Geom. Funct. Anal.
https://doi.org/10.1007/s00039-023-00641-y
GAFA Geometric And Functional Analysis

ON THE RANDOM CHOWLA CONJECTURE

Oleksiy Klurman, Ilya D. Shkredov and Max Wenqiang Xu

Abstract. We show that for a Steinhaus random multiplicative function $f : \mathbb{N} \to \mathbb{D}$ and any polynomial $P(x) \in \mathbb{Z}[x]$ of $\deg P \geq 2$ which is not of the form $w(x+c)^d$ for some $w \in \mathbb{Z}$, $c \in \mathbb{Q}$, we have

$$\frac{1}{\sqrt{N}} \sum_{n \leq N} f(P(n)) \xrightarrow{d} \mathcal{CN}(0,1),$$

where $\mathcal{CN}(0,1)$ is the standard complex Gaussian distribution with mean 0 and variance 1. This confirms a conjecture of Najnudel in a strong form. We further show that there almost surely exist arbitrary large values of $x \geq 1$, such that

$$\left| \sum_{n \leq x} f(P(n)) \right| \gg_P \sqrt{x} (\log \log x)^{1/2},$$

for any polynomial $P(x) \in \mathbb{Z}[x]$ with $\deg P \geq 2$, which is not a product of linear factors (over $\mathbb{Q}$). This matches the bound predicted by the law of the iterated logarithm. Both of these results are in contrast with the well-known case of linear polynomial $P(n) = n$, where the partial sums are known to behave in a non-Gaussian fashion and the corresponding sharp fluctuations are speculated to be $O(\sqrt{x}(\log \log x)^{\frac{1}{4}+\varepsilon})$ for any $\varepsilon > 0$.

1 Introduction

The main focus of the present paper is to take yet another look at one of the two most studied models of random multiplicative functions. Let $(f(p))_{p \text{ prime}}$ be a sequence of independent uniformly distributed on the unit circle $\{|z| = 1\}$ random variables. A Steinhaus random multiplicative function is given by $f(n) = \prod_{p^\beta || n} f(p)^\beta$. Similarly, let $(f(p))_{p \text{ prime}}$ be a sequence of independent random variables taking values ±1 with probability 1/2, then a Rademacher random multiplicative function is given by $f(n) := \prod_{p|n} f(p)$ for all n which are square-free, and $f(n) = 0$ otherwise. In 1944, Wintner [Win44] introduced Rademacher random multiplicative functions to model the behaviour of the Möbius function $\mu(n)$, whereas Steinhaus random multiplicative functions are intended to model randomly selected Dirichlet characters $\chi(n)$, and

Published online: 18 April 2023

Archemidean characters n^{it} $(t \in \mathbb{R})$. We refer the reader to [GS01, Section 2] and the introduction of [Har20, Har21] for a meticulous overview of this subject.

A classical question of interest which attracted a lot of attention is to understand the distribution and the sizes of the partial sums $\sum_{n\leq x} f(n)$. The fundamental difficulty stems from the fact that the values $f(n)$ and $f(m)$ are not independent whenever $(m,n) > 1$ and thus the corresponding sums cannot be directly treated using tools for independent random variables.

1.1 Distribution results. It is a natural guiding conjecture that $\frac{1}{\sqrt{N}}\sum_{n\leq N} f(n) \xrightarrow{d} \mathcal{CN}(0,1)$ in the Steinhaus case and $\frac{1}{\sqrt{N}}\sum_{n\leq N} f(n) \xrightarrow{d} \mathcal{N}(0,1)$ in the Rademacher case, where "$\xrightarrow{d}$" stands for convergence in distribution and $\mathcal{N}(0,1)$ and $\mathcal{CN}(0,1)$ stand for standard real and complex Gaussian distribution respectively. But Chatterjee suggested that this conjecture should not hold. Chatterjee's conjecture (expressed in [Hou11]), was proved by Harper [Har13b], using an intricate conditioning argument. It is now a direct consequence of a more recent breakthrough work by Harper [Har20] on Helson's conjecture that in fact $\frac{1}{\sqrt{N}}\sum_{n\leq N} f(n) \xrightarrow{d} 0$ in both cases. Interestingly, if one restricts to several natural subsums, Chatterjee and Soundararajan [CS12], Harper [Har13b] and Hough [Hou11] established central limit theorems. It remains a deep mystery whether appropriately normalized partial sums $\sum_{n\leq N} f(n)$ have a limiting distribution as $N \to \infty$.

The problem considered in this note is motivated by the celebrated conjecture of Chowla [Cho65], which states that for the Liouville (or the Möbius) function λ and any polynomial $P(x) \in \mathbb{Z}[x]$, which is not of the form $P(x) = cg^2(x)$ for some $g \in \mathbb{Z}[x]$,

$$\sum_{n\leq x} \lambda(P(n)) = o(x).$$

The case $\deg(P) = 1$ corresponds to the prime number theorem but the general case is widely open for any polynomial with $\deg P \geq 2$. Some remarkable progress has been recently made in the case $P(x) = \prod_{k=1}^{n}(a_k x + b_k)$ and $a_i, b_i \in \mathbb{Z}$ (albeit with a logarithmic weight) in the case of rather general multiplicative functions (so–called Elliott's conjecture, see [Ell92, MRT15] and [Tao16]) thanks to the combination of several works by Tao [Tao16], Matomäki-Radziwiłł-Tao [MRT15], Tao and Teräväinen [TT18], and more recently by Helfgott and Radziwiłł[HR21]. A weaker statement, that $\lambda(P(n))$ changes sign infinitely often has been obtained by Cassaigne-Ferenczi-Mauduit-Rivat-Sárközy [CFM00], Borwein-Choi-Ganguly [BCG13], and more recently by Teräväinen [Ter20] for a special class of polynomials $P(x) \in \mathbb{Z}[x]$.

Prior to our work, we are aware of no unconditional results for Chowla's conjecture in the context of random multiplicative functions for any polynomial of $\deg P \geq 2$. It has been previously speculated and Najnudel [Naj20] conjectured that if $P(x) = x(x+1)$ (and more generally, if $P(x) = \prod_{i=1}^{k}(x+a_i)^{m_i}$) then the convergence in distribution

$$\frac{1}{\sqrt{N}} \sum_{n\leq N} f(n)f(n+1) \xrightarrow{d} \mathcal{CN}(0,1),$$

must hold for f being a Steinhaus random multiplicative function and reformulated this conjecture *in terms of showing that certain family of Diophantine equations possess only trivial solutions*. Such a family naturally arises while computing $2k$-th moment of the left hand side for arbitrarily large values of $k \geq 1$. Our first result is an unconditional version of a central limit theorem which works for general polynomials $P(x) \in \mathbb{Z}[x]$. To keep our notations consistent, we may assume $f(-n) := f(n)$ for all $n \in \mathbb{N}$ and $f(0) = 0$.

Theorem 1.1. *Let f be a Steinhaus random multiplicative function. Then for any polynomial $P(x)$ in $\mathbb{Z}[x]$ with $\deg P \geq 2$ which is not of the form $P(x) = w(x+c)^d$ for some $w \in \mathbb{Z}$, $c \in \mathbb{Q}$, as $N \to \infty$,*

$$\frac{1}{\sqrt{N}} \sum_{n\leq N} f(P(n)) \xrightarrow{d} \mathcal{CN}(0,1).$$

This result is optimal since in the case $P(x) = w(x+c)^d$, for some $w \in \mathbb{Z}$, $c \in \mathbb{Q}$ we have $\frac{1}{\sqrt{N}} \sum_{n\leq N} f(w(x+c)^d) \xrightarrow{d} 0$, after noticing that f^d is also a Steinhaus random multiplicative function and appealing to the results in [Har20]. It is worth mentioning that the same proof allows us to deduce central limit theorems for various sparse subsums. For example, without much additional effort, one could show that when $n = p$ are primes,

$$\frac{1}{\sqrt{\pi(N)}} \sum_{p\leq N} f(P(p)) \xrightarrow{d} \mathcal{CN}(0,1). \tag{1.1}$$

See Remark 2.5 for further discussions on how to establish (1.1).

1.2 Large fluctuations.

A classical question in probability is to understand the largest fluctuations of the sums of independent random variables. If, say, $\{\xi_k\}_{k=1}^{\infty}$ is a sequence of independent Steinhaus random variables, then according to Khintchine's law of the iterated logarithm, we almost surely have

$$\limsup_{x\to\infty} \frac{|\sum_{k\leq x} \xi_k|}{\sqrt{2x\log\log x}} = 1. \tag{1.2}$$

An important feature is that (1.2) exhibits the magnitude of the global fluctuations (that is $\sqrt{x\log\log x}$) which is substantially larger than the expected size of the partial sums at any given point x (of the order $\sqrt{x}$).

In the case of random multiplicative functions this subject has a long and rich history. In a pioneering paper Wintner [Win44] studied random Dirichlet series and in the Rademacher case was able to exhibit an almost sure bound $\sum_{n\le x} f(n) = O(x^{1/2+\varepsilon})$ and moreover, almost surely $\sum_{n\le x} f(n) = O(x^{1/2-\varepsilon})$ is false. Erdős (unpublished but stated in [Erd85]) claimed that almost surely one has the bound $O(\sqrt{x}(\log x)^A)$ and one almost surely does not have $O(\sqrt{x}(\log x)^{-B})$ for some constants $A, B > 0$. In a beautiful and rather influential work, Halász [Hal83] proved an almost sure bound $O(\sqrt{x}\exp(A\sqrt{\log\log x \log\log\log x}))$ and that one almost surely does not have $O(\sqrt{x}\exp(-B\sqrt{\log\log x \log\log\log x}))$ for some positive constants A, B. Thirty years later, Lau, Tenenbaum and Wu [LTW13] (see also related work [Bas12]) sharpened the analysis of hypercontractive inequalities in Halász's argument, establishing an almost sure upper bound $O(\sqrt{x}(\log\log x)^{2+\varepsilon})$. On the other hand, Harper [Har13] used Gaussian process machinery to study the suprema of random Euler products, showing that almost surely $O(\sqrt{x}/(\log\log x)^{5/2+\varepsilon})$ is false. The latter results may be seen as approximations to the law of the iterated logarithm however quantitatively substantially weaker. In a recent breakthrough, answering a question of Halász and proving an old conjecture of Erdős, Harper [Har21] showed that if f is a Steinhaus (or Rademacher) random multiplicative function, then almost surely $|\sum_{n\le x} f(n)| \ge \sqrt{x}(\log\log x)^{1/4-\varepsilon}$ holds for a sequence of arbitrary large values of $x \ge 1$. Remarkably, this furnishes the first bound that grows faster than $\sqrt{x}$ and moreover the exponent $1/4$ is speculated to be sharp (see also [Har20, Mas22]).

We establish a lower bound of the size $\sqrt{x\log\log x}$ matching the one predicted by the Khintchine's type law of the iterated logarithm.

Corollary 1.2. *Let f be a Steinhaus random multiplicative function. Then for any polynomial $P(x)$ in $\mathbb{Z}[x]$ with $\deg P \ge 2$ which is not a product of linear factors (over $\mathbb{Q}$), there almost surely exists arbitrarily large x such that*

$$\left|\sum_{n\le x} f(P(n))\right| \gg_{\deg P} \sqrt{x\log\log x}. \tag{1.3}$$

In fact, we prove a more general local version and then apply the standard Borel–Cantelli type argument to deduce Corollary 1.2.

Theorem 1.3. *Let f be a Steinhaus random multiplicative function and let $P(x)$ be a polynomial in $\mathbb{Z}[x]$ with $d = \deg P \ge 2$ which is not a product of linear factors (over $\mathbb{Q}$). Then uniformly for all large X,*

$$\max_{X\le x\le X^{(\log X)^2}} \frac{1}{\sqrt{x}}\left|\sum_{n\le x} f(P(n))\right| \ge c_d\sqrt{\log\log X} \tag{1.4}$$

with probability $1 - O(\frac{1}{(\log\log X)^{0.02}})$ for some fixed $c_d > 0$ depending on d.

We conclude this section by mentioning that in the deterministic case, a well-known conjecture of Gonek [Ng04] predicts the sharp upper bound

$\sum_{n\le x}\mu(n) = O(\sqrt{x}(\log\log\log x)^{5/4})$. In view of Theorem 1.1 and Corollary 1.2 it seems reasonable to expect that the largest fluctuations of the Chowla type sums $\sum_{n\le x}\mu(P(n))$ are of the order $\sqrt{x\log\log x}$ for any admissible polynomial $P(x)\in\mathbb{Z}[x]$ of $\deg P\ge 2$.

1.3 Outline of the proofs.

A standard point of departure in establishing central limit theorems is a computation of higher (integral) moments:

$$\mathbb{E}\left|\sum_{n\le N} f(P(n))\right|^{2k} = \sum_{1\le n_i,m_i\le N} \mathbb{1}_{P(n_1)P(n_2)\ldots P(n_k)=P(m_1)P(m_2)\ldots P(m_k)}.$$

The latter naturally leads to a consideration of the *higher multiplicative energies* of the polynomial images, which is interesting on its own right (see Section 2 for the discussion). This seems to be a difficult problem as far as general polynomials $P\in\mathbb{Z}[x]$ are concerned for any $k\ge 3$. [1]

To overcome this obstacle and prove Theorem 1.1, we take advantage of the crucial feature that the partial sums $\sum_{n\le N} f(P(n))$ exhibit the structure of a *martingale difference sequence*. Such an observation has been previously utilized by several authors including Harper [Har13b] and Lau, Tenenbaum and Wu [LTW13] in the context of studying non-Gaussian behaviour of $\sum_{n\le N} f(n)$. After applying a complex-valued version of McLeish's martingale central limit theorem (which we will state in Section 2), fortunately, the case $k=2$ suffices to accomplish our modest task. To this end for subsets $A\subseteq\mathbb{R}$, we introduce a multiplicative energy [TV06] of the set $\mathsf{E}^\times(A) := \#\{(a_1,a_2,a_3,a_4)\in A \;:\; a_1a_2=a_3a_4\ne 0\}$ and prove the following. Let $[N]=\{1,2,\ldots,N\}$, where $N\in\mathbb{Z}$, $N>0$.

Proposition 1.4. *Let $N\ge 1$ be a positive integer and $P(x)\in\mathbb{Z}[x]$ be a polynomial with degree $d\ge 2$ and $P(x)\ne w(x+c)^d$ for any $w\in\mathbb{Z}$, $c\in\mathbb{Q}$. Then, for $d>2$ we have the bounds*

$$\mathsf{E}^\times(P([N])) = 2N^2 + O_d(N^{2-\frac{1}{2(2d-1)}+o_d(1)})$$

and for $d=2$

$$\mathsf{E}^\times(P([N])) = 2N^2 + O(N^{5/3+o(1)}).$$

Proposition 1.4 will be immediately deduced from a more general Theorem 3.2, with the key input in the proof coming from the use of a celebrated result of Bombieri–Pila [BP89] bounding the number of integral points on curves.

The proof of Theorem 1.3 heavily relies on several probabilistic results (in the form established in [Har21]). Roughly speaking, in the linear case $P(n)=n$, Harper establishes a multivariate Gaussian approximation for the sums $\sum_{X<p\le x} f(p)\sum_{n\le x/p} f(n)$ conditional on all the values $(f(p))_{p\le X}$, sampled at a well spaced sequence of ($\asymp \log X$) points $X^{8/7}\le x\le X^{4/3}$, thus making the inner sums fixed. Normal

[1] In a recent preprint [WX22], the authors have made progress towards this question.

approximation and normal comparison results (Lemmas 4.3 and 4.4 respectively) are then used to produce large fluctuations. The main bulk of the work goes into analyzing the sizes of conditional variances and covariances using techniques from multiplicative chaos. In our case, we first use a conditioning argument but the set of primes to condition on is chosen more judiciously (in the spirit of a greedy algorithm). The key consequence of such a conditioning argument together with Proposition 1.4 is that the "essential" parts of the random sums at different scales become *independent* with the *conditional variance being roughly of size x* with very high probability. Such independence here simplifies analysis of the covariance structure and makes the normal comparison Lemma 4.4 easy to apply. The fact that "typical" conditional variance is of size $\asymp x$ might explain why our bounds match those predicted by (1.2). These two features are both different from the linear case studied in [Har21]. The main arithmetic input we use to construct such a set of primes to condition on is that for any polynomial $P \in \mathbb{Z}[x]$, which is not a product of linear factors, the set of $n \leq X$ with the largest prime factor $P^{+}(P(n)) \gg_d n \log n$, has positive density (see Lemma 4.2 due to Maynard and Rudnick [MR21]). Such "very large" primes are clearly not available in the linear case.

1.4 Organization of the paper and future work. We prove Theorem 1.1 in Section 2 with the crucial energy bounds deferred to be proved in Section 3. In Section 4, we prove Theorem 1.3 and Corollary 1.2. The situation with Rademacher random multiplicative functions is more delicate. With additional effort, the methods of the present paper also work in that case for P belonging to a wide class of polynomials. However for the case of P with $\deg P \geq 3$ even the existence of a positive proportion of square-free values of $P(n)$ is only known under the assumption of the ABC conjecture thanks to the work of Granville [Gra98]. We have decided to keep the presentation here relatively simple focusing on the main ideas rather than the generality of the results. In future work, we shall return to the study of the Rademacher case (both unconditionally and conditional on the ABC conjecture).

2 Proof of Theorem 1.1

We begin with the following preparatory observations. Given a polynomial $P(x) \in \mathbb{Z}[x]$ (as in Theorems 1.1 and 1.3) with a positive leading coefficient, there exists a constant $N_0 := N_0(P)$ such that for all $n \geq N_0$, the values

$$P(n+1) > P(n) > P(N_0) > 0. \tag{2.1}$$

Similar monotonicity property clearly holds when $P(x)$ has a negative leading coefficient.

We now note that the limiting distribution of the partial sums $\sum_{n\leq N} \frac{1}{\sqrt{N}} f(P(n))$ is the same as of $\sum_{N_0\leq n\leq N} \frac{1}{\sqrt{N}} f(P(n))$, since we have a pointwise bound

$$\sum_{n\leq N_0} \frac{1}{\sqrt{N}} f(P(n)) \leq \frac{N_0}{\sqrt{N}} = o_{N\to+\infty}(1). \tag{2.2}$$

Therefore, changing $P(x)$ to $P(x+N_0)$ if necessary, we may assume throughout the paper that (2.1) holds for all $n \geq 1$. Inspired by the work of Harper [Har13b], our key observation here is that the partial sums of random multiplicative function possess the structure of a martingale difference sequence.

Definition 2.1 *(Martingale difference sequence). Let $Z_1, Z_2, \dots, Z_N$ be a sequence of complex-valued random variables. Suppose $\mathbb{E}[Z_1] = 0$ and for all $1 \leq i \leq N-1$,*

$$\mathbb{E}[Z_{i+1} | Z_1, \dots, Z_i] = 0.$$

Then $(Z_i)_{i \leq N}$ form a Martingale difference sequence.

We next introduce the following complex-valued version of a classical result due to McLeish [McL74], recently developed in [SX22, Theorem 2.4]. This is particularly suitable for proving central limit theorems in the Steinhaus setting.

Lemma 2.2 (Complex-valued version of McLeish's theorem). *Let Z_1, …, Z_N be a martingale difference sequence of complex-valued random variables, and put $S_N = \sum_{n=1}^{N} Z_n$. Assume that $\mathbb{E}[|Z_n|^4]$ exists for each n. Then for any fixed real numbers t_1 and t_2 we have, with $t^2 = (t_1^2 + t_2^2)/2$,*

$$\mathbb{E}[e^{it_1 Re(S_N) + it_2 Im(S_N)}] = e^{-t^2/2} + O\Big(e^{t^2}\Big(\sum_{n=1}^{N} \mathbb{E}[|Z_n|^4]\Big)^{\frac{1}{4}}\Big) + O\Big(e^{t^2}\Big(\mathbb{E}\Big[\Big(\sum_{n=1}^{N} |Z_n|^2 - 1\Big)^2\Big]\Big)^{\frac{1}{2}}\Big)$$
$$+ O\Big(e^{t^2} \max_{\phi \in [0,2\pi]} \Big(\mathbb{E}\Big[\Big(\sum_{n=1}^{N} (e^{-i\phi} Z_n^2 + e^{i\phi} \overline{Z_n}^2)\Big)^2\Big]\Big)^{\frac{1}{2}}\Big).$$

In order to apply Lemma 2.2 to our setting, we let $P^+(m)$ be the largest prime factor of a positive integer m and consider

$$M_p = M_p(N) := \frac{1}{\sqrt{N}} \sum_{\substack{n \leq N \\ P^+(P(n)) = p}} f(P(n)). \tag{2.3}$$

By our initial reduction, we have that (2.1) holds for all $n \geq 1$, and consequently, for any pair $1 \leq m, n \leq N$, we have the orthogonality relation

$$\mathbb{E}[f(P(m))\overline{f(P(n))}] = \mathbb{1}_{P(m)=P(n)} = \mathbb{1}_{n=m}. \tag{2.4}$$

It follows that

$$\mathbb{E}[M_p | f(q) : q < p] = 0,$$

yielding that $(M_p)_p$ form a martingale difference sequence. Lemma 2.2, in turn, readily implies the following result.

Lemma 2.3. *Let f be a Steinhaus random multiplicative function, and let*

$$\mathcal{A} = \mathcal{A}(N) = \{P(n) : 1 \leq n \leq N\}.$$

Suppose the following conditions hold:

(1) *We have*

$$\#\{(m_1,m_2,n_1,n_2)\in\mathcal{A}^4 : m_1m_2=n_1n_2, P^+(m_i)=P^+(n_i), m_i \neq n_i,\ 1\le i\le 2\} = o_{N\to+\infty}(N^2).$$

(2) *We have*

$$\#\{(m_1,m_2,n_1,n_2)\in\mathcal{A}^4 : m_1m_2=n_1n_2, P^+(m_1)=P^+(m_2) = P^+(n_1)=P^+(n_2)\} = o_{N\to+\infty}(N^2).$$

Then, as $N\to+\infty$, we have

$$\frac{1}{\sqrt{N}}\sum_{n\le N} f(P(n)) \xrightarrow{d} \mathcal{CN}(0,1).$$

Proof of Lemma 2.3. The largest prime factor of $P(x)$ is CN^d for some constant C and d is the degree of the polynomial. Consider $S_N=\sum_{p\le CN^d} M_p$, where M_p are defined in (2.3). As noted before, $\{M_p\}_p$ forms a martingale difference sequence. Therefore, we may apply Lemma 2.2 to evaluate $\mathbb{E}[e^{it_1\mathrm{Re}(S_N)+it_2\mathrm{Im}(S_N)}]$. Observe that

$$\sum_{p\le N}\mathbb{E}[|M_p|^4] = \frac{1}{|\mathcal{A}|^2}\sum_{p\le CN^d}\ \sum_{\substack{m_1,m_2,n_1,n_2\in\mathcal{A}\\ P^+(m_1)=P^+(m_2)=P^+(n_1)=P^+(n_2)=p}} \mathbb{E}[f(m_1m_2)\overline{f(n_1n_2)}]$$
$$= \frac{1}{|\mathcal{A}|^2}\sum_{\substack{m_1,m_2,n_1,n_2\in\mathcal{A}\\ P^+(m_1)=P^+(m_2)=P^+(n_1)=P^+(n_2)\\ m_1m_2=n_1n_2}} 1,$$

which is $o(1)$ by our assumption (2). Thus the first error term in Lemma 2.2 becomes $o(e^{t^2})$. We next compute

$$\mathbb{E}\Big[\Big(\sum_{p\le CN^d}|M_p|^2-1\Big)^2\Big] = \sum_{p,q\le CN^d}\mathbb{E}[|M_p|^2|M_q|^2] - 2\sum_{p\le CN^d}\mathbb{E}[|M_p|^2]+1. \quad (2.5)$$

Separating the terms $m_1=n_1$ and $m_2=n_2$, and using assumption (1) to bound the remaining terms, we end up with

$$\sum_{p,q\le CN^d}\mathbb{E}[|M_p|^2|M_q|^2] = \frac{1}{|\mathcal{A}|^2}\sum_{p,q\le CN^d}\ \sum_{\substack{m_1,n_1\in\mathcal{A}\\ P^+(n_1)=P^+(m_1)=p}}\ \sum_{\substack{m_2,n_2\in\mathcal{A}\\ P^+(m_2)=P^+(n_2)=q}} \mathbb{E}[f(m_1m_2)\overline{f(n_1n_2)}]$$
$$= \frac{1}{|\mathcal{A}|^2}\sum_{\substack{m_1,n_1,m_2,n_2\in\mathcal{A}\\ m_1m_2=n_1n_2\\ P^+(m_1)=P^+(n_1)\\ P^+(m_2)=P^+(n_2)}} 1 = \frac{1}{|\mathcal{A}|^2}\Big(\sum_{n\in\mathcal{A}}1\Big)^2+o(1)=1+o(1).$$

Further the subtracted term in (2.5) is

$$\sum_{p\le CN^d}\mathbb{E}[|M_p|^2] = \frac{1}{|\mathcal{A}|}\sum_{p\le CN^d}\ \sum_{\substack{m,n\in\mathcal{A}\\ P^+(m)=P^+(n)=p}}\mathbb{E}[f(m)\overline{f(n)}] = \frac{1}{|\mathcal{A}|}\sum_{n\in\mathcal{A}}1=1,$$

so that (2.5) gives acceptable contribution of $o(1)$. Therefore the second error term while using Lemma 2.2 can also be bounded by $o(e^{t^2})$.

We are left to handle the third error term in Lemma 2.2, which involves the maximum over $\phi \in [0,2\pi]$ of

$$\mathbb{E}\Big[\Big(\sum_{p\le CN^d}(e^{-i\phi}M_p^2+e^{i\phi}\overline{M}_p{}^2)\Big)^2\Big]=\sum_{p,q\le CN^d}\mathbb{E}\Big[(e^{-i\phi}M_p^2+e^{i\phi}\overline{M}_p{}^2)(e^{-i\phi}M_q^2+e^{i\phi}\overline{M}_q{}^2)\Big].$$

If $p\neq q$ then upon expanding we get

$$\mathbb{E}[M_p^2M_q^2]=\mathbb{E}[M_p^2\overline{M}_q{}^2]=\mathbb{E}[\overline{M}_p^2M_q^2]=\mathbb{E}[\overline{M}_p^2\overline{M}_q{}^2]=0.$$

From our treatment of the first error term, it follows that the terms with $p=q$ contribute

$$\ll\sum_{p\le CN^d}\mathbb{E}\Big[|M_p|^4\Big]=o(1).$$

Thus the third error term appearing in Lemma 2.2 is $o(1)$ and we have finally showed that

$$\mathbb{E}[e^{it_1\mathrm{Re}(S_N)+it_2\mathrm{Im}(S_N)}]=e^{-t^2}+o(1).$$

Note that e^{-t^2} is the Fourier transform of a standard Gaussian which in turn imply the desired convergence in distribution. □

We use "$\ll_d$" and "$\ll_P$" to denote that the implicit constant depends on the degree d or polynomial P respectively.

Proof of Theorem 1.1 assuming Proposition 1.4. It suffices to check that both conditions of Lemma 2.3 are satisfied. To check condition (1), we need to show that

$$\begin{aligned}&\#\{(m_1,m_2,n_1,n_2)\in[N]^4:P(m_1)P(m_2)=P(n_1)P(n_2),\{m_1,m_2\}\neq\{n_1,n_2\}\}\\&\quad=o_{N\to+\infty}(N^2).\end{aligned}$$

The latter immediately follows from Proposition 1.4. The second condition can be written as

$$\begin{aligned}&\#\{(m_1,m_2,n_1,n_2)\in[N]^4:\ P^+(P(m_1))=P^+(P(m_2))=P^+(P(n_1))=P^+(P(n_2)),\\&\quad\text{and } P(m_1)P(m_2)=P(n_1)P(n_2)\}=o(N^2).\end{aligned}$$

Using Proposition 1.4, we conclude that the off-diagonal contribution is $o(N^2)$. To estimate the diagonal contribution, we distinguish between three ranges: $p\le\log\log N$, $\log\log N\le p\le N$ or $p>N$.

We first consider the case $p\le\log\log N$. Since $P(n)\in\{1,\dots,P(N)\}$ for $n\le N$, we trivially bound the contribution in this case by $\ll\psi(P(N),\log\log N)^4$, where $\psi(x,y)$ is the number of y-smooth integers between 1 and x. By using the estimate for the smooth numbers (e.g. see [Gra08, (1.18)]) and the fact that $P(N) = O(N^d)$, we have the bounds

$$\ll_P \psi(P(N), \log\log N)^4 \ll_P (\log N)^{O_P(\log\log N)} = o_P(N).$$

Next, we consider the case $\log\log N \leq p \leq N$. Notice that the number of $n \leq p$ such that $p|P(n)$ is at most $d = \deg P$ for each fixed p and consequently, the number of *diagonal* solutions is at most

$$\ll_P \sum_{\log\log N \leq p \leq N} \frac{N^2}{p^2} = o_P(N^2).$$

Finally, if $p > N$ we notice that for each fixed $n \leq N$ with large $N \geq 1$, there are at most $O_d(1)$ primes $p \geq N$ with $p|P(n)$ and therefore there is in total $O_d(N)$ number of pairs (p, n) such that $p|P(n)$ and $p \geq N$. Combining with the fact that for each $p \geq N$ there are at most $O_d(1)$ integers n such that $p|P(n)$, it follows that the number of diagonal solutions in this regime is $O_d(N)$ which is negligible. This concludes the proof. □

Remark 2.4. Lemma 2.3 and Theorem 3.2 could be used to establish a quantitative rate of convergence in our Theorem 1.1. We leave the details to the interested reader.

Remark 2.5. We now sketch the proof to the claim in (1.1), i.e. establishing similar results when n ranges over the primes. Due to the power saving N^c for some $c > 0$, in Proposition 1.4 for the off-diagonal contribution, it is enough to show that the number of diagonal solutions to $P(p_1)P(p_2) = P(p_3)P(p_4)$ with all $P(p_i)$ having the same largest prime factors is $o(\pi(N)^2)$. We follow similar lines as above and split the estimation into three cases. When $p \leq \log\log N$ or $p \geq N$, the same bounds used in the proof of Theorem 1.1 are sufficient for our purposes. Finally, for $\log\log N \leq p \leq N$, we have an estimate

$$\ll_P \sum_{\log\log N \leq p \leq N} \frac{N^2}{(\log \frac{2N}{p})^2 p^2} = o_d(\pi(N)^2).$$

3 Energy bounds and paucity phenomena

The main purpose of this section is to give a proof of Theorem 3.2 which concerns the paucity phenomenon of polynomial sequence and directly implies Proposition 1.4. Our task is to calculate the number of integral points on the variety

$$\mathcal{V}_P = \{(x_1, x_2, x_1', x_2') \in [N] \ : \ P(x_1)P(x_2) = P(x_1')P(x_2')\} \tag{3.1}$$

and more generally, we consider the variety

$$\mathcal{V}_P = \{(x_1, \ldots, x_k, x_1', \ldots, x_k') \in \mathbb{F}^{2k} \ : \ P(x_1)\ldots P(x_k) = P(x_1')\ldots P(x_k')\}, \tag{3.2}$$

where $\mathbb{F}$ is an arbitrary field and $k \geq 2$. We aim to obtain a paucity result, that is, the number of the "non–trivial" solutions is negligible relative to the "trivial" ones. There are two basic questions concerning $\mathcal{V}_P$. The first one is related to the definition of the "trivial" solutions. The points $(x_1, \dots, x_k, x'_1, \dots, x'_k)$ with $\{x_1, \dots, x_k\} = \{x'_1, \dots, x'_k\}$ clearly belong to $\mathcal{V}_P$ and one can hope that those contribute the main term to $\mathsf{E}^{\times}(P[N])$. The other natural choice comes from the "trivial" curves lying on (3.2) which are of the form $P(x_i) = P(x'_j)$, $i, j \in [k]$. To this end, one can show that if the curve $P(x) - P(y) = 0$ is irreducible, then it contains a negligible number of points. The question of its reducibility is more subtle and it is known [Fri70, Theorem 1] that the polynomial $\phi(x, y) := \frac{P(x)-P(y)}{x-y}$ is absolutely irreducible unless $P(x)$ is decomposable, that is of the form $h(r(x))$ for some polynomials $h(x), r(x)$. See [DLS61, DS64, Sch85] for further discussion of this notion. Other examples come from the families $P(x) = ax^d + b$ and $P(x) = cT_d(x)$, where $T_d(x)$ is the Chebychev polynomial of the first kind. On the other hand, if $P(x) = h(r(x))$, then solutions $r(x_i) = r(x'_j)$ can be treated as "trivial" and $r(x_i) = r(x'_j)$ and $x_i = x'_j$ have approximately the same number of solutions.

The second question is concerned with the low–dimensional subvarieties of $\mathcal{V}_P$. Typically, such subvarieties contain the main mass of the solutions (see [Hea02]). Fortunately, we will be able to get away by considering just one–dimensional subvarieties and consequently we need to understand lines belonging to $\mathcal{V}_P$.

Let $\mathbb{F}$ be an algebraically closed field (later we apply the case $\mathbb{F} = \mathbb{C}$ only), $k \geq 2$ be an integer, $P(x) \in \mathbb{F}[x]$, $P(x) \neq \omega(x+c)^d$ for any $\omega, c \in \mathbb{F}$.

Let $\mathcal{Z}_P$ denote the set of distinct zeros of $P(x)$. Since $P(x) \neq \omega(x+c)^d$, it follows that $|\mathcal{Z}_P| > 1$. Let $l_1, \dots, l_k, l'_1, \dots, l'_k$ be non–vertical and non–horizontal linear transforms and suppose that

$$\{(l_1(t), \dots, l_k(t), l'_1(t), \dots, l'_k(t)) \ : \ t \in \mathbb{F}\} \subseteq \mathcal{V}_P. \tag{3.3}$$

We are interested in describing nontrivial families of lines $L = \{l_1, \dots, l_k\}$, $L' = \{l'_1, \dots, l'_k\}$, that is $L \neq L'$, satisfying (3.3).

Example 1. We call polynomial $P_\beta(x)$ a generalized even polynomial if $P_\beta(x) = g(x - \beta/2)$, where $g(x)$ is an even polynomial. Clearly $P_\beta(x) = P_\beta(\beta - x)$ which produces a large family of nontrivial lines. In our case, we confine ourselves with positive variables and thus such an obstacle could be easily treated.

Example 2. Let us turn our attention to the following construction from [Pra04, Section 18.2.2]. Let $d_0 = \deg P \geq 3$, $\alpha^{d_0} = 1$, $\alpha \neq \pm 1$ and let

$$P_{\alpha,\beta}(x) = a_0 \left(x + \frac{\beta}{\alpha - 1} \right)^{d_0} + c, \qquad \text{where } a_0 \in \mathbb{F} \setminus \{0\} \text{ and } c \in \mathbb{F}. \tag{3.4}$$

Then $P_{\alpha,\beta}(x) = P_{\alpha,\beta}(\alpha x + \beta)$ and upon taking different pair α_*, β_*, $\alpha_*^d = 1$ such that $\frac{\beta}{\alpha-1} = \frac{\beta_*}{\alpha_*-1}$, we obtain $P_{\alpha,\beta}(x) = P_{\alpha_*,\beta_*}(x)$. Consequently, for any $t \in \mathbb{F}$ one has

$$P_{\alpha,\beta}(t)P_{\alpha_*,\beta_*}(t) = P_{\alpha,\beta}(\alpha t+\beta)P_{\alpha_*,\beta_*}(\alpha_* t+\beta_*).$$

Moreover, one can further consider $P(x) := h(P_{\alpha,\beta}(x))$ for any $h \in \mathbb{F}[x]$ to increase the number of possible "nontrivial subvarieties".

To overcome the difficulties mentioned above, we formulate a simple finiteness result on lines (3.3) (see Lemma 3.1 below). We begin by introducing generalizations of polynomials (3.4). Let $\mathcal{I}$ be the family of all possible products of such polynomials,

$$\mathcal{I} := \{P_1P_2\cdots P_j : j \in \mathbb{N}, P_i \text{ satisfy (3.4) for all } 1 \le i \le j\}. \tag{3.5}$$

We observe that $r \in \mathcal{I}$ if and only if the set $\mathcal{Z}_r$ is a shift of a union of some concentric regular polygons. Notice that $\mathcal{I}$ contains all generalized even polynomials. We now consider polynomials of the form $a_0\prod_{j=1}^{d}(x+b_0-\rho^j)$ where ρ is *not* a root of unity and, more generally define the family

$$a_0\prod_{i=1}^{d_1}(x+b_1-\rho_1^i)\cdots\prod_{i=1}^{d_s}(x+b_s-\rho_s^i)\cdot r(x), \tag{3.6}$$

where $s \ge 0$ is an integer, $r \in \mathcal{I}$, $d_1+\cdots+d_s+\deg(r) = \deg(P)$, $a_0 \in \mathbb{F}\backslash\{0\}$, $b_1,\ldots,b_s \in \mathbb{F}$ and $\rho_1,\ldots,\rho_s \in \mathbb{F}$ are *not* roots of unity. One can check that for any polynomial P belonging to the family (3.6), the set $\mathcal{Z}_P$ consists of the union of at most s shifts of geometric progressions and a shift of a union of some concentric regular polygons. Finally, let $\mathcal{L}(\mathcal{Z}_P)$ be the set of all lines, generated by $\mathcal{Z}_P \times \mathcal{Z}_P$. Since an arbitrary line is determined by any two points of $\mathcal{Z}_P \times \mathcal{Z}_P$ we have that $|\mathcal{L}(\mathcal{Z}_P)| \le \binom{|\mathcal{Z}_P|^2}{2}$, and the line $l(t) = t$ always belongs to $\mathcal{L}(\mathcal{Z}_P)$.

Lemma 3.1. *Let $\mathbb{F}$ be an algebraically closed field, and $P(x) \in \mathbb{F}[x]$ such that $P(x) \neq \omega(x+c)^d$ for any $\omega, c \in \mathbb{F}$ and $d \ge 2$. Let $\mathcal{Z}_P$ be the set of distinct zeros of $P(x)$ in $\mathbb{F}$ with $|\mathcal{Z}_P| > k \ge 2$ and let the lines $l_1,\ldots,l_k,l'_1,\ldots,l'_k$ satisfy (3.3). Then,*

(1) *For any $i \in [k]$ there exists $j \in [k]$ such that $l'_i \circ l_j^{-1} \in \mathcal{L}(\mathcal{Z}_P)$.*
(2) *If $P \notin \mathcal{I}$ and $l'_i \neq l_j$, then there exists $j_* \in [k]$, $j_* \neq j$ such that the graph of $l'_i \circ l_{j_*}^{-1}$ intersects $\mathcal{Z}_P \times \mathcal{Z}_P$.*
(3) *Let s be an integer, $s \ge k-1$, $l'_i \neq l_j$ and suppose that $P(x)$ is not of the form (3.6). Then there exists $\tilde{j}_* \in [k]$, $\tilde{j}_* \neq j$ such that $l'_i \circ l_{\tilde{j}_*}^{-1} \in \mathcal{L}(\mathcal{Z}_P)$.*

Proof. From the definition of $\mathcal{V}_P$ it follows that for any $i \in [k]$, if $l'_i(t) \in \mathcal{Z}_P$ for some t, then there exists $j \in [k]$ with $l_j(t) \in \mathcal{Z}_P$. Hence $(l'_i \circ l_j^{-1})(z_1) = z_2$ for some $z_1, z_2 \in \mathcal{Z}_P$ and the graph of $l'_i \circ l_j^{-1}$ intersects $\mathcal{Z}_P \times \mathcal{Z}_P$. Since $|\mathcal{Z}_P| > k$, we have $|\mathcal{Z}_P|$ distinct zeros of P and $|\mathcal{Z}_P|$ pairs of lines (l'_i, l_j), $j \in [k]$. By the pigeonhole principle there is a line l_j^{-1} such that the graph of $l'_i \circ l_j^{-1}$ intersects $\mathcal{Z}_P \times \mathcal{Z}_P$ in at least two points and hence it belongs to $\mathcal{L}(\mathcal{Z}_P)$. This concludes the proof of 1).

To prove (2) and (3), we consider $l := l'_i \circ l_j^{-1}$ such that $l(t)$ is not an identical map. We begin by considering all affine transformations $x \to \alpha x + \beta$ fixing $\mathcal{Z}_P$, in other words, one has $\alpha \mathcal{Z}_P + \beta = \mathcal{Z}_P$. By shifting, we have $\alpha \mathcal{Z}'_P = \mathcal{Z}'_P$ for the set $\mathcal{Z}'_P := \mathcal{Z}_P - \beta/(\alpha - 1)$ and one can see that $\mathcal{Z}'_P$ is a geometric progression with step α. Since $|\mathcal{Z}'_P| = |\mathcal{Z}_P| > 1$, it follows that α is a root of unity and by shifting again, if necessary, we arrive at the conclusion that $P \in \mathcal{I}$. Since by our assumption, $P \notin \mathcal{I}$ (equivalently, there are no non-identical affine transformations fixing $\mathcal{Z}_P$), we see that $l(\mathcal{Z}_P) \neq \mathcal{Z}_P$ and thus there exists $j_* \in [k]$, $j_* \neq j$ such that the graph of $l'_i \circ l_{j_*}^{-1}$ intersects $\mathcal{Z}_P \times \mathcal{Z}_P$ and we have proved (2). Now more generally, we have seen that $\mathcal{Z}'_P$ is a union of α-invariant sets and a non-invariant part. Split the non-invariant part as a union of non-invariant geometric progressions with step α. Since $P(x)$ is not of the form (3.6), the number of such progressions must be at least $s+1$. Consequently, $|l(\mathcal{Z}_P) \cap \mathcal{Z}_P| < |\mathcal{Z}_P| - s$. By our assumption $s \geq k-1$ and applying the pigeonhole principle again we find $\tilde{j}_* \in [k]$, $\tilde{j}_* \neq j$ such that $l'_i \circ l_{\tilde{j}_*}^{-1} \in \mathcal{L}(\mathcal{Z}_P)$. This completes the proof. □

It will be convenient to formulate our energy results with variables constrained to certain arithmetic progressions. To this end, for positive numbers $q < N/2$ and non–negative $0 \leq a < q$ we let $[N]_{a,q}$ to denote the set of $x \in [N]$ such that $x \equiv a \pmod{q}$. In particular, for $q = 1$, $a = 0$ we have $[N]_{a,q} = [N]$. Let $\mathrm{d}(a) = 1$ for $a = 0$ and $\mathrm{d}(a) = 0$ otherwise.

Theorem 3.2. *Let $P(x) \in \mathbb{Z}[x]$ with $\deg(P) = d \geq 2$, and let N, $q < 2^{-1} N^{\frac{1}{2(1+\mathrm{d}(a))}}$ be positive integers, $0 \leq a < q$. If $P(x) \neq \omega(x+c)^d$ for any choice of $\omega \in \mathbb{Z}$, $c \in \mathbb{Q}$, then for $d > 2$*

$$\mathsf{E}^{\times}(P([N]_{a,q})) - \frac{2N^2}{q^2} \ll \frac{N^{2 - \frac{1}{2(2d-1)} + o_d(1)}}{q^{2 + \frac{\mathrm{d}(a)}{2d-1}}}, \tag{3.7}$$

and for $d = 2$ the following holds

$$\mathsf{E}^{\times}(P([N]_{a,q})) - \frac{2N^2}{q^2} \ll \frac{N^{5/3 + o_d(1)}}{q^{2 + \frac{\mathrm{d}(a)}{3}}}. \tag{3.8}$$

We remark that the terms $2N^2/q^2$ correspond to the diagonal solutions and thus Theorem 3.2 yields a power saving for the off-diagonal contribution. The main ingredient in our proof is the following celebrated result due to Bombieri and Pila [BP89, Theorem 5].

Lemma 3.3. (Bombieri–Pila). *Let $\mathcal{C}$ be an absolutely irreducible curve (over the rationals) with degree $d \geq 2$ and $N \geq \exp(d^6)$. Then the number of integral points on $\mathcal{C}$ and inside a square $[0, N] \times [0, N]$ does not exceed*

$$N^{\frac{1}{d}} \exp(12\sqrt{d \log N \log \log N}).$$

Having Lemmas 3.1 and 3.3 at our disposal, we are ready to obtain the main result of this section. The proof of our Theorem 3.2 involves a series of case considerations, so we split it into some steps to help the reader.

Proof of Theorem 3.2. Let $\tau := \max_{n\in[|P(N)|^2]} \tau(n) = N^{o_d(1)}$ where $\tau(n)$ is the number of divisors of n.

Preliminary reduction. We begin with the following simple observation. By our assumption $P(x) \neq \omega(x+c)^d$ for any $\omega \in \mathbb{Z}$, $c \in \mathbb{Q}$ and hence performing a rational change of variables one can assume that $P(x) = x^d + g(x)$, where $g \in \mathbb{Q}[x]$, $\lambda \neq 0$ is the leading coefficient of g and $\deg g = m \leq d-2$. Now our variable x runs over a rational shift of $[N]_{a,q}$, say, $\frac{u[N]_{a,q}+v}{w}$ with $u, v, w \in \mathbb{Z}$ and $u, w > 0$. We multiply $P(x)$ by w^d which clearly does not change the multiplicative energy and thus we may assume that $x \in u[N]_{a,q} + v$. Of course, after the multiplication our function $g(x)$ changes but with some abuse of the notation we use the same letter g for the obtained new function. Since $u[N]_{a,q} + v \subset [uN]_{au+v,uq}$ it suffices to estimate the off diagonal contribution in $\mathsf{E}^{\times}(P([uN]_{au+v,uq}))$.

Main argument. In view of the above, changing $q \to uq$ and $a \to au+v$ if necessary, we need to estimate the number of solutions $x, y, X, Y \in [N]_{a,q}$ to the equation

$$(x^d + g(x))(y^d + g(y)) = (X^d + g(X))(Y^d + g(Y)) \tag{3.9}$$

or, in other words,

$$\begin{aligned}(XY)^d - (xy)^d = {} & (x^d g(y) + y^d g(x) + g(x)g(y)) \\ & -(X^d g(Y) + Y^d g(X) + g(X)g(Y)).\end{aligned} \tag{3.10}$$

The choice $\{x, y\} = \{X, Y\}$ corresponds to $\frac{2N^2}{q^2} + O(N/q)$ solutions of the last equation and thanks to the condition $q < 2^{-1}N^{\frac{1}{2(1+\mathrm{d}(a))}}$ we see that the term $O(N/q)$ is negligible when compared to $N^{2-\frac{1}{2(2d-1)}+o_d(1)} \cdot q^{-2-\frac{\mathrm{d}(a)}{2d-1}}$. Now let $\Delta \leq N/q$ be a parameter to be chosen later. We may assume that all variables take the form $qk+a$, where $k \geq \Delta$. Indeed, the contribution of the case where this does not hold is at most $4\Delta\tau N/q$ solutions.

Main argument: introducing two new variables s and t. Let $s = XY - xy$ and $t = X + Y \in [2, 2N]$ and notice that s, $t - 2a$ are divisible by q (if $a = 0$, then s is divisible by q^2). If $s = 0$ and $g(x) = \lambda x^m$, $m \leq d-2$ (the case of constant g corresponds to $d = 2$), then we obtain just trivial solutions of our equation. Indeed, we have $xy = XY$ and Equation (3.9) implies $x^{d-m} + y^{d-m} = X^{d-m} + Y^{d-m}$. It follows that $\{x, y\} = \{X, Y\}$. Thus, without loss of generality we first assume that $s > 0$ and write

$$(XY)^d - (xy)^d = (xy+s)^d - (xy)^d = s\sum_{j=0}^{d-1}\binom{d}{j}(xy)^j s^{d-j-1}. \tag{3.11}$$

From (3.10) and the fact that $s > 0$ it follows that $|s|(\Delta q)^{2d-2} \ll N^{d+m} \le N^{2d-2}$ and hence

$$|s| \ll (N/\Delta q)^{2d-2}. \tag{3.12}$$

From the binomial formula we get

$$(\alpha+\beta)^n = \alpha^n + \beta^n + \sum_{j=1}^{n-1} \binom{n}{j} \alpha^j \beta^{n-j}$$

and by induction we can write symmetric polynomial as $\alpha^n + \beta^n = f_n(\alpha\beta, \alpha+\beta)$, where $f_n \in \mathbb{Z}[z,w]$, $f_n(z,w) = w^n + \tilde{f}_n(z,w)$, $\deg_z \tilde{f}_n = \deg_w \tilde{f}_n = n-2$, $n > 2$ and $f_2(z,w) = w^2 - 2z$ for $n = 2$. We need this information about $f_n(z,w)$ below. Now we now fix variables s, t and define

$$P_{s,t}(xy) := -(X^d g(Y) + Y^d g(X) + g(X)g(Y)) - ((XY)^d - (xy)^d).$$

Using (3.11) and writing $\lambda' = \lambda$ if $m = d-2$ and zero otherwise, we get

$$P_{s,t}(xy) = (2\lambda' - sd)(xy)^{d-1} + \tilde{P}_{s,t}(xy),$$

where $\deg \tilde{P}_{s,t} \le d-2$.

Hence (3.10), with s, t being fixed takes the form

$$\sigma := G(x,y) + P_{s,t}(xy) = 0, \tag{3.13}$$

where $G(x,y) := x^d g(y) + y^d g(x) + g(x)g(y)$.

Main argument: application of Bombieri–Pila in the case of absence of linear factors. We first consider the case that polynomial $G(x,y) + P_{s,t}(xy)$ has no linear factors over $\mathbb{C}$. If $G(x,y) + P_{s,t}(xy)$ is absolutely irreducible, then by Lemma 3.3 we have at most $N^{1/d+o(1)}$ solutions in x, y for $d > 2$. For $d = 2$ it is easy to check directly that the number of solutions is $N^{o(1)}$ (basically, it follows from the fact that any non–linear quadratic equation can be reduced either to a Pell's equation or to a hyperbolic equation and thus to the question about the upper bounds on the divisor function in $\mathbb{Z}$ or in $\mathbb{Z}[i]$). In general, considering absolutely irreducible factors of $G(x,y) + P_{s,t}(xy)$ and recalling that there are no linear factors, we apply Lemma 3.3 with $d = 2$ to bound the total number of solutions in s, t, x, y by

$$O\left(\frac{\Delta N \tau}{q} + \frac{N}{q} \cdot \frac{(N/\Delta q)^{2d-2}}{q^{1+\mathrm{d}(a)}} N^{1/2+o(1)}\right). \tag{3.14}$$

Indeed, the number of possible values of t is $O(N/q)$ and by definition of s we know that $q^{1+\mathrm{d}(a)}$ divides s. Combining (3.12) and choosing Δ to satisfy $\Delta^{2d-1} = N^{2d-2+1/2}/q^{2d-1+\mathrm{d}(a)}$, we obtain

$$O\left(\frac{N^{2-\frac{1}{2(2d-1)}+o_d(1)}}{q^{2+\frac{\mathrm{d}(a)}{2d-1}}}\right). \tag{3.15}$$

solutions.

Main argument: linear factors. We now turn to the case when polynomial $G(x,y)+P_{s,t}(xy)$ has linear factors over $\mathbb{C}$. In this case, for some $\alpha,\beta,\gamma\in\mathbb{C}$, one can write (3.13) as

$$\sigma=(\alpha x+\beta y+\gamma)F(x,y) \tag{3.16}$$

with (by "…" we denote lower order terms)

$$F(x,y)=\alpha^{-1}x^{d-1}g(y)+\beta^{-1}y^{d-1}g(x)+\dots, \tag{3.17}$$

where it is easy to check that the case $\alpha=0$ or $\beta=0$ is not possible. We claim that if $2\lambda'-sd\neq 0$, then $g(x)=\lambda x^m$, $m=d-2$, $\lambda'=\lambda$ and $\gamma=0$ (the case $s=0$, $g=\lambda x^m$ corresponds to the trivial solutions and was considered before). Indeed, since $\deg\tilde{P}_{s,t}\leq d-2$ and $\deg g\leq d-2$, from the definition of σ and (3.17) we get

$$\begin{aligned} &g(x)g(y)+(2\lambda'-sd)(xy)^{d-1}+\tilde{P}_{s,t}(xy)\\ &\quad=\alpha\beta^{-1}y^{d-1}xg(x)+\beta\alpha^{-1}x^{d-1}yg(y)\\ &\qquad+\gamma(\alpha^{-1}x^{d-1}g(y)+\beta^{-1}y^{d-1}g(x))+\dots \end{aligned} \tag{3.18}$$

and hence if $g(x)$ has lower order terms we would not be able to compensate $y^{d-1}xg(x)$, $x^{d-1}yg(y)$ via the left–hand side of (3.18). Without loss of generality, we may assume that our linear factor in (3.16) is $x-(\gamma-\beta y)$ (with some abuse of notation, we have changed the definition of β and γ). Consequently, we arrive at

$$X^2-tX-\beta y^2+\gamma y+s=0. \tag{3.19}$$

Linear factors I: Equation (3.19) *also has linear factors.* If the last equation has a linear factor we necessarily have $\gamma^2=\beta(t^2-4s)$ (it can be seen by computing discriminant of the conic (3.19)). If $\gamma=0$, then $t^2=4\,s$ and since $q<2^{-1}N^{\frac{1}{2(1+\mathrm{d}(a))}}$, we have that the number of solutions is $O(N/q\cdot(N/\Delta q)^{2d-2}N^{o_d(1)})$, giving negligible contribution. If $\gamma\neq 0$, then there are two possibilities: $s=0$ and $s=2\lambda'/d$. Without loss of generality assume that our linear factor is $X-(wy+r)$. Substituting the last equation into (3.19), we derive $w^2=\beta$ and $r(t-r)=s$. Thus we have four lines: $x=\gamma-\beta y$, $X=wy+r$, $Y=t-r-wy$ and $y=y$.

Suppose $|\mathcal{Z}_P|>2$ and $r\notin\{0,t\}$. Applying part (1) of Lemma 3.1 with $\mathbb{F}=\mathbb{C}$ and line $y=y$, we deduce that there is finite number of possibilities for w and for either r or $t-r$. Since $r(t-r)=s$, we obtain finite number of possibilities for t, provided $r\neq 0,t$. The latter is possible only if $s=0$. Since $y\in[N]_{q,a}$, we get $O(N/q)$ solutions.

Next we consider the case $|\mathcal{Z}_P|>2$ and $r\in\{0,t\}$. We observe that we must have $s=0$ and there is a finite number of possibilities for w. In the case $r=0$ (similar argument works for $r=t$) our lines are $x=\gamma-\beta y$, $X=wy$, $Y=t-wy$ and $y=y$. If $w=1$, then since $\gamma^2=\beta(t^2-4\,s)$ and $\beta=w^2$, we have $\beta=1$, $\gamma=t$ and we obtain just trivial solutions (the case $\gamma=-t$ is impossible). Assuming now that $w \neq 1$ (recall that $t \neq 0$ and hence $t - wy = Y \neq y$) we apply part (3) of Lemma 3.1 with $k = 2$, $s = 1$. This concludes the proof provided $P(x)$ is not of the form (3.6). Next, suppose that $P(x)$ is of the form (3.6). If, additionally $P(x) \in \mathcal{I}$, we recall that $t \neq 0$ and $\gamma \neq 0$, and therefore w must be a root of unity. Since our solutions are non-negative rationals we must have $w = 1$ and this case has already been considered. If $P(x) \notin \mathcal{I}$, then part (2) of Lemma 3.1 implies that the line $Y = t - wy$ determines a point in $\mathcal{Z}_P \times \mathcal{Z}_P$. Thus we have finite number of possibilities for t.

It remains to consider the case $|\mathcal{Z}_P| = 2$. In this case we have system of equations $r(t-r) = s$, $w^2 = \beta$, $\gamma^2 = \beta(t^2 - 4\,s)$ and two intersections of our lines with $\mathcal{Z}_P \times \mathcal{Z}_P$, namely, $wz_1 + r = z_2$, $t - r - wz_3 = z_4$ or $wz_1 + r = z_2$, $wz_3 + r = z_4$ or $t - r - wz_1 = z_2$, $t - r - wz_3 = z_4$, where z_i run over $\mathcal{Z}_P$. In either case, w and all other variables are determined in a unique way (there are $O(1)$ choices for γ). The case of permutations corresponds to $s = 0$, $r = 0, t$ and $\gamma = t$ which produces a finite number of lines.

Linear factors II: Equation (3.19) *has no linear factors.* To conclude the proof, we note that if (3.19) has no linear factors, then it has $N^{o_d(1)}$ solutions. The number of possibilities for s and t is $\frac{N}{q^{2+\mathrm{d}(a)}} \cdot (N/\Delta q)^{2d-2}$ and thus we obtain a bound which is better than (3.15). This completes the proof. □

Remark 3.4. The condition that $P(x) \neq \omega(x + c)^d$ for any pair $\omega \in \mathbb{Z}$, $c \in \mathbb{Q}$ is clearly necessary. If, say, $P(-x) = P(x)$, then we find non–trivial solutions of the form $(-z, w, z, w)$, where $z, w \in [-N, N]$ are *integers* (not necessary positive). Nevertheless, as one can see from the proof if P is a generalized even polynomial, then we have a similar asymptotic formula for $\mathsf{E}^{\times}(P([N]_{a,q}))$ albeit with a different main term corresponding to the "generalized" trivial solutions.

Remark 3.5. It is worth mentioning that our exponent 5/3 in (3.8) coincides with that of Hooley's from [Hoo96], where the author considers equation $x^d + y^d = X^d + Y^d$ with $x, y \in [N]$. More general binary forms are considered in [Hoo86]. A special (multiplicative) form of our Equation (3.9) makes the calculations simpler.

Remark 3.6. The argument of Theorem 3.2 is rather general and one can consider the common energy $\mathsf{E}^{\times}(P([N]), S, P([N]), S)$ and even the energy $\mathsf{E}^{\times}(S_1, S_2, S_3, S_4)$ for sufficiently large sets $S_i \subseteq [N]$. Furthermore, a similar argument works for the equation

$$P_1(x)P_1(y) = P_2(X)P_2(Y), \qquad x, y, X, Y \in [N]_{a,q}, \tag{3.20}$$

where $P_1, P_2 \in \mathbb{Z}[x]$, $\deg(P_1) = \deg(P_2) = d > 1$, $P_1(x), P_2(x) \neq \omega(x + c)^d$ for any $\omega \in \mathbb{Z}$, $c \in \mathbb{Q}$ and P_1, P_2 have the same leading coefficients. These observations will play an important role in our future work on the analogs of Theorems 1.1 and 1.3 for Rademacher random multiplicative functions.

4 Large fluctuations: proofs of Corollary 1.2 and Theorem 1.3

In order to prove Corollary 1.2 we first show that this is directly implied by the local version, that is Theorem 1.3, in the spirit of Harper's work [Har21]. We begin by recalling the first Borel–Cantelli lemma.

Lemma 4.1 (The first Borel–Cantelli Lemma). *Let $\{E_n\}_{n\geq 1}$ be any sequence of events. Then*

$$\sum_{n=1}^{+\infty} \mathbb{P}(E_n) < +\infty \implies \mathbb{P}(\limsup_{n\to+\infty} E_n) = 0,$$

where

$$\limsup_{n\to+\infty} E_n = \bigcap_{n=1}^{+\infty} \bigcup_{m=n}^{+\infty} E_m.$$

Proof of Corollary 1.2 assuming Theorem 1.3. Let $W(X) = (\log\log X)^{0.02}$. Theorem 1.3 implies that the probability that (1.4) fails is at most $O(1/W(X))$. Summing over a suitable sparse set of X-values, we can guarantee that the series of exceptional probabilities converges and thus by Lemma 4.1, almost surely, only finitely many X in the chosen sparse set make the events (1.4) fail. Consequently, there exists arbitrary large x, for which (1.3) holds. □

The rest of this section is devoted to the proof of Theorem 1.3. We recall the following result borrowed from the work of Maynard and Rudnick [MR21, Corollary 3.2] (in fact, they attribute this Corollary to Granville).

Lemma 4.2. *Let $P(x) \in \mathbb{Z}[x]$ be a polynomial with $d = \deg P \geq 2$ which is not a product of linear factors (over $\mathbb{Q}$). Then for a positive proportion of integers n,*

$$^+(P(n)) \geq \frac{1}{2d^2} \cdot n \log n.$$

We mention a simple and direct consequence of Lemma 4.2, needed for our applications. Let $\rho = \rho(d)$ be the positive proportion in the lemma above. Then for any x large enough, we have

$$\#\left\{ n \leq x : P^+(P(n)) \geq \frac{1}{2d^2} \cdot n \log n \right\} \geq \rho x/2.$$

This implies that

$$\#\left\{ \rho x/100 < n \leq x : P^+(P(n)) \geq \frac{1}{2d^2} n \log n \right\} \geq \rho x \left(\frac{1}{2} - \frac{1}{100} \right) \geq \rho x/3.$$

Since $\log(\rho x/100) \geq \frac{1}{2}\log x$ for large $x \geq 1$, it follows that

$$\#\left\{ \rho x/100 \leq n \leq x : P^+(P(n)) \geq \frac{\rho}{400d^2} x \log x \right\} \geq \rho x/3. \tag{4.1}$$

Next we collect two important probabilistic tools (see [Har13a, Har21, RR09, HL18]).

Lemma 4.3 (Normal approximation result). *Suppose that $m \geq 1$, and that $\mathcal{R}$ is a finite nonempty set. Suppose that for each $1 \leq i \leq m$ and $h \in \mathcal{R}$ we are given a deterministic coefficient $c(i,h) \in \mathbb{C}$. Finally, suppose that $(V_i)_{1 \leq i \leq m}$ is a sequence of independent, mean zero, complex-valued random variables, and let $Y = (Y_h)_{h \in \mathcal{R}}$ be the $\#\mathcal{R}$-dimensional random vector with components $Y_h := \Re(\sum_{i=1}^m c(i,h)V_i)$. If $Z = (Z_h)_{h \in \mathcal{R}}$ is a multivariate normal random vector with the same mean vector and covariance matrix as Y, then for any $u \in \mathbb{R}$ and any small $\eta > 0$ we have*

$$\mathbb{P}(\max_{h \in \mathcal{R}} Y_h \leq u) \leq \mathbb{P}(\max_{h \in \mathcal{R}} Z_h \leq u + \eta)$$
$$+ O\left(\frac{1}{\eta^2} \sum_{g,h \in \mathcal{R}} \sqrt{\sum_{i=1}^m |c(i,g)|^2 |c(i,h)|^2 \mathbb{E}[|V_i|^4]} + \frac{1}{\eta^3} \sum_{i=1}^m \mathbb{E}[|V_i|^3] (\sum_{h \in \mathcal{R}} |c(i,h)|)^3\right).$$

Lemma 4.4 (Normal comparison result). *Suppose that $n \geq 2$, and that $\varepsilon \geq 0$ is sufficiently small (i.e. less than a certain small absolute constant). Let $X_1, ..., X_n$ be mean zero, variance one, jointly normal random variables, and suppose $\mathbb{E}[X_i X_j] \leq \varepsilon$ whenever $i \neq j$. Then for any $100\varepsilon \leq \delta \leq 1/100$ (say), we have*

$$\mathbb{P}(\max_{1 \leq i \leq n} X_i \leq \sqrt{(2-\delta)\log n}) \leq e^{-\Theta(n^{\delta/20}/\sqrt{\log n})} + n^{-\delta^2/50\varepsilon}.$$

Proof of Theorem 1.3. We recall, that as in the Theorem 1.1, we may assume (2.1) holds, i.e. all polynomial values are distinct and positive. Let X be large and $x_i = X^{i(\log 3i)^2}$ for all $1 \leq i \leq \log X$ such that all points belong to $[X, X^{2\log X (\log\log X)^2}]$. We aim to show that with probability $1 - O(\frac{1}{W(X)})$ where $W(X) \to +\infty$ as $X \to +\infty$, one has

$$\max_{1 \leq i \leq \log X} \frac{|\sum_{n \leq x_i} f(P(n))|}{\sqrt{x_i \log\log x_i}} \gg 1, \tag{4.2}$$

where the implicit absolute constant is independent of X. To analyze (4.2), we use a conditioning argument. Instead of simply conditioning on small primes (as has been done before), we condition on all primes which are outside of the union of the following sets $\mathcal{A}_i$.

Step 1: construction of sets $\mathcal{A}_i$. Recall the definition of constant $\rho = \rho(d)$ in (4.1). We first define set $\mathcal{E}_i$:

$$\begin{aligned}\mathcal{E}_i &:= \mathcal{G}_i \backslash \mathcal{B}_i \\ &:= \left\{p \geq \frac{x_i \log x_i}{400 d^2 \rho^{-1}} : \exists\ n \leq x_i\ s.t.\ p | P(n)\right\} \\ &\quad \backslash \left\{p \geq \frac{x_i \log x_i}{400 d^2 \rho^{-1}} : \exists\ n \leq x_{i-1}\ s.t.\ p | P(n)\right\}.\end{aligned} \tag{4.3}$$

We claim that for large enough X,

$$\frac{\rho}{4d} x_i \leq |\mathcal{E}_i| \leq d x_i. \tag{4.4}$$

Indeed, we first apply (4.1) to conclude that there exists a constant $C > 0$ such that

$$|\mathcal{G}_i| \geq \rho x_i/3d.$$

This is because that for each $n \leq x_i$, the number of primes $\{p \geq \frac{\rho}{400d^2} x_i \log x_i : p|P(n)\}$ is bounded by $\leq d$ for sufficiently large $x_i \geq X$. The subtracted set $\mathcal{B}_i$ has cardinality at most $dx_{i-1} = o(x_i)$, yielding $|\mathcal{E}_i| \geq \frac{\rho}{4d} x_i$ for sufficiently large X. The other inequality is immediate as $|\mathcal{G}_i| \leq dx_i$.

We further pick a large subset $\mathcal{A}_i \subset \mathcal{E}_i$ such that no two distinct primes in $\mathcal{A}_i$ both divide $P(n)$ for some $n \leq x_i$. This is done using a greedy algorithm as explained below. For each $n \leq x_i$ and $p \in \mathcal{E}_i$, put

$$\mathcal{M}(n) := \{p \in \mathcal{E}_i : p|P(n)\} \quad \text{and} \quad \mathcal{N}(p) := \{n \leq x_i : p|P(n)\}.$$

We similarly define the set

$$\mathcal{M}(\mathcal{N}(p)) := \bigcup_{n \in \mathcal{N}(p)} \mathcal{M}(n).$$

We next greedily select elements to $\mathcal{A}_i \subset \mathcal{E}_i$ as follows. We first pick the smallest prime from $\mathcal{E}_i$ and label it as $p_1 \in \mathcal{A}_i$. Note that $P(n) = O(n^d)$ and for $n \leq x_i < p_1$ we have $|\mathcal{N}(p_1)| \leq d$ and $|\mathcal{M}(n)| \leq d$ for each $n \in \mathcal{N}(p_1)$. Consequently,

$$|\mathcal{M}(\mathcal{N}(p_1))| \leq d \cdot d = d^2.$$

We now pick $p_2 \in \mathcal{A}_i$ to be the smallest prime in

$$\mathcal{E}_i \backslash \mathcal{M}(\mathcal{N}(p_1)),$$

and repeat this iterative procedure. This produces a set $\mathcal{A}_i$ of size

$$|\mathcal{A}_i| \geq |\mathcal{E}_i|/d^2,$$

which together with (4.4) yields

$$\frac{x_i}{4\rho^{-1} d^3} < |\mathcal{A}_i| \leq dx_i.$$

In summary, we have chosen sets $\mathcal{A}_i$ of primes for $1 \leq i \leq \log X$ such that

(1) $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for all $i \neq j$.
(2) $|\mathcal{A}_i| \asymp_d x_i$.
(3) There does not exist $p, q \in \mathcal{A}_i$ with $p \neq q$ such that $pq|P(n)$ for some $n \leq x_i$.

Let

$$\mathcal{A} := \bigcup_{1 \leq i \leq \log X} \mathcal{A}_i. \tag{4.5}$$

Step 2: splitting the sum and treating negligible part. We split the sum into the following three pieces, depending on how many prime factors in $\mathcal{A}_i$ that $P(n)$ has. Let $\sum_{n\le x_i} f(P(n)) = S_{i,1} + S_{i,2} + S_{i,3}$ where

$$\begin{aligned} S_{i,1} &:= \sum_{p\in\mathcal{A}_i} \sum_{\substack{n\le x_i \\ p|P(n) \\ q\neq p: q|P(n) \implies q\notin\mathcal{A}}} f(P(n)), \\ S_{i,2} &:= \sum_{\substack{n\le x_i \\ \exists p\in\mathcal{A}\setminus\mathcal{A}_i \\ p|P(n)}} f(P(n)), \\ S_{i,3} &:= \sum_{\substack{n\le x_i \\ p|P(n) \implies p\notin\mathcal{A}}} f(P(n)). \end{aligned}$$

By the decomposition above, we have

$$\mathbb{P}\left(\max_{1\le i\le \log X} \frac{|\sum_{n\le x_i} f(P(n))|}{\sqrt{x_i \log\log x_i}} \gg 1\right) \ge \mathbb{P}(E_1\cap E_2) \ge \mathbb{P}(E_1) - (1-\mathbb{P}(E_2)) \quad (4.6)$$

where E_1 and E_2 are the events (with appropriately chosen absolute constants)

$$\begin{aligned} E_1 &:= \max_{1\le i\le \log X} \frac{|S_{i,1}+S_{i,3}|}{\sqrt{x_i \log\log x_i}} \gg 1, \\ E_2 &:= \max_{1\le i\le \log X} \frac{|S_{i,2}|}{\sqrt{x_i}(\log\log x_i)^{0.01}} \ll 1. \end{aligned}$$

Our plan now goes as follows: we first use union bounds to show that the event E_2 happens with probability close to 1. A more subtle task is to show that $\mathbb{P}(E_1)$ is close to 1. In order to do so, we use a conditioning argument. We first show that with probability close to 1 the sum $S_{i,3}$ is small (which only depends on $f(p)$ for $p \notin \mathcal{A}$). In particular, with probability close to 1, we can find a large random subset of indexes $\mathcal{R} \subset \{1, 2, \dots, \lfloor \log X \rfloor\}$ such that for those $i \in \mathcal{R}$ all the corresponding $S_{i,3}$ are small. We then condition on all $f(p)$ with $p \notin \mathcal{A}$ (now $\mathcal{R}$ is fixed) and thus the sum $S_{i,1}$ transforms into a sum of independent variables with certain weights. The latter puts us in the position of applying Lemma 4.4 to produce large fluctuations.

We first estimate $\mathbb{P}(E_2)$. Using the orthogonality together with (2.1), we have

$$\begin{aligned} \mathbb{E}[|S_{i,2}|^2] &= \#\{n \le x_i : \exists\, p \in \mathcal{A}\setminus\mathcal{A}_i \;\text{ s.t. }\; p|P(n)\} \\ &= \#\{n \le x_i : \exists\, p \in \bigcup_{1\le j\le i-1} \mathcal{A}_j \;\text{ s.t. }\; p|P(n)\}. \end{aligned} \quad (4.7)$$

The second equality above follows from the definition in (4.3). Recall that $x_j = X^{j(\log 3j)^2}$ and we can further bound the quantity (4.7),

$$\begin{aligned}
&\ll \sum_{p\in\bigcup_{1\le j\le i-1}\mathcal{A}_j}\left(\frac{x_i}{p}+1\right)\\
&\ll x_i\sum_{1\le j\le i-1}\sum_{p\in\mathcal{A}_j}\frac{1}{p}+\sum_{1\le j\le i-1}|\mathcal{A}_j|\\
&\ll x_i\sum_{1\le j\le i-1}\frac{1}{\log x_j}+\frac{x_i}{X}\\
&\ll \frac{x_i}{\log X}.
\end{aligned}$$

Using Markov's inequality, the event $|S_{i,2}| \gg \sqrt{x_i}(\log\log X)^{0.1}$ occurs with probability at most $O(1/\log X(\log\log X)^{0.2})$. Applying union bound for all $1 \le i \le \log X$ we get that with appropriately chosen implicit constants,

$$\mathbb{P}(E_2) = \mathbb{P}\left(\max_{1\le i\le \log X}\frac{|S_{i,2}|}{\sqrt{x_i}(\log\log X)^{0.1}} \ll 1\right) = 1 - O(1/(\log\log X)^{0.2}). \quad (4.8)$$

Step 3: creating large fluctuations. We next estimate $\mathbb{P}(E_1)$. To deal with $S_{i,3}$ for $1 \le i \le \log X$, we show that with probability $1 - O(\frac{1}{(\log\log X)^{0.02}})$, there exists a large random subset $\mathcal{R} \subset \{1,2,3,\cdots,\lfloor\log X\rfloor\}$ with $|\mathcal{R}| \ge 0.99\log X$ such that for every $i \in \mathcal{R}$,

$$|S_{i,3}| = |\sum_{\substack{n\le x_i\\ p|P(n)\implies p\notin\mathcal{A}}} f(P(n))| \ll \sqrt{x_i}(\log\log x_i)^{0.01}. \quad (4.9)$$

Indeed, using the second-moment estimate and Markov's inequality, the expected number of points x_i with $1 \le i \le \log X$ for which (4.9) fails is

$$\mathbb{E}[\#\{i \le \log X : (4.9)\text{ fails}\}] \ll \frac{\log X}{(\log\log X)^{0.02}},$$

and the claim follows. By our construction, each of the sums $S_{i,3}$ is independent of the values $f(p)$ for $p \in \mathcal{A}$. From now on, we condition on all variables $f(p)$ with $p \notin \mathcal{A}$ and thus the set $\mathcal{R}$ is fixed. We aim to apply Lemma 4.3 to understand the maximum of $S_{i,1}$ over $i \in \mathcal{R}$. Since by our construction all $\mathcal{A}_i$ are disjoint for different choices of i, we crucially have that $S_{i,1}$'s are independent. We next apply Lemma 4.3. The independent random variables V_j here are indexed by the primes and

$$V_p := \frac{1}{\sqrt{x_k}}\sum_{\substack{n\le x_k\\ p|P(n)\\ q\ne p:q|P(n)\implies q\notin\mathcal{A}}} f(P(n)), \quad (4.10)$$

where $k = k(p)$ is uniquely determined by $p \in \mathcal{A}_k$. Let $\tilde{\mathbb{P}}$ denote the conditional probability and let $\tilde{\mathbb{E}}$ be the conditional expectation (conditioned on all values of $f(p)$ with $p \notin \mathcal{A}$). Lemma 4.3 implies that for any $u \in \mathbb{R}$ and small $\eta > 0$, we have

$$\tilde{\mathbb{P}}\left(\max_{i\in\mathcal{R}} \mathfrak{Re}\frac{S_{i,1}}{\sqrt{x_i}} \leq u\right) \leq \mathbb{P}(\max_{i\in\mathcal{R}} Z_i \leq u+\eta) \\ + O\left(\frac{1}{\eta^2}\sum_{i,j\in\mathcal{R}}\sqrt{\sum_{p\in\mathcal{A}_i\cap\mathcal{A}_j}\tilde{\mathbb{E}}[|V_p|^4]}\right) + O\left(\frac{1}{\eta^3}\sum_{\substack{i\in\mathcal{R}\\ p\in\mathcal{A}_i}}\tilde{\mathbb{E}}[|V_p|^3]|\mathcal{R}|^3\right), \tag{4.11}$$

where Z_i are jointly normal random variables with mean zero

$$\mathbb{E}[Z_i] := \tilde{\mathbb{E}}\left[\mathfrak{R}\frac{S_{i,1}}{\sqrt{x_i}}\right] = 0.$$

We define

$$[a]_t := \text{the largest factor of } a \text{ that is coprime to } t.$$

e.g. $[100]_5 = 4$. For every $i \in \mathcal{R}$, the variance is

$$\mathbb{E}[Z_i^2] := \tilde{\mathbb{E}}[(\mathfrak{R}S_{i,1})^2] = \frac{1}{2x_i}\sum_{p\in\mathcal{A}_i}\sum_{k\geq 1}\left|\sum_{\substack{n\leq x_i\\ p^k||P(n)\\ q\neq p: q|P(n)\implies q\notin\mathcal{A}}} f([P(n)]_p)\right|^2.$$

Since $\mathcal{A}_i \cap \mathcal{A}_j$ is empty unless $i = j$, the first "big Oh" term simplifies to

$$O\left(\frac{1}{\eta^2}\sum_{i\in\mathcal{R}}\sqrt{\sum_{p\in\mathcal{A}_i}\tilde{\mathbb{E}}[|V_p|^4]}\right). \tag{4.12}$$

Notice that for each fixed $p \in \mathcal{A}_i$, the number of $n \leq x_i$ such that $p|P(n)$ is $\leq d$ (since $p \gg x_i \log x_i$). Consequently, for $p \in \mathcal{A}_i$ we have a trivial pointwise bound

$$|V_p| = \left|\frac{1}{\sqrt{x_i}}\sum_{\substack{n\leq x_i\\ p|P(n)\\ q\neq p: q|P(n)\implies q\notin\mathcal{A}}} f(P(n))\right| \ll_d \frac{1}{\sqrt{x_i}}. \tag{4.13}$$

Plugging (4.13) into (4.11), and noticing that $x_1 \geq X$, the error terms in (4.11) are at most

$$\ll \frac{1}{\eta^2}\frac{|\mathcal{R}|}{X} + \frac{1}{\eta^3}\frac{|\mathcal{R}|^4}{\sqrt{X}} \ll \eta^{-3}X^{-1/2+\varepsilon},$$

for any given $\varepsilon > 0$. Thus we have

$$\tilde{\mathbb{P}}\left(\max_{i\in\mathcal{R}} \mathfrak{Re}\frac{S_{i,1}}{\sqrt{x_i}} \le u\right) \le \mathbb{P}\left(\max_{i\in\mathcal{R}} Z_i \le u+\eta\right) + O(\eta^{-3}X^{-1/2+\varepsilon}). \tag{4.14}$$

Step 4: analyzing Gaussian model. From now on, we only need to focus on Z_i with $i \in \mathcal{R}$ and we aim to bound the probability $\mathbb{P}(\max_{i\in\mathcal{R}} Z_i \le u+\eta)$ for appropriately chosen u, η. To this end, we first show that there exist constants $m, c > 0$, such that with probability (over the realizations of $f(p)$ for $p \notin \mathcal{A}$) at least $1 - O(\frac{1}{X^c})$ one has $\min_{i\in\mathcal{R}} \mathbb{E}[Z_i^2] \ge m$. Then we will apply Lemma 4.4 to establish the estimate.

We use the following notation

$$\mathcal{T}_{i,p} := \{n \le x_i : p|P(n), \text{ and there does not exist } q \in \mathcal{A} \text{ such that } q \ne p \text{ and } q|P(n)\},$$

and

$$\mathcal{T}_{i,p,k} := \{n \in \mathcal{T}_{i,p} : p^k || P(n)\}.$$

Over all realizations of $f(p)$ with $p \notin \mathcal{A}$ which we conditioned on before, the expected value of $\mathbb{E}[Z_i^2]$ is

$$\begin{aligned}\mu_i := \mathbb{E}[\mathbb{E}[Z_i^2]] &= \frac{1}{2x_i}\sum_{p\in\mathcal{A}_i}\sum_{k\ge1} \#\{(m,n)\in\mathcal{T}_{i,p,k}^2 : [P(m)]_p = [P(n)]_p\} \\ &= \frac{1}{2x_i}\sum_{p\in\mathcal{A}_i}\sum_{k\ge1} \#\{(m,n)\in\mathcal{T}_{i,p,k}^2 : P(m) = P(n)\}.\end{aligned} \tag{4.15}$$

The second equality follows from the definition of $\mathcal{T}_{i,p,k}$. Since polynomial values $P(n)$ are all distinct, we further have that

$$\mu_i = \frac{1}{2x_i}\sum_{p\in\mathcal{A}_i} \#\{n \in \mathcal{T}_{i,p}\} \gg_d 1, \tag{4.16}$$

where the last inequality follows from the definition of $\mathcal{A}_i$. Indeed the number of $n \le x_i$ which have prime factors in $\mathcal{A}_i$ is $\gg_d x_i$ and those n for which $P(n)$ also has some prime factor $q \in \cup_{1\le k\le i-1}\mathcal{A}_k$ is $o(x_i)$ (see the computation in (4.7)).

Our final ingredient is the following concentration result, which essentially follows from the energy estimates proved in Section 3. We have

$$\begin{aligned}&\mathbb{E}[(\mathbb{E}[Z_i^2])^2] \\ &\quad= \frac{1}{4x_i^2}\sum_{\substack{p_1,p_2\in\mathcal{A}_i\\ k_1,k_2\ge1}} \#\{(m_1,n_1,m_2,n_2)\in\mathcal{T}_{i,p_1,k_1}^2\times\mathcal{T}_{i,p_2,k_2}^2 \\ &\qquad : [P(m_1)]_{p_1}[P(m_2)]_{p_2} = [P(n_1)]_{p_1}[P(n_2)]_{p_2}\} \\ &\quad= \frac{1}{4x_i^2}\sum_{\substack{p_1,p_2\in\mathcal{A}_i\\ k_1,k_2\ge1}} \#\{(m_1,n_1,m_2,n_2)\in\mathcal{T}_{i,p_1,k_1}^2\times\mathcal{T}_{i,p_2,k_2}^2 : P(m_1)P(m_2) = P(n_1)P(n_2)\}.\end{aligned} \tag{4.17}$$

The second equality above follows from the definition of $\mathcal{T}_{i,p,k}$. Now we are ready to compute the variance of $\mathbb{E}[Z_i^2]$:

$$\begin{aligned}\mathbb{E}[(\mathbb{E}[Z_i^2]-\mu_i)^2] &= \mathbb{E}[(\mathbb{E}[Z_i^2])^2]-\mu_i^2 \\ &= \frac{1}{4x_i^2}\sum_{\substack{p_1,p_2\in\mathcal{A}_i\\ k_1,k_2\geq 1}} \#\{(m_1,n_1,m_2,n_2)\in\mathcal{T}_{i,p_1,k_1}^2\times\mathcal{T}_{i,p_2,k_2}^2 : P(m_1)P(m_2) \\ &= P(n_1)P(n_2)\} - \left(\frac{1}{2x_i}\sum_{p\in\mathcal{A}_i}\sum_{k\geq 1}\#\{(m,n)\in\mathcal{T}_{i,p,k}^2 : P(m)=P(n)\}\right)^2.\end{aligned} \tag{4.18}$$

We analyze the difference in (4.18) as follows. Firstly, for the case $p_1=p_2=p$, as $|\mathcal{T}_{i,p}|\ll_d 1$, we know that the number of such quadruples (m_1,m_2,n_1,n_2) is at most $\ll_d |\mathcal{A}_i|\ll x_i$ and thus contributes $\ll_d \frac{1}{x_i}$ (which is negligible) after normalized by $1/4x_i^2$ in the above difference. For the general case $p_1\neq p_2$, the contribution to the difference in (4.18) is $\ll\frac{1}{4x_i^2}$ times the number of quadruples $(m_1,n_1,m_2,n_2)\in[N]^4$ with

$$P(m_1)P(m_2)=P(n_1)P(n_2)\quad\text{but}\quad\{P(m_1),P(m_2)\}\neq\{P(n_2),P(n_2)\}.$$

We now invoke the result of Proposition 1.4 to conclude that there exists a constant $c'>0$ such that the contribution in this case to the difference in (4.18) is $\ll\frac{1}{x_i^{c'}}$. Combining the above discussions together, we have that the quantity in (4.18) is $\ll\frac{1}{x_i^{c'}}$. By using Chebyshev's inequality, it follows that the exceptional probability, i.e.

$$\mathbb{P}(\mathbb{E}[Z_i^2]<\mu_i/2)=O(1/x_i^{c'}).$$

Write

$$V_i:=\min_{i\in\mathcal{R}}\mathbb{E}[Z_i^2].$$

Taking union bounds over all $i\in\mathcal{R}$ (here $|\mathcal{R}|\leq\log X$) and using (4.16), we conclude that there exist positive constants c and m such that over all realizations of $(f(p))$ that have been conditioned on, with probability at least $1-O\left(\frac{1}{X^c}\right)$,

$$V_i\geq m>0. \tag{4.19}$$

For any $u\in\mathbb{R}$ and small η, we have, by the definition of V_i,

$$\mathbb{P}(\max_{i\in\mathcal{R}}Z_i\leq u+\eta)\leq\mathbb{P}\left(\max_{i\in\mathcal{R}}\frac{Z_i}{\sqrt{\mathbb{E}[Z_i^2]}}\leq\frac{u+\eta}{\sqrt{V_i}}\right). \tag{4.20}$$

Since $\mathcal{A}_i$'s are disjoint, we have

$$\mathbb{E}[Z_iZ_j]=0,\qquad\text{if } i\neq j. \tag{4.21}$$

Let $u=\sqrt{m\log\log X}$. We apply Lemma 4.4 and (4.19) with $\delta=1/100$ and $\varepsilon=1/X$ to get that the right hand side of (4.20) is

$$\le \mathbb{P}\left(\max_{i\in\mathcal{R}}\frac{Z_i}{\sqrt{\mathbb{E}[Z_i^2]}}\le\frac{\sqrt{m\log\log X}+\eta}{\sqrt{m}}\right)\ll e^{-\Theta((\log X)^{1/3000})}. \tag{4.22}$$

Plugging (4.20) and (4.22) into (4.14) and choosing η to be a fixed constant, we derive

$$\tilde{\mathbb{P}}\left(\max_{i\in\mathcal{R}}\mathfrak{Re}\frac{S_{i,1}}{\sqrt{x_i}}\le\sqrt{m\log\log X}\right)\ll e^{-\Theta((\log X)^{1/3000})}.$$

Since $\log\log x_i\ll\log\log X^{i(\log 3i)^2}\ll\log\log X$ for all $i\le\log X$, the latter inequality can be rewritten as

$$\tilde{\mathbb{P}}\left(\max_{i\in\mathcal{R}}\mathfrak{Re}\frac{S_{i,1}}{\sqrt{x_i\log\log x_i}}\gg 1\right)\ge 1-O\left(e^{-\Theta((\log X)^{1/3000})}\right), \tag{4.23}$$

for an appropriately chosen small absolute constant. Since the probability of existence of $\mathcal{R}$ satisfying (4.9) is at least $1-O(\frac{1}{(\log\log x)^{0.02}})$ and (4.19) holds with probability $1-O(\frac{1}{x^c})$, we combine these with the error term in (4.23) to arrive at the estimate

$$\mathbb{P}(E_1)\ge 1-O\left(\frac{1}{(\log\log x)^{0.02}}\right). \tag{4.24}$$

Inserting (4.8), (4.24) into (4.6) concludes the proof. □

Acknowledgments

The authors are very grateful to Adam Harper and Alexander Mangerel for the insightful comments on an earlier draft of this manuscript. We would like to warmly thank anonymous referees for a careful reading of the paper and for the numerous very helpful comments and corrections. O.K. and I.S. would like to thank Max Planck Institute for Mathematics (Bonn) for providing excellent working conditions and support. O.K. would also like to express his gratitude to Mittag-Leffler Institute for mathematical research for providing stimulating working environment and support. I.S. was supported by the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-02-2022-872). M.W.X. is supported by the Cuthbert C. Hurd Graduate Fellowship in the Mathematical Sciences (Stanford). O.K. and I.S. are grateful for the FRG support of the Heilbronn Institute for Mathematical Research.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

[Bas12] J. Basquin. Sommes friables de fonctions multiplicatives aléatoires. *Acta Arith.*, (3)152 (2012), 243–266

[BP89] E. Bombieri and J. Pila. The number of integral points on arcs and ovals. *Duke Math. J.*, (2)59 (1989), 337–357

[BCG13] P. Borwein, S.K.K. Choi, and H. Ganguli. Sign changes of the Liouville function on quadratics. *Canad. Math. Bull.*, 56(2) (2013), 251–257

[CFM00] J. Cassaigne, S. Ferenczi, C. Mauduit, J. Rivat, and A. Sárközy. On finite pseudorandom binary sequences. IV. The Liouville function. II. *Acta Arith.*, (4)95 (2000), 343–359

[CS12] S. Chatterjee and K. Soundararajan. Random multiplicative functions in short intervals. *Int. Math. Res. Not. IMRN*, (3) (2012), 479–492

[Cho65] S. Chowla. *The Riemann Hypothesis and Hilbert's Tenth Problem. Mathematics and Its Applications*, Vol. 4. Gordon and Breach Science Publishers, New York-London-Paris, (1965).

[DLS61] H. Davenport, D.J. Lewis, and A. Schinzel. Equations of the form $f(x) = g(y)$. *Quart. J. Math. Oxford Ser. (2)*, 12 (1961), 304–312

[DS64] H. Davenport and A. Schinzel. Two problems concerning polynomials. *J. Reine Angew. Math.*, (215)214 (1964), 386–391

[Ell92] P.D.T.A. Elliott. On the correlation of multiplicative functions. *Notas Soc. Mat. Chile*, (1)11 (1992), 1–11

[Erd85] P. Erdős. Some applications of probability methods to number theory. In: *Mathematical Statistics and Applications, Vol. B (Bad Tatzmannsdorf, 1983)*, pp. 1–18. Reidel, Dordrecht, (1985).

[Fri70] M. Fried. On a conjecture of Schur. *Michigan Math. J.*, 17 (1970), 41–55

[Gra98] A. Granville. *ABC* allows us to count squarefrees. *Internat. Math. Res. Notices*, (19) (1998), 991–1009

[Gra08] A. Granville. Smooth numbers: computational number theory and beyond. In: *Algorithmic Number Theory: Lattices, Number Fields, Curves and Cryptography, volume 44 of Math. Sci. Res. Inst. Publ.*, pp. 267–323. Cambridge University Press, Cambridge, (2008).

[GS01] A. Granville and K. Soundararajan. Large character sums. *J. Amer. Math. Soc.*, (2)14 (2001), 365–397

[Hal83] G. Halász. On random multiplicative functions. In: *Hubert Delange colloquium (Orsay, 1982), Volume 83 of Publ. Math. Orsay*, pp. 74–96. Univ. Paris XI, Orsay, (1983).

[Har13] A.J. Harper. Bounds on the suprema of Gaussian processes, and omega results for the sum of a random multiplicative function. *Ann. Appl. Probab.*, (2)23 (2013), 584–616

[Har13a] A.J. Harper. A note on the maximum of the Riemann zeta function, and log-correlated random variables. *arXiv e-prints*, page arXiv:1304.0677, (2013a).

[Har13b] A.J. Harper. On the limit distributions of some sums of a random multiplicative function. *J. Reine Angew. Math.*, 678 (2013b), 95–124

[Har20] A.J. Harper. Moments of random multiplicative functions, I: Low moments, better than squareroot cancellation, and critical multiplicative chaos. *Forum Math. Pi*, 8 (2020), e1, 95

[Har21] A.J. Harper. Almost Sure Large Fluctuations of Random Multiplicative Functions. *International Mathematics Research Notices*, 11 (2021), rnab299.

[HL18] A.J. Harper and Y. Lamzouri. Orderings of weakly correlated random variables, and prime number races with many contestants. *Probab. Theory Related Fields*, (3-4)170 (2018), 961–1010

[Hea02] D.R. Heath-Brown. The density of rational points on curves and surfaces. *Ann. of Math. (2)*, (2)155 (2002), 553–595

[HR21] H.A. Helfgott and M. Radziwiłł. Expansion, divisibility and parity. *arXiv e-prints*, page arXiv:2103.06853, (2021).

[Hoo86] C. Hooley. On binary quartic forms. *J. Reine Angew. Math.*, 366 (1986), 32–52

[Hoo96] C. Hooley. On another sieve method and the numbers that are a sum of two hth powers. II. *J. Reine Angew. Math.*, 475 (1996), 55–75

[Hou11] B. Hough. Summation of a random multiplicative function on numbers having few prime factors. *Math. Proc. Cambridge Philos. Soc.*, (2)150 (2011), 193–214

[LTW13] Y.-K. Lau, G. Tenenbaum, and J. Wu. On mean values of random multiplicative functions. *Proc. Amer. Math. Soc.*, (2)141 (2013), 409–420

[Mas22] D. Mastrostefano. An almost sure upper bound for random multiplicative functions on integers with a large prime factor. *Electron. J. Probab.*, 27 (2022), Paper No. 32, 1–21

[MRT15] K. Matomäki, M. Radziwiłł, and T. Tao. An averaged form of Chowla's conjecture. *Algebra Number Theory*, (9)9 (2015), 2167–2196

[MR21] J. Maynard and Z. Rudnick. A lower bound on the least common multiple of polynomial sequences. *Riv. Math. Univ. Parma (N.S.)*, (1)12 (2021), 143–150

[McL74] D.L. McLeish. Dependent central limit theorems and invariance principles. *Ann. Probability*, 2 (1974), 620–628

[Naj20] J. Najnudel. On consecutive values of random completely multiplicative functions. *Electron. J. Probab.*, 25 (2020), Paper No. 59, 28

[Ng04] N. Ng. The distribution of the summatory function of the Möbius function. *Proc. London Math. Soc. (3)*, (2)89 (2004) 361–389

[Pra04] V.V. Prasolov. *Polynomials, volume 11 of Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, (2004). Translated from the 2001 Russian second edition by Dimitry Leites.

[RR09] G. Reinert and A. Röllin. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.*, (6)37 (2009), 2150–2173

[Sch85] A. Schinzel. Reducibility of polynomials in several variables. II. *Pacific J. Math.*, (2)118 (1985), 531–563

[SX22] K. Soundararajan and M.W. Xu. Central limit theorems for random multiplicative functions. *arXiv e-prints*, page arXiv:2212.06098, (2022).

[Tao16] T. Tao. The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. *Forum Math. Pi*, 4 (2016) e8, 36

[TT18] T. Tao and J. Teräväinen. Odd order cases of the logarithmically averaged Chowla conjecture. *J. Théor. Nombres Bordeaux*, (3)30 (2018), 997–1015

[TV06] T. Tao and V. Vu. *Additive combinatorics, Volume 105 of Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, (2006).

[Tao16] T. Tao. The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. In: *Forum of Mathematics, Pi, Volume 4*. University Press, (2016).

[Ter20] J. Teräväinen. On the Liouville function at polynomial arguments. *arXiv e-prints*, page arXiv:2010.07924, (2020).

[WX22] V.Y. Wang and M.W. Xu. Paucity phenomena for polynomial products (2022). arXiv:2211.02908.

[Win44] A. Wintner. Random factorizations and Riemann's hypothesis. *Duke Math. J.*, 11 (1944), 267–275

O. Klurman
School of Mathematics, University of Bristol, Bristol, UK.
lklurman@gmail.com

I. D. Shkredov
London Institute for Mathematical Sciences, 21 Albemarle St., London, UK.
ilya.shkredov@gmail.com
and
IITP RAS, Bolshoy Karetny per. 19, Moscow, Russia.
and
Institute of Physics, Mathematics and Information Technology, Immanuel Kant Baltic Federal University, Kaliningrad, Russia.

M. W. Xu
Department of Mathematics, Stanford University, Stanford, CA, USA.
maxxu@stanford.edu

Received: March 29, 2022
Revised: December 15, 2022
Accepted: January 23, 2023