elements of I_q. This simplifies our analysis, and only has the effect of possibly reducing the exponent of $\log x$ in Theorem 1.

(3) (Clean up stage) Thirdly, we choose b modulo primes $q \in (x/2, x]$ to ensure that the remaining elements $m \in S_{[2,x/2]}(b, \mathcal{I}) \cap [1, y]$ do not lie in $S_{[2,x]}(b, \mathcal{I}) \cap [1, y]$ by matching a unique prime $q = q(m)$ with $|I_q| \geqslant 1$ to each element m and setting $b \equiv m \pmod q$. (Again we use the single element $0 \in I_q$.)

Following the argument used to show (1.5), and using (1.7), we have $S_{[2,x]}(b, \mathcal{I}) \cap [1, y] = \emptyset$ after Stage (3) provided that the size of the sifted set following Stage (2), $S_{[2,x/2]}(b', \mathcal{I}) \cap [1, y]$ (where $b' \in \mathbb{Z}/P_{[2,x/2]}\mathbb{Z}$) does not exceed $x/(4C_0 \log x)$. Hence, as a first reduction, it suffices to prove that there is some b for which

$$|S_{[2,x/2]}(b, \mathcal{I}) \cap [1, y]| \leqslant \frac{x}{4C_0 \log x}. \tag{2.3}$$

Recall that $S_J(b, \mathcal{I})$ is P_J-periodic with density $\sigma_J(\mathcal{I})$. It follows that

$$|S_J(b, \mathcal{I}) \cap [x, x + y]| = \left(1 + O\left(\frac{P_J}{y}\right)\right) \sigma_J(\mathcal{I}) y \tag{2.4}$$

for any interval $[x, x + y] \subset \mathbb{R}$.

After stage (1), from (1.6) we see that the expected size of $|S_{[2,z]}(b, \mathcal{I}) \cap [1, y]|$ is $\asymp \frac{y}{\log z} \sim \frac{y}{\log x}$. A random, uniform choice of b modulo primes $q \in (z, x/2]$ would not reduce the residual set very much further, and would lead to a version of Theorem 1 with a gap of size $\asymp x$. Instead, we use a greedy algorithm to select $b \equiv b_q \pmod q$. If q has size $q \asymp y/H$ for some H (which will lie in the range $(\log x)^{\delta_1} \ll H \ll (\log x)^{(1-\varepsilon)/M}$ by (2.1) and (2.2)), then $(b_q \bmod q) \cap [1, y]$ is a set of size about H. However, by considering the initial portion $S_{[2,H^M]}(b, \mathcal{I})$ (say) of the sieving process, one can see (e.g. using the large sieve [8, Lemma 7.5 and Cor. 9.9] or Selberg's sieve [10, Sec. 1.2]) that the size of the intersection $(b_q \bmod q) \cap S_{[2,H^M]}(b, \mathcal{I}) \cap [1, y]$ must be somewhat smaller, namely of size

$$\ll \sigma_{[2,H]} H \asymp \frac{H}{\log H}.$$

However, because of the substantial freedom to choose b_q (amongst q different choices), one might hope that no further size reduction occurs when one sieves up to z instead of H^M. Namely, one can hope to choose b_q such that

$$(b_q \bmod q) \cap S_{[2,H^M]}(b, \mathcal{I}) \cap [1, y] = (b_q \bmod q) \cap S_{[2,z]}(b, \mathcal{I}) \cap [1, y]. \tag{2.5}$$

Each individual choice of b_q might only be expected to obey this condition with probability that looks (very) roughly like $\sigma_{(H^M,z]}^{H\sigma_{[2,H]}}$, but with our choice of parameters and (1.6), this quantity is substantially larger than $1/q$, and so there should be many possibilities for b_q for each q. If one chooses each of the b_q independently of each other, standard computations then suggest that we will achieve (2.3) if $x = y(\log_2 y)^{-c_0/C_0}$ for small enough $c_0 > 0$. To succeed with the stronger relation (2.1), we use a hypergraph covering lemma of Pippenger-Spencer type introduced in [7].

It is convenient to separately consider the primes $q \in (z, x/2]$ in dyadic blocks. Let

$$\mathfrak{H} := \left\{ H \in 2^{\mathbb{Z}} : 2(\log x)^{\delta_1} = \frac{2y}{x} \leqslant H \leqslant \frac{y}{2z} = \frac{1}{2}(\log x)^{(1-\varepsilon)/M} \right\} \tag{2.6}$$

be the set of relevant dyadic scales H. From (1.2) we have

$$\sum_{y/2H<q\leqslant y/H} |I_q| = \frac{(1+o(1))y}{2H\log x} \quad (x \to \infty)$$

uniformly in H and hence (by (1.1)) we can find a collection $\mathcal{Q}_H$ of primes q in $(y/2H, y/H]$ with $|I_q| \geqslant 1$ of cardinality

$$|\mathcal{Q}_H| = \left\lceil \frac{(1-\varepsilon)y}{2C_0 H \log x} \right\rceil. \tag{2.7}$$

Fix these sets $\mathcal{Q}_H$ and let

$$\mathcal{Q} = \bigcup_{H\in\mathfrak{H}} \mathcal{Q}_H.$$

With H fixed, we will examine separately the effect of the sieving by primes in $[2, H^M]$ and by the primes in $(H^M, z]$. We denote by $\mathbf{b}$ a random residue class from $\mathbb{Z}/P\mathbb{Z}$, chosen with uniform probability, where we adopt the abbreviations

$$P = P_{[2,z]}, \quad \sigma = \sigma_{[2,z]}(\mathcal{I}), \quad \mathbf{S} = S_{[2,z]}(\mathbf{b}, \mathcal{I}) \tag{2.8}$$

as well as the projections

$$P_1 = P_{[2,H^M]}, \quad \sigma_1 = \sigma_{[2,H^M]}(\mathcal{I}), \quad \mathbf{b}_1 \equiv \mathbf{b} \pmod{P_1}, \quad \mathbf{S}_1 = S_{[2,H^M]}(\mathbf{b}, \mathcal{I}) \tag{2.9}$$

and

$$P_2 = P_{(H^M,z]}, \quad \sigma_2 = \sigma_{(H^M,z]}(\mathcal{I}), \quad \mathbf{b}_2 \equiv \mathbf{b} \pmod{P_2}, \quad \mathbf{S}_2 = S_{(H^M,z]}(\mathbf{b}, \mathcal{I}) \tag{2.10}$$

with the convention that $\mathbf{b}_1 \in \mathbb{Z}/P_1\mathbb{Z}$ and $\mathbf{b}_2 \in \mathbb{Z}/P_2\mathbb{Z}$. Thus, $\mathbf{b}_1$ and $\mathbf{b}_2$ are each uniformly distributed, and independent of each other. We also have the obvious relations

$$P = P_1P_2, \quad \sigma = \sigma_1\sigma_2, \quad \mathbf{S} = \mathbf{S}_1 \cap \mathbf{S}_2.$$

For each $q \in \mathcal{Q}_H$ and $n \in [-y, y]$, define the random set

$$\mathbf{AP}_H(q,n) := \{n + qh : 1 \leqslant h \leqslant H\} \cap \mathbf{S}_1 \tag{2.11}$$

that describes the portion of the progression $n \pmod q$ that survives the sieving process up to H^M. Central to our argument is the weight function

$$\boldsymbol{\lambda}(q,n) := \begin{cases} \sigma_2^{-|\mathbf{AP}_H(q,n)|} & \text{if } \mathbf{AP}_H(q,n) \subset \mathbf{S}_2 \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

Informally, $\boldsymbol{\lambda}(q,n)$ then isolates those n with the (somewhat unlikely) property that the portion $\mathbf{AP}_H(q,n)$ of the arithmetic progression $n \bmod q$ that survives the sieving process up to H^M, in fact also survives the sieving process all the way up to z. This additional survival is only expected to occur with probability about $\sigma_2^{|\mathbf{AP}_H(q,n)|}$, and the weight is given to exactly counteract this probability, so that we anticipate $\boldsymbol{\lambda}(q,n)$ to be about 1 on average over n. In particular, $\boldsymbol{\lambda}(q,n)$ will be concentrated on those n for which $\mathbf{AP}_H(q,n)$ is large.

We need the weights to satisfy the following bounds, for some specific choice b of $\mathbf{b}$. For this particular choice, which is now deterministic rather than random, we use the non-boldface notation S, $\lambda(q,n)$, etc.

Theorem 3 (Second reduction)**.** *Suppose that $M > 4 + \delta_1$ and $\varepsilon > 0$ is sufficiently small. There exists an integer b and a set $\mathcal{Q}' \subset \mathcal{Q}$ so that*

(i) *we have*

$$|S \cap [1,y]| \leqslant 2\sigma y, \tag{2.13}$$

(ii) *for all $q \in \mathcal{Q}'$, one has*

$$\sum_{-y<n\leqslant y} \lambda(q,n) = \left(1 + O\left(\frac{1}{(\log x)^{\delta_1(1+\varepsilon)}}\right)\right) 2y, \tag{2.14}$$

(iii) *for all but at most $\frac{x}{8C_0 \log x}$ elements n of $S \cap [1,y]$, one has*

$$\sum_{q\in\mathcal{Q}'} \sum_{h\leqslant H_q} \lambda(q, n-qh) = \left(C_2 + O\left(\frac{1}{(\log x)^{\delta_1(1+\varepsilon)}}\right)\right) 2y \tag{2.15}$$

for some quantity C_2 independent of n with

$$\frac{5}{4}\log 5 \leqslant C_2 \leqslant 10, \tag{2.16}$$

where H_q denotes the largest power of two less than y/q.

Remark 5. Axiom (ii) asserts that $\lambda(q,n)$ behaves like $O(1)$ on average in some sense. However, when n is drawn from the smaller set $S \cap [1,y]$ (which has density $\approx \sigma$ in $[1,y]$, as indicated by (2.4)), the quantity $\lambda(q, n-qh)$ appearing in Axiom (iii) is biased to be a bit larger (in our construction, it will eventually behave like $\frac{\log y}{\log(y/q)}$ on the average). It is this bias that ultimately allows us to gain somewhat over the trivial bound of $\gg x$ on the gap size in Theorem 1.

We conclude this section by deducing (2.3) (and hence Theorem 1) from Theorem 3. Let V be the set of all $n \in S \cap [1,y]$ such that (2.15) holds; Axiom (iii) then asserts that V contains all but at most $\frac{x}{8C_0 \log x}$ elements of $S \cap [1,y]$. For each $q \in \mathcal{Q}'$, we choose a random integer $\mathbf{n}_q$ with probability density function

$$\mathbb{P}(\mathbf{n}_q = n) = \frac{\lambda(q,n)}{\sum_{-y<n'\leqslant y} \lambda(q,n')}; \tag{2.17}$$

note from Axiom (ii) that the denominator is non-zero, so that this is a well-defined probability distribution. We will not need to assume any independence hypotheses on the $\mathbf{n}_q$. For each $q \in \mathcal{Q}'$, we then define the random subset $\mathbf{e}_q$ of V by the formula

$$\mathbf{e}_q := V \cap \{\mathbf{n}_q + hq : 1 \leqslant h \leqslant H_q\}. \tag{2.18}$$

We will construct a further random subset $\mathbf{e}'_q$ of V for each $q \in \mathcal{Q}'$ so that the essential support of $\mathbf{e}'_q$ is contained in that of $\mathbf{e}_q$ union the empty set. That is to say, for any non-empty subset A of

V, the probability $\mathbb{P}(\mathbf{e}'_q = A)$ is only non-zero when $\mathbb{P}(\mathbf{e}_q = A)$ is non-zero. Furthermore, these random subsets will satisfy

$$\mathbb{E}\Big|V \setminus \bigcup_{q\in\mathcal{Q}'} \mathbf{e}'_q\Big| \leqslant \frac{x}{8C_0 \log x}. \tag{2.19}$$

Then by Markov's inequality, one can find, for each q, a set e_q in the essential support of $\mathbf{e}'_q$ (and hence either empty or in the essential support of $\mathbf{e}_q$) such that

$$|V \setminus \bigcup_{q\in\mathcal{Q}} e_q| \leqslant \frac{x}{8C_0 \log x}.$$

By construction, for each $q \in \mathcal{Q}'$ there is a number n_q such that

$$e_q \subset \{n \in V : n \equiv n_q \pmod q\}.$$

Taking $b \equiv n_q \pmod q$ for all $q \in \mathcal{Q}'$, we find that

$$\big|S_{[2,x/2]}(b,\mathcal{I}) \cap [1,y]\big| \leqslant \frac{x}{8C_0 \log x} + \frac{x}{8C_0 \log x} = \frac{x}{4C_0 \log x},$$

as required for (2.3).

It remains to locate random subsets $\mathbf{e}'_q$ obeying (2.19). A naive first attempt would be to just take each $\mathbf{e}'_q$ to be a copy of $\mathbf{e}_q$ in such a fashion that the $\mathbf{e}'_q$ are independent in q. However, if one does this, then the quantity $|V \setminus \bigcup_{q\in\mathcal{Q}'} e_q|$ will only be smaller than $|V|$ by a factor of about $\exp(-C_1)$, largely thanks to (2.15). This is not sufficient for our application (although a modification of this approach will give Theorem 1 with the quantity $x(\log x)^{1/\exp(CC_0)}$ replaced by $x(\log_2 x)^{c/C_0}$ for some absolute constant $c > 0$). Instead, we use the following hypergraph covering lemma, which is a minor variant of [7, Corollary 4]:

Lemma 2.1 (Hypergraph covering lemma)**.** *Let y be a large number, and let $V, \mathcal{A}$ be finite sets with $|\mathcal{A}| \leqslant y$ and $|V| \leqslant y$. For each $q \in \mathcal{A}$, let $\mathbf{e}_q$ be a random subset of V satisfying the size bound*

$$|\mathbf{e}_q| \leqslant r := (\log y)^{1/2}. \tag{2.20}$$

Also assume the following:

- *(Sparsity) For all $q \in \mathcal{A}$ and $v \in V$, one has*

$$\mathbb{P}(v \in \mathbf{e}_q) \leqslant y^{-1/2-1/100}. \tag{2.21}$$

- *(Uniform covering) For all $v \in V$, one has*

$$\Big|\sum_{q\in\mathcal{A}} \mathbb{P}(v \in \mathbf{e}_q) - C_2\Big| \leqslant \frac{\eta}{500} \tag{2.22}$$

for some quantities C_2 and η, independent of v, satisfying

$$\frac{5}{4}\log 5 \leqslant C_2 \leqslant 10 \tag{2.23}$$

and

$$(\log y)^{-\frac{\log 5}{4\log 10}} \leqslant \eta < 1. \tag{2.24}$$

- *(Small codegrees) For any distinct $v, v' \in V$, one has*

$$\sum_{q\in\mathcal{A}} \mathbb{P}(v, v' \in \mathbf{e}_q) \leqslant \exp\{-(\log y)^{4/5}\}. \tag{2.25}$$

Then there is a collection of random subsets $\mathbf{e}'_q$ of V for each $q \in \mathcal{A}$, with the essential support of $\mathbf{e}'_q$ contained in that of $\mathbf{e}_q$ together with the empty set, such that

$$\mathbb{E}|V \setminus \bigcup_{q\in\mathcal{A}} \mathbf{e}'_q| \ll \eta|V|. \tag{2.26}$$

This lemma is proven using almost exactly the same argument used to prove [7, Corollary 4] (after some minor changes of notation); we defer the proof to an appendix. We will apply the lemma with $\mathcal{A} = \mathcal{Q}'$ and with

$$\eta = \frac{1}{(\log x)^{\delta_1} \log_2 x}.$$

Observe from (2.13), (1.6), (2.1) that

$$\eta|V| \ll \frac{1}{(\log x)^{\delta_1} \log_2 x} \frac{y}{\log z} \ll \frac{y}{(\log x)^{1+\delta_1} \log_2 x} \ll \frac{x}{\log x \log_2 x}.$$

Hence, (2.19) follows if x is large enough. Thus, it suffices to verify the hypotheses (2.20), (2.21), (2.22), (2.25) of the lemma.

Note that if $q \in \mathcal{Q}'$, then from (2.18) and (2.6) we have

$$|\mathbf{e}_q| \leqslant H_q \leqslant \frac{y}{z} = (\log x)^{(1-\varepsilon)/M} \leqslant (\log y)^{1/M}$$

which gives (2.20). Similarly, for $n \in V$ and $q \in \mathcal{Q}'$, we have from (2.18), (2.17), (2.12), that

$$\begin{aligned}\mathbb{P}(n \in \mathbf{e}_q) &= \sum_{1\leqslant h\leqslant H_q} \mathbb{P}(\mathbf{n}_q = n - hq) \\ &\ll \frac{1}{y} \sum_{1\leqslant h\leqslant H_q} \lambda(q, n - hq) \\ &\ll \frac{1}{y} H_q \sigma_2^{-H_q} \ll \frac{1}{y^{9/10}}\end{aligned}$$

which gives (2.21) for y large enough.

Applying (2.18), (2.17), (2.14), and (2.15) successively yields

$$\begin{aligned}\sum_{q\in\mathcal{A}} \mathbb{P}(v \in \mathbf{e}_q) &= \sum_{q\in\mathcal{Q}'} \sum_{h\leqslant H_q} \mathbb{P}(\mathbf{n}_q = v - hq) \\ &= C_2 + O((\log x)^{-\delta_1-\varepsilon}),\end{aligned}$$

and (2.22) follows. We now turn to (2.25). Observe from (2.18) that for distinct $v, v' \in V$, one can only have $v, v' \in \mathbf{e}_q$ if q divides $v - v'$. Since $|v - v'| \leqslant 2y$ and $q \geqslant z > \sqrt{2y}$, there is at most one q for which this is the case, and (2.25) now follows from (2.21). This concludes the derivation of (2.3) from Theorem 3.

The proof of Theorem 3 depends on various first and second moment estimations of the weights, which are given in the next sections.

3. CORRELATIONS

We deduce Theorem 3 from the following moment calculations.

Theorem 4 (Third reduction)**.** *Assume that $M \geqslant 2$. We have the following:*

(i) *One has*

$$\mathbb{E}|\mathbf{S} \cap [1,y]|^j = \left(1 + O\left(\frac{1}{(\log x)^{1-\varepsilon}}\right)\right)(\sigma y)^j \qquad (j = 0,1,2). \tag{3.1}$$

(ii) *For every $H \in \mathfrak{H}$,*

$$\mathbb{E}\sum_{q\in\mathcal{Q}_H}\left(\sum_{-y<n\leqslant y}\boldsymbol{\lambda}(q,n)\right)^j = \left(1 + O\left(\frac{1}{H^{M-2}}\right)\right)(2y)^j|\mathcal{Q}_H| \qquad (j = 0,1,2).$$

(iii) *For every $H \in \mathfrak{H}$,*

$$\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\left(\sum_{q\in\mathcal{Q}_H}\sum_{h\leqslant H}\boldsymbol{\lambda}(q,n-qh)\right)^j = \left(1 + O\left(\frac{1}{H^{M-2}}\right)\right)\left(\frac{|\mathcal{Q}_H|H}{\sigma_2}\right)^j \sigma y \tag{3.2}$$

for $j = 0,1,2$.

Note that for every $n \in [1,y]$ and $h \leqslant H$ we have $n - qh \in [-y,y]$, so the quantity in (3.2) is well-defined. As with the previous theorem, the quantity $\lambda(q,n)$ behaves like 1 on the average when n is drawn from $[-y,y] \cap \mathbb{Z}$, but for n drawn from $\mathbf{S} \cap [1,y]$ (in particular, $n \in \mathbf{S}_2$), the quantity $\lambda(q,n-qh)$ is now biased to have an average value of approximately σ_2^{-1} because $n-qh+qh = n$ is automatically in $\mathbf{S}_2$; recall the definition (2.12) of $\lambda(q,n-qh)$.

Deduction of Theorem 3 from Theorem 4. We draw $\mathbf{b}$ uniformly at random from $\mathbb{Z}/P\mathbb{Z}$. It will suffice to generate a random set $\boldsymbol{\mathcal{Q}}'$ and a random function $\boldsymbol{\lambda}$ such that the conclusions of Theorem 3 (with b replaced by $\mathbf{b}$) hold with positive probability.

From Theorem 4(i) we have

$$\mathbb{E}\big||\mathbf{S}\cap[1,y]| - \sigma y\big|^2 \ll \frac{(\sigma y)^2}{(\log x)^{1-\varepsilon}}.$$

Hence by Chebyshev's inequality, we see that

$$\mathbb{P}\left(|\mathbf{S}\cap[1,y]| \leqslant 2\sigma y\right) = 1 - O((\log x)^{\varepsilon-1}). \tag{3.3}$$

Let $H \in \mathfrak{H}$. From Theorem 4(ii) we have

$$\mathbb{E}\sum_{q\in\mathcal{Q}_H}\left(\sum_{-y<n\leqslant y}\boldsymbol{\lambda}(q,n) - 2y\right)^2 \ll \frac{y^2|\mathcal{Q}_H|}{H^{M-2}}. \tag{3.4}$$

Now let $\boldsymbol{\mathcal{Q}}'_H$ be the subset of $q \in \mathcal{Q}_H$ with the property that

$$\Big|\sum_{-y<n\leqslant y}\boldsymbol{\lambda}(q,n) - 2y\Big| \leqslant \frac{y}{H^{1+\varepsilon}}. \tag{3.5}$$

It follows from (3.4) and (3.5) that

$$\mathbb{E}|\mathcal{Q}_H \backslash \boldsymbol{\mathcal{Q}}'_H| \ll \frac{|\mathcal{Q}_H|}{H^{M-4-2\varepsilon}}. \tag{3.6}$$

By Markov's inequality, it follows that with probability $1-O(H^{-\varepsilon})$, one has

$$|\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H| \ll \frac{|\mathcal{Q}_H|}{H^{M-4-3\varepsilon}}. \tag{3.7}$$

Assume now that $M > 4+3\varepsilon$, that is, the exponent in the denominator in (3.7) is positive. Since $\sum_{H\in\mathfrak{H}} H^{-\varepsilon} \ll (y/x)^{-\varepsilon} \ll (\log x)^{-\delta_1\varepsilon}$, with probability $1-O((\log x)^{-\delta_1\varepsilon})$ the relation (3.7) holds for every $H\in\mathfrak{H}$ simultaneously. We now set

$$\boldsymbol{\mathcal{Q}}' := \bigcup_{H\in\mathfrak{H}} \boldsymbol{\mathcal{Q}}'_H.$$

Then, on the probability $1-o(1)$ event that (3.7) holds for every H and that (3.3) holds, items (i) (2.13) and (ii) (2.14) of Theorem 3 follow upon recalling (3.5) and the lower bound $H \gg (\log x)^{\delta_1}$.

We work on part (iii) of Theorem 3 using Theorem 4(iii) in a similar fashion to previous arguments. We have

$$\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\left|\sum_{q\in\mathcal{Q}_H}\sum_{h\leqslant H}\boldsymbol{\lambda}(q,n-qh)-\frac{|\mathcal{Q}_H|H}{\sigma_2}\right|^2 \ll \frac{1}{H^{M-2}}\left(\frac{|\mathcal{Q}_H|H}{\sigma_2}\right)^2\sigma y.$$

If we let $\mathcal{E}_H$ denote the set of $n\in\mathbf{S}\cap[1,y]$ such that

$$\left|\sum_{q\in\mathcal{Q}_H}\sum_{h\leqslant H}\boldsymbol{\lambda}(q,n-qh)-\frac{|\mathcal{Q}_H|H}{\sigma_2}\right| \geqslant \frac{|\mathcal{Q}_H|H}{\sigma_2 H^{(M-2)/2-\varepsilon}}, \tag{3.8}$$

then

$$\mathbb{E}|\mathcal{E}_H| \ll \frac{\sigma y}{H^{\varepsilon}}.$$

By Markov's inequality, we conclude that $|\mathcal{E}_H|\leqslant \sigma y/H^{\varepsilon/2}$ with probability $1-O(H^{-\varepsilon/2})$.

We next estimate the contribution from "bad" primes $q\in\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H$. For any $h\leqslant H$, by Cauchy-Schwarz we have

$$\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\sum_{q\in\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H}\boldsymbol{\lambda}(q,n-hq) \leqslant \big(\mathbb{E}|\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H|\big)^{1/2}\left(\mathbb{E}\sum_{q\in\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H}\big|\sum_n\boldsymbol{\lambda}(q,n-hq)\big|^2\right)^{1/2}$$

and by the triangle inequality, (3.4) and (3.6),

$$\begin{aligned}\mathbb{E}\sum_{q\in\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H}\big|\sum_n\boldsymbol{\lambda}(q,n-hq)\big|^2 &\leqslant 2\mathbb{E}\sum_{q\in\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H}\left(\big|\sum_n\boldsymbol{\lambda}(q,n-hq)-2y\big|^2+4y^2\right)\\ &\ll \frac{y^2|\mathcal{Q}_H|}{H^{M-4-2\varepsilon}}.\end{aligned}$$

Therefore, by (3.6) and summing over $h\leqslant H$,

$$\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\sum_{q\in\mathcal{Q}_H\backslash\boldsymbol{\mathcal{Q}}'_H}\sum_{h\leqslant H}\boldsymbol{\lambda}(q,n-hq) \ll \frac{y|\mathcal{Q}_H|}{H^{M-5-2\varepsilon}}.$$

Let $\mathcal{E}'_H$ denote the set of $n \in \mathbf{S} \cap [1, y]$ so that

$$\sum_{q \in \mathcal{Q}_H \setminus \boldsymbol{\mathcal{Q}}'_H} \sum_{h \leqslant H} \boldsymbol{\lambda}(q, n - hq) \geqslant \frac{|\mathcal{Q}_H| H}{H^{(1+\varepsilon)\delta_1} \sigma_2}. \tag{3.9}$$

Then

$$\mathbb{E}|\mathcal{E}'_H| \ll \frac{y H^{\delta_1(1+\varepsilon)} \sigma_2}{H^{M-4-2\varepsilon}} \ll \sigma y \frac{\log H}{H^{M-4-\delta_1-3\varepsilon}}.$$

By Markov's inequality, $|\mathcal{E}'_H| \leqslant \sigma y / H^{\varepsilon}$ with probability $1 - O(1/H^{M-4-\delta_1-5\varepsilon})$.

Now suppose that

$$M > 4 + \delta_1,$$

ε is small enough so that $M - 4 - \delta_1 - 5\varepsilon > \varepsilon$, and that we are in the event that (3.3) holds, and that for every H, we have (3.7), $|\mathcal{E}_H| \leqslant \sigma y / H^{\varepsilon/2}$ and $|\mathcal{E}'_H| \leqslant \sigma y / H^{1+\varepsilon}$. This simultaneous event happens with positive probability on account of $\sum_{H \in \mathfrak{H}} H^{-\eta} \ll (\log x)^{-\delta_1 \eta}$ for any $\eta > 0$. As mentioned before, items (i) and (ii) of Theorem 3 hold. Now let

$$\mathcal{N} = \mathbf{S} \cap [1, y] \setminus \bigcup_{H \in \mathfrak{H}} (\mathcal{E}_H \cup \mathcal{E}'_H).$$

The number of exceptional elements satisfies

$$\left| \bigcup_{H \in \mathfrak{H}} (\mathcal{E}_H \cup \mathcal{E}'_H) \right| \ll \frac{\sigma y}{(\log x)^{\delta_1(1+\varepsilon)}},$$

which is smaller than $\frac{x}{8C_0 \log x}$ for large x. It remains to verify (2.15) for $n \in \mathcal{N}$. Since $n \notin \mathcal{E}_H$ and $n \notin \mathcal{E}'_H$ for every H, the inequalities opposite to those in (3.8) and (3.9) hold, and we obtain

$$\begin{aligned} \sum_{q \in \boldsymbol{\mathcal{Q}}'} \sum_{h \leqslant H_q} \boldsymbol{\lambda}(q, n - qh) &= \sum_{H \in \mathfrak{H}} \sum_{q \in \boldsymbol{\mathcal{Q}}'_H} \sum_{h \leqslant H} \boldsymbol{\lambda}(q, n - qh) \\ &= \left(1 + O\left(\frac{1}{(\log x)^{(1+\varepsilon)\delta_1}}\right)\right) C_2 \times 2y \end{aligned}$$

where

$$C_2 := \frac{1}{2y} \sum_{H \in \mathfrak{H}} \frac{|\mathcal{Q}_H| H}{\sigma_2}.$$

From (2.7), we see that C_2 does not depend on n (C_2 depends only of x). Using (1.6), (2.7), (2.2), (2.1), we thus have, as $x \to \infty$,

$$\begin{aligned} C_2 &= \frac{1+o(1)}{2y} \sum_{H \in \mathfrak{H}} \frac{\log z}{\log(H^M)} \, \frac{(1-\varepsilon)y}{2C_0 H \log x} H \\ &= \frac{1-\varepsilon+o(1)}{4MC_0} \sum_{H \in \mathfrak{H}} \frac{1}{\log H} \\ &= \frac{1-\varepsilon+o(1)}{4MC_0} \sum_{2(\log x)^{\delta_1} \leqslant 2^j \leqslant \frac{1}{2}(\log x)^{(1-\varepsilon)/M}} \frac{1}{j \log 2} \\ &= (1-\varepsilon+o(1)) \frac{\log\left(\frac{1-\varepsilon}{M\delta_1}\right)}{4MC_0 \log 2} \\ &= (1-\varepsilon+o(1)) \left(\frac{1.4}{\log 2} + \frac{\log(1-\varepsilon)}{4MC_0 \log 2} \right) \geqslant \frac{5}{4} \log 5, \end{aligned}$$

giving (2.16) for small enough ε. □

It remains to establish Theorem 4. This is the objective of the next section of the paper.

4. Computing correlations

On account of (2.2) and (2.6), we have $H^M \leqslant (\log x)^{1-\varepsilon}$ and consequently

$$P_1 \leqslant \exp\left[O((\log x)^{1-\varepsilon})\right]. \tag{4.1}$$

Hence, by (2.4), $\mathbf{S}_1$ has good behavior in intervals of length larger than P_1, regardless of the choice of $\mathbf{b}_1$.

In our verification of the claims in Theorem 4, we will frequently need to compute k-point correlations of the form

$$\mathbb{P}\left(n_1, \dots, n_k \in \mathbf{S}_2\right)$$

for various integers $n_1, \dots, n_k$ (not necessarily distinct). Heuristically, since $\mathbf{S}_2$ has density σ_2, we expect these correlations to be approximately equal to σ_2^k for typical choices of $n_1, \dots, n_k$. Unfortunately, there is some fluctuation from this prediction, most obviously when two or more of the $n_1, \dots, n_k$ are equal, but also if the reductions $n_i \pmod p, n_j \pmod p$ for some prime $p \in (H^M, z]$ have the same difference as two elements of I_p. Fortunately we can control these fluctuations to be small on average. To formalize this statement we need some notation. Let $\mathcal{D}_H \subset \mathbb{N}$ denote the collection of squarefree numbers d, all of whose prime factors lie in $(H^M, z]$. This set includes 1, but we will frequently remove 1 and work instead with $\mathcal{D}_H \backslash \{1\}$. For each $d \in \mathcal{D}_H$, let $I_d \subset \mathbb{Z}/d\mathbb{Z}$ denote the collection of residue classes $a \bmod d$ such that $a \bmod p \in I_p$ for all $p \mid d$. In particular we may form the difference set $I_d - I_d \subset \mathbb{Z}/d\mathbb{Z}$ of I_d. For any integer m and any parameter $A > 0$, we define the error function

$$E_A(m) := \sum_{d \in \mathcal{D}_H \backslash \{1\}} \frac{A^{\omega(d)}}{d} 1_{m \pmod d \in I_d - I_d}, \tag{4.2}$$

where we recall that $\omega(d)$ is the number of prime factors of ω. The quantity $E_A(m)$ looks complicated, but in practice it will be quite small as soon as we are able to perform any sort of averaging in m. We also observe that E_A is an even function: $E_A(-m) = E_A(m)$.

We then have

Lemma 4.1. *Let $1 \leqslant k \leqslant 10H$, and suppose that $n_1, \dots, n_k$ are integers (possibly with repetition), and that $M \geqslant 2$. Then we have*

$$\mathbb{P}(n_1, \dots, n_k \in \mathbf{S}_2) = \left(1 + O\left(\frac{k^2}{H^M}\right) + O\left(\frac{1}{k^2} \sum_{1 \leqslant i < j \leqslant k} E_{C_0 k^2}(n_i - n_j)\right)\right) \sigma_2^k.$$

Proof. For each prime $p \in (H^M, z]$, let $\mathbf{b}_{2,p} \in \mathbb{Z}/p\mathbb{Z}$ be the reduction of $\mathbf{b}_2$ modulo p, thus each $\mathbf{b}_{2,p}$ is uniformly distributed in $\mathbb{Z}/p\mathbb{Z}$ and the $\mathbf{b}_{2,p}$ are independent in p. Let N_p denote the set of residue classes $\{n_1 \pmod p, \dots, n_k \pmod p\}$. By the Chinese remainder theorem, we thus have

$$\begin{aligned}
\mathbb{P}(n_1, \dots, n_k \in \mathbf{S}_2) &= \prod_{p \in (H^M, z]} \mathbb{P}(n_1 \pmod p, \dots, n_k \pmod p \notin \mathbf{b}_{2,p} + I_p) \\
&= \prod_{p \in (H^M, z]} (1 - \mathbb{P}(\mathbf{b}_{2,p} \in N_p - I_p)) \\
&= \prod_{p \in (H^M, z]} \left(1 - \frac{|N_p - I_p|}{p}\right) \\
&= \sigma_2^k \prod_{p \in (H^M, z]} \left(1 - \frac{|N_p - I_p|}{p}\right) \left(1 - \frac{|I_p|}{p}\right)^{-k}.
\end{aligned}$$

We may crudely estimate the size of the difference set $N_p - I_p$ by

$$k|I_p| \geqslant |N_p - I_p| \geqslant k|I_p| - |I_p| \sum_{1 \leqslant i < j \leqslant k} 1_{n_i - n_j \pmod p \in I_p - I_p}.$$

Since $|I_p| \leqslant C_0$, $k \leqslant 10H$, and $H^M < p \leqslant z \leqslant y$, we obtain

$$\begin{aligned}
\left(1 - \frac{|N_p - I_p|}{p}\right) \left(1 - \frac{|I_p|}{p}\right)^{-k} &= \exp\left(-\frac{|N_p - I_p|}{p} + k\frac{|I_p|}{p} + O\left(\frac{k^2 C_0^2}{p^2}\right)\right) \\
&= M_p \exp\left(O\left(\frac{k^2}{p^2}\right)\right),
\end{aligned}$$

where

$$1 \leqslant M_p \leqslant \exp\left(C_0 \sum_{1 \leqslant i < j \leqslant k} \frac{1_{n_i - n_j \pmod p \in I_p - I_p}}{p}\right).$$

We have

$$\prod_{p \in (H^M, z]} \exp\left(O\left(\frac{k^2}{p^2}\right)\right) = \exp\left(O(k^2/H^M)\right) = 1 + O\left(\frac{k^2}{H^M}\right).$$

By the arithmetic mean-geometric mean inequality, we have

$$\begin{aligned}\prod_{p\in(H^M,z]} M_p &\leqslant \prod_{1\leqslant i<j\leqslant k}\prod_{p\in(H^M,z]} \exp\left\{C_0\frac{1_{n_i-n_j \pmod p\in I_p-I_p}}{p}\right\}\\ &\leqslant \frac{2}{k^2-k}\sum_{1\leqslant i<j\leqslant k}\prod_{p\in(H^M,z]} \exp\left\{C_0\left(\frac{k^2-k}{2}\right)\frac{1_{n_i-n_j \pmod p\in I_p-I_p}}{p}\right\}\\ &\leqslant \frac{2}{k^2-k}\sum_{1\leqslant i<j\leqslant k}\prod_{p\in(H^M,z]} \left(1+C_0k^2\frac{1_{n_i-n_j \pmod p\in I_p-I_p}}{p}\right)\\ &= \frac{2}{k^2-k}\sum_{1\leqslant i<j\leqslant k}\big(1+E_{C_0k^2}(n_i-n_j)\big)\\ &= 1+\frac{2}{k^2-k}\sum_{1\leqslant i<j\leqslant k} E_{C_0k^2}(n_i-n_j). \qquad \square\end{aligned}$$

To estimate the average contribution of the errors $E_{C_0k^2}(n_i-n_j)$ appearing in the above lemma, we will use the following estimate.

Lemma 4.2. *Suppose that $(m_t)_{t\in T}$ is a sequence of integers indexed by a finite set T, obeying the bounds*

$$\sum_{t\in T} 1_{m_t\equiv a \pmod d} \ll \frac{X}{\phi(d)}+Y \tag{4.3}$$

for some $X,Y>0$ and all $d\in\mathcal{D}_H\backslash\{1\}$ and $a\in\mathbb{Z}/d\mathbb{Z}$. Then, for any $0<A$ satisfying $AC_0^2\leqslant H^M$ and any integer j, one has

$$\sum_{t\in T} E_A(m_t+j) \ll X\frac{A}{H^M}+Y\exp\left(AC_0^2\log\log y\right).$$

In practice, Y will be much smaller than X, and the first term on the right-hand side will dominate.

Proof. From the Chinese remainder theorem and (1.1), we see that for any $d\in\mathcal{D}_H$, we have

$$|I_d| = \prod_{p\,|\,d}|I_p| \leqslant C_0^{\omega(d)}.$$

In particular, the difference set $I_d-I_d\subset\mathbb{Z}/d\mathbb{Z}$ obeys the bound

$$|I_d-I_d| \leqslant C_0^{2\omega(d)}.$$

From (4.2), (4.3) we thus have

$$\begin{aligned}\sum_{t\in T} E_A(m_t+j) &= \sum_{d\in\mathcal{D}_H\backslash\{1\}}\frac{A^{\omega(d)}}{d}\sum_{a\in I_d-I_d}\#\{t\in T: m_t+j\equiv a \pmod d\}\\ &\ll \sum_{d\in\mathcal{D}_H\backslash\{1\}}\frac{(AC_0^2)^{\omega(d)}}{d}\left(\frac{X}{\phi(d)}+Y\right).\end{aligned}$$

From Euler products and Mertens' theorem we have

$$\sum_{d\in\mathcal{D}_H}\frac{(AC_0^2)^{\omega(d)}}{d}=\prod_{p\in(H^M,z]}(1+AC_0^2/p)\leqslant\exp\{AC_0^2\log\log y\}$$

and

$$\sum_{d\in\mathcal{D}_H}\frac{(AC_0^2)^{\omega(d)}}{d\phi(d)}=\prod_{p\in(H^M,z]}(1+AC_0^2/p^2)\leqslant\exp\{AC_0^2/H^M\}\leqslant 1+O(A/H^M).\qquad\square$$

Proof of Theorem 4. We start with part (i), that is, (3.1). The $j=0$ case is trivial, so we consider the case $j=1$. By linearity of expectation and (1.4), we have

$$\mathbb{E}|\mathbf{S}\cap[1,y]|=\sum_{n\leqslant y}\mathbb{P}(n\in\mathbf{S}).$$

Since the set S is periodic with period P and has density σ, the summands here are all equal to σ, and the claim follows. Now we consider the $j=2$ case. Here we set $H=\max(\mathfrak{H})\asymp y/z\asymp(\log x)^{(1-\varepsilon)/M}$. By linearity of expectation and the independent splitting (2.9) and (2.10), we get

$$\begin{aligned}\mathbb{E}|\mathbf{S}\cap[1,y]|^2&=\sum_{n_1,n_2\leqslant y}\mathbb{P}\left(n_1,n_2\in\mathbf{S}\right)\\&=\sum_{n_1,n_2\leqslant y}\mathbb{P}\left(n_1,n_2\in\mathbf{S}_1\right)\mathbb{P}\left(n_1,n_2\in\mathbf{S}_2\right).\end{aligned}$$

Observe that the probability $\mathbb{P}\left(n_1,n_2\in\mathbf{S}_1\right)$ depends only on the reductions $\ell_1:\equiv n_1\pmod{P_1}$, $\ell_2:\equiv n_2\pmod{P_1}$. Also, applying Lemma 4.1, we have

$$\mathbb{P}(n_1,n_2\in\mathbf{S}_2)=\left(1+O\left(\frac{1}{H^M}\right)+O\left(E_{C_0}(n_1-n_2)\right)\right)\sigma_2^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}|\mathbf{S}\cap[1,y]|^2=\sigma_2^2\left(1+O(1/H^M)\right)&\sum_{1\leqslant\ell_1,\ell_2\leqslant P_1}\mathbb{P}\left(\ell_1,\ell_2\in\mathbf{S}_1\right)\left(\frac{y^2}{P_1^2}+O\left(\frac{y}{P_1}\right)\right)+\\&+O\left(\sigma_2^2\sum_{1\leqslant\ell_1,\ell_2\leqslant P_1}\mathbb{P}\left(\ell_1,\ell_2\in\mathbf{S}_1\right)\sum_{\substack{1\leqslant n_1,n_2\leqslant y\\ n_1\equiv\ell_1\pmod{P_1}\\ n_2\equiv\ell_2\pmod{P_1}}}E_{C_0}(n_1-n_2)\right).\end{aligned}\tag{4.4}$$

By the definition (2.9),

$$\sum_{1\leqslant\ell_1,\ell_2\leqslant P_1}\mathbb{P}\left(\ell_1,\ell_2\in\mathbf{S}_1\right)=\mathbb{E}\left|\mathbf{S}_1\cap[1,P_1]\right|^2=(\sigma_1P_1)^2.\tag{4.5}$$

Next, fix $\ell_1,\ell_2\in\mathbb{Z}/P_1\mathbb{Z}$. Direct counting shows that for any n_1, natural number d and residue class $a \bmod d$, we have

$$\#\{n_2\leqslant y:n_2\equiv\ell_2\pmod{P_1},n_1-n_2\equiv a\pmod{d}\}\ll\frac{y}{dP_1}+1\leqslant\frac{y}{\phi(d)P_1}+1.$$

Applying Lemma 4.2 to the inner sum over n_2, we deduce that

$$\sum_{\substack{1\leqslant n_1,n_2\leqslant y\\ n_1\equiv \ell_1 \pmod{P_1}\\ n_2\equiv \ell_2 \pmod{P_1}}} E_{C_0}(n_1-n_2) \ll \left(\frac{y}{P_1}\right)^2 \frac{1}{H^M} + \frac{y}{P_1}\exp\left(O(C_0^3\log\log y)\right) \ll \frac{y^2}{P_1^2 H^M} \tag{4.6}$$

using (4.1). Inserting the bounds (4.5) and (4.6) into (4.4), and recalling (4.1) again, completes the proof of the case $j=2$ of part (i). We note that $H^M \asymp (\log x)^{1-\varepsilon}$.

We now begin the proof of Theorem 4(ii). The case $j=0$ is trivial, so we turn attention to the $j=1$ claim:

$$\mathbb{E}\sum_{q\in\mathcal{Q}_H}\sum_{-y\leqslant n\leqslant y}\boldsymbol{\lambda}(q,n) = \left(1+O\left(\frac{1}{H^{M-2}}\right)\right)(2y)|\mathcal{Q}_H|. \tag{4.7}$$

The left-hand expands as

$$\mathbb{E}\sum_{q\in\mathcal{Q}_H}\sum_{-y\leqslant n\leqslant y}\frac{1_{\mathbf{AP}_H(q,n)\subset\mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}_H(q,n)|}}.$$

Recalling the splitting (2.9) and (2.10), that $\mathbf{b}_1$ and $\mathbf{b}_2$ are independent, and consequently that $\mathbf{AP}_H(q,n)$ and $\mathbf{S}_2$ are independent (since the sets $\mathbf{AP}_H(q,n)$ defined in (2.11) are determined by $\mathbf{S}_1$). The above expression then equals

$$\sum_{q\in\mathcal{Q}_H}\sum_{-y\leqslant n\leqslant y}\sum_{b_1}\mathbb{P}(\mathbf{b}_1=b_1)\sigma_2^{-|AP_H(q,n)|}\mathbb{P}(AP_H(q,n)\subset\mathbf{S}_2).$$

If we then apply Lemma 4.1, with $\mathbf{b}_1$ fixed, and ignore the condition $n+qh\in\mathbf{S}_1$ in the error term, we find that this equals

$$\sum_{q\in\mathcal{Q}_H}\sum_{-y\leqslant n\leqslant y}\left(1+O\left(\frac{1}{H^{M-2}}\right)+O\left(\frac{1}{H^2}\sum_{1\leqslant h,h'\leqslant H:h\neq h'}E_{C_0H^2}(qh-qh')\right)\right).$$

Clearly it suffices to show that

$$\sum_{q\in\mathcal{Q}_H}E_{C_0H^2}(qh-qh') \ll \frac{|\mathcal{Q}_H|}{H^{M-2}}$$

for any distinct $1\leqslant h,h'\leqslant H$. For future reference we will show the more general estimate

$$\sum_{q\in\mathcal{Q}_H}E_{C_0H^2}(qh-qh'+k) \ll \frac{|\mathcal{Q}_H|}{H^{M-2}} \tag{4.8}$$

uniformly for any integer k.

Fix $1\leqslant h,h'\leqslant H$. If $d\in\mathcal{D}_H\backslash\{1\}$ and $a \bmod d$ is a residue class, all the prime divisors of d are larger than $H^M > H \geqslant |h-h'|$; meanwhile, q is larger than z and is hence coprime to d. Thus the relation $qh-qh'\equiv a \pmod d$ only holds for q in at most one residue class modulo d, and hence by the Brun-Titchmarsh inequality we have

$$\#\{q\in\mathcal{Q}_H : qh-qh'\equiv a \pmod d\} \ll \frac{y/H}{\phi(d)\log y}$$

when (say) $d \leqslant \sqrt{y}$. For $d > \sqrt{y}$, we discard the requirement that q be prime, and obtain the crude bound

$$\#\{q \in \mathcal{Q}_H : qh - qh' \equiv a \pmod d\} \ll \frac{y/H}{\sqrt{y}}.$$

Thus for all d we have

$$\#\{q \in \mathcal{Q}_H : qh - qh' \equiv a \pmod d\} \ll \frac{y}{H\phi(d)\log y} + \frac{\sqrt{y}}{H}$$

and hence by Lemma 4.2, we may bound the left-hand side of (4.8) by

$$\begin{aligned} &\ll \frac{y}{H\log y}\frac{H^2}{H^M} + \frac{\sqrt{y}}{H}\exp\left(O(H^2C_0^3\log\log y)\right) \\ &\ll |\mathcal{Q}_H|H^{2-M} + \frac{\sqrt{y}}{H}\exp\left(O(H^2C_0^3\log\log y)\right), \end{aligned}$$

and the claim (4.8) follows from (2.7), noting that $M \geqslant 2$ implies that $H^2 \leqslant (\log x)^{1-\varepsilon}$.

Now we turn to the $j = 2$ case of Theorem 4(ii), which is

$$\mathbb{E}\sum_{q\in\mathcal{Q}_H}\left(\sum_{-y\leqslant n\leqslant y}\boldsymbol{\lambda}(q,n)\right)^2 = \left(1 + O\left(\frac{1}{H^{M-2}}\right)\right)(2y)^2|\mathcal{Q}_H|.$$

The left-hand side may be expanded as

$$\mathbb{E}\sum_{q\in\mathcal{Q}_H}\sum_{-y\leqslant n_1,n_2\leqslant y}\frac{1_{\mathbf{AP}_H(q,n_1)\cup\mathbf{AP}_H(q,n_2)\subset\mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}_H(q,n_1)|+|\mathbf{AP}_H(q,n_2)|}}.$$

Applying Lemma 4.1 and noting that $\mathbf{S}_2$ is independent of $\mathbf{AP}_H(q,n_1)$ and $\mathbf{AP}_H(q,n_2)$, we may write this as

$$\begin{aligned} \sum_{q\in\mathcal{Q}_H}\sum_{-y\leqslant n_1,n_2\leqslant y}&\left(1 + O\left(\frac{1}{H^{M-2}}\right) + \right. \\ &\left. + O\left(\frac{1}{H^2}\sum_{h,h'\leqslant H}\left(E_{4C_0H^2}(qh-qh')1_{h\neq h'} + E_{4C_0H^2}(n_1+qh-n_2-qh')\right)\right)\right). \end{aligned} \tag{4.9}$$

Using (4.8), we obtain an acceptable main term and error terms for everything except for the summands with $h = h'$. For any fixed n_2, any $d \geqslant 1$ and $a \bmod d$,

$$\#\{-y \leqslant n_1 \leqslant y : n_1 - n_2 \equiv a \pmod d\} \ll \frac{y}{d} + 1$$

so by Lemma 4.2, we have

$$\sum_{-y\leqslant n_1,n_2\leqslant y}E_{4C_0H^2}(n_1-n_2) \ll y^2\frac{H^2}{H^M} + y\exp\left(O(H^2C_0^3\log\log y)\right)\frac{y^2}{H^{M-2}}.$$

This completes the proof of the $j = 2$ case of part (ii).

Now we establish Theorem 4(iii). The $j = 0$ case follows from the $j = 1$ case of part (i) (that is, (3.1)), so we turn to the $j = 1$ case, which is

$$\mathbb{E} \sum_{n\in\mathbf{S}\cap[1,y]} \sum_{q\in\mathcal{Q}_H} \sum_{h\leqslant H} \boldsymbol{\lambda}(q, n - qh) = \left(1 + O\left(\frac{1}{H^{M-2}}\right)\right) |\mathcal{Q}_H| H\sigma_1 y.$$

It suffices to show that for each $h \leqslant H$, one has

$$\mathbb{E} \sum_{n\in\mathbf{S}\cap[1,y]} \sum_{q\in\mathcal{Q}_H} \boldsymbol{\lambda}(q, n - qh) = \left(1 + O\left(\frac{1}{H^{M-2}}\right)\right) |\mathcal{Q}_H| \sigma_1 y.$$

The left-hand side can be expanded as

$$\mathbb{E} \sum_{n\in\mathbf{S}\cap[1,y]} \sum_{q\in\mathcal{Q}_H} \frac{1_{\mathbf{AP}(q,n-qh)\subset\mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}(q,n-qh)|}}.$$

By (2.9), the constraint $n \in \mathbf{S}\cap[1, y]$ implies that $n \in \mathbf{S}_1\cap[1, y]$. Conversely, if $n \in \mathbf{S}_1\cap[1, y]$, then $n \in \mathbf{AP}(q, n-qh)$, and the condition $n \in \mathbf{S}$ is subsumed in the condition that $\mathbf{AP}(q, n-qh) \subset \mathbf{S}_2$. Thus we may replace the constraint $n \in \mathbf{S} \cap [1, y]$ here with $n \in \mathbf{S}_1 \cap [1, y]$ and rewrite the above expression as

$$\mathbb{E} \sum_{n\in\mathbf{S}_1\cap[1,y]} \sum_{q\in\mathcal{Q}_H} \frac{1_{\mathbf{AP}(q,n-qh)\subset\mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}(q,n-qh)|}}.$$

Applying Lemma 4.1 and using the independence of $\mathbf{S}_2$ and $\mathbf{S}_1$, $\mathbf{AP}(q, n - qh)$, we may write this as

$$\mathbb{E} \sum_{n\in\mathbf{S}_1\cap[1,y]} \sum_{q\in\mathcal{Q}_H} \left(1 + O\left(\frac{1}{H^{M-2}}\right) + O\left(\frac{1}{H^2} \sum_{h,h'\leqslant H:h\neq h'} E_{C_0H^2}(qh - qh')\right)\right).$$

From (2.4) and (4.1), we have

$$\mathbb{E}|\mathbf{S}_1 \cap [1, y]| = \left(1 + O\left(\frac{1}{H^M}\right)\right) \sigma_1 y, \tag{4.10}$$

and the claim now follows from (4.8).

Finally, we establish the $j = 2$ case of Theorem 4(iii), which expands as

$$\mathbb{E} \sum_{n\in\mathbf{S}\cap[1,y]} \sum_{q_1,q_2\in\mathcal{Q}_H} \sum_{h_1,h_2\leqslant H} \boldsymbol{\lambda}(q_1, n-q_1h_1)\boldsymbol{\lambda}(q_2, n-q_2h_2) = \left(1 + O\left(\frac{1}{H^{M-2}}\right)\right) |\mathcal{Q}_H|^2 H^2 \frac{\sigma_1}{\sigma_2} y.$$

It suffices to show that for any $1 \leqslant h_1, h_2 \leqslant H$, one has

$$\mathbb{E} \sum_{n\in\mathbf{S}\cap[1,y]} \sum_{q_1,q_2\in\mathcal{Q}_H} \boldsymbol{\lambda}(q_1, n - q_1h_1)\boldsymbol{\lambda}(q_2, n - q_2h_2) = \left(1 + O\left(\frac{1}{H^{M-2}}\right)\right) |\mathcal{Q}_H|^2 \frac{\sigma_1}{\sigma_2} y.$$

We can expand the left-hand side as

$$\mathbb{E} \sum_{n\in\mathbf{S}\cap[1,y]} \sum_{q_1,q_2\in\mathcal{Q}_H} \frac{1_{\mathbf{AP}_H(q_1,n-q_1h_1)\cup\mathbf{AP}_H(q_2,n-q_2h_2)\subset\mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}_H(q_1,n-q_1h_1)|+|\mathbf{AP}_H(q_2,n-q_2h_2)|}}$$

As in the $j = 1$ case, we may replace the constraint $n \in \mathbf{S} \cap [1, y]$ here with $n \in \mathbf{S}_1 \cap [1, y]$. Next, we observe that the set

$$\mathbf{AP}_H(q_1, n - q_1h_1) \cup \mathbf{AP}_H(q_2, n - q_2h_2)$$

contains at most $|\mathbf{AP}_H(q_1, n-q_1h_1)|+|\mathbf{AP}_H(q_2, n-q_2h_2)|-1$ distinct elements, as n is common to both of the sets $\mathbf{AP}_H(q_1, n - q_1h_1), \mathbf{AP}_H(q_2, n - q_2h_2)$. Thus if we apply Lemma 4.1 (noting that $\mathbf{S}_2$ is independent of $\mathbf{S}_1, \mathbf{AP}_H(q_1, n - q_1h_1)$ and $\mathbf{AP}_H(q_2, n - q_2h_2)$) after eliminating the duplicate constraint, we may write the preceding expression as

$$\sigma_2^{-1}\mathbb{E} \sum_{n\in\mathbf{S}_1\cap[1,y]} \sum_{q_1,q_2\in\mathcal{Q}_H} \left(1 + O\left(\frac{1}{H^{M-2}} + \frac{E'(q_1) + E'(q_2) + E''(q_1, q_2)}{H^2}\right)\right)$$

where

$$E'(q) := \sum_{h,h'\leqslant H:h\neq h'} E_{4C_0H^2}(qh - qh')$$

and

$$E''(q_1, q_2) := \sum_{h_1',h_2'\leqslant H:h_1\neq h_1',h_2\neq h_2'} E_{4C_0H^2}(q_1h_1' - q_1h_1 - q_2h_2' + q_2h_2).$$

Apart from the $E''(q_1, q_2)$ term, which is new, all of the above terms can be shown to be acceptable by the $j = 1$ analysis. Thus (using (4.10)) it suffices to show that

$$\sum_{q_1,q_2\in\mathcal{Q}_H} E_{4C_0H^2}(q_1h_1' - q_1h_1 - q_2h_2' + q_2h_2) \ll \frac{1}{H^{M-2}}|\mathcal{Q}_H|^2$$

for each $h_1', h_2' \leqslant H$ with $h_1' \neq h_1$, $h_2' \neq h_2$). But this follows from (4.8) (applied with q replaced by q_1 and k replaced by $-q_2h_2' + q_2h_2$, and then summing in q_2). This completes the proof of (iii), and hence of Theorem 4.

□

APPENDIX A. PROOF OF THE COVERING LEMMA

In this appendix we prove Lemma 2.1. Our main tool will be the following general hypergraph covering lemma from [7]:

Theorem A (Probabilistic covering)**.** *There exists an absolute constant $C_3 \geqslant 1$ such that the following holds. Let $D, r, A \geqslant 1$, $0 < \kappa \leqslant 1/2$, and let $m \geqslant 0$ be an integer. Let $\delta > 0$ satisfy the smallness bound*

$$\delta \leqslant \left(\frac{\kappa^A}{C_3 \exp(AD)}\right)^{10^{m+2}}. \tag{A.1}$$

Let $I_1, \dots, I_m$ be disjoint finite non-empty sets, and let V be a finite set. For each $1 \leqslant j \leqslant m$ and $i \in I_j$, let $\mathbf{e}_i$ be a random finite subset of V. Assume the following:

- *(Edges not too large) Almost surely for all $j = 1, \dots, m$ and $i \in I_j$, we have*

$$\#\mathbf{e}_i \leqslant r; \tag{A.2}$$

- *(Each sieve step is sparse) For all $j = 1, \dots, m$, $i \in I_j$ and $v \in V$,*

$$\mathbb{P}(v \in \mathbf{e}_i) \leqslant \frac{\delta}{|I_j|^{1/2}}; \tag{A.3}$$

- *(Very small codegrees) For every $j = 1, \dots, m$, and distinct $v_1, v_2 \in V$,*

$$\sum_{i \in I_j} \mathbb{P}(v_1, v_2 \in \mathbf{e}_i) \leqslant \delta \tag{A.4}$$

- *(Degree bound) If for every $v \in V$ and $j = 1, \dots, m$ we introduce the normalized degrees*

$$d_{I_j}(v) := \sum_{i \in I_j} \mathbb{P}(v \in \mathbf{e}_i) \tag{A.5}$$

and then recursively define the quantities $P_j(v)$ for $j = 0, \dots, m$ and $v \in V$ by setting

$$P_0(v) := 1 \tag{A.6}$$

and

$$P_{j+1}(v) := P_j(v) \exp(-d_{I_{j+1}}(v)/P_j(v)) \tag{A.7}$$

for $j = 0, \dots, m-1$ and $v \in V$, then we have

$$d_{I_j}(v) \leqslant D P_{j-1}(v) \qquad (1 \leqslant j \leqslant m, v \in V) \tag{A.8}$$

and

$$P_j(v) \geqslant \kappa \qquad (0 \leqslant j \leqslant m, v \in V). \tag{A.9}$$

Then there are random variables $\mathbf{e}'_i$ for each $i \in \bigcup_{j=1}^m I_j$ with the following properties:

(a) *For each $i \in \bigcup_{j=1}^m I_j$, the essential support of $\mathbf{e}'_i$ is contained in the essential support of $\mathbf{e}_i$, union the empty set singleton $\{\emptyset\}$. In other words, almost surely $\mathbf{e}'_i$ is either empty, or is a set that $\mathbf{e}_i$ also attains with positive probability.*

(b) *For any $0 \leqslant J \leqslant m$ and any finite subset e of V with $\#e \leqslant A - 2rJ$, one has*

$$\mathbb{P}\left(e \subset V \setminus \bigcup_{j=1}^{J} \bigcup_{i \in I_j} \mathbf{e}'_i\right) = \left(1 + O_{\leqslant}(\delta^{1/10^{J+1}})\right) P_J(e) \tag{A.10}$$

where

$$P_j(e) := \prod_{v \in e} P_j(v). \tag{A.11}$$

Proof. See [7, Theorem 3]. □

To derive Lemma 2.1 from Theorem A, we repeat the proof of [7, Corollary 4]. Let the notation and hypotheses be as in Lemma 2.1. In addition, we adopt a variant of the $O-$ notation: $f = O_{\leqslant}(g)$ means that $|f| \leqslant g$; that is, the implied constant is 1.

Let

$$m = \left\lceil \frac{\log(1/\eta)}{\log 5} \right\rceil; \qquad \text{so that} \quad 1 \leqslant m \leqslant \frac{\log_2 y}{3 \log 10} + 1. \tag{A.12}$$

By (2.23), $C_2 \geqslant \frac{5}{4}\log 5$ and thus we may find disjoint intervals $\mathscr{I}_1, \dots, \mathscr{I}_m$ in $[0,1]$ with length

$$|\mathscr{I}_j| = \frac{5^{1-j}\log 5}{C_2} \tag{A.13}$$

for $j = 1, \dots, m$. Let $\vec{\mathbf{t}} = (\mathbf{t}_q)_{q\in\mathcal{A}}$ be a tuple of elements $\mathbf{t}_q$ of $[0,1]$ drawn uniformly and independently at random for each $q \in \mathcal{A}$ (independently of the $\mathbf{e}_q$), and define the random sets

$$I_j = I_j(\vec{\mathbf{t}}) := \{q \in \mathcal{A} : \mathbf{t}_q \in \mathscr{I}_j\}$$

for $j = 1, \dots, m$. These sets are clearly disjoint.

We will verify (for a suitable choice of $\vec{\mathbf{t}}$) the hypotheses of Theorem A with the indicated sets I_j and random variables $\mathbf{e}_q$, and with suitable choices of parameters $D, r, A \geqslant 1$ and $0 < \kappa \leqslant 1/2$.

Let $v \in V$, $1 \leqslant j \leqslant m$ and consider the independent random variables $(\mathbf{X}_q^{(v,j)}(\vec{\mathbf{t}}))_{q\in\mathcal{A}}$, where

$$\mathbf{X}_q^{(v,j)}(\vec{\mathbf{t}}) = \begin{cases} \mathbb{P}(v \in \mathbf{e}_q) & \text{if } q \in I_j(\vec{\mathbf{t}}) \\ 0 & \text{otherwise.} \end{cases}$$

By (2.22), (A.13), and (A.12), we have for every $1 \leqslant j \leqslant m$ and $v \in V$ that

$$\begin{aligned} \sum_{q\in\mathcal{A}} \mathbb{E}\mathbf{X}_q^{(v,j)}(\vec{\mathbf{t}}) &= \sum_{q\in\mathcal{A}} \mathbb{P}(v \in \mathbf{e}_q)\mathbb{P}(q \in I_j(\vec{\mathbf{t}})) \\ &= |\mathscr{I}_j| \sum_{q\in\mathcal{A}} \mathbb{P}(v \in \mathbf{e}_q) \\ &= 5^{1-j}\log 5 + O_{\leqslant}\left(\frac{5^{1-j}}{100\cdot 5^m}\right). \end{aligned}$$

By (2.21), we have $|\mathbf{X}_q^{(n,j)}(\vec{\mathbf{t}})| \leqslant y^{-1/2-1/100}$ for all q, and hence by Hoeffding's inequality,

$$\begin{aligned} \mathbb{P}\left(\Big|\sum_{q\in\mathcal{A}} (\mathbf{X}_q^{(v,j)}(\vec{\mathbf{t}}) - \mathbb{E}\mathbf{X}_q^{(v,j)}(\vec{\mathbf{t}}))\Big| \geqslant \frac{1}{y^{1/200}}\right) &\leqslant 2\exp\left\{-2\frac{y^{-1/100}}{y^{-1-1/50}\cdot|\mathcal{A}|}\right\} \\ &= 2\exp\left\{-2y^{1/50}\right\}. \end{aligned}$$

By a union bound, the bound $|V| \leqslant y$ and (A.12), there is a deterministic choice $\vec{t}$ of $\vec{\mathbf{t}}$ (and hence $I_1, \dots, I_m$) such that for *every* $v \in V$ and *every* $j = 1, \dots, m$, we have

$$\Big|\sum_{q\in\mathcal{A}} (\mathbf{X}_q^{(v,j)}(\vec{t}) - \mathbb{E}\mathbf{X}_q^{(v,j)}(\vec{\mathbf{t}}))\Big| < \frac{1}{y^{1/200}}.$$

We fix this choice $\vec{t}$ (so that the I_j are now deterministic), and we conclude that

$$\begin{aligned} \sum_{q\in I_j} \mathbb{P}(v \in \mathbf{e}_q) &= \sum_{q\in\mathcal{A}} \mathbf{X}_q^{(v,j)}(\vec{t}) \\ &= 5^{1-j}\log 5 + O_{\leqslant}\left(\frac{5^{1-j}}{100\cdot 5^m} + \frac{1}{y^{1/20}}\right) \\ &= 5^{1-j}\log 5 + O_{\leqslant}\left(\frac{5^{1-j}}{90\cdot 5^m}\right) \end{aligned} \tag{A.14}$$

uniformly for all $j = 1, \dots, m$, and all $v \in V$. In particular, all sets I_j are nonempty.

Set

$$\delta := \exp\{-(\log y)^{4/5}\} \tag{A.15}$$

and observe from (2.21) and the bound $|I_j| \leqslant |\mathcal{A}| \leqslant y$ that the sparsity condition (A.3) holds. Also, the small codegree condition (2.25) implies the small codegree condition (A.4).

From (A.5), (A.14) and (2.24), we now have

$$d_{I_j}(v) = (1 + O_{\leqslant}(\mu))5^{-j+1}\log 5$$

for all $v \in V$, $1 \leqslant j \leqslant m$ and $\mu = \frac{1}{90\cdot 5^m}$. A routine induction using (A.6), (A.7) then shows (for y sufficiently large) that

$$P_j(v) = (1 + O_{\leqslant}(4^j\mu))5^{-j} \quad (0 \leqslant j \leqslant m). \tag{A.16}$$

In particular we have

$$d_{I_j}(v) \leqslant DP_{j-1}(v) \qquad (1 \leqslant j \leqslant m)$$

for some absolute constant D, and

$$P_j(v) \geqslant \kappa \qquad (0 \leqslant j \leqslant m),$$

where

$$\kappa \gg 5^{-m}.$$

We now set

$$A := 2rm + 1.$$

By (2.24) and (2.20), one has

$$A \ll (\log y)^{1/2}\log_2 y$$

and so

$$\frac{\kappa^A}{C_3\exp(AD)} \gg \exp\left(-O\left((\log y)^{1/2}(\log_2 y)^2\right)\right).$$

By (A.12), $10^m \leqslant 10(\log y)^{1/3}$ and hence by (A.15), we see that

$$\delta^{1/10^{m+2}} \leqslant \exp\left\{-\frac{(\log y)^{0.55}}{1000}\right\},$$

and hence (A.1) is satisfied if y is large enough. Thus all the hypotheses of Theorem A have been verified for this choice of parameters. Applying this Theorem A and using (A.16), one thus obtains random variables $\mathbf{e}'_p$ for $p \in \bigcup_{j=1}^m I_j$ whose essential range is contained in the essential range of $\mathbf{e}_p$ together with $\emptyset$, such that

$$\mathbb{P}\left(n \notin \bigcup_{j=1}^m \bigcup_{q\in I_j} \mathbf{e}'_q\right) \ll 5^{-m} \ll \eta \tag{A.17}$$

for all $n \in V$. By linearity of expectation this gives (2.26) as required.

REFERENCES

[1] V. Bouniakowsky, *Nouveaux théorèmes relatifs á la distinction des nombres premiers et á la d'ecomposition des entiers en facteurs*, Mém. Acad. Sc. St. Pétersbourg **6** (1857), 305–329.

[2] N. G. de Bruijn, *On the number of positive integers $\leqslant x$ and free of prime factors $> y$*, Nederl. Acad. Wetensch. Proc. Ser. A. **54** (1951) 50–60.

[3] A. C. Cojocaru and M. R. Murty, *An introduction to Sieve Methods and their Applications*, Cambridge University Press, 2006.

[4] H. Cramér, *Some theorems concerning prime numbers*, Ark. Mat. Astr. Fys. **15** (1920), 1–33.

[5] H. Cramér, *On the order of magnitude of the difference between consecutive prime numbers,* Acta Arith. **2** (1936), 396–403.

[6] K. Ford. B. Green, S. Konyagin, T. Tao, *Large gaps between consecutive prime numbers*, Annals of Math. **183** (2016), 935–974.

[7] K. Ford. B. Green, S. Konyagin, J. Maynard, T. Tao, *Long gaps between primes*, preprint.

[8] J. Friedlander and H. Iwaniec, *Opera de Cribro*, Amer. Math. Soc., 2010.

[9] H. Halberstam and H.-E. Richert, *Sieve Methods*, Academic Press, London, 1974.

[10] C. Hooley, *Applications of sieve methods to the theory of numbers*, Cambridge Tracts in Mathematics, No. 70, Cambridge University Press, 1976.

[11] H. Iwaniec, *On the problem of Jacobsthal*, Demonstratio Math. **11** (1978), 225–231.

[12] J. C. Lagarias, A. M. Odlyzko, *Effective versions of the Chebotarev density theorem*, Algebraic number fields: L-functions and Galois properties (Proc. Sympos., Univ. Durham, Durham, 1975), Academic Press, 1977, pp. 409–464.

[13] E. Landau, *Neuer Beweis des Primzahlsatzes und Beweis des Primidealsatzes*, Mathematische Annalen. **56**, No. 4, (1903), 645–670.

[14] J. Maynard, *Large gaps between primes*, Annals of Math. **183** (2016), 915–933.

[15] C. Sanna, M. Szikszai, *A coprimality condition on consecutive values of polynomials*, Bull. London Math. Soc. (2017), DOI: 10.1112/blms.12078. Also see arXiv:1704.01738v1.

(Corresponding author) DEPARTMENT OF MATHEMATICS, 1409 WEST GREEN STREET, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, URBANA, IL 61801, USA

E-mail address: ford@math.uiuc.edu

STEKLOV MATHEMATICAL INSTITUTE, 8 GUBKIN STREET, MOSCOW, 119991, RUSSIA

E-mail address: konyagin@mi.ras.ru

MATHEMATICAL INSTITUTE, RADCLIFFE OBSERVATORY QUARTER, WOODSTOCK ROAD, OXFORD OX2 6GG, ENGLAND

E-mail address: james.alexander.maynard@gmail.com

MATHEMATICS DEPARTMENT, DARTMOUTH COLLEGE, HANOVER, NH 03755, USA

E-mail address: carl.pomerance@dartmouth.edu

DEPARTMENT OF MATHEMATICS, UCLA, 405 HILGARD AVE, LOS ANGELES CA 90095, USA

E-mail address: tao@math.ucla.edu