SMALL GAPS BETWEEN PRIMES

1. Introduction

Recently Goldston, Pintz, Yıldırım [to appear] proved that

\[
\liminf_{n \to \infty} \frac{p_{n+1} - p_n}{\log p_n} = 0
\]  

(1)

where \( \{p_n\} \) denotes the sequence of primes in their natural order. In fact they are able to prove a good deal more than this. For example they obtain [submitted] an explicit upper bound

\[
p_{n+} - p_n \ll (\log p_n)^{1/2} (\log \log p_n)^2
\]

(2)

which is satisfied for infinitely many \( n \), and they are able to show that the difference \( p_{n+1} - p_n \) is infinitely often bounded under the assumption of an unproven but plausible hypothesis concerning the level of distribution of the primes \( p \leq x \) into arithmetic progressions, namely that the Bombieri–Vinogradov theorem [Chapter xx] holds for moduli \( q \leq Q \) with \( Q = x^\theta \) for some \( \theta > \frac12 \). Their principal idea is to use artefacts from sieve theory, especially the Selberg sieve, not directly in the form of a sieve but as a means to enhance terms of special interest, particularly those terms with relatively few prime factors. As a preliminary observation consider the starting point for the Selberg upper bound sieve in the form

\[
\sum_{a \in \mathcal{A}} \left( \sum_{\substack{q \leq R \ q \mid a}} \Lambda_q \right)^2
\]

and recall that one is planning to minimise this under the assumptions that \( \Lambda_1 = 1 \) and that

\[
A_d = \sum_{a \in \mathcal{A}} 1 \quad \text{where} \quad d \mid a
\]

can be approximated by an expression of the form

\[
\frac{Xg(d)}{d}
\]

where \( X \) is a good approximation to \( A_1 \) and \( g \) is multiplicative. The minimising choice of \( \Lambda_q \) is given by

\[
\Lambda_q = \mu(q) \frac{S(R, q)}{S(R, 1)} \prod_{p \mid q} \left( \frac{p}{p - g(p)} - 1 \right)
\]
where

\[ S(R,q) = \sum_{r \leq R/q, (r,q)=1} \mu(q)^2 \prod_{p|q} \frac{g(p)}{p - g(p)}. \]

Typically one applies this when the sieve is of dimension \( k \), e.g.

\[ \sum_{p \leq y} g(p) \frac{\log p}{p} = k \log y + O(1). \]

Under this kind of condition one might expect that

\[ S(R,q) \sim C (\log R)^k \prod_{p|q} \frac{p - g(p)}{p} \]

and so \( \Lambda_q \) could be replaced by

\[ \Lambda_q = \mu(q) \frac{\log^k (R/q)}{\log^k R} \]

Indeed this is correct, and whilst there is some loss in precision in the final conclusion there is one significant advantage, namely that this choice of \( \Lambda_q \) can be applied quite effectively to any sieving question where the dimension is \( k \). One might add that the factor \( \log^{-k} R \) can be considered as a normalising factor, and as it occurs in every term one can pursue the analysis with this factor omitted.

Let

\[ \mathbf{h} = h_1, \ldots, h_k \]

denote a \( k \)-tuple of integers satisfying

\[ 1 \leq h_j \leq H. \]

and consider

\[ P(n; \mathbf{h}) = (n + h_1) \ldots (n + h_k). \]

The successful line of attack is based on exploring the existence of more than one prime amongst the factors \( n + h_j \) and in order to enhance the chance of this occurring it is natural to consider

\[ \Lambda_R(n; \mathbf{h}, l) = \sum_{q \leq R \atop q \mid P(n, \mathbf{h})} \mu(q) \frac{(\log R/q)^m}{m!} \]

where \( m > k \) and \( l = m - k \). It will be seen that there a natural reason for the \( m! \) also. Of course one assumes that

\[ k \geq 2. \]
The likelihood of discovering primes in the $k$-tuple $n + h_1, \ldots, n + h_k$ depends on the avoidance of the zero residue class modulo $p$ for all primes $p$. A measure of this is the singular series

$$
\mathcal{S}_k(h) = \prod_p \left( 1 - \frac{\nu_p(h)}{p} \right) \left( 1 - \frac{1}{p} \right)^{-k}
$$

(7)

where $\nu_p(h)$ is the number of distinct residue classes modulo $p$ amongst the $h$. The case $k = 2, h_1 = 0, h_2 = 2$ is the twin prime constant (see Chapter [x]). This expression was first studied in many special cases by Hardy and Littlewood [1923]. See also Vaughan [1997], especially Chapter 3. Clearly $\nu_p(h) \leq k$ and the singular series can only be convergent when $\nu_p(h) = k$ for the majority of primes $p$. This is only possible if the $h_1, \ldots, h_k$ are distinct, and then it will hold whenever $p$ is large enough. It is useful, therefore, to define $H$ to be the set of $h$ such that $h_1, \ldots, h_k$ are distinct and satisfy $h_j \leq H$.

2. Averages of the sifting function

The heart of the proof of (1) is contained in the following two theorems.

**Theorem 1.** Suppose that $k \geq 2$, $l \geq 1$, $H \leq \log N$ and $N^{1/8} \leq R \leq N^{1/4}$ and $h \in \mathcal{H}$. Then

$$
\sum_{n \leq N} \Lambda_R(n; h, l)^2 = \frac{(2l)!}{(k + 2l)!} \mathcal{S}_k(h) N(\log R)^{k + 2l} + O \left( N(\log N)^{k + 2l - 1}(\log \log N)^2 \right).
$$

It is readily seen from the proof that the exponent $1/4$ here could be replaced by anything smaller than $1/2$.

It is useful to define the function

$$
\theta(n) = \begin{cases} 
\log n & n \text{ prime,} \\
0 & \text{otherwise.}
\end{cases}
$$

**Theorem 2.** Suppose that $k \geq 2$, $l \geq 1$, $H \leq \log N$, $h \in \mathcal{H}$ and $1 \leq h_0 \leq H$. Then there is a number $B(k, l)$ such that whenever $N^{1/8} \leq R \leq N^{1/4}(\log N)^{-B(k, l)}$ one has the following. If $h_0 \not\in h$ and $h^* = h \cup \{h_0\}$, then

$$
\sum_{n \leq N} \theta(n + h_0) \Lambda_R(n; h, l)^2 = \frac{(2l)!}{(k + 2l)!} \mathcal{S}_{k+1}(h^*) N(\log R)^{k + 2l} + O \left( N(\log N)^{k + 2l - 1}(\log \log N)^2 \right).
$$

If $k \geq 3$ and $h_0 \in h$, then

$$
\sum_{n \leq N} \theta(n + h_0) \Lambda_R(n; h, l)^2 = \frac{(2l + 2)!}{(k + 2l + 1)!} \mathcal{S}_k(h) N(\log R)^{k + 2l + 1} + O \left( N(\log N)^{k + 2l}(\log \log N)^2 \right).
$$
The treatments of these theorems is similar and can be reduced fairly quickly to a uniform treatment. In the case of Theorem 1 the sum in question is

\[
\sum_{q \leq R} \sum_{r \leq R} \mu(q) \mu(r) \frac{(\log R/q)^m}{m!} \frac{(\log R/r)^m}{m!} \sum_{n=1}^{N} f([q,r]).
\]

Let \( \rho(q) \) denote the number of solutions of the congruence \( P(n; h) \equiv 0 \pmod{q} \). Then \( \rho \) is a multiplicative function, \( \rho(p) = \nu_p(h) \) and so \( \rho(p) = k \) \( \downarrow D \) where \( D = \prod_{1 \leq i < j \leq k} |h_j - h_i| \), and generally \( 0 \leq \rho(p) \leq k \). Thus the sum over \( n \) is

\[
N \frac{\rho(q)}{q} + \vartheta \rho(q)
\]

where \( |\vartheta| \leq 1 \). Thus the sum becomes

\[
N \sum_{q \leq R} \sum_{r \leq R} \mu(q) \mu(r) \frac{(\log R/q)^m}{m!} \frac{(\log R/r)^m}{m!} f([q,r]) + O(E)
\]

where

\[
E = \sum_{q \leq R} \sum_{r \leq R} \mu(q)^2 \mu(r)^2 \frac{(\log R/q)^m}{m!} \frac{(\log R/r)^m}{m!} \rho([q,r]).
\]

Plainly \( \frac{(\log R/q)^m}{m!} \frac{(\log R/r)^m}{m!} \rho([q,r]) \ll R^\varepsilon \) and so \( E \ll R^{2+\varepsilon} \) and in view of the hypothesis \( R \leq N^{1/4} \) this is easily absorbed into the claimed error bound. Thus it remains to deal with

\[
\sum_{q \leq R} \sum_{r \leq R} \mu(q) \mu(r) \frac{(\log R/q)^m}{m!} \frac{(\log R/r)^m}{m!} f([q,r])
\]

where

\[
f(p) = \frac{k}{p} \quad \downarrow D
\]

where

\[
D = \prod_{1 \leq i < j \leq k} |h_j - h_i|
\]

and

\[
0 \leq f(p) = \nu_p(h) \leq \frac{k}{p} \leq \min \left( \frac{k}{p}, 1 \right)
\]

generally.

In Theorem 2, regardless of whether \( h_0 \) is in \( h \) or not the initial sum is

\[
\sum_{q \leq R} \sum_{r \leq R} \mu(q) \mu(r) \frac{(\log R/q)^m}{m!} \frac{(\log R/r)^m}{m!} \sum_{n=1}^{N} \theta(n + h_0).
\]
For a given \( u = [q, r] \) the \( n \) with \( (n + h_0, q) > 1 \) contribute \( \ll (\log u) \log(2N) \) to the sum over \( n \). Thus the total contribution from such \( n \) is

\[
\ll (\log R)(\log 2N) \left( \sum_{q \leq R} (\log R/q)^m \right)^2 \ll R^2 (\log N)^2
\]

which is more than acceptable as part of the error terms in Theorem 2. Also the sum over the remaining \( n \) can be replaced by the sum

\[
\sum_{n=1}^{N} \theta(n)
\]

\[
\text{with } \sum_{(n, u) = 1 \atop u \mid P(n - h_0; h)} (\sum_{q \leq R} (\log R/q)^m)^2 \ll R^2 (\log N)^2
\]

which is more than acceptable as part of the error terms in Theorem 2. Also the sum over the remaining \( n \) can be replaced by the sum

\[
\sum_{v=1}^{u} \vartheta(N; u, v).
\]

(12)

Let

\[
\sigma(u) = \sum_{v=1}^{u} \vartheta(N; u, v) \quad \text{with } \vartheta(N; u, v) = \sum_{(v, u) = 1 \atop u \mid P(r - h_0; h)} (\sum_{q \leq R} (\log R/q)^m)^2 \ll R^2 (\log N)^2
\]

This is a multiplicative function of \( u \), and \( \sigma(p) \) is the number of solutions of \( P(r - h_0; h) \equiv 0 \pmod{p} \) with \( 1 \leq v \leq p - 1 \). If \( h_0 \in \mathfrak{h} \), say \( h_0 = h_j \), then the possible solution \( v \equiv h_0 - h_j \) is excluded by the condition \( (p, v) = 1 \). Thus \( \sigma(p) = \nu_p(\mathfrak{h}) - 1 \) and so \( \sigma(p) \leq k - 1 \). On the other hand, when \( p \nmid \prod_{1 \leq j \leq k} |h_j - h_i| \), we have \( \sigma(p) = k - 1 \). If \( h_0 \notin \mathfrak{h} \), then \( \sigma(p) = \nu_p(\mathfrak{h}) - 1 \) and so \( \sigma(p) \leq k \) and when \( p \nmid \prod_{0 \leq i < j \leq k} |h_j - h_i| \) we have \( \sigma(p) = k \).

In either case we can replace \( \vartheta(N; u, v) \) in (12) by \( N/\phi(u) \) with a total error in our original sum of at most

\[
\ll \left( \sum_{u \leq R^2} \mu(u)^2 t(u) d_k(u) \max_{1 \leq v \leq u \atop (v, u) = 1} |E(N; u, v)| \right) \sum_{u \leq R^2} \max_{1 \leq v \leq u \atop (v, u) = 1} |E(N; u, v)|.
\]

where

\[
E(N; u, v) = \vartheta(N; u, v) - \frac{N}{\phi(u)}
\]

and \( t(u) \) is the number of choices of \( q, r \) with \( [q, r] = u \). For squarefree \( u \), \( t(u) \leq d_3(u) \). By an application of the Cauchy–Schwarz inequality the square of the sum above is

\[
\ll \left( \sum_{u \leq R^2} \mu(u)^2 t(u)^2 d_k(u)^2 u^{-1} N \log N \right) \sum_{u \leq R^2} \max_{1 \leq v \leq u \atop (v, u) = 1} |E(N; u, v)|.
\]
and by the Bombieri–Vinogradov theorem it follows that if \( R \leq N^{1/4}(\log N)^{-B(k,l)} \) for suitable \( B(k,l) \), then the expression in (13) is

\[
\ll N
\]

which is again acceptable. Thus once more it remains to deal with (8) where now either \( h_0 \in h \),

\[
f(p) = \frac{k - 1}{p - 1} \quad (p \nmid D)
\]

where

\[
D = \prod_{1 \leq i < j \leq k} |h_j - h_i|
\]

and

\[
0 \leq f(p) = \frac{\nu_p(h) - 1}{p - 1} \leq \min \left( \frac{k - 1}{p - 1}, 1 \right)
\]

generally, or \( h_0 \notin h \),

\[
f(p) = \frac{k}{p - 1} \quad (p \nmid D)
\]

where

\[
D = \prod_{0 \leq i < j \leq k} |h_j - h_i|
\]

and

\[
0 \leq f(p) = \frac{\nu_p(h^*) - 1}{p - 1} \leq \min \left( \frac{k}{p - 1}, 1 \right)
\]

generally.

In the sum (8) we put \( a = (q,r) \) and then replace \( q \) and \( r \) by \( aq \) and \( ar \) so that now \( (q,r) = 1 \), \( aq \leq R \) and \( ar \leq R \). Now we replace the condition \( (q,r) = 1 \) by \( \sum_{b|(q,r)} \mu(b) \) and then replace \( q \) and \( r \) by \( bq \) and \( br \) so that \( abq \leq R \) and \( abr \leq R \). Thus, as \( f \) is multiplicative and for a non-zero contribution \( abq \) and \( abr \) are squarefree, the expression in (8) becomes

\[
\sum_a f(a) \sum_b \mu(b) f(b)^2 \left( \sum_{q \leq R/(ab)} \mu(abq) f(q) \frac{(\log \frac{R}{abq})^m}{m!} \right)^2
\]

\[
= \sum_n \mu(n)^2 f(n) \sum_{b|n} \mu(b) f(b) \left( \sum_{q \leq R/n} \mu(q) f(q) \frac{(\log \frac{R}{q})^m}{m!} \right)^2
\]

We now treat the innermost sum here in the cases when \( f(p) = p/k \) or \( f(p) = k/(p - 1) \) for large \( p \), i.e (11) or (19) hold. The remaining case will follow essentially by replacing \( k \) by \( k - 1 \).
Lemma 3. Suppose that \( k \in \mathbb{N}, l \in \mathbb{N}, m = k + l, n \in \mathbb{N}, D \in \mathbb{N}, X \geq 1 \) and \( g \) is a multiplicative function such that \( g(p) = k/p \) whenever \( p \nmid D \), or \( g(p) = k/(p-1) \) when \( p \nmid D \), and that \( 0 \leq g(p) \leq \min\left(k/(p-1), 1\right) \) generally. Then

\[
\sum_{q \leq X} \mu(q) g(q) \frac{(\log X/q)^m}{m!} = G_n \frac{(\log X)^l}{l!} + E
\]

where

\[
G_n = \left( \prod_{p \mid n} (1 - 1/p)^{-k} \right) \prod_{p \nmid n} ((1 - g(p))(1 - 1/p)^{-k})
\]

and

\[
E \ll_m (\log X)^{l-1}(\log D) \prod_{p \mid n} (1 + kp^{-3/4}).
\]

We also have

\[
G_n \ll_m (\log D) \prod_{p \mid n} (1 + kp^{-3/4}).
\]

Suppose in addition that \( g(p) \leq 1 - \delta \) for every prime \( p \) where \( \delta \) is a positive number depending at most on \( k \). Then

\[
\sum_{n \leq X} \mu(n)^2 \left( \prod_{p \mid n} \frac{g(p)}{1 - g(p)} \right) \frac{(\log X/n)^{2l}}{(2l)!} = G^* \frac{(\log X)^{k+2l}}{(k+2l)!} + E^*
\]

where

\[
G^* = \prod_{p} (1 - g(p))^{-1}(1 - 1/p)^k
\]

and

\[
E^* \ll_m (\log X)^{k+2l-1} \log 2D.
\]

Proof. When \( 1 \leq X \leq 2 \) the main conclusions follow easily from the bounds for \( G_n \) and \( G^* \), so it suffices to prove the lemma when \( X \geq 2 \). We first consider the first part of the lemma. For any \( \theta > 0 \) and \( Y > 0 \),

\[
\frac{1}{2\pi i} \int_{\theta - i\infty}^{\theta + i\infty} \frac{Y^s}{s^{m+1}} ds = \begin{cases} \frac{(\log Y)^m}{m!} & \text{when } Y \geq 1, \\ 0 & \text{when } 0 \leq Y < 1. \end{cases}
\]

The series

\[
F_n(s) = \sum_{\substack{q = 1 \\ (q,n) = 1}}^{\infty} \frac{\mu(q)g(q)}{q^s}
\]
converges absolutely and locally uniformly in the half-plane $\Re s > 0$. Hence the sum in question can be rewritten as

$$\sum_{q \leq X, (q,n)=1} \mu(q) g(q) \frac{(\log X/q)^m}{m!} = \frac{1}{2\pi i} \int_{\theta-i\infty}^{\theta+i\infty} F_n(s) \frac{X^s}{s^{m+1}} ds. \quad (21)$$

Moreover, for $\Re s > 0$ we have

$$F_n(s) = \zeta(s+1)^{-k} G_n(s)$$

where

$$G_n(s) = \left( \prod_{p|n} (1 - p^{-1-s})^{-k} \right) \prod_{p \nmid n} ((1 - g(p)p^{-s})(1 - p^{-1-s})^{-k}).$$

Since $g(p) = k/p$ or $k/(p-1)$ for large $p$ the product defining $G_n$ converges absolutely and locally uniformly for $\Re s > -\frac{1}{2}$. Thus the integrand is analytic throughout that half–plane except at $s = 0$, where it has a pole of order $m + 1 - k = l + 1$, and at the zeros of $\zeta(s+1)$ (if there are any). The residue of the integrand at 0 is

$$\sum_{t=0}^{l} \frac{G_n^{(t)}(0)}{t!} \mathcal{P}_{l-t}(\log X) \quad (22)$$

where $\mathcal{P}_u$ is a polynomial of degree $u$ with leading coefficient $1/u!$ and its coefficients depend at most on $k$ and $l$.

Suppose that $\Re s = \sigma \geq -\alpha \geq -1/4$. Then it follows that

$$\sum_{p \nmid n} \log |(1 - g(p)p^{-s}|1 - p^{-1-s}|^{-k} \leq \sum_{p|D, p \nmid n} p^{\alpha - 1} + c_1(k)$$

whence

$$|G_n(s)| \leq M(k, \alpha) \prod_{p|n} (1 + kp^{-3/4})$$

in that half–plane, where

$$M(k, \alpha) = c(k) \exp \left( k \sum_{p|D} p^{\alpha - 1} \right).$$

Hence, by Cauchy’s inequalities for the derivatives of an analytic function we have

$$|G_n^{(t)}(0)| \leq \alpha^{-t} M(k, \alpha) \prod_{p|n} (1 + kp^{-3/4}).$$
We show that
\[ G_n(s) \ll (\log \log 4D)^k \prod_{p | n} (1 + kp^{-3/4}) \] (23)
in the region \( \Re s \geq -\min(1/4, 1/\log \log 4D) \) and that for \( 0 \leq t \leq l \),
\[ |G_n^{(t)}(0)| \ll (\log \log 4D)^{t+k} \prod_{p | n} (1 + kp^{-3/4}). \] (24)

If \( D \leq e^e \), say, then we have at once that \( G_n(s) \ll \prod_{p | n} (1 + kp^{-3/4}) \) in the half-plane \( \sigma \geq -1/4 \) and \( G_n^{(t)}(0) \ll \prod_{p | n} (1 + kp^{-3/4}) \) when \( t \leq l \). Thus we can suppose that \( D > e^e \). Let \( w = \omega(D) \). Then \( k \sum_{p | D} p^{\alpha-1} \leq k \sum_{p \leq p_w} p^{\alpha-1} \) where \( p_w \) is the \( w \)-th prime.

Now \( \sum_{p \leq p_w} p^{\alpha-1} = \sum_{j=0}^{\infty} \sum_{p \leq p_w} \left( \frac{\alpha \log p}{j!} \right)^j p^{-1} \) The term \( j = 0 \) is \( \leq \log \log p_w + c \) for some constant \( c \). Each term with \( j \geq 1 \) is \( \leq \left( \frac{\alpha \log p_w}{j!} \right)^j \left( \log p_w + c' \right) \). Thus \( k \sum_{p \leq p_w} p^{\alpha-1} \leq k \log \log p_w + O(\exp(\alpha \log p_w)) \), and since \( \vartheta(p_w) \leq \log D \) and so \( p_w \ll \log D \), when we take \( \alpha = 1/\log \log D \) we obtain \( k \sum_{p | D} p^{\alpha-1} \leq k \log \log D + O(1) \). Thus we have the desired conclusion for \( G_n(s) \) and for \( G_n^{(t)}(0) \).

We are now in a position to deal with the vertical path in (21). We take \( \theta = 1/\log(2X) \), and for suitable choices of \( T \geq 3 \) and a positive constant \( c \) replace the finite line segment \( \{ \theta - iT, \theta + iT \} \) by the line segments joining \( \{ \theta - iT, -\frac{1}{c \log T} - iT, -\frac{1}{c \log T} + iT, \theta + iT \} \). Here \( c \) is chosen so that we stay well within the zero–free region for \( \zeta(s+1) \) and we have \( \zeta(s+1)^{-1} \ll \log(2 + |t|) \) on these line segments. See, for example, Montgomery and Vaughan [2006], Theorem 6.7. In the process we pick up the residue of the integrand at \( s = 0 \) given by (22), and in view of (24) this is \( G_n(0) \frac{(\log X)^j}{t!} \) plus an acceptable error. Note that \( G_n(0) = G_n \).

We choose \( T = c' \log 4D \) where \( c' \) is a suitable constant. Thus (23) is valid throughout our new path. The vertical line segments with \( |3s| \geq T \) contribute
\[
\ll T^{-m} (\log T)^k (\log \log 4D)^k \prod_{p | n} (1 + kp^{-3/4}) \ll (\log \log 4D)^k \prod_{p | n} (1 + kp^{-3/4}),
\]
and the horizontal paths with \( |t| = T \) contribute at most a similar amount. The vertical line segments with \( 1 \leq |t| \leq T \) and \( \sigma = -1/(c \log T) \) contribute
\[
\ll X^{-1/(c \log T)} (\log \log 4D)^k \prod_{p | n} (1 + kp^{-3/4}) \ll (\log \log 4D)^k \prod_{p | n} (1 + kp^{-3/4})
\]
and on the part with \( |t| \leq 1 \) and \( \sigma = -1/(c \log T) \), again by Theorem 6.7 of Montgomery and Vaughan [2006] we have \( \zeta(s+1)^{-1} s^{-1} \ll 1 \) and so the integrand is
\[
\ll |s|^{-l-1} X^{-1/(c \log T)} (\log \log 4D)^k \prod_{p | n} (1 + kp^{-3/4}) \ll (\log \log 4D)^{k+l+1} \prod_{p | n} (1 + kp^{-3/4}).
\]
This completes the proof of the first part of the lemma.

The left hand side in the second part of the lemma is

\[ \frac{1}{2\pi i} \int_{\theta - i\infty}^{\theta + i\infty} \sum_{n=1}^{\infty} \mu(n)^2 n^{-s} \left( \prod_{p|n} \frac{g(p)}{1 - g(p)} \right) \frac{X^s}{s^{2l+1}} ds. \]

The Dirichlet series here can be rewritten as

\[ \zeta(s+1)^k G^*(s) \]

where

\[ G^*(s) = \prod_p \left( 1 + \frac{g(p)}{p^s(1 - g(p))} \right) (1 - p^{-s-1})^k. \]

Thus the function \( G^*(s) \) can be treated in the same way as \( G_n(s) \) (with \( n = 1 \)). The main difference now is that the integrand has a pole at 0 of order \( k + 2l + 1 \). However we may choose a similar path to that above, and on it, \( \zeta(s+1) \ll \log(2 + |t|) \). Thus we may proceed as before. Now we obtain the desired conclusion with

\[ E^* \ll (\log X)^{k+2l-1} \left( \log \log 4D \right)^{2l+k+1}. \]

This completes the proof of the lemma.

The sum on the right in (20) is

\[ \sum_{n \in \mathcal{N}} f(n) \left( \prod_{p|n} (1 - f(p)) \right) \left( \sum_{q \leq R/n \atop (q,n)=1} \mu(q) f(q) \frac{(\log R/n)^m}{m!} \right)^2 \]

where \( \mathcal{N} \) is the set of squarefree natural numbers not exceeding \( R \) such that for each prime divisor \( p \) of \( n \) we have \( f(p) < 1 \). Now we apply Lemma 3. The square of the innermost sum is

\[ G_n^2 \left( \frac{(\log R/n)^{2l}}{(2l)!} \right)^2 + O \left( (\log R)^{2l-1} (\log H)^2 \prod_{p|n} (1 + kp^{-3/4})^2 \right). \]

The error term contributes

\[ \ll \sum_{n \leq R} d_k(n) \phi(n)^{-1} (\log R)^{2l-1} (\log \log N)^2 \prod_{p|n} (1 + 2kp^{-3/4}) \]

to the sum over \( n \). The product is bounded by \( \sum_{u|n} d_{2k}(u) u^{-3/4} \) and it then follows that the total is

\[ \ll (\log R)^{k+2l-1} (\log \log N)^2 \]
which is good enough. The main term is
\[
\sum_{n \in \mathbb{N}} f(n) \left( \prod_{p|n} (1 - f(p)) \right) \frac{(\log R/n)^{2l}}{(l!)^2} \left( \prod_{p|n} (1 - 1/p)^{-k} \right)^2 \prod_{p|n} ((1 - f(p))^2(1 - 1/p)^{-2k})
\]
\[
= \left( \prod_{p} (1 - f(p))(1 - 1/p)^{-k} \right)^2 \sum_{n \in \mathbb{N}} f(n) \prod_{p|n} (1 - f(p))^{-1} \frac{(\log R/n)^{2l}}{(l!)^2}.
\]
We now show that this is
\[
\left( \prod_{p} (1 - f(p))(1 - 1/p)^{-k} \right) \left( \frac{2l}{l} \right) \frac{(\log R)^{k+2l}}{(k+2l)!} + O((\log R)^{k+2l-1}(\log \log N)^2).
\]
Since this expansion and the previous one are both 0 (in the leading term) when \(f(p) = 1\) for some \(p\), it suffices to establish the latter one when \(f(p) < 1\) for all \(p\). In that case, by (11) and (19), \(f(p) \leq k/(k+1)\) and we can apply the last part of Lemma 3. This gives
\[
\sum_{n \leq R} \mu(n)^2 \left( \prod_{p|n} \frac{f(p)}{1 - f(p)} \right) \frac{(\log R/n)^{2l}}{(l!)^2} = \left( \frac{2l}{l} \right) \frac{(\log R)^{k+2l}}{(k+2l)!} \prod_{p} (1 - f(p))^{-1}(1 - 1/p)^k + O((\log R)^{k+2l-1} \log H)
\]
and combined with the bound \(G_1 \ll \log D\) of Lemma 3 this gives the desired estimate.

When \(f\) is given by (11) we obtain Theorem 1. When \(f\) is given by (19) we have
\[
(1 - f(p))(1 - 1/p)^{-k} = (1 - \nu_p(h^*)/(1 - 1/p)^{-k-1}
\]
and so
\[
\prod_{p} (1 - f(p))(1 - 1/p)^{-k} = \mathfrak{G}_{k+1}(h^*)
\]
which establishes the first part of Theorem 2.

It remains to deal with the case when \(h_0 \in h\), so that (16) holds. Then \(f(p) = (\nu_p(h) - 1)/(p - 1) = (k - 1)/(p - 1)\) for large \(p\). Thus for large \(p\) we have \(f(p) = k'/p - 1\) with \(k' = k - 1\). Now our \(m = k + l = k' + l + 1\), so we also have to replace \(l\) by \(l = l + 1\).

Also, generally we have \(f(p) \leq k'/(p - 1)\) so the above analysis, and in particular Lemma 3, can be applied in the same way (since \(k \geq 3\) we have \(k' \geq 2\)). Then we obtain
\[
\sum_{n \leq N} \theta(n+h_0) \Lambda_R(n; h, l)^2 = \left( \frac{2l'}{l'} \right) \prod_{p} (1 - f(p))(1 - 1/p)^{-k'} \frac{(\log R)^{k'+2l'}}{(k'+2l')!} N(\log R)^{k'+2l'} + O(N(\log N)^{k'+2l'-1}(\log \log N)^2).
\]
Here \(k' + 2l' = k + 2l + 1\) and, by (16),
\[
(1 - f(p))(1 - 1/p)^{-k'} = \frac{p - \nu_p(h)}{p - 1}(1 - 1/p)^{1-k} = (1 - \nu_p(h)/p)(1 - 1/p)^{-k}
\]
and the second part of Theorem 2 follows.
3 Averages of the singular series

It is necessary to have some control of the size of $\mathcal{S}_k(h)$ and the simplest way is to average it.

**Theorem 4 (Gallagher).** Suppose that $k \geq 2$. Then

$$
\sum_{h \in \mathcal{H}} \mathcal{S}_k(h) = H^k + O(H^{k-1+\varepsilon}).
$$

**Proof.** By (7) we have

$$
\mathcal{S}_k(h) = \sum_{q=1}^{\infty} \mu(q)^2 F(q; h)
$$

where $F(q; h)$ is a multiplicative function of $q$ and

$$
F(p; h) = \left(1 - \frac{\nu_p(h)}{p}\right) \left(1 - \frac{1}{p}\right)^{-k} - 1.
$$

When $\nu_p(h) = k$ we have

$$
|F(p; h)| \leq \frac{C_k}{p^2}
$$

and otherwise

$$
|F(p; h)| \leq \frac{C_k}{p}
$$

where $C_k$ is a suitable positive number. Let $D = \prod_{1 \leq i < j \leq k} |h_j - h_i|$, so that $D \leq H^{k(k-1)/2}$. Then for squarefree $q,$

$$
|F(q; h)| \leq q^{-2} C_k^\omega(q) (D, q) \ll \varepsilon q^{\varepsilon-2} (D, q).
$$

For convenience we introduce the parameter $Q \geq 1$ which is at our disposal. Then

$$
\sum_{q>Q} \mu(q)^2 |F(q; h)| \ll \sum_{r|D} r \sum_{q>Q} q^{\varepsilon-2} \ll \sum_{r|D} r^{\varepsilon-1} \sum_{t>Q/r} t^{\varepsilon-2} \ll Q^{\varepsilon-1} d(D).
$$

Hence

$$
\sum_{q>Q} \mu(q)^2 |F(q; h)| \ll Q^{\varepsilon-1} H^{\varepsilon}.
$$

There is a different representation of $F(q; h)$ which is useful when $1 < q \leq Q$. Let

$$
G(q; h) = \sum_{(a, q) = 1}^q c_q(a_2) \cdots c_q(a_k) c_q(-a_2 - \cdots - a_k) e((a_2 h_2 - h_1) + \cdots + a_k (h_k - h_1))/q
$$

(26)
where \( a = a_2, \ldots, a_k \) and the sum is over \( a \) satisfying \( 1 \leq a_j \leq q \) and \( (a_2, \ldots, a_k, q) = 1 \), and where \( c_q(a) \) denotes Ramanujan’s sum

\[
c_q(a) = \sum_{\substack{r=1 \\ (r,q)=1}}^{q} e(ar/q).
\]

It can be verified that \( G(q; h) \) is a multiplicative function of \( q \). Moreover if \( p^2 | q \), then corresponding to each summand in \( G \) there is an \( a_j \) such that \( p \nmid a_j \), and then \( c_q(a_j) = 0 \). Thus \( G \) has its support on the squarefree numbers. Moreover

\[
G(p; h) = G^*(p; h) - (p - 1)^k
\]

where

\[
G^*(p; h) = \sum_{a_2=1}^{p} \cdots \sum_{a_k=1}^{p} c_p(a_2) \cdots c_p(a_k) c_p(-a_2 - \cdots - a_k) e\left((a_2(h_2 - h_1) + \cdots + a_k(h_k - h_1))/p\right).
\]

This is

\[
p^{k-1} M
\]

where \( M \) is the number of solutions of \( r_1 + h_1 \equiv r_2 + h_2 \equiv \ldots \equiv r_k + h_k \) (mod \( p \)) with \( 1 \leq r_i \leq p - 1 \). Suppose exactly \( j \) of the \( h \) are distinct modulo \( p \), so \( j = \nu_p(h) \). For sake of argument we may suppose that \( h_1, \ldots, h_j \) are distinct modulo \( p \) and then \( M \) is the number of solutions of \( r_1 \equiv r_2 + h_2 - h_1 \equiv \ldots \equiv r_j + h_j - h_1 \) (mod \( p \)) with \( 1 \leq r_i \leq p - 1 \). Clearly \( 0, h_2 - h_1, \ldots, h_j - h_1 \) are distinct modulo \( p \) and the condition on \( r_1 \) for solubility is that \( r_1 \) is excluded from these residue classes. Since then the remaining \( r_i \) are determined it follows that \( M = p - j = p - \nu_p(h) \). Hence

\[
G(p; h) = p^{k-1}(p - \nu_p(h)) - (p - 1)^k = (p - 1)^k F(p; h).
\]

Hence

\[
\mu(q)^2 F(q; h) = \frac{G(q; h)}{\phi(q)^k}.
\]

(27)

The case \( k = 2 \) is somewhat special so we treat that first. By (26),

\[
G(q; h) = \sum_{\substack{a=1 \\ (a,q)=1}}^{q} \mu(q)^2 e(a(h_1 - h_2)/q)
\]
and so $\sum_{h \in \mathcal{H}} G(q; h) = \mu(q)^2 \sum_{h_2 \leq H} \sum_{q}^{\infty} \sum_{a=1}^{q} \sum_{h_1 \leq H} e(a(h_1 - h_2)/q)$. The innermost sum is $\ll ||a/q||^{-1}$ and we have $\sum_{a=1}^{q-1} ||a/q||^{-1} \ll q \log q$. Thus

$$\sum_{h \in \mathcal{H}} \sum_{1 < q \leq Q} \frac{G(q, h)}{\phi(q)^2} \ll HQ^\varepsilon.$$ 

The case $k = 2$ of the theorem now follows from (25) and (27) with $Q = H$.

For the rest of the proof we suppose that $k \geq 3$. Crudely

$$|G(q; h)| \leq G^*(q)$$

where

$$G^*(q) = \sum_{(a,q) = 1} |c_q(a_2) \ldots c_q(a_k)c_q(-a_2 - \cdots - a_k)|$$

and this is also a multiplicative function of $q$ (with its support on the square free numbers). Consider the $k$ numbers $a_2, \ldots, a_k, a_2 - a_3 - \cdots - a_k$. When $(a, p) = 1$ at least two of these numbers are not multiples of $p$. Moreover in $G^*(p)$ the terms with exactly $j$ of the $a_2, \ldots, a_k, a_2 + \cdots + a_k$ divisible by $p$ contribute $(p-1)^j$ and since the $a_2, \ldots, a_k, a_2 + \cdots + a_k$ are linearly dependent the number of such terms is at most $\binom{k}{j}(p-1)^{k-1-j}$. Hence $G^*(p) \leq 2^{k}(p-1)^{k-1}$ and $G^*(q)\phi(q)^{-k} \ll q^{\varepsilon-1}$. Hence

$$\sum_{h \in \mathcal{H}} \sum_{1 < q \leq Q} \frac{G(q; h)}{\phi(q)^k} = \sum_{h \in [1,H]^k} \sum_{1 < q \leq Q} \frac{G(q; h)}{\phi(q)^k} \ll H^{k-1} \sum_{q \leq Q} q^{\varepsilon-1}$$

and so

$$\sum_{h \in \mathcal{H}} \sum_{1 < q \leq Q} \frac{G(q; h)}{\phi(q)^k} - \sum_{h \in [1,H]^k} \sum_{1 < q \leq Q} \frac{G(q; h)}{\phi(q)^k} \ll H^{k-1}Q^\varepsilon. \quad (28)$$

Returning to (26) when $q > 1$ at least two of $a_2, \ldots, a_k, a_2 - \cdots - a_k$ are non-zero (mod $q$). If we pick any two such of the $a_i$ and call them $b_1, b_2$ the remaining $a_i$ can be listed in the form $b_3, \ldots, b_{k-1}, -b_1 - b_2 - \cdots - b_{k-1}$. If $b_1, b_2$ are among the $a_i$, then this is obvious. If one of $b_1, b_2$ is $-a_2 - \cdots - a_k$ then we can write any one of the $a_i$ not amongst the $b_1, b_2$ in the form $-b_1 - b_2 - s$ (mod $q$) where $s$ is the sum of the remaining $a_i$. Thus

$$\sum_{h \in [1,H]^k} G(q, h) \ll$$

$$H^{k-2} \sum_{b_1=1}^{q-1} \frac{|c_q(b_1)|}{\|b_1/q\|} \sum_{b_2=1}^{q-1} \frac{|c_q(b_2)|}{\|b_2/q\|} \sum_{b \in [1,q]^{k-3}} |c_q(b_3) \ldots c_q(b_{k-1})c_q(b_1 + \cdots + b_{k-1})|$$
where \( b = b_3, \ldots, b_{k-1} \) and where the sum over \( b \) is taken to be \(|c_q(b_1 + b_2)|\) when \( k = 3 \). In general this sum does not exceed

\[
\phi(q) \left( \sum_{b=1}^{q} |c(b)| \right)^{k-3}
\]

Since \(|c_q(b)| \leq (q, b)\) the sum here is at most

\[
\sum_{r|q} r\phi(q/r) \leq d(q)q.
\]

Similarly

\[
\sum_{b=1}^{q-1} \frac{|c_q(b)|}{\|b/q\|} \leq \sum_{r|q} r^{q/r-1} \sum_{a=1}^{q/r-1} \|a/(q/r)\|^{-1} \ll d(q)q \log q.
\]

Therefore

\[
\sum_{h \in [1,H]} \sum_{1 \leq q \leq Q} \frac{G(q, h)}{\phi(q)^k} \ll H^{k-2}Q^{1+\varepsilon}.
\]

Hence, by (25) and (28) the choice \( Q = H \) secures the theorem.

4 The main theorem

**Theorem 5.** Suppose that \( \varepsilon \) is a sufficiently small positive number. There are \( k = k(\varepsilon), \ l = l(\varepsilon) \) such that if \( N > N_0(\varepsilon), \ R = N^{\frac{1}{4} - \varepsilon}, \ H = 10\varepsilon \log N \) and

\[
S = \sum_{n=2N}^{2N} \left( \sum_{1 \leq h_0 \leq H} \left( \theta(n + h_0) - \log(3N) \right) \sum_{h \in H} \Lambda_R(n; h, l)^2, \right)
\]

then

\[
S > 0.
\]

The positivity of \( S \) implies that for arbitrarily large \( N \) there are \( n \in [N + 1, 2N] \) such that

\[
\sum_{1 \leq h_0 \leq H} \theta(n + h_0) > \log(3N).
\]

Since \( \theta(n + h_0) \leq \log(2N + h_0) < \log(3N) \) there must be two values of \( h_0 \) in \([1, H]\) such that \( \theta(n + h_0) > 0 \), i.e. \( n + h_0 \) is prime. Hence there are primes \( p', p \) with \( N < p < p' \leq 3N \) such that

\[
\frac{p' - p}{\log p} \leq \frac{H}{\log N} < 10\varepsilon.
\]
Corollary 6. We have
\[
\liminf_{n \to \infty} \frac{p_{n+1} - p_n}{\log p_n} = 0.
\]

Proof of Theorem 5. By Theorems 1 and 2,
\[
S = S_1 + S_2 - S_3 + O_{k,l} \left( N (\log N)^{2k+2l} (\log \log N)^2 \right)
\]
where
\[
S_1 = \frac{k}{(k + 2l + 1)!} \binom{2l + 2}{l + 1} N (\log R)^{k+2l+1} \sum_{h \in \mathcal{H}} \mathcal{S}_k(h),
\]
\[
S_2 = \frac{1}{(k + 2l)!} \binom{2l}{l} N (\log R)^{k+2l} \sum_{h^* \in \mathcal{H}^*} \mathcal{S}_{k+1}(h^*),
\]
\[
S_3 = \frac{1}{(k + 2l)!} \binom{2l}{l} N (\log 3N)(\log R)^{k+2l} \sum_{h \in \mathcal{H}} \mathcal{S}_k(h)
\]
and \(\mathcal{H}^*\) is the set of \(h = h_0, \ldots, h_k\) with the \(h_j\) distinct and satisfying \(1 \leq h_j \leq H\). By Theorem 4 with \(R = N^{1/4-\varepsilon}\) and \(H = 10\varepsilon \log N\) this gives
\[
S = S_4 + O_{k,l} \left( N (\log N)^{2k+2l} (\log \log N)^2 \right)
\]
where
\[
S_4 = \frac{1}{(k + 2l)!} \binom{2l}{l} \left( \frac{2k(2l + 1)}{(k + 2l + 1)(l + 1)} \left( \frac{1}{4} - \varepsilon \right) + 10\varepsilon - 1 \right) NH^k (\log N)(\log R)^{k+2l}.
\]
The choice \(k = l^2\), \(l = \lceil 1/\varepsilon \rceil\) gives
\[
\frac{2k(2l + 1)}{(k + 2l + 1)(l + 1)} \left( \frac{1}{4} - \varepsilon \right) + 10\varepsilon - 1 = \frac{7}{2} \varepsilon + O(\varepsilon^2).
\]

References

Goldston, D. A., Pintz, J., Yildirim, C. Y. submitted, Primes in tuples II.

RCV: Department of Mathematics, McAllister Building, Pennsylvania State University, University Park, PA 16802-6401, U.S.A.
E-mail address: rvvaughan@math.psu.edu