TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 00, Number 0, Pages 000–000
S 0002-9947(XX)0000-0

LONG STRINGS OF CONSECUTIVE COMPOSITE VALUES OF POLYNOMIALS

KEVIN FORD, MIKHAIL R. GABDULLIN

ABSTRACT. We show that for any polynomial $f : \mathbb{Z} \to \mathbb{Z}$ with positive leading coefficient and irreducible over $\mathbb{Q}$, if x is large enough then there is a string of $(\log x)(\log\log x)^{1/835}$ consecutive integers $n \in [1, x]$ for which $f(n)$ is composite. This improves the result in [6], which has the exponent of $\log\log x$ being constant depending on f which can be exponentially small in the degree of f.

1. INTRODUCTION

The first author, together with Konyagin, Maynard, Pomerance and Tao, showed in [6] that a general "sieved set" contains long gaps. More precisely, for each prime p consider a set I_p of residue classes modulo p, and call the collection of all sets I_p a *sieving system*. As in [6], assume the following regularity conditions:

(a) We have $|I_p| \leqslant p - 1$ for all p;
(b) $|I_p|$ is bounded; there is a $B \in \mathbb{N}$ with $|I_p| \leqslant B$ for all p;
(c) $|I_p|$ has average value 1, in the sense that

$$\prod_{p \leqslant x} \left(1 - \frac{|I_p|}{p}\right) \sim \frac{C_1}{\log x} \qquad (x \to \infty), \tag{1.1}$$

for some constant $C_1 > 0$.

(d) There is a $\rho > 0$, so that the density of primes with $|I_p| \geqslant 1$ equals ρ, that is,

$$\lim_{x\to\infty} \frac{|\{p \leqslant x : |I_p| \geqslant 1\}|}{x/\log x} = \rho.$$

Now define the *sieved set*

$$S_x := \mathbb{Z} \setminus \bigcup_{p \leqslant x} I_p,$$

so that S_x is a periodic set with period equal to the product of the primes $p \leqslant x$. The main theorem from [6] states that if (a)–(d) hold, then for any $\varepsilon > 0$ and large x, the set S_x contains a gap of size $x(\log x)^{C(\rho)-\varepsilon}$, where

$$C(\rho) := \sup\Big\{\delta > 0 : \frac{6 \cdot 10^{2\delta}}{\log(1/(2\delta))} < \rho\Big\}. \tag{1.2}$$

In particular, $C(\rho)$ decays exponentially in $1/\rho$. One of the principal applications of this result in [6] is to finding long strings of consecutive composite values of polynomial sequences. Consider a polynomial $f : \mathbb{Z} \to \mathbb{Z}$ of degree $d \geqslant 1$, with positive leading coefficient and irreducible over $\mathbb{Q}$. Let $I_p = \varnothing$ for $p \leqslant d$ and for p dividing the leading coefficient of f, and

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

otherwise. Note that the polynomial need not have integer coefficients. Indeed, by Pólya's theorem [11], f is integer valued at integers if and only if f has the form $f(x) = \sum_{j=0}^{d} a_j \binom{x}{j}$ with every $a_j \in \mathbb{Z}$. In particular,

Received by the editors October 23, 2024.
2020 *Mathematics Subject Classification.* Primary 11N35, 11N32, 11B05.
Keywords and phrases: gaps, prime values of polynomials, sieves.
KF was supported by National Science Foundation grant DMS-2301264.

$d!f(y) \in \mathbb{Z}[y]$ and thus the sieving system is well-defined. We call the collection of all sets I_p a *polynomial sieving system.*

By Lagrange's theorem, $|I_p| \leqslant d < p$ for all $p > d$ not dividing the leading coefficient of f, and hence (a) and (b) hold. Item (c) holds by Landau's Prime Ideal Theorem [9] (see also [4, pp. 35–36]), while (d) follows from the Chebotarev Density Theorem [3] (see also [8]), with $\rho = \rho(f)$ equal to $C/|G|$, where G is the Galois group of f, a subgroup of the symmetric group on d objects, and C is the number of elements of G having at least one fixed point. We have $\rho(f) \geqslant 1/d$ always, and sometimes $\rho(f) = 1/d$.

Examples. When $f(x) = x^2 + 1$, we have $|I_2| = 0$, $|I_p| = 2$ for all $p \equiv 1 \pmod 4$ and $|I_p| = 0$ for $p \equiv 3 \pmod 4$. Thus, $\rho(f) = 1/2$. Similarly, if $d = 2^k$ for a positive integer k and $f(x) = x^d + 1$, then $|I_p| = d$ for $p \equiv 1 \pmod{2^{k+1}}$ and $|I_p| = 0$ for all other odd primes p, hence $\rho(f) = 1/d = 1/2^k$.

The connection with sieved sets comes from the obvious relation

$$\{n \in \mathbb{N} : f(n) > x, f(n) \text{ prime}\} \subset S_x.$$

Now let X be large and set $x := \frac{1}{2}\log X$. The period of the set S_x is $X^{1/2+o(1)}$ by the Prime Number Theorem. Thus, writing g_x for the longest gap in S_x, the interval $[X/2, X]$ contains a gap in S_x of length g_x. For large X, $f(n) > x$ for all $n \in (X/2, X]$, and we conclude that

$$\max\{m : X/2 < n < n + m \leqslant X \text{ and } f(n), \dots, f(n+m) \text{ composite}\} \geqslant g_{(1/2)\log X}. \tag{1.3}$$

From the main theorem of [6] we see that the left side of (1.3) is $\geqslant (\log X)(\log\log X)^{C(\rho(f))-o(1)}$. In particular, we have [6, Corollary 1], which states that the left side of (1.3) is at least of size $\geqslant (\log X)(\log\log X)^{C(1/d)-o(1)}$. The exponent here decays exponentially in d (roughly $C(1/d) \approx e^{-6d}$).

It is still an open conjecture (of Bunyakovsky [2]) that there are infinitely many integers n for which $f(n)$ is prime. Moreover it is believed (see the conjecture of Bateman and Horn [1]) that the density of these prime values on $[X/2, X]$ is $\asymp_f 1/\log X$, and so the gaps above would be unusually large compared to the average gap of size $\asymp_f \log X$.

Our main result is a stronger lower bound on the length of strings of consecutive composite values of polynomials, with exponent of $\log\log x$ being independent of f.

Theorem 1. *Let f be as above. For all $\varepsilon > 0$ and large X, there are $n, n+m \in (X/2, X]$ with $m \geqslant (\log X)(\log\log X)^{C(1)-\varepsilon}$ and with $f(n), \dots, f(n+m)$ all composite.*

Numerically, $1/C(1) = 834.109\ldots$. In particular, when $f(x) = x$, our bound falls well short of the best known lower bound for the maximal gap between the primes below x from [5], which is

$$\gg (\log X)(\log\log X)\frac{\log\log\log\log X}{\log\log\log X}.$$

However, as noted in [6], the methods used to find large gaps between primes do not apply to gaps in more general sieved sets.

Our proof is based on the method developed in [6], but with one important difference. In [6], only one of the elements of I_q is utilized for large q (for those q with $|I_q| \geqslant 1$), whereas in this paper we utilize all of the set I_q. This introduces a number of complications, which we get around using special properties of polynomial sequences. Our methods do not apply for all of the sieved sets considered in [6], but they do generalize to sieved sets for which the sizes of the I_p have a limiting distribution and for which the difference sets $I_p - I_p := \{a - a' : a, a' \in I_p\}$, interpreted as subsets of $\mathbb{Z}$, do not have large overlap. To state our general theorem, we introduce further conditions, Hypotheses (e), (f) and (g) (here and throughout the paper, the symbols p and q always denote primes):

(e) For each $\nu \in \{1, \dots, B\}$, the density of those p with $|I_p| = \nu$ exists. That is, for some non-negative real numbers ρ_ν, $1 \leqslant \nu \leqslant B$, we have

$$\lim_{x\to\infty} \frac{\#\{p \leqslant x : |I_p| = \nu\}}{x/\log x} = \rho_\nu.$$

(f) For non-zero v, define

$$N(v) = \#\{p : v \bmod p \in I_p - I_p\}.$$

Then, for all $v \geqslant 1$, we have $N(v) \ll v^{0.49}$.

(g) There are positive constants c_1, c_2 such that the following holds. Let $u \geqslant 10$, and for each prime q with $|I_q| \geqslant 1$, let m_q be a nonzero integer with $m_q \bmod q \in (I_q - I_q)$. If $|w| \leqslant u$ and $k \geqslant 1$, then

$$\#\big\{q : |I_q| \geqslant 1, 0 < |m_q| \leqslant u, m_q + w \neq 0, N(m_q + w) \geqslant k\big\} \ll u(\log u)^{c_1} e^{-c_2 k}.$$

Hypothesis (e) is stronger than Hypothesis (d) and will replace it. Furthermore, (e) implies that the average of $|I_p|$, over $p \leqslant x$, is asymptotically $\rho_1 + 2\rho_2 + \cdots + B\rho_B$, which, by the weak average assumption (c), equals 1.

Theorem 2. *Consider any sieving system satisfying conditions (a)–(c) and (e)–(g) above. For any $\varepsilon > 0$ and large enough x, S_x has a gap of size at least $x(\log x)^{C(1)-\varepsilon}$.*

Clearly, Theorem 1 follows from Theorem 2, provided that we verify (e), (f) and (g) in the case of polynomial sieving systems. This verification is accomplished in the next section. The following sections are devoted to the proof of Theorem 2. As noted, the main new idea is to utilize all of the sets I_q for large q, which is encoded in a certain weight function; see (3.12) for specifics. Hypothesis (f) will be needed at the end of Section 4 and near the end of Section 6, while hypothesis (g) will be needed for the proof of the crucial Lemma 6.3.

2. Verifying the hypotheses of Theorem 2 for polynomial sieving systems

Item (e) is an immediate corollary of the Chebotarev density theorem. In fact, ρ_ν is precisely the proportion of elements of the Galois group of f which have exactly ν fixed points.

To verify (f) and (g), we introduce an auxiliary polynomial F which has roots which are the differences of the roots of f. By Pólya's theorem [11], f is integer valued at integers if and only if f has the form $f(x) = \sum_{j=0}^{d} a_j \binom{x}{j}$ with every $a_j \in \mathbb{Z}$. In particular, there exists a minimal positive integer $t|d!$ such that $tf \in \mathbb{Z}[x]$. Let also $r_1, ..., r_d$ be the complex roots of f. Writing

$$\tilde{f}(x) := tf(x) = cx^d + c_{d-1}x^{d-1} + \cdots + c_1 x + c_0 = c\prod_{i=1}^{d}(x - r_i),$$

where $c = c_d, c_0, \ldots, c_{d-1} \in \mathbb{Z}$, we define the polynomial

$$F(x) = c^{d^2+d} \prod_{1 \leqslant i,j \leqslant d} \Big(x - (r_i - r_j)\Big) = c^{d^2} \prod_{i=1}^{d} \tilde{f}(x + r_i), \tag{2.1}$$

so that $\deg F = d^2$. We will need the following properties of F.

Lemma 2.1. *The polynomial F obeys the following properties:*

(i) $F \in \mathbb{Z}[x]$;
(ii) *If $f(a) \equiv f(b) \equiv 0 \pmod q$ for some integers a and b, then $F(a-b) \equiv 0 \pmod q$;*
(iii) $F(l) \neq 0$ *for any $l \in \mathbb{Z}\backslash\{0\}$.*

Proof. Our proofs utilize the Fundamental Theorem of Symmetric Polynomials (FTSP) [10, p.20, Theorem (2.4)], which states that any symmetric polynomial $P \in \mathbb{Z}[\mathbf{u}]$, with $\mathbf{u} = (u_1, \ldots, u_k)$, is equal to a polynomial in $e_1(\mathbf{u}), \ldots, e_k(\mathbf{u})$ with integer coefficients, where e_j is the j-th elementary symmetric polynomial. In particular, by the definition of $\tilde{f}$, $c_j = (-1)^{j+d} c e_{d-j}(\mathbf{r})$ for each $0 \leqslant j \leqslant d-1$. Thus $e_j(cr_1, \ldots, cr_d) \in \mathbb{Z}$ for all $1 \leqslant j \leqslant d$.

We start with the first claim. By (2.1),

$$F(x) = \prod_{i=1}^{d} \Big[\sum_{j=0}^{d} c^{d-j}(cx + cr_i)^j c_j \Big],$$

whose coefficients are evidently symmetric polynomials in $(cr_1, \ldots, cr_d)$ with integer coefficients. By FTSP, $F \in \mathbb{Z}[x]$.

Now we turn to the second claim. Fix $a \in \mathbb{Z}$. We have

$$\tilde{f}(x) = (x - a)g(x) + \tilde{f}(a)$$

for some polynomial $g \in \mathbb{Z}[x]$ of degree $d-1$ depending on a, and therefore by (2.1),

$$F(x) = c^{d^2} \prod_{i=1}^{d} (x + r_i - a) g(x + r_i) + \tilde{f}(a) h(x),$$

where, by another application of FTSP, $h \in \mathbb{Z}[x]$. A similar argument shows that

$$c^{d^2-1} \prod_{i=1}^{d} g(x + r_i) \in \mathbb{Z}[x]$$

as well. Thus, for any $b \in \mathbb{Z}$,

$$F(a-b) = c^{d^2-1} (-1)^d \tilde{f}(b) \prod_{i=1}^{d} g(a - b + r_i) + \tilde{f}(a) h(a-b).$$

Therefore, $f(a) \equiv f(b) \equiv 0 \pmod q$ implies $\tilde{f}(a) \equiv \tilde{f}(b) \equiv 0 \pmod q$ and hence $q | F(a-b)$ for such q, as needed.

For the third claim, let us assume for a contradiction that $F(l) = 0$ for some integer $l \neq 0$. It means that there is $r_0 \in \mathbb{C}$ so that $f(r_0 + l) = f(r_0) = 0$. But then the polynomial

$$g(x) = f(x + l) - f(x)$$

also vanishes at the point $x = r_0$ and $g \not\equiv 0$ (otherwise $r_0 + kl$ would be zero of f for any k, so that $f \equiv 0$). Clearly, we also have $\deg g < \deg f$. But this is impossible, since f is irreducible and thus the minimal polynomial of r_0 is $\tilde{f}/c$. □

We also need a classic theorem of Erdős [7] about the average size of the number of divisors of polynomials. As is usual, $\tau(n)$ stands for the number of positive divisors of n.

Lemma 2.2. *For any irreducible polynomial $g \in \mathbb{Z}[x]$,*

$$\sum_{k \leqslant x} \tau(|g(k)|) \ll x \log x.$$

Let $\omega(n)$ denote the number of distinct prime factors of the nonzero integer n. If q is prime and $v \bmod q \in I_q - I_q$, Lemma 2.1 (ii) implies that $q | F(v)$. Lemma 2.1 (iii) implies that $F(v) \neq 0$ if $v \neq 0$. Since $|F(v)| \ll v^{d^2}$, $\omega(F(v)) \ll \log v + 1$ and this proves (f).

Now we verify (g). If $m_q \mod q \in I_q - I_q$ then $q | F(m_q)$ by Lemma 2.1 (ii), and if $m_q \neq 0$ then $F(m_q) \neq 0$ by Lemma 2.1 (iii). We let $m = m_q + w$. Thus, if m satisfies $0 < |m| \leqslant 2u$, $|w| \leqslant u$ and $m \neq w$, there are $O(\log u)$ primes dividing $F(m - w)$; that is, $O(\log u)$ primes q with $m_q + w = m$. Also, if $m \bmod p \in I_p - I_p$ implies that $p | F(m)$. Also F is the product of at most d^2 irreducible factors, say $F = F_1 \dots F_s$ with each F_i irreducible. Hence, if $N(m) \geqslant k$ then there are at least k distinct primes p dividing $F(m)$, and therefore, for some i, at least k/s distinct primes dividing $F_i(m)$. Hence, using Lemma 2.2,

$$\begin{aligned}
\#\{u^{1/2} < q \leqslant u : N(m_q + w) \geqslant k\} &\ll \#\{0 < |m| \leqslant 2u : N(m) \geqslant k\} \log u \\
&\leqslant (\log u) \sum_{i=1}^{s} \#\{0 < |m| \leqslant 2u : \omega(F_i(m)) \geqslant k/s\} \\
&\leqslant (\log u) \sum_{i=1}^{s} 2^{-k/s} \sum_{0 < |m| \leqslant 2u} \tau(F_i(m)) \\
&\ll 2^{-k/d^2} u \log^2 u.
\end{aligned}$$

In the last inequality, the implied constant in the $\ll$ depends on d. This proves (g), with $c_1 = 2$ and $c_2 = (\log 2)/d^2$, and completes the verification of the hypotheses of Theorem 2 for polynomial sieving systems.

3. NOTATION AND BASIC SETUP

We use notation similar to that of [6], with the most important change being the modification of the weight function $\boldsymbol{\lambda}$ (see (3.11) and (3.12) below). Throughout the proof, we will use positive parameters K, ξ, M which we describe below; one may think of them as being fixed for most of the time (in fact, it is only the end of Section 4 where the exact choice of them is important). The implied constants in O, $\ll$ and related order estimates may depend on these parameters, as well as on the constants B, C_1, ρ, c_1, c_2 in conditions (a)–(c) and (e)–(g), and the implied constants in the $\ll$ bounds in (f) and (g). We will rely on probabilistic methods; boldface symbols such as $\mathbf{S}$, $\boldsymbol{\lambda}$, $\mathbf{n}$, etc. will denote random variables (sets, functions, numbers, etc.), and the corresponding non-boldface symbols S, λ, n will denote deterministic counterparts of these variables.

For a fixed $\delta \in (1/10^3, C(1))$, we define

$$y = \lceil x(\log x)^\delta \rceil \tag{3.1}$$

and

$$z = \frac{y \log\log x}{(\log x)^{1/2}}. \tag{3.2}$$

As in [6], our goal is to find a number b so that $S_x + b$ has no elements in $[1, y]$, which will show that S_x has a gap of size at least y. This is accomplished in three stages:

(1) (Uniform random stage) First, we choose b modulo $P(z)$ 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 primes in $(z, x/2]$ randomly, but dependent on the choice of b modulo p for $p \leqslant z$. 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 $\{b_q + a + kq : k \in \mathbb{Z}, a \bmod q \in I_q\} \cap [1, y]$ knocks out nearly as many elements of the random set $(S_z + b) \cap [1, y]$ as possible. Unlike the argument in [6], we make use of all of the elements of I_q in this stage. This is the source of our improved theorems.
(3) (Clean up stage) Thirdly, we choose b modulo primes $q \in (x/2, x]$ to ensure that the remaining elements $m \in (S_{x/2} + b) \cap [1, y]$ do not lie in $(S_x + b) \cap [1, y]$ by matching a unique prime $q = q(m)$ with $|I_q| \geqslant 1$ to each element m and setting $b \equiv m \pmod q$. Here we do use only a single element of I_q, whereas using all of I_q would not improve our theorem at all.

To handle the Greedy stage (2), we divide the primes in $(z, x/2]$ into subsets, where the primes in each subset are about the same size and with $|I_q|$ is constant. Primes with rare values of $|I_q|$ will play an insignificant role in our arguments, thus we define

$$\mathcal{N} = \{1 \leqslant \nu \leqslant B : \rho_\nu > 0\},$$

so that, by the remarks following Hypothesis (g),

$$\sum_{\nu \in \mathcal{N}} \nu \rho_\nu = 1. \tag{3.3}$$

For example, for the polynomial sieving system with $f(x) = x^2 + 1$, we have $\mathcal{N} = \{2\}$ since $\rho_1 = 0$.

Let $\xi > 1$ be a real number (which we will finally choose to be close to 1), and define the set of scales

$$\mathfrak{H} = \left\{ H \in \{1, \xi, \xi^2, ...\} : \frac{2y}{x} \leqslant H \leqslant \frac{y}{\xi z} \right\}$$

so that

$$2(\log x)^\delta \leqslant H \leqslant \frac{y}{z} = \frac{(\log x)^{1/2}}{\log\log x} \qquad (H \in \mathfrak{H}). \tag{3.4}$$

For each $H \in \mathfrak{H}$ and $\nu \in \mathcal{N}$, let

$$\mathcal{Q}_{H,\nu} = \left\{ q \in \left(\frac{y}{\xi H}, \frac{y}{H} \right] : |I_q| = \nu \right\}.$$

Hypothesis (e) implies that for each fixed H, ν we have the asymptotic

$$|\mathcal{Q}_{H,\nu}| \sim \rho_\nu (1 - 1/\xi) \frac{y}{H \log x} \qquad (x \to \infty). \tag{3.5}$$

Note that if we denote by ρ the density of primes p with $|I_p| \geqslant 1$, then by (e),

$$\rho = \lim_{x\to\infty} \frac{\#\{p \leqslant x : |I_p| \geqslant 1\}}{x/\log x} = \sum_{\nu\in\mathcal{N}} \rho_\nu. \tag{3.6}$$

Let also

$$\mathcal{Q}_H := \left\{ q \in \Big(\frac{y}{\xi H}, \frac{y}{H}\Big] : |I_q| \in \mathcal{N} \right\} = \bigcup_{\nu\in\mathcal{N}} \mathcal{Q}_{H,\nu}$$

and, for $\nu \in \mathcal{N}$,

$$\mathcal{Q}^\nu := \bigcup_{H\in\mathfrak{H}} Q_{H,\nu},$$

so that

$$\mathcal{Q} := \bigcup_{H\in\mathfrak{H}} \mathcal{Q}_H = \bigcup_{\nu\in\mathcal{N}} \mathcal{Q}^\nu.$$

We note that for all $q \in \mathcal{Q}$, $z < q \leqslant x/2$. Further, for each $q \in \mathcal{Q}$, let H_q be the unique H such that $q \in \mathcal{Q}_H$, which is equivalent to

$$\frac{y}{\xi H_q} < q \leqslant \frac{y}{H_q}.$$

Let also M be a number with

$$6 < M \leqslant 7,$$

which we will eventually take to be very close to 6. We use the notation

$$S_z = \mathbb{Z}\backslash \bigcup_{p\leqslant z} I_p,$$

and

$$S_{z,x} = \mathbb{Z}\backslash \bigcup_{z<p\leqslant x} I_p,$$

and also adopt the abbreviations

$$P = P(z) = \prod_{p\leqslant z} p, \quad \sigma = \sigma(z) := \prod_{p\leqslant z} \left(1 - \frac{|I_p|}{p}\right), \quad \mathbf{S} = S_z + \mathbf{b}, \tag{3.7}$$

where $\mathbf{b}$ is a residue class chosen uniformly at random from $\mathbb{Z}/P(z)\mathbb{Z}$; so, $\mathbf{S}$ is a random shift of S_z. For a fixed $H \in \mathfrak{H}$, we also define

$$P_1 = \prod_{p\leqslant H^M} p, \quad \sigma_1 = \sigma(H^M), \quad \mathbf{b}_1 \equiv \mathbf{b} \pmod{P_1}, \quad \mathbf{S}_1 = S_{H^M} + \mathbf{b}_1, \tag{3.8}$$

and

$$P_2 = \prod_{H^M<p\leqslant z} p, \quad \sigma_2 = \frac{\sigma(z)}{\sigma(H^M)}, \quad \mathbf{b}_2 \equiv \mathbf{b} \pmod{P_2}, \quad \mathbf{S}_2 = S_{H^M,z} + \mathbf{b}_2, . \tag{3.9}$$

Obviously, for each $H \in \mathfrak{H}$,

$$P = P_1P_2, \quad \sigma = \sigma_1\sigma_2, \quad \mathbf{S} = \mathbf{S}_1 \cap \mathbf{S}_2. \tag{3.10}$$

Note that all the quantities defined in (3.8) and (3.9) depend on H and M; however, we will not indicate this dependence for brevity (the values of H and M will always be clear from context).

Finally, let $\nu_q = |I_q|$ for primes q. For primes $q \in \mathcal{Q}^\nu$, let

$$I_q = \{a_{1,q} \bmod q, ..., a_{\nu,q} \bmod q\} \qquad (a_{i,q} \in [1, q] \cap \mathbb{Z},\ 1 \leqslant i \leqslant \nu).$$

We set

$$\mathbf{AP}(J; q, n) = \left(\bigsqcup_{i=1}^{\nu_q} \{n + a_{i,q} + hq : 1 \leqslant h \leqslant J\} \right) \cap \mathbf{S}_1, \tag{3.11}$$

this being a significant departure from [6]. Here $\mathbf{AP}(J;q,n)$ is a portion of ν_q residue classes modulo q. Also define

$$\boldsymbol{\lambda}(H;q,n) = \frac{\mathbb{1}_{\mathbf{AP}(KH;q,n)\subset \mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}(KH;q,n)|}}, \tag{3.12}$$

where K is a positive integer which will be chosen large enough, and $\mathbb{1}_X$ is the indicator function of a statement X. So, for each $q \in \mathcal{Q}$, the weights $\boldsymbol{\lambda}(H;q,n)$ are random functions which depend on $\mathbf{b}$. Heuristically, $\boldsymbol{\lambda}(H;q,n)$ has mean approximately 1, since the probability that a given set Y lies in $\mathbf{S}_2$ is about $\sigma_2^{|Y|}$; see Lemma 6.1 below for a precise statement.

3.1. **General notational conventions.** The notation $f = O(g)$ and $f \ll g$ mean that f/g is bounded. The notation $f = O_{\leqslant}(g)$ means that $|f|/g \leqslant 1$. The notation $o(1)$ stands for a function tending to zero as $x \to \infty$, at a rate which may depend only on the parameters ξ, K, M and δ, which we consider to be fixed. The notation $f \sim g$ means $f = g + o(1)$. As is usual, $\omega(n)$ is the number of distinct prime factors of n.

4. Reduction to concentration of $\lambda(H;q,n)$

In this section we deduce Theorem 2 from the following statement. Recall that $\mathbf{S} = S_z + \mathbf{b}$ with $\mathbf{b}$ chosen uniformly at random from $\mathbb{Z}/P(z)\mathbb{Z}$.

Theorem 3. *Let $\delta < C(1)$, $M > 6$, $\xi > 1$, $K > 0$, $0 < \varepsilon < \frac{1}{7}(M-6)$, and assume that x is large enough depending on δ, M, ξ, K, and ε. Then there exist a choice of $\mathbf{b} \pmod{P(z)}$ and subsets $\mathcal{R}^\nu \subseteq \mathcal{Q}^\nu$ for $\nu \in \mathcal{N}$ so that:*

(i) *one has*

$$|S \cap [1,y]| \leqslant 2\sigma y; \tag{4.1}$$

(ii) *for all $q \in \mathcal{R} = \bigcup_{\nu\in\mathcal{N}} \mathcal{R}^\nu$ one has*

$$\sum_{-(K+1)y<n\leqslant y} \lambda(H_q;q,n) = \left(1 + O_{\leqslant}\left(\frac{1}{(\log x)^{\delta(1+\varepsilon)}}\right)\right)(K+2)y; \tag{4.2}$$

(iii) *for each $\nu \in \mathcal{N}$ and $i \in \{1,...,\nu\}$, all but at most $\frac{\rho x}{8B^2\log x}$ elements n of $S\cap[1,y]$ obey*

$$\sum_{q\in\mathcal{R}^\nu}\sum_{h\leqslant KH_q} \lambda(H_q;q,n-a_{i,q}-qh) = \left(C_{2,\nu} + O_{\leqslant}\left(\frac{2}{(\log x)^{\delta(1+\varepsilon)}}\right)\right)(K+2)y, \tag{4.3}$$

where $C_{2,\nu}$ is independent of n and i with

$$C_{2,\nu} \sim \frac{K\rho_\nu}{(K+2)M}\frac{1-1/\xi}{\log\xi}\log\left(1/(2\delta)\right) \quad (x\to\infty). \tag{4.4}$$

We now commence with the deduction of Theorem 2 from Theorem 3. Let V be the set of elements n of $S\cap[1,y]$ for which (4.3) holds for all $\nu\in\mathcal{N}$ and $i = 1,...,\nu$. For each $q\in\mathcal{R}$, we define the random integer $\mathbf{n}_q$ by the distribution

$$\mathbb{P}(\mathbf{n}_q = n) = \frac{\lambda(H_q;q,n)}{\sum_{-(K+1)y<n'\leqslant y}\lambda(H_q;q,n')} \qquad (-(K+1)y < n \leqslant y).$$

For $q\in\mathcal{R}$, define the random set

$$\mathbf{e}_q = \bigsqcup_{i=1}^{\nu_q} \mathbf{e}_{i,q},$$

where

$$\mathbf{e}_{i,q} = V \cap \{\mathbf{n}_q + a_{i,q} + hq : 1 \leqslant h \leqslant KH_q\}, \quad i = 1,...,\nu_q. \tag{4.5}$$

Note that $\mathbf{e}_{i,q}$ and $\mathbf{e}_{j,q}$ are disjoint for $i \neq j$, since all $a_{i,q}$ are distinct modulo q. To be able to make a clean-up stage, we need to find a choice n_q of $\mathbf{n}_q$ (which corresponds to a choice e_q of the random sets $\mathbf{e}_q$) for each $q \in \mathcal{R}$, so that for b satisfying $b \equiv n_q \pmod q$ for $q \in \mathcal{R}$, the estimate

$$\Big|(S_{x/2}+b)\cap[1,y]\Big| \leqslant \frac{\rho x}{4\log x} \tag{4.6}$$

holds. Then by (3.6), the number of primes $p \in (x/2, x]$ with $|I_p| \geqslant 1$ is $\sim \frac{\rho x}{2\log x}$ for x large, which guarantees that clean-up stage is possible. To be precise, we may match each element $m \in (S_{x/2}+b)\cap[1,y]$ with a unique prime $p \in (x/2, x]$ with $|I_p| \geqslant 1$, and choose $b \equiv m - a_{1,p} \pmod p$ for each such pair. Then $(S_x + b)\cap[1,y] = \varnothing$, as desired.

Since $S_{x/2} + b$ avoids the $\nu_q = |I_q|$ residues $n_q + a_{1,q}, ..., n_q + a_{\nu_q,q}$ modulo q, we have, by the definition of the set V and (4.5),

$$\begin{aligned}\Big|(S_{x/2}+b)\cap[1,y]\Big| &\leqslant \Big|\big((S_{x/2}+b)\cap[1,y]\big)\backslash V\Big| + \Big|V\backslash\bigcup_{q\in\mathcal{R}} e_q\Big| \\ &\leqslant \sum_{\nu\in\mathcal{N}}\sum_{i=1}^{\nu}\frac{\rho x}{8B^2\log x} + \Big|V\backslash\bigcup_{q\in\mathcal{R}} e_q\Big| \\ &\leqslant \frac{\rho x}{4\log x},\end{aligned}$$

thus verifying (4.6), provided that

$$\Big|V\backslash\bigcup_{q\in\mathcal{R}} e_q\Big| \leqslant \frac{\rho x}{8\log x}. \tag{4.7}$$

To show (4.7), we need the following hypergraph covering lemma, which is Lemma 3.1 of [6].

Lemma 4.1 (Hypergraph covering lemma)**.** *Suppose that $0 < \delta \leqslant 1/2$ and $K_0 \geqslant 1$, and let $y \geqslant y_0(\delta, K_0)$ with $y_0(\delta, K_0)$ sufficiently large, and let V be a finite set with $|V| \leqslant y$. Let $1 \leqslant s \leqslant y$, and suppose that $\mathbf{e}_1, ..., \mathbf{e}_s$ are random subsets of V satisfying the following:*

$$|\mathbf{e}_i| \leqslant \frac{K_0(\log y)^{1/2}}{\log\log y} \quad (1 \leqslant i \leqslant s), \tag{4.8}$$

$$\mathbb{P}(v \in \mathbf{e}_i) \leqslant y^{-1/2-1/100} \quad (v \in V, 1 \leqslant i \leqslant s), \tag{4.9}$$

$$\sum_{i=1}^{s}\mathbb{P}(v, v' \in \mathbf{e}_i) \leqslant y^{-1/2} \quad (v, v' \in V, v \neq v'), \tag{4.10}$$

$$\Big|\sum_{i=1}^{s}\mathbb{P}(v \in \mathbf{e}_i) - C_2\Big| \leqslant \eta \quad (v \in V), \tag{4.11}$$

where C_2 and η satisfy

$$10^{2\delta} \leqslant C_2 \leqslant 100, \quad \eta \geqslant \frac{1}{(\log y)^{\delta}\log\log y}. \tag{4.12}$$

Then there are subsets e_i of V, $1 \leqslant i \leqslant s$, with e_i being in the support of $\mathbf{e}_i$ for every i, and such that

$$\Big|V\backslash\bigcup_{i=1}^{s} e_i\Big| \leqslant C_3\eta|V|, \tag{4.13}$$

where C_3 is an absolute constant.

We apply this lemma with $s = |\mathcal{R}|$, $\{\mathbf{e}_i : i = 1, ..., s\} = \{\mathbf{e}_q : q \in \mathcal{R}\}$, $K_0 = BK$, and

$$\eta = \frac{\rho}{20C_1C_3(\log x)^\delta}.$$

By (1.1) and (3.2), we have $\sigma = \sigma(z) \sim C_1/\log z \sim C_1/\log x$. The conclusion of the lemma (with the bound (4.1)) implies that there is a choice of sets e_q with

$$\left|V \setminus \bigcup_{q\in\mathcal{R}} e_q\right| \leqslant C_3\eta|V| = \frac{\rho|V|}{20C_1(\log x)^\delta} \leqslant \frac{\rho\sigma y}{10C_1(\log x)^\delta} \leqslant \frac{\rho x}{8\log x},$$

which is enough for (4.7), so that we are left to verify the conditions of the lemma. First of all, by (3.4), we have

$$|\mathbf{e}_q| = \sum_{i=1}^{\nu_q} |\mathbf{e}_{i,q}| \leqslant \nu_q K H_q \leqslant \frac{BKy}{z} \leqslant \frac{BK(\log x)^{1/2}}{\log\log x} \leqslant \frac{BK(\log y)^{1/2}}{\log\log y},$$

which gives us (4.8). For each $n \in V$ and $q \in \mathcal{R}$, (4.2) together with the trivial bound $\lambda(H_q; q, n) \leqslant \sigma_2^{-KBH_q} \leqslant y^{o(1)}$ (from (3.4), (3.11) and (3.12)) gives us

$$\begin{aligned}\mathbb{P}(n \in \mathbf{e}_q) &= \sum_{1\leqslant h\leqslant KH_q} \sum_{i=1}^{\nu_q} \mathbb{P}(\mathbf{n}_q = n - a_{i,q} - hq) \\ &\ll \frac{1}{y} \sum_{1\leqslant h\leqslant KH_q} \sum_{i=1}^{\nu_q} \lambda(n - a_{i,q} - hq) \\ &\ll y^{-0.999},\end{aligned} \tag{4.14}$$

which verifies (4.9). Now we turn to (4.11). From (4.2), (4.3), and (4.4), followed by an application of (3.3), we obtain

$$\sum_{q\in\mathcal{R}} \mathbb{P}(n \in \mathbf{e}_q) = \sum_{\nu\in\mathcal{N}} \sum_{q\in\mathcal{R}^\nu} \mathbb{P}(n \in \mathbf{e}_q) = \sum_{\nu\in\mathcal{N}} \nu C_{2,\nu} + O((\log x)^{-\delta(1+\varepsilon)}) =$$

$$= \sum_{\nu\in\mathcal{N}} \nu\rho_\nu \frac{K(1-1/\xi)}{(K+2)M\log\xi} \log(1/(2\delta)) + o(1) = C_2 + o(1),$$

where we define

$$C_2 = \frac{K(1-1/\xi)}{(K+2)M\log\xi} \log(1/(2\delta)).$$

Recalling that $\delta < C(1)$ together with the definition (1.2) of $C(1)$, we see that C_2 is at least $10^{2\delta}$ provided that $M - 6$ and $\xi - 1$ are sufficiently small in terms of δ, K is sufficiently large in terms of δ, $0 < \varepsilon < \frac{1}{7}(M-6)$, and x is large enough depending on $\delta, M, \xi, K, \varepsilon$. Also $C_2 \leqslant 100$ due to $\delta \geqslant 10^{-3}$. Thus, (4.11) follows.

It remains to check that (4.10) holds. We take any distinct $v, v' \in V$ and see that

$$\sum_{q\in\mathcal{R}} \mathbb{P}(v, v' \in \mathbf{e}_q) \leqslant \sum_{q\in\mathcal{R}} \left(\sum_{i=1}^{\nu_q} \mathbb{P}(v, v' \in \mathbf{e}_{i,q}) + \sum_{\substack{1\leqslant i,j\leqslant\nu_q \\ i\neq j}} \mathbb{P}(v \in \mathbf{e}_{i,q}, v' \in \mathbf{e}_{j,q}) \right).$$

If both v, v' both belong to some $\mathbf{e}_{i,q}$, then q divides $v - v'$; but $0 < |v - v'| \leqslant Ky$ and q is a prime greater than $z > y^{3/4}$, hence there is at most one such q. Further, if $v \in \mathbf{e}_{i,q}$ and $v' \in \mathbf{e}_{j,q}$ for some q and $i \neq j$, then $v - v' \equiv a_{i,q} - a_{j,q} \pmod q$ and hence $v - v' \bmod q \in I_q - I_q$. By hypothesis (f) and the bound $|v - v'| \leqslant Ky$, the number of such q is $\ll y^{0.49}$. Thus, by (4.14),

$$\sum_{q\in\mathcal{R}} \mathbb{P}(v, v' \in \mathbf{e}_q) \ll y^{0.49} \cdot \max_{v,i,q} \mathbb{P}(v \in \mathbf{e}_{i,q}) \ll y^{-0.509},$$

which gives (4.10).

Thus, we verified the conditions of Lemma 4.1, and (4.7) follows. This completes the proof of Theorem 2 assuming Theorem 3.

5. Concentration of $\lambda(H;q,n)$

In this section we reduce Theorem 3 to the following assertion.

Theorem 4. *Let $M>2$, $K>0$, and $\xi>1$. Then*
(i) *One has*

$$\mathbb{E}\Big|\mathbf{S}\cap[1,y]\Big| = \sigma y; \quad \mathbb{E}\Big|\mathbf{S}\cap[1,y]\Big|^2 = \left(1+O\left(\frac{1}{\log y}\right)\right)(\sigma y)^2; \tag{5.1}$$

(ii) *For every $H\in\mathfrak{H}$, every $\nu\in\mathcal{N}$ and $j\in\{0,1,2\}$,*

$$\mathbb{E}\sum_{q\in\mathcal{Q}_{H,\nu}}\left(\sum_{-(K+1)y<n\leqslant y}\boldsymbol{\lambda}(H;q,n)\right)^j = \left(1+O\left(\frac{\log H}{H^{M-2}}\right)\right)\big((K+2)y\big)^j|\mathcal{Q}_{H,\nu}|; \tag{5.2}$$

(iii) *For every $H\in\mathfrak{H}$, every $\nu\in\mathcal{N}$, $i\in\{1,...,\nu\}$, and $j\in\{0,1,2\}$,*

$$\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\left(\sum_{q\in\mathcal{Q}_{H,\nu}}\sum_{h\leqslant KH}\boldsymbol{\lambda}(H;q,n-a_{i,q}-qh)\right)^j = \left(1+O\left(\frac{\log H}{H^{M-2}}\right)\right)\left(\frac{|\mathcal{Q}_{H,\nu}|\cdot\lfloor KH\rfloor}{\sigma_2}\right)^j\sigma y. \tag{5.3}$$

Deduction of Theorem 3 from Theorem 4. Fix $\delta<C(1)$, $M>6$, $\xi>1$, $K>0$, and also $0<\varepsilon<\frac{1}{7}(M-6)$. From Theorem 4 (i) we have

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

Hence by Chebyshev's inequality, we see that

$$\mathbb{P}(|\mathbf{S}\cap[1,y]|\leqslant 2\sigma y) = 1-O(1/\log x), \tag{5.4}$$

showing that (4.1) in Theorem 3 holds with probability $1-o(1)$.

Now we work on parts (ii) and (iii) of Theorem 3. For each $H\in\mathfrak{H}$ and $\nu\in\mathcal{N}$, we have from (5.2)

$$\mathbb{E}\sum_{q\in\mathcal{Q}_{H,\nu}}\left(\sum_{-(K+1)y<n\leqslant y}\boldsymbol{\lambda}(H;q,n)-(K+2)y\right)^2 \ll \frac{y^2|\mathcal{Q}_{H,\nu}|}{H^{M-2-\varepsilon}}. \tag{5.5}$$

Now let $\boldsymbol{\mathcal{R}}_{H,\nu}$ be the (random) set of $q\in\mathcal{Q}_{H,\nu}$ for which

$$\left|\sum_{-(K+1)y<n\leqslant y}\boldsymbol{\lambda}(H;q,n)-(K+2)y\right| \leqslant \frac{y}{H^{1+\varepsilon}}. \tag{5.6}$$

By estimating the left-hand side of (5.5) from below by the sum over $q\in\mathcal{Q}_{H,\nu}\backslash\boldsymbol{\mathcal{R}}_{H,\nu}$, we find that

$$\mathbb{E}|\mathcal{Q}_{H,\nu}\backslash\boldsymbol{\mathcal{R}}_{H,\nu}| \ll \frac{|\mathcal{Q}_{H,\nu}|}{H^{M-4-3\varepsilon}}. \tag{5.7}$$

We let

$$\boldsymbol{\mathcal{R}}^{\nu} = \bigcup_{H\in\mathfrak{H}}\boldsymbol{\mathcal{R}}_{H,\nu},$$

and then Theorem 3 (ii) follows from the lower bound on H given in (3.4).

We now turn to the condition (iii) of Theorem 3. Similarly to (5.6), for each $H \in \mathfrak{H}$, $\nu \in \mathcal{N}$, and $i \in \{1, ..., \nu\}$, from (5.3) we have

$$\mathbb{E} \sum_{n \in \mathbf{S} \cap [1,y]} \left(\sum_{q \in \mathcal{Q}_{H,\nu}} \sum_{h \leqslant KH} \boldsymbol{\lambda}(H; q, n - a_{i,q} - qh) - \frac{|\mathcal{Q}_{H,\nu}| \cdot \lfloor KH \rfloor}{\sigma_2} \right)^2 \ll \frac{1}{H^{M-2-\varepsilon}} \left(\frac{|\mathcal{Q}_{H,\nu}| \cdot \lfloor KH \rfloor}{\sigma_2} \right)^2 \sigma y. \tag{5.8}$$

Let $\boldsymbol{\mathcal{E}}_{H,\nu,i}$ be the set of $n \in \mathbf{S} \cap [1, y]$ such that

$$\left| \sum_{q \in \mathcal{Q}_{H,\nu}} \sum_{h \leqslant KH} \boldsymbol{\lambda}(H; q, n - a_{i,q} - qh) - \frac{|\mathcal{Q}_{H,\nu}| \cdot \lfloor KH \rfloor}{\sigma_2} \right| \geqslant \frac{|\mathcal{Q}_{H,\nu}| \cdot \lfloor KH \rfloor}{\sigma_2 H^{1+\varepsilon}}. \tag{5.9}$$

Then, since $M > 6$ and $\varepsilon < (M-6)/7$, (5.8) implies that

$$\mathbb{E}|\boldsymbol{\mathcal{E}}_{H,\nu,i}| \ll \frac{\sigma y}{H^{M-4-3\varepsilon}} \ll \frac{\sigma y}{H^2},$$

and, hence, $|\boldsymbol{\mathcal{E}}_{H,\nu,i}| \leqslant \sigma y / H^{1+\varepsilon}$ with probability $1 - O(H^{-1+\varepsilon})$.

Now we estimate the contribution from "bad" primes $q \in \mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}$. For any $h \leqslant KH$, we get from Cauchy-Schwarz inequality for vector functions

$$\mathbb{E} \sum_{q \in \mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}} \sum_{n \in \mathbf{S} \cap [1,y]} \boldsymbol{\lambda}(H; q, n - a_{i,q} - qh) \leqslant (\mathbb{E}|\mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}|)^{1/2} \left(\mathbb{E} \sum_{q \in \mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}} \left| \sum_{-(K+1)y < n \leqslant y} \boldsymbol{\lambda}(H; q, n) \right|^2 \right)^{1/2},$$

where we extended the range of summation of $\boldsymbol{\lambda}(H; q, \cdot)$ to the larger interval $(-(K+1)y, y]$ (note that $a_{i,q} + qh \leqslant q + Ky \leqslant (K+1)y$ and the weights $\boldsymbol{\lambda}(H; q, \cdot)$ are non-negative). Further, by the triangle inequality, (5.5) and (5.7),

$$\mathbb{E} \sum_{q \in \mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}} \left| \sum_{-(K+1)y < n \leqslant y} \boldsymbol{\lambda}(H; q, n) \right|^2 \leqslant 2\mathbb{E} \sum_{q \in \mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}} \left(\left| \sum_{-(K+1)y < n \leqslant y} \boldsymbol{\lambda}(H; q, n) - (K+2)y \right|^2 + (K+2)^2 y^2 \right) \ll \frac{y^2 |\mathcal{Q}_{H,\nu}|}{H^{M-4-3\varepsilon}}.$$

Combining two latter estimates, using (5.7) again, and summing over all $h \leqslant KH$, we get

$$\mathbb{E} \sum_{n \in \mathbf{S} \cap [1,y]} \sum_{q \in \mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}} \sum_{h \leqslant KH} \boldsymbol{\lambda}(H; q, n - a_{i,q} - qh) \ll \frac{y |\mathcal{Q}_{H,\nu}|}{H^{M-5-3\varepsilon}}.$$

Let $\boldsymbol{\mathcal{F}}_{H,\nu,i}$ be the set of $n \in \mathbf{S} \cap [1, y]$ such that

$$\sum_{q \in \mathcal{Q}_{H,\nu} \backslash \boldsymbol{\mathcal{R}}_{H,\nu}} \sum_{h \leqslant KH} \boldsymbol{\lambda}(H; q, n - a_{i,q} - qh) \geqslant \frac{|\mathcal{Q}_{H,\nu}| \cdot \lfloor KH \rfloor}{\sigma_2 H^{1+\varepsilon}}. \tag{5.10}$$

Then

$$\mathbb{E}|\boldsymbol{\mathcal{F}}_{H,\nu,i}| \ll \frac{\sigma_2 y}{H^{M-5-4\varepsilon}} \ll \frac{\sigma y \log H}{H^{M-5-4\varepsilon}},$$

and, by Markov's inequality,

$$|\boldsymbol{\mathcal{F}}_{H,\nu,i}| \leqslant \frac{\sigma y}{H^{1+\varepsilon}}$$

with probability $1 - O(H^{-(M-6-6\varepsilon)})$. Since $\varepsilon < (M-6)/7$, we have $M - 6 - 6\varepsilon > \varepsilon$, and the last probability becomes $1 - O(H^{-\varepsilon})$.

Since $\sum_{H\in\mathfrak{H}} H^{-\varepsilon} \ll (\log x)^{-\delta\varepsilon}$, with probability $1-o(1)$ we have that for all $H\in\mathfrak{H}$, $\nu\in\mathcal{N}$, and $i\in\{1,...,\nu\}$ that both sets $\boldsymbol{\mathcal{E}}_{H,\nu,i}, \boldsymbol{\mathcal{F}}_{H,\nu,i}$ have size at most $(\sigma y)H^{-1-\varepsilon}$.

Now we make a choice of $b \pmod{P(z)}$. We consider the event that $|\mathbf{S}\cap[1,y]|\leqslant 2\sigma y$ and that for each H,ν,i, the sets $\boldsymbol{\mathcal{E}}_{H,\nu,i}, \boldsymbol{\mathcal{F}}_{H,\nu,i}$ have size at most $(\sigma y)H^{-1-\varepsilon}$. By the above discussion, this event holds with probability at least $1-o(1)$ as $x\to\infty$ (so this probability is at least, say, $1/2$ whenever x is large enough depending on δ, M, ξ, K, and ε). From now, we fix a $b \bmod P(z)$ such that it is so, and thus all of our random sets and weights become deterministic. With this choice of b we verify condition (iii) in Theorem 3.

For fixed $\nu\in\mathcal{N}$ and $i\in\{1,...,\nu\}$, we set

$$\mathcal{M}_{\nu,i} = \Big(S\cap[1,y]\Big)\setminus \bigcup_{H\in\mathfrak{H}} \left(\mathcal{E}_{H,\nu,i}\cup\mathcal{F}_{H,\nu,i}\right).$$

Now we verify (4.3) with given ν and i for $n\in\mathcal{M}_{\nu,i}$. By (3.4), $\sum_{H\in\mathfrak{H}} H^{-1-\varepsilon} \ll (\log x)^{-(1+\varepsilon)\delta}$, and so the number of exceptional elements satisfies

$$\left|\bigcup_{H\in\mathfrak{H}} (\mathcal{E}_{H,\nu,i}\cup\mathcal{F}_{H,\nu,i})\right| \ll \frac{\sigma y}{(\log x)^{(1+\varepsilon)\delta}},$$

which, by (3.2), is smaller than $\frac{\rho x}{8B^2\log x}$ for large x. We fix arbitrary $n\in\mathcal{M}_{\nu,i}$. For such n, the inequalities (5.9) and (5.10) both fail, and therefore for each $H\in\mathfrak{H}$,

$$\sum_{q\in\mathcal{R}_{H,\nu}} \sum_{h\leqslant KH} \lambda(H;q,n-a_{i,q}-qh) = \left(1+O_{\leqslant}\left(\frac{2}{(\log x)^{(1+\varepsilon)\delta}}\right)\right)\frac{|\mathcal{Q}_{H,\nu}|\cdot\lfloor KH\rfloor}{\sigma_2}.$$

Summing over all $H\in\mathfrak{H}$, we have

$$\sum_{q\in\mathcal{R}^{\nu}} \sum_{h\leqslant KH_q} \lambda(H_q;q,n-a_{i,q}-qh) = \left(1+O_{\leqslant}\left(\frac{2}{(\log x)^{(1+\varepsilon)\delta}}\right)\right) C_{2,\nu}(K+2)y$$

with (recall that σ_2 depends on H)

$$C_{2,\nu} = \frac{1}{(K+2)y}\sum_{H\in\mathfrak{H}} \frac{|\mathcal{Q}_{H,\nu}|\cdot\lfloor KH\rfloor}{\sigma_2}.$$

Note that $C_{2,\nu}$ depends on x, K, M, ξ, and δ, but not on n. Since

$$\lfloor KH\rfloor = KH\big(1+O(1/H)\big) = KH\big(1+O(\log x)^{-\delta}\big)$$

and

$$\sigma_2^{-1} = \prod_{H^M<p\leqslant z} (1-|I_p|/p)^{-1} \sim \frac{\log z}{M\log H},$$

we get, using (3.5),

$$C_{2,\nu} \sim \frac{K}{(K+2)y}\cdot\rho_\nu(1-1/\xi)\sum_{H\in\mathfrak{H}} \frac{y/H}{\log x}\cdot\frac{H\log z}{M\log H} \sim \frac{\rho_\nu K(1-1/\xi)}{M(K+2)}\sum_{H\in\mathfrak{H}}\frac{1}{\log H},$$

as $x\to\infty$. Recalling the definition of $\mathfrak{H}$, we see that

$$C_{2,\nu} \sim \frac{\rho_\nu K(1-1/\xi)}{M(K+2)\log\xi}\sum_j \frac{1}{j},$$

where j runs over the interval

$$\frac{\delta\log\log x}{\log\xi} \leqslant j \leqslant \frac{(1/2+o(1))\log\log x}{\log\xi}.$$

We thus obtain

$$C_{2,\nu} \sim \frac{\rho_\nu K(1-1/\xi)}{M(K+2)\log\xi}\log(1/(2\delta)), \quad x\to\infty,$$

and the claim (4.3) follows. □

It remains to establish Theorem 4. This is the aim of the last section of the paper.

6. Computing correlations

In this section we prove Theorem 4. The claim (i) is exactly (5.1) and (5.2) of Theorem 3 from [6]. To verify the claims (ii) and (iii), we must rework the argument from [6] using our new weight function $\boldsymbol{\lambda}$ from (3.12). For $H \in \mathfrak{H}$, let $\mathcal{D}_H$ be the collection of square-free numbers D, all of whose prime divisors lie in $(H^M, z]$. For each $D \in \mathcal{D}_H$, let $I_D \subset \mathbb{Z}/D\mathbb{Z}$ be defined as $I_D = \bigcap_{p|D} I_p$. Further, for $A > 0$, let

$$E_A(m;H) = \big(\mathbb{1}_{m\neq 0}\big) \sum_{D\in\mathcal{D}_H\setminus\{1\}} \frac{A^{\omega(D)}}{D} \mathbb{1}_{m \bmod D \in I_D - I_D}. \tag{6.1}$$

Note that $E_A(m;H) = E_A(-m;H)$ for all $m \in \mathbb{Z}$. Also, this notation differs slightly from that in [6], in that we include the factor $\mathbb{1}_{m\neq 0}$ here. Our notation then makes in unnecessary to explicitly exclude the case $m = 0$ from summations.

We need the following lemmas, which are Lemma 5.1 and Lemma 5.2 of [6], respectively.

Lemma 6.1. *Let $10 < H < z^{1/M}$, $1 \leqslant l \leqslant BKH$, and $\mathcal{U} \subset \mathcal{V}$ be two finite sets of integers with $|\mathcal{V}| = l$. Then*

$$\mathbb{P}(\mathcal{U} \subset \mathbf{S}_2) = \sigma_2^{|\mathcal{U}|}\left(1 + O\Big(|\mathcal{U}|^2 H^{-M} + l^{-2} \sum_{v,v'\in\mathcal{V}} E_{2l^2B}(v - v'; H)\Big)\right).$$

We note that in Lemma 5.1 of [6] the slightly different range $1 \leqslant l \leqslant 10KH$ is stated, but actually the same proof gives the above Lemma 6.1.

Lemma 6.2. *Let $10 < H < z^{1/M}$, $0 < AB^2 \leqslant H^M$ and $(m_t)_{t\in T}$ be a finite sequence such that*

$$\sum_{t\in T} \mathbb{1}_{m_t \equiv a \pmod D} \ll \frac{X}{\varphi(D)} + R \tag{6.2}$$

for some $X, R > 0$, and all $D \in \mathcal{D}_H\setminus\{1\}$ and $a \in \mathbb{Z}/D\mathbb{Z}$. Then for any integer j

$$\sum_{t\in T} E_A(m_t + j; H) \ll \frac{XA}{H^M} + R\exp(AB^2 \log\log y).$$

For the rest of the paper we use the notation

$$A = A(H) = 8B^3K^2H^2$$

for the brevity, and also for prime q, define the set

$$\mathcal{A}_q = \Big\{a_{i,q} - a_{j,q} \,\big|\, 1 \leqslant i \leqslant j \leqslant \nu_q\Big\}.$$

Recalling that $a_{i,q} \in [1, q]$, we see that $\mathcal{A}_q \subset [1 - q, q - 1]$.

We will need the following bound, which is where we deploy Hypothesis (g).

Lemma 6.3. *Let $\nu \in \mathcal{N}$ and $H \in \mathfrak{H}$. For $q \in \mathcal{Q}_{H,\nu}$, suppose that $m_q \bmod q \in I_q - I_q$ with $0 < |m_q| \leqslant x\log x$, and suppose that $w \in \mathbb{Z}$ with $|w| \leqslant x\log x$. Then*

$$\sum_{q\in\mathcal{Q}_{H,\nu}} E_{A(H)}(m_q + w; H) \ll \frac{|\mathcal{Q}_{H,\nu}|\log H}{H^{M-2}}.$$

Proof. If $m_q \bmod D \in I_D - I_D$, then $m_q \bmod p \in I_p - I_p$ for each $p|D$. Thus, if $m_q + w \neq 0$ then

$$\begin{aligned} E_A(m_q + w; H) &= \prod_{\substack{m_q + w \bmod p \in I_p - I_p \\ H^M < p \leqslant z}} \left(1 + \frac{A}{p}\right) - 1 \\ &\leqslant \exp\left(A \sum_{\substack{m_q + w \bmod p \in I_p - I_p \\ H^M < p \leqslant z}} \frac{1}{p}\right) - 1. \end{aligned} \tag{6.3}$$

Recall the notation $N(m)$ from Hypothesis (f). Thus, the number of primes p with $H^M < p \leqslant z$ and with $m_q + w \bmod p \in I_p - I_p$ is at most $N(m_q + w)$. Let c_3 be a sufficiently large constant, depending on c_1 and c_2 from Hypothesis (g), and let

$$\widetilde{\mathcal{Q}}_{H,\nu} = \big\{q \in \mathcal{Q}_{H,\nu} : N(m_q + w) \leqslant c_3 \log H\big\}.$$

Clearly, for $q \in \widetilde{\mathcal{Q}}_{H,\nu}$ we have

$$\sum_{\substack{m_q+w \bmod p \in I_p - I_p \\ H^M < p \leqslant z}} \frac{1}{p} \leqslant \frac{c_3 \log H}{H^M}.$$

Therefore, using the fact that $A = O(H^2)$,

$$\sum_{q \in \widetilde{\mathcal{Q}}_{H,\nu}} E_A(m_q + w; H) \ll |\widetilde{\mathcal{Q}}_{H,\nu}|\Big(\exp(O((\log H)H^{-(M-2)})) - 1\Big) \ll \frac{|\mathcal{Q}_{H,\nu}| \log H}{H^{M-2}}.$$

Using Hypothesis (g), for some positive constants c_1, c_2 (depending only on the sieving system),

$$\begin{aligned}\sum_{q \in \mathcal{Q}_{H,\nu} \setminus \widetilde{\mathcal{Q}}_{H,\nu}} E_A(m_q + w; H) &\leqslant \sum_{k > c_3 \log H} \#\big\{q \in \mathcal{Q}_{H,\nu} : m_q + w \neq 0,\ N(m_q + w) = k\big\} e^{Ak/H^M} \\ &\ll x(\log x)^{c_1+1} \sum_{k > c_3 \log H} e^{-c_2 k} \exp\Big(O(kH^{-(M-2)})\Big) \\ &\ll x(\log x)^{c_1+1} e^{-c_2 c_3 \log H} \\ &\ll |\mathcal{Q}_{H,\nu}| H^{-(M-2)},\end{aligned}$$

if c_3 is large enough, using the lower bound $H \geqslant (\log x)^\delta$ from (3.4) and the asymptotic (3.5); recall also that $6 < M \leqslant 7$. This concludes the proof. □

Now we fix $H \in \mathfrak{H}$ and $\nu \in \mathcal{N}$ for the rest of the paper. We start with the proof of part (ii) in the case $j = 1$ (the case $j = 0$ being trivial), which is

$$\mathbb{E} \sum_{q \in \mathcal{Q}_{H,\nu}} \sum_{-(K+1)y < n \leqslant y} \boldsymbol{\lambda}(H; q, n) = \left(1 + O\left(\frac{\log H}{H^{M-2}}\right)\right)(K+2)y|\mathcal{Q}_{H,\nu}|. \tag{6.4}$$

By (3.12), the left-hand side expands as

$$\mathbb{E} \sum_{q \in \mathcal{Q}_{H,\nu}} \sum_{-(K+1)y < n \leqslant y} \frac{\mathbb{1}_{\mathbf{AP}(KH;q,n) \subset \mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}(KH;q,n)|}}.$$

Recall that, according to the definitions (3.8) and (3.9), $\mathbf{b}_1$ and $\mathbf{b}_2$ are independent, and so are $\mathbf{AP}(KH; q, n)$ and $\mathbf{S}_2$. With $\mathbf{b}_1$ fixed, $\mathbf{AP}(KH; q, n)$ is also fixed and we will denote it as $AP(KH; q, n)$. Then the above expression equals

$$\sum_{q \in \mathcal{Q}_{H,\nu}} \sum_{-(K+1)y < n \leqslant y} \sum_{b_1 \bmod P_1} \frac{\mathbb{P}(\mathbf{b}_1 = b_1)}{\sigma_2^{|\operatorname{AP}(KH;q,n)|}} \mathbb{P}(\operatorname{AP}(KH; q, n) \subset \mathbf{S}_2).$$

For fixed q, n, and b_1, we apply Lemma 6.1 to the sets $\mathcal{U} = \operatorname{AP}(KH; q, n)$ and

$$\mathcal{V} = \bigsqcup_{i=1}^{\nu} \{n + a_{i,q} + qh : 1 \leqslant h \leqslant KH\},$$

so that $l = |\mathcal{V}| = \nu\lfloor KH \rfloor \asymp H$. Since $E_A(m;H)$ is an increasing function of A, we find that the left-hand side of (6.4) is equal to

$$\sum_{q\in\mathcal{Q}_{H,\nu}} \sum_{-(K+1)y<n\leqslant y} \Big[1 + O\big(H^{-(M-2)}\big)\Big] + \\ + O\bigg(yH^{-2} \sum_{q\in\mathcal{Q}_{H,\nu}} \sum_{a\in\mathcal{A}_q} \sum_{1\leqslant h,h'\leqslant KH} E_{A(H)}(a+qh-qh';H)\bigg). \tag{6.5}$$

We note that $|a + q(h - h')| \leqslant (K + 1)y \leqslant x \log x$ for $a \in \mathcal{A}_q$ and large x. When h and h' are fixed, we apply Lemma 6.3 with $w = 0$ and $m_q = a + qh - qh'$, where we've chosen one of the $O(1)$ elements $a \in \mathcal{A}_q$ for each q. Thus, we see that the second line in (6.5) is

$$\ll (yH^{-2})H^2|\mathcal{Q}_{H,\nu}| \cdot (\log H)H^{-M+2} = \frac{y|\mathcal{Q}_{H,\nu}| \log H}{H^{M-2}}.$$

This proves the $j = 1$ case of part (ii) in Theorem 4, that is, (6.4).

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

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

The left-hand side is expanded as

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

For fixed q, n_1, n_2, we will apply Lemma 6.1 with

$$U = \mathbf{AP}(KH;q,n_1) \cup \mathbf{AP}(KH;q,n_2)$$

and

$$V = V_1 \cup V_2,$$

where

$$V_j = \bigsqcup_{i=1}^{\nu} \{n_j + a_{i,q} + qh : 1 \leqslant h \leqslant KH\}, \quad j = 1, 2.$$

We first estimate the contribution of the triples (n_1, n_2, q) for which V_1 and V_2 have non-empty intersection. This implies that $(n_1 - n_2) \bmod q \in \mathcal{A}_q$, and, hence, there are $O(yH)$ such pairs n_1, n_2 for each q. Each of them contributes at most $\sigma_2^{-2BKH} = y^{o(1)}$, so the total contribution of such triples is $O(y^{1+o(1)}|\mathcal{Q}_{H,\nu}|)$, which is negligible. Thus we may restrict our attention to those triples (n_1, n_2, q) for which the sets V_1 and V_2 do not intersect; let us call these triples *good*. In particular, for any good triple (n_1, n_2, q), the sets $\mathbf{AP}(KH;q,n_1)$ and $\mathbf{AP}(KH;q,n_2)$ also do not intersect. Then it is enough to show that

$$\mathbb{E} \sum_{q\in\mathcal{Q}_{H,\nu}} \sum_{\substack{-(K+1)y<n_1,n_2\leqslant y \\ (n_1,n_2,q)\text{ good}}} \frac{\mathbb{1}_{\mathbf{AP}(KH;q,n_1)\sqcup\mathbf{AP}(KH;q,n_2)\subset\mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}(KH;q,n_1)|+|\mathbf{AP}(KH;q,n_2)|}} \\ = \left(1 + O\left(\frac{\log H}{H^{M-2}}\right)\right)(K+2)^2y^2|\mathcal{Q}_{H,\nu}|. \tag{6.6}$$

Arguing as in the case $j = 1$, we see that the left-hand side of (6.6) equals

$$\sum_{q\in\mathcal{Q}_{H,\nu}} \sum_{\substack{-(K+1)y<n_1,n_2\leqslant y \\ (n_1,n_2,q)\text{ good}}} \Big(1 + O\Big(\frac{1}{H^{M-2}}\Big)\Big) + \\ + O\bigg(\frac{1}{H^2} \sum_{q\in\mathcal{Q}_{H,\nu}} \sum_{-(K+1)y<n_1,n_2\leqslant y} R_0(n_1,n_2,q)\bigg), \tag{6.7}$$

where

$$\begin{aligned} R_0(n_1,n_2,q) &:= \sum_{v,v'\in V} E_{A(H)}(v-v';H) \\ &\ll \sum_{1\leqslant h,h'\leqslant KH} \sum_{a\in\mathcal{A}_q} E_{A(H)}(n_1-n_2+a+q(h-h');H) \\ &\ll \sum_{\substack{1\leqslant h,h'\leqslant KH \\ a\in\mathcal{A}_q \\ a\neq 0 \text{ or } h\neq h'}} E_{A(H)}(n_1-n_2+a+q(h-h');H) + H\,E_{A(H)}(n_1-n_2) \\ &= R_1(n_1,n_2,q) + H\,E_{A(H)}(n_1-n_2), \end{aligned}$$

say. Recalling that all but $O(yH)$ pairs (n_1,n_2) are good, we get the main term $(K+2)^2y^2|\mathcal{Q}_{H,\nu}|$ from the first line of (6.7), with an acceptable error term.

We next estimate the contribution from $R_1(n_1,n_2,q)$. With n_1,n_2,h_1,h_2 fixed and also fixing one of the $O(1)$ choices for $a\in A_q$ for each $q\in\mathcal{Q}_{H,\nu}$, we estimate $\sum_{q\in\mathcal{Q}_{H,\nu}} R_1(n_1,n_2,q)$ using Lemma 6.3 with $w=n_1-n_2$ and $m_q=a+q(h-h')$. Since either $a\neq 0$ or $h\neq h'$, we have $m_q\neq 0$. Also, for large x, $|w|\leqslant x\log x$ and $|m_q|\leqslant x\log x$. Therefore,

$$\sum_{n_1,n_2}\sum_{q\in\mathcal{Q}_{H,\nu}} R_1(n_1,n_2,q) \ll H^2y^2\,\frac{|\mathcal{Q}_{H,\nu}|\log H}{H^{M-2}},$$

which is acceptable for (6.6). To estimate the contribution from $E_{A(H)}(n_1-n_2;H)$, we apply Lemma 6.2, by first fixing n_2 and observing that (6.2) holds with $X=y$ and $R=1$. Therefore, recalling that $A(H)\ll H^2$,

$$\begin{aligned} \sum_{q\in\mathcal{Q}_{H,\nu}}\sum_{-(K+1)y<n_1,n_2\leqslant y} E_{A(H)}(n_1-n_2;H) &\ll |\mathcal{Q}_{H,\nu}|\,y\biggl(\frac{y}{H^{M-2}}+e^{AB^2\log\log y}\biggr) \\ &\ll \frac{y^2|\mathcal{Q}_{H,\nu}|}{H^{M-2}}, \end{aligned}$$

which is also acceptable for (6.6). This gives (6.6), as desired, completing the $j=2$ case of (ii).

Proof of (iii). Fix $H\in\mathfrak{H}$, $\nu\in\mathcal{N}$ and $1\leqslant i\leqslant\nu$. The case $j=0$ follows from part (i), so we focus on the case $j=1$, which states

$$\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\sum_{q\in\mathcal{Q}_{H,\nu}}\sum_{h\leqslant KH}\boldsymbol{\lambda}(H;q,n-a_{i,q}-qh) = \biggl(1+O\biggl(\frac{\log H}{H^{M-2}}\biggr)\biggr)|\mathcal{Q}_{H,\nu}|\cdot\lfloor KH\rfloor\sigma_1 y.$$

It is enough to show that, for any $h\leqslant KH$,

$$\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\sum_{q\in\mathcal{Q}_{H,\nu}}\boldsymbol{\lambda}(H;q,n-a_{i,q}-qh) = \biggl(1+O\biggl(\frac{\log H}{H^{M-2}}\biggr)\biggr)|\mathcal{Q}_{H,\nu}|\sigma_1 y. \tag{6.8}$$

According to (3.12), the left-hand side is equal to

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

By (3.10), the condition $n\in\mathbf{S}\cap[1,y]$ implies that $n\in\mathbf{S}_1\cap[1,y]$. On the other hand, if $n\in\mathbf{S}_1$, then $n\in\mathbf{AP}(KH;q,n-a_{i,q}-qh)$, and thus the condition $n\in\mathbf{S}_2$ is contained in the condition $\mathbf{AP}(KH;q,n-a_{i,q}-qh)\subset\mathbf{S}_2$. So the left-hand side of (6.8) can be rewritten as

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

Recalling that $\mathbf{S}_2$ is independent of $\mathbf{S}_1$ and of $\mathbf{AP}(KH; q, n-a_{i,q}-qh)$, we may apply Lemma 6.1 as before and find that the left-hand side of (6.8) is

$$\mathbb{E}\sum_{n\in\mathbf{S}_1\cap[1,y]}\sum_{q\in\mathcal{Q}_{H,\nu}}\left(1+O\left(\frac{1}{H^{M-2}}+H^{-2}\sum_{a\in\mathcal{A}_q}\sum_{h',h''\leqslant KH}E_{A(H)}(a+qh'-qh'')\right)\right).$$

Recall that $\mathbb{E}|\mathbf{S}_1\cap[1,y]|=\sigma_1 y$ by Theorem 4 (i). Thus, we see that (6.8) follows from Lemma 6.3, applied with n, h', h'' fixed, $w=0$, some choice of $a\in\mathcal{A}_q$ for each q, and $m_q=a+qh'-qh''$.

Now we turn to the case $j=2$ of (iii), which states

$$\sum_{h_1,h_2\leqslant KH}\mathbb{E}\sum_{n\in\mathbf{S}\cap[1,y]}\sum_{q_1,q_2\in\mathcal{Q}_{H,\nu}}\boldsymbol{\lambda}(H;q_1,n-a_{i,q_1}-q_1h_1)\boldsymbol{\lambda}(H;q_2,n-a_{i,q_2}-q_2h_2)$$
$$=\left(1+O\left(\frac{\log H}{H^{M-2}}\right)\right)|\mathcal{Q}_{H,\nu}|^2\cdot\lfloor KH\rfloor^2\frac{\sigma_1}{\sigma_2}y. \tag{6.9}$$

Arguing as in the $j=1$ case, the left-hand side equals

$$\sum_{h_1,h_2\leqslant KH}\mathbb{E}\sum_{n\in\mathbf{S}_1\cap[1,y]}\sum_{q_1,q_2\in\mathcal{Q}_{H,\nu}}\frac{\mathbb{1}_{\mathbf{AP}(KH;q_1,n-a_{i,q_1}-q_1h_1)\cup\mathbf{AP}(KH;q_2,n-a_{i,q_2}-q_2h_2)\subset\mathbf{S}_2}}{\sigma_2^{|\mathbf{AP}(KH;q_1,n-a_{i,q_1}-q_1h_1)|+|\mathbf{AP}(KH;q_2,n-a_{i,q_2}-q_2h_2)|}}; \tag{6.10}$$

Note that here we again replace the condition $n\in\mathbf{S}\cap[1,y]$ by $n\in\mathbf{S}_1\cap[1,y]$ for the same reason as in $j=1$ case. Further, by (5.1), the contribution from $q_1=q_2$ is

$$\ll H^2\sigma_2^{-2BKH}|\mathcal{Q}_{H,\nu}|\sigma_1 y\ll|\mathcal{Q}_{H,\nu}|y^{1+o(1)},$$

which, by (3.5), is an acceptable error term.

We call a pair $(q_1,q_2)\in\mathcal{Q}_{H,\nu}^2$ with $q_1\neq q_2$ *good*, if for all $\mathbf{S}_1$, all $n\in\mathbf{S}_1\cap[1,y]$ and all $h_1,h_2\leqslant KH$ we have

$$\{n\}=\mathbf{AP}(KH;q_1,n-a_{i,q_1}-q_1h_1)\cap\mathbf{AP}(KH;q_2,n-a_{i,q_2}-q_2h_2),$$

and call (q_1,q_2) *bad* otherwise; recall that for any $n\in\mathbf{S}_1\cap[1,y]$, n lies in both $\mathbf{AP}(KH;q_1,n-a_{i,q_1}-q_1h_1)$ and $\mathbf{AP}(KH;q_2,n-a_{i,q_2}-q_2h_2)$. We need to estimate the number of bad pairs. First of all, if a pair (q_1,q_2) is bad then there is a choice of h_1,h_2 so that both sets

$$\bigsqcup_{j_1=1}^{\nu}\{a_{j_1,q_1}-a_{i,q_1}+q_1(h_1''-h_1):h_1''\leqslant KH\}$$

and

$$\bigsqcup_{j_2=1}^{\nu}\{a_{j_2,q_2}-a_{i,q_2}+q_2(h_2''-h_2):h_2''\leqslant KH\}$$

contain the same nonzero number, say, n_0. Fix q_2, j_2, h_2 and h_2'' so that

$$n_0=a_{j_2,q_2}-a_{i,q_2}+q_2(h_2''-h_2).$$

Then we have $n_0 \bmod q_1\in I_{q_1}-I_{q_1}$. By Hypothesis (f), the number of such q_1 is $O(y^{0.49})$. Therefore, the number of bad pairs (q_1,q_2) is $\ll y^{1.49}H^2\ll y^{1.5}$. Since each of them contributes $y^{1+o(1)}$ to the left side of (6.10), the contribution from these bad pairs is negligible.

It remains to estimate the contribution to (6.10) from good pairs (q_1,q_2). Note that if (q_1,q_2) is a good pair, then, for any $\mathbf{S}_1,h_1,h_2,n$ the set

$$\mathbf{AP}(KH;q_1,n-a_{i,q_1}-q_1h_1)\cup\mathbf{AP}(KH;q_2,n-a_{i,q_2}-q_2h_2)$$

has size $|\mathbf{AP}(KH; q_1, n - a_{i,q_1} - q_1h_1)| + |\mathbf{AP}(KH; q_2, n - a_{i,q_2} - q_2h_2)| - 1$. Then, as before, we can apply Lemma 6.1 to rewrite the terms in (6.10) corresponding to good (q_1, q_2) as

$$\frac{\sigma y}{\sigma_2^2} \sum_{(q_1,q_2)\text{ good}} \sum_{h_1,h_2\leqslant KH} \left(1 + O\Big(\frac{1}{H^{M-2}}\Big)\right) + \\ + O\left(\frac{\sigma y}{\sigma_2^2 H^2} \sum_{q_1,q_2\in\mathcal{Q}_{H,\nu}} \Big(H^2E'(q_1) + H^2E'(q_2) + E''(q_1,q_2)\Big)\right), \tag{6.11}$$

where again we used that $\mathbb{E}|\mathbf{S}_1 \cap [1, y]| = \sigma_1 y$ from (5.1), that $\sigma_1/\sigma_2 = \sigma/\sigma_2^2$ and where we define

$$E'(q) = \sum_{a\in\mathcal{A}_q} \sum_{h,h'\leqslant KH} E_{A(H)}(a + qh - qh'; H)$$

and

$$E''(q_1, q_2) = \sum_{1\leqslant h_1,h_2\leqslant KH} \sum_{\substack{a_1\in\mathcal{A}_{q_1}\\ a_2\in\mathcal{A}_{q_2}}} \sum_{h_1',h_2'\leqslant KH} E_{A(H)}(a_1 - a_2 + q_1h_1' - q_1h_1 - q_2h_2' + q_2h_2; H).$$

As the number of bad pairs (q_1, q_2) is very small, the first line of (6.11) produces the main term in (5.3) with an acceptable error.

By Lemma 6.3 with $w = 0$,

$$\sum_{q_1,q_2\in\mathcal{Q}_{H,\nu}} (E'(q_1) + E'(q_2)) \ll H^2\,\frac{|\mathcal{Q}_{H,\nu}|^2 \log H}{H^{M-2}}. \tag{6.12}$$

For the sum on $E''(\cdot)$, if we have $a_1 = a_2 = h_1' - h_1 = h_2' - h_2 = 0$ then the summand is zero for any q_1, q_2. Consider now the summands with either $a_1 \neq 0$ or $h_1 \neq h_1'$. Fix $h_1, h_2, h_1', h_2', q_2, a_2$ and also a choice $a_1 \in \mathcal{A}_{q_1}$ for each $q_1 \in \mathcal{Q}_{H,\nu}$. Apply Lemma 6.3 to the sum over q_1, with $w = -a_2 - q_2h_2' + q_2h_2$ and $m_q = a_1 + q_1(h_1' - h_1)$ so that $m_q \neq 0$. A similar argument handles the case when $a_2 \neq 0$ or $h_2 \neq h_2'$, that is, fixing q_1, a_1 and summing over q_2, and we conclude that

$$\sum_{q_1,q_2\in\mathcal{Q}_{H,\nu}} E''(q_1, q_2) \ll H^4\,\frac{|\mathcal{Q}_{H,\nu}|^2 \log H}{H^{M-2}}. \tag{6.13}$$

Inserting (6.12) and (6.13) into (6.11) establishes the desired bound (6.9).

This completes the proof of the case $j = 2$, and Theorem 4 (iii) follows.

□

REFERENCES

[1] P. T. Bateman and R. A. Horn, *A heuristic asymptotic formula concerning the distribution of prime numbers*, Math. Comp. **16** (1962), 363–367.

[2] 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.

[3] N. Tschebotareff, *Die Bestimmung der Dichtigkeit einer Menge von Primzahlen, welche zu einer ge gebenen Substitutionsklasse gehören*, Mathematische Annalen **95** (1) (1926), 191–228.

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

[5] K. Ford, B. Green, S. Konyagin, J. Maynard, T. Tao, *Long gaps between primes*, Journal of the American Mathematical Society **31** (2018), 65–105.

[6] K. Ford, S. V. Konyagin, J. Maynard, C. Pomerance, T. Tao, *Long gaps in sieved sets*, J. European Math. Soc. (JEMS) **23**, 667–700 (2021). *Corrigendum:* J. European Math. Soc. (JEMS) **25** (2023), no. 6, 2483–2485.

[7] P. Erdős, *On the sum* $\sum_{k=1}^{x} d(f(k))$, J. London Math. Soc. **27**, 1952, pp. 7–15.

[8] J. C. Lagarias and 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.

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

[10] I. G. Macdonald, *Symmetric functions and Hall polynomials*, Oxford university press, 1998.

[11] G. Pólya, *Über ganzwertige ganze Funktionen*, Rend. Circ. Mat. Palermo **40** (1915), 1–16.

DEPARTMENT OF MATHEMATICS, 1409 WEST GREEN STREET, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, URBANA, IL 61801, USA
Email address: ford126@illinois.edu

DEPARTMENT OF MATHEMATICS, 1409 WEST GREEN STREET, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, URBANA, IL 61801, USA; STEKLOV MATHEMATICAL INSTITUTE, GUBKINA STR., 8, MOSCOW, 119991, RUSSIA
Email address: gabdullin.mikhail@yandex.ru, mikhailg@illinois.edu