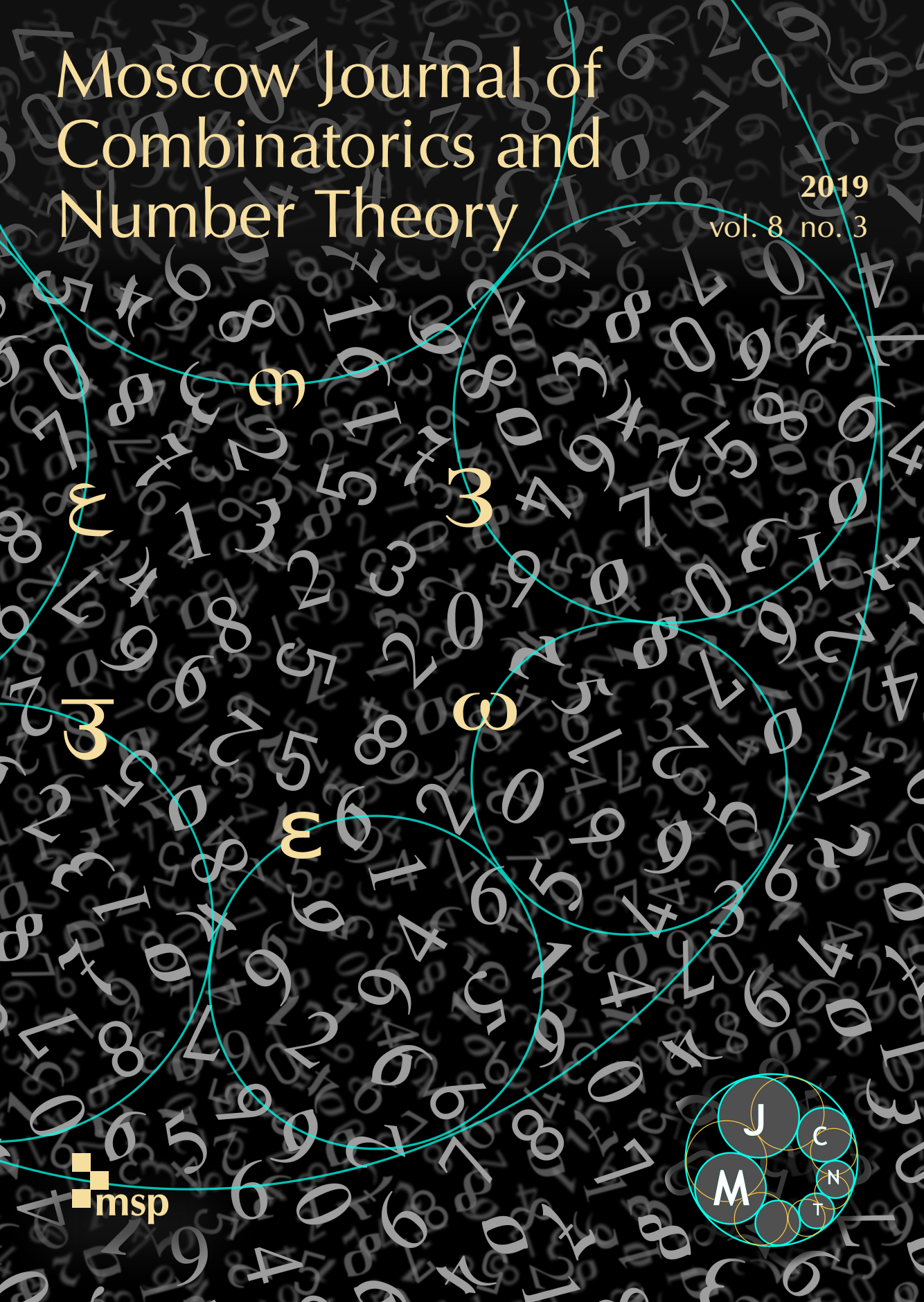
Moscow Journal of Combinatorics and Number Theory

2019
vol. 8 no. 3



msp

Moscow Journal of Combinatorics and Number Theory

msp.org/moscow

See inside back cover or msp.org/moscow for submission instructions.

The subscription price for 2019 is US $310/year for the electronic version, and $365/year (+$20, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Moscow Journal of Combinatorics and Number Theory (ISSN 2640-7361 electronic, 2220-5438 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

MJCNT peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 3, 2019
dx.doi.org/10.2140/moscow.2019.8.193

Integer complexity: the integer defect

Harry Altman

Define $\|n\|$ to be the *complexity* of n, the smallest number of 1s needed to write n using an arbitrary combination of addition and multiplication. John Selfridge showed that $\|n\| \geq 3\log_3 n$ for all n, leading this author and Zelinsky to define the *defect* of n, $\delta(n)$, to be the difference $\|n\| - 3\log_3 n$. Meanwhile, in the study of addition chains, it is common to consider $s(n)$, the number of small steps of n, defined as $\ell(n) - \lfloor \log_2 n \rfloor$, an integer quantity, where $\ell(n)$ is the length of the shortest addition chain for n. So here we analogously define $D(n)$, the *integer defect* of n, an integer version of $\delta(n)$ analogous to $s(n)$. Note that $D(n)$ is not the same as $\lceil \delta(n) \rceil$.

We show that $D(n)$ has additional meaning in terms of the defect well-ordering we considered in 2015, in that $D(n)$ indicates which powers of ω the quantity $\delta(n)$ lies between when one restricts to n with $\|n\|$ lying in a specified congruence class modulo 3. We also determine all numbers n with $D(n) \leq 1$, and use this to generalize a result of Rawsthorne (1989).

1. Introduction

The *complexity* of a natural number n, denoted by $\|n\|$, is the least number of 1s needed to write it using any combination of addition and multiplication, with the order of the operations specified using parentheses grouped in any legal nesting. For instance, $n = 11$ has a complexity of 8, since it can be written using eight 1s as

$$11 = (1+1+1)(1+1+1)+1+1,$$

but not with any fewer than eight. More formally, $\|n\|$ is the number of leaves in the smallest arithmetic formula for n, using addition and multiplication as the gates and 1 as the leaves. This notion was implicitly introduced by Kurt Mahler and Jan Popken [1953], and was later popularized by Richard Guy [1986; 1987, p. 965; 1989, p. 905; 2004, pp. 399–400].

Integer complexity is approximately logarithmic; it satisfies the bounds

$$3\log_3 n = \frac{3}{\log 3}\log n \leq \|n\| \leq \frac{3}{\log 2}\log n, \quad n > 1.$$

The lower bound can be deduced from the results of Mahler and Popken, and was explicitly proved by John Selfridge [Guy 1986]. It is attained with equality for $n = 3^k$ for all $k \geq 1$. The upper bound can be obtained by writing n in binary and finding a representation using Horner's algorithm. It is not sharp, and the constant $3/\log 2$ can be improved for large n [Zelinsky $\geq$ 2019].

MSC2010: 11A67.
Keywords: integer complexity, well-ordering, arithmetic formulas.

Based on the lower bound, this author and Zelinsky [Altman and Zelinsky 2012] introduced the notion of the *defect* of n, denoted by $\delta(n)$, which is the difference $\|n\| - 3\log_3 n$. Subsequent work [Altman 2015] showed that the set of defects is in fact a well-ordered subset of the real line, with order type ω^ω.

However, it is worth considering the result of Selfridge in more detail:

Theorem 1.1 (Selfridge). *For any $k \geq 1$, let $E(k)$ be the largest number that can be made with k 1s, i.e., the largest n with $\|n\| \leq k$. Then*:

(1) *If $k = 1$, then $E(k) = 1$.*

(2) *If $k \equiv 0 \pmod 3$, then $E(k) = 3^{k/3}$.*

(3) *If $k \equiv 1 \pmod 3$ and $k > 1$, then $E(k) = 4 \cdot 3^{(k-4)/3}$.*

(4) *If $k \equiv 2 \pmod 3$, then $E(k) = 2 \cdot 3^{(k-2)/3}$.*

(This result is also a special case of the results of [Mahler and Popken 1953].) From this one can of course derive the lower bound $\|n\| \geq 3\log_3 n$, but what if one wanted an integer version? We make the following definition:

Definition 1.2. Given a natural number n, we define $L(n)$ to be the largest k such that $E(k) \leq n$.

With this, we define:

Definition 1.3. For a natural number n, we define the *integer defect* of n, denoted by $D(n)$, to be the difference $\|n\| - L(n)$.

Because of Theorem 1.1, $L(n)$ is quite easy to compute (see Proposition 3.8), and hence if one knows $\|n\|$ then $D(n)$ is also easy to compute. Note that while we consider $D(n)$ to be an integer analogue of $\delta(n)$, it is not in general equal to $\lceil \delta(n) \rceil$; see Theorem 3.12 for the precise relation. However it is not immediately obvious that $D(n)$ has any actual significance. But, in fact, the integer defect of a number tells you about its position in the well-ordering of defects.

Remark 1.4. $L(k)$ is not the best lower bound we can get from Theorem 1.1; that would instead be the smallest k such that $E(k) \geq n$, which we might denote $L'(n)$. ($L'(n)$ could also be defined as the minimum of $\|m\|$ over all $m \geq n$.) For reasons that will become clear later, though, we will prefer to discuss L rather than L'. In any case, $L'(n) = L(n) + 1$ unless $n = E(k)$ for some k, in which case $L'(n) = L(n) = k$, so one can easily convert any results expressed in the one formulation to the other. One could consider a similar $D'(n)$ as well, but we will not do that either.

1A. ***The sets $\mathscr{D}^0$, $\mathscr{D}^1$, and $\mathscr{D}^2$ and the main result.*** As has been noted above, if we define $\mathscr{D}$ to be the set of all defects, then as a subset of the real line this set is well-ordered and has order type ω^ω. However, more specific theorems are proved in [Altman 2015]. We will need the following definition:

Definition 1.5. If a is a congruence class modulo 3, we define

$$\mathscr{D}^a = \{\delta(n) : \|n\| \equiv a \pmod 3,\ n \neq 1\}.$$

Remark 1.6. The number $n = 1$ is excluded from $\mathscr{D}^1$ because it is dissimilar to other numbers whose complexity is congruent to 1 modulo 3. Unlike other numbers which are 1 modulo 3, the number 1 cannot be written as $3j + 4$ for some $j \geq 0$, and so the largest number that can be made with a single 1 is simply 1, rather than $4 \cdot 3^j$.

In fact the sets $\mathscr{D}^a$ for $a = 0, 1, 2$ are disjoint, and so together with $\{1\}$ form a partition of $\mathscr{D}$.

Moreover in [Altman 2015] it was proved:

Theorem 1.7. *For $a = 0, 1, 2$, the sets $\mathscr{D}^a$ are all well-ordered, each with order type ω^ω.*

$D(n)$ will tell us about the position of $\delta(n)$ in these sets, the $\mathscr{D}^a$. We show:

Theorem 1.8 (Main theorem). *Let $n > 1$ be a natural number. Let ζ be the order type of $\mathscr{D}^{\|n\|} \cap [0, \delta(n))$. Then $D(n)$ is equal to the smallest k such that $\zeta < \omega^k$.*

As mentioned above, $D(n)$ is easy to compute, so this theorem gives a way to easily compute around where $\delta(n)$ falls in the ordering on $\mathscr{D}^a$.

We will also prove a version of this theorem for the stable integer defect; see Sections 1D and 3.

It is worth comparing this theorem to what was already known. It was proved in [Altman 2015] that the limit of the initial ω^k elements of $\mathscr{D}$ is equal to k. This raises the question — just what is the limit of the initial ω^k elements of $\mathscr{D}^a$? It was further shown in [Altman 2015] that when $k \equiv a \pmod 3$ this limit is equal to k, but what about otherwise?

In this paper we will answer this question:

Theorem 1.9. *The limit of the initial ω^k elements of $\mathscr{D}^a$ is equal to k if $k - a \equiv 0 \pmod 3$, it is equal to $k + \delta(2)$ if $k - a \equiv 1 \pmod 3$, and it is equal to $k + 2\delta(2)$ if $k - a \equiv 2 \pmod 3$.*

In fact, Theorem 1.9 will be used to prove Theorem 1.8. See Section 4 for more general statements. Further generalizations will appear in a future paper [Altman and Arias de Reyna ≥ 2019].

1B. ***Generalizing Rawsthorne's theorem.*** We know how to compute $E(k)$, the highest number of complexity at most k (or exactly k), but what about the next highest? This question was answered by Daniel Rawsthorne [1989]:

Theorem 1.10 (Rawsthorne). *For any $k \geq 8$, the highest number of complexity at most k other than $E(k)$ itself is $\frac{8}{9}E(k)$, and this number has complexity exactly k.*

In this paper we generalize this result. First, a definition:

Definition 1.11. Given $r \geq 0$ and $k \geq 1$, we define $E_r(k)$ to be the r-th largest number of complexity at most k. We will 0-index here, so that by definition $E_0(k) = E(k)$, and Theorem 1.10 gives a formula for $E_1(k)$.

Then, with this, we show:

Theorem 1.12. *Given $r \geq 0$, and a, a congruence class modulo 3, there exists $K_{r,a} > 1$ and $h_{r,a} \in \mathbb{Q}$ such that for $k \geq K_{r,a}$, with $k \equiv a \pmod 3$, we have $E_r(k) = h_{r,a}E(k)$, and these $h_{r,a}$ and $K_{r,a}$ are as given by Tables 1, 2, and 3. Moreover, for such r and k, we have $E_r(k) > E(k-1)$ and therefore $\|E_r(k)\| = k$ (and thus for such r and k, $E_r(k)$ is not just the r-th largest number with complexity at most k, but the r-th largest number with complexity exactly k).*

Note that Tables 1, 2, and 3 do not list the regular pattern in the $h_{r,a}$ until such point as $K_{r,a}$ also becomes regular; for tables based solely on $h_{r,a}$, see Tables 4, 5, and 6.

What does Theorem 1.12 have to do with integer defect? Well, the numbers $h_{r,a}E(k)$ appearing in this theorem are almost exactly the numbers n with $D(n) \leq 1$; see Proposition 5.6 for a precise statement.

r	$h_{r,0}$	$K_{r,0}$
0	1	3
1	8/9	6
2	64/81	12
3	7/9	12
4	20/27	12
5	19/27	12
6	512/729	18
7	56/81	18
8	55/81	18
9	164/243	18
10	163/243	18
(for $n \geq 6$) $2n-1$	$2/3+2/3^n$	$3n$
(for $n \geq 6$) $2n$	$2/3+1/3^n$	$3n$

Table 1. Table of h_r and K_r for $k \equiv 0 \pmod 3$.

r	$h_{r,2}$	$K_{r,2}$
0	1	2
1	8/9	8
2	5/6	8
3	64/81	14
4	7/9	14
5	20/27	14
6	13/18	14
7	19/27	14
8	512/729	20
9	56/81	20
10	37/54	20
11	55/81	20
12	164/243	20
13	109/162	20
14	163/243	20
(for $n \geq 6$) $3n-3$	$2/3+2/3^n$	$3n+2$
(for $n \geq 6$) $3n-2$	$2/3+1/(2\cdot 3^{n-1})$	$3n+2$
(for $n \geq 6$) $3n-1$	$2/3+1/3^n$	$3n+2$

Table 2. Table of h_r and K_r for $k \equiv 2 \pmod 3$.

After all, by Theorem 1.8, the numbers n with $D(n) \leq 1$ are precisely those n whose $\delta(n)$ lie in the initial ω of $\mathscr{D}^{\|n\|}$. So if one fixes a particular k, then going down the set of n with $\|n\| = k$ corresponds to going up the set of defects $\delta(n)$ of n with $\|n\| = k$, and assuming k is large enough relative to how far

r	$h_{r,1}$	$K_{r,1}$
0	1	4
1	8/9	10
2	5/6	10
3	64/81	16
4	7/9	16
5	41/54	16
(for $n \geq 4$) $n+2$	$3/4+1/(4\cdot 3^n)$	$3n+4$

Table 3. Table of h_r and K_r for $k \equiv 1 \pmod 3$ with $k > 1$.

up or down you want to go, this is just looking at $\mathscr{D}^k$. And if we count up one at a time, then — again, assuming k is sufficiently large relative to how far out we count — we will stay within the initial ω of $\mathscr{D}^k$. So with a classification of numbers n such that $D(n) \leq 1$, one can determine the $E_r(k)$. (Indeed, one can also do the reverse.)

Note that Theorem 1.10 also works for $k = 6$, so if one wants to break it down by the residue of k modulo 3, one could say it works for $k \geq 6$ with $k \equiv 0 \pmod 3$, for $k \geq 8$ with $k \equiv 2 \pmod 3$, and for $k \geq 10$ with $k \equiv 1 \pmod 3$. (Indeed, this is what we have done in Tables 1, 2, and 3.) Note how all three of these correspond to k exactly large enough for $E(k)$ to be divisible by 9, as per the last part of Theorem 1.12.

One thing worth noting here is that the formulae for $E_0(k)$ and $E_1(k)$, as originally proven by Selfridge and Rawsthorne respectively, were both originally proven directly by induction on k, whereas here we have proven Theorem 1.12 by a different method, namely, analysis of defects (although this analysis of defects in turn depends on Rawsthorne's formula for $E_1(k)$ to serve as a base case; see [Altman and Zelinsky 2012]). This raises the question of whether a similar inductive proof for general $E_r(k)$ could be done now that the formulae for them are known. (In fact this author originally proved these formulae by a different method entirely, that of analyzing certain transformations of expression, so other methods certainly are possible.)

1C. ***Low-defect polynomials and numbers of low defect.*** In order to prove Theorem 1.8, we make use of the idea of low-defect polynomials from [Altman 2015; 2016]. A low-defect polynomial is a particular type of multilinear polynomial; see Section 2 for details. In [Altman 2015] it is proved that, given any positive real number s, one can write down a finite set of low-defect polynomials $\mathcal{S}$ such that every number n with $\delta(n) < s$ can be written in the form $f(3^{n_1}, \ldots, 3^{n_d})3^{n_{d+1}}$ for some $f \in \mathcal{S}$, and that, moreover, such an n can always be represented "efficiently" in such a fashion. Additionally, one can choose $\mathcal{S}$ such that for any $f \in \mathcal{S}$, one has $\deg f \leq s$. (Note that the degree of a low-defect polynomial is always equal to the number of variables it is in, since low-defect polynomials are multilinear and always include a term containing all the variables.)

Using this fact about low-defect polynomials, this author proved in [Altman 2015] that the set $\mathscr{D}$ is well-ordered with order type ω^ω, as well as the more specific Theorem 1.7 mentioned above, and other results mentioned above such as that the limit of the initial ω^k defects is equal to k. However, this is not enough to prove the more specific theorems shown in this paper, such as Theorem 1.9. But in [Altman

2016] an improvement was shown, that we can in fact take $\mathcal{S}$ such that for all $f \in \mathscr{T}$, one has $\delta(f) \le s$; here $\delta(f)$ is a number that bounds $\delta(n)$ above for any n represented by f in the fashion described above; again, see Section 2 for more on this.

On top of that, it was shown in [Altman 2016] that $\delta(f) \ge \deg f + \delta(m)$, where m is the leading coefficient of f. Putting this together, one gets the inequality

$$\deg f + \delta(m) \le s.$$

It is this stronger inequality that allows us to prove Theorem 1.8, where the inequality $\deg f \le s$ would not be enough. To see why this inequality is so helpful, say we are given s and we pick $\mathcal{S}$ as described above. Then if $f \in \mathcal{S}$, one of two things must be true: either $\deg f < \lfloor s \rfloor$, in which case f does not make much of a contribution to $\mathscr{D} \cap [0, s)$ compared to polynomials of higher degree, or $\deg f = \lfloor s \rfloor$, in which case $\delta(m)$ is at most the fractional part of s, a number which is less than 1. Since there are only finitely many defects below any given number less than 1, this puts substantial constraints on m and therefore on f, in ways that the weaker inequality $\deg f \le s$ does not. This allows us to prove Theorem 1.9.

Note that the method we use to turn the results of [Altman 2016] into Theorem 1.8 actually has much more power than we use in this paper, but an exploration of the full power of this method would take us too far away from the subject of $D(n)$, and so will be detailed in a future paper [Altman and Arias de Reyna ≥ 2019].

1D. ***A quick note on stabilization.*** An important property satisfied by integer complexity is the phenomenon of stabilization. Because one has $\|3^k\| = 3k$ for $k > 1$, as well as $\|2 \cdot 3^k\| = 2+3k$ and $\|4 \cdot 3^k\| = 4+3k$, one might hope that in general the equation $\|3n\| = \|n\| + 3$ holds for all $n > 1$. Unfortunately that is not the case; for instance, for $n = 107$, one has $\|107\| = 16$, but $\|321\| = 18$. Another counterexample is $n = 683$, for which one has $\|683\| = 22$, but $\|2049\| = 23$. There are even cases where $\|3n\| < \|n\|$, such as $n = 4721323$, which has $\|3n\| = \|n\| - 1$.

And yet the initial hope is not entirely in vain. In [Altman and Zelinsky 2012], it was proved:

Theorem 1.13. *For any natural number n, there exists $K \ge 0$ such that, for any $k \ge K$,*

$$\|3^k n\| = 3(k - K) + \|3^K n\|.$$

Based on this, we define:

Definition 1.14. A number m is called *stable* if $\|3^k m\| = 3k + \|m\|$ holds for every $k \ge 0$. Otherwise it is called *unstable*.

So, we can restate Theorem 1.13 by saying, for any n, there is some K such that $3^K n$ is stable.

This allows us to define stable or stabilized analogues of many of the concepts and discussed above, and prove stabilized analogues of the theorems discussed in Section 1A. See Sections 2A and 3 for the relevant definitions, and Section 4 for the versions of the main theorems generalized to cover the stabilized case as well.

1E. ***Discussion: comparison to addition chains.*** In order to make sense of Theorem 1.8, it is helpful to introduce an analogy to addition chains, a different notion of complexity which is similar in spirit but different in detail. An *addition chain* for n is defined to be a sequence $(a_0, a_1, \ldots, a_r)$ such that $a_0 = 1$, $a_r = n$, and, for any $1 \le k \le r$, there exist $0 \le i, j < k$ such that $a_k = a_i + a_j$; the number r is called the length of the addition chain. The shortest length among addition chains for n, called the

addition chain length of n, is denoted by $\ell(n)$. Addition chains were introduced by H. Dellac [1894] and reintroduced by A. Scholz [1937]; extensive surveys on the topic can be found in [Knuth 1998, Section 4.6.3, pp. 461–485] and [Subbarao 1989].

The notion of addition chain length has obvious similarities to that of integer complexity; each is a measure of the resources required to build up the number n starting from 1. Both allow the use of addition, but integer complexity supplements this by allowing the use of multiplication, while addition chain length supplements this by allowing the reuse of any number at no additional cost once it has been constructed. (That is to say, while integer complexity is a formula model, addition chains are a circuit model.) Furthermore, both measures are approximately logarithmic; the function $\ell(n)$ satisfies

$$\log_2 n \le \ell(n) \le 2\log_2 n.$$

A difference worth noting is that $\ell(n)$ is actually known to be asymptotic to $\log_2 n$, as was proved by Brauer [1939], but the function $\|n\|$ is not known to be asymptotic to $3\log_3 n$; the value of the quantity $\limsup_{n\to\infty} \|n\|/\log n$ remains unknown.

Nevertheless, there are important similarities between integer complexity and addition chains. As mentioned above, the set of all integer complexity defects is a well-ordered subset of the real numbers, with order type ω^ω. We might also define the notion of *addition chain defect*, defined by

$$\delta^\ell(n) := \ell(n) - \log_2 n;$$

for as shown in [Altman 2018b], the well-ordering theorem for integer complexity has an analogue for addition chains:

Theorem 1.15 (Addition chain well-ordering theorem). *Let $\mathscr{D}^\ell$ denote the set $\{\delta^\ell(n) : n \in \mathbb{N}\}$. Then considered as a subset of the real numbers, $\mathscr{D}^\ell$ is well-ordered and has order type ω^ω.*

More commonly, however, it is not $\delta^\ell(n)$ that has been studied, but rather $s(n)$, the number of *small steps* of n, which is defined to be $\ell(n) - \lfloor \log_2 \rfloor$, or equivalently $\lceil \delta^\ell(n) \rceil$. The quantity $D(n)$ that we introduce seems to play a role in integer complexity similar to $s(n)$ in the study of addition chains. Now, unlike with $s(n)$ and $\delta^\ell(n)$, $D(n)$ is not simply $\lceil \delta(n) \rceil$; for instance, $D(56) = 1$ even though $\delta(56) > 1$. (Although Theorem 3.12 will show how $D(n)$ is in a certain sense almost $\lceil \delta(n) \rceil$.) But, there are further analogies.

Analogous to Theorem 1.13, we have (from [Altman 2018b]) the following:

Theorem 1.16. *For any natural number n, there exists $K \ge 0$ such that, for any $k \ge K$,*

$$\ell(2^k n) = (k - K) + \ell(2^K n).$$

So we define a number n to be *ℓ-stable* if, for any k, one has $\ell(2^k n) = k + \ell(n)$; then Theorem 1.16 says that, for any n, there is some K such that $2^K n$ is ℓ-stable. This allows us to formulate a stabilized version of the previous analogy — and of the ones to follow.

In [Altman 2018b], this author conjectured:

Conjecture 1.17. *For each whole number k, $\mathscr{D}^\ell \cap [0, k]$ has order type ω^k.*

In other words, this conjecture states that the limit of the initial ω^k addition chain defects is equal to k. If true, this would mean that $s(n)$ plays the same role for $\mathscr{D}^\ell$ as $D(n)$ does for the $\mathscr{D}^a$, that $s(n)$ is the smallest k such that the order type of $\mathscr{D}^\ell \cap [0, \delta^\ell(n))$ is less than ω^k.

Similarly, based on conjectures in [Altman 2018b], one gets analogies between $D_{\mathrm{st}}(n)$ and $s_{\mathrm{st}}(n)$ and how they determine position in $\mathscr{D}^a_{\mathrm{st}}$ and $\mathscr{D}^\ell_{\mathrm{st}}$, respectively; see Section 3 for definitions of these.

It is worth noting here one important difference between these two cases: in the integer complexity case, we need to split things into congruence classes modulo 3 based on $\|n\|$. This has no analogue in the addition chain case. This comes from a difference in certain fundamental inequalities that these quantities obey. Integer complexity obeys $\|3n\| \le \|n\| + 3$, with equality if and only if $\delta(3n) = \delta(n)$. The addition chain analogue of this is that one has $\ell(2n) \le \ell(n) + 1$, with equality if and only if $\delta^\ell(2n) = \delta^\ell(n)$. The result [Altman 2018b; Altman and Zelinsky 2012] is that if we have two numbers m and n with $\delta^\ell(n) = \delta^\ell(m)$, then one must have $m = 2^k n$ for some $k \in \mathbb{Z}$, and if we have two numbers m and n with $\delta(n) = \delta(m)$, then one must have $m = 3^k n$ for some $k \in \mathbb{Z}$. However in the latter case we must also have $\|m\| \equiv \|n\| \pmod 3$; this is why the sets $\mathscr{D}^a$ are disjoint. In the addition chain case there is no such congruence requirement; $\ell(n)$ and $\ell(m)$ need only be congruent modulo 1, which is no requirement at all, so splitting up $\mathscr{D}^\ell$ in a similar manner does not make sense. The set $\mathscr{D}^\ell$ already covers the one and only congruence class that exists in the addition chain case.

But it is not only our primary theorem but also our secondary theorem here that has an analogue for addition chains, and in this case the analogy does not rely on any conjectures. While the hypothesis that the order type of $\mathscr{D}^\ell \cap [0, k]$ is equal to ω^k remains a conjecture, that this holds for $k \le 2$ — and in particular that it holds for $k = 1$ — was proven in [Altman 2018b]. This means that just as we can look at the first ω elements of each $\mathscr{D}^a$ in order to determine the r-th-highest number of complexity k, we can look at the first ω elements of $\mathscr{D}^\ell$ to determine the r-th-highest number of addition chain length k (or at most k, which in these cases is the same thing). (Again, here k must be sufficiently large relative to r. Also, again here we are using the convention that r starts at 0 rather than 1.)

Specifically, it is an easy corollary of the classification of numbers with $s(n) \le 1$ [Gioia et al. 1962] that:

Theorem 1.18. *For $k \ge r + 1$ (or for $k \ge 0$ when $r = 0$), the r-th-largest number of addition chain length k is $(1/2 + 1/2^{r+1})2^k$.*

Obviously here the fraction $1/2 + 1/2^{r+1}$ plays the role of the h_r, and $r + 1$ plays the role of K_r; unlike with integer complexity, there are no irregularities here, just a single straightforward infinite family. (And note how the analogue of the K_r increases in what is mostly steps of 1, rather than mostly steps of 3 like the actual K_r, because once again with addition chains there is only one congruence class.) For more on the analogy between integer complexity and addition chains, particularly with regard to their sets of defects, one may see [Altman 2016].

Before we continue, it is worth noting that one might also consider a computational model where one allows addition, multiplication, *and* the free reuse of already-constructed intermediates; that is to say, circuits with addition and multiplication, rather than formulas. These are referred to as addition-multiplication chains, and one may see [Bahig 2008] for more on them.

2. Integer complexity, well-ordering, and low-defect polynomials

In this section we summarize the results of [Altman 2015; 2016; Altman and Zelinsky 2012] that we will need later regarding the defect $\delta(n)$, the stable complexity $\|n\|_{\mathrm{st}}$ and stable defect $\delta_{\mathrm{st}}(n)$ described below, and low-defect polynomials.

2A. ***The defect and stability.*** First, some basic facts about the defect:

Theorem 2.1. *We have*:

(1) *For all* n, $\delta(n) \geq 0$.

(2) *For* $k \geq 0$, $\delta(3^k n) \leq \delta(n)$, *with equality if and only if* $\|3^k n\| = 3k + \|n\|$. *The difference* $\delta(n) - \delta(3^k n)$ *is a nonnegative integer.*

(3) *A number* n *is stable if and only if for any* $k \geq 0$, $\delta(3^k n) = \delta(n)$.

(4) *If the difference* $\delta(n) - \delta(m)$ *is rational, then* $n = m3^k$ *for some integer* k *(and so* $\delta(n) - \delta(m) \in \mathbb{Z}$*).*

(5) *Given any* n, *there exists* k *such that* $3^k n$ *is stable.*

(6) *For a given defect* α, *the set* $\{m : \delta(m) = \alpha\}$ *has either the form* $\{n3^k : 0 \leq k \leq L\}$ *for some* n *and* L, *or the form* $\{n3^k : 0 \leq k\}$ *for some* n. *The latter occurs if and only if* α *is the smallest defect among* $\delta(3^k n)$ *for* $k \in \mathbb{Z}$.

(7) *If* $\delta(n) = \delta(m)$, *then* $\|n\| = \|m\| \pmod 3$.

(8) $\delta(1) = 1$, *and for* $k \geq 1$, $\delta(3^k) = 0$. *No other integers occur as* $\delta(n)$ *for any* n.

(9) *If* $\delta(n) = \delta(m)$ *and* n *is stable, then so is* m.

Proof. Parts (1)–(8), except part (3), are just Theorem 2.1 from [Altman 2015]. Part (3) is Proposition 12 from [Altman and Zelinsky 2012], and part (9) is Proposition 3.1 from [Altman 2015]. □

We will want to consider the set of all defects:

Definition 2.2. We define the *defect set* $\mathscr{D}$ to be $\{\delta(n) : n \in \mathbb{N}\}$, the set of all defects.

We also defined $\mathscr{D}^a$, for a a congruence class modulo 3, in Definition 1.5 earlier.

The paper [Altman 2015] also defined the notion of a *stable defect*:

Definition 2.3. We define a *stable defect* to be the defect of a stable number, and define $\mathscr{D}_{\text{st}}$ to be the set of all stable defects. Also, for a a congruence class modulo 3, we define $\mathscr{D}^a_{\text{st}} = \mathscr{D}^a \cap \mathscr{D}_{\text{st}}$.

Because of part (9) of Theorem 2.1, this definition makes sense; a stable defect α is not just one that is the defect of some stable number, but one for which any n with $\delta(n) = \alpha$ is stable. Stable defects can also be characterized by the following proposition from [Altman 2015]:

Proposition 2.4. *A defect* α *is stable if and only if it is the smallest* $\beta \in \mathscr{D}$ *such that* $\beta \equiv \alpha \pmod 1$.

We can also define the *stable defect* of a given number, which we denote by $\delta_{\text{st}}(n)$.

Definition 2.5. For a positive integer n, define the *stable defect* of n, denoted by $\delta_{\text{st}}(n)$, to be $\delta(3^k n)$ for any k such that $3^k n$ is stable. (This is well-defined as if $3^k n$ and $3^\ell n$ are stable; then $k \geq \ell$ implies $\delta(3^k n) = \delta(3^\ell n)$, and $\ell \geq k$ implies this as well.)

Note that the statement "α is a stable defect", which earlier we were thinking of as "$\alpha = \delta(n)$ for some stable n", can also be read as the equivalent statement "$\alpha = \delta_{\text{st}}(n)$ for some n".

Similarly we have the stable complexity:

Definition 2.6. For a positive integer n, define the *stable complexity* of n, denoted by $\|n\|_{\text{st}}$, to be $\|3^k n\| - 3k$ for any k such that $3^k n$ is stable.

We then have the following facts relating the notions of $\|n\|$, $\delta(n)$, $\|n\|_{\text{st}}$, and $\delta_{\text{st}}(n)$:

Proposition 2.7. *We have*:

(1) $\delta_{\mathrm{st}}(n) = \min_{k\geq 0} \delta(3^k n)$.

(2) $\delta_{\mathrm{st}}(n)$ *is the smallest* $\alpha \in \mathscr{D}$ *such that* $\alpha \equiv \delta(n) \pmod 1$.

(3) $\|n\|_{\mathrm{st}} = \min_{k\geq 0}(\|3^k n\| - 3k)$.

(4) $\delta_{\mathrm{st}}(n) = \|n\|_{\mathrm{st}} - 3\log_3 n$.

(5) $\delta_{\mathrm{st}}(n) \leq \delta(n)$, *with equality if and only if* n *is stable.*

(6) $\|n\|_{\mathrm{st}} \leq \|n\|$, *with equality if and only if* n *is stable.*

(7) $\|3n\|_{\mathrm{st}} = \|n\|_{\mathrm{st}} + 3$.

(8) *If* $\delta_{\mathrm{st}}(n) = \delta_{\mathrm{st}}(m)$, *then* $\|n\|_{\mathrm{st}} \equiv \|m\|_{\mathrm{st}} \pmod 3$.

Proof. Statements (1)–(6) are just Propositions 3.5, 3.7, and 3.8 from [Altman 2015]. Statement (7) follows from the definition of stable complexity; if $3^k n$ is stable, then $\|3n\|_{\mathrm{st}} = \|3^k n\| - 3(k-1) = \|3^k n\| - 3k + 3 = \|n\|_{\mathrm{st}} + 3$. To prove statement (8), note that if $\delta_{\mathrm{st}}(n) = \delta_{\mathrm{st}}(m)$, then by statement (2) one has $\delta(n) \equiv \delta(m) \pmod 1$, and so by Theorem 2.1, one has that $n = m3^k$ for some $k \in \mathbb{Z}$, and so $\|n\|_{\mathrm{st}} = \|m\|_{\mathrm{st}} + 3k$. □

Note, by the way, that just as $\mathscr{D}_{\mathrm{st}}$ can be characterized either as defects $\delta(n)$ with n stable or as defects $\delta_{\mathrm{st}}(n)$ for any n, $\mathscr{D}^a_{\mathrm{st}}$ can be characterized either as defects $\delta(n)$ with n stable and $\|n\| \equiv a \pmod 3$, or as defects $\delta_{\mathrm{st}}(n)$ for any n with $\|n\|_{\mathrm{st}} \equiv a \pmod 3$.

Three defects that will be particularly important in this paper are the smallest three defects:

Proposition 2.8. $\mathscr{D} \cap [0, 2\delta(2)] = \{0, \delta(2), 2\delta(2)\}$.

Proof. Proposition 37 from [Altman and Zelinsky 2012] tells us that the only leaders with defect less than $3\delta(2)$ are 3, 2, and 4, which respectively have defects 0, $\delta(2)$, and $2\delta(2)$. □

2B. ***Low-defect polynomials.*** As has been mentioned in Section 1C, we are going to represent the set of numbers with defect at most r by substituting powers of 3 into certain multilinear polynomials we call *low-defect polynomials*. We will associate with each one a "base complexity" to form a *low-defect pair*. In this section we will review the basic properties of these polynomials. First, their definition:

Definition 2.9. We define the set $\mathscr{P}$ of *low-defect pairs* as the smallest subset of $\mathbb{Z}[x_1, x_2, \ldots] \times \mathbb{N}$ such that:

(1) For any constant polynomial $k \in \mathbb{N} \subseteq \mathbb{Z}[x_1, x_2, \ldots]$ and any $C \geq \|k\|$, we have $(k, C) \in \mathscr{P}$.

(2) Given (f_1, C_1) and (f_2, C_2) in $\mathscr{P}$, we have $(f_1 \otimes f_2, C_1 + C_2) \in \mathscr{P}$, where, if f_1 is in d_1 variables and f_2 is in d_2 variables,

$$(f_1 \otimes f_2)(x_1, \ldots, x_{d_1+d_2}) := f_1(x_1, \ldots, x_{d_1}) f_2(x_{d_1+1}, \ldots, x_{d_1+d_2}).$$

(3) Given $(f, C) \in \mathscr{P}$, $c \in \mathbb{N}$, and $D \geq \|c\|$, we have $(f \otimes x_1 + c, C + D) \in \mathscr{P}$, where $\otimes$ is as above.

The polynomials obtained this way will be referred to as *low-defect polynomials*. If (f, C) is a low-defect pair, C will be called its *base complexity*. If f is a low-defect polynomial, we will define its

absolute base complexity, denoted by $\|f\|$, to be the smallest C such that (f, C) is a low-defect pair. We will also associate to a low-defect polynomial f the *augmented low-defect polynomial*

$$\hat{f} = f \otimes x_1;$$

if f is in d variables, this is fx_{d+1}.

So, e.g., $(3x_1+1)x_2+1$ is a low-defect polynomial, as are $(3x_1+1)(3x_2+1)$, $(3x_1+1)(3x_2+1)x_3+1$, and $2\big((73(3x_1+1)x_2+6)(2x_3+1)x_4+1\big)$. In this paper we will only concern ourselves with low-defect pairs (f, C) where $C = \|f\|$, so in the remainder of what follows, we will mostly dispense with the formalism of low-defect pairs and just discuss low-defect polynomials.

Note that the degree of a low-defect polynomial is also equal to the number of variables it uses; see Proposition 2.10. Also note that augmented low-defect polynomials are never themselves low-defect polynomials; as we will see in a moment (Proposition 2.10), low-defect polynomials always have nonzero constant term, whereas augmented low-defect polynomials always have zero constant term. We can also observe that low-defect polynomials are in fact read-once polynomials as discussed in for instance [Volkovich 2016].

Note that we do not really care about what variables a low-defect polynomial is in — if we permute the variables of a low-defect polynomial or replace them with others, we will still regard the result as a low-defect polynomial. From this perspective, the meaning of $f \otimes g$ could be simply regarded as "relabel the variables of f and g so that they do not share any, then multiply f and g". Helpfully, the $\otimes$ operator is associative not only with this more abstract way of thinking about it, but also in the concrete way it was defined above.

In [Altman 2015] were proved the following propositions about low-defect polynomials:

Proposition 2.10. *Suppose f is a low-defect polynomial of degree d. Then f is a polynomial in the variables $x_1, \ldots, x_d$, and it is a multilinear polynomial, i.e., it has degree 1 in each of its variables. The coefficients are nonnegative integers. The constant term is nonzero, and so is the coefficient of $x_1 \cdots x_d$, which we will call the **leading coefficient** of f.*

Proof. This is Proposition 4.2 from [Altman 2015]. □

Proposition 2.11. *If f is a low-defect polynomial of degree d, then*

$$\|f(3^{n_1}, \ldots, 3^{n_d})\| \leq \|f\| + 3(n_1 + \cdots + n_d),$$
$$\|\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})\| \leq \|f\| + 3(n_1 + \cdots + n_{d+1}).$$

Proof. This is a combination of Proposition 4.5 and Corollary 4.12 from [Altman 2015]. □

The above proposition motivates the following definition:

Definition 2.12. Given a low-defect polynomial f (say of degree d) and a number N, we will say that f *efficiently* 3*-represents* N if there exist nonnegative integers $n_1, \ldots, n_d$ such that

$$N = f(3^{n_1}, \ldots, 3^{n_d}) \quad \text{and} \quad \|N\| = \|f\| + 3(n_1 + \cdots + n_d).$$

We will say $\hat{f}$ efficiently 3-represents N if there exist $n_1, \ldots, n_{d+1}$ such that

$$N = \hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}}) \quad \text{and} \quad \|N\| = \|f\| + 3(n_1 + \cdots + n_{d+1}).$$

More generally, we will also say f 3-represents N if there exist nonnegative integers $n_1, \dots, n_d$ such that $N = f(3^{n_1}, \dots, 3^{n_d})$, and similarly for $\hat{f}$.

Note that previous papers [Altman 2015; 2016; 2018a] instead spoke of a low-defect pair (f, C) efficiently 3-representing a number N; however, as mentioned in those papers, it is only possible for some (f, C) to efficiently 3-represent a number N if in fact $C = \|f\|$, so there is no loss here.

In keeping with the name, numbers 3-represented by low-defect polynomials, or their augmented versions, have bounded defect. Let us make some definitions first:

Definition 2.13. Given a low-defect polynomial f we define $\delta(f)$, the defect of f, to be $\|f\| - 3\log_3 m$, where m is the leading coefficient of f.

Definition 2.14. Given a low-defect polynomial f of degree d, we define

$$\delta_f(n_1, \dots, n_d) = \|f\| + 3(n_1 + \cdots + n_d) - 3\log_3 f(3^{n_1}, \dots, 3^{n_d}).$$

Then we have:

Proposition 2.15. *Let f be a low-defect polynomial of degree d, and let the numbers $n_1, \dots, n_{d+1}$ be nonnegative integers*:

(1) *We have*

$$\delta(\hat{f}(3^{n_1}, \dots, 3^{n_{d+1}})) \le \delta_f(n_1, \dots, n_d),$$

and the difference is an integer.

(2) *We have*

$$\delta_f(n_1, \dots, n_d) \le \delta(f),$$

and if $d \ge 1$, this inequality is strict.

(3) *The function δ_f is strictly increasing in each variable, and*

$$\delta(f) = \sup_{n_1,\dots,n_d} \delta_f(n_1, \dots, n_d).$$

Proof. This is a combination of Proposition 4.9 and Corollary 4.14 from [Altman 2015] and Proposition 2.15 from [Altman 2016]. □

Importantly, the set of defects coming from a low-defect polynomial of degree r has order type approximately ω^r; if rather than the actual defects we use δ_f, then this is exact. More formally:

Proposition 2.16. *Let f be a low-defect polynomial of degree d. Then*:

(1) *The image of δ_f is a well-ordered subset of $\mathbb{R}$, with order type ω^d.*

(2) *The set of $\delta(N)$ for all N 3-represented by the augmented low-defect polynomial $\hat{f}$ is a well-ordered subset of $\mathbb{R}$, with order type at least ω^d and at most $\omega^d(\lfloor\delta(f)\rfloor + 1) < \omega^{d+1}$. The same is true if f is used instead of the augmented version $\hat{f}$.*

Proof. This is a combination of Propositions 6.2 and 6.3 from [Altman 2015]. □

The second part of the above proposition follows from the first by means of theorems about cutting and pasting of well-ordered sets, ultimately due to [Carruth 1942]. In particular:

Proposition 2.17. *We have*:

(1) *If S is a well-ordered set and $S = S_1 \cup \cdots \cup S_n$, and S_1 through S_n all have order type less than ω^k, then so does S.*

(2) *If S is a well-ordered set of order type ω^k and $S = S_1 \cup \cdots \cup S_n$, then at least one of S_1 through S_n also has order type ω^k.*

Proof. One may see [Carruth 1942] or [de Jongh and Parikh 1977] for proofs of these. □

We will need in particular the following variant:

Proposition 2.18. *Suppose α is an ordinal and S is a well-ordered set which can be written as a finite union $S_1 \cup \cdots \cup S_k$ such that*:

(1) *The S_i all have order types at most ω^α.*

(2) *If a set S_i has order type ω^α, it is cofinal in S.*

Then the order type of S is at most ω^α. In particular, if at least one of the S_i has order type ω^α, then S has order type ω^α.

Proof. A proof of this can be found in [Altman 2018b], where it is Proposition 5.4. □

As was noted above, we have $\delta(f(3^{n_1}, \ldots, 3^{n_d})) \le \delta_f(n_1, \ldots, n_d)$. Importantly, though, for certain low-defect polynomials f, namely, those with $\delta(f) < \deg f + 1$, we can show that equality holds for "most" choices of $(n_1, \ldots, n_d)$ in a certain sense.

Specifically:

Proposition 2.19. *Let f be a low-defect polynomial of degree d with $\delta(f) < d + 1$. Define its "exceptional set" to be*

$$S := \{(n_1, \ldots, n_d) : \|f(3^{n_1}, \ldots, 3^{n_d})\|_{\text{st}} < \|f\| + 3(n_1 + \cdots + n_d)\}.$$

Then the set $\{\delta(f(3^{n_1}, \ldots, 3^{n_d})) : (n_1, \ldots, n_d) \in S\}$ has order type less than ω^d, and therefore so does the set $\{\delta(\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})) : (n_1, \ldots, n_d) \in S\}$. In particular, for $a \not\equiv \|f\| \pmod 3$, the set

$$\{\delta(\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})) : (n_1, \ldots, n_{d+1}) \in \mathbb{Z}_{\ge 0}^{d+1}\} \cap \mathscr{D}^a$$

has order type less than ω^d. Meanwhile, the set

$$\{\delta(f(3^{n_1}, \ldots, 3^{n_d})) : (n_1, \ldots, n_d) \notin S\}$$

has order type at least ω^d, and thus so does the set

$$\{\delta(f(3^{n_1}, \ldots, 3^{n_d})) : (n_1, \ldots, n_d) \in \mathbb{Z}_{\ge 0}^{d}\} \cap \mathscr{D}_{\text{st}}^{\|f\|};$$

moreover, the supremum of this latter set is equal to $\delta(f)$.

Proof. Most of this is direct from Proposition 7.2 from [Altman 2015]; the only parts not covered in the statement there are the statement about $\{\delta(\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})) : (n_1, \ldots, n_d) \in S\}$, the statement regarding $a \not\equiv \|f\| \pmod 3$, and the final statement.

The first of these follows directly from the first part, because

$$\delta(\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})) \le \delta(f(3^{n_1}, \ldots, 3^{n_d}))$$

with the difference being an integer, and that integer can certainly be no more than

$$\delta(f(3^{n_1},\dots,3^{n_d}))\le\delta(f).$$

Thus the set

$$\{\delta(\hat{f}(3^{n_1},\dots,3^{n_{d+1}})):(n_1,\dots,n_{d+1})\in\mathbb{Z}_{\ge0}^{d+1}\}\cap\mathscr{D}^a$$

can be covered by finitely many translates of $\{\delta(f(3^{n_1},\dots,3^{n_d})):(n_1,\dots,n_d)\in S\}$ and therefore by Proposition 2.17 has order type less than ω^d.

For the statement about

$$\{\delta(\hat{f}(3^{n_1},\dots,3^{n_{d+1}})):(n_1,\dots,n_{d+1})\in\mathbb{Z}_{\ge0}^{d+1}\}\cap\mathscr{D}^a,$$

with $a\not\equiv C\pmod 3$, if $\|\hat{f}(3^{n_1},\dots,3^{n_d})\|\equiv a\not\equiv\|f\|\pmod 3$, then in particular this means that

$$\|\hat{f}(3^{n_1},\dots,3^{n_{d+1}})\|\ne\|f\|+3(n_1+\cdots+n_{d+1}),$$

which means that

$$\|\hat{f}(3^{n_1},\dots,3^{n_{d+1}})\|<\|f\|+3(n_1+\cdots+n_{d+1}),$$

and therefore that

$$\|f(3^{n_1},\dots,3^{n_d})\|_{\mathrm{st}}<\|f\|+3(n_1+\cdots+n_d),$$

i.e., that $(n_1,\dots,n_d)\in S$. Applying what was proved in the previous paragraph now proves the statement.

As for the final statement, the set $\{\delta(f(3^{n_1},\dots,3^{n_d})):(n_1,\dots,n_d)\in\mathbb{Z}_{\ge0}^d\}\cap\mathscr{D}_{\mathrm{st}}^{\|f\|}$ contains $\delta_f(\mathbb{N}^d\setminus S)$ (one may see the proof in [Altman 2015]) which in turn contains $\delta_f(\mathbb{N}^d)\setminus\delta_f(S)$. Since the image of δ_f has order type ω^d while $\delta_f(S)$ has order type less than ω^d — similarly to above, this follows by the initial statement and Proposition 2.17 — it follows that $\delta_f(\mathbb{N}^d)\setminus\delta_f(S)$ has order type ω^d and thus is cofinal in the image of δ_f, and thus has supremum $\delta(f)$, and the same is true of the larger set $\{\delta(f(3^{n_1},\dots,3^{n_d})):(n_1,\dots,n_d)\in\mathbb{Z}_{\ge0}^d\}\cap\mathscr{D}_{\mathrm{st}}^{\|f\|}$, which is also bounded above by $\delta(f)$. □

Finally, one more property of low-defect polynomials we will need is the following:

Proposition 2.20. *Let f be a low-defect polynomial, and suppose that a is the leading coefficient of f. Then $\|f\|\ge\|a\|+\deg f$. In particular, $\delta(f)\ge\delta(a)+\deg f$.*

Proof. This is Proposition 3.24 from [Altman 2016]. □

With this, we have the basic properties of low-defect polynomials.

Remark 2.21. Note that one reason nothing is lost here by discarding the formalism of low-defect pairs is that the low-defect pairs (f,C) we will (implicitly) concern ourselves with in this paper are ones that satisfy $C-3\log_3 m<\deg f+1$, where m is the leading coefficient of f. However, by Proposition 2.20,

$$\deg f\le\delta(f)\le C-3\log_3 m<\deg f+1;$$

thus $C-\|f\|=(C-3\log_3 m)-\delta(f)<1$ and so $C=\|f\|$. Thus if we were to use low-defect pairs, we would only be using pairs where $C=\|f\|$, so we lose nothing by making this assumption.

2C. *Good coverings.* We need one more set of definitions before we can state the theorem that will be used as the basis of the proof of the main theorem. We define:

Definition 2.22. A natural number n is called a *leader* if it is the smallest number with a given defect. By part (6) of Theorem 2.1, this is equivalent to saying that either $3 \nmid n$, or, if $3 \mid n$, then $\delta(n) < \delta(n/3)$, i.e., $\|n\| < 3 + \|n/3\|$.

Let us also define:

Definition 2.23. For any real $s \geq 0$, define the set of *s-defect numbers* A_s to be

$$A_s := \{n \in \mathbb{N} : \delta(n) < s\}.$$

Define the set of *s-defect leaders* B_s to be

$$B_s := \{n \in A_s : n \text{ is a leader}\}.$$

These sets are related by the following proposition from [Altman 2015]:

Proposition 2.24. *For every $n \in A_s$, there exists a unique $m \in B_s$ and $k \geq 0$ such that $n = 3^k m$ and $\delta(n) = \delta(m)$; then $\|n\| = \|m\| + 3k$.*

Because of this, if we want to describe the set A_s, it suffices to describe the set B_s. Now we can define:

Definition 2.25. For a real number $s \geq 0$, a finite set $\mathcal{S}$ of low-defect polynomials will be called a *good covering* for B_s if every $n \in B_s$ can be efficiently 3-represented by some polynomial in $\mathcal{S}$ (and hence every $n \in A_s$ can be efficiently represented by some $\hat{f}$ with $f \in \mathcal{S}$) and if for every $f \in \mathcal{S}$, we have $\delta(f) \leq s$, with this being strict if $\deg f = 0$.

This allows us to state the main theorem from [Altman 2016]:

Theorem 2.26. *For any real number $s \geq 0$, there exists a good covering of B_s.*

Proof. This is Theorem 4.9 from [Altman 2016] rewritten in terms of Definition 2.25, and using low-defect polynomials instead of pairs. (Any low-defect pairs (f, C) with $C > \|f\|$, can be filtered out of a good covering, since such a pair can never efficiently 3-represent anything.) □

Note that by Proposition 2.20, if f is in a good covering of B_s with leading coefficient m, we must have $\delta(m) + \deg f \leq s$.

3. The integer defect

In this section we state some basic facts about $D(n)$, what it means, and how it may be computed.

Let us start by giving another interpretation of what $D(n)$ means:

Proposition 3.1. *For a natural number n,*

$$D(n) = |\{k : n < E(k) \leq E(\|n\|)\}|.$$

That is to say, $D(n)$ measures how far down n is among numbers with complexity $\|n\|$, measured by how many values of E one passes as one counts downwards towards n from the largest number also having complexity $\|n\|$.

Proof. By definition, $L(n)$ is the largest k such that $E(k) \leq n$. Since $E(k)$ is strictly increasing, the number of k such that $n < E(k) \leq E(\|n\|)$ is equal to the difference $\|n\| - L(n)$, i.e., $D(n)$. □

So for instance, one has that $D(n) = 0$ if and only if n is of the form $E(k)$ for some k, i.e., n is the largest number of its complexity; while $D(n) \leq 1$ if and only if $n > E(\|n\| - 1)$, i.e., n is greater than all numbers of lower complexity. Numbers n with $D(n) \leq 1$ will be discussed more in Section 5.

As for properties of the integer defect, it behaves largely analogously to the real defect:

Proposition 3.2. *We have*:

(1) *For all* n, $D(n) \geq 0$.

(2) *For all* $n > 1$, $L(3n) = L(n) + 3$.

(3) *For* $n > 1$ *and* $k \geq 0$, *one has* $D(3^k n) \leq D(n)$, *with equality if and only if* $\|3^k n\| = 3k + \|n\|$.

(4) *A number* $n > 1$ *is stable if and only if for any* $k \geq 0$, $D(3^k n) = D(n)$.

Proof. Statement (1) is just the statement that $L(n) \leq \|n\|$; this follows from the definition of $L(n)$ as $E(\|n\|) \geq n$ and so (as $E(k)$ is increasing) one must have $L(n) \leq \|n\|$. And once statement (2) is established, statements (3) and (4) then follow from that and may be proved in exactly the same way their analogous statements in Theorem 2.1 are proved. This leaves just statement (2) to be proved. Note that, for any $k > 1$, $E(k+3) = 3E(k)$. Therefore, for any $k > 1$, $E(k+3) \leq 3n$ if and only if $E(k) \leq n$, and so $L(3n) = L(n) + 3$; the only possible exception to this would be if one had $L(n) = 1$, which happens only when $n = 1$. □

Note that while the theorem that for any n there is some k such that $3^k n$ is stable was originally proven using the defect $\delta(n)$, it could also just as well be proven using the integer defect $D(n)$.

We can also of course define a stable variant of $D(n)$:

Definition 3.3. For a positive integer n, we define the stable integer defect of n, denoted by $D_{\mathrm{st}}(n)$, to be $D(3^k n)$ for any k such that $3^k n$ is stable.

Note that Proposition 3.2 shows that this is well-defined.

Proposition 3.4. *We have*:

(1) $D_{\mathrm{st}}(n) = \min_{k \geq 0} D(3^k n)$.

(2) *For* $n > 1$, $D_{\mathrm{st}}(n) = \|n\|_{\mathrm{st}} - L(n)$.

(3) $D_{\mathrm{st}}(n) \leq D(n)$, *with equality if and only if* n *is stable or* $n = 1$.

(4) *For* $n > 1$, $D(n) - D_{\mathrm{st}}(n) = \delta(n) - \delta_{\mathrm{st}}(n) = \|n\| - \|n\|_{\mathrm{st}}$.

Proof. With the exception of (4), of which no analogue has previously been mentioned, these all follow from Proposition 3.2 and their proofs are exactly analogous to those of the statements in Proposition 2.7; meanwhile (4) follows immediately from (2) and the definition of $D(n)$. □

We then also have the analogue of Proposition 3.1:

Proposition 3.5. *For a natural number* $n > 1$,

$$D_{\mathrm{st}}(n) = |\{k : n < E(k) \leq E(\|n\|_{\mathrm{st}})\}|.$$

Proof. Once again, by definition, $L(n)$ is the largest k such that $E(k) \le n$. And since $E(k)$ is strictly increasing, the number of k such that $n < E(k) \le E(\|n\|_{\text{st}})$ is equal to the difference $\|n\|_{\text{st}} - L(n)$, which by Proposition 3.4 is $D_{\text{st}}(n)$. □

Remark 3.6. It may seem strange that 1 needs to be excluded, given that its special status goes away when stabilized. However, $\|1\|_{\text{st}} = 0$, and $E(0)$ is not defined, so $n = 1$ must still be excluded from the theorem statement.

Note, by the way:

Proposition 3.7. *For any natural number n, $D(n) = 0$ if and only if $D_{\text{st}}(n) = 0$.*

Proof. It is immediate that a number n with $D(n) = 0$ is stable and so has $D_{\text{st}}(n) = 0$ (unless $n = 1$, in which case one still has $D_{\text{st}}(n) = 0$). For the reverse, a number n has $D_{\text{st}}(n) = 0$ if and only if there is some k such that $D(3^k n) = 0$. However, as the numbers n with $D(n) = 0$ are precisely those numbers of the form 3^k, $2 \cdot 3^k$, and $4 \cdot 3^k$, we see that if n has $D_{\text{st}}(n) = 0$, it must itself be of one of these forms, and thus have $D(n) = 0$. □

See Corollaries 5.2 and 5.3 for related statements.

Having discussed what $D(n)$ is and how it acts, let us finally discuss how it may be computed. The quantity $D(n)$ is just the difference $\|n\| - L(n)$. We know how to compute $\|n\|$, although not necessarily quickly; see [Arias de Reyna and van de Lune 2014] for the currently best-known algorithm for computing complexity, and [Cordwell et al. 2019] for the best-known bounds on its runtime. But the other half, computing $L(n)$, is very simple and can be done much quicker, because it is given by the following formula:

Proposition 3.8. *For a natural number n,*

$$L(n) = \max\left\{3\lfloor \log_3 n \rfloor,\ 3\left\lfloor \log_3\left(\tfrac{1}{2}n\right)\right\rfloor + 2,\ 3\left\lfloor \log_3\left(\tfrac{1}{4}n\right)\right\rfloor + 4,\ 1\right\}.$$

Proof. The quantity $L(n)$ is by definition the largest k such that $E(k) \le n$. The largest such k congruent to 0 modulo 3 is $3\lfloor \log_3 n \rfloor$ (so long as this quantity is positive; otherwise there is none), the largest such k congruent to 2 modulo 3 is $3\left\lfloor \log_3\left(\frac{1}{2}n\right)\right\rfloor + 2$ (with the same caveat), the largest such $k > 1$ congruent to 1 modulo 3 is $3\left\lfloor \log_3\left(\frac{1}{4}n\right)\right\rfloor + 4$ (again with the same caveat), and of course the largest such k equal to 1 is 1. So the largest of these is $L(n)$ (and any of them that are not valid positive and thus not a valid k will not affect the maximum). □

Let us make here a definition that will be useful later:

Definition 3.9. For a natural number n, define $R(n) = n/E(\|n\|)$. We also define $R_{\text{st}}(n)$ to be $R(3^k n)$ for any k such that $3^k n$ is stable, or equivalently (for $n > 1$) as $n/E(\|n\|_{\text{st}})$.

This is easily related to the defect, as was done in an earlier paper [Altman 2015]:

Proposition 3.10. *We have, for $n > 1$,*

$$\delta(n) = \begin{cases} -3\log_3 R(n) & \text{if } \|n\| \equiv 0 \pmod 3, \\ -3\log_3 R(n) + 2\,\delta(2) & \text{if } \|n\| \equiv 1 \pmod 3, \\ -3\log_3 R(n) + \delta(2) & \text{if } \|n\| \equiv 2 \pmod 3, \end{cases}$$

and the same relation (without the $n > 1$ restriction) holds between $R_{\text{st}}(n)$, $\|n\|_{\text{st}}$, and $\delta_{\text{st}}(n)$.

Proof. The relation between $R(n)$ and $\delta(n)$ is just Proposition A.3 from [Altman 2015], and the proof for the stable case is exactly analogous. □

Now we see that in addition to being easy to compute $L(n)$, it is also simple to determine $D(n)$ from $\delta(n)$, at least if we know the value of $\|n\|$ modulo 3, which technically is implicit in $\delta(n)$. First, a definition:

Definition 3.11. Let a be a congruence class modulo 3 and k be a whole number. Define

$$t_a(k) = \begin{cases} k & \text{if } k \equiv a \pmod 3, \\ k + \delta(2) & \text{if } k \equiv a+1 \pmod 3, \\ k + 2\delta(2) & \text{if } k \equiv a+2 \pmod 3. \end{cases}$$

Now:

Theorem 3.12. *Let $n > 1$ be a natural number. Then $D(n)$ is equal to the smallest k such that $\delta(n) \le t_{\|n\|}(k)$. Moreover, if n is any natural number, $D_{\mathrm{st}}(n)$ is equal to the smallest k such that $\delta_{\mathrm{st}}(n) \le t_{\|n\|_{\mathrm{st}}}(k)$.*

Since two numbers with the same defect also have the same complexity modulo 3 (and $\delta(n) = 1$ if and only if $n = 1$), and the analogous statement is also true of stable complexity and defect, in particular we have that if $\delta(n) = \delta(m)$ then $D(n) = D(m)$, and if $\delta_{\mathrm{st}}(n) = \delta_{\mathrm{st}}(m)$ then $D_{\mathrm{st}}(n) = D_{\mathrm{st}}(m)$.

Note in addition that since $\delta(n) = \delta(m)$ implies $\delta_{\mathrm{st}}(n) = \delta_{\mathrm{st}}(m)$ (see statement (2) in Proposition 2.7) one has that if $\delta(n) = \delta(m)$ then $D_{\mathrm{st}}(n) = D_{\mathrm{st}}(m)$.

Theorem 3.12 makes precise how $D(n)$ is "almost $\lceil \delta(n) \rceil$". It is, as was noted in the Introduction, not the same, but it is the smallest k such that $\delta(n) \le t_{\|n\|}(k)$, where $t_{\|n\|}(k)$ may not be exactly k but never differs from it by more than $2\delta(2) < 0.215$.

Proof. We prove only the nonstabilized case, as the stabilized case is exactly analogous. We assume $n > 1$.

From Proposition 3.1, we can see that $D(n)$ is determined by $R(n)$ and the value of $\|n\|$ modulo 3. Specifically,

$$D(n) = \left| \left\{ k : R(n) < \frac{E(k)}{E(\|n\|)} \le 1 \right\} \right|,$$

so $D(n)$ is the number of values of $E(k)/E(\|n\|)$ in $(R(n), 1]$. What are the values of this? They can be obtained as products of values $E(k)/E(k+1)$; this is equal to $\frac{2}{3}$ when $k \equiv 1$ or $2 \pmod 3$ (for $k > 1$) and to $\frac{3}{4}$ when $k \equiv 0 \pmod 3$.

Thus, if $\|n\| \equiv 0 \pmod 3$, $D(n)$ will increase whenever $R(n)$ passes a value of the sequence $1, \frac{2}{3}, \frac{4}{9}, \frac{1}{3}, \frac{2}{9}, \frac{4}{27}, \frac{1}{9}, \ldots$; if $\|n\| \equiv 1 \pmod 3$, whenever it passes a value of the sequence $1, \frac{3}{4}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{9}, \ldots$; and if $\|n\| \equiv 2 \pmod 3$, whenever it passes a value of the sequence $1, \frac{2}{3}, \frac{1}{2}, \frac{1}{3}, \frac{2}{9}, \frac{1}{6}, \frac{1}{9}, \ldots$. (These sequences are just the sequences obtained by taking products of one of the three shifts of the periodic sequence $\frac{2}{3}, \frac{2}{3}, \frac{3}{4}, \frac{2}{3}, \frac{2}{3}, \frac{3}{4}, \ldots$; note that regardless of which shift is used, the repeating part of the sequence always has a product of $\frac{1}{3}$, and so the product sequences will always consist of three interwoven geometric sequences each with ratio $\frac{1}{3}$.)

It just remains, then, to convert these values of $R(n)$ to their equivalents in defects, which can be done with Proposition 3.10. Once this is done one finds that the values of $\delta(n)$ where $D(n)$ increases are precisely those listed in the definition of $t_{\|n\|}$, which completes the proof. □

Theorem 3.12 will form half the proof of Theorem 1.8, and its stable analogue, Theorem 4.2; it tells us that the values of $D(n)$ "switch over" when $\delta(n)$ is of the form k, $k+\delta(2)$, or $k+2\delta(2)$ depending on the congruence class of $k-\|n\|$ modulo 3. The other half the proof is, of course, Theorem 1.9 (and its stable analogue, Theorem 4.1), which will tell us that these changeover points are exactly the limits of the initial ω^k defects in $\mathscr{D}^a$ (or $\mathscr{D}^a_{\mathrm{st}}$).

4. The order interpretation of $D(n)$

In this section we aim to prove Theorem 1.9 using the methods described in Section 1C; combined with Theorem 3.12 from the previous section, this will prove Theorem 1.8. Really, we want to prove generalizations:

Theorem 4.1. *For any $k \geq 0$ and a, a congruence class modulo 3, the order type of $\mathscr{D}^a \cap [0, t_a(k)]$ and the order type of $\mathscr{D}^a_{\mathrm{st}} \cap [0, t_a(k)]$ are both equal to ω^k.*

Theorem 4.2. *Let $n > 1$ be a natural number. Let ζ be the order type of $\mathscr{D}^{\|n\|} \cap [0, \delta(n))$. Then $D(n)$ is equal to the smallest k such that $\zeta < \omega^k$. The same is true if we replace $\delta(n)$ by $\delta_{\mathrm{st}}(n)$, $\mathscr{D}^{\|n\|}$ by $\mathscr{D}^{\|n\|_{\mathrm{st}}}_{\mathrm{st}}$, and $D(n)$ by $D(n)_{\mathrm{st}}$.*

Note that the proofs in this section will rely heavily on the results in Sections 2B and 2C. Before we prove these, though, we will need a slight elaboration on Proposition 2.19:

Proposition 4.3. *Let f be a low-defect polynomial of degree d with $\delta(f) < d+1$. Then the order type of the set of all $\delta(N)$ for n 3-represented by $\hat{f}$ is exactly ω^d.*

Proof. By Proposition 2.19, $\{\delta(\hat{f}(3^{n_1}, \ldots, 3^{n_d})) : (n_1, \ldots, n_d) \in S\}$ has order type less than ω^d. Meanwhile, also by Proposition 2.19, the set

$$\{\delta(f(3^{n_1}, \ldots, 3^{n_d})) : (n_1, \ldots, n_d) \notin S\}$$

has order type at least ω^d, and is cofinal in $[0, \delta(f))$ (or $[0, \delta(f)]$ if $\deg f = 0$) and therefore in the set of all $\delta(N)$ for n 3-represented by $\hat{f}$. But in fact, for $(n_1, \ldots, n_d) \notin S$, one has $\delta(\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})) = \delta_f(n_1, \ldots, n_d)$, and so this set (even when $f(3^{n_1}, \ldots, 3^{n_d})$ is replaced by $\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})$) is a subset of the image of δ_f, which by Proposition 2.16 has order type ω^d. So the conditions of Proposition 2.18 apply, and the union of these two sets, the set of all $\delta(n)$ for N 3-represented by $\hat{f}$, has order type at most ω^d. We already know by Proposition 2.16 it has order type at least ω^d, so this proves the claim. □

We now prove the main theorems of this section.

Proof of Theorem 4.1. We need to show that the order type of $\mathscr{D}^a \cap [0, t_a(k)]$ and the order type of $\mathscr{D}^a_{\mathrm{st}} \cap [0, t_a(k)]$ are both equal to ω^k. This proof breaks down into two parts, an upper bound and a lower bound. Since $\mathscr{D}^a_{\mathrm{st}} \subseteq \mathscr{D}^a$, it suffices to prove the upper bound for $\mathscr{D}^a \cap [0, t_a(k)]$ and the lower bound for $\mathscr{D}^a_{\mathrm{st}} \cap [0, t_a(k)]$.

We begin with the upper bound. First, we observe that $t_a(k)$ is not itself an element of $\mathscr{D}^a$ for any $k > 0$. We can see this as neither $k+\delta(2)$ nor $k+2\delta(2)$ is a defect for any $k > 0$ (such a defect would have to come from some number n satisfying $3^\ell n = 2$ or $3^\ell n = 4$ for $\ell > 0$, which is impossible), and

similarly no nonzero integer is a defect except $k = 1$, which though an element of $\mathscr{D}$ is by definition excluded from all three $\mathscr{D}^a$. Thus

$$\mathscr{D}^a \cap [0, t_a(k)] = \mathscr{D}^a \cap [0, t_a(k))$$

and we may concern ourselves with the order type of the latter.

Now we take a good covering $\mathcal{S}$ of $B_{t_a(k)}$ as per Theorem 2.26. For any $f \in \mathcal{S}$ with leading coefficient m, we have the inequality $\delta(m) + \deg f \leq \delta(f) \leq t_a(k)$. In particular, for any $f \in \mathcal{S}$, we have $\deg f \leq \lfloor t_a(k) \rfloor = k$.

Suppose now that $\deg f = k$; then there is more we can say. For in this case, we have $\delta(m) \leq t_a(k) - k \leq 2\delta(2)$. Thus $\delta(m) \in \{0, \delta(2), 2\delta(2)\}$ by Proposition 2.8. Note that by their respective definitions, $\delta(f) \equiv \delta(m) \pmod 1$, and, as noted above, $\delta(f) \geq \deg f = k$, and so

$$\delta(f) = k + \delta(m) \in \{k, k + \delta(2), k + 2\delta(2)\}.$$

Note that $\delta(f) = k + \delta(m)$ means that

$$k + \|m\| - 3\log_3 m = \|f\| - 3\log_3 m$$

and therefore $\|f\| = k - \|m\|$. Moreover, if $\delta(m) = 0$, then m is of the form 3^ℓ (for some $\ell > 0$) and $\|m\| = 3\ell$, while if $\delta(m) = \delta(2)$ then m is of the form $2 \cdot 3^\ell$ with $\|m\| = 2 + 3\ell$, and if $\delta(m) = 2\delta(2)$ then m is of the form $4 \cdot 3^\ell$ with $\|m\| = 4 + 3\ell$; from this we can conclude that, modulo 3,

$$\|f\| \equiv \begin{cases} k & \text{if } \delta(f) = k, \\ k-2 & \text{if } \delta(f) = k + \delta(2), \\ k-1 & \text{if } \delta(f) = k + 2\delta(2). \end{cases}$$

Now, let

$$T_f = \{\delta(\hat{f}(3^{n_1}, \ldots, 3^{n_{d+1}})) : n_1, \ldots, n_{d+1} \geq 0\} \cap \mathscr{D}^a,$$

where $d = \deg f$. Then by the assumption that $\mathcal{S}$ is a good covering of $B_{t_a(k)}$, we have

$$\mathscr{D}^a \cap [0, t_a(k)) = \bigcup_{f \in \mathcal{S}} T_f.$$

We want to show that the conditions of Proposition 2.18 hold for the sets T_f, so that we can conclude that $\mathscr{D}^a \cap [0, t_a(k))$ has order type at most ω^k. If $\deg f < k$, then, by Proposition 2.16, T_f has order type less than ω^k, and thus so does $T_f \cap \mathscr{D}^a$. Meanwhile, if $\deg f = k$, then since $\delta(f) \leq t_a(k) < k + 1$, we can apply Proposition 4.3 to conclude that the set of $\delta(N)$ for N 3-represented by $\hat{f}$ has order type ω^k. However, if $\delta(f) \neq t_a(k)$, then by the previous paragraph and Proposition 2.19, we see that while this has order type ω^k, T_f, which is its intersection with $\mathscr{D}^a$, has order type less than ω^k.

It remains to check, then, that when $\deg f = k$ and $\delta(f) = t_a(k)$, that the set T_f is cofinal in $\bigcup_{f \in \mathcal{S}} T_f = \mathscr{D}^a \cap [0, t_a(k))$, or in other words, simply that it is cofinal in $[0, t_a(k))$. But this follows from Proposition 2.19, which in fact goes further and states that $T_f \cap \mathscr{D}^a_{\text{st}}$ is cofinal in $[0, \delta(f)) = [0, t_a(k))$.

Thus, applying Proposition 2.18, we conclude that $\mathscr{D}^a \cap [0, t_a(k))$ has order type at most ω^k. This proves the upper bound.

To prove the lower bound, let us consider the low-defect polynomial

$$f = (\cdots((mx_1 + 1)x_2 + 1)\cdots)x_k + 1$$

(for a particular m to be chosen shortly) which has $\|f\| = \|m\| + k$. (The upper bound on $\|f\|$ is immediate and the lower bound follows from Proposition 2.20.) For the value of m, we take

$$m = \begin{cases} 3 & \text{if } k - a \equiv 0 \pmod 3, \\ 4 & \text{if } k - a \equiv 2 \pmod 3, \\ 2 & \text{if } k - a \equiv 1 \pmod 3, \end{cases}$$

so that $\|m\| \equiv a - k \pmod 3$ and $\|f\| \equiv a \pmod 3$, meaning $\mathscr{D}_{\mathrm{st}}^{\|f\|} = \mathscr{D}_{\mathrm{st}}^{a}$.

Then $\delta(f) = t_a(k)$ and so in particular $\delta(f) < k+1$, meaning once again we can apply Proposition 2.19 to conclude that the set

$$\{\delta(f(3^{n_1}, \ldots, 3^{n_k})) : (n_1, \ldots, n_k) \in \mathbb{Z}_{\geq 0}^k\} \cap \mathscr{D}_{\mathrm{st}}^{\|f\|}$$

has order type at least ω^k. Since this set is bounded above by $\delta(f) = t_a(k)$, and $\mathscr{D}_{\mathrm{st}}^{\|f\|} = \mathscr{D}_{\mathrm{st}}^{a}$, we conclude that the order type of $\mathscr{D}_{\mathrm{st}}^{a} \cap [0, t_a(k))$ is at least ω^k. □

In particular this encompasses Theorem 1.9.

Proof of Theorem 1.9. This is just a rephrasing of Theorem 4.1 with the application to $\mathscr{D}_{\mathrm{st}}^{a}$ omitted. □

Having proven Theorem 4.1, we can now combine it with Theorem 3.12 to obtain Theorems 4.2 and 1.8:

Proof of Theorem 4.2. By Theorem 3.12, $D(n)$ is equal to the smallest k such that $\delta(n) \leq t_{\|n\|}(k)$. However, since the order type of $\mathscr{D}^{\|n\|} \cap [0, t_{\|n\|}(k))$ is equal to ω^k, one has that $\zeta < \omega^k$ if and only if $\delta(n) < t_{\|n\|}(k)$. Thus $D(n)$ is equal to the smallest k such that $\zeta < \omega^k$. The proof for the stabilized version is similar. □

Proof of Theorem 1.8. This is just the special case of Theorem 4.2 where we only consider $\delta(n)$ and not $\delta_{\mathrm{st}}(n)$. □

5. Numbers n with $D(n) \leq 1$

In the previous section we showed that the numbers with integral defect at most k correspond to the initial ω^k defects in each of $\mathscr{D}^0$, $\mathscr{D}^1$, and $\mathscr{D}^2$. In this section we take a closer look at the initial ω, the numbers with integral defect at most 1, and use this to generalize Theorem 1.10.

Let us start by listing all the numbers with integral defect at most 1:

Theorem 5.1. *A natural number n satisfies $D(n) \leq 1$ if and only if it can be written in one of the following forms*:

(1) 1, *of complexity* 1.

(2) $2^a 3^k$ *for* $a \leq 10$, *of complexity* $2a + 3k$ (*for* a, k *not both zero*).

(3) $2^a(2^b 3^\ell + 1)3^k$ *for* $a + b \leq 2$, *of complexity* $2(a+b) + 3(\ell + k) + 1$ (*for* b, ℓ *not both zero*).

Proof. By Theorem 3.12, any n with $D(n) \leq 1$ must have $\delta(n) \leq 1 + 2\delta(2)$. Theorem 31 from [Altman and Zelinsky 2012] gives a classification of all numbers n with $\delta(n) < 12\delta(2)$, together with their complexities; since $12\delta(2) > 1 + 2\delta(2)$, any n with $D(n) \leq 1$ may be found among these. (One may also use the algorithms from [Altman 2018a] to find such a classification.) It is then a straightforward matter to determine which of the n listed there have $D(n) \leq 1$. □

This has an important corollary:

Corollary 5.2. *For any natural number n, $D(n) = 1$ if and only if $D_{\mathrm{st}}(n) = 1$.*

Proof. From Theorem 5.1, we see that if $D(n) \leq 1$ then we also have $D(3^k n) \leq 1$, and if $D(3^k n) \leq 1$ then we have $D(n) \leq 1$; this shows that $D(n) \leq 1$ if and only if $D_{\mathrm{st}}(n) \leq 1$. Combining this with Proposition 3.7 proves the claim. □

From this we can conclude:

Corollary 5.3. *For any natural number $n > 1$, if $D(n) \leq 2$ then n is stable (and so $D_{\mathrm{st}}(n) \leq 2$).*

Proof. If $D(n) = 0$ or $D(n) = 1$, this is Proposition 3.7 or Corollary 5.2, respectively. If $D(n) = 2$, then for any $k \geq 0$, if we had $D(3^k n) < 2$, then, by Proposition 3.7 and Corollary 5.2, we would have $D(n) < 2$, contrary to assumption; thus $D(3^k n) = 2$ for all $k \geq 0$, i.e., n is stable (by Proposition 3.4). □

Note that the converse, that if $D_{\mathrm{st}}(n) \leq 2$ then $D(n) \leq 2$, does not hold; for instance, we can consider 107, which has $D_{\mathrm{st}}(107) = 2$ but $D(107) = 3$, or 683, which has $D_{\mathrm{st}}(683) = 2$ but $D(683) = 4$. (It is easy to verify that these numbers have stable integer defect at most 2 because $D(321) = D(2049) = 2$; that these numbers do then have stable integer defect equal to 2 and not any lower can then be inferred from Corollary 5.3. Alternately, the stable complexity, and thus stable integer defect, may be computed with the algorithms from [Altman 2018a].)

However, for our purposes, the most important consequence of Corollary 5.2 is the following rephrasing of it:

Proposition 5.4. *Let $k > 1$ be a natural number and suppose h is a value of R corresponding to a defect in the initial ω of $\mathscr{D}^k$. Then if $hE(k)$ is a natural number n, one has $\|n\| = k$, and, moreover, $n > E(k-1)$.*

Proof. Suppose $hE(k)$ is a natural number n. We must have $n > 1$ because having $h = 1/E(k)$ for $k > 1$ would by Proposition 3.10 correspond to a defect which is a nonzero integer, and these (by Proposition 2.7) do not exist.

Then there is, by definition of h, some number $m > 1$ with $\|m\| \equiv k \pmod 3$ and $R(m) = h$, i.e., $m = hE(\|m\|)$. Since $\|m\| \equiv k \pmod 3$ we see that $m = n3^\ell$ for some $\ell \in \mathbb{Z}$, where $\ell = (\|m\| - k)/3$. But also we have $D(m) \leq 1$. Therefore, whether $\ell \geq 0$ or $\ell \leq 0$, we must have $D_{\mathrm{st}}(n) \leq 1$, and so, by Proposition 3.7 and Corollary 5.2, we have $D(n) \leq 1$. Then by Proposition 3.2, we have $\|m\| = \|n\| + 3\ell$. From the definition of ℓ we also have $\|m\| = k + 3\ell$ and thus we conclude that $\|n\| = k$. And since $D(n) = 1$ this means (by Proposition 3.1) that $n > E(k-1)$. □

We can now prove Theorem 1.12:

Proof of Theorem 1.12. Suppose we want to determine the r-th largest number of complexity k. This is equivalent to determining the r-th largest value of $R(n) = n/E(k)$ that occurs among numbers n of complexity k, which is equivalent to determining the r-th smallest defect $\delta(n)$ that occurs among numbers n of complexity k.

Now, we can easily determine the initial values $\alpha_0, \ldots, \alpha_r$ of $\mathscr{D}^k$; let $h_0, \ldots, h_r$ be the corresponding values of the function R, as given by Proposition 3.10. (For instance, for a way of getting $h_0, \ldots, h_r$ directly rather than going by means of defects, one may take the numbers n given in Theorem 5.1, group them by the residues of $\|n\|$ modulo 3, and then sort them in decreasing order by $R(n)$; note that the values of $R(n)$ obtained this way for any one congruence class of $\|n\|$ modulo 3 will have reverse order

r		h	corresponding leader
0		1	$3 = 3^1 2^0 = 2^1 3^0 + 1$
1		8/9	$8 = 2^3 = 2^1(3^1+1)$
2		64/81	$64 = 2^6$
3		7/9	$7 = 2^1 3^1 + 1$
4		20/27	$20 = 2^1(3^2+1)$
5		19/27	$19 = 2^1 3^2 + 1$
6		512/729	$512 = 2^9$
(for $n \geq 4$)	$2n-1$	$2/3+2/3^n$	$2^1(3^{n-1}+1)$
(for $n \geq 4$)	$2n$	$2/3+1/3^n$	$2^1 3^{n-1}+1$

Table 4. Table of h_r for $k \equiv 0 \pmod 3$.

r		h	corresponding leader
0		1	$2 = 2^1$
1		8/9	$16 = 2^4 = 2^2(3^1+1)$
2		5/6	$5 = 2^2 3^0 + 1$
3		64/81	$128 = 2^7$
4		7/9	$14 = 2^1(2^1 3^1+1)$
5		20/27	$40 = 2^2(3^2+1)$
6		13/18	$13 = 2^2 3^1 + 1$
7		19/27	$38 = 2^1(2^1 3^2+1)$
8		512/729	$1024 = 2^{10}$
(for $n \geq 4$)	$3n-3$	$2/3+2/3^n$	$2^2(3^{n-1}+1)$
(for $n \geq 4$)	$3n-2$	$2/3+1/(2\cdot 3^{n-1})$	$2^2 3^{n-1}+1$
(for $n \geq 4$)	$3n-1$	$2/3+1/3^n$	$2^1(2^1 3^{n-1}+1)$

Table 5. Table of h_r for $k \equiv 2 \pmod 3$.

type ω.) One may see Tables 4, 5, and 6 for tables of the resulting values of h. Then certainly, the r-th largest number of complexity k is at most $h_r E(k)$, because the set of values of $R(n)$ occurring for n with $\|n\| = k$ is a subset of the values of $R(n)$ occurring for $n > 1$ with $\|n\| \equiv k \pmod 3$. However, it will only be exactly the r-th largest number of complexity k if all of h_1 through h_r do indeed occur for some n with $\|n\| = k$.

But, by Proposition 5.4, this is equivalent to just requiring that all of the numbers $h_0 E(k), \ldots, h_r E(k)$ are indeed whole numbers (and moreover when this does occur one will have $h_i E(k) > E(k-1)$). In other words, this is the same as requiring

$$k \geq \begin{cases} -3 \min_{s \leq r} v_3(h_s) & \text{if } k \equiv 0 \pmod 3, \\ -3 \min_{s \leq r} v_3(h_s) + 4 & \text{if } k \equiv 1 \pmod 3, \\ -3 \min_{s \leq r} v_3(h_s) + 2 & \text{if } k \equiv 2 \pmod 3. \end{cases}$$

r	h	corresponding leader
0	1	$4 = 2^2 = 3^1+1$
1	8/9	$32 = 2^5$
2	5/6	$10 = 3^2+1$
3	64/81	$256 = 2^8$
(for $n \geq 2$) $n+2$	$3/4+1/(4\cdot 3^n)$	$3^{n+1}+1$

Table 6. Table of h_r for $k \equiv 1 \pmod 3$ with $k > 1$.

So we have our $h_{r,a}$, and we can take $K_{r,a}$ to be given by this formula. (Although since for $K_{0,0}$ it may not may make much sense to take $K_{0,0} = 0$, one may wish to take $K_{0,0} = 3$ instead, as we have done in Table 1.)

Combining this with Tables 4, 5, and 6 yields Tables 1, 2, and 3, and proves the theorem. □

Remark 5.5. While in the proof of Theorem 1.12 we have referred to facts proved in Section 4, none of the techniques deployed in that section are necessary for the proof. For instance, one can easily verify the values of the $\overline{\mathscr{D}^a}(\omega)$ by directly determining the initial ω elements without needing to determine it for all ω^k; indeed Tables 4, 5, and 6 essentially do this directly from Theorem 5.1.

As a final note, it is worth making formal a statement mentioned in Section 1B, that the numbers $hE(k)$ coming from Theorem 1.12 are almost exactly the n with $D(n) \leq 1$:

Proposition 5.6. *A number n has $D(n) \leq 1$ if and only if there are some $\ell \geq 0$, $k \geq 1$, and $r \geq 0$ such that $k \geq K_{r,k}$ and $3^\ell n = h_{r,k}E(k)$.*

Proof. We already know that if $k \geq K_{r,k}$ then, if we let $m = h_{r,k}E(k)$, that $m > E(k-1) = E(\|m\|-1)$, i.e., $D(m) \leq 1$, and so if $m = 3^\ell n$, then $D(n) \leq 1$ by Corollary 5.2.

Conversely, if $D(n) \leq 1$, let $h = R(n)$; then by the construction of the $h_{r,a}$ in the proof of Theorem 1.12, and the fact that the values of $R(n)$ for numbers n with $\|n\|$ in a fixed congruence class modulo 3 have reverse order type ω, there is some r such that $h = h_{r,\|n\|}$. We may then take any $k \geq K_{r,\|n\|}$ with $k \equiv \|n\| \pmod 3$; then $3^\ell n = h_{r,\|n\|}E(k) = h_{r,k}E(k)$ for $\ell = (k - \|n\|)/3$. □

Acknowledgements

Work of the author was supported by NSF grants DMS-0943832 and DMS-1101373.

References

[Altman 2015] H. Altman, "Integer complexity and well-ordering", *Michigan Math. J.* **64**:3 (2015), 509–538. MR Zbl

[Altman 2016] H. Altman, "Integer complexity: representing numbers of bounded defect", *Theoret. Comput. Sci.* **652** (2016), 64–85. MR Zbl

[Altman 2018a] H. Altman, "Integer complexity: algorithms and computational results", *Integers* **18** (2018), art. id. #45.

[Altman 2018b] H. Altman, "Internal structure of addition chains: well-ordering", *Theoret. Comput. Sci.* **721** (2018), 54–69. MR Zbl

[Altman and Arias de Reyna ≥ 2019] H. Altman and J. Arias de Reyna, "Integer complexity, stability, and self-similarity", in preparation.

[Altman and Zelinsky 2012] H. Altman and J. Zelinsky, "Numbers with integer complexity close to the lower bound", *Integers* **12**:6 (2012), 1093–1125. MR Zbl

[Arias de Reyna and van de Lune 2014] J. Arias de Reyna and J. van de Lune, "Algorithms for determining integer complexity", preprint, 2014. arXiv

[Bahig 2008] H. M. Bahig, "On a generalization of addition chains: addition-multiplication chains", *Discrete Math.* **308**:4 (2008), 611–616. MR Zbl

[Brauer 1939] A. Brauer, "On addition chains", *Bull. Amer. Math. Soc.* **45** (1939), 736–739. MR Zbl

[Carruth 1942] P. W. Carruth, "Arithmetic of ordinals with applications to the theory of ordered Abelian groups", *Bull. Amer. Math. Soc.* **48** (1942), 262–271. MR Zbl

[Cordwell et al. 2019] K. Cordwell, A. Epstein, A. Hemmady, S. J. Miller, E. Palsson, A. Sharma, S. Steinerberger, and Y. N. T. Vu, "On algorithms to calculate integer complexity", *Integers* **19** (2019), art. id. A12. MR

[Dellac 1894] H. Dellac, "Réponse 49", *L'Intermédiaire des Mathématiciens* **1** (1894), 163–164.

[Gioia et al. 1962] A. A. Gioia, M. V. Subba Rao, and M. Sugunamma, "The Scholz–Brauer problem in addition chains", *Duke Math. J.* **29** (1962), 481–487. MR Zbl

[Guy 1986] R. K. Guy, "Unsolved problems: some suspiciously simple sequences", *Amer. Math. Monthly* **93**:3 (1986), 186–190. MR

[Guy 1987] R. K. Guy, "Unsolved Problems: monthly unsolved problems, 1969–1987", *Amer. Math. Monthly* **94**:10 (1987), 961–970. MR Zbl

[Guy 1989] R. K. Guy, "Unsolved problems come of age", *Amer. Math. Monthly* **96**:10 (1989), 903–909. Zbl

[Guy 2004] R. K. Guy, *Unsolved problems in number theory*, 3rd ed., Springer, 2004. MR Zbl

[de Jongh and Parikh 1977] D. H. J. de Jongh and R. Parikh, "Well-partial orderings and hierarchies", *Nederl. Akad. Wetensch. Proc. Ser. A* **80**:3 (1977), 195–207. MR Zbl

[Knuth 1998] D. E. Knuth, *The art of computer programming, II: Seminumerical algorithms*, 3rd ed., Addison-Wesley, Reading, MA, 1998. MR Zbl

[Mahler and Popken 1953] K. Mahler and J. Popken, "On a maximum problem in arithmetic", *Nieuw Arch. Wiskunde* (3) **1** (1953), 1–15. MR Zbl

[Rawsthorne 1989] D. A. Rawsthorne, "How many 1's are needed?", *Fibonacci Quart.* **27**:1 (1989), 14–17. MR Zbl

[Scholz 1937] A. Scholz, "Aufgabe 253", *Jahresber. Deutsch. Math.-Ver.* **47**:2 (1937), 41–42.

[Subbarao 1989] M. V. Subbarao, "Addition chains: some results and problems", pp. 555–574 in *Number theory and applications* (Banff, AB, 1988), edited by R. A. Mollin, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **265**, Kluwer, Dordrecht, 1989. MR Zbl

[Volkovich 2016] I. Volkovich, "Characterizing arithmetic read-once formulae", *ACM Trans. Comput. Theory* **8**:1 (2016), art. id. 2. MR Zbl

[Zelinsky ≥ 2019] J. Zelinsky, "An upper bound on integer complexity", in preparation.

Received 2 Aug 2018. Revised 7 Apr 2019.

HARRY ALTMAN:

harry.j.altman@gmail.com
University of Michigan, Ann Arbor, MI, United States

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 3, 2019
dx.doi.org/10.2140/moscow.2019.8.219

Generalized simultaneous approximation to *m* linearly dependent reals

Leonhard Summerer

In order to analyse the simultaneous approximation properties of m reals, the parametric geometry of numbers studies the joint behaviour of the successive minima functions with respect to a one-parameter family of convex bodies and a lattice defined in terms of the m given reals. For simultaneous approximation in the sense of Dirichlet, the linear independence over $\mathbb{Q}$ of these reals together with 1 is equivalent to a certain nice intersection property that any two consecutive minima functions enjoy. This paper focusses on a slightly generalized version of simultaneous approximation where this equivalence is no longer in place and investigates conditions for that intersection property in the case of linearly dependent irrationals.

1. Introduction

In Diophantine approximation the simultaneous approximation to $m := n-1$ real numbers $\xi_1, \dots, \xi_m$ has a long tradition, starting with Dirichlet who proved the existence of nontrivial solutions $(x, y_1, \dots, y_m) \in \mathbb{Z}^n$ to the system

$$\begin{aligned} |x| &\le e^q, \\ |\xi_1 x - y_1| &\le e^{-q/m}, \\ &\vdots \\ |\xi_m x - y_m| &\le e^{-q/m} \end{aligned} \tag{$\star$}$$

for any parameter $q > 0$. In other words, if $\mathcal{B}(q)$ consists of points $(p_0, p_1, \dots, p_m)$ with $|p_0| \le e^q$, $|p_i| \le e^{-q/m}$ for $1 \le i \le m$, and $\Lambda = \Lambda(\boldsymbol{\xi})$ the lattice of points $(x, \xi_1 x - y_1, \dots, \xi_m x - y_m)$ with $(x, y_1, \dots, y_m) \in \mathbb{Z}^n$, Dirichlet's theorem asserts that there is a nonzero lattice point in $\mathcal{B}(q)$, i.e., that the first minimum $\lambda_1(q)$ with respect to $\mathcal{B}(q)$ and Λ is at most 1.

Lately, the successive minima functions $\lambda_1(q), \dots, \lambda_n(q)$ have been intensively studied within the framework of parametric geometry of numbers, culminating in a fundamental paper of D. Roy [2015] in which he reduces the problem of describing the joint spectrum of a family of exponents of Diophantine approximation relative to ($\star$) to combinatorial analysis. A main tool for the investigation of the successive minima functions is the following result from [Schmidt and Summerer 2009]:

Proposition 1.1. *Suppose $1, \xi_1, \dots, \xi_m$ are linearly independent over $\mathbb{Q}$ and let $\lambda_i(q)$ denote the successive minima with respect to $\Lambda(\boldsymbol{\xi})$ and $\mathcal{B}(q)$. Then for every $s < n$ there exist arbitrarily large values of q for which $\lambda_s(q) = \lambda_{s+1}(q)$.*

The author was supported by FWF grant I 3466-N35.

MSC2010: 11H06, 11J13.

Keywords: parametric geometry of numbers, successive minima, simultaneous approximation.

An analogous result holds in the more general situation where a system of exponents $(\nu_0, -\nu_1, \dots, -\nu_m)$ with $\nu_i > 0$ for $1 \le i \le m$ and $\nu_0 - \nu_1 - \cdots - \nu_m = 0$ is considered (see [Schmidt and Summerer 2009], page 72, Corollary 2.2). Here we normalize to the case $\nu_0 = 1$ so that $\nu_1 + \cdots + \nu_m = 1$ and denote by $\mathcal{B}^\nu(q)$ the box of points $(p_0, p_1, \dots, p_m)$ defined by $|p_0| \le e^q$, $|p_i| \le e^{-\nu_i q}$ for $1 \le i \le m$. This modifies the initial system to

$$\begin{aligned} |x| &\le e^q, \\ |\xi_1 x - y_1| &\le e^{-\nu_1 q}, \\ &\vdots \\ |\xi_m x - y_m| &\le e^{-\nu_m q}. \end{aligned} \tag{$\star\star$}$$

When $A = \{i_1 < \cdots < i_s\} \subseteq \{1, \dots, m\}$, let $\pi_A : \mathbb{R}^n \to \mathbb{R}^s$ be the map with

$$\pi_A((p_0, p_1, \dots, p_m)) = (p_{i_1}, \dots, p_{i_s}) \in \mathbb{R}^s.$$

Proposition 1.1 and its generalization to successive minima with respect to $\Lambda(\boldsymbol{\xi})$ and $\mathcal{B}^\nu(q)$ were proved in [Schmidt and Summerer 2009] by showing that the assumption of Theorem 1.1, page 69 of that paper is fulfilled for $\Lambda(\xi)$ and $\mathcal{B}^\nu(q)$ if $1, \xi_1, \dots, \xi_m$ are linearly independent over $\mathbb{Q}$. For the convenience of the reader we state this result here in the present notation:

Theorem 1.2. *Suppose for every s-dimensional space S spanned by lattice points (i.e., points of Λ), there is some $A \subseteq \{1, \dots, m\}$ of cardinality s with $\pi_A(S) = \mathbb{R}^s$. Then there are arbitrarily large values of q with $\lambda_s(q) = \lambda_{s+1}(q)$.*

The question of whether the condition in Theorem 1.2 and the condition of linear independence of $1, \xi_1, \dots, \xi_m$ in Proposition 1.1 are also necessary to guarantee that for given s we have arbitrarily large values of q with $\lambda_s(q) = \lambda_{s+1}(q)$ (in the cases ($\star$) and ($\star\star$)) was the major motivation for the subsequent investigations. Regarding the set of exponents, we will without loss of generality suppose that

$$0 < \nu_1 \le \nu_2 \le \cdots \le \nu_m \tag{1-0}$$

in addition to $\nu_1 + \cdots + \nu_m = 1$.

It will follow from our exposition that in the standard simultaneous approximation case ($\star$) where $\nu_i = 1/m$ we have $\lambda_{n-1}(q) = \lambda_n(q)$ for some arbitrarily large q if and only if the linear independence condition is satisfied, in particular:

Corollary 1.3. *Suppose $\xi_1, \xi_2, \dots, \xi_m$ are real numbers with $\xi_k = \xi_{k+1}$ for some $1 \le k \le m$, and $\xi_1, \dots, \xi_{k-1}, \xi_{k+1}, \dots, \xi_m$ together with 1 are linearly independent over $\mathbb{Q}$, and let $\lambda_i(q)$, $1 \le i \le n$, denote the successive minima with respect to $\Lambda(\boldsymbol{\xi})$ and $\mathcal{B}(q)$. Then $\lambda_{n-1}(q) < \lambda_n(q)$ for all sufficiently large q.*

On the other hand, if $\xi_k = \xi_{k+1}$ and $\mathcal{B}(q)$ is replaced by $\mathcal{B}^\nu(q)$ with ν_m sufficiently large compared to ν_{m-1}, the situation may be different. In fact, for $\xi_1 = \xi_2$ in the three-dimensional case (i.e., $m = 2$) we will give a bound for ν_2 that guarantees $\lambda_2(q) = \lambda_3(q)$ for some arbitrarily large q in Section 4. All these particular cases of simultaneous approximation to linearly dependent reals fit in the general situation where for some k and r the real numbers $\xi_k, \xi_{k+1}, \dots, \xi_{k+r-1}$ are linear combinations of $1, \xi_{k+r}, \dots, \xi_m$ with rational coefficients. For this setting we will state conditions that guarantee that $\lambda_{n-r}(q) = \lambda_{n-r+1}(q)$ in Section 2. The proof of this result will be given in Section 3.

2. Basic notation and statement of the main result

We fix some exponents $(1, -\nu_1, \dots, -\nu_m)$ with $\nu_1 + \cdots + \nu_m = 1$ satisfying (1-0) in (⋆⋆) and write $\mathcal{B}(q)$ briefly for the body introduced as $\mathcal{B}^{\nu}(q)$ in the Introduction. Moreover we choose $r \in \{1, \dots, m-1\}$ and $k \in \{1, \dots, m-1-r\}$, set $s := n - r$ and define the sets $B := \{k, \dots, k+r-1\}$, $C := \{0, 1, \dots, m\} \setminus B$, $D := \{0, k+r, \dots, m\}$ with cardinalities

$$|B| = r, \quad |C| = s, \quad |D| = s - k + 1,$$

as well as $C' := C \setminus \{0\}$, $D' := D \setminus \{0\}$. Also let

$$\nu_B := \sum_{i \in B} \nu_i, \quad \nu_{C'} := \sum_{i \in C'} \nu_i,$$

so that $\nu_B + \nu_{C'} = 1$.

We will now consider the case of linearly dependent components ξ_i, more precisely the case where

$$\xi_j = \mathcal{L}_j(1, \xi_1, \dots, \xi_m) \quad \text{for } j \in B, \tag{2-0}$$

with r linear forms

$$\mathcal{L}_j(p_0, p_1, \dots, p_m) = \sum_{i \in D} c_i^{(j)} p_i$$

with rational coefficients $c_i^{(j)}$ so that $\xi_j = c_0^{(j)} + \sum_{i \in D'} c_i^{(j)} \xi_i$. Further put

$$c^{(j)} := \sum_{i \in D} |c_i^{(j)}| \quad \text{as well as} \quad c := \max\bigl(1, \max_{j \in B} c^{(j)}\bigr),$$

and let d be the least common denominator of the $c_i^{(j)}$ with $j \in B$, $i \in D$. Note that d as well as c depend only on the coefficients of the system (2-0).

To any m-tuple $(\xi_1, \dots, \xi_m)$ we had already associated the lattice $\Lambda = \Lambda(\boldsymbol{\xi})$ of points $p(\boldsymbol{x}) := (x, \xi_1 x - y_1, \dots, \xi_m x - y_m)$, with $\boldsymbol{x} := (x, y_1, \dots, y_m) \in \mathbb{Z}^n$, and the successive minima $\lambda_1(q), \dots, \lambda_n(q)$ with respect to $\mathcal{B}(q)$. We will write $L_i(q) = \log(\lambda_i(q))$ for $i = 1, \dots, n$ so that by Minkowski's second theorem

$$L_1(q) + \cdots + L_n(q) \le 0. \tag{2-1}$$

Now let S be the s-dimensional subspace of $\mathbb{R}^n$ spanned by the lattice points with $y_j = \mathcal{L}_j(x, y_1, \dots, y_m)$ for $j \in B$. Further we write S^C for the s-dimensional space of points with coordinates η_i, where $i \in C$, and let $\Lambda^C \subseteq S^C$ denote the s-dimensional lattice $\pi_C(\Lambda)$ consisting of points

$$(x, \xi_1 x - y_1, \dots, \xi_{k-1} x - y_{k-1}, \xi_{k+r} x - y_{k+r}, \dots, \xi_m x - y_m),$$

with $(x, y_1, \dots, y_{k-1}, y_{k+r}, \dots, y_m) \in \mathbb{Z}^s$. Let $\mathcal{B}^C(q) \subseteq S^C$ be the box with

$$|\eta_0| \le e^q, \qquad |\eta_i| \le e^{-\nu_i q} \quad (i \in C').$$

This box has volume $2^s e^{q - \nu_C q} = 2^s e^{\nu_B q}$. We will also need the successive minima $\lambda_j^C(q)$ as well as their logarithms $L_j^C(q)$, $1 \le j \le s$, that are defined in terms of $\mathcal{B}^C(q)$ and Λ^C. Minkowski's second theorem then implies

$$-\nu_B q - n \log n < L_1^C(q) + \cdots + L_s^C(q). \tag{2-2}$$

Note that in the present situation the condition of Theorem 1.2 is not fulfilled for the s-dimensional subspace S defined above. In fact, for any $A \subset \{1, \dots, m\}$ of cardinality s we have $|A^c| = r$ and A^c contains 0. Now S is the span of lattice points with $y_j = \mathcal{L}_j(x, y_1, \dots, y_m)$ for $j \in B$ and in view of (2-0) these lattice points have

$$\xi_j x - y_j = \mathcal{L}_j(0, x\xi_1 - y_1, \dots, x\xi_m - y_m), \quad j \in B.$$

This may be interpreted as a system of r linear equations among the $p_i = x\xi_i - y_i$, with $i \in B \cup D'$. As $0 \notin B \cup D'$, at most $r-1$ of these indices are not in A. It follows that the p_i with $i \in (B \cup D') \cap A$ satisfy at least $r - (r-1)$ linear relations; hence the projection $\pi_A : S \to \mathbb{R}^s$ is not surjective.

However it will turn out that the condition is not necessary for the conclusion $\lambda_s(q) = \lambda_{s+1}(q)$ for arbitrarily large q. More precisely we will show:

Theorem 2.1. *Let $\xi_1, \xi_2, \dots, \xi_m$ be real numbers satisfying (2-0) and $s = n - r$ as already defined.*

(a) *The relation*

$$L_s^C(q) \le v_k q - \log c - 2\log d - 1 \tag{2-3}$$

implies $L_s(q) < L_{s+1}(q)$. If (2-3) holds for every large q, and $\{\xi_i : i \in C'\}$ together with 1 are linearly independent over $\mathbb{Q}$, then for each $j < s$ there are arbitrarily large values of q with $L_j(q) = L_{j+1}(q)$.

(b) *Assume that (2-3) is fulfilled for certain arbitrarily large q and that for some (other) arbitrarily large q we have*

$$L_s^C(q) \ge v_B q + n^2. \tag{2-4}$$

Then there exist arbitrarily large q with $L_s(q) = L_{s+1}(q)$.

In the special case where (2-0) is reduced to

$$\xi_k = \cdots = \xi_{k+r}, \tag{2-5}$$

we have $\mathcal{L}_j(1, \xi_1, \dots, \xi_m) = \xi_{k+r}$ so that $c_{k+r}^{(j)} = 1$ for $j = k, \dots, k+r-1$ and all other coefficients are zero so that obviously $c = d = 1$. As (2-5) clearly implies $\xi_{k+l} = \cdots = \xi_{k+r}$ for any $l \in \{1, \dots, r\}$, we may as well apply the above results with $\widetilde{B} := \{k+l, \dots, k+r-1\}$ and $\widetilde{C} := \{0, 1, \dots, m\} \setminus \widetilde{B}$. In this way we see that the relation

$$L_{s+l}^{\widetilde{C}}(q) \le v_{k+l} q - 1 \tag{2-6}$$

implies $L_{s+l}(q) < L_{s+l+1}(q)$ and that the fact (2-6) is fulfilled for certain arbitrarily large q together with

$$L_s^{\widetilde{C}}(q) \ge v_{\widetilde{B}} q + n^2 \tag{2-7}$$

for some other arbitrarily large q guarantees that there exist arbitrarily large q with $L_{s+l}(q) = L_{s+l+1}(q)$.

These results highlight the interest of considering parametric geometry of numbers in a more general context than the classical simultaneous approximation problem as initiated in [Schmidt and Summerer 2009] and investigated in much more detail in [Schmidt ≥ 2019].

3. Deduction of Theorem 2.1

Assume that (2-0) holds for $\xi_1, \xi_2, \ldots, \xi_m$ and keep all notation as introduced in Section 2. For points $p(\boldsymbol{x})$ in $\Lambda \cap S$ with $\pi_C(p(\boldsymbol{x})) \in \mathcal{B}^C(q)$ we get for $j \in B$

$$\begin{aligned}|\xi_j x - y_j| &= |\mathcal{L}_j(1, \xi_1, \ldots, \xi_m)x - \mathcal{L}_j(x, y_1, \ldots, y_m)| \\ &\le |c_{k+r}^{(j)}|\,|\xi_{k+r}x - y_{k+r}| + \cdots + |c_m^{(j)}|\,|\xi_m x - y_m| \\ &\le |c_{k+r}^{(j)}|e^{-\nu_{k+r}q} + \cdots + |c_m^{(j)}|e^{-\nu_m q} \\ &\le c^{(j)}e^{-\nu_{k+r}q} \\ &\le c^{(j)}e^{-\nu_j q} \end{aligned} \tag{3-0}$$

for large q in view of (1-0). Hence by the definition of c we have $p(\boldsymbol{x}) \in c\mathcal{B}(q)$. So if $\lambda\mathcal{B}^C(q)$ contains s linearly independent points of Λ^C, then $c\lambda\mathcal{B}(q)$ contains s linearly independent points $p(\boldsymbol{x})$ where $\boldsymbol{x} \in d^{-1}\mathbb{Z}^n$ and thus $dc\mathcal{B}(q)$ contains s linearly independent points $p(\boldsymbol{x})$ of $\Lambda \cap S$. It follows that $\lambda_s(q) \le dc\lambda_s^C(q)$ and consequently

$$L_s(q) \le L_s^C(q) + \log c + \log d. \tag{3-1}$$

In combination with (2-3) that we assume in (a), (3-1) yields

$$L_s(q) \le \nu_k q - \log d - 1. \tag{3-2}$$

On the other hand, points in Λ outside S have $y_{j_0} \ne \mathcal{L}_j(x, y_1, \ldots, y_m)$ for at least one $j_0 \in B$, so that $|\mathcal{L}_j(x, y_1, \ldots, y_m) - y_{j_0}| \ge d^{-1}$. This implies

$$\begin{aligned}|\xi_{j_0}x - y_{j_0}| &= |\mathcal{L}_{j_0}(1, \xi_1, \ldots, \xi_m)x - y_{j_0}| \\ &= |\mathcal{L}_{j_0}(1, \xi_1, \ldots, \xi_m)x - \mathcal{L}_{j_0}(x, y_1, \ldots, y_m) + \mathcal{L}_{j_0}(x, y_1, \ldots, y_m) - y_{j_0}| \\ &\ge |\mathcal{L}_{j_0}(x, y_1, \ldots, y_m) - y_{j_0}| - |\mathcal{L}_{j_0}(1, \xi_1, \ldots, \xi_m)x - \mathcal{L}_{j_0}(x, y_1, \ldots, y_m)| \\ &\ge d^{-1} - c^{(j_0)}e^{-\nu_{j_0}q}\end{aligned}$$

and hence $|\xi_{j_0}x - y_{j_0}| \ge d^{-1} - ce^{-\nu_k q}$ by the definition of c and (1-0). Denoting by $\lambda_{\boldsymbol{x}}(q)$ the least $\lambda > 0$ with $p(\boldsymbol{x}) \in \lambda\mathcal{B}(q)$ and writing $L_{\boldsymbol{x}}(q) = \log \lambda_{\boldsymbol{x}}(q)$, we thus have

$$\lambda_{\boldsymbol{x}}(q) = \inf_{\boldsymbol{x} \in \lambda\mathcal{B}(q)} \lambda \ge d^{-1}e^{\nu_k q} - c$$

for $p(\boldsymbol{x}) \in \Lambda \setminus S$, so that any lattice point outside S has

$$L_{\boldsymbol{x}}(q) > \nu_k q - \log d - 1 \tag{3-3}$$

for sufficiently large q, so that certainly

$$L_{s+1}(q) > \nu_k q - \log d - 1. \tag{3-4}$$

Together (3-2) and (3-4) imply $L_s(q) < L_{s+1}(q)$, i.e., the first assertion of (a).

To prove the second assertion of (a) and part (b) we introduce the function

$$G(q) := \min_{\boldsymbol{x}\in\Lambda\setminus S} L_{\boldsymbol{x}}(q),$$

which by (3-3) satisfies

$$G(q) > \nu_k q - \log d - 1, \tag{3-5}$$

and is continuous and piecewise linear. In particular, for those q for which (2-3) holds we have $L_s^C(q) < G(q)$ and thus

$$L_j(q) = L_j^C(q) \tag{3-6}$$

for all $j \le s$.

Now assume that (2-3), hence (3-6), holds for all large q. If $\{\xi_i : i \in C'\}$ together with 1 are linearly independent over $\mathbb{Q}$ then Proposition 1.1 applied to simultaneous approximation of $\{\xi_i : i \in C'\}$, i.e., successive minima defined with respect to Λ^C and $\mathcal{B}^C(q)$, implies the existence of arbitrarily large q with $L_j^C(q) = L_{j+1}^C(q)$ for any $j < s$. In combination with (3-6) the second assertion of (a) follows.

In general, given any q, at least one of $L_1(q), \dots, L_{s+1}(q)$ will stem from a point $p(\boldsymbol{x})$ outside S, say $L_l(q) = L_{\boldsymbol{x}}(q)$ with $p(\boldsymbol{x}) \notin S$, where l is chosen minimal subject to this property. Note that the definition of l implies that (3-6) now holds for $i = 1, \dots, l-1$.

If $l = s+1$, it follows from (2-2) that

$$L_1^C(q) + \cdots + L_s^C(q) > -\nu_B q - n^2 \tag{3-7}$$

and by the definition of G combined with (2-4)

$$L_{s+1}(q) = G(q) > L_s^C(q) > \nu_B q + n^2 \tag{3-8}$$

holds for certain arbitrarily large $q = q_0$. Together (3-6)–(3-8) would imply

$$L_1(q_0) + \cdots + L_{s+1}(q_0) > 0,$$

and as $0 < L_{s+1}(q_0) \le L_{s+2}(q_0) + \cdots + L_n(q_0)$ this would contradict (2-1).

If $l \le s$ then (2-1) yields

$$L_1(q) + \cdots + L_{l-1}(q) + (n-l+1)G(q) \le 0,$$

which can be rephrased as

$$\begin{aligned}(n-l+1)G(q) &\le -L_1(q) - \cdots - L_{l-1}(q)\\ &= -L_1^C(q) - \cdots - L_{l-1}^C(q) && \text{(by (3-6))}\\ &< L_l^C(q) + \cdots + L_s^C(q) + \nu_B q + n^2 && \text{(by (2-2))}\\ &\le (s+1-l)L_s^C(q) + \nu_B q + n^2.\end{aligned}$$

For $q = q_0$ with (2-4) this yields $(n-l+1)G(q) \le (s-l+2)L_s^C(q_0)$; therefore

$$G(q_0) < \frac{s-l+2}{n-l+1} L_s^C(q_0) \le L_s^C(q_0)$$

for some arbitrarily large q_0 since $s \le n-1$ by definition. By assumption there are also arbitrarily large q_1 with (2-3) for which we have $L_s^C(q_1) < G(q_1)$, as already noticed. Since L_s^C as well as G are continuous, there will be some q in (q_0, q_1) with

$$L_s^C(q) = G(q). \tag{3-9}$$

Since S has dimension s, we have $L_{s+1}(q) \ge G(q)$ for every q. There are s linearly independent lattice points $p(\boldsymbol{x})$ in S with $L_{\boldsymbol{x}}(q) \le L_s^C(q)$, as well as a lattice point $\boldsymbol{x} \notin S$ with $L_{\boldsymbol{x}}(q) = G(q)$, so that by (3-9) we have $L_{s+1}(q) \le G(q)$; hence $L_{s+1}(q) = G(q)$. Also there are fewer than s independent lattice points $p(\boldsymbol{x})$ with $L_{\boldsymbol{x}}(q) < L_s^C(q)$ so that $L_s(q) = L_s^C(q)$. Therefore $L_s(q) = L_{s+1}(q)$; hence (b) is proved.

4. Another version of Theorem 2.1

In order to apply Theorem 2.1 it is essential to be able to check whether the conditions (2-3) and (2-4) are fulfilled for the given ξ_i and the given exponents. For this purpose, let us first replace the functions $L_s^C(q)$ defined with respect to $\mathcal{B}^C(q)$ by functions $\hat{L}_s^C(q)$ defined with respect to a set $\hat{\mathcal{B}}^C(q)$ of volume 2^s.

Define ρ and σ by

$$\rho(s - \nu_B) = s \quad \text{and} \quad \sigma = \rho - 1. \tag{4-0}$$

For $i \in C$ set $\mu_i := \rho\nu_i + \sigma$ so that

$$\begin{aligned}\sum_{i\in C}\mu_i &= \rho\nu_C + (s-1)\sigma \\ &= \rho(1 - \nu_B + s - 1) + 1 - s = \rho(s-\nu_B) - s + 1 = 1\end{aligned}$$

by (4-0). The box $\hat{\mathcal{B}}^C(q)$ is now defined by

$$|\eta_0| \le e^q, \qquad |\eta_i| \le e^{-\mu_i q} \quad (i \in C'),$$

which may also be written as

$$|\eta_0| \le e^{-\sigma q + \rho q}, \qquad |\eta_i| \le e^{-\sigma q - \rho\nu_i q} \quad (i \in C').$$

Thus $\hat{\mathcal{B}}^C(q)$ is $e^{-\sigma q}\mathcal{B}^C(\rho q)$. The corresponding quantities $\hat{L}_j^C(q)$ for $1 \le j \le s$ have

$$\hat{L}_j^C(q) = \sigma q + L_j^C(\rho q).$$

Therefore (2-3) becomes

$$\begin{aligned}\hat{L}_s^C(q) &\le \sigma q + \rho\nu_k q - \log c - 2\log d - 1 \\ &= (\rho(1+\nu_k) - 1)q - \log c - 2\log d - 1 \\ &= \frac{s\nu_k + \nu_B}{s - \nu_B} q - \log c - 2\log d - 1.\end{aligned}$$

Moreover (2-4) becomes

$$\hat{L}_s^C(q) \ge \sigma q + \rho\nu_B q + n^2 = (\rho(1+\nu_B) - 1)q + n^2 = \frac{(s+1)\nu_B}{s-\nu_B} q + n^2.$$

We may thus rewrite Theorem 2.1 as:

Corollary 4.1. *Let $\xi_1, \xi_2, \dots, \xi_m$ be real numbers satisfying (2-0).*

(a) *The relation*

$$\hat{L}_s^C(q) \le \frac{s\nu_k + \nu_B}{s - \nu_B} q - \log c - 2\log d - 1 \tag{4-1}$$

implies $L_s(q) < L_{s+1}(q)$. If (4-1) holds for every large q, and $\{\xi_i : i \in C'\}$ together with 1 are linearly independent over $\mathbb{Q}$, then for each $j < s$ there are arbitrarily large values of q with $L_j(q) = L_{j+1}(q)$.

(b) *Assume that (4-1) is fulfilled for certain arbitrarily large q and that for some (other) arbitrarily large q we have*

$$\hat{L}_s^C(q) \ge \frac{(s+1)\nu_B}{s - \nu_B} q + n^2. \tag{4-2}$$

Then there exist arbitrarily large q with $L_s(q) = L_{s+1}(q)$.

In this reformulation of the main result, the conditions to check, i.e., (4-1) and (4-2), are concerned with the functions $\hat{L}_i^C(q)$, whose behaviour is rather well understood in the case where they stem from a classical simultaneous approximation problem in lower dimension, hence when all $\mu_i,\ i \in C$ are equal, which amounts to all $\nu_i,\ i \in C$, are equal.

In particular, when all ν_i are equal this leads to the deduction Corollary 1.3: (2-0) reduces to the equation $\xi_k = \xi_{k+1}$, which is of the form (2-5) and we have $B = \{k\}$; hence $C' = \{1, \dots, k-1, k+1, \dots, m\}$ and thus $s = n - 1 = m$. Moreover in the case of classical simultaneous approximation one has $\nu_i = 1/m$ for $i = 1, \dots, m$ so that relation (4-1) reads

$$\hat{L}_m^C(q) \le \frac{1 + 1/m}{m - 1/m} q - 1 = \frac{1}{m-1} q - 1. \tag{4-3}$$

We claim that this relation holds for all sufficiently large q, so that assertion (a) of Corollary 4.1 yields $L_m(q) = L_{n-1}(q) < L_n(q)$ for all large q. Indeed for the simultaneous approximation of $m-1$ linearly independent reals, here these are $\xi_1, \dots, \xi_{k-1}, \xi_{k+1}, \dots, \xi_m$, one always has $\hat{L}_m^C(q) < q/(m-1) - g(q)$ for some function g tending to infinity (see [Schmidt and Summerer 2009], page 77, equation (4.9)), which implies (4-3).

Our next example deals with a case where not all the ν_i are identical and shows the existence of $\xi_1, \dots, \xi_m$ and exponents $\nu_1, \dots, \nu_m$ for which the intersection properties of the successive minima functions with respect to $\mathcal{B}^\nu(q)$ differ from those with respect to $\mathcal{B}(q)$.

We consider the case $m = 2$ of simultaneous approximation to (ξ, ξ), where ξ is an irrational number with $\omega(\xi) > 1$. Here $\omega(\xi)$ is the supremum of all η such that there are arbitrarily large values of Q for which $|\xi x - y| \le Q^{-\eta}$ has a nontrivial integer solution (x, y) with $|x| \le Q$. Then the (single) approximation constant

$$\bar{\varphi}_2(\xi) = \frac{\omega - 1}{\omega + 1}$$

(as defined in [Schmidt and Summerer 2013], page 3) has $\bar{\varphi}_2(\xi) > 0$. By Corollary 1.3 applied in the case $\xi_1 = \xi_2 = \xi$, i.e., for classical simultaneous approximation to (ξ, ξ), we have $\lambda_1(q) = \lambda_2(q)$ for some arbitrarily large q since ξ is irrational, whereas $\lambda_2(q) < \lambda_3(q)$ for all sufficiently large q.

We claim that this will not be the case for approximation relative to exponents (ν_1, ν_2) provided ν_2 is sufficiently large.

Corollary 4.2. *Let ξ be an irrational number with $\bar{\varphi}_2(\xi) > 0$ and let (ν_1, ν_2) be a system of exponents with*

$$\nu_2 > \frac{3 - \bar{\varphi}_2(\xi)}{3 + \bar{\varphi}_2(\xi)}.$$

Then for $s \in \{1, 2\}$ there exist arbitrarily large $q = q(s)$ with $L_s(q) = L_{s+1}(q)$.

Proof. For $s = 1$ this is clear by the irrationality of ξ. So let $s = 2$ and apply Corollary 4.1 with $B = \{1\}$ and $C = \{2\}$ so that $s = 2$ and $\nu_B = 1 - \nu_2$. Note that by the definition of $\bar{\varphi}_2(\xi)$ and $\hat{\mathcal{B}}^C(q)$ we have $\limsup_{q\to\infty} \hat{L}_2^C(q)/q = \bar{\varphi}_2(\xi)$.

Moreover $c = d = 1$ so that (4-1) reads

$$\hat{L}_2^C(q) \le \frac{3 - 3\nu_2}{1 + \nu_2} q - 1,$$

which is certainly fulfilled for some arbitrarily large q as $3 - 3\nu_2 > 0$ and $\liminf_{q\to\infty} \hat{L}_2^C(q)/q = 0$ for single approximation.

On the other hand (4-2) becomes

$$\hat{L}_2^C(q) \ge \frac{3 - 3\nu_2}{1 + \nu_2} q + n^2,$$

which is fulfilled for certain arbitrarily large q provided

$$\frac{3 - 3\nu_2}{1 - +nu_2} < \bar{\varphi}_2(\xi) \quad \Longleftrightarrow \quad \nu_2 > \frac{3 - \bar{\varphi}_2(\xi)}{3 + \bar{\varphi}_2(\xi)}.$$

So part (b) of Corollary 4.1 implies $L_2(q) = L_3(q)$ for some arbitrarily large q as desired. □

It remains to say a few words on the case where the ν_i, $i \in C$, are distinct. Then the μ_i will be as well and it is not clear how to check conditions (4-1) and (4-2) when the functions $\hat{L}_s^C(q)$ do not stem from classical simultaneous approximation. However in [Schmidt ≥ 2019] a very precise description of the possible behaviour of the successive minima functions defined with respect to $\Lambda(\xi)$ and $\mathcal{B}^\nu(q)$ is sketched. In order to show the existence of real numbers for which those successive minima functions follow a prescribed behaviour, an appropriate analogue of Roy's results [2015, Theorem 1.3, Corollary 1.4] for generalized systems of exponents would be needed. This would considerably broaden the range of applications of the results in this paper.

References

[Roy 2015] D. Roy, "On Schmidt and Summerer parametric geometry of numbers", *Ann. of Math.* (2) **182**:2 (2015), 739–786. MR Zbl

[Schmidt ≥ 2019] W. M. Schmidt, "On parametric geometry of numbers", preprint. To appear in *Acta Arith.*

[Schmidt and Summerer 2009] W. M. Schmidt and L. Summerer, "Parametric geometry of numbers and applications", *Acta Arith.* **140**:1 (2009), 67–91. MR Zbl

[Schmidt and Summerer 2013] W. M. Schmidt and L. Summerer, "Diophantine approximation and parametric geometry of numbers", *Monatsh. Math.* **169**:1 (2013), 51–104. MR Zbl

Received 17 Oct 2018. Revised 20 Mar 2019.

LEONHARD SUMMERER:

leonhard.summerer@univie.ac.at
Faculty of Mathematics, University of Vienna, Vienna, Austria

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 3, 2019
dx.doi.org/10.2140/moscow.2019.8.229

On the distribution of values of Hardy's Z-functions in short intervals II: The q-aspect

Ramdin Mawia

We continue our investigations regarding the distribution of positive and negative values of Hardy's Z-functions $Z(t,\chi)$ in the interval $[T, T+H]$ when the conductor q and T both tend to infinity. We show that for $q \leqslant T^{\eta}$, $H = T^{\vartheta}$, with $\vartheta > 0$, $\eta > 0$ satisfying $\frac{1}{2}+\frac{1}{2}\eta < \vartheta \leqslant 1$, the Lebesgue measure of the set of values of $t \in [T, T+H]$ for which $Z(t,\chi) > 0$ is $\gg (\varphi(q)^2/4^{\omega(q)}q^2)H$ as $T \to \infty$, where $\omega(q)$ denotes the number of distinct prime factors of the conductor q of the character χ, and φ is the usual Euler totient. This improves upon our earlier result. We also include a corrigendum for the first part of this article.

1. Introduction

Let χ be a primitive Dirichlet character of conductor $q > 1$. In this paper, we continue our investigations in [Mawia 2017] concerning the distribution of positive and negative values of Hardy's Z-function $Z(t,\chi)$ for t in the interval $[T, T+H]$. Let us recall some basic facts concerning the function $Z(t,\chi)$. First of all, Hardy's Z-function $Z(t,\chi)$ corresponding to the Dirichlet L-function $L(s,\chi)$ is defined by

$$Z(t,\chi) := \Psi\big(\tfrac{1}{2}+it,\chi\big)^{-1/2} L\big(\tfrac{1}{2}+it,\chi\big),$$

where

$$\Psi(s,\chi) = \mathfrak{w}(\chi)\left(\frac{\pi}{q}\right)^{s-1/2} \frac{\Gamma((1-s+\mathfrak{a})/2)}{\Gamma((s+\mathfrak{a})/2)}$$

is the factor from the functional equation $L(s,\chi) = \Psi(s,\chi)L(1-s,\bar{\chi})$. Here the quantity $\mathfrak{a} = (1-\chi(-1))/2$ measures the parity of the character χ and the number

$$\mathfrak{w}(\chi) = \frac{\tau(\chi)}{i^{\mathfrak{a}}\sqrt{q}}, \quad \text{with } \tau(\chi) = \sum_{a \pmod q} \chi(a) e\left(\frac{a}{q}\right),$$

is called the root number of χ. It is immediately seen from the definition that $Z(t,\chi)$ is a real-valued function of the real variable t and that $|Z(t,\chi)| = \big|L\big(\frac{1}{2}+it,\chi\big)\big|$, so that the real zeros of $Z(t,\chi)$ correspond to the zeros of $L(s,\chi)$ on the critical line. One of the main interests of Hardy's Z-functions comes from this fact. For a brief introduction to Hardy's Z-function for $\zeta(s)$, see [Karatsuba and Voronin 1992, III, §4]; a more comprehensive theory is developed in [Ivić 2013].

Although it is a well-known tool in the computations and study of the distributions of zeros of the Riemann zeta function (see for example [Karatsuba 1981]), research into its fine structures is a rather recent development, due mainly to results of Ivić [2004; 2010; 2017a; 2017b; 2017c], Korolev [2007;

MSC2010: 11M06, 11M26.
Keywords: Hardy's function, Hardy–Selberg function, Dirichlet L-function, value distribution.

2008; 2017] and Jutila [2009; 2011]. Our interest here, following [Gonek and Ivić 2017] and later [Mawia 2017] is in the distribution of positive and negative values of $Z(t, \chi)$. In [Mawia 2017], we proved the following result:

Theorem. *Consider a fixed primitive Dirichlet character χ of conductor $q > 1$. For $2 \leqslant H \leqslant T$, let*

$$I_+(T, H; \chi) = \{T \leqslant t \leqslant T + H : Z(t, \chi) > 0\},$$
$$I_-(T, H; \chi) = \{T \leqslant t \leqslant T + H : Z(t, \chi) < 0\}.$$

Fix $0 < \epsilon < \frac{1}{4}$ and let $T^{3/4+\epsilon} \leqslant H \leqslant T$. Then we have

$$\mu(I_+(T, H; \chi)) \gg H \quad \text{and} \quad \mu(I_-(T, H; \chi)) \gg H,$$

where μ is the Lebesgue measure on the line.

It should be remarked that the constants in the above result depend on the conductor q, as will be clear in the course of this paper. Our main concern here will be to take into account the variation of the conductor q and see to what extent results of the above type hold true when q is allowed to vary. For example, one may ask questions of the following type: does the above result remain valid when q is not held fixed but allowed to vary, say up to $q \leqslant T$, or for that matter, as $q \to \infty$ independently of T? The main result we prove in this article is the following.

Theorem 1. *Let $q \leqslant T^{\eta}$, $H = T^{\vartheta}$ with $\frac{1}{2} + \frac{1}{2}\eta < \vartheta \leqslant 1$. Then, as $T \to \infty$ we have*

$$\mu(I_\pm(T, H; \chi)) \gg \frac{\varphi(q)^2}{4^{\omega(q)} q^2} H$$

uniformly in q, T and H, where the implied constants depend only on η, ϑ.

Thus, this result says that the main result of [Mawia 2017] holds for q almost as big as T and improves the interval as well by reducing $\frac{3}{4}$ to $\frac{1}{2}$. Note in particular that when q tends to infinity through integers with a bounded number of prime factors, $\varphi(q)^2/4^{\omega(q)}q^2 \asymp 1$ and hence $\mu(I_\pm(T, H; \chi)) \gg H$ as $T \to \infty$ and $q \to \infty$ along integers with a bounded number of prime factors, say $q \to \infty$ via primes, for example.

To prove this result, although the main argument for the result is the same as in [Mawia 2017], we have to explicitly show the dependence on q of the error terms in all our lemmas there, and that makes the argument more delicate. The improvement on the interval comes mainly from an application of the first derivative bound rather than the second derivative bound for exponential integrals in Lemmas 7 and 8. Following [Gonek and Ivić 2017], our main tool will be a study of various mollified integrals of $Z(t, \chi)$, using the mollifiers $B_X(s, \chi)$ analogous to the ones introduced for $\zeta(s)$ by Selberg [1942]. An obvious merit of this mollifier may be illustrated as follows. By following the steps of [Gonek and Ivić 2017] but without using the mollifiers they used, and applying the well-known bounds

$$\int_T^{2T} Z(t)\,\mathrm{d}t \ll T^{7/8}, \quad \int_T^{2T} |Z(t)|\,\mathrm{d}t \gg T, \quad \int_T^{2T} Z(t)^2\,\mathrm{d}t = \int_T^{2T} \left|\zeta\left(\tfrac{1}{2} + it\right)\right|^2 \mathrm{d}t \asymp T \log T,$$

one only obtains $\mu(I_+(T, 2T)) \gg T/\log T$, which is rather weak. In this light, it would be interesting to use other mollifiers and see what they yield. However, the mollifiers introduced by Selberg seem to be most effective in several circumstances. The function $F(t) = Z(t)\left|B_X\left(\frac{1}{2} + it\right)\right|^2$, which is a mollification

of $Z(t)$, and is the main object in our lemmas for proving the above theorem, is also called the Hardy–Selberg function [Karatsuba and Voronin 1992, III, §5], and we shall follow this nomenclature introduced by Karatsuba [1984].

Once the lemmas on the mollified integrals are proved, the proof of the main result, Theorem 1 above, is the same as in [Gonek and Ivić 2017] (which is repeated in [Mawia 2017, §2]), so we content ourselves here with a precise formulation and proof of the requisite lemmas.

This paper is organised as follows. In Section 2, we restate and prove some lemmas on approximations of $L(s, \chi)$ by Dirichlet polynomials, and on the coefficients $\alpha_n(\chi)$ of our mollifiers. The three main lemmas essential for proving the above theorem are then proved in Section 3. Also, throughout this paper, we use, without comment, bounds on several sums which are either standard or which appeared already in [Mawia 2017], although they may not be obvious.

Corrigendum to [Mawia 2017]. There are a couple of minor errors in [Mawia 2017] which are not difficult to correct. They stem from an erroneous statement of the Riemann–Siegel formula for $Z(t, \chi)$, which is equation (5) in Lemma 3, p. 39. The error is that the variable n in the sum for $\Theta(t, \chi)$ should run up to $n \leqslant \sqrt{qt/(2\pi)}$, not up to $n \leqslant \sqrt{qT/(2\pi)}$. The correct statement is given in (2-1) of Lemma 2 below. A consequence of this error is that our treatment of the integrals in equation (10), p. 44 was erroneous. This problem is addressed in Lemma 9 below. Next, immediately before equation (11) on p. 45, "the first O-term in (3)" and "the first two integrals in (3)" should be replaced by "the first O-term in (10)" and "the first two integrals in (10)" respectively. Finally, from equation (12), p. 45 onwards the conditions $kn = lm$ and $kn \neq lm$ in the summation should be replaced by the conditions $km = ln$ and $km \neq ln$ respectively, and the fraction kn/lm should be replaced throughout by ln/km. As is clear, one does not need to change the subsequent argument.

2. Preliminary lemmas

As in [Mawia 2017], our main tool will be a form of the "Riemann–Siegel formula" for $Z(t, \chi)$, a version of which is stated in Lemma 3 in [Mawia 2017]; the proof of the "approximate functional equation" from which this can be derived is given in [Chandrasekharan and Narasimhan 1963; Suetuna 1932; Tchudakoff 1947]. The proof of the original Riemann–Siegel formula can be found in [Siegel 1932] (see also [Siegel 1966, pp. 275–310]). A more precise form for the terms of the asymptotic series for the remainder in the approximate functional equation for $L(s, \chi)$ is given in [Siegel 1943]. It should be remarked here that sharper forms of this formula are available, at least for the case $q = 1$ (see for example [Borwein et al. 2000; Gabcke 1979]); the approximate formula [Karatsuba and Voronin 1992, III, §4, Theorem 1] is also of interest, and similar formulas exist for the derivatives of $Z(t)$ as well [Karatsuba 1981]. We need a version which is uniform in both q- and t-aspects. Such an approximate functional equation for $L(s, \chi)$ is proved in [Lavrik 1968] (Corollary 1 to Theorem 1), from which it is easy to deduce the following corresponding result for $Z(t, \chi)$:

Lemma 2. *We have the following approximate equation for $Z(t, \chi)$:*

$$Z(t, \chi) = \Theta(t, \chi) + \overline{\Theta}(t, \chi) + O((q/t)^{1/4} \log(2t)), \tag{2-1}$$

where

$$\Theta(t, \chi) = \mathfrak{w}(\chi)^{-1/2} e^{(\pi i \mathfrak{a}/4) - (\pi i/8)} \left(\frac{qt}{2\pi e}\right)^{it/2} \sum_{n \leqslant \sqrt{qt/(2\pi)}} \frac{\chi(n)}{n^{1/2+it}}. \tag{2-2}$$

Here and in the following, we will restrict ourselves to $t > 0$, since the values of the Z-function for $L(s, \chi)$ at $t < 0$ correspond to the values of the Z-function for $L(s, \bar{\chi})$ at $t > 0$; precisely, we have $Z(-t, \chi) = Z(t, \bar{\chi})$.

As has already been noted in the corrigendum above, the expression (5) of Lemma 3 in [Mawia 2017] is erroneous and should be replaced by the above expression (2-1). As a result we have to modify our argument for proving [Mawia 2017, Lemma 8] accordingly; see Lemma 9 below.

Next, we shall need the following approximate functional equation for $L(s, |\chi|)$, whose proof is similar to the analogous formula for $\zeta(s)$.

Lemma 3. *Let $x > 1$ be a positive real number. We have*

$$L(s, |\chi|) = \sum_{n \leqslant x} \frac{|\chi(n)|}{n^s} - \frac{\varphi(q)}{q} \frac{x^{1-s}}{1-s} + O(2^{\omega(q)}|s|x^{-\sigma})$$

uniformly for $\sigma \geqslant \sigma_0 > 0$. Also, as $s \to 1$, we have

$$(s-1)L(s, |\chi|) = \frac{\varphi(q)}{q} + O(2^{\omega(q)}|s-1|).$$

Here, $\omega(q)$ denotes the number of distinct prime factors of q.

Proof. For $\sigma > 1$, we have

$$L(s, |\chi|) = \sum_{n \leqslant x} \frac{|\chi(n)|}{n^s} + \sum_{n > x} \frac{|\chi(n)|}{n^s}.$$

One writes the sum $\sum_{x < n \leqslant y} |\chi(n)| n^{-s}$ as an integral $\int_x^y u^{-s}\, \mathrm{d}A(u)$, where $A(u) = \sum_{n \leqslant u} |\chi(n)|$ counts the number of positive integers $\leqslant u$ which are coprime to q. Integrating by parts, we get that

$$\sum_{x < n \leqslant y} \frac{|\chi(n)|}{n^s} = \frac{A(y)}{y^s} - \frac{A(x)}{x^s} + \frac{\varphi(q)}{q} \frac{y^{1-s} - x^{1-s}}{1-s} s + s \int_x^y \frac{A(u) - \varphi(q)u/q}{u^{s+1}}\, \mathrm{d}u.$$

Letting $y \to \infty$, the resulting equation is valid for $\sigma \geqslant \sigma_0 > 0$. Noting that $A(u) = \varphi(q)u/q + O(2^{\omega(q)})$, the result follows. We observe here that in the formulas in this lemma, $2^{\omega(q)} \leqslant d(q) \ll q^{\epsilon}$. Note that $2^{\omega(q)} \ll 1$ if we let q tend to infinity along positive integers with a bounded number of distinct prime factors, which is the case for $q \to \infty$ via primes, for example. □

We remark here that we do not need any condition on the relative size of the imaginary part t of s and the length x of the approximating Dirichlet polynomial, unlike in the case of the slightly more refined result for $\zeta(s)$ found, for instance, in [Karatsuba and Voronin 1992, III, §2] or [Ivić 1985, §1.5], where the condition $|t| \leqslant \pi x$ is required.

We will again need a couple of lemmas regarding the Dirichlet coefficients of $L(s, \chi)^{-1/2}$, which we denote by $\alpha_n(\chi)$, following [Selberg 1942]. Recall that it may be explicitly expressed as

$$\alpha_n(\chi) = (-1)^{e_1 + \cdots + e_k} \binom{\frac{1}{2}}{e_1} \cdots \binom{\frac{1}{2}}{e_k} \chi(n), \tag{2-3}$$

where $n = p_1^{e_1} \cdots p_k^{e_k}$ is the factorisation of n. It is a multiplicative sequence satisfying $|\alpha_n(\chi)| \leqslant 1$ for all $n \geqslant 1$. Although not necessary for our applications, it is easily seen from the above explicit expression

that its sign is essentially the same as $\mu(\mathrm{rad}(n))$, where $\mathrm{rad}(n)$ is the radical of n, the product of the distinct prime factors of n. Precisely, we have $\alpha_n(\chi) = \mu(\mathrm{rad}(n))|\alpha_n(\chi)|\chi(n)$. We recall other notation already used in [Mawia 2017]. For $1 \leqslant n \leqslant X$ let

$$\beta_n \equiv \beta_n(\chi) = \alpha_n(\chi)\left(1 - \frac{\log n}{\log X}\right)$$

and define the mollifying functions $B_X(s, \chi)$ by $B_X(s, \chi) = \sum_{n \leqslant X} \beta_n n^{-s}$, which correspond to the sums $\eta(t)$ used by Selberg [1942]. We write

$$B_X(s, \chi)^2 = \sum_{n \leqslant X^2} b_n(\chi) n^{-s}, \tag{2-4}$$

where $b_n(\chi) = \sum_{d \mid n} \beta_d(\chi)\beta_{n/d}(\chi)$ with the sum running through those divisors d of n which satisfy both $d \leqslant X$ and $n/d \leqslant X$; note that $|b_n(\chi)| \leqslant d(n)$, where $d(n)$ is the usual divisor function.

We shall need the following analogue of Lemma 11 in [Selberg 1942].

Lemma 4. *Let ϱ be a positive integer and $0 \leqslant \gamma \leqslant 1$, write*

$$f(s) = f(s; \gamma, \varrho, \chi) = \frac{1}{\sqrt{(s + i\gamma)L(1 + s + i\gamma, |\chi|)}} \prod_{p \mid \varrho} (1 - |\chi(p)|p^{-1-s-i\gamma})^{-1/2}$$

and let, for $r = 2, 3$,

$$f(s) = \sum_{j=0}^{r-1} \frac{f^{(j)}(0)}{j!} s^j + s^r R_r(s).$$

Then for $s = it$, $-2 \leqslant t \leqslant 2$ and $j = 0, 1, 2$, we have

$$f^{(j)}(0) \ll \frac{2^{\omega(q)} q^{1/2}}{\varphi(q)^{1/2}} \prod_{p \mid \varrho} (1 + |\chi(p)|p^{-3/4}),$$

$$R_r(it) \ll \frac{2^{\omega(q)} q^{1/2}}{\varphi(q)^{1/2}} \prod_{p \mid \varrho} (1 + |\chi(p)|p^{-3/4}).$$

Using this, one can prove the following, which is an analogue of [Selberg 1942, Lemma 12].

Lemma 5. *Let $1 \leqslant d \leqslant X$, $0 \leqslant \gamma \leqslant 2^{-\omega(q)}\varphi(q)/(q\sqrt{\log X})$ and ϱ be a positive integer. Then, for $r = 1, 2$,*

$$\sum_{\substack{n \leqslant X/d \\ (n, \varrho) = 1}} \frac{\alpha_n(|\chi|)}{n^{1+i\gamma}} \left(\log \frac{X}{dn}\right)^r = c_r \sqrt{\frac{q\gamma}{\varphi(q)}} \prod_{p \mid \varrho} (1 - |\chi(p)|p^{-1-i\gamma})^{-1/2} \left(\log \frac{X}{d}\right)^r + O\left(\frac{2^{\omega(q)} q^{1/2}}{\varphi(q)^{1/2}} (\log X)^{r-1/2} \prod_{p \mid \varrho} (1 + |\chi(p)|p^{-3/4})\right), \tag{2-5}$$

where c_r is an absolute constant. Here and in the following, (m, n) denotes the gcd *of two integers m and n.*

We shall also need the following lemma, analogous to Lemma 13 in [Selberg 1942].

Lemma 6. *Let $X \geqslant 3$, $1 \leqslant d \leqslant X$ and $0 \leqslant \gamma \leqslant q^{\epsilon}/\log X$ and let ϱ be a positive integer. Then*

$$\sum_{\substack{n \leqslant X/d \\ (n,\varrho)=1}} \frac{\alpha_n(|\chi|)}{n}\left(\log\frac{X}{dn}\right)\frac{\sin(\gamma \log nd)}{\gamma} \ll \frac{2^{\omega(q)}q^{1/2}}{\varphi(q)^{1/2}}(\log X)^{3/2}\prod_{p\,|\,\varrho}(1+|\chi(p)|p^{-3/4}),$$

where, for $\gamma = 0$, the left-hand side should be interpreted as its limit as $\gamma \to 0^+$.

Proof. For γ in this range, the O-term in Lemma 5 dominates the main term in absolute value, so we get, for $r = 1, 2$,

$$\sum_{\substack{n \leqslant X/d \\ (n,\varrho)=1}} \frac{\alpha_n(|\chi|)}{n^{1+i\gamma t}}\left(\log\frac{X}{dn}\right)^r = O\left(\frac{2^{\omega(q)}q^{1/2}}{\varphi(q)^{1/2}}(\log X)^{r-1/2}\prod_{p\,|\,\varrho}(1+|\chi(p)|p^{-3/4})\right)$$

for $0 \leqslant t \leqslant 1$. Multiplying the formula for $r = 1$ by $\log X$ and subtracting the equation for $r = 2$, we obtain

$$\sum_{\substack{n \leqslant X/d \\ (n,\varrho)=1}} \frac{\alpha_n(|\chi|)}{n^{1+i\gamma t}}\left(\log\frac{X}{dn}\right)\log dn = O\left(\frac{2^{\omega(q)}q^{1/2}}{\varphi(q)^{1/2}}(\log X)^{3/2}\prod_{p\,|\,\varrho}(1+|\chi(p)|p^{-3/4})\right).$$

Multiplying the formula by $d^{-i\gamma t}$ and integrating with respect to t between 0 and 1 and taking the real part, we get the lemma. □

3. Lemmas on the Hardy–Selberg function

In this section, we study the integrals of $Z(t, \chi)$ mollified using $B_X\left(\frac{1}{2}+it, \chi\right)$, namely the integrals of the function $Z(t, \chi)\left|B_X\left(\frac{1}{2}+it, \chi\right)\right|^2$, which, as we remarked earlier, is also called the Hardy–Selberg function, and was introduced and made use of in [Karatsuba 1984] to study the distribution of zeros of $\zeta(s)$ in short intervals of the critical line. Note that it depends on the parameter X as well. Our first lemma concerns the integral $\int_T^{T+H} Z(t, \chi)\left|B_X\left(\frac{1}{2}+it, \chi\right)\right|^2 \mathrm{d}t$, which should be $o(H)$, expecting some cancellation due to the sign changes of $Z(t, \chi)$, and in view of the fact that $B_X\left(\frac{1}{2}+it, \chi\right)$ looks like the partial sum of $L\left(\frac{1}{2}+it, \chi\right)^{-1/2}$ up to length X. The following lemma confirms this for a suitable range of H, and it supersedes Lemma 6 in [Mawia 2017].

Lemma 7. *Let $q \leqslant T^{\eta}$, $X = T^{\theta}$, $H = T^{\vartheta}$, with $0 < \eta, \theta, \vartheta \leqslant 1$. Then, as $T \to \infty$,*

$$\int_T^{T+H} Z(t, \chi)\left|B_X\left(\tfrac{1}{2}+it, \chi\right)\right|^2 dt = O(q^{1/4}T^{1/2}X\log^2 T + q^{1/4}T^{-1/4}HX\log X)$$

and hence the integral is $o(H)$ for $\frac{1}{2}+\frac{1}{4}\eta+\theta < \vartheta \leqslant 1$, $\eta+4\theta < 1$.

Proof. Note first that this integral is $\ll d(q)q^{3/16}(\log q)^2 H \max(H^{1/4}, T^{1/4})X$ for all values of q and H, using Heath-Brown's hybrid bound [1978, Theorem 1]. This is certainly interesting for small values of H, but is bigger than our bound for H in the above range. This is only natural, as any application of known bounds on $L\left(\frac{1}{2}+it, \chi\right)$ does not take into account the variations of $Z(t, \chi)$.

The proof of the lemma follows the same lines as in [Mawia 2017], where we use the definition of $Z(t, \chi)$ and move the contour to the segment going from $\frac{1}{2}+iT$ to $c+iT$ to $c+i(T+H)$ to $\frac{1}{2}+i(T+H)$,

with $c = 1 + 1/\log T$. For the integral on the horizontal segments, we use the trivial convexity bound [Iwaniec and Kowalski 2004, (5.20)] for $L(s, \chi)$, which is

$$L(\sigma + it, \chi) \ll (qt)^{(1-\sigma)/2+\epsilon} \quad (0 \leqslant \sigma \leqslant 1,\ t \geqslant 1),$$

and the inequalities (easily seen from equations (8) and (9) in [Mawia 2017])

$$B_X(s, \chi)B_X(1-s, \bar{\chi}) \ll X \log X, \quad \Psi(s, \chi)^{-1/2} \ll (qT)^{(\sigma-1/2)/2}$$

for s in the horizontal segments, so that the overall contribution of the integral along the horizontal segments is

$$\ll \int_{1/2}^{c} (qT)^{(1-\sigma)/2+\epsilon}(X \log X)(qT)^{(\sigma-1/2)/2}\, d\sigma \ll X(qT)^{1/4+\epsilon}.$$

Note that a bound sharper than the convexity bound (such as the Burgess bound or the hybrid bounds of Heath-Brown [1978; 1980]) will improve this lemma, but not the main theorem, as the main barrier comes from Lemma 9. The integral along the vertical segment is

$$\mathfrak{w}(\chi)^{-1/2} \int_T^{T+H} \Bigg(\sum_{n\geqslant 1} \chi(n) n^{-c-it}\Bigg)\Bigg(\sum_{k\leqslant X} \beta_k(\chi) k^{-c-it}\Bigg)\Bigg(\sum_{\ell\leqslant X} \beta_\ell(\bar{\chi})\ell^{c+it-1}\Bigg) \times \left(\frac{qt}{2\pi}\right)^{(c+it-1/2)/2} e^{-i(t+\pi(1-2\mathfrak{a})/4)/2}(1 + O(1/t))\, dt,$$

of which the O-term has a contribution

$$\ll \int_T^{T+H} (\log^2 T) X^{1-(1/\log T)} (qt)^{(c-1/2)/2} t^{-1}\, dt \ll q^{1/4} X H T^{-3/4} \log^2 T.$$

Here, we have used the bounds

$$\left|\sum_{n\geqslant 1} \chi(n) n^{-c-it}\right| \leqslant L(c, |\chi|) = \zeta(c) \prod_{p\,|\,q} (1 - p^{-c}) \ll \log T,$$
$$B_X(c + it, \chi) \ll \log T, \quad B_X(1 - c - it, \bar{\chi}) \ll X.$$

Omitting the constant coefficients, the remaining expression can be rewritten as

$$\Sigma := \sum_{\substack{n\geqslant 1\\ k,\ell\leqslant X}} \frac{\chi(n)\beta_k(\chi)\beta_\ell(\bar{\chi})\ell^{c-1}}{(nk)^c} \int_T^{T+H} \left(\frac{qt}{2\pi}\right)^{(c-1/2)/2} \exp\left(\frac{it}{2} \log\left(\frac{qt\ell^2}{2\pi e n^2 k^2}\right)\right) dt.$$

To simplify matters, we will henceforward write $1/4$ in place of $(c - 1/2)/2$ and 1 in place of c, since in any case $(c - 1/2)/2 = 1/4 + 1/(2 \log T)$ and the error incurred in our sums due to this simplification is clearly negligible. Let us further use the following notation:

$$F(t) = \left(\frac{qt}{2\pi}\right)^{1/4}, \quad f(t) = \frac{t}{2} \log\left(\frac{qt\ell^2}{2\pi e n^2 k^2}\right), \quad \tau = \sqrt{\frac{qT}{2\pi}},$$
$$\tau_0 = \sqrt{\frac{q(T - T^{3/4})}{2\pi}}, \quad \tau_1 = \sqrt{\frac{q(T + H)}{2\pi}}, \quad \tau_2 = \sqrt{\frac{q(T + H + T^{3/4})}{2\pi}}.$$

For each $k, \ell \leqslant X$ and $n < \tau_0 \ell/k$ (resp. $n > \tau_2 \ell/k$), the derivative $f'(t)$ of $f(t)$ has no zeros, is monotonic, and satisfies $f'(t) \geqslant \alpha T^{-1/4}$ (resp. $f'(t) \leqslant -\beta T^{-1/4}$) for some absolute constants $\alpha, \beta > 0$, so using the first derivative test for exponential integrals [Ivić 1985, Lemma 2.1], we immediately see that the above sum for n in either one of these ranges is $\ll q^{1/4} T^{1/2} X \log^2 T$. For $\tau_0 \ell/k \leqslant n \leqslant \tau_2 \ell/k$, the derivative $f'(t)$ vanishes at $t_0 = 2\pi n^2 k^2/q\ell^2$, and t_0 may or may not lie in $[T, T+H]$. Let a and b be real numbers with $T - T^{3/4} \leqslant a \leqslant T \leqslant b \leqslant T + H + T^{3/4}$. Assume $a \leqslant t_0 \leqslant b$ (without loss). Following the proof of Lemma 2 from Chapter III, Section 1.3 of [Karatsuba and Voronin 1992] we get

$$\begin{aligned}\int_a^b F(t)\exp(if(t))\,\mathrm{d}t = \frac{2\sqrt{2}\pi}{q^{1/2}}\left(\frac{nk}{\ell}\right)^{3/2}\exp\left(-\pi i\frac{n^2k^2}{q\ell^2}\right) + O(qT)^{1/4}\\ + O\left(q^{-1/2}\left(\frac{nk}{\ell}\right)^{3/2}\min\left\{1, \frac{1}{\sqrt{f(t_0)-f(a)}}\right\}\right)\\ + O\left(q^{-1/2}\left(\frac{nk}{\ell}\right)^{3/2}\min\left\{1, \frac{1}{\sqrt{f(b)-f(t_0)}}\right\}\right).\end{aligned}$$

It follows that the integral is always $\ll q^{-1/2}(nk/\ell)^{3/2} + (qT)^{1/4}$. Therefore, the part of Σ contributed by n in the range $\tau\ell/k \leqslant n \leqslant \tau_1\ell/k$ (in which case $T \leqslant t_0 \leqslant T+H$) is

$$\begin{aligned}&\ll \sum_{k,\ell\leqslant X}\sum_{\tau\ell/k\leqslant n\leqslant\tau_1\ell/k}\left\{\frac{(nk)^{1/2}}{\sqrt{q}\ell^{3/2}} + \frac{q^{1/4}T^{1/4}}{nk}\right\}\\ &\ll q^{1/4}T^{-1/4}HX\log X + q^{1/4}T^{-3/4}HX^2 \ll q^{1/4}T^{-1/4}HX\log X.\end{aligned}$$

Finally, we have to consider the ranges $\tau_0\ell/k \leqslant n < \tau\ell/k$ and $\tau_1\ell/k < n \leqslant \tau_2\ell/k$ (in which two cases t_0 does not lie in $[T, T+H]$). Since the two cases are analogous, we only treat the first. We note that

$$\int_T^{T+H} F(t)\exp(if(t))\,\mathrm{d}t = \left(\int_{t_0}^{T+H} - \int_{t_0}^{T}\right)F(t)\exp(if(t))\,\mathrm{d}t \ll q^{-1/2}\left(\frac{nk}{\ell}\right)^{3/2} + (qT)^{1/4}.$$

Hence the contribution of this range to Σ is

$$\begin{aligned}&\ll \sum_{k,\ell\leqslant X}\sum_{\tau_0\ell/k\leqslant n<\tau\ell/k}\left\{\frac{(nk)^{1/2}}{\sqrt{q}\ell^{3/2}} + \frac{(qT)^{1/4}}{nk}\right\}\\ &\ll q^{1/4}T^{1/2}X\log X + q^{1/4}X\log X \ll q^{1/4}T^{1/2}X\log X.\end{aligned}$$

□

Remark. The referee has pointed out that one may directly apply Lemma 2 from Chapter III, Section 1.3 of [Karatsuba and Voronin 1992] in the treatment of the integral $\int_a^b F(t)\exp(if(t))$ in the last part of the above proof, since for n in the given range, we have $t_0 = 2\pi n^2k^2/q\ell^2 \asymp T$.

Next, the following lemma replaces Lemma 7 in [Mawia 2017] when $q \to \infty$ with T.

Lemma 8. *Let $q \leqslant T^\eta$, $X = T^\theta$, $H = T^\vartheta$, with $0 < \eta, \theta, \vartheta \leqslant 1$, such that $\frac{1}{4} + \frac{1}{4}\eta + \theta < \vartheta \leqslant 1$. Then we have*

$$\int_T^{T+H} |Z(t,\chi)|\left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^2 dt \geqslant H + O(q^{1/4}T^{1/4}X) \quad (T\to\infty).$$

Proof. The proof is as in [Mawia 2017], but to take q into account, we use a different approximation for $L\left(\frac{1}{2}+it,\chi\right)$. Using Corollary 1 to Theorem 1 in [Lavrik 1968], we have the approximation

$$L\left(\tfrac{1}{2}+it,\chi\right)=\xi(t,\chi)+\Psi\left(\tfrac{1}{2}+it,\chi\right)\bar{\xi}(t,\chi)+O\left(\left(\frac{q}{T}\right)^{1/4}\log(2T)\right),$$

valid for $T\leqslant t\leqslant 2T$, where $\Psi\left(\frac{1}{2}+it,\chi\right)$ is the usual factor from the functional equation (mentioned in the first paragraph of the Introduction) and

$$\xi(t,\chi)=\sum_{n\leqslant\sqrt{qt/(2\pi)}}\frac{\chi(n)}{n^{1/2+it}}.$$

Using this, we get

$$\begin{aligned}\int_T^{T+H}|Z(t,\chi)|\left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^2\,\mathrm{d}t&=\int_T^{T+H}\left|L\left(\tfrac{1}{2}+it,\chi\right)\right|\left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^2\,\mathrm{d}t\\&\geqslant\left|\int_T^{T+H}L\left(\tfrac{1}{2}+it,\chi\right)B_X\left(\tfrac{1}{2}+it,\chi\right)^2\,\mathrm{d}t\right|\\&=\left|\int_T^{T+H}\xi(t,\chi)B_X\left(\tfrac{1}{2}+it,\chi\right)^2\,\mathrm{d}t\right.\\&\qquad\left.+\int_T^{T+H}\Psi\left(\tfrac{1}{2}+it,\chi\right)\bar{\xi}(t,\chi)B_X\left(\tfrac{1}{2}+it,\chi\right)^2\,\mathrm{d}t\right|\\&\quad+O\left(\left(\frac{q}{T}\right)^{1/4}\log(2T)\int_T^{T+H}\left|\sum_{n\leqslant X}\beta_n(\chi)n^{-1/2-it}\right|^2\,\mathrm{d}t\right).\end{aligned}$$

Let us first treat the O-term. Using the mean value theorem for Dirichlet polynomials (see Theorem 5.2 of [Ivić 1985]), the integral inside the O-term is easily seen to be

$$H\sum_{n\leqslant X}\frac{|\beta_n(\chi)|^2}{n}+O\left(\sum_{n\leqslant X}|\beta_n(\chi)|^2\right)=O(H\log X)+O(X)=O(H\log X).$$

Hence the O-term is $((q/T)^{1/4}\log(2T)H\log X)$. We now estimate the integral

$$\int_T^{T+H}\xi(t,\chi)B_X\left(\tfrac{1}{2}+it,\chi\right)^2\,\mathrm{d}t.$$

Using the expansion (2-4) for $B_X\left(\frac{1}{2}+it,\chi\right)^2$, we see that this integral is (we write $\tau_1=\sqrt{q(T+H)/2\pi}$ and $T(m)=\max(T,2\pi m^2/q)$)

$$\begin{aligned}&=\int_T^{T+H}\sum_{m\leqslant\sqrt{qt/(2\pi)}}\chi(m)m^{-1/2-it}\sum_{n\leqslant X^2}b_n(\chi)n^{-1/2-it}\,\mathrm{d}t\\&=H+\sum_{\substack{m\leqslant\tau_1;n\leqslant X^2\\mn>1}}\frac{\chi(m)b_n(\chi)}{\sqrt{mn}}\int_{T(m)}^{T+H}(mn)^{-it}\,\mathrm{d}t\\&=H+O\left(\sum_{\substack{m\leqslant\tau_1;n\leqslant X^2\\mn>1}}\frac{d(n)}{\sqrt{mn}\log(mn)}\right)=H+O(\sqrt{\tau_1}X).\end{aligned}$$

Next we look at the integral $\int_T^{T+H} \Psi\big(\frac{1}{2}+it,\chi\big)\bar{\xi}(t,\chi)B_X\big(\frac{1}{2}+it,\chi\big)^2\,\mathrm{d}t$. First, applying Stirling's approximation, we have

$$\frac{\Gamma\big(\frac{1}{4}-\frac{1}{2}it+\frac{1}{2}\mathfrak{a}\big)}{\Gamma\big(\frac{1}{4}+\frac{1}{2}it+\frac{1}{2}\mathfrak{a}\big)}=\left(\frac{2e}{t}\right)^{it}e^{\pi i(1/4-\mathfrak{a}/2)}\left(1+O\left(\frac{1}{T}\right)\right)$$

for $T\leqslant t\leqslant 2T$. Using this, we see that

$$\int_T^{T+H}\Psi\big(\tfrac{1}{2}+it,\chi\big)\bar{\xi}(t,\chi)B_X\big(\tfrac{1}{2}+it,\chi\big)^2\,\mathrm{d}t$$
$$=\mathfrak{w}(\chi)\sum_{m\leqslant\tau_1;n\leqslant X^2}\frac{\bar{\chi}(m)b_n(\chi)}{\sqrt{mn}}\int_{T(m)}^{T+H}\left(\frac{2\pi em}{qnt}\right)^{it}\mathrm{d}t+O\left(\frac{H}{T}\sqrt{\tau_1}X\log X\right).$$

We will apply a first derivative bound for exponential integrals [Ivić 1985, Lemma 2.1]. Write $f(t)=t\log(qnt/(2\pi em))$. Then $f'(t)=\log(qnt/(2\pi m))$, which is monotonic. Since

$$\frac{qnt}{2\pi m}\geqslant\frac{qT}{2\pi\tau_1}\gg(qT)^{1/2},$$

we see that $f'(t)\gg_\eta\log T$, as $q\leqslant T^\eta$. Therefore, the first derivative test gives $\int_{T(m)}^{T+H}\exp(if(t))\,\mathrm{d}t\ll 1/\log T$. Consequently,

$$\mathfrak{w}(\chi)\sum_{m\leqslant\tau_1;n\leqslant X^2}\frac{\bar{\chi}(m)b_n(\chi)}{\sqrt{mn}}\int_{T(m)}^{T+H}\left(\frac{2\pi em}{qnt}\right)^{it}\mathrm{d}t\ll\frac{1}{\log T}\sum_{m\leqslant\tau_1;n\leqslant X^2}\frac{d(n)}{\sqrt{mn}}$$
$$\ll\frac{\sqrt{\tau_1}}{\log T}X\log X\ll q^{1/4}T^{1/4}X.$$

The result follows. □

The following lemma is the most difficult part of the proof, and, as remarked earlier, the proof we gave in [Mawia 2017] had a minor error as the approximate functional equation we used was erroneous. But as will become clear, the proof remains substantially the same. To further reduce the length of the interval in the main theorem, it will be necessary to improve this lemma.

Lemma 9. *Let $q\leqslant T^\eta$, $X=T^\theta$, with $0<\theta<\frac{1}{8}$ and $H=T^\vartheta$ with $\frac{1}{2}+\frac{1}{2}\eta+2\theta<\vartheta\leqslant 1$. Then*

$$\int_T^{T+H}Z(t,\chi)^2\big|B_X\big(\tfrac{1}{2}+it,\chi\big)\big|^4\,dt\ll\frac{4^{\omega(q)}q^2}{\varphi(q)^2}H\quad(T\to\infty).$$

Proof. We would like to point out first that this bound is much sharper in all aspects than we would get from a direct application of the Burgess bound or the hybrid bounds of [Heath-Brown 1978]. For, using Theorem 1 of [Heath-Brown 1978], we have

$$\int_T^{T+H}Z(t,\chi)^2\big|B_X\big(\tfrac{1}{2}+it,\chi\big)\big|^4\,\mathrm{d}t\ll d(q)^2q^{3/8}(\log q)^4H\max(H^{1/2},T^{1/2})X^2,$$

which is always bigger than our bound, for H in the stated range. As remarked earlier, this is due to the fact that the variations of $Z(t,\chi)$ are not taken into account when one applies bounds on $\big|L\big(\frac{1}{2}+it,\chi\big)\big|$. However, the merit of this bound is that it holds for *any* values of q and H without restriction.

We will now start the proof of the lemma. We will write, as before, $\tau = \sqrt{qT/(2\pi)}$. Using the approximate equation (2-1) of Lemma 2, we have

$$\begin{aligned}\int_T^{T+H} & Z(t,\chi)^2 \left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^4 \mathrm{d}t \\ &= \int_T^{T+H} \Theta(t,\chi)^2 \left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^4 \mathrm{d}t \\ &\quad + \int_T^{T+H} \overline{\Theta}(t,\chi)^2 \left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^4 \mathrm{d}t + 2\int_T^{T+H} |\Theta(t,\chi)|^2 \left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^4 \mathrm{d}t \\ &\quad + O\left(\left(\frac{q}{T}\right)^{1/4} \log(2T) \int_T^{T+H} |\Theta(t,\chi)| \left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^4 \mathrm{d}t\right) \\ &\quad + O\left(\left(\frac{q}{T}\right)^{1/2} \log^2(2T) \int_T^{T+H} \left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^4 \mathrm{d}t\right). \end{aligned} \tag{3-1}$$

Using the trivial bound $\left|B_X\left(\frac{1}{2}+it,\chi\right)\right| \leqslant \sum_{n\leqslant X} 1/\sqrt{n} \ll \sqrt{X}$, we see that the two O-terms are respectively

$$O\left(\left(\frac{q}{T}\right)^{1/4} \log(2T) X^2 \int_T^{T+H} |\Theta(t,\chi)|\,\mathrm{d}t\right) \quad \text{and} \quad O\left(\left(\frac{q}{T}\right)^{1/2} H X^2 \log^2(2T)\right).$$

At this point, we apply the Cauchy–Schwarz inequality and get

$$\int_T^{T+H} |\Theta(t,\chi)|\,\mathrm{d}t \leqslant \sqrt{H}\left(\int_T^{T+H} \left|\sum_{n\leqslant\sqrt{qt/(2\pi)}} \frac{\chi(n)}{n^{1/2+it}}\right|^2 \mathrm{d}t\right)^{1/2} \tag{3-2}$$

using the expression (2-2). The error in [Mawia 2017] lies in the subsequent application of the mean value theorem for Dirichlet polynomials at this point, which is not applicable in this form, since the summation is up to $n \leqslant \sqrt{qt/(2\pi)}$, and not up to $n \leqslant \sqrt{qT/(2\pi)}$ (t being the variable of integration, whereas T is the lower limit of the integral). However, this is easy to rectify. One method to correct this error is to use an analogue of Lemma 6 in [Selberg 1942], where the sum runs up to $n \leqslant \sqrt{T/(2\pi)}$ but the error is of size $T^{-3/20}$ instead of $T^{-1/4}$. We however follow another route. Let us write for brevity $T(m,n) = \max(T, 2\pi m^2/q, 2\pi n^2/q)$ and $\tau_1 = \sqrt{q(T+H)/(2\pi)}$. We then have

$$\begin{aligned}\int_T^{T+H} \left|\sum_{n\leqslant\sqrt{qt/(2\pi)}} \frac{\chi(n)}{n^{1/2+it}}\right|^2 \mathrm{d}t &= \sum_{m,n\leqslant\tau_1} \frac{\chi(m)\bar{\chi}(n)}{\sqrt{mn}} \int_{T(m,n)}^{T+H} \left(\frac{n}{m}\right)^{it} \mathrm{d}t \\ &= H\sum_{n\leqslant\tau} \frac{|\chi(n)|}{n} + \sum_{\substack{m,n\leqslant\tau \\ m\neq n}} \frac{\chi(m)\bar{\chi}(n)}{\sqrt{mn}\log(n/m)} \left(\left(\frac{n}{m}\right)^{i(T+H)} - \left(\frac{n}{m}\right)^{iT}\right) \\ &\quad + \sum_{\tau<n\leqslant\tau_1} \frac{|\chi(n)|}{n}\left(T+H-\frac{2\pi n^2}{q}\right) \\ &\quad + \sum_{\substack{\tau<m,n\leqslant\tau_1 \\ m\neq n}} \frac{\chi(m)\bar{\chi}(n)}{\sqrt{mn}\log(n/m)} \left(\left(\frac{n}{m}\right)^{i(T+H)} - \left(\frac{n}{m}\right)^{iT(m,n)}\right). \end{aligned}$$

Obviously, the first and third terms are $\ll H\log(qT)$. It is also easy to see that the second and last terms in the above expression are $\ll (qT)^{1/2}\log(qT)$ (see Lemma 1 in [Selberg 1942]). It follows from (3-2) that $\int_T^{T+H}|\Theta(t,\chi)|\,\mathrm{d}t \ll H\sqrt{\log(qT)}$ and the first O-term in (3-1) is $O((q/T)^{1/4}HX^2\sqrt{\log(qT)}\log(2T))$.

The main difficulty is in the treatment of the first three terms on the right side of (3-1). Since the first two integrals are bounded in absolute value by the third integral, it is enough to look at the third integral, namely

$$J=\int_T^{T+H}|\Theta(t,\chi)|^2\left|B_X\left(\tfrac{1}{2}+it,\chi\right)\right|^4\mathrm{d}t.$$

Using the expressions (2-4) and (2-1), we have

$$\begin{aligned}J&=\sum_{\substack{k,\ell\leqslant\tau_1\\ m,n\leqslant X^2}}\frac{\chi(k)\bar\chi(\ell)b_m(\chi)b_n(\bar\chi)}{\sqrt{k\ell mn}}\int_{T(k,\ell)}^{T+H}\left(\frac{\ell n}{km}\right)^{it}\mathrm{d}t\\&=H\sum_{\substack{k,\ell\leqslant\tau;m,n\leqslant X^2\\ km=\ell n}}\frac{\chi(k)\bar\chi(\ell)b_m(\chi)b_n(\bar\chi)}{\sqrt{k\ell mn}}\qquad(3\text{-}3)\\&\quad+\left(\sum\nolimits_1+\sum\nolimits_2+\sum\nolimits_3\right)\frac{\chi(k)\bar\chi(\ell)b_m(\chi)b_n(\bar\chi)}{\sqrt{k\ell mn}}(T+H-T(k,\ell))\qquad(3\text{-}4)\\&\quad+\sum_{\substack{k,\ell\leqslant\tau_1;m,n\leqslant X^2\\ km\neq\ell n}}\frac{\chi(k)\bar\chi(\ell)b_m(\chi)b_n(\bar\chi)}{i\sqrt{k\ell mn}\log(\ell n/km)}\left(\left(\frac{\ell n}{km}\right)^{i(T+H)}-\left(\frac{\ell n}{km}\right)^{iT(k,\ell)}\right),\qquad(3\text{-}5)\end{aligned}$$

where

$$\sum\nolimits_1=\sum_{\substack{\tau<k,\ell\leqslant\tau_1;m,n\leqslant X^2\\ km=\ell n}},\quad\sum\nolimits_2=\sum_{\substack{k\leqslant\tau<\ell\leqslant\tau_1;m,n\leqslant X^2\\ km=\ell n}},\quad\sum\nolimits_3=\sum_{\substack{\ell\leqslant\tau<k\leqslant\tau_1;m,n\leqslant X^2\\ km=\ell n}}.$$

We shall treat the above sums one by one, starting from the easiest, namely (3-5), to the hardest, that is, (3-3). Using the fact that $|b_n(\chi)|\leqslant d(n)\ll T^\epsilon$ for $n\leqslant\tau_1X^2$, and applying [Selberg 1942, Lemma 1], we see that the sum in (3-5) above is $O(q^{1/2}T^{1/2+\epsilon}X^2)$ for any $\epsilon>0$.

Let us now look at the first sum in (3-4), namely the sum enclosed by $\sum_1$. We want to show that it is $\ll H$. This sum is essentially the same as the sum in (3-3), except that the range of k, ℓ is different. Although we can adapt the proof given in [Mawia 2017] that the sum in (3-3) is $\ll 1$, we give a simpler proof for this sum when $\vartheta<1$. Since $T+H-T(k,\ell)\leqslant H$ for k,ℓ in the given range, it is enough to show that

$$\sum_{\substack{\tau<k,\ell\leqslant\tau_1;m,n\leqslant X^2\\ km=\ell n}}\frac{d(m)d(n)}{\sqrt{k\ell mn}}\ll 1.$$

Note first that

$$\sum_{\substack{\tau<k,\ell\leqslant\tau_1;m,n\leqslant X^2\\ km=\ell n}}\frac{d(m)d(n)}{\sqrt{k\ell mn}}=\sum_{m,n\leqslant X^2}\frac{d(m)d(n)}{mn}g\sum_{\tau(m,n)<\nu\leqslant\tau_1(m,n)}\frac{1}{\nu},\qquad(3\text{-}6)$$

where $\tau(m,n)=g\tau/\min(m,n)$, $\tau_1(m,n)=g\tau_1/\max(m,n)$ and $g=(m,n)$ in each term. Since the sum on the right is symmetric in m and n, it is enough to look at the sum for $m\leqslant n\leqslant X^2$, which we do

henceforward. Observe that the inner sum is zero unless $m/n > \tau/\tau_1$, in which case, the inner sum is of size $\ll \log(m\tau_1/(n\tau)) + (m+n)/(g\tau)$ (the term $\log(m\tau_1/(n\tau))$ comes from the main term of the harmonic sum, whereas the term $(m+n)/(g\tau)$ comes from the error term) and in the outer sum, for each $n \leqslant X^2$, the variable m ranges over $n\tau/\tau_1 < m \leqslant n$. Using the fact that $d(m)d(n) \ll T^{\epsilon}$, we see that the sum (3-6) is

$$\begin{aligned}
&\ll T^{\epsilon} \sum_{n\leqslant X^2} \frac{1}{n} \sum_{n\tau/\tau_1<m\leqslant n} \frac{g}{m} \log \frac{m\tau_1}{n\tau} + \sum_{m,n\leqslant X^2} \frac{d(m)d(n)}{\tau}\left(\frac{1}{m}+\frac{1}{n}\right)\\
&\ll HT^{-1+\epsilon} \sum_{n\leqslant X^2} \sum_{n\tau/\tau_1<m\leqslant n} \frac{g}{mn} + \frac{X^2\log^3 X}{\tau}\\
&\ll H^2T^{-2+\epsilon}(\log X)^3 + \frac{X^2\log^3 X}{\tau} \ll 1
\end{aligned}$$

for small enough ϵ, as long as $\vartheta < 1$ and $\theta < \frac{1}{4} + \frac{1}{4}\eta$ (this latter condition is satisfied since $\theta < \frac{1}{8}$ by assumption). This completes the proof when $\vartheta < 1$. When $\vartheta = 1$, we can adapt the treatment of the sum (3-3), given in the ensuing paragraphs. It is to be noted that, although similar in appearance, the sum (3-4) needs a much less delicate treatment than the sum in (3-3) due to the restriction $\tau < k$, $\ell \leqslant \tau_1$.

Next we look at the sums enclosed by $\sum_2$ and $\sum_3$ in (3-4). Since the two sums are similar, it suffices to look at one of them, say $\sum_2$. Comparing it with $\sum_1$, we expect it to be roughly of the same size, as the range of ℓ is the same although k runs through a set of smaller integers, and we have exactly the same terms in the two sums. We observe that the sum $\sum_2$ is smaller in absolute value than

$$H \sum_{m,n\leqslant X^2} \frac{d(m)d(n)}{mn} g \sum_{g\tau/m<\nu\leqslant \min\{g\tau/n, g\tau_1/m\}} \frac{1}{\nu},$$

by following the same steps as in the treatment of $\sum_1$. This sum can now be estimated as (3-6).

It remains to treat the sum in (3-3). This sum was shown in [Mawia 2017] to be $\ll 1$ when q is fixed; the case when q is not fixed will be treated shortly. We first remark that in the treatment given of this sum in [Mawia 2017], the positions of k and l have to be interchanged throughout, so that the integral in the first line of equation (12) on p. 45 should be $\int_T^{T+H} (ln/km)^{it}\,\mathrm{d}t$, and the conditions $kn = lm$ and $kn \neq lm$ should be replaced throughout by the conditions $km = ln$ and $km \neq ln$ respectively.

We now consider the same sum when q is not fixed. Note first that

$$\sum_{\substack{k,\ell\leqslant\tau; m,n\leqslant X^2\\ km=\ell n}} \frac{\chi(k)\bar{\chi}(\ell)b_m(\chi)b_n(\bar{\chi})}{\sqrt{k\ell mn}} = \sum_{m,n\leqslant X^2} \frac{b_m(|\chi|)b_n(|\chi|)}{mn} g \sum_{\nu\leqslant\tau_0} \frac{|\chi(\nu)|}{\nu}, \tag{3-7}$$

where $\tau_0 = g\tau/\max(m,n)$ and $g = (m,n)$ in each term. As in [Mawia 2017], let us first define, for $\gamma \geqslant 0$,

$$S(\gamma) = \Re\left\{\tau^{i\gamma} \sum_{n_j\leqslant X} \frac{\beta_{n_1}(|\chi|)\beta_{n_2}(|\chi|)\beta_{n_3}(|\chi|)\beta_{n_4}(|\chi|)}{n_1n_2n_3n_4} \frac{g^{1+i\gamma}}{(n_1n_3)^{i\gamma}} \sum_{\nu\leqslant\tau_1} \frac{|\chi(\nu)|}{\nu^{1+i\gamma}}\right\},$$

where we have written $\tau_1 = g\tau/\max(n_1n_2, n_3n_4)$ with g now standing for (n_1n_2, n_3n_4). Note that the right side of (3-7) is equal to $S(0)$. We will show that $S(\gamma) = O(4^{\omega(q)}q^2/\varphi(q)^2)$ for $0 < \gamma \leqslant 1/\log T$; it will follow by continuity that $S(0) = O(4^{\omega(q)}q^2/\varphi(q)^2)$ as well. As remarked in [Mawia 2017], the

proof is essentially contained in the proof that $K(\gamma) = O(1)$ in [Selberg 1942, pp. 27–31]. Following the same reduction steps as in [Mawia 2017], we arrive at the expression

$$S(\gamma) = \Re\{\tau^{i\gamma} L(1+i\gamma, |\chi|) S_2(\gamma)\} + O(|S_2(\gamma)| \log T) + O(2^{\omega(q)} (qT)^{-1/2} X^2 \log^4 T), \tag{3-8}$$

with

$$S_2(\gamma) = \sum_{n_j \leqslant X} \frac{\beta_{n_1}(|\chi|)\beta_{n_2}(|\chi|)\beta_{n_3}(|\chi|)\beta_{n_4}(|\chi|)}{n_1 n_2 n_3 n_4} \frac{g^{1+i\gamma}}{(n_1 n_3)^{i\gamma}}. \tag{3-9}$$

Note that we have used Lemma 3 in the course of the reduction. Let us look at the first O-term in (3-8). We showed in [Mawia 2017] that $S_2(\gamma)$ is equal to

$$S_2(\gamma) = \sum_{d \leqslant X^2} \varphi_{i\gamma}(d) S_3(\gamma; d)^2, \tag{3-10}$$

where

$$S_3(\gamma; d) = \sum_{\substack{m,n \leqslant X \\ d \mid mn}} \frac{\beta_m(|\chi|)\beta_n(|\chi|)}{m n^{1+i\gamma}}$$

(see the second displayed equation on p. 48), and that $S_3(\gamma; d)$ is further equal to

$$S_3(\gamma; d) = \frac{1}{\log^2 X} \sum_{\substack{d_1, d_2 \leqslant X^2 \\ d \mid d_1 d_2 \mid d^\infty}} \frac{\alpha_{d_1}(|\chi|)\alpha_{d_2}(|\chi|)}{d_1 d_2^{1+i\gamma}} \left\{ \sum_{\substack{n \leqslant X/d_1 \\ (n,d)=1}} \frac{\alpha_n(|\chi|)}{n} \log \frac{X}{d_1 n} \right\} \left\{ \sum_{\substack{n \leqslant X/d_2 \\ (n,d)=1}} \frac{\alpha_n(|\chi|)}{n^{1+i\gamma}} \log \frac{X}{d_2 n} \right\}, \tag{3-11}$$

where, in the above sum, the only primes dividing d_1, d_2 are the prime factors of d, which is often symbolically written as $d_1 \mid d^\infty$, $d_2 \mid d^\infty$. We will now apply Lemma 5 with $r = 1$. Note that for $0 \leqslant \gamma \leqslant 1/\log X$, the O-term dominates the "main" term in (2-5). Using this fact in (3-11), we see that

$$\begin{aligned} S_3(\gamma; d) &\ll \frac{2^{\omega(q)} q}{\varphi(q) \log X} \sum_{\substack{d_1, d_2 \leqslant X^2 \\ d \mid d_1 d_2 \mid d^\infty}} \frac{1}{d_1 d_2} \prod_{p \mid d} (1 + |\chi(p)| p^{-3/4})^2 \\ &\ll \frac{2^{\omega(q)} q}{d \varphi(q) \log X} \prod_{p \mid d} (1 + p^{-3/4})^3. \end{aligned} \tag{3-12}$$

Using this in (3-10), we arrive at

$$\begin{aligned} S_2(\gamma) &\ll \frac{4^{\omega(q)} q^2}{\varphi(q)^2 \log^2 X} \sum_{d \leqslant X^2} \frac{|\varphi_{i\gamma}(d)|}{d^2} \prod_{p \mid d} (1 + p^{-3/4})^6 \\ &\ll \frac{4^{\omega(q)} q^2}{\varphi(q)^2 \log X}. \end{aligned}$$

Plugging this into (3-8), we have

$$S(\gamma) = \Re\{\tau^{i\gamma} L(1+i\gamma, |\chi|) S_2(\gamma)\} + O\left(\frac{4^{\omega(q)} q^2}{\varphi(q)^2}\right).$$

The treatment of the first term here is somewhat delicate. Using Lemma 3, we have $L(1+i\gamma, |\chi|) = \varphi(q)/qi\gamma + O(2^{\omega(q)})$, so that

$$S(\gamma) = \Im\left\{\tau^{i\gamma}\frac{\varphi(q)}{q\gamma}S_2(\gamma)\right\} + O\left(\frac{4^{\omega(q)}q^2}{\varphi(q)^2}\right). \tag{3-13}$$

Using the fact that the inequality $|\Im(ab^2)| = |y(u^2-v^2)+2yuv| \leqslant |\Im a||b|^2 + 2|a||b||\Im b|$ holds for $a = x+iy,\ b = u+iv$, we have, in view of (3-10) and (3-13),

$$|S(\gamma)| \leqslant \sum_{d\leqslant X^2}\left|\Im\left(\frac{\tau^{i\gamma}\varphi_{i\gamma}(d)\varphi(q)}{q\gamma}\right)\right||S_3(\gamma;d)|^2 + 2\sum_{d\leqslant X^2}|\varphi_{i\gamma}(d)||S_3(\gamma;d)|\left|\Im\left(\frac{S_3(\gamma;d)}{\gamma}\right)\right| + O\left(\frac{4^{\omega(q)}q^2}{\varphi(q)^2}\right). \tag{3-14}$$

Let us look at the terms one by one. First of all, we observe that

$$\begin{aligned}\left|\Im\left(\frac{\tau^{i\gamma}\varphi_{i\gamma}(d)\varphi(q)}{q\gamma}\right)\right| &= \left|\Im\left(\frac{\tau^{i\gamma}\varphi(q)}{q\gamma}d^{1+i\gamma}\sum_{c\,|\,d}\frac{\mu(c)}{c^{1+i\gamma}}\right)\right| \\ &= d\frac{\varphi(q)}{q}\left|\sum_{c\,|\,d}\frac{\mu(c)}{c}\frac{\sin(\gamma\log(\tau d/c))}{\gamma}\right| \\ &\leqslant d\frac{\varphi(q)}{q}\log(\tau d)\prod_{p\,|\,d}(1+p^{-3/4}).\end{aligned}$$

Using the bound (3-12) for $S_3(\gamma;d)$, we see that the first term in (3-14) is

$$\ll \frac{2^{\omega(q)}q\log(\tau X^2)}{\varphi(q)\log^2 X}\sum_{d\leqslant X^2}\frac{1}{d}\prod_{p\,|\,d}(1+p^{-3/4})^7 \ll \frac{2^{\omega(q)}q}{\varphi(q)}. \tag{3-15}$$

Similarly, using Lemmas 5 and 6 and the expression

$$\Im\left(\frac{S_3(\gamma;d)}{\gamma}\right) = \frac{1}{\log^2 X}\sum_{\substack{d_1,d_2\leqslant X\\ d\,|\,d_1d_2\,|\,d^\infty}}\frac{\alpha_{d_1}(|\chi|)\alpha_{d_2}(|\chi|)}{d_1d_2}\left\{\sum_{\substack{n\leqslant X/d_1\\(n,d)=1}}\frac{\alpha_n(|\chi|)}{n}\log\frac{X}{d_1n}\right\}\left\{\sum_{\substack{n\leqslant X/d_2\\(n,d)=1}}\frac{\alpha_n(|\chi|)}{n}\log\frac{X}{d_2n}\frac{\sin(\gamma\log(d_2n))}{\gamma}\right\},$$

we get that the second term in (3-14) is

$$\ll \frac{2^{\omega(q)}q}{\varphi(q)}. \tag{3-16}$$

It follows from (3-14), (3-15) and (3-16) that $S(\gamma) \ll 4^{\omega(q)}q^2/\varphi(q)^2$. □

Acknowledgement

The author acknowledges support from INSPIRE Faculty Award IFA17-MA115 of the Department of Science and Technology (DST), Government of India, and would like to thank the Theoretical Statistics

and Mathematics Unit, Indian Statistical Institute, Kolkata, for hosting him and for providing excellent working environment. The author also thanks K. Mallesham and O. Ramaré for helpful discussions, S. Ganguly for encouragement, and the anonymous referee for a detailed reading of the manuscript which has led to the correction of some errors in the original manuscript, and an improvement in the overall presentation. Finally, the author thanks the editorial board, especially Maxim Korolev, for the efficient handling of the manuscript.

References

[Borwein et al. 2000] J. M. Borwein, D. M. Bradley, and R. E. Crandall, "Computational strategies for the Riemann zeta function", *J. Comput. Appl. Math.* **121**:1-2 (2000), 247–296. MR Zbl

[Chandrasekharan and Narasimhan 1963] K. Chandrasekharan and R. Narasimhan, "The approximate functional equation for a class of zeta-functions", *Math. Ann.* **152** (1963), 30–64. MR

[Gabcke 1979] W. Gabcke, *Neue Herleitung und explizite Restabschätzung der Riemann-Siegel-Formel*, Dissertation zur Erlangung des Doktorgrades, Georg-August-Universität Göttingen, 1979, http://hdl.handle.net/11858/00-1735-0000-0022-6013-8. Zbl

[Gonek and Ivić 2017] S. M. Gonek and A. Ivić, "On the distribution of positive and negative values of Hardy's Z-function", *J. Number Theory* **174** (2017), 189–201. MR Zbl

[Heath-Brown 1978] D. R. Heath-Brown, "Hybrid bounds for Dirichlet L-functions", *Invent. Math.* **47**:2 (1978), 149–170. MR Zbl

[Heath-Brown 1980] D. R. Heath-Brown, "Hybrid bounds for Dirichlet L-functions, II", *Quart. J. Math. Oxford Ser.* (2) **31**:122 (1980), 157–167. MR Zbl

[Ivić 1985] A. Ivić, *The Riemann zeta-function: the theory of the Riemann zeta-function with applications*, John Wiley & Sons, New York, 1985. MR Zbl

[Ivić 2004] A. Ivić, "On the integral of Hardy's function", *Arch. Math.* (*Basel*) **83**:1 (2004), 41–47. MR Zbl

[Ivić 2010] A. Ivić, "On some problems involving Hardy's function", *Cent. Eur. J. Math.* **8**:6 (2010), 1029–1040. MR Zbl

[Ivić 2013] A. Ivić, *The theory of Hardy's Z-function*, Cambridge Tracts in Mathematics **196**, Cambridge University Press, 2013. MR Zbl

[Ivić 2017a] A. Ivić, "On a cubic moment of Hardy's function with a shift", pp. 99–112 in *Exploring the Riemann zeta function*, edited by H. Montgomery et al., Springer, 2017. MR

[Ivić 2017b] A. Ivić, "On certain moments of Hardy's function $Z(t)$ over short intervals", *Mosc. J. Comb. Number Theory* **7**:2 (2017), 59–73. MR Zbl

[Ivić 2017c] A. Ivich, "Hardy's function $Z(t)$: results and unsolved problems", *Trudy Mat. Inst. Steklova* **296** (2017), 111–122. In Russian; translated in *Proc. Steklov Inst. Math.* **296**:1 (2017), 104–114. MR Zbl

[Iwaniec and Kowalski 2004] H. Iwaniec and E. Kowalski, *Analytic number theory*, American Mathematical Society Colloquium Publications **53**, American Mathematical Society, Providence, RI, 2004. MR Zbl

[Jutila 2009] M. Jutila, "Atkinson's formula for Hardy's function", *J. Number Theory* **129**:11 (2009), 2853–2878. MR Zbl

[Jutila 2011] M. Jutila, "An asymptotic formula for the primitive of Hardy's function", *Ark. Mat.* **49**:1 (2011), 97–107. MR Zbl

[Karatsuba 1981] A. A. Karatsuba, "On the distance between consecutive zeros of the Riemann zeta function that lie on the critical line", *Trudy Mat. Inst. Steklova* 157 (1981), 49–63. In Russian; translated in *Proc. Steklov Inst. Math.* **157** (1983), 51–66. MR Zbl

[Karatsuba 1984] A. A. Karatsuba, "Zeros of the function $\zeta(s)$ on short intervals of the critical line", *Izv. Akad. Nauk SSSR Ser. Mat.* **48**:3 (1984), 569–584. In Russian; translated in *Math. USSR-Izv.* **24**:3 (1984), 523–537. MR Zbl

[Karatsuba and Voronin 1992] A. A. Karatsuba and S. M. Voronin, *The Riemann zeta-function*, De Gruyter Expositions in Mathematics **5**, Walter de Gruyter, Berlin, 1992. MR Zbl

[Korolev 2007] M. A. Korolev, "On the integral of the Hardy function $Z(t)$", *Dokl. Akad. Nauk* **413**:5 (2007), 599–602. In Russian; translated in *Dokl. Math.* **75**:2 (2007), 295–298. MR Zbl

[Korolev 2008] M. A. Korolëv, "On the integral of the Hardy function $Z(t)$", *Izv. Ross. Akad. Nauk Ser. Mat.* **72**:3 (2008), 19–68. In Russian; translated in *Izv. Math.* **72**:3 (2008), 429–478. MR Zbl

[Korolev 2017] M. A. Korolëv, "On the Riemann–Siegel formula for the derivatives of the Hardy function", *Algebra i Analiz* **29**:4 (2017), 53–81. In Russian; translated in *St. Petersburg Math. J.* **29**:4 (2018), 581–601. MR Zbl

[Lavrik 1968] A. F. Lavrik, "The approximate functional equation for Dirichlet L-functions", *Trudy Moskov. Mat. Obšč.* **18** (1968), 91–104. In Russian. MR Zbl

[Mawia 2017] R. Mawia, "On the distribution of values of Hardy's Z-function in short intervals", *Mosc. J. Comb. Number Theory* **7**:3 (2017), 34–50. MR

[Selberg 1942] A. Selberg, "On the zeros of Riemann's zeta-function", *Skr. Norske Vid. Akad. Oslo I.* **1942**:10 (1942), 59 pp. Reprinted as pp. 85–141 of his *Collected papers*, vol. I, Springer, 1989. MR

[Siegel 1932] C. L. Siegel, "Über Riemanns Nachlaß zur analytischen Zahlentheorie", *Quellen Studien B.* **2** (1932), 45–80. Zbl

[Siegel 1943] C. L. Siegel, "Contributions to the theory of the Dirichlet L-series and the Epstein zeta-functions", *Ann. of Math.* (2) **44** (1943), 143–172. MR Zbl

[Siegel 1966] C. L. Siegel, *Gesammelte Abhandlungen*, Band I, Springer, 1966. MR Zbl

[Suetuna 1932] Z. Suetuna, "Über die approximative Funktionalgleichung für Dirichletsche L-Funktionen", *Jpn. J. Math.* **9** (1932), 111–116. Zbl

[Tchudakoff 1947] N. Tchudakoff, "On Goldbach–Vinogradov's theorem", *Ann. of Math.* (2) **48** (1947), 515–545. MR Zbl

Received 10 Nov 2018. Revised 7 May 2019.

RAMDIN MAWIA:

ramdinm71@gmail.com
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 3, 2019
dx.doi.org/10.2140/moscow.2019.8.247

On products of shifts in arbitrary fields

Audie Warren

We adapt the approach of Rudnev, Shakan, and Shkredov (2018) to prove that in an arbitrary field $\mathbb{F}$, for all $A \subset \mathbb{F}$ finite with $|A| < p^{1/4}$ if $p := \mathrm{Char}(\mathbb{F})$ is positive, we have

$$|A(A+1)| \gg \frac{|A|^{11/9}}{(\log|A|)^{7/6}}, \quad |AA| + |(A+1)(A+1)| \gg \frac{|A|^{11/9}}{(\log|A|)^{7/6}}.$$

This improves upon the exponent of $\frac{6}{5}$ given by an incidence theorem of Stevens and de Zeeuw.

1. Introduction and main result

For finite $A \subseteq \mathbb{F}$, we define the *sumset* and *product set* of A as

$$A + A = \{a + b : a, b \in A\}, \quad AA = \{ab : a, b \in A\}.$$

It is an active area of research to show that one of these sets must be large relative to A. The central conjecture in this area is the following.

Conjecture 1 (Erdős–Szemerédi). *For all $\epsilon > 0$, and for all $A \subseteq \mathbb{Z}$ finite, we have*

$$|AA| + |A + A| \gg |A|^{2-\epsilon}.$$

The notation $X \ll Y$ is used to hide absolute constants; i.e., $X \ll Y$ if and only if there exists an absolute constant $c > 0$ such that $X \ll cY$. If $X \ll Y$ and $Y \ll X$ we write $X \asymp Y$. We will let p denote the characteristic of $\mathbb{F}$ throughout (p may be zero). Due to the possible existence of finite subfields in $\mathbb{F}$, extra restrictions on $|A|$ relative to p must be imposed if p is positive; *all such conditions can be ignored if $p = 0$.*

Although Conjecture 1 is stated over the integers, it can be considered over fields, the real numbers being of primary interest. Current progress over $\mathbb{R}$ places us at an exponent of $\frac{4}{3} + c$ for some small c, due to Shakan [2018], building on [Konyagin and Shkredov 2015; Solymosi 2009]. Incidence geometry, and in particular the Szemerédi–Trotter theorem, are tools often used to prove such results in the real numbers.

Conjecture 1 can also be considered over arbitrary fields $\mathbb{F}$. Over arbitrary fields we replace the Szemerédi–Trotter theorem with a point-plane incidence theorem of [Rudnev 2018], which was used by Stevens and de Zeeuw [2017] to derive a point-line incidence theorem. An exponent of $\frac{6}{5}$ was proved in 2014 by Roche-Newton, Rudnev, and Shkredov [Roche-Newton et al. 2016]. An application of the Stevens–de Zeeuw theorem also gives this exponent of $\frac{6}{5}$ for Conjecture 1, so that $\frac{6}{5}$ became a threshold to be broken.

MSC2010: 11B75, 68R05.
Keywords: growth, sum-product estimates, energy.

The $\frac{6}{5}$ threshold has recently been broken; see [Shakan and Shkredov 2018; Rudnev et al. 2018; Chen et al. 2018]. The following theorem was proved by Rudnev, Shakan, and Shkredov and is the current state-of-the-art bound.

Theorem 2 [Rudnev et al. 2018]. *Let $A \subset \mathbb{F}$ be a finite set. If $\mathbb{F}$ has positive characteristic p, assume $|A| < p^{18/35}$. Then we have*

$$|A+A| + |AA| \gg |A|^{11/9-o(1)}.$$

Another way of considering the sum-product phenomenon is to consider the set $A(A+1)$, which we would expect to be quadratic in size. This encapsulates the idea that a translation of a multiplicatively structured set should destroy its structure, which is a main theme in sum-product questions. Study of growth of $|A(A+1)|$ began in [Garaev and Shen 2010]; see also [Jones and Roche-Newton 2013; Zhelezov 2015; Mohammadi 2018]. Current progress for $|A(A+1)|$ comes from an application of the Stevens–de Zeeuw theorem, giving the same exponent of $\frac{6}{5}$. In this paper we use the multiplicative analogue of ideas in [Rudnev et al. 2018] to prove the following theorem.

Theorem 3. *Let $A, B, C, D \subset \mathbb{F}$ be finite with the conditions*

$$|C(A+1)||A| \le |C|^3, \quad |C(A+1)|^2 \le |A||C|^3, \quad |B| \le |D|, \quad |A|, |B|, |C|, |D| < p^{1/4}.$$

Then we have

$$|AB|^8 |C(A+1)|^2 |D(B-1)|^8 \gg \frac{|B|^{13}|A|^5|D|^3|C|}{(\log|A|)^{17}(\log|B|)^4}.$$

In our applications of this theorem we have $|A| = |B| = |C| = |D|$ so that the first three conditions are trivially satisfied. The conditions involving p could likely be improved; however, for sake of exposition we do not attempt to optimise these. The main proof closely follows [Rudnev et al. 2018] (in the multiplicative setting), the central difference being a bound on multiplicative energies in terms of products of shifts. An application of Theorem 3 beats the threshold of $\frac{6}{5}$, matching the $\frac{11}{9}$ appearing in Theorem 2. Specifically, we have:

Corollary 4. *Let $A \subseteq \mathbb{F}$ be finite, with $|A| < p^{1/4}$. Then*

$$|A(A+1)| \gg \frac{|A|^{11/9}}{(\log|A|)^{7/6}}, \quad |AA| + |(A+1)(A+1)| \gg \frac{|A|^{11/9}}{(\log|A|)^{7/6}}.$$

Corollary 4 can be seen by applying Theorem 3 with $B = A+1$, $C = A$ and $D = A+1$ for the first result, and $B = -A$, $D = C = A+1$ for the second result.

2. Preliminary results

We require some preliminary theorems. The first is the point-line incidence theorem of Stevens and de Zeeuw.

Theorem 5 [Stevens and de Zeeuw 2017]. *Let A and B with $|A| \ge |B|$ be finite subsets of a field $\mathbb{F}$, and let L be a set of lines. Assuming $|L||B| \ll p^2$ and $|B||A|^2 \le |L|^3$, we have*

$$I(A \times B, L) \ll |A|^{1/2}|B|^{3/4}|L|^{3/4} + |L|.$$

Note that as $|A| \geq |B|$, we have $|A|^{1/2}|B|^{3/4} \leq |A|^{3/4}|B|^{1/2}$; in particular with the same conditions we have the above result with the exponents of A and B swapped. Because of this, the condition $|A| \geq |B|$ is only needed to specify the second two conditions. We may therefore restate Theorem 5 as:

Theorem 6. *Let A and B be finite subsets of a field $\mathbb{F}$, and let L be a set of lines. Assuming*

$$|L| \min\{|A|, |B|\} \ll p^2 \quad \text{and} \quad |A||B| \max\{|A|, |B|\} \leq |L|^3,$$

we have

$$I(A \times B, L) \ll \min\{|A|^{1/2}|B|^{3/4}, |A|^{3/4}|B|^{1/2}\}|L|^{3/4} + |L|.$$

This second formulation will be how we apply Theorem 5. Before stating the next two theorems we require some definitions. For $x \in \mathbb{F}$ we define the *representation function*

$$r_{A/D}(x) = \left|\left\{(a, d) \in A \times D : \frac{a}{d} = x\right\}\right|.$$

Note that for all x we have $r_{A/D}(x) \leq \min\{|A|, |D|\}$. This is seen as fixing one of a, d in the equation $a/d = x$ necessarily determines the other. The set A/D in this definition can be changed to any other combination of sets, changing the fraction a/d in the definition to match. For $n \in \mathbb{R}^+$, we define the n-th moment *multiplicative energy* of sets $A, D \subseteq \mathbb{F}$ as

$$E_n^*(A, D) = \sum_x r_{A/D}(x)^n.$$

When $n = 2$ we shall simply write $E^*(A, D)$, and when $A = D$ we write $E_n^*(A) := E_n^*(A, A)$. By considering that we have $a/a = 1$ for all $a \in A$, we have the trivial lower bound $E_n^*(A) \geq |A|^n$. When n is in fact a natural number, $E_n^*(A, D)$ can be considered as the number of solutions to

$$\frac{a_1}{d_1} = \frac{a_2}{d_2} = \cdots = \frac{a_n}{d_n}, \quad a_i \in A,\ d_i \in D,$$

giving the trivial upper bound $E_n^*(A, D) \leq |A|^n|D|$ by fixing a_1 to a_n and then choosing a single d_i, which necessarily determines all other d_i.

We use Theorem 6 to prove two further results. The first is a bound on the fourth-order multiplicative energy relative to products of shifts.

Theorem 7. *For all finite nonempty $A, C, D \subset \mathbb{F}$ with*

$$|A|^2|C(A+1)| \leq |D||C|^3, \quad |A||C(A+1)|^2 \leq |D|^2|C|^3, \quad |A||C||D|^2 \ll p^2,$$

we have

$$E_4^*(A, D) \ll \min\left\{\frac{|C(A+1)|^2|D|^3}{|C|}, \frac{|C(A+1)|^3|D|^2}{|C|}\right\} \log|A|.$$

The second result is similar, but for the second moment multiplicative energy.

Theorem 8. *For all finite and nonempty $A, C, D \subset \mathbb{F}$ with*

$$|A|^2|C(A+1)| \leq |D||C|^3, \quad |A||C(A+1)|^2 \leq |D|^2|C|^3, \quad |A||C||D| \min\{|C|, |D|\} \ll p^2,$$

we have

$$E^*(A, D) \ll \frac{|C(A+1)|^{3/2}|D|^{3/2}}{|C|^{1/2}} \log|A|.$$

The set $A+1$ appearing in these theorems can be changed to any translate $A+\lambda$ for $\lambda \neq 0$ by noting that $|C(A+1)| = |C(\lambda A+\lambda)|$ and renaming $A' = \lambda A$. For our purposes, we will use $\lambda = \pm 1$.

Proof of Theorem 7. Without loss of generality, we can assume that $0 \notin A, C, D$. We begin by proving

$$E_4^*(A, D) \ll \frac{|C(A+1)|^2|D|^3}{|C|}\log|A|.$$

Define the set

$$S_\tau := \{x \in A/D : \tau \leq r_{A/D}(x) < 2\tau\}.$$

By a dyadic decomposition, there is some τ with

$$|S_\tau|\tau^4 \ll E_4^*(A, D) \ll |S_\tau|\tau^4 \log|A|.$$

Note that $\tau \leq \min\{|A|, |D|\}$. Take an element $t \in S_\tau$. It has τ representations in A/D, so there are τ ways to write $t = a/d$ with $a \in A,\ d \in D$. For all $c \in C$, we have

$$t = \frac{a}{d} = \frac{1}{d}\left(\frac{ac+c-c}{c}\right) = \frac{1}{d}\left(\frac{\alpha}{c} - 1\right),$$

where $\alpha = c(a+1) \in C(A+1)$. This shows that we have $|S_\tau|\tau|C|$ incidences between the lines

$$L = \{l_{d,c} : d \in D, c \in C\}, \quad l_{d,c} \text{ given by } y = \frac{1}{d}\left(\frac{x}{c} - 1\right),$$

and the point set $P = C(A+1) \times S_\tau$. Under the conditions $|D||C|\min\{|S_\tau|, |C(A+1)|\} \ll p^2$ and $|S_\tau||C(A+1)|\max\{|S_\tau|, |C(A+1)|\} \leq |D|^3|C|^3$, we have

$$|S_\tau|\tau|C| \leq I(P, L) \ll |C(A+1)|^{1/2}|S_\tau|^{3/4}|C|^{3/4}|D|^{3/4} + |D||C|.$$

The conditions are satisfied under the assumptions $|D||A||C|\min\{|D|, |C|\} \ll p^2$, $|A|^2|C(A+1)| \leq |D||C|^3$, and $|A||C(A+1)|^2 \leq |D|^2|C|^3$. Assuming that the leading term is dominant, we have

$$|S_\tau|\tau^4|C| \ll |C(A+1)|^2|D|^3$$

so that as $E_4^*(A, D)/\log|A| \ll |S_\tau|\tau^4$, we have

$$E_4^*(A, D) \ll \frac{|C(A+1)|^2|D|^3}{|C|}\log|A|.$$

We therefore assume the leading term is not dominant. Suppose $|D||C|$ is dominant so that

$$|C(A+1)|^{1/2}|S_\tau|^{3/4}|C|^{3/4}|D|^{3/4} \leq |D||C|. \tag{1}$$

Multiplying by τ^3 and simplifying, we have

$$|C(A+1)|^2\frac{E_4^*(A, D)^3}{\log|A|^3} \ll |C(A+1)|^2|S_\tau|^3\tau^{12} \leq |D||C|\tau^{12} \quad \Longrightarrow \quad E_4^*(A, D) \ll \frac{|D|^{1/3}|C|^{1/3}\tau^4}{|C(A+1)|^{2/3}}\log|A|.$$

The result now follows if

$$\frac{|D|^{1/3}|C|^{1/3}\tau^4}{|C(A+1)|^{2/3}} \ll \frac{|C(A+1)|^2|D|^3}{|C|}.$$

We must therefore prove the result in the case that this is not true; we will prove the result under the assumption

$$\frac{|C(A+1)|^2|D|^3}{|C|} \leq \frac{|D|^{1/3}|C|^{1/3}\tau^4}{|C(A+1)|^{2/3}},$$

which gives (using $\tau \leq |A|$)

$$|D|^8|C|^4|A|^4 \leq |D|^8|C(A+1)|^8 \leq \tau^{12}|C|^4 \leq |A|^{12}|C|^4,$$

so that we have $|D| \leq |A|$. We then have (using $|C(A+1)| \geq |C|^{1/2}|A|^{1/2}$)

$$|D||C| \geq |C(A+1)|^{1/2}|S_\tau|^{3/4}|C|^{3/4}|D|^{3/4} \geq |C(A+1)|^{1/2}|C|^{3/4}|D|^{3/4} \geq |A|^{1/4}|C||D|^{3/4} \geq |D||C|,$$

so that the two terms are in fact balanced and the result follows.

Secondly, we prove that

$$E_4^*(A, D) \ll \frac{|C(A+1)|^3|D|^2}{|C|}\log|A|.$$

To do this, we swap the roles of D and S_τ from above. We define the line set and point set by

$$L = \{l_{t,c} : t \in S_\tau, c \in C\}, \quad P = C(A+1) \times D.$$

Any incidence from the previous point and line sets remains an incidence for the new ones, via

$$t = \frac{1}{d}\left(\frac{\alpha}{c} - 1\right) \iff d = \frac{1}{t}\left(\frac{\alpha}{c} - 1\right).$$

Under the conditions

$$|S_\tau||C|\min\{|D|, |C(A+1)|\} \ll p^2, \quad |D||C(A+1)|\max\{|D|, |C(A+1)|\} \leq |S_\tau|^3|C|^3, \tag{2}$$

we have

$$|S_\tau|\tau|C| \leq I(P, L) \ll |C(A+1)|^{3/4}|S_\tau|^{3/4}|C|^{3/4}|D|^{1/2} + |S_\tau||C|.$$

If the leading term dominates, the result follows from $|S_\tau|\tau^4 \gg E_4^*(A, D)/\log|A|$. Assume the leading term is not dominant; that is,

$$|C(A+1)|^3|D|^2 \leq |S_\tau||C|.$$

Then by using $|S_\tau| \leq |A||D|$ and $|A|, |C| \leq |C(A+1)|$ we have

$$|A||C|^2|D|^2 \leq |C(A+1)|^3|D|^2 \leq |S_\tau||C| \leq |A||D||C|,$$

so that $|C| = |D| = 1$ and the result is trivial by $E_4^*(A, D) \leq |A||D|^4 \leq |A|$.

We now check the conditions (2) for using Theorem 5. The first condition in (2) is satisfied if $|A||C||D|^2 \ll p^2$, which is true under our assumptions. The second depends on $\max\{|D|, |C(A+1)|\}$, which we assume is $|D|$ (if not the first term in Theorem 7 gives stronger information, which we have already proved). Assuming the second condition does not hold, we have

$$|S_\tau|^3|C|^3 < |D|^2|C(A+1)|.$$

Multiplying by τ^{12} and bounding $\tau \le |A|$, we get

$$E_4^*(A, D) \ll \frac{|A|^4 |D|^{2/3} |C(A+1)|^{1/3}}{|C|} \log |A|. \tag{3}$$

We may now assume the bound

$$\frac{|C(A+1)|^3 |D|^2}{|C|} \le \frac{|A|^4 |D|^{2/3} |C(A+1)|^{1/3}}{|C|}. \tag{4}$$

Indeed, if we were to have

$$\frac{|A|^4 |D|^{2/3} |C(A+1)|^{1/3}}{|C|} < \frac{|C(A+1)|^3 |D|^2}{|C|}$$

then we may apply this bound in (3) and the result follows. Assuming (4), we have

$$|A|^8 |D|^4 \le |C(A+1)|^8 |D|^4 \le |A|^{12}.$$

So that $|D| \le |A|$. In turn, this implies $|A| \ge |D| \ge |C(A+1)| \ge |A|$, so that $|A| = |C(A+1)| = |D|$. Returning to (3), this gives

$$E_4^*(A, D) \ll \frac{|A|^4 |D|^{2/3} |C(A+1)|^{1/3}}{|C|} \log |A| = \frac{|C(A+1)|^3 |D|^2}{|C|} \log |A|,$$

and the result is proved. □

Proof of Theorem 8. The proof follows similarly to that of Theorem 7. We again define the lines and points

$$L = \{l_{d,c} : d \in D, c \in C\}, \quad l_{d,c} \text{ given by } y = \frac{1}{d}\left(\frac{x}{c} - 1\right), \qquad P = C(A+1) \times S_\tau,$$

where in this case the set S_τ is rich with respect to $E^*(A, D)$, so that

$$|S_\tau| \tau^2 \ll E^*(A, D) \ll |S_\tau| \tau^2 \log |A|.$$

With the conditions $|A||C||D| \min\{|D|, |C|\} \ll p^2$ and $|S_\tau||C(A+1)| \max\{|S_\tau|, |C(A+1)|\} \le |D|^3 |C|^3$ (which are satisfied under our assumptions), we have, by Theorem 6,

$$|S_\tau| \tau |C| \le I(P, L) \ll |S_\tau|^{1/2} |C(A+1)|^{3/4} |D|^{3/4} |C|^{3/4} + |D||C|.$$

If the leading term dominates, we have

$$|S_\tau| \tau^2 \ll \frac{|C(A+1)|^{3/2} |D|^{3/2}}{|C|^{1/2}}$$

and the result follows from $E^*(A, D)/\log |A| \ll |S_\tau| \tau^2$. We therefore assume that the leading term does not dominate; that is,

$$|S_\tau|^{1/2} |C(A+1)|^{3/4} |D|^{3/4} |C|^{3/4} \le |D||C|.$$

Multiplying through by τ and squaring, we get the bound

$$E^*(A, D) \ll \frac{|D|^{1/2}|C|^{1/2}\tau^2}{|C(A+1)|^{3/2}} \log |A|. \tag{5}$$

Much as before, we may now assume the bound

$$\frac{|D|^{3/2}|C(A+1)|^{3/2}}{|C|^{1/2}} \leq \frac{|D|^{1/2}|C|^{1/2}\tau^2}{|C(A+1)|^{3/2}}, \tag{6}$$

as assuming otherwise yields the result via (5). The bound (6) then gives

$$|D||C(A+1)|^3 \leq |C|\tau^2.$$

Bounding $\tau \leq |A|$ and $|C||A|^2 \leq |C(A+1)|^3$, we have $|D| = 1$. Similarly, bounding $\tau^2 \leq |A||D|$ and $|C(A+1)|^3 \geq |C|^2|A|$, we find $|C| = 1$, so that the result is trivial. □

3. Proof of Theorem 3

We follow a multiplicative analogue of the argument in [Rudnev et al. 2018]. Without loss of generality we may assume $A, B \subseteq \mathbb{F}^*$. For some $\delta > 0$, define a popular set of products as

$$P := \left\{ x \in AB : r_{AB}(x) \geq \frac{|A||B|}{|AB|\delta} \right\}.$$

Let $P^c := AB \setminus P$. Note that by writing

$$|\{(a, b) \in A \times B : ab \in P\}| + |\{(a, b) \in A \times B : ab \in P^c\}| = |A||B|$$

and noting that

$$|\{(a, b) \in A \times B : ab \in P^c\}| < |P^c|\frac{|A||B|}{|AB|\delta} \leq \frac{|A||B|}{\delta},$$

we have

$$|\{(a, b) \in A \times B : ab \in P\}| \geq \left(1 - \frac{1}{\delta}\right)|A||B|.$$

We also define a popular subset of A with respect to P as

$$A' := \left\{a \in A : |\{b \in B : ab \in P\}| \geq \tfrac{2}{3}|B|\right\}.$$

We have

$$|\{(a, b) \in A \times B : ab \in P\}| = \sum_{a \in A'} |\{b : ab \in P\}| + \sum_{a \in A \setminus A'} |\{b : ab \in P\}| \geq \left(1 - \frac{1}{\delta}\right)|A||B|. \tag{7}$$

Suppose that $|A \setminus A'| = c|A|$ for some $c \geq 0$, so that $|A'| = (1 - c)|A|$. Noting that

$$\sum_{a \in A'} |\{b : ab \in P\}| \leq (1 - c)|A||B|, \qquad \sum_{a \in A \setminus A'} |\{b : ab \in P\}| \leq \frac{2c}{3}|A||B|,$$

we have by (7)

$$(1 - c)|A||B| + \frac{2c}{3}|A||B| \geq \left(1 - \frac{1}{\delta}\right)|A||B| \Longrightarrow c \leq \frac{3}{\delta},$$

so that $|A'| \geq (1 - 3/\delta)|A|$.

We use a multiplicative version of Lemma 8 in [Rudnev et al. 2018]. The proof we present is an expanded version of the proof present in that paper.

Lemma 9. *For all finite $A \subset \mathbb{F}$, there exists $A_1 \subseteq A$ with $|A_1| \gg |A|$ such that*

$$E^*_{4/3}(A'_1) \gg E^*_{4/3}(A_1).$$

Proof. We give an algorithm which shows such a subset exists, as otherwise we have a contradiction. We recursively define

$$A_i = A'_{i-1}, \quad A_0 = A, \quad i \leq \log|A|,$$

where A'_i is defined relative to A_i. Using the same arguments as above, we have $|A'_i| \geq (1 - 3/\delta)|A_i|$. We shall set $\delta = \log|A|$. We have the chain of inequalities

$$|A_i| = |A'_{i-1}| \geq \left(1 - \frac{3}{\log|A|}\right)|A_{i-1}| \geq \cdots \geq \left(1 - \frac{3}{\log|A|}\right)^i |A|.$$

Note that assuming $|A| \geq 16$ (if this is not true then the result is trivial), we have

$$\left(1 - \frac{3}{\log|A|}\right)^i \geq \left(1 - \frac{3}{\log|A|}\right)^{\log|A|} \geq \left(\frac{1}{4}\right)^4$$

since the function $(1 - 3/z)^z$ is increasing for $z > 3$. We now have

$$|A_i| \geq \left(\frac{1}{4}\right)^4 |A| \gg |A|$$

at all steps i. We assume that at all steps, we have

$$E^*_{4/3}(A'_i) < \frac{E^*_{4/3}(A_i)}{4},$$

as otherwise we have $E^*_{4/3}(A'_i) \gg E^*_{4/3}(A_i)$ and we are done. After $\log|A|$ steps, we have a set A_k with

$$|A_k| \gg |A|, \quad E^*_{4/3}(A'_k) < \frac{E^*_{4/3}(k)}{4} < \frac{E^*_{4/3}(A_{k-1})}{16} < \cdots < \frac{E^*_{4/3}(A)}{4^{\log|A|}}.$$

But then we have

$$E^*_{4/3}(A) > E^*_{4/3}(A'_k)4^{\log|A|} \gg |A|^{4/3+2} = |A|^{10/3},$$

which is a contradiction. Therefore at some step we have an A_i satisfying the lemma. □

We now return to the proof of Theorem 3, with $\delta = \log|A|$ applied in the definition of P. We apply Lemma 9 to A to find a large subset $A_1 \subset A$ with $E^*_{4/3}(A'_1) \gg E^*_{4/3}(A_1)$, $|A_1| \gg |A|$. Noting that proving the result for A_1 implies it for A, we shall rename A_1 as A for simplicity.

We use a dyadic decomposition to find a set $Q \subset A'/A'$ such that

$$|Q|\Delta^{4/3} \ll E^*_{4/3}(A') \ll |Q|\Delta^{4/3}\log|A|$$

for some $\Delta > 0$.

We will bound the size of the set

$$N = \left\{(a, a', b, b') \in (A')^2 \times B^2 : \frac{a}{a'} \in Q,\ ab, ab', a'b, a'b' \in P\right\}.$$

By summing over all $a, a' \in A'$ with $a/a' \in Q$, we have

$$|N| = \sum_{\substack{a,a' \in A' \\ a/a' \in Q}} |\{b \in B : ab, a'b \in P\}|^2$$

and we see that as $|\{b \in B : ab \in P\}| \geq \frac{2}{3}|B|$ for all $a \in A'$, by considering the intersection of $\{b \in B : ab \in P\}$ and $\{b \in B : a'b \in P\}$, we have $|\{b \in B : ab, a'b \in P\}| \geq \frac{1}{3}|B|$ for all $a, a' \in A'$. Using that elements $q \in Q$ have at least Δ representations in A'/A', we have $|N| \geq \frac{1}{9}|B|^2|Q|\Delta$.

We now find an upper bound on $|N|$. Define an equivalence relation on $A^2 \times B^2$ via

$$(a, a', b, b') \sim (c, c', d, d') \quad \Longleftrightarrow \quad \text{there exists } \lambda \text{ such that } a = \lambda c,\ a' = \lambda c',\ b = \frac{d}{\lambda},\ b' = \frac{d'}{\lambda}.$$

Note that the conditions

$$\frac{a}{a'} \in Q, \quad ab, a'b, ab', a'b' \in P \tag{8}$$

are invariant in the class (i.e., if one class element satisfies these conditions, then they all do), as λ cancels in each condition. Let X denote the set of equivalence classes $[a, a', b, b']$, where the conditions (8) are satisfied. We can bound $|N|$ by the sum of the size of each equivalence class $[a, a', b, b']$ in X:

$$|N| \leq \sum_X |[a, a', b, b']|.$$

By the Cauchy–Schwarz inequality and completing the sum over all equivalence classes, we have

$$|Q|^2\Delta^2|B|^4 \ll |N|^2 \leq |X| \sum_{[a,a',b,b']} |[a, a', b, b']|^2. \tag{9}$$

We must now bound the two quantities on the right-hand side of this equation. We first claim that

$$\sum_{[a,a',b,b']} |[a, a', b, b']|^2 \leq \sum_x r_{A/A}(x)^2 r_{B/B}(x)^2. \tag{10}$$

To see this, note that the left-hand side of (10) counts pairs of elements of equivalence classes. Take any two elements $(a, a', b, b'),\ (c, c', d, d') \in A^2 \times B^2$ from the same equivalence class. By definition, we may write $(c, c', d, d') = (\lambda a, \lambda a', b/\lambda, b'/\lambda)$. As $0 \notin A, B$, the 8-tuple $(a, a', b, b', c, c', d, d')$ satisfies

$$\lambda = \frac{c}{a} = \frac{c'}{a'} = \frac{b}{d} = \frac{b'}{d'}$$

for some $\lambda \in \mathbb{R}$, and thus corresponds to a contribution to the quantity $r_{A/A}(\lambda)^2 r_{B/B}(\lambda)^2$, and thus also corresponds to a contribution to the sum $\sum_x r_{A/A}(x)^2 r_{B/B}(x)^2$. We also see that different pairs from equivalence classes necessarily give different 8-tuples, and so the claim is proved. We use Cauchy–Schwarz on the right-hand side of (10) to bound it by a product of fourth energies:

$$\sum_x r_{A/A}(x)^2 r_{B/B}(x)^2 \leq E_4^*(A)^{1/2} E_4^*(B)^{1/2}.$$

We use Theorem 7 to bound these energies. We bound via

$$E_4^*(A) \ll \frac{|C(A+1)|^2|A|^3}{|C|}\log|A|, \quad E_4^*(B) \ll \frac{|D(B-1)|^2|B|^3}{|D|}\log|B|,$$

with conditions

$$|C(A+1)||A| \le |C|^3, \quad |C(A+1)|^2 \le |A||C|^3, \quad |A|^3|C| \ll p^2,$$
$$|D(B-1)||B| \le |D|^3, \quad |D(B-1)|^2 \le |B||D|^3, \quad |B|^3|D| \ll p^2,$$

which are all satisfied under our assumptions. Returning to (9), we now have

$$|Q|^2\Delta^2|B|^4 \ll |X|\frac{|C(A+1)||A|^{3/2}|D(B-1)||B|^{3/2}}{|C|^{1/2}|D|^{1/2}}(\log|A|\log|B|)^{1/2}. \tag{11}$$

We now bound $|X|$, the number of equivalence classes where the conditions (8) are satisfied. Note that any (a,a',b,b') belonging to an equivalence class in X maps to a solution of the equation

$$w = \frac{s}{t} = \frac{u}{v}, \tag{12}$$

with $w \in Q,\ s,t,u,v \in P$, by taking $w = a/a',\ s = ab,\ t = a'b,\ u = ab',\ v = a'b'$. Note that taking two solutions (a,a',b,b') and (c,c',d,d') that are *not* from the same equivalence class necessarily gives us two different solutions to (12) via the map above. Therefore we may bound $|X|$ by the number of solutions to (12).

$$|X| \le \left|\left\{(w,s,t,u,v) \in Q\times P^4 : w = \frac{s}{t} = \frac{u}{v}\right\}\right| = \left|\left\{(s,t,u,v) \in P^4 : \frac{s}{t} = \frac{u}{v} \in Q\right\}\right|.$$

The popularity of P allows us to bound this by

$$|X| \le \frac{|AB|^4(\log|A|)^4}{|A|^4|B|^4}\left|\left\{(a_1,a_2,a_3,a_4,b_1,b_2,b_3,b_4) \in A^4\times B^4 : \frac{a_1b_1}{a_2b_2} = \frac{a_3b_3}{a_4b_4} \in Q\right\}\right|.$$

We dyadically pigeonhole the set BA/A in relation to the number of solutions to $r/a = r'/a' \in Q$, with $r,r' \in BA/A,\ a,a' \in A$, to find popular subsets $R_1, R_2 \subseteq BA/A$ in terms of these solutions. We have

$$|X| \le \frac{|AB|^4(\log|A|)^4}{|A|^4|B|^4}\sum_{i=1}^{2\log|A|}\sum_{\substack{x\in AB/A\\ 2^i\le r_{AB/A}(x)<2^{i+1}}} r_{AB/A}(x)\left|\left\{(a_3,a_4,b_1,b_3,b_4) \in A^2\times B^3 : \frac{x}{b_1} = \frac{a_3b_3}{a_4a_4} \in Q\right\}\right|.$$

We use the pigeonhole principle to give us $\Delta_1 > 0$ and $R_1 \subseteq AB/A$ such that

$$|X| \ll \Delta_1\frac{|AB|^4(\log|A|)^5}{|A|^4|B|^4}\left|\left\{(r_1,a_3,a_4,b_2,b_3,b_4) \in R_1\times A^2\times B^3 : \frac{r_1}{b_2} = \frac{a_3b_3}{a_4b_4} \in Q\right\}\right|.$$

We perform a similar dyadic decomposition to get $\Delta_1' > 0$ and $R_2 \subseteq AB/A$ such that

$$|X| \ll \Delta_1\Delta_1'\frac{|AB|^4(\log|A|)^6}{|A|^4|B|^4}\left|\left\{(r_1,r_2,b_2,b_4) \in R_1\times R_2\times B^2 : \frac{r_1}{b_2} = \frac{r_2}{b_4} \in Q\right\}\right|.$$

These decompositions now allow us to bound via fourth energies, as follows:

$$\begin{aligned}|X| &\ll \Delta_1\Delta_1'\frac{|AB|^4(\log|A|)^6}{|A|^4|B|^4}\left|\left\{(r_1,r_2,b_2,b_4)\in R_1\times R_2\times B^2 : \frac{r_1}{b_2}=\frac{r_2}{b_4}\in Q\right\}\right|\\ &= \Delta_1\Delta_1'\frac{|AB|^4(\log|A|)^6}{|A|^4|B|^4}\sum_{q\in Q} r_{R_1/B}(q)r_{R_2/B}(q)\\ &\le \Delta_1\Delta_1'\frac{|AB|^4(\log|A|)^6}{|A|^4|B|^4}\left(\sum_{q\in Q} r_{R_1/B}(q)^2\right)^{1/2}\left(\sum_{q\in Q} r_{R_2/B}(q)^2\right)^{1/2}\\ &\le \Delta_1\Delta_1'|Q|^{1/2}\frac{|AB|^4(\log|A|)^6}{|A|^4|B|^4}E_4^*(B,R_1)^{1/4}E_4^*(B,R_2)^{1/4}, \end{aligned} \tag{13}$$

where the third and fourth lines follow from applications of the Cauchy–Schwarz inequality. We will now show that given $|B||D||R_i|^2 \ll p^2$ and $|B|\le|D|$ (which are true under our assumptions), we have

$$E_4^*(B,R_i) \ll \frac{|D(B-1)|^3|R_i|^2}{|D|}\log|B|. \tag{14}$$

Firstly, with the additional conditions

$$|B|^2|D(B-1)| \le |R_i||D|^3, \quad |B||D(B-1)|^2 \le |R_i|^2|D|^3 \tag{15}$$

we may bound these fourth energies by Theorem 7 to get (14). We can therefore assume one of these conditions does not hold.

Firstly, suppose that $|B|^2|D(B-1)| > |R_i||D|^3$. We will use the trivial bound

$$E_4^*(B,R_i) \le |R_i|^4|B|.$$

Note that it would be enough to prove

$$E_4^*(B,R_i) \le \frac{|D(B-1)|^3|R_i|^2}{|D|},$$

which would follow from

$$|R_i|^4|B| \le \frac{|D(B-1)|^3|R_i|^2}{|D|}, \tag{16}$$

which is true if and only if $|R_i|^2|B||D| \le |D(B-1)|^3$. Using our assumed bound $|B|^2|D(B-1)| > |R_i||D|^3$, we know

$$|R_i|^2|B||D| < \frac{|B|^5|D(B-1)|^2}{|D|^5}.$$

By the assumption $|B|\le|D|$, we have

$$|R_i|^2|B||D| < \frac{|B|^5|D(B-1)|^2}{|D|^5} \le |D(B-1)|^3,$$

and so by (16) the bound on the fourth energy holds.

Now assume the second condition from (15) does not hold; that is, $|B||D(B-1)|^2 > |R_i|^2|D|^3$. Again, we use the trivial bound

$$E_4^*(B, R_i) \le |R_i|^4|B|.$$

We have

$$|R_i|^4|B| \le \frac{|D(B-1)|^3|R_i|^2}{|D|} \quad \Longleftrightarrow \quad |R_i|^2|B||D| \le |D(B-1)|^3,$$

so it is enough to prove $|R_i|^2|B||D| \le |D(B-1)|^3$, as before. Using the assumption $|B||D(B-1)|^2 > |R_i|^2|D|^3$, we have

$$|R_i|^2|B||D| < \frac{|B|^2|D(B-1)|^2}{|D|^2}$$

and it follows from our assumption $|B| \le |D|$ that

$$\frac{|B|^2|D(B-1)|^2}{|D|^2} \le |D(B-1)|^3.$$

Therefore we have $|R_i|^2|B||D| < |D(B-1)|^3$ and so the bound on the fourth energy holds. Returning to (13), we use (14) to bound $|X|$ as

$$\begin{aligned} |X| &\ll \Delta_1\Delta_1'|Q|^{1/2}\frac{|AB|^4(\log|A|)^6}{|A|^4|B|^4}E_4^*(B, R_1)^{1/4}E_4^*(B, R_2)^{1/4} \\ &\ll \Delta_1\Delta_1'|R_1|^{1/2}|R_2|^{1/2}|Q|^{1/2}\frac{|AB|^4|D(B-1)|^{3/2}}{|A|^4|B|^4|D|^{1/2}}(\log|A|)^6(\log|B|)^{1/2}. \end{aligned} \tag{17}$$

As $|R_i|\Delta_i \le \sum_{x\in R_i} r_{BA/A}(x)$, the product $|R_1|^{1/2}|R_2|^{1/2}\Delta_1\Delta_1'$ can be bounded by

$$|R_1|^{1/2}|R_2|^{1/2}\Delta_1\Delta_1' \le \Bigg(\sum_{x\in R_1} r_{BA/A}(x)^2 \sum_{x\in R_2} r_{BA/A}(x)^2\Bigg)^{1/2},$$

where it is important to note that $r_{BA/A}(x)$ gives a triple (b, a, a'). For $i = 1, 2$, we have

$$\sum_{x\in R_i} r_{BA/A}(x)^2 \le \left|\left\{(a_1, a_2, a_3, a_4, b_1, b_2) \in A^4 \times B^2 : \frac{b_1a_1}{a_2} = \frac{b_2a_3}{a_4}\right\}\right|.$$

Following a similar dyadic decomposition as before, we find a pair of subsets $S_1, S_2 \subseteq A/A$ with respect to these solutions, and some $\Delta_2, \Delta_2' > 0$ with

$$\begin{aligned} \sum_{x\in R_i} r_{BA/A}(x)^2 &\ll \Delta_2\Delta_2'(\log|A|)^2\big|\{(s_1, s_2, b_1, b_2) \in S_1 \times S_2 \times B^2 : s_1b_1 = s_2b_2\}\big| \\ &\le \Delta_2\Delta_2'(\log|A|)^2\sum_x r_{S_1B}(x)r_{S_2B}(x) \\ &\le \Delta_2\Delta_2'(\log|A|)^2E^*(B, S_1)^{1/2}E^*(B, S_2)^{1/2}, \end{aligned}$$

where the third inequality is given by the Cauchy–Schwarz inequality. We will use an argument similar to that above to prove that with the two conditions $|B||D||S_i|\min\{|D|, |S_i|\} \ll p^2$ and $|B| \le |D|$ (which

are satisfied under our assumptions), we have

$$E^*(B, S_i) \ll \frac{|S_i|^{3/2}|D(B-1)|^{3/2}}{|D|^{1/2}} \log |B|. \tag{18}$$

Under the extra conditions

$$|B|^2|D(B-1)| \le |S_i||D|^3, \quad |B||D(B-1)|^2 \le |S_i|^2|D|^3 \tag{19}$$

we can bound this energy by Theorem 8 to get (18). We therefore assume the first condition from (19) does not hold; that is, $|B|^2|D(B-1)| > |S_i||D|^3$. We bound the energy via the trivial estimate

$$E^*(B, S_i) \le |B||S_i|^2.$$

It is now enough to show that

$$|B||S_i|^2 \le \frac{|S_i|^{3/2}|D(B-1)|^{3/2}}{|D|^{1/2}}, \quad \text{which is true if and only if} \quad |B||D|^{1/2}|S_i|^{1/2} \le |D(B-1)|^{3/2}.$$

Using our assumption $|B|^2|D(B-1)| > |S_i||D|^3$, we have

$$|B||D|^{1/2}|S_i|^{1/2} < \frac{|B|^2|D(B-1)|^{1/2}}{|D|}.$$

Our assumption that $|B| \le |D|$ then gives

$$\frac{|B|^2|D(B-1)|^{1/2}}{|D|} \le |B||D(B-1)|^{1/2} \le |D(B-1)|^{3/2},$$

so that $|B||D|^{1/2}|S_i|^{1/2} < |D(B-1)|^{3/2}$, and the bound (18) holds. Next we assume that the second condition in (19) does not hold; that is, $|B||D(B-1)|^2 > |S_i|^2|D|^3$. We again use the trivial bound

$$E^*(B, S_i) \le |B||S_i|^2.$$

Comparing this to our desired bound, we have

$$|B||S_i|^2 \le \frac{|S_i|^{3/2}|D(B-1)|^{3/2}}{|D|^{1/2}} \quad \Longleftrightarrow \quad |B||D|^{1/2}|S_i|^{1/2} \le |D(B-1)|^{3/2},$$

so that the desired bound would follow from the second inequality above. Using our assumption that $|B||D(B-1)|^2 > |S_i|^2|D|^3$, we know

$$|B||D|^{1/2}|S_i|^{1/2} < \frac{|B|^{5/4}|D(B-1)|^{1/2}}{|D|^{1/4}},$$

and by our assumption that $|B| \le |D|$, we have

$$\frac{|B|^{5/4}|D(B-1)|^{1/2}}{|D|^{1/4}} \le |D(B-1)|^{3/2},$$

so that we have $|B||D|^{1/2}|S_i|^{1/2} < |D(B-1)|^{3/2}$ as needed.

In all cases the bound on $E^*(B, S_i)$ holds, so that we find

$$\begin{aligned}[|R_1|^{1/2}|R_2|^{1/2}\Delta_1\Delta_1']^2 &\ll \Delta_2^2\Delta_2'^2 E^*(B,S_1)E^*(B,S_2)(\log|A|)^4\\ &\ll \frac{\Delta_2^2\Delta_2'^2|S_1|^{3/2}|S_2|^{3/2}|D(B-1)|^3}{|D|}(\log|A|)^4(\log|B|)^2\\ &\le \frac{E^*_{4/3}(A)^3|D(B-1)|^3}{|D|}(\log|A|)^4(\log|B|)^2,\end{aligned}$$

where the final inequality follows as Δ_2 and Δ_2' correspond to representations of elements of S_1 and S_2 in A/A, so that

$$|S_1|^{3/2}\Delta_2^2 = (|S_1|\Delta_2^{4/3})^{3/2} \le \left(\sum_x r_{A/A}(x)^{4/3}\right)^{3/2} \le E^*_{4/3}(A)^{3/2},$$

and similarly for S_2. Combining the bounds (11), (17), and the above, we have

$$|Q|^{3/2}\Delta^2|B|^{13/2}|A|^{5/2}|D|^{3/2}|C|^{1/2} \ll |AB|^4|C(A+1)||D(B-1)|^4E^*_{4/3}(A)^{3/2}(\log|A|)^{17/2}(\log|B|)^2,$$

which simplifies to

$$E^*_{4/3}(A')^3|B|^{13}|A|^5|D|^3|C| \ll |AB|^8|C(A+1)|^2|D(B-1)|^8E^*_{4/3}(A)^3(\log|A|)^{17}(\log|B|)^4.$$

We know by Lemma 9 that $E_{4/3}(A') \gg E_{4/3}(A)$, so we have

$$|B|^{13}|A|^5|D|^3|C| \ll |AB|^8|C(A+1)|^2|D(B-1)|^8(\log|A|)^{17}(\log|B|)^4$$

as needed. □

Acknowledgements

The author was supported by the Austrian Science Fund (FWF) Project P 30405-N32. The author would also like to thank Oliver Roche-Newton and Misha Rudnev for helpful conversations. I would also like to thank the anonymous referee for several helpful comments.

References

[Chen et al. 2018] C. Chen, B. Kerr, and A. Mohammadi, "A new sum-product estimate in prime fields", preprint, 2018. arXiv

[Garaev and Shen 2010] M. Z. Garaev and C.-Y. Shen, "On the size of the set $A(A+1)$", *Math. Z.* **265**:1 (2010), 125–132. MR Zbl

[Jones and Roche-Newton 2013] T. G. F. Jones and O. Roche-Newton, "Improved bounds on the set $A(A+1)$", *J. Combin. Theory Ser. A* **120**:3 (2013), 515–526. MR Zbl

[Konyagin and Shkredov 2015] S. V. Konyagin and I. D. Shkredov, "On sum sets of sets having small product set", *Tr. Mat. Inst. Steklova* **290** (2015), 304–316. In Russian; translated in *Proc. Steklov Inst. Math.* **290**:1 (2015), 288–299. MR Zbl

[Mohammadi 2018] A. Mohammadi, "On growth of the set $A(A+1)$ in arbitrary finite fields", preprint, 2018. arXiv

[Roche-Newton et al. 2016] O. Roche-Newton, M. Rudnev, and I. D. Shkredov, "New sum-product type estimates over finite fields", *Adv. Math.* **293** (2016), 589–605. MR Zbl

[Rudnev 2018] M. Rudnev, "On the number of incidences between points and planes in three dimensions", *Combinatorica* **38**:1 (2018), 219–254. MR Zbl

[Rudnev et al. 2018] M. Rudnev, G. Shakan, and I. Shkredov, "Stronger sum-product inequalities for small sets", preprint, 2018. arXiv

[Shakan 2018] G. Shakan, "On higher energy decompositions and the sum-product phenomenon", *Math. Proc. Cambridge Philos. Soc.* (online publication July 2018).

[Shakan and Shkredov 2018] G. Shakan and I. D. Shkredov, "Breaking the 6/5 threshold for sums and products modulo a prime", preprint, 2018. arXiv

[Solymosi 2009] J. Solymosi, "Bounding multiplicative energy by the sumset", *Adv. Math.* **222**:2 (2009), 402–408. MR Zbl

[Stevens and de Zeeuw 2017] S. Stevens and F. de Zeeuw, "An improved point-line incidence bound over arbitrary fields", *Bull. Lond. Math. Soc.* **49**:5 (2017), 842–858. MR Zbl

[Zhelezov 2015] D. Zhelezov, "On additive shifts of multiplicative almost-subgroups in finite fields", preprint, 2015. arXiv

Received 6 Dec 2018. Revised 30 Apr 2019.

AUDIE WARREN:

audie.warren@oeaw.ac.at
Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

msp

Moscow Journal of Combinatorics and Number Theory
Vol. 8, No. 3, 2019
dx.doi.org/10.2140/moscow.2019.8.263

The mean square discrepancy in the circle problem

Steven M. Gonek and Alex Iosevich

We study the mean square of the error term in the Gauss circle problem. A heuristic argument based on the consideration of off-diagonal terms in the mean square of the relevant Voronoi-type summation formula leads to a precise conjecture for the mean square of this discrepancy.

1. Introduction

Let $r(n)$ denote the number of representations of the integer n as a sum of two squares of integers and let

$$P(x) = \sum_{n \le x}{}' r(n) - \pi x + 1, \tag{1-1}$$

where the prime superscript on the summation means that $r(x)$ is counted with weight $\frac{1}{2}$ if x is an integer. Finding the best estimate of the discrepancy $P(x)$ is known as Gauss' circle problem. It is trivial that $P(x) \ll x^{1/2}$, and it is conjectured that $P(x) \ll x^{1/4+\epsilon}$, where here and throughout ϵ denotes a small positive number that may be different at each occurrence. In the opposite direction, G. H. Hardy [1915; 1916b] proved that $P(x) = \Omega_+(x^{1/4})$ and $P(x) = \Omega_-((x \log x)^{1/4})$, and this has been improved slightly by a number of mathematicians; for example, see [Soundararajan 2003]. Here the notation $f(x) = \Omega_+(g(x))$ means there is a sequence of real numbers $x_n \to \infty$ and a positive constant c such that $f(x_n) \ge c|g(x_n)|$ for all n. Similarly, $f(x) = \Omega_-(g(x))$ means there is a sequence $x_n \to \infty$ and a positive constant c such that $f(x_n) \le -c|g(x_n)|$ for all n.

In spite of more than a century of effort, for example, by Sierpiński [1906], van der Corput [1923], Kolesnik [1985], Iwaniec and Mozzochi [1988], and Huxley [2003], Gauss's circle problem has resisted solution. In an attempt to understand it better, mathematicians have considered several variants of the problem that exploit the fact that the average of $P(x)$ is easier to analyze. For example, there has been considerable interest in the mean square of the discrepancy

$$\int_0^X P(x)^2\,dx.$$

It is known that

$$\int_0^X P(x)^2\,dx = CX^{3/2} + Q(X), \tag{1-2}$$

The work of Gonek was partially supported by NSF grant DMS-1200582. The work of Iosevich was partially supported by NSF Grant DMS-1045404.

MSC2010: 11P21.

Keywords: lattice points, circle problem, discrepancy estimates.

where

$$C = \frac{1}{3\pi^2}\sum_{n=1}^{\infty}\frac{r^2(n)}{n^{3/2}} = \frac{16}{3\pi^2}\frac{\zeta_{\mathbb{Q}(i)}\left(\frac{3}{2}\right)^2}{\zeta(3)}(1+2^{-3/2})^{-1} = 1.69396\ldots \tag{1-3}$$

and $Q(X)$ is a function that is $o(X^{3/2})$. H. Cramér [1922] proved that $Q(X) \ll X^{5/4+\epsilon}$, E. Landau [1923] that $Q(X) \ll X^{1+\epsilon}$, A. Walfisz [1927] that $Q(X) \ll X\log^3 X$, and I. Kátai [1965] that $Q(X) \ll X\log^2 X$; see also the work of E. Preissmann [1988] for another proof. W. G. Nowak [2004] proved the estimate

$$Q(X) \ll X(\log X)^{3/2}\log\log X,$$

and Y.-K. Lau and K.-M. Tsang [2009] proved that

$$Q(X) \ll X\log X\log\log X, \tag{1-4}$$

the best estimate to date of which we are aware.

Our goal here is to conjecture a precise formula for $Q(X)$. We will then use this to determine how large and how small $Q(X)$ can be, and to uncover a previously unobserved phenomenon described below.

Conjecture. *There is a constant $0 < \vartheta < 1$ such that as $X \to \infty$,*

$$Q(X) = C(X)X - X + O(X^{\vartheta}), \tag{1-5}$$

where

$$C(X) = \lim_{N\to\infty}\frac{1}{2\pi^3}\sum_{1\le m,n\le N}\frac{r(n)r(m)\cos(2\pi(\sqrt{m}+\sqrt{n})\sqrt{X})}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})}.$$

That is,

$$\int_0^X P(x)^2\,dx = CX^{3/2} + C(X)X - X + O(X^{\vartheta}), \tag{1-6}$$

where C is given by (1-3).

The phenomenon referred to above is the presence of the slowly oscillating function $C(X)$ in (1-6). This points to why it is so difficult to determine the exact size of $Q(X)$. A. Ivić [1996; 2001] has shown that the Laplace transform of $P(x)^2$ is

$$\int_0^{\infty} P(x)^2 e^{-x/T}\,dx = \frac{1}{4}\left(\frac{T}{\pi}\right)^{3/2}\sum_{n=1}^{\infty} r^2(n)n^{-3/2} - T + O(T^{2/3+\epsilon}).$$

Comparing this with (1-6), we see that the Laplace transform does not "see" the oscillating term $C(X)$.

It is not obvious that the limit defining $C(X)$ exists. We prove this in:

Proposition 1. *For $X \ge 0$, $N \ge 1$ let*

$$C_N(X) = \frac{1}{2\pi^3}\sum_{1\le m,n\le N}\frac{r(n)r(m)\cos(2\pi(\sqrt{m}+\sqrt{n})\sqrt{X})}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})}. \tag{1-7}$$

Then for $X \ge 1$,

$$C(X) = \lim_{N\to\infty} C_N(X)$$

exists. Moreover, for $X \ge 1$ we have

$$|C(X) - C_N(X)| \ll \frac{X\log N}{N^{1/4}}. \tag{1-8}$$

In Section 9 we shall use Proposition 1 to prove:

Theorem 2. *As* $X \to \infty$,

$$|C(X)| \le \Big(\frac{16}{\pi} + o(1)\Big) \log X. \tag{1-9}$$

Hence, if the Conjecture is true,

$$|Q(X)| \le \Big(\frac{16}{\pi} + o(1)\Big) X \log X. \tag{1-10}$$

The upper bound (1-10) suggests that (1-4) is too large by a factor of at least $\log\log X$. However, we suspect that even (1-10) is larger than the true upper bound. It is possible that the lower bound for $Q(X)$ provided by the next theorem is closer to the actual upper bound.

Theorem 3. *We have*

$$\limsup_{X\to\infty} \frac{C(X)}{\log\log X} \ge \frac{1}{2\pi}.$$

Hence, if the Conjecture is true, then

$$Q(X) = \Omega_+(X \log\log X). \tag{1-11}$$

In view of the Conjecture one might ask whether one can model $C_N(X)$ by the sum

$$\frac{1}{2\pi^3} \sum_{m,n\le N} \frac{r(n)r(m)\cos(2\pi X_{m,n})}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})} \quad (X^4 \le N \le X^A),$$

where the $X_{m,n}$ are independent identically distributed random variables. This would be reasonable if the approximately $N^2/2$ numbers $\{\sqrt{m}+\sqrt{n}\}_{m\le n\le N}$ were linearly independent over the rationals. Using estimates of H. L. Montgomery and A. Odlyzko [1988] for large deviations of sums of random variables, one could then show that this sum is likely to be no larger than $O(\log\log X)$. Unfortunately, the numbers $\{\sqrt{m}+\sqrt{n}\}_{m\le n}$ are highly linearly dependent over the rationals in the sense that a relatively sparse subset of these numbers spans the set. For example, the $2N-1$ numbers $\sqrt{n}+\sqrt{2}$, $\sqrt{n}+\sqrt{3}$ $(n = 1, 2, \dots, N)$ allow us to write an arbitrary one of the approximately N^2 elements $\sqrt{k}+\sqrt{l}$ as $(\sqrt{k}+\sqrt{2})+(\sqrt{l}+\sqrt{3})-(\sqrt{2}+\sqrt{3})$. Thus, such a model might not be very accurate.

Our method may also be applied to other well-known problems. For example, it may be used to conjecture a formula for the term $F(X)$ in

$$\int_0^X \Delta(x)^2\,dx = \frac{X^{3/2}}{6\pi^2} \sum_{n=1}^{\infty} \frac{d(n)^2}{n^{3/2}} + F(X), \tag{1-12}$$

the mean square of the error term in the Dirichlet divisor problem, where

$$\Delta(x) = \sum_{n\le x}{}' d(n) - x(\log x + 2\gamma - 1) - \tfrac{1}{4},$$

$d(n) = \sum_{d|n} 1$, and γ is Euler's constant. This will be the subject of a forthcoming paper by the first author and Dr. Fan Ge. The important feature shared by the circle and divisor problems that makes our method applicable is, of course, the existence of a Voronoi-type summation formula for their error terms.

2. Proof of Proposition 1

Before proving Proposition 1, we gather several formulae and a lemma.

Hardy [1916a] proved that for $x > 0$

$$P(x) = x^{1/2} \sum_{n=1}^{\infty} \frac{r(n)}{n^{1/2}} J_1(2\pi\sqrt{nx}), \tag{2-1}$$

where J_1 is a Bessel function of the first kind. Using the approximation

$$J_1(u) = \Big(\frac{2}{\pi u}\Big)^{1/2} \Big(\cos\Big(u - \frac{3\pi}{4}\Big) - \frac{3}{8u}\sin\Big(u - \frac{3\pi}{4}\Big)\Big) + O(u^{-5/2}), \tag{2-2}$$

valid for $u \geq 1$, we deduce that

$$P(x) = \frac{x^{1/4}}{\pi} \sum_{n=1}^{\infty} \frac{r(n)}{n^{3/4}} \cos\Big(2\pi\sqrt{nx} - \frac{3\pi}{4}\Big) - \frac{3x^{-1/4}}{16\pi^2} \sum_{n=1}^{\infty} \frac{r(n)}{n^{5/4}} \sin\Big(2\pi\sqrt{nx} - \frac{3\pi}{4}\Big) + O(x^{-3/4}) \tag{2-3}$$

for $x \geq 1$. From this we obtain

$$P_1(x) := \int_0^x P(u)\,du = \frac{x^{3/4}}{\pi^2} \sum_{n=1}^{\infty} \frac{r(n)}{n^{5/4}} \sin\Big(2\pi\sqrt{nx} - \frac{3\pi}{4}\Big) + O(x^{1/4}) \tag{2-4}$$

for $x \geq 1$. Note that

$$P_1(x) \ll x^{3/4} \tag{2-5}$$

for $x \geq 1$ follows immediately from this.

Lemma 4. *Let $1 \leq A < B$ and let $0 < \epsilon < 1$. Then uniformly for $x \geq 1$ and for y real,*

$$\sum_{A \leq k \leq B} \frac{r(k)}{k^{3/4}} \cos(x\sqrt{k} + y) \ll \frac{x^2}{A^{1/2}} + \frac{1}{A^{1/4}} + x^{\epsilon - 1/2} \min\Big(\frac{x}{A^{1/2}\|(x/2\pi)^2\|}, \log\frac{B}{A}\Big). \tag{2-6}$$

The implied constant depends at most on ϵ.

Proof. Denote the left-hand side of (2-6) by S. Then by (1-1) we may write

$$S = \pi \int_A^B u^{-3/4} \cos(x\sqrt{u} + y)\,du + \int_{A^-}^B u^{-3/4} \cos(x\sqrt{u} + y)\,dP(u) = S_1 + S_2.$$

Now

$$\begin{aligned} S_1 &= \frac{2\pi}{x} \int_A^B u^{-1/4}\,d(\sin(x\sqrt{u} + y)) \\ &= \frac{2\pi \sin(x\sqrt{u} + y)}{xu^{1/4}}\Big|_A^B + \frac{\pi}{2x} \int_A^B \frac{\sin(x\sqrt{u} + y)}{u^{5/4}}\,du \ll A^{-1/4}x^{-1}. \end{aligned}$$

Using $P(u) \ll u^{1/2}$, we see that

$$\begin{aligned} S_2 &= \frac{P(u)\cos(x\sqrt{u} + y)}{u^{3/4}}\Big|_{A^-}^B + \int_A^B \frac{P(u)}{u^{7/4}}\big(\tfrac{3}{4}\cos(x\sqrt{u} + y) + \tfrac{1}{2}x\sqrt{u}\sin(x\sqrt{u} + y)\big)\,du \\ &= \frac{x}{2} \int_{A^-}^B \frac{P(u)}{u^{5/4}} \sin(x\sqrt{u} + y)\,du + O(A^{-1/4}). \end{aligned}$$

Thus, for $x \geq 1$, we have

$$S = \frac{x}{2} \int_{A^-}^{B} \frac{P(u)}{u^{5/4}} \sin(x\sqrt{u} + y)\, du + O(A^{-1/4}).$$

Let $P_1(u)$ be as in (2-4). Then by (2-5)

$$\begin{aligned} S &= \frac{x}{2} \int_{A^-}^{B} \frac{\sin(x\sqrt{u} + y)}{u^{5/4}}\, dP_1(u) + O(A^{-1/4}) \\ &= \frac{x}{2}\left(P_1(u) \frac{\sin(x\sqrt{u} + y)}{u^{5/4}} \Big|_{A^-}^{B} + \int_A^B P_1(u) \left(\frac{\frac{5}{4}\sin(x\sqrt{u} + y)}{u^{9/4}} - \frac{x}{2} \frac{\cos(x\sqrt{u} + y)}{u^{7/4}} \right) du \right) + O(A^{-1/4}) \\ &= -\left(\frac{x}{2}\right)^2 \int_A^B P_1(u) \frac{\cos(x\sqrt{u} + y)}{u^{7/4}}\, du + O(xA^{-1/2}) + O(A^{-1/4}). \end{aligned}$$

Next we insert the formula for $P_1(x)$ from (2-4) into the last integral. The $O(u^{1/4})$ term in (2-4) contributes $O(x^2/A^{1/2})$, so we obtain

$$S = -\frac{x^2}{4\pi^2} \sum_{n=1}^{\infty} \frac{r(n)}{n^{5/4}} \int_A^B \frac{\sin\left(2\pi\sqrt{nu} - \frac{3\pi}{4}\right) \cos(x\sqrt{u} + y)}{u}\, du + O(x^2 A^{-1/2}) + O(A^{-1/4}).$$

Writing the numerator in the integrand as a sum of two sines, we have

$$S = -\frac{x^2}{8\pi^2} \sum_{n=1}^{\infty} \frac{r(n)}{n^{5/4}} (I_1(n) + I_2(n)) + O(x^2 A^{-1/2}) + O(A^{-1/4}), \tag{2-7}$$

where

$$\begin{aligned} I_1(n) &= \int_A^B u^{-1} \sin\left((x + 2\pi\sqrt{n})\sqrt{u} + y - \frac{3\pi}{4} \right) du, \\ I_2(n) &= \int_A^B u^{-1} \sin\left((-x + 2\pi\sqrt{n})\sqrt{u} - y - \frac{3\pi}{4} \right) du. \end{aligned}$$

Integration by parts shows that

$$\begin{aligned} I_1(n) &\ll \frac{1}{A^{1/2}(x + 2\pi\sqrt{n})}, \\ I_2(n) &\ll \frac{1}{A^{1/2}|x - 2\pi\sqrt{n}|}. \end{aligned}$$

We also trivially have $I_2(n) \ll \log B/A$. It follows that

$$x^2 \sum_{n=1}^{\infty} \frac{r(n) I_1(n)}{n^{5/4}} \ll \frac{x}{A^{1/2}}$$

and

$$x^2 \sum_{n=1}^{\infty} \frac{r(n) I_2(n)}{n^{5/4}} \ll x^2 \sum_{n=1}^{\infty} \frac{r(n)}{n^{5/4}} \min\left(\frac{1}{A^{1/2}|x - 2\pi\sqrt{n}|}, \log\frac{B}{A} \right).$$

The terms in the last sum with $n \le \frac{1}{2}(x/2\pi)^2$ or $n > 2(x/2\pi)^2$ contribute

$$\ll \frac{x^2}{A^{1/2}}\left(\sum_{n\le \frac{1}{2}(x/2\pi)^2} + \sum_{n>2(x/2\pi)^2}\right)\frac{r(n)(x+2\pi\sqrt{n})}{n^{5/4}|x^2-4\pi^2 n|}$$

$$\ll \frac{x^2}{A^{1/2}}\left(\frac{1}{x}\sum_{n\le \frac{1}{2}(x/2\pi)^2}\frac{r(n)}{n^{5/4}} + \sum_{n>2(x/2\pi)^2}\frac{r(n)}{n^{7/4}}\right)$$

$$\ll \frac{x^2}{A^{1/2}}(x^{-1}+x^{-3/2}\log x) \ll \frac{x}{A^{1/2}}.$$

The remaining terms give

$$\ll \frac{x^2}{A^{1/2}} \sum_{\substack{\frac{1}{2}(x/2\pi)^2<n\le 2(x/2\pi)^2 \\ n\ne[(x/2\pi)^2],\,[(x/2\pi)^2]+1}} \frac{r(n)}{n^{5/4}}\frac{x}{|(x/2\pi)^2-n|} + x^{\epsilon-1/2}\min\left(\frac{x}{A^{1/2}\|(x/2\pi)^2\|}, \log\frac{B}{A}\right)$$

$$\ll \frac{x^{1/2+\epsilon}}{A^{1/2}} + x^{\epsilon-1/2}\min\left(\frac{x}{A^{1/2}\|(x/2\pi)^2\|}, \log\frac{B}{A}\right).$$

Thus, from (2-7) we conclude that

$$S \ll \frac{x^2}{A^{1/2}} + \frac{1}{A^{1/4}} + x^{\epsilon-1/2}\min\left(\frac{x}{A^{1/2}\|(x/2\pi)^2\|}, \log\frac{B}{A}\right).$$

This completes the proof of Lemma 4. □

We now prove Proposition 1. By definition

$$C_N(X) = \frac{1}{4\pi^3}\sum_{m\le N}\frac{r(m)^2\cos(4\pi\sqrt{X}\sqrt{m})}{m^2} + \frac{1}{\pi^3}\sum_{1\le m<n\le N}\frac{r(m)\,r(n)\,\cos(2\pi\sqrt{X}(\sqrt{m}+\sqrt{n}))}{(mn)^{3/4}\,(\sqrt{m}+\sqrt{n})}.$$

The second sum on the right equals

$$\sum_{1<n\le N} r(n)n^{-5/4}\sum_{1\le m\le n-1}\frac{r(m)\cos(2\pi\sqrt{X}(\sqrt{m}+\sqrt{n}))}{m^{3/4}(1+\sqrt{m/n})}.$$

Thus, for $M > N$ we have

$$|C_M(X)-C_N(X)| \le \sum_{N<m\le M}\frac{r(m)^2}{m^2} + \sum_{N<n\le M}\frac{r(n)}{n^{5/4}}\left|\sum_{1\le m\le n-1}\frac{r(m)\cos(2\pi\sqrt{X}(\sqrt{m}+\sqrt{n}))}{m^{3/4}(1+\sqrt{m/n})}\right|.$$

By partial summation, the sum over m within absolute values equals

$$\frac{S_{n-1}}{1+\sqrt{(n-1)/n}} + \sum_{1\le m\le n-2} S_m\left(\frac{1}{1+\sqrt{m/n}} - \frac{1}{1+\sqrt{(m+1)/n}}\right),$$

where

$$S_m = S_m(n) = \sum_{1\le k\le m}\frac{r(k)\cos(2\pi\sqrt{X}(\sqrt{k}+\sqrt{n}))}{k^{3/4}}.$$

Thus, for $X \geq 1$ we have

$$\sum_{1\leq m\leq n-1} \frac{r(m)\cos(2\pi\sqrt{X}(\sqrt{m}+\sqrt{n}))}{m^{3/4}(1+\sqrt{m/n})}$$
$$\ll \max_{1\leq m\leq n-1} |S_m(n)| \cdot \left(\frac{1}{1+\sqrt{(n-1)/n}} + \sum_{1\leq m\leq n-2}\left(\frac{1}{1+\sqrt{m/n}} - \frac{1}{1+\sqrt{(m+1)/n}}\right)\right)$$
$$\leq \max_{1\leq m\leq n-1} |S_m(n)|.$$

Now by Lemma 4 with $x = 2\pi\sqrt{X}$, $y = 2\pi\sqrt{nX}$, $A = 1$, and $B = m-1$, we see that

$$S_m(n) \ll_\epsilon X + X^{\epsilon-1/4}\log m$$

for any $0 < \epsilon < 1$. Therefore, since $r(n) \ll d(n)$, we find that for $M > N$,

$$|C_M(X) - C_N(X)| \ll \sum_{N<m\leq M} \frac{r(m)^2}{m^2} + \sum_{N<n\leq M} \frac{r(n)}{n^{5/4}}\left|\sum_{1\leq m\leq n-1} \frac{r(m)\cos(2\pi\sqrt{X}(\sqrt{m}+\sqrt{n}))}{m^{3/4}(1+\sqrt{m/n})}\right|$$
$$\ll_\epsilon \frac{(\log N)^3}{N} + \sum_{N<n\leq M} \frac{d(n)}{n^{5/4}}\left(X + \frac{\log n}{X^{1/4-\epsilon}}\right)$$
$$\ll_\epsilon \frac{(\log N)^3}{N} + \frac{X\log N}{N^{1/4}} + \frac{(\log N)^2}{N^{1/4}X^{1/4-\epsilon}} \ll \frac{X(\log N)}{N^{1/4}}.$$

In the last inequality, we have taken $\epsilon = \frac{1}{8}$, which allows us to make the implied constant absolute. It now follows from Cauchy's criterion that $\lim_{N\to\infty} C_N(X)$ exists as $N \to \infty$. The second assertion of Proposition 1 follows immediately from the last inequality.

3. Beginning of the argument

To avoid technical difficulties, we shall estimate

$$\mathcal{I}(X) = \int_{X/2}^{X} P(x)^2\,dx$$

rather than

$$\int_0^X P(x)^2\,dx,$$

and then add the results for $(X/2, X]$, $(X/4, X/2]$, $(X/8, X/4]\ldots$. Hardy [1916a] proved that

$$P(x) = x^{1/2}\sum_{n=1}^{\infty} \frac{r(n)}{n^{1/2}} J_1(2\pi\sqrt{nx}) \quad (x > 0), \tag{3-1}$$

where

$$J_1(y) = \frac{1}{\pi}\int_0^\pi \cos(nx - y\sin x)\,dx$$

is a Bessel function of the first kind. To estimate $\mathcal{I}(X)$ we use a truncated version of (3-1) due to Ivić [1996], which we state as:

Lemma 5. *Let $X \ge 2$ and $X \le N \le X^A$ with $A > 1$. For $x > 0$ define $R(x, N)$ by*

$$P(x) = x^{1/2} \sum_{n \le N} \frac{r(n)}{n^{1/2}} J_1(2\pi\sqrt{nx}) + R(x, N). \tag{3-2}$$

Then for $X/2 \le x \le X$ and any $\epsilon > 0$, we have

$$R(x, N) \ll \begin{cases} x^{\epsilon} & \text{always}, \\ \left(\frac{x}{N}\right)^{1/2} \frac{x^{\epsilon}}{\|x\|} + x^{3/4}\left(\frac{x}{N}\right)^{1/2} + \left(\frac{x}{N}\right)^{1/4} & \text{if } x \notin \mathbb{Z}. \end{cases}$$

If we take $X/2 \le x \le X$ and impose the condition that $X^4 \le N \le X^A$ in Lemma 5, then

$$R(x, N) \ll \min\left\{x^{\epsilon},\ \left(\frac{x}{N}\right)^{1/2} \frac{x^{\epsilon}}{\|x\|} + \left(\frac{x}{N}\right)^{1/4}\right\}$$

and we easily see that

$$\int_{X/2}^{X} |R(x, N)|^2\, dx \ll X^{3/2+\epsilon} N^{-1/2}.$$

From (1-2) and the known bounds for $Q(X)$, the mean square of the main term in (3-2) over $[X/2, X]$ is $O(X^{3/2})$. Thus, by the Cauchy–Schwarz inequality the contribution of the cross term to the mean square of (3-2) is $O(X^{3/2+\epsilon/2} N^{-1/4})$. Thus, if $N \ge X^4$,

$$\mathcal{I}(X) = \int_{X/2}^{X} \left(x^{1/2} \sum_{n \le N} \frac{r(n)}{n^{1/2}} J_1(2\pi\sqrt{nx})\right)^2 dx + O(X^{1/2+\epsilon}).$$

Of course, if we take N even larger, the error term will be smaller.

Next we use the approximation

$$J_1(u) = \left(\frac{2}{\pi u}\right)^{1/2} \left(\cos\left(u - \frac{3\pi}{4}\right) - \frac{3}{8u} \sin\left(u - \frac{3\pi}{4}\right)\right) + O(u^{-5/2}), \tag{3-3}$$

which is valid for $u \ge 1$, and find that

$$\begin{aligned} \mathcal{I}(X) = \frac{1}{\pi^2} \int_{X/2}^{X} \Bigg(x^{1/4} \sum_{n \le N} \frac{r(n)}{n^{3/4}} \cos\left(2\pi\sqrt{nx} - \frac{3\pi}{4}\right) - \frac{3}{16\pi x^{1/4}} \sum_{n \le N} \frac{r(n)}{n^{5/4}} \sin\left(2\pi\sqrt{nx} - \frac{3\pi}{4}\right) \\ + O\left(\frac{1}{x^{3/4}} \sum_{n \le N} \frac{r(n)}{n^{7/4}}\right)\Bigg)^2 dx + O(X^{1/2+\epsilon}) \\ = \frac{1}{\pi^2} \int_{X/2}^{X} (A_1(x) - A_2(x) + A_3(x))^2\, dx + O(X^{1/2+\epsilon}). \end{aligned} \tag{3-4}$$

The sums in $A_2(x)$ and $A_3(x)$ are $O(1)$ uniformly in N and x so

$$\int_{X/2}^{X} A_2(x)^2\, dx \ll X^{1/2} \quad \text{and} \quad \int_{X/2}^{X} A_3(x)^2\, dx \ll X^{-1/2}.$$

By (1-2) we therefore have

$$\int_{X/2}^{X} A_1(x)^2\, dx \ll X^{3/2}.$$

From these estimates and the Cauchy–Schwarz inequality we now have

$$\mathcal{I}(X) = \frac{1}{\pi^2}\int_{X/2}^{X} A_1(x)^2 - 2A_1(x)A_2(x)\,dx + O(X^{1/2+\epsilon}). \tag{3-5}$$

Note that if we were to apply the Cauchy–Schwarz inequality to the integral of $A_1(x)A_2(x)$, we would obtain the estimate $O(X)$ for this term, which is the expected size of the lowest-order term in our main term. By isolating this term and treating it with a little more care, we shall show that it is in fact $O(X^{1/2}\log X)$.

The sums in the definitions of $A_1(x)$ and $A_2(x)$ contain trigonometric rather than Bessel functions. This makes it convenient to again work with integrals over $[0, X]$ rather than $[X/2, X]$, and then to take the difference of the results for $[0, X]$ and $[0, X/2]$ at the end of the argument. To proceed we use the identity

$$\cos a\ \cos b = \Re\tfrac{1}{2}[\exp(i(a-b)) + \exp(i(a+b))]$$

and obtain

$$\begin{aligned}\frac{1}{\pi^2}\int_0^X A_1(x)^2\,dx &= \Re\frac{1}{2\pi^2}\sum_{m,n\le N}\frac{r(n)r(m)}{(mn)^{3/4}}\int_0^X x^{1/2}\exp(2\pi i\sqrt{x}(\sqrt{n}-\sqrt{m}))\,dx\\ &\quad + \Re\frac{i}{2\pi^2}\sum_{m,n\le N}\frac{r(n)r(m)}{(mn)^{3/4}}\int_0^X x^{1/2}\exp(2\pi i\sqrt{x}(\sqrt{n}+\sqrt{m}))\,dx\\ &= \mathrm{I}(X) + \mathrm{J}(X).\end{aligned} \tag{3-6}$$

When $y \neq 0$, a substitution and two integrations by parts shows that

$$\int_0^X x^{1/2}\exp(2\pi i y\sqrt{x})\,dx = e^{2\pi i y\sqrt{X}}\left(\frac{X}{i\pi y} + \frac{\sqrt{X}}{\pi^2 y^2} - \frac{1}{i2\pi^3 y^3}\right) + \frac{1}{i2\pi^3 y^3}.$$

When $y = 0$ the integral equals $\frac{2}{3}X^{3/2}$ trivially. Thus

$$\begin{aligned}\mathrm{I}(X) &= \frac{X^{3/2}}{3\pi^2}\sum_{n\le N}\frac{r(n)^2}{n^{3/2}} + \frac{1}{2\pi^2}\sum_{m\neq n\le N}\frac{r(n)r(m)}{(mn)^{3/4}}\mathrm{i}_X(\sqrt{m}-\sqrt{n}),\\ \mathrm{J}(X) &= \frac{1}{2\pi^2}\sum_{m,n\le N}\frac{r(n)r(m)}{(mn)^{3/4}}\mathrm{j}_X(\sqrt{m}+\sqrt{n}),\end{aligned}$$

where

$$\begin{aligned}\mathrm{i}_X(y) &= \frac{X\sin(2\pi y\sqrt{X})}{\pi y} + \frac{\sqrt{X}\cos(2\pi y\sqrt{X})}{\pi^2 y^2} - \frac{\sin(2\pi y\sqrt{X})}{2\pi^3 y^3},\\ \mathrm{j}_X(y) &= \frac{X\cos(2\pi y\sqrt{X})}{\pi y} - \frac{\sqrt{X}\sin(2\pi y\sqrt{X})}{\pi^2 y^2} + \frac{1-\cos(2\pi y\sqrt{X})}{2\pi^3 y^3}.\end{aligned} \tag{3-7}$$

Similarly, using

$$\cos a\ \sin b = -\Im\tfrac{1}{2}[\exp(i(a-b)) - \exp(i(a+b))],$$

we obtain

$$\frac{1}{\pi^2}\int_0^X A_1(x)A_2(x)\,dx = -\Im\frac{3}{32\pi^3}\sum_{m,n\le N}\frac{r(n)r(m)}{n^{3/4}m^{5/4}}\int_0^X \exp(2\pi i\sqrt{x}(\sqrt{n}-\sqrt{m}))\,dx + \Im\frac{3i}{32\pi^3}\sum_{m,n\le N}\frac{r(n)r(m)}{n^{3/4}m^{5/4}}\int_0^X \exp(2\pi i\sqrt{x}(\sqrt{n}+\sqrt{m}))\,dx$$
$$= \mathrm{K}(X)+\mathrm{L}(X). \tag{3-8}$$

When $y\neq 0$,

$$\int_0^X \exp(2\pi i y\sqrt{x})\,dx = e^{2\pi i y\sqrt{X}}\left(\frac{\sqrt{X}}{i\pi y}+\frac{1}{2\pi^2y^2}\right)-\frac{1}{2\pi^2y^2},$$

whereas when $y=0$ the integral equals X. Thus

$$\mathrm{K}(X)=\frac{3}{32\pi^3}\sum_{m\neq n\le N}\frac{r(n)r(m)}{n^{3/4}m^{5/4}}\mathrm{k}_X(\sqrt{n}-\sqrt{m}),$$
$$\mathrm{L}(X)=\frac{3}{32\pi^3}\sum_{m,n\le N}\frac{r(n)r(m)}{n^{3/4}m^{5/4}}\mathrm{l}_X(\sqrt{n}+\sqrt{m}),$$

where

$$\mathrm{k}_X(y)=\frac{\sqrt{X}\cos(2\pi y\sqrt{X})}{\pi y}-\frac{\sin(2\pi y\sqrt{X})}{2\pi^2y^2},$$
$$\mathrm{l}_X=\frac{\sqrt{X}\sin(2\pi y\sqrt{X})}{\pi y}+\frac{\cos(2\pi y\sqrt{X})-1}{2\pi^2y^2}.$$

In Sections 4–7 we estimate $\mathrm{I}(X)$, $\mathrm{J}(X)$, $\mathrm{K}(X)$, and $\mathrm{L}(X)$. Our main terms come from $\mathrm{I}(X)$, which also requires the lengthiest treatment.

4. Calculation of $\mathrm{I}(X)$

We write

$$\mathrm{I}(X)=\frac{X^{3/2}}{3\pi^2}\sum_{n\le N}\frac{r(n)^2}{n^{3/2}}+\frac{1}{2\pi^2}\sum_{m\neq n\le N}\frac{r(n)r(m)}{(mn)^{3/4}}\mathrm{i}_X(\sqrt{m}-\sqrt{n})=\mathrm{I}_{\mathrm{D}}+\mathrm{I}_{\mathrm{O}}. \tag{4-1}$$

Setting

$$C=\frac{1}{3\pi^2}\sum_{n=1}^{\infty}\frac{r(n)^2}{n^{3/2}},$$

we see that since $r(n)\ll_\epsilon n^\epsilon$ for any $\epsilon>0$ and $N\ge X^4$,

$$\mathrm{I}_{\mathrm{D}}=CX^{3/2}+O_\epsilon(X^{3/2}N^{-1/2+\epsilon})=CX^{3/2}+O(1). \tag{4-2}$$

To evaluate $\mathrm{I}_{\mathrm{O}}(X)$ we write $m=n+h$ and use the symmetry in m and n to see that

$$\mathrm{I}_{\mathrm{O}}=\frac{1}{\pi^2}\sum_{n<N}\sum_{1\le h\le N-n}\frac{r(n)r(n+h)}{(n(n+h))^{3/4}}\mathrm{i}_X(\sqrt{n+h}-\sqrt{n}).$$

We replace $\sqrt{n+h}-\sqrt{n}$ by the approximation $h/(2\sqrt{n})$, and replace $n+h$ by n in the denominator to obtain

$$\mathrm{I_O} \approx \frac{1}{\pi^2}\sum_{n<N}\sum_{1\le h\le N-n}\frac{r(n)r(n+h)}{n^{3/2}}\mathrm{i}_X(h/(2\sqrt{n})); \tag{4-3}$$

here and below $A \approx B$ means that $A = B + E$, with E an error term of order less than B. This is the first place in the argument where we have abandoned rigor. F. Chamizo [1999, Corollary 5.3] has shown that

$$\sum_{n\le x} r(n)r(n+h) = c(h)x + E(x,h),$$

where

$$c(h) = \frac{8(-1)^h}{h}\left(\sum_{d|h}(-1)^d d\right) \tag{4-4}$$

and

$$E(x,h) \ll_\epsilon x^{145/196+\epsilon}$$

uniformly for $h \le x$. This suggests that we may replace $r(n)r(n+h)$ in (4-3) by $c(h)$. We shall ignore the error terms; in a rigorous analysis, these would swamp our expected main terms. However, it is plausible to assume that the error terms for various h are independent and largely cancel one another. Supposing this to be the case and replacing the sum over n by an integral, we find that

$$\mathrm{I_O} \approx \frac{1}{\pi^2}\int_0^N\left(\sum_{h=1}^{\infty} c(h)\mathrm{i}_X(h/(2\sqrt{u}))\right)\frac{du}{u^{3/2}}.$$

From the definition of $c(h)$ in (4-4) we see that

$$\begin{aligned}\mathrm{I_O} &\approx \frac{8}{\pi^2}\int_0^N\left(\sum_{h=1}^{\infty}\frac{(-1)^h}{h}\left(\sum_{d|h}(-1)^d d\right)\mathrm{i}_X(h/(2\sqrt{u}))\right)\frac{du}{u^{3/2}}\\ &= \frac{8}{\pi^2}\int_0^N\sum_{k=1}^{\infty}\sum_{d=1}^{\infty}\frac{(-1)^{d(k+1)}}{k}\mathrm{i}_X(dk/(2\sqrt{u}))\frac{du}{u^{3/2}}.\end{aligned} \tag{4-5}$$

By (3-7) the sum over d is

$$\begin{aligned}&\sum_{d=1}^{\infty}(-1)^{d(k+1)}\mathrm{i}_X(dk/(2\sqrt{u}))\\ &\quad= \frac{2X\sqrt{u}}{k}\sum_{d=1}^{\infty}(-1)^{d(k+1)}\frac{\sin(\pi dk\sqrt{X/u})}{\pi d} + \frac{4u\sqrt{X}}{k^2}\sum_{d=1}^{\infty}(-1)^{d(k+1)}\frac{\cos(\pi dk\sqrt{X/u})}{\pi^2 d^2}\\ &\qquad - \frac{4u^{3/2}}{k^3}\sum_{d=1}^{\infty}(-1)^{d(k+1)}\frac{\sin(\pi dk\sqrt{X/u})}{\pi^3 d^3}.\end{aligned}$$

Using (A-7) of the Appendix to express these sums in terms of Bernoulli polynomials, we find that if k is odd, the right-hand side equals

$$-\frac{2X\sqrt{u}}{k}B_1(\{k\sqrt{X/4u}\}) + \frac{4u\sqrt{X}}{k^2}B_2(\{k\sqrt{X/4u}\}) - \frac{8u^{3/2}}{3k^3}B_3(\{k\sqrt{X/4u}\}),$$

and if k is even, it equals

$$-\frac{2X\sqrt{u}}{k}B_1\big(\{k\sqrt{X/4u}+\tfrac12\}\big)+\frac{4u\sqrt{X}}{k^2}B_2\big(\{k\sqrt{X/4u}+\tfrac12\}\big)-\frac{8u^{3/2}}{3k^3}B_3\big(\{k\sqrt{X/4u}+\tfrac12\}\big).$$

Inserting these into (4-5), we obtain

$$\begin{aligned}\mathrm{I_O}&\approx\frac{8}{\pi^2}\int_0^N\sum_{\substack{k=1\\k\text{ odd}}}^{\infty}\Big(-\frac{2X\sqrt{u}}{k^2}B_1(\{k\sqrt{X/4u}\})+\frac{4u\sqrt{X}}{k^3}B_2(\{k\sqrt{X/4u}\})-\frac{8u^{3/2}}{3k^4}B_3(\{k\sqrt{X/4u}\})\Big)\frac{du}{u^{3/2}}\\&\quad+\frac{8}{\pi^2}\int_0^N\sum_{\substack{k=1\\k\text{ even}}}^{\infty}\Big(-\frac{2X\sqrt{u}}{k^2}B_1(\{k\sqrt{X/4u}+\tfrac12\})+\frac{4u\sqrt{X}}{k^3}B_2\big(\{k\sqrt{X/4u}+\tfrac12\}\big)\\&\qquad\qquad-\frac{8u^{3/2}}{3k^4}B_3\big(\{k\sqrt{X/4u}+\tfrac12\}\big)\Big)\frac{du}{u^{3/2}}\\&=\mathrm{I_{O,odd}}+\mathrm{I_{O,even}}.\end{aligned}\tag{4-6}$$

In both integrals we make the substitution $x=k\sqrt{X/4u}$ so that $2\sqrt{u}/k=\sqrt{X}/x$ and $du/u^{3/2}=-4\,dx/(k\sqrt{X})$. We then find that

$$\mathrm{I_{O,odd}}=\frac{32X}{\pi^2}\sum_{\substack{k=1\\k\text{ odd}}}^{\infty}\frac{1}{k^2}\int_{k\sqrt{X/4N}}^{\infty}\Big(-\frac{B_1(\{x\})}{x}+\frac{B_2(\{x\})}{x^2}-\frac{B_3(\{x\})}{3x^3}\Big)\,dx,$$

$$I_{\mathrm{O,even}}=\frac{32X}{\pi^2}\sum_{\substack{k=1\\k\text{ even}}}^{\infty}\frac{1}{k^2}\int_{k\sqrt{X/4N}}^{\infty}\Big(-\frac{B_1(\{x+\frac12\})}{x}+\frac{B_2(\{x+\frac12\})}{x^2}-\frac{B_3(\{x+\frac12\})}{3x^3}\Big)\,dx.$$

4.1. ***Calculation of $I_{\mathrm{O,odd}}$.*** We have

$$\mathrm{I_{O,odd}}=\frac{32X}{\pi^2}\sum_{\substack{k=1\\k\text{ odd}}}^{\infty}\frac{1}{k^2}\int_{k\sqrt{X/4N}}^{\infty}\Big(-\frac{B_1(\{x\})}{x}+\frac{B_2(\{x\})}{x^2}-\frac{B_3(\{x\})}{3x^3}\Big)\,dx.$$

We split the sum over k into two sums according to whether $k\le 2\sqrt{N/X}$ or $k>2\sqrt{N/X}$. For terms in the second sum we have

$$\int_{k\sqrt{X/4N}}^{\infty}\frac{B_j(\{x\})}{x^j}\,dx\ll(k\sqrt{X/4N})^{-j}.$$

Hence, the contribution from the second sum is

$$\ll X\sum_{j=1}^{3}\sum_{k>2\sqrt{N/X}}\frac{1}{k^{j+2}}(N/X)^{j/2}\ll X^{3/2}N^{-1/2}.$$

Thus

$$\mathrm{I}_{O,\mathrm{odd}}=\frac{32X}{\pi^2}\sum_{\substack{k\le2\sqrt{N/X}\\k\text{ odd}}}\frac{1}{k^2}\int_{k\sqrt{X/4N}}^{\infty}\Big(-\frac{B_1(\{x\})}{x}+\frac{B_2(\{x\})}{x^2}-\frac{B_3(\{x\})}{3x^3}\Big)\,dx+O(X^{3/2}N^{-1/2}).$$

In the remaining sum $k\sqrt{X/4N} \le 1$ and we split the integral into two parts, one over $[k\sqrt{X/4N}, 1]$ and the other over $[1, \infty)$. By (A-6) of the Appendix the first integral is

$$\int_{k\sqrt{X/4N}}^{1} \left(\frac{\frac{1}{2} - x}{x} + \frac{x^2 - x + \frac{1}{6}}{x^2} - \frac{x^3 - \frac{3}{2}x^2 + \frac{1}{2}x}{3x^3} \right) dx = -\tfrac{1}{3} + O(k\sqrt{X/N}).$$

By Lemma 8

$$\int_1^{\infty} \left(-\frac{B_1(\{x\})}{x} + \frac{B_2(\{x\})}{x^2} - \frac{B_3(\{x\})}{3x^3} \right) dx = -\left(\tfrac{1}{2}\log 2\pi - 1\right) + \left(\log 2\pi - \tfrac{11}{6}\right) - \tfrac{1}{3}\left(\tfrac{3}{2}\log 2\pi - \tfrac{11}{4}\right) = \tfrac{1}{12}.$$

Hence

$$\mathrm{I}_{\mathrm{O,odd}} = \frac{32X}{\pi^2} \sum_{\substack{k \le 2\sqrt{N/X} \\ k \text{ odd}}} \frac{1}{k^2}\left(-\tfrac{1}{4} + O(k\sqrt{X/N})\right) + O(X^{3/2}N^{-1/2}).$$

The contribution of the O-term in the sum is $O(X^{3/2}N^{-1/2}\log(N/X))$. Hence

$$\begin{aligned} \mathrm{I}_{\mathrm{O,odd}} &= -\frac{8X}{\pi^2} \sum_{\substack{k \le 2\sqrt{N/X} \\ k \text{ odd}}} \frac{1}{k^2} + O(X^{3/2}N^{-1/2}\log(N/X)) \\ &= -\frac{8X}{\pi^2} \sum_{\substack{k=1 \\ k \text{ odd}}}^{\infty} \frac{1}{k^2} + O(X^{3/2}N^{-1/2}\log(N/X)). \end{aligned} \tag{4-7}$$

4.2. ***Calculation of*** $\mathbf{I_{O,even}}$**.** The treatment of $\mathrm{I}_{\mathrm{O,even}}$ is similar to that of $\mathrm{I}_{\mathrm{O,odd}}$ so we will skip some of the details. We have

$$\mathrm{I}_{\mathrm{O,even}} = \frac{32X}{\pi^2} \sum_{\substack{k=1 \\ k \text{ even}}}^{\infty} \frac{1}{k^2} \int_{k\sqrt{X/4N}}^{\infty} \left(-\frac{B_1(\{x+\frac{1}{2}\})}{x} + \frac{B_2(\{x+\frac{1}{2}\})}{x^2} - \frac{B_3(\{x+\frac{1}{2}\})}{3x^3} \right) dx.$$

Our first step is to split the sum over k according to whether $k \le \sqrt{N/X}$ or $k > \sqrt{N/X}$ (note that the division for odd k was at $2\sqrt{N/X}$). As before, the total contribution from the tail is $X^{3/2}N^{-1/2}$. Thus,

$$I_{\mathrm{O,even}} = \frac{32X}{\pi^2} \sum_{\substack{k \le \sqrt{N/X} \\ k \text{ even}}} \frac{1}{k^2} \int_{k\sqrt{X/4N}}^{\infty} \left(-\frac{B_1(\{x+\frac{1}{2}\})}{x} + \frac{B_2(\{x+\frac{1}{2}\})}{x^2} - \frac{B_3(\{x+\frac{1}{2}\})}{3x^3} \right) dx + O(X^{3/2}N^{-1/2}).$$

Observe that for each k in the sum we have $k\sqrt{X/4N} \le \frac{1}{2}$. We may therefore split the integral over the intervals $\left[k\sqrt{X/4N}, \frac{1}{2}\right]$ and $\left[\frac{1}{2}, \infty\right)$. By (A-6) the first integral equals

$$\begin{aligned} &\int_{k\sqrt{X/4N}}^{1/2} \left(-\frac{x}{x} + \frac{\left(x+\frac{1}{2}\right)^2 - \left(x+\frac{1}{2}\right) + \frac{1}{6}}{x^2} - \frac{\left(x+\frac{1}{2}\right)^3 - \frac{3}{2}\left(x+\frac{1}{2}\right)^2 + \frac{1}{2}\left(x+\frac{1}{2}\right)}{3x^3} \right) dx \\ &\qquad = -\int_{k\sqrt{X/4N}}^{1/2} \tfrac{1}{3}\, dx = -\tfrac{1}{6} + O(k\sqrt{X/N}). \end{aligned}$$

By Lemma 8

$$\int_{1/2}^{\infty}\Big(-\frac{B_1(\{x+\frac12\})}{x}+\frac{B_2(\{x+\frac12\})}{x^2}-\frac{B_3(\{x+\frac12\})}{3x^3}\Big)\,dx$$
$$=-\big(-\tfrac12+\tfrac12\log 2\big)+\big(-\tfrac23+\log 2\big)-\tfrac13\big(-1+\tfrac32\log 2\big)=\tfrac16.$$

Hence,

$$\mathrm{I}_{\mathrm{O,even}}=\frac{32X}{\pi^2}\sum_{\substack{k\le\sqrt{N/X}\\ k\text{ even}}}\frac{1}{k^2}\big(-\tfrac16+\tfrac16+O(k\sqrt{X/N})\big)+O(X^{3/2}N^{-1/2})$$
$$\ll X^{3/2}N^{-1/2}\log(N/X). \tag{4-8}$$

4.3. ***Completion of the calculation of*** $\mathrm{I}(X)$**.** By (4-6), (4-7), and (4-8), we see that

$$\mathrm{I}_{\mathrm{O}}=-\frac{8X}{\pi^2}\sum_{\substack{k=1\\ k\text{ odd}}}^{\infty}\frac{1}{k^2}+O(X^{3/2}N^{-1/2}\log(N/X)).$$

Combining this with (4-1) and (4-2), we see that

$$\mathrm{I}(X)=CX^{3/2}-\frac{8X}{\pi^2}\sum_{\substack{k=1\\ k\text{ odd}}}^{\infty}\frac{1}{k^2}+O(X^{3/2}N^{-1/2}\log(N/X)),$$

where

$$C=\frac{1}{3\pi^2}\sum_{n=1}^{\infty}\frac{r(n)^2}{n^{3/2}}.$$

It is easy to see that

$$\sum_{\substack{k=1\\ k\text{ odd}}}^{\infty}\frac{1}{k^2}=\frac34\zeta(2)=\frac{\pi^2}{8}.$$

Using this and the assumption that $X^4\le N\le X^A$, we find that

$$\mathrm{I}(X)=CX^{3/2}-X+O(X^{1/2}\log X). \tag{4-9}$$

5. Calculation of $\mathrm{J}(X)$

Our treatment of $\mathrm{J}(X)$ is easier. We have

$$\mathrm{J}(X)=\frac{1}{2\pi^2}\sum_{m,n\le N}\frac{r(n)r(m)}{(mn)^{3/4}}\mathrm{j}_X(\sqrt m+\sqrt n),$$

where

$$\mathrm{j}_X(y)=\frac{X\cos(2\pi y\sqrt X)}{\pi y}-\frac{\sqrt X\sin(2\pi y\sqrt X)}{\pi^2y^2}+\frac{1-\cos(2\pi y\sqrt X)}{2\pi^3y^3}.$$

The second and third terms of j_X contribute

$$\ll\sum_{m,n\le N}\frac{r(n)r(m)}{(mn)^{3/4}}\bigg(\frac{\sqrt X}{(\sqrt m+\sqrt n)^2}+\frac{1}{(\sqrt m+\sqrt n)^3}\bigg).$$

Since $(\sqrt{m}+\sqrt{n})^2 > 2\sqrt{mn}$, this is

$$\ll \sqrt{X} \sum_{m,n\le N} \frac{r(n)r(m)}{(mn)^{5/4}} + \sum_{m,n\le N} \frac{r(n)r(m)}{(mn)^{3/2}} \ll \sqrt{X}.$$

Hence, recalling (1-7),

$$\begin{aligned} \mathrm{J}(X) &= \frac{X}{2\pi^3} \sum_{m,n\le N} \frac{r(n)r(m)}{(mn)^{3/4}} \frac{\cos(2\pi(\sqrt{m}+\sqrt{n})\sqrt{X})}{(\sqrt{m}+\sqrt{n})} + O(X^{1/2}) \\ &= XC_N(X) + O(X^{1/2}). \end{aligned} \tag{5-1}$$

6. Estimation of $\mathrm{K}(X)$

We have

$$\mathrm{K}(X) = \frac{3}{32\pi^3} \sum_{m\ne n\le N} \frac{r(n)r(m)}{n^{3/4}m^{5/4}} \mathrm{k}_X(\sqrt{n}-\sqrt{m}),$$

where

$$\mathrm{k}_X(y) = \frac{\sqrt{X}\cos(2\pi y\sqrt{X})}{\pi y} - \frac{\sin(2\pi y\sqrt{X})}{2\pi^2 y^2}.$$

Thus

$$\mathrm{K}(X) \ll \sum_{m\ne n\le N} \frac{r(n)r(m)}{n^{3/4}m^{5/4}} \left(\frac{\sqrt{X}}{|\sqrt{n}-\sqrt{m}|} + \frac{1}{(\sqrt{n}-\sqrt{m})^2} \right).$$

We split the sum over m and n according to whether $m < n$ or $m > n$ so that

$$\mathrm{K}(X) \ll \left(\sum_{n<N} \sum_{n<m\le N} + \sum_{m<N} \sum_{m<n\le N} \right) \cdots = K_1 + K_2.$$

In K_1 we write $m = n+h$ and further split the sum over h as

$$\begin{aligned} K_1 &= \sum_{n<N} \left(\sum_{1\le h\le n/2} + \sum_{n/2<h\le N-n} \right) \frac{r(n)r(n+h)}{n^{3/4}(n+h)^{5/4}} \left(\frac{\sqrt{X}}{|\sqrt{n+h}-\sqrt{n}|} + \frac{1}{(\sqrt{n+h}-\sqrt{n})^2} \right) \\ &= K_{11} + K_{12}. \end{aligned}$$

In K_{11} we use $\sqrt{n+h}-\sqrt{n} \gg h/\sqrt{n}$ and $r(n) \ll n^{\epsilon}$ and find for ϵ small enough that

$$\begin{aligned} K_{11} &\ll \sum_{n<N} \sum_{1\le h\le n/2} \frac{r(n)r(n+h)}{n^{3/4}(n+h)^{5/4}} \left(\frac{\sqrt{nX}}{h} + \frac{n}{h^2} \right) \\ &\ll \sqrt{X} \sum_{n<N} \frac{1}{n^{3/2-2\epsilon}} \sum_{1\le h\le n/2} \frac{1}{h} + \sum_{n<N} \frac{1}{n^{1-2\epsilon}} \sum_{1\le h\le n/2} \frac{1}{h^2} \\ &\ll \sqrt{X} + N^{2\epsilon} \ll \sqrt{X}. \end{aligned}$$

In K_{12} we use the inequalities $\sqrt{n+h}-\sqrt{n} \gg \sqrt{h}$, $(n+h)^{5/4} \gg h^{5/4}$ and $r(n) \ll n^{\epsilon}$ and find that

$$\begin{aligned} K_{12} &\ll \sum_{n<N} \sum_{n/2<h\le N-n} \frac{1}{n^{3/4-\epsilon}h^{5/4-\epsilon}}\left(\sqrt{\frac{X}{h}}+\frac{1}{h}\right) \\ &\ll \sqrt{X}\sum_{n<N}\frac{1}{n^{3/4-\epsilon}} \sum_{n/2<h\le N-n}\frac{1}{h^{7/4-\epsilon}} + \sum_{n<N}\frac{1}{n^{3/4-\epsilon}} \sum_{n/2<h\le N-n}\frac{1}{h^{9/4-\epsilon}} \\ &\ll \sqrt{X}\sum_{n<N}\frac{1}{n^{3/2-2\epsilon}} + \sum_{n<N}\frac{1}{n^{2-2\epsilon}} \ll \sqrt{X}. \end{aligned}$$

Hence, $K_1 = K_{11} + K_{12} = O(\sqrt{X})$.

We treat K_2 in the same way and find that it is also $O(\sqrt{X})$. Thus

$$\mathrm{K}(X) \ll \sqrt{X}.$$

7. Estimation of L(X)

We treat $\mathrm{L}(X)$ as we did the last two terms in $\mathrm{J}(X)$. We have

$$\mathrm{L}(X) = \frac{3}{32\pi^3}\sum_{m,n\le N}\frac{r(n)r(m)}{n^{3/4}m^{5/4}}\,1_X(\sqrt{n}+\sqrt{m}),$$

where

$$1_X = \frac{\sqrt{X}\sin(2\pi y\sqrt{X})}{\pi y} + \frac{\cos(2\pi y\sqrt{X})-1}{2\pi^2 y^2}.$$

Using the inequality $(\sqrt{m}+\sqrt{n}) > 2\sqrt{mn}$, we find that

$$\begin{aligned} \mathrm{L}(X) &\ll \sum_{m,n\le N}\frac{r(n)r(m)}{n^{3/4}m^{5/4}}\left(\frac{\sqrt{X}}{(mn)^{1/4}}+\frac{1}{(mn)^{1/2}}\right) \\ &\ll \sqrt{X}\sum_{n\le N}\frac{r(n)}{n}\sum_{m\le N}\frac{r(m)}{m^{3/2}} + 1 \\ &\ll \sqrt{X}\log X. \end{aligned}$$

8. Completion of the argument for the Conjecture

By (3-6), (4-9), and (5-1)

$$\frac{1}{\pi^2}\int_0^X A_1(x)^2\,dx = \mathrm{I}(X)+\mathrm{J}(X) = CX^{3/2} - X + C_N(X) + O(X^{1/2+\epsilon}).$$

By (3-8) and the estimates of the last two sections

$$\frac{1}{\pi^2}\int_0^X A_1(x)A_2(x)\,dx = K(X)+L(X) \ll \sqrt{X}\log X.$$

It now follows from (3-5) that

$$\mathcal{I}(X) = \int_{X/2}^{X} P(x)^2\,dx = C(X^{3/2} - (X/2)^{3/2}) - X/2 + (C_N(X) - C_N(X/2)) + O(X^{1/2+\epsilon}),$$

where

$$C = \frac{1}{3\pi^2}\sum_{n=1}^{\infty} r^2(n) n^{-3/2}.$$

We add this result (with the same value of $N \in [X^4, X^A]$) for the intervals $(X/2, X]$, $(X/4, X/2], \ldots,$ $(X/2^r, X/2^{r-1}]$, where $r = [3\log X/4\log 2]$. Then

$$X^{1/4} \le \frac{X}{2^r} < 2X^{1/4}.$$

Now $(X^{1/4})^{3/2} = X^{3/8} \ll X^{1/2+\epsilon}$, so

$$X/2^r C_N(X/2^r) \ll X/2^r \log N \ll X^{1/2+\epsilon}.$$

Hence

$$\int_{X/2^r}^{X} P(x)^2\,dx = CX^{3/2} - X + C_N(X) + O(X^{1/2+\epsilon}). \tag{8-1}$$

Finally,

$$\int_0^{X/2^r} P(x)^2 \ll (X/2^r)^{3/2} \ll X^{3/8},$$

so the integral on the left-hand side of (8-1) may be extended over the entire interval $[0, X]$. This completes the argument for the Conjecture.

9. Proof of Theorem 2

For $N \ge 1$ and $x \in \mathbb{R}$ we have

$$C_N(x^2) = \frac{1}{2\pi^3}\sum_{m,n\le N} \frac{r(n)r(m)\cos(2\pi(\sqrt{m}+\sqrt{n})x)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})}.$$

Clearly

$$\max_{x\in\mathbb{R}} |C_N(x^2)| = C_N(0) = \frac{1}{2\pi^3}\sum_{m,n\le N} \frac{r(n)r(m)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})}. \tag{9-1}$$

Lemma 6. *For $N \ge 2$ we have*

$$\frac{2}{\pi}\log N + O(1) \le C_N(0) \le \frac{4}{\pi}\log N + O(1).$$

Proof. Let

$$\sum_{n\le x} r(n) = \pi x + E(x),$$

where $E(x) \ll x^{1/3}$. By Riemann–Stieltjes integration, for $y < z$ we have

$$\sum_{y<n\le z} \frac{r(n)}{n^\sigma} = \pi\frac{z^{1-\sigma} - y^{1-\sigma}}{1-\sigma} + \frac{E(u)}{u^\sigma}\Big|_y^z + \sigma\int_y^z \frac{E(u)}{u^{1+\sigma}}\,du.$$

Here the case $\sigma = 1$ is interpreted as a limit. In particular,

$$\sum_{n\le z}\frac{r(n)}{n} = \pi\log z + O(1) \tag{9-2}$$

and

$$\sum_{y<n\le z}\frac{r(n)}{n^{\sigma}} = \pi\frac{z^{1-\sigma}-y^{1-\sigma}}{1-\sigma} + O(\max(y^{1/3-\sigma}, z^{1/3-\sigma})). \tag{9-3}$$

Using this and the symmetry in m and n in the double sum defining $C_N(x^2)$, we find that

$$C_N(0) = \frac{1}{\pi^3}\sum_{1\le n<N}\frac{r(n)}{n^{3/4}}\Bigg(\sum_{n<m\le N}\frac{r(m)}{m^{3/4}(\sqrt{m}+\sqrt{n})}\Bigg) + O(1).$$

Now

$$\frac{1}{2}\sum_{n<m\le N}\frac{r(m)}{m^{5/4}} \le \sum_{n<m\le N}\frac{r(m)}{m^{3/4}(\sqrt{m}+\sqrt{n})} \le \sum_{n<m\le N}\frac{r(m)}{m^{5/4}},$$

so by (9-3) we have

$$2\pi n^{-1/4} + O(n^{-11/12}) \le \sum_{n<m\le N}\frac{r(m)}{m^{3/4}(\sqrt{m}+\sqrt{n})} \le 4\pi n^{-1/4} + O(n^{-11/12}).$$

Thus,

$$C_N(0) \le \frac{1}{\pi^3}\sum_{1\le n<N}\frac{r(n)}{n^{3/4}}(4\pi n^{-1/4} + O(n^{-11/12})) = \frac{4}{\pi}\log N + O(1).$$

Similarly,

$$C_N(0) \ge \frac{2}{\pi}\log N + O(1).$$

This completes the proof of the lemma. □

To prove Theorem 2, we note that by (1-8)

$$|C(X)| = |C_N(X)| + O\Bigg(\frac{X\log N}{N^{1/4}}\Bigg).$$

Thus, by (9-1) and Lemma 6

$$|C(X)| \le \frac{4}{\pi}\log N + O\Bigg(\frac{X\log N}{N^{1/4}}\Bigg) + O(1).$$

Taking $N = X^4\log X$, say, we obtain

$$|C(X)| \le \Big(\frac{16}{\pi} + o(1)\Big)\log X.$$

This proves the first assertion of Theorem 2. The second follows from this and the Conjecture.

10. Proof of Theorem 3

We base the proof of Theorem 3 on a variant of a lemma of [Soundararajan 2003].

For each $\boldsymbol{n} = (m, n) \in \mathbb{Z}^2$ let $a_{\boldsymbol{n}}$ and $\lambda_{\boldsymbol{n}}$ be nonnegative real numbers with the $\lambda_{\boldsymbol{n}}$ arranged in nondecreasing order. Assume that $\sum_{\boldsymbol{n}} a_{\boldsymbol{n}} < \infty$ and set

$$F(x) = \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} \cos(2\pi\lambda_{\boldsymbol{n}} x).$$

Lemma 7. *Let $L \geq 2$ be an integer and let $\Lambda \geq 2$ be a real number. Let $\mathcal{M}$ be a subset of the double indices $\boldsymbol{n}$ for which $\lambda_{\boldsymbol{n}} \leq \Lambda/2$, and let M be the cardinality of $\mathcal{M}$. Then for any real number $Y \geq 2$ there exists an x such that $Y/2 \leq x \leq (6L)^{M+1}Y$ and*

$$F(x) \geq \frac{1}{8} \sum_{\boldsymbol{n}\in\mathcal{M}} a_{\boldsymbol{n}} - \frac{1}{L-1} \sum_{\substack{\boldsymbol{n} \\ \lambda_{\boldsymbol{n}}\leq\Lambda}} a_{\boldsymbol{n}} - \frac{4}{\pi^2\Lambda Y} \sum_{\boldsymbol{n}} a_{\boldsymbol{n}}. \tag{10-1}$$

Proof. Let $K(u) = (\sin \pi u/\pi u)^2$ be Fejér's kernel and let $k(y) = \max(0, 1-|y|)$ be its Fourier transform. Then the Fourier transform of $\Lambda K(\Lambda u)$ is $k(y/\Lambda) = \max(0, 1-|y|/\Lambda)$. Consider the integral

$$\begin{aligned}
\int_{-\infty}^{\infty} \Lambda K(\Lambda u) F(u+x)\, du &= \frac{1}{2} \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} \int_{-\infty}^{\infty} \Lambda K(\Lambda u)(e^{2\pi i\lambda_{\boldsymbol{n}}(x+u)} + e^{-2\pi i\lambda_{\boldsymbol{n}}(x+u)})\, du \\
&= \frac{1}{2} \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} e^{2\pi i\lambda_{\boldsymbol{n}} x} k(-\lambda_{\boldsymbol{n}}/\Lambda) + \frac{1}{2} \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} e^{-2\pi i\lambda_{\boldsymbol{n}} x} k(\lambda_{\boldsymbol{n}}/\Lambda) \\
&= \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} \cos(2\pi\lambda_{\boldsymbol{n}} x) k(\lambda_{\boldsymbol{n}}/\Lambda),
\end{aligned}$$

where the last equality holds because $k(-y) = k(y)$. Defining

$$G(x) = \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} \cos(2\pi\lambda_{\boldsymbol{n}} x) k(\lambda_{\boldsymbol{n}}/\Lambda),$$

we may write this as

$$\int_{-\infty}^{\infty} \Lambda K(\Lambda u) F(u+x)\, du = G(x).$$

Since F, G, and K are real-valued functions, and K and Λ are nonnegative, we find next that

$$\begin{aligned}
G(x) &= \int_{-Y/2}^{Y/2} \Lambda K(\Lambda u) F(u+x)\, du + \int_{|u|>Y/2} \Lambda K(\Lambda u) F(u+x)\, du \\
&\leq \max_{|u|\leq Y/2} F(u+x) \int_{-\infty}^{\infty} \Lambda K(\Lambda u)\, du + \int_{|u|>Y/2} \frac{1}{\pi^2\Lambda u^2} |F(u+x)|\, du \\
&\leq \max_{|u|\leq Y/2} F(u+x) + \frac{4}{\pi^2\Lambda Y} \left(\sum_{\boldsymbol{n}} a_{\boldsymbol{n}} \right).
\end{aligned} \tag{10-2}$$

Now let $\mathcal{M}$ be a subset of indices $\boldsymbol{n}$ with $\lambda_{\boldsymbol{n}} \leq \Lambda/2$, and let M be its cardinality. By Dirichlet's theorem there is an x_0 with $Y \leq x_0 \leq (6L)^M Y$ such that $\|x_0\lambda_{\boldsymbol{n}}\| \leq 1/(6L)$ for each $\boldsymbol{n} \in \mathcal{M}$. Consider

$$\sum_{|l|\leq L} G(x_0 l) k(l/L) = \frac{1}{2} \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} k(\lambda_{\boldsymbol{n}}/\Lambda) \sum_{|l|\leq L} k(l/L) \cos(2\pi\lambda_{\boldsymbol{n}} x_0 l).$$

The sum over l equals

$$\frac{1}{L}\left(\frac{\sin(\pi L\lambda_{\boldsymbol{n}}x_0)}{\sin(\pi\lambda_{\boldsymbol{n}}x_0)}\right)^2,$$

which is nonnegative. We may therefore drop any terms we wish to from the sum over $\boldsymbol{n}$ to obtain a lower bound. Moreover, for each $\boldsymbol{n}\in\mathcal{M}$, $\cos(2\pi\lambda_{\boldsymbol{n}}x_0l)\ge\cos(2\pi l/(6L))\ge\cos(\pi/3)\ge\frac{1}{2}$, so

$$\sum_{|l|\le L}k(l/L)\cos(2\pi\lambda_{\boldsymbol{n}}x_0l)\ge\frac{1}{2}\sum_{|l|\le L}k(l/L)=\frac{L}{2}.$$

Thus

$$\sum_{|l|\le L}G(x_0l)k(l/L)\ge\frac{L}{4}\sum_{\boldsymbol{n}\in\mathcal{M}}a_{\boldsymbol{n}}k(\lambda_{\boldsymbol{n}}/\Lambda).$$

Since $G(-x_0l)=G(x_0l)$, there is an $1\le l_0\le L$ such that

$$G(0)+2G(x_0l_0)\sum_{1\le l\le L}k(l/L)\ge\frac{L}{4}\sum_{\boldsymbol{n}\in\mathcal{M}}a_{\boldsymbol{n}}k(\lambda_{\boldsymbol{n}}/\Lambda).$$

The sum over l equals $(L-1)/2$, so we see that for this l_0

$$G(x_0l_0)\ge\frac{1}{4}\sum_{\boldsymbol{n}\in\mathcal{M}}a_{\boldsymbol{n}}k(\lambda_{\boldsymbol{n}}/\Lambda)-\frac{1}{L-1}\sum_{\substack{\boldsymbol{n}\\ \lambda_{\boldsymbol{n}}\le\Lambda}}a_{\boldsymbol{n}}.$$

From this and (10-2) we obtain

$$\max_{|u|\le Y/2}F(u+x_0l_0)\ge\frac{1}{4}\sum_{\boldsymbol{n}\in\mathcal{M}}a_{\boldsymbol{n}}k(\lambda_{\boldsymbol{n}}/\Lambda)-\frac{1}{L-1}\sum_{\substack{\boldsymbol{n}\\ \lambda_{\boldsymbol{n}}\le\Lambda}}a_{\boldsymbol{n}}-\frac{4}{\pi^2\Lambda Y}\sum_{\boldsymbol{n}}a_{\boldsymbol{n}}.$$

Now $\mathcal{M}\subset[0,\Lambda/2]$, so $k(\lambda_{\boldsymbol{n}}/\Lambda)\ge\frac{1}{2}$. Furthermore, for $|u|\le Y/2$ we have

$$x_0l_0+u\ge Y-Y/2=Y/2,$$
$$x_0l_0+u\le(6L)^MYL+X/2\le(6L)^{M+1}Y.$$

Thus, there is an x with $Y/2\le x\le(6L)^{M+1}Y$ such that

$$F(x)\ge\frac{1}{8}\sum_{\boldsymbol{n}\in\mathscr{M}}a_{\boldsymbol{n}}-\frac{1}{L-1}\sum_{\substack{\boldsymbol{n}\\ \lambda_{\boldsymbol{n}}\le\Lambda}}a_{\boldsymbol{n}}-\frac{4}{\pi^2\Lambda Y}\sum_{\boldsymbol{n}}a_{\boldsymbol{n}}.$$

This completes the proof of Lemma 7. □

We now prove Theorem 3. Let

$$C_N(X)=\frac{1}{2\pi^3}\sum_{m,n\le N}\frac{r(n)r(m)\cos(2\pi(\sqrt{m}+\sqrt{n})\sqrt{X})}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})}$$

with $X^4\le N\le X^A$. We replace X by x^2 and will first show that for all large Z, there exists an x with $Z^{1/2}\le x\le Z^{3/2}$ such that

$$C_N(x^2)\ge\left(\frac{8}{\pi}+o(1)\right)\log\log x. \tag{10-3}$$

Since $Z\le X\le Z^3$, we also need $N\ge(Z^3)^4=Z^{12}$. We take $N=[Z^{12}(\log Z)^4]$.

Write

$$C_N(x^2) = \frac{1}{2\pi^3} F(x) = \frac{1}{2\pi^3} \sum_{\boldsymbol{n}} a_{\boldsymbol{n}} \cos(2\pi \lambda_{\boldsymbol{n}} x),$$

with $\lambda_{\boldsymbol{n}} = \lambda_{m,n} = \sqrt{m} + \sqrt{n},\ m, n \le N$, and

$$a_{\boldsymbol{n}} = a_{m,n} = \frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})}.$$

We apply Lemma 7 to $F(x)$ with

$$\mathcal{M} = \{(m,n) : \lambda_{m,n} \le \Lambda/2\}$$

and Λ to be determined later. Observe that $M = |\mathcal{M}| \ll \Lambda^4$. If $L, Y \ge 2$ with L an integer, then by Lemma 7 there is an x with $Y/2 \le x \le (6L)^{M+1}Y$ such that

$$\begin{aligned}\sum_{m,n\le N} &\frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})} \cos(2\pi(\sqrt{m}+\sqrt{n})x) \\ &\ge \frac{1}{8} \sum_{\boldsymbol{n}\in\mathcal{M}} \frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})} - \frac{1}{L-1} \sum_{\substack{\boldsymbol{n} \\ \lambda_{m,n}\le\Lambda}} \frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})} \\ &\qquad - \frac{4}{\pi^2 \Lambda Y} \sum_{\boldsymbol{n}} \frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})}. \qquad (10\text{-}4)\end{aligned}$$

By arguments similar to those in the proof of Lemma 6 we have

$$\begin{aligned}\sum_{\substack{\boldsymbol{n} \\ \lambda_{m,n}\le\Lambda}} \frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})} &\le 2 \sum_{\sqrt{n}\le\Lambda-1} \frac{r(n)}{n^{3/4}} \sum_{\sqrt{n}<\sqrt{m}\le\Lambda-\sqrt{n}} \frac{r(m)}{m^{5/4}} + O(1) \\ &\le 2 \sum_{n\le(\Lambda-1)^2} \frac{r(n)}{n^{3/4}} \sum_{n<m\le(\Lambda-\sqrt{n})^2} \frac{r(m)}{m^{5/4}} + O(1) \\ &\le 8\pi \sum_{n\le(\Lambda-1)^2} \frac{r(n)}{n} + O(1) \\ &\le 16\pi^2 \log\Lambda + O(1).\end{aligned}$$

As in the last section, half of this, namely $8\pi^2 \log\Lambda + O(1)$, is a lower bound. Applying similar reasoning to all three sums in (10-4), we find that there is an x such that $Y/2 \le x \le (6L)^{M+1}Y$ for which

$$\sum_{m,n\le N} \frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})} \cos(2\pi(\sqrt{m}+\sqrt{n})x) \ge \pi^2 \log\Lambda - O\left(\frac{\log\Lambda}{L}\right) - O\left(\frac{\log Y}{\Lambda Y}\right).$$

We now choose $Y = 2Z^{1/2},\ L = [\log Z]$. Then

$$(6L)^{M+1} \le \exp(O(\Lambda^4 \log\log Z)).$$

We need this to be at most $Z/2$, and it will be if we take $\Lambda = (\log Z)^{1/4}/(\log\log Z)^2$ with Z sufficiently large. With these choices of the parameters, we see that there exists an x with $Z^{1/2} \le x \le Z^{3/2}$ such that

$$C_N(x^2) = \frac{1}{2\pi^3} \sum_{m,n\le N} \frac{r(m)r(n)}{(mn)^{3/4}(\sqrt{m}+\sqrt{n})} \cos(2\pi(\sqrt{m}+\sqrt{n})x) \ge (1+o(1))\frac{1}{2\pi}\log\log Z.$$

Recalling that $X = x^2$, we find that there is an $X \in [Z, Z^3]$ such that

$$C_N(X) \ge (1+o(1))\frac{1}{2\pi}\log\log X.$$

Since $N = [Z^{12}(\log Z)^4] \gg X^{12}$ and $Z \ge X^{1/3}$, we see from (1-8) that

$$C(X) = C_N(X) + O\left(\frac{X\log N}{N^{1/4}}\right) = C_N(X) + O(1). \tag{10-5}$$

Theorem 3 now follows.

Appendix: Facts about Bernoulli polynomials

We collect the formulas we need about Bernoulli polynomials $B_k(x)$ here. Appendix B of [Montgomery and Vaughan 2007] is a convenient reference.

The first three Bernoulli polynomials are

$$B_1(x) = x - \tfrac{1}{2}, \quad B_2(x) = x^2 - x + \tfrac{1}{6}, \quad \text{and} \quad B_3(x) = x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x. \tag{A-6}$$

If we let $\{x\}$ denote the fractional part of the real number x and replace x by $\{x\}$ in B_j, the resulting functions are periodic with period 1. They therefore have Fourier series expansions, and these are given by

$$B_j(\{x\}) = -\frac{j!}{(2\pi i)^j} \sum_{\substack{n=-\infty \\ n\ne 0}}^{\infty} \frac{e^{2\pi i n x}}{n^j} \quad (j = 1, 2, 3).$$

These hold for all real x, except in the case of B_1 we need $x \notin \mathbb{Z}$. With this stipulation we immediately have

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\sin(2\pi n x)}{\pi n} &= -B_1(\{x\}) = \tfrac{1}{2} - \{x\}, \\
\sum_{n=1}^{\infty} \frac{\cos(2\pi n x)}{\pi^2 n^2} &= B_2(\{x\}) = \{x\}^2 - \{x\} + \tfrac{1}{6}, \\
\sum_{n=1}^{\infty} \frac{\sin(2\pi n x)}{\pi^3 n^3} &= \tfrac{2}{3}B_3(\{x\}) = \tfrac{2}{3}\{x\}^3 - \{x\}^2 + \tfrac{1}{3}\{x\}.
\end{aligned} \tag{A-7}$$

We collect the other formulas we need in:

Lemma 8. *We have*

$$\int_1^{\infty} \frac{B_1(\{x\})}{x}\,dx = \tfrac{1}{2}\log 2\pi - 1, \quad \int_1^{\infty} \frac{B_2(\{x\})}{x^2}\,dx = \log 2\pi - \tfrac{11}{6},$$
$$\int_1^{\infty} \frac{B_3(\{x\})}{x^3}\,dx = \tfrac{3}{2}\log 2\pi - \tfrac{11}{4},$$

and

$$\int_{1/2}^{\infty} \frac{B_1(\{x+\frac{1}{2}\})}{x}\,dx = -\tfrac{1}{2}+\tfrac{1}{2}\log 2, \qquad \int_{1/2}^{\infty} \frac{B_2(\{x+\frac{1}{2}\})}{x^2}\,dx = -\tfrac{2}{3}+\log 2,$$
$$\int_{1/2}^{\infty} \frac{B_3(\{x+\frac{1}{2}\})}{x^3}\,dx = -1+\tfrac{3}{2}\log 2.$$

Proof. From [Montgomery and Vaughan 2007, p. 503 and exercise 23 on p. 508],[1] we find that

$$\int_{1}^{\infty} \frac{B_2(\{x\})}{x^2}\,dx = \log 2\pi - \tfrac{11}{6}. \tag{A-8}$$

We integrate the left-hand side by parts and find that

$$\int_{1}^{\infty} \frac{B_2(\{x\})}{x^2}\,dx = \frac{B_2(\{x\})}{-x}\bigg|_1^{\infty} + \int_{1}^{\infty} \frac{B_2'(\{x\})}{x}\,dx.$$

Now $B_k'(x) = kB_{k-1}(x)$ for $k \ge 1$, see [Montgomery and Vaughan 2007, p. 495], and $B_2(0) = \frac{1}{6}$ by (A-6). Hence, by (A-8),

$$\int_{1}^{\infty} \frac{B_1(\{x\})}{x}\,dx = \tfrac{1}{2}\log 2\pi - 1.$$

Similarly, we find that

$$\int_{1}^{\infty} \frac{B_3(\{x\})}{x^3}\,dx = \tfrac{3}{2}\int_{1}^{\infty} \frac{B_2(\{x\})}{x^2}\,dx = \tfrac{3}{2}\log 2\pi - \tfrac{11}{4}.$$

This establishes the first three formulas of the lemma.

The second set of formulas can be treated in the same way. However, we have not found a ready reference for the value of

$$\int_{1/2}^{\infty} \frac{B_1(\{x+\frac{1}{2}\})}{x}\,dx = \int_{1}^{\infty} \frac{B_1(\{x\})}{x-\frac{1}{2}}\,dx,$$

so we derive it from scratch.

By Riemann–Stieltjes integration

$$\begin{aligned}
&\sum_{1\le n\le N} \log\big(n-\tfrac{1}{2}\big)\\
&= \int_1^N \log\big(x-\tfrac{1}{2}\big)\,dx + \int_{1^-}^N \log\big(x-\tfrac{1}{2}\big)\,d\big([x]-x+\tfrac{1}{2}\big)\\
&= \big(\big(N-\tfrac{1}{2}\big)\log\big(N-\tfrac{1}{2}\big)-\big(N-\tfrac{1}{2}\big)\big)-\big(\tfrac{1}{2}\log\tfrac{1}{2}-\tfrac{1}{2}\big)+\big([x]-x+\tfrac{1}{2}\big)\log\big(x-\tfrac{1}{2}\big)\big|_{1^-}^N - \int_1^N \frac{[x]-x+\frac{1}{2}}{x-\frac{1}{2}}\,dx\\
&= N\log\big(N-\tfrac{1}{2}\big)-N+1+\int_1^N \frac{B_1(\{x\})}{x-\frac{1}{2}}\,dx.
\end{aligned} \tag{A-9}$$

[1] Note that $B_1(x)$ in exercise 23(a) should read $B_1(\{x\})$.

On the other hand,

$$\sum_{1\le n\le N}\log\bigl(n-\tfrac12\bigr)=\log\biggl(\prod_{1\le n\le N}\frac{2n-1}{2}\biggr)=\log\biggl(\frac{(2N)!}{2^{2N}N!}\biggr).$$

By Stirling's formula, $\log n!=n\log n-n+\frac12\log 2\pi n+O(1/n)$, so

$$\begin{aligned}\sum_{1\le n\le N}\log\bigl(n-\tfrac12\bigr)&=\bigl(2N\log 2N-2N+\tfrac12\log 4\pi N\bigr)-2N\log 2-\bigl(N\log N-N+\tfrac12\log 2\pi N\bigr)+O(1/N)\\&=N\log N-N+\tfrac12\log 2+O(1/N).\end{aligned}$$

Combining this and (A-9) we obtain

$$\begin{aligned}\int_1^N\frac{B_1(\{x\})}{x-\frac12}\,dx&=\bigl(N\log N-N+\tfrac12\log 2+O(1/N)\bigr)-\bigl(N\log\bigl(N-\tfrac12\bigr)-N+1\bigr)\\&=N\log\frac{N}{N-\frac12}+\tfrac12\log 2-1+O(1/N).\end{aligned}$$

Letting $N\to\infty$, we deduce that

$$\int_{1/2}^\infty\frac{B_1\bigl(\{x+\frac12\}\bigr)}{x}\,dx=\int_1^\infty\frac{B_1(\{x\})}{x-\frac12}\,dx=-\tfrac12+\tfrac12\log 2.$$

We now argue as above to find the value of the remaining two integrals. Integration by parts using $B_j'(x)=jB_{j-1}(x)$ reveals that

$$\int_{1/2}^\infty\frac{B_j\bigl(\{x+\frac12\}\bigr)}{x^j}\,dx=\frac{B_j(\{1\})2^{j-1}}{(j-1)}+\frac{j}{j-1}\int_{1/2}^\infty\frac{B_{j-1}\bigl(\{x+\frac12\}\bigr)}{x^{j-1}}\,dx.$$

Thus, using (A-6) we find that

$$\begin{aligned}\int_{1/2}^\infty\frac{B_3\bigl(\{x+\frac12\}\bigr)}{x^3}\,dx&=\tfrac32\int_{1/2}^\infty\frac{B_2\bigl(\{x+\frac12\}\bigr)}{x^2}\,dx,\\\int_{1/2}^\infty\frac{B_2(\{x+\frac12\})}{x^2}\,dx&=\tfrac13+2\int_{1/2}^\infty\frac{B_1(\{x+\frac12\})}{x}\,dx\end{aligned}$$

Thus, the B_2 integral equals $-\frac23+\log 2$ and the B_3 integral equals $-1+\frac32\log 2$. This completes the proof of the lemma. □

Acknowledgement

The authors are most grateful to Dr. Fan Ge, who carefully read the manuscript and made a number of helpful suggestions.

References

[Chamizo 1999] F. Chamizo, "Correlated sums of $r(n)$", *J. Math. Soc. Japan* **51**:1 (1999), 237–252. MR Zbl

[van der Corput 1923] J. G. van der Corput, "Neue zahlentheoretische Abschätzungen", *Math. Ann.* **89**:3-4 (1923), 215–254. MR Zbl

[Cramér 1922] H. Cramér, "Über zwei Sätze des Herrn G. H. Hardy", *Math. Z.* **15**:1 (1922), 201–210. MR JFM

[Hardy 1915] G. H. Hardy, "On the expression of a number as the sum of two squares", *Quart. J. Math* **46** (1915), 263–283. JFM

[Hardy 1916a] G. H. Hardy, "The average order of the arithmetical functions $P(x)$ and $\Delta(x)$", *Proc. London Math. Soc.* (2) **15** (1916), 192–213. MR Zbl

[Hardy 1916b] G. H. Hardy, "On Dirichlet's divisor problem", *Proc. London Math. Soc.* (2) **15** (1916), 1–25. MR Zbl

[Huxley 2003] M. N. Huxley, "Exponential sums and lattice points, III", *Proc. London Math. Soc.* (3) **87**:3 (2003), 591–609. MR Zbl

[Ivić 1996] A. Ivić, "The Laplace transform of the square in the circle and divisor problems", *Studia Sci. Math. Hungar.* **32**:1-2 (1996), 181–205. MR Zbl

[Ivić 2001] A. Ivić, "A note on the Laplace transform of the square in the circle problem", *Studia Sci. Math. Hungar.* **37**:3-4 (2001), 391–399. MR Zbl

[Iwaniec and Mozzochi 1988] H. Iwaniec and C. J. Mozzochi, "On the divisor and circle problems", *J. Number Theory* **29**:1 (1988), 60–93. MR Zbl

[Kátai 1965] I. Kátai, "The number of lattice points in a circle", *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* **8** (1965), 39–60. In Russian. MR Zbl

[Kolesnik 1985] G. Kolesnik, "On the method of exponent pairs", *Acta Arith.* **45**:2 (1985), 115–143. MR Zbl

[Landau 1923] E. Landau, "Über die Gitterpunkte in einem Kreise (vierte Mitteilung)", *Nachr. Ges. Wiss. Göttingen, Math.-Naturwiss. Kl.* **1923** (1923), 58–65.

[Lau and Tsang 2009] Y.-K. Lau and K.-M. Tsang, "On the mean square formula of the error term in the Dirichlet divisor problem", *Math. Proc. Cambridge Philos. Soc.* **146**:2 (2009), 277–287. MR Zbl

[Montgomery and Odlyzko 1988] H. L. Montgomery and A. M. Odlyzko, "Large deviations of sums of independent random variables", *Acta Arith.* **49**:4 (1988), 427–434. MR Zbl

[Montgomery and Vaughan 2007] H. L. Montgomery and R. C. Vaughan, *Multiplicative number theory, I: Classical theory*, Cambridge Studies in Advanced Mathematics **97**, Cambridge University Press, 2007. MR Zbl

[Nowak 2004] W. G. Nowak, "Lattice points in a circle: an improved mean-square asymptotics", *Acta Arith.* **113**:3 (2004), 259–272. MR Zbl

[Preissmann 1988] E. Preissmann, "Sur la moyenne quadratique du terme de reste du problème du cercle", *C. R. Acad. Sci. Paris Sér. I Math.* **306**:4 (1988), 151–154. MR Zbl

[Sierpiński 1906] W. Sierpiński, "O pewnem zagadnieniu z rachunku funkcyj asymptotycznych", *Prace Matematyczno-Fizyczne* **17** (1906), 77–118.

[Soundararajan 2003] K. Soundararajan, "Omega results for the divisor and circle problems", *Int. Math. Res. Not.* **2003**:36 (2003), 1987–1998. MR Zbl

[Walfisz 1927] A. Walfisz, "Teilerprobleme", *Math. Z.* **26**:1 (1927), 66–88. MR Zbl

Received 11 Dec 2018. Revised 10 Mar 2019.

STEVEN M. GONEK:
gonek@math.rochester.edu
Department of Mathematics, University of Rochester, Rochester, NY, United States

ALEX IOSEVICH:
iosevich@math.rochester.edu
Department of Mathematics, University of Rochester, Rochester, NY, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the submission page.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles are usually in English or French, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not refer to bibliography keys. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and a Mathematics Subject Classification for the article, and, for each author, affiliation (if appropriate) and email address.

Format. Authors are encouraged to use LaTeX and the standard amsart class, but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should normally be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages — Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc. — allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with as many details as you can about how your graphics were generated.

Bundle your figure files into a single archive (using zip, tar, rar or other format of your choice) and upload on the link you been provided at acceptance time. Each figure should be captioned and numbered so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript with all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables.

White Space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Moscow Journal of Combinatorics and Number Theory
vol. 8 no. 3 2019

Integer complexity: the integer defect 193
HARRY ALTMAN

Generalized simultaneous approximation to m linearly dependent reals 219
LEONHARD SUMMERER

On the distribution of values of Hardy's Z-functions in short intervals, II: The q-aspect 229
RAMDIN MAWIA

On products of shifts in arbitrary fields 247
AUDIE WARREN

The mean square discrepancy in the circle problem 263
STEVEN M. GONEK and ALEX IOSEVICH