CENTRAL LIMIT THEOREMS FOR RANDOM MULTIPLICATIVE FUNCTIONS

KANNAN SOUNDARARAJAN AND MAX WENQIANG XU

To Peter Sarnak on the occasion of his seventieth birthday

Abstract. A Steinhaus random multiplicative function f is a completely multiplicative function obtained by setting its values on primes $f(p)$ to be independent random variables distributed uniformly on the unit circle. Recent work of Harper shows that $\sum_{n\le N} f(n)$ exhibits "more than square-root cancellation," and in particular $\frac{1}{\sqrt{N}}\sum_{n\le N} f(n)$ does not have a (complex) Gaussian distribution. This paper studies $\sum_{n\in\mathcal{A}} f(n)$, where $\mathcal{A}$ is a subset of the integers in $[1, N]$, and produces several new examples of sets $\mathcal{A}$ where a central limit theorem can be established. We also consider more general sums such as $\sum_{n\le N} f(n)e^{2\pi in\theta}$, where we show that a central limit theorem holds for any irrational θ that does not have extremely good Diophantine approximations.

1. Introduction

In recent years there has been a lot of progress in understanding the behavior of random multiplicative functions. One motivation for studying such functions is that understanding these may help shed light on functions of interest in number theory such as Dirichlet characters or the Liouville and Möbius functions. Two natural models for random multiplicative functions are (1) the Steinhaus model of a random completely multiplicative function $f(n)$ where the values $f(p)$ (on primes p) are chosen independently and uniformly from the unit circle, and (2) the Rademacher model of a multiplicative function $f(n)$ taking values ± 1 on square-free integers and 0 on integers having a square factor, with $f(p)$ chosen independently and uniformly from $\{-1, 1\}$.

A fundamental question is to understand the distribution of the partial sums $\sum_{n\le N} f(n)$ for random multiplicative functions f (either in the Steinhaus case or in the Rademacher case). Since the values of f at integers satisfy dependency relations, it is a challenging problem to understand this distribution. A breakthrough result of Harper [8] established that typically $\sum_{n\le N} f(n)$ is $o(\sqrt{N})$. Note that $\sqrt{N}$ is the size of the standard deviation of $\sum_{n\le N} f(n)$, and thus Harper's result (which confirmed a conjecture of Helson [9]) exhibits "more than square-root cancellation" in such partial sums.

One of our goals in this paper is to explore the distribution of partial sums of random multiplicative functions when restricted to subsets $\mathcal{A}$ of $[1, N]$. We shall give criteria and several examples of sets $\mathcal{A}$ where such partial sums satisfy a central limit theorem. For simplicity, we describe our results in the Steinhaus setting, and sketch briefly (in Section 9)

corresponding results in the Rademacher case. We begin with a simplified criterion for $\mathcal{A}$ where a central limit theorem holds (see Theorem 3.1 for a more precise, but more technical, result).

Theorem 1.1. *Let N be large, and let $\mathcal{A}$ be a subset of $[1,N]$ with size*

$$|\mathcal{A}| \geq N\exp(-\tfrac{1}{3}\sqrt{\log N \log\log N}). \tag{1.1}$$

Suppose that there exists a subset $\mathcal{S}\subset\mathcal{A}$ with size $|\mathcal{S}| = (1+o(1))|\mathcal{A}|$ satisfying the following criterion:

$$\#\{(s_1,s_2,s_3,s_4)\in\mathcal{S}^4:\quad s_1s_2 = s_3s_4\} = (2+o(1))|\mathcal{S}|^2. \tag{1.2}$$

Then, as f ranges over random multiplicative functions in the Steinhaus model, the quantity

$$\frac{1}{\sqrt{|\mathcal{A}|}}\sum_{n\in\mathcal{A}} f(n)$$

is distributed like a standard complex normal random variable with mean 0 and variance 1.

Here and below, when we say "distributed like" we mean convergence in distribution as the parameter N tends to infinity. Recall that a real normal random variable W with mean 0 and variance σ^2 is given by

$$\mathbb{P}(W\leq t) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{x^2}{2\sigma^2}}dx.$$

A standard complex normal random variable Z with mean 0 and variance 1 is given by $Z = X+iY$ where X and Y are two independent real normal random variables with mean 0 and variance $\frac{1}{2}$.

In Theorem 1.1 the condition (1.1) is very mild and usually we are interested in much larger sets $\mathcal{A}$. It can also be weakened further, as in our more precise version Corollary 3.2 below. The criterion in (1.2) is more important, and may be viewed as a fourth moment condition. The quantity in (1.2) may be thought of as the *multiplicative energy* of the set $\mathcal{S}$, defined by

$$E_\times(\mathcal{S}) = \#\{(s_1,s_2,s_3,s_4)\in\mathcal{S}^4:\quad s_1s_2 = s_3s_4\}. \tag{1.3}$$

Since there are always trivial solutions $s_1 = s_3$ and $s_2 = s_4$ (or $s_1 = s_4$ and $s_2 = s_3$), we see that $E_\times(\mathcal{S})\geq 2|\mathcal{S}|(|\mathcal{S}|-1)+|\mathcal{S}| = 2|\mathcal{S}|^2-|\mathcal{S}|$. Thus the condition (1.2) asks for the multiplicative energy to be as small as it possibly can be. We also point out the flexibility in choosing a subset $\mathcal{S}$ of $\mathcal{A}$ which can be quite useful because even if $|\mathcal{S}|\sim|\mathcal{A}|$ it can happen that $E_\times(\mathcal{A})$ is much bigger than $E_\times(\mathcal{S})$.

There are several results in the literature establishing a central limit theorem for random multiplicative functions restricted to suitable subsets $\mathcal{A}$ of $[1,N]$. For instance Hough [12] considered the set $\mathcal{A}$ of integers with exactly k prime factors for any fixed k, and Harper [7] extended this to allow any $k = o(\log\log N)$. Short intervals $[x,x+y]$ were considered in the work of Chatterjee and Soundararajan [3] (to be precise, they worked in the Rademacher setting) who showed that a central limit theorem holds provided $y = o(x/\log x)$ (and with

a technical condition that $y \geq x^{\frac{1}{5}} \log x$ should be suitably large, so that $[x, x+y]$ contains many square-free integers). Recent work of Klurman, Shkredov and Xu [13] considers the set of polynomial values $\{P(n) : 1 \leq n \leq N\}$ where $P(x) \in \mathbb{Z}[x]$ is a polynomial not of the form $w(x+c)^d$ for integers w, c and d. Using Theorem 1.1 we give several new examples of subsets $\mathcal{A}$ where a central limit theorem holds.

Corollary 1.2. *Let x and y be large, with $y \leq x/(\log x)^{\alpha+\epsilon}$ where $\alpha = 2\log 2 - 1$, and $\epsilon > 0$ is arbitrary. Then, as f varies over Steinhaus random multiplicative functions, the quantity*

$$\frac{1}{\sqrt{y}} \sum_{x \leq n \leq x+y} f(n),$$

is distributed like a standard complex normal random variable with mean 0 and variance 1.

Corollary 1.2 improves upon the result established by Chatterjee and Soundararajan [3], where y was required to be $o(x/\log x)$. The range $y = o(x/\log x)$ is the threshold at which the fourth moment of $\sum_{x \leq n \leq x+y} f(n)$ matches that of a Gaussian, and it was tentatively suggested in [3] that this might be the largest range in which a central limit theorem holds. Corollary 1.2 shows that larger values of y are permissible; the source of the improvement is the flexibility in Theorem 1.1 of choosing an appropriate density 1 subset $\mathcal{S}$ of $[x, x+y]$ on which the fourth moment can be shown to match that of a Gaussian. It is conceivable that the central limit theorem holds for still larger values of y, and perhaps for all $y \leq x/(\log x)^{\epsilon}$ for any $\epsilon > 0$. On the other hand, Harper's work [8] shows that $\frac{1}{\sqrt{y}} \sum_{x \leq n \leq x+y} f(n)$ is not Gaussian if $y \geq x/(\log\log x)^{\frac{1}{2}-\epsilon}$ (and indeed some modifications to his ideas permit even the wider range $y \geq x/\exp((\log\log x)^{\frac{1}{2}-\epsilon})$).

Our second example concerns integers that are sums of two squares (one could consider more general sifted sequences, and we comment on this briefly in §5). While we are unable to treat the set of integers in $[1, x]$ that are sums of two squares, we come close to this and can treat the sums of two squares in any short interval $[x, x+y]$ so long as $x^{\frac{1}{3}} < y = o(x)$. Here our interest is in allowing y as large as possible, and the lower bound on y is imposed in order to guarantee that $[x, x+y]$ contains the expected number of integers (namely about $y/\sqrt{\log x}$) that are sums of two squares (see Hooley [11] who shows that the exponent $\frac{1}{3}$ may be replaced with the best available result on the circle problem).

Corollary 1.3. *Let x and y be large with y in the range $x^{\frac{1}{3}} < y = o(x)$. Let $\mathcal{A}$ denote the set of integers in $[x, x+y]$ that are the sum of two squares. In the given range of y, we have $|\mathcal{A}| \asymp \frac{y}{\sqrt{\log x}}$, and moreover, as f varies over Steinhaus random multiplicative functions, the quantity*

$$\frac{1}{\sqrt{|\mathcal{A}|}} \sum_{n \in \mathcal{A}} f(n)$$

is distributed like a standard complex normal random variable with mean 0 and variance 1.

If we restrict a Steinhaus random multiplicative function to its values on primes, then these values are independent random variables and clearly the central limit theorem holds.

Our next example shows that the central limit theorem also holds for the set of shifted primes $p+k$, for any fixed non-zero integer k.

Corollary 1.4. *Let k be any fixed non-zero integer. Let N be large, and let $\mathcal{A}$ denote the set of integers of the form $p+k$ in $[1,N]$. As f varies over Steinhaus random multiplicative functions, the quantity*

$$\frac{1}{\sqrt{|\mathcal{A}|}}\sum_{n\in\mathcal{A}} f(n)$$

is distributed like a standard complex normal random variable with mean 0 and variance 1.

The examples in the three corollaries above give subsets of size $N/(\log N)^{\beta+\epsilon}$ in $[1,N]$ for which a central limit theorem holds, with the largest subset being the short interval result of Corollary 1.2 which permits $\beta = 2\log 2 - 1$. What is the largest subset $\mathcal{A}$ of $[1,N]$ for which we can establish a central limit theorem for $\sum_{n\in\mathcal{A}} f(n)$? As we indicated earlier, we expect (but cannot prove) that a central limit theorem holds for short intervals $[x, x+y]$ with $y = x/(\log x)^{\epsilon}$. The largest set that we have been able to find is given in the next corollary, and this set is related to the Erdős multiplication table problem. The construction is partly random, and based largely on the work of Ford [5].

Corollary 1.5. *For all large N, there exists a subset $\mathcal{A}$ of $[1,N]$ with*

$$|\mathcal{A}| \geq \frac{N}{(\log N)^{\theta}(\log\log N)^{7}}, \qquad \text{where } \theta = 1 - \frac{1+\log\log 4}{\log 4} = 0.0430\ldots$$

such that, for random Steinhaus multiplicative functions f, the quantity

$$\frac{1}{\sqrt{|\mathcal{A}|}}\sum_{n\in\mathcal{A}} f(n)$$

is distributed like a standard complex normal random variable with mean 0 and variance 1.

Our proof of Theorem 1.1 applies more generally to weighted sums $\sum_{n\leq N} a_n f(n)$ for suitable sequences of complex numbers a_n (indeed the precise version Theorem 3.1 is formulated for these more general sums). We highlight a particularly interesting case when $a_n = e(n\theta)$ (here $e(n\theta)$ denotes as usual $e^{2\pi i n\theta}$). For a large class of irrational numbers θ we show that $\sum_{n\leq N} f(n)e(n\theta)$ has a Gaussian distribution.

Theorem 1.6. *Let θ denote an irrational number such that for some positive constant $C = C(\theta)$ and all $q\in\mathbb{N}$ we have*

$$\|q\theta\| := \min_{n\in\mathbb{Z}} |q\theta - n| \geq C\exp(-q^{\frac{1}{50}}). \tag{1.4}$$

If N is sufficiently large (in terms of θ), then as f varies over Steinhaus random multiplicative functions, the quantity

$$\frac{1}{\sqrt{N}}\sum_{n\leq N} e(n\theta)f(n)$$

is distributed like a standard complex normal random variable with mean 0 and variance 1.

The condition (1.4) imposed on θ holds for almost all irrational numbers θ, and includes all algebraic irrationals, as well as transcendental numbers such as e and π that are known not to be Liouville numbers. We have made no attempt to optimize the exponent $\frac{1}{50}$ appearing in the criterion (1.4), and there is certainly scope for improving it. On the other hand, from Harper's work it follows that Theorem 1.6 cannot hold for rational θ, as well as θ that permit extremely good rational approximations, so that some version of criterion (1.4) is necessary. Finally, we point out recent related work of Benatar, Nishry, and Rodgers [1] who consider, for a random Steinhaus multiplicative function f, distribution questions concerning $\sum_{n\leq N} f(n)e(n\theta)$ as θ varies in $\mathbb{R}/\mathbb{Z}$.

Briefly, the paper is organized as follows. Section 2 develops the martingale central limit theorem in a quantitative form, based on the work of McLeish [16]. Section 3 makes an initial application of these results to the setting of random multiplicative functions. In particular, we derive there our main technical result Theorem 3.1, and deduce the simplified Theorem 1.1 stated above. Section 4 is devoted to the distribution of random multiplicative functions in short intervals, and we establish Corollary 1.2 there. Corollary 1.3 is treated in Section 5. In both Sections 4 and 5 a crucial role is played by the flexibility of being able to choose dense subsets $\mathcal{S}$ of $\mathcal{A}$ in Theorem 3.1. Corollary 1.4 admits a short proof based on a simple upper bound sieve, which is presented in Section 6. Section 7 gives the proof of Corollary 1.5, and the construction is based on the work of Ford [5] on extremal product sets. Section 8 deals with Theorem 1.6, and the work of Montgomery and Vaughan [17] on exponential sums with multiplicative functions plays a key role here. Finally, Section 9 ends with a brief discussion of corresponding results in the setting of Rademacher random multiplicative functions.

Acknowledgments. We thank Adam Harper for helpful discussions and comments on an earlier version of the paper. We are also grateful to Louis Gaudet for raising a question during the second author's graduate student seminar at AIM, which led us to Corollary 1.3. Thanks are also due to the referee for a careful reading. K.S. is partially supported through a grant from the National Science Foundation, and a Simons Investigator Grant from the Simons Foundation. M.W.X. is partially supported by the Cuthbert C. Hurd Graduate Fellowship in the Mathematical Sciences, Stanford.

2. McLeish's martingale central limit theorem

In this section we give a quantitative version of the martingale central limit theorem, which we will apply to the study of random multiplicative functions. We follow the treatment in McLeish [16], but adding some quantification. The short proof is included for completeness, and in the hope that it may be useful to readers more familiar with analytic number theory than probability.

Let $X_1, \dots, X_N$ denote a martingale difference sequence of real valued random variables. That is, we suppose that

$$\mathbb{E}[X_1] = 0,$$

and for $1 \leq n \leq N-1$,

$$\mathbb{E}[X_{n+1}|X_1, \dots, X_n] = 0,$$

where $\mathbb{E}[X|Y]$ denotes the conditional expectation of X given Y. Define

$$S_N = X_1 + \ldots + X_N,$$

and our goal is to show that, under suitable conditions, S_N behaves like a real Gaussian with mean 0 and variance 1. We will achieve this by computing the Fourier transform $\mathbb{E}[e^{itS_N}]$ and showing that it approximates $e^{-t^2/2}$ (which is the Fourier transform of a standard real Gaussian). We begin with a simple lemma, which will be key to the result.

Lemma 2.1. *Let $y_1, \ldots, y_N$ be real numbers, and define K (in $[1, N]$) to be the largest integer such that*

$$\sum_{n=1}^{K-1} y_n^2 < 2.$$

Then for any real number t we have

$$e^{it(y_1+\ldots+y_N)} = \prod_{n=1}^{K}(1+ity_n)e^{-\frac{t^2}{2}} + O\Big(e^{t^2}\max_{n=1}^{N}|y_n|\Big) + O\Big(e^{t^2}\min\Big(1, \Big|\sum_{n=1}^{N} y_n^2 - 1\Big|\Big)\Big).$$

Proof. Throughout the proof, the following elementary observations will be useful. For any real number x we have

$$|1+ix| = (1+x^2)^{\frac{1}{2}} \le e^{x^2/2}, \tag{2.1}$$

and, by a Taylor expansion,

$$e^{ix} = (1+ix)e^{-x^2/2}\exp(O(|x|^3)). \tag{2.2}$$

First let us consider the case $\sum_{n=1}^{N} y_n^2 \ge 2$, where the remainder terms in the lemma are clearly $\gg e^{t^2}(1+\max_{n=1}^{N}|y_n|)$. On the other hand, note that by (2.1) and the definition of K

$$\begin{aligned}\Big|e^{it(y_1+\ldots+y_N)} - \prod_{n=1}^{K}(1+ity_n)e^{-\frac{t^2}{2}}\Big| &\le 1 + (1+|ty_K|)e^{-\frac{t^2}{2}}\prod_{n=1}^{K-1}|1+ity_n| \\ &\le 1 + (1+|ty_K|)e^{-\frac{t^2}{2}}\exp\Big(\frac{1}{2}\sum_{n=1}^{K-1}t^2y_n^2\Big) \\ &\le 1 + (1+|ty_K|)e^{t^2/2}.\end{aligned}$$

Since $e^{t^2} \gg 1 + te^{t^2/2}$, the result follows in this case.

Consider then the complementary case $\sum_{n=1}^{N} y_n^2 < 2$, so that $K = N$. Here (2.1) gives

$$\Big|\prod_{n=1}^{N}(1+ity_n)\Big| \le \exp\Big(\sum_{n=1}^{N}\frac{t^2y_n^2}{2}\Big) \le e^{t^2},$$

so that the result holds if $\max_{n=1}^{N}|y_n| \ge e^{-\frac{t^2}{2}}$, or if $\sum_{n=1}^{N} y_n^2 \ge \frac{3}{2}$. Assume therefore that $\max_{n=1}^{N}|y_n| \le e^{-\frac{t^2}{2}}$, and that $\sum_{n=1}^{N} y_n^2 \le \frac{3}{2}$.

Now the Taylor approximation (2.2) gives

$$e^{it(y_1+\ldots+y_N)} = \prod_{n=1}^{N}(1+ity_n)\exp\Big(-\frac{1}{2}\sum_{n=1}^{N}t^2y_n^2 + O\Big(\sum_{n=1}^{N}|ty_n|^3\Big)\Big).$$

Since

$$|t|^3\sum_{n=1}^{N}|y_n|^3 \le |t|^3\Big(\max_n|y_n|\Big)\sum_{n=1}^{N}y_n^2 \le 2|t|^3\max_n|y_n|,$$

and this is $\ll 1$ by our assumption, we conclude that

$$\begin{aligned} e^{it(y_1+\ldots+y_N)} &= \prod_{n=1}^{N}(1+ity_n)\exp\Big(-\frac{1}{2}\sum_{n=1}^{N}t^2y_n^2\Big)\Big(1+O\Big(|t|^3\max_n|y_n|\Big)\Big)\\ &= \prod_{n=1}^{N}(1+ity_n)\exp\Big(-\frac{1}{2}\sum_{n=1}^{N}t^2y_n^2\Big) + O\Big(|t|^3\max_n|y_n|\Big). \end{aligned}$$

Since

$$\Big|\prod_{n=1}^{N}(1+ity_n)\Big| \le \exp\Big(\frac{t^2}{2}\sum_{n=1}^{N}y_n^2\Big) \le \exp\Big(\frac{3}{4}t^2\Big),$$

and

$$\Big|\exp\Big(-\frac{t^2}{2}\sum_{n=1}^{N}y_n^2\Big) - \exp\Big(-\frac{t^2}{2}\Big)\Big| \le \frac{t^2}{2}\Big|\sum_{n=1}^{N}y_n^2 - 1\Big|,$$

the desired estimate follows. □

Proposition 2.2. *Let X_n be a real valued martingale difference sequence, and put $S_N = X_1+\ldots+X_N$. Assume that $\mathbb{E}[\max_{n=1}^{N}|X_n|]$ exists. Then for any $t\in\mathbb{R}$ we have*

$$\mathbb{E}[e^{itS_N}] = e^{-t^2/2} + O\Big(e^{t^2}\Big(\mathbb{E}[\max_{n=1}^{N}|X_n|] + \mathbb{E}\Big[\min\Big(1,\Big|\sum_{n=1}^{N}X_n^2-1\Big|\Big)\Big]\Big)\Big).$$

Proof. Define a new sequence of random variables $\widetilde{X_n}$ by

$$\widetilde{X_n} = X_n\mathbb{I}\Big[\sum_{j=1}^{n-1}X_j^2 \le 2\Big].$$

From its definition one sees that $\widetilde{X_n}$ is also a martingale difference sequence.

Further, if K is the largest integer in $[1,N]$ with $\sum_{n=1}^{K-1}X_n^2\le 2$, then note that $\widetilde{X}_n = X_n$ for $n\le K$, and $\widetilde{X}_n = 0$ for $n>K$. From Lemma 2.1 it follows that

$$e^{itS_N} = \prod_{n=1}^{N}(1+it\widetilde{X}_n)e^{-\frac{t^2}{2}} + O\Big(e^{t^2}\Big(\max_n|X_n| + \min\Big(1,\Big|\sum_{n=1}^{N}X_n^2-1\Big|\Big)\Big)\Big).$$

Take expectations on both sides, and note that the martingale property gives

$$\mathbb{E}\Big[\prod_{n=1}^{N}(1+it\widetilde{X_n})\Big] = 1.$$

The proposition follows. □

From Proposition 2.2 we extract the following quantitative result where the conditions are stronger than necessary but easier to check in practice.

Theorem 2.3. *Let X_n be a real valued martingale difference sequence, and put $S_N = X_1 + \ldots + X_N$. Suppose that $\mathbb{E}[X_n^4]$ exists for each n. Then for any real number t we have*

$$\mathbb{E}[e^{itS_N}] = e^{-t^2/2} + O\Big(e^{t^2}\Big(\sum_{n=1}^{N}\mathbb{E}[X_n^4]\Big)^{\frac14}\Big) + O\Big(e^{t^2}\Big(\mathbb{E}\Big[\Big(\sum_{n=1}^{N}X_n^2-1\Big)^2\Big]\Big)^{\frac12}\Big).$$

Proof. Hölder's inequality gives

$$\mathbb{E}[\max_{n=1}^{N}|X_n|] \le \mathbb{E}\Big[\Big(\sum_{n=1}^{N}X_n^4\Big)^{\frac14}\Big] \le \Big(\mathbb{E}\Big[\sum_{n=1}^{N}X_n^4\Big]\Big)^{\frac14} = \Big(\sum_{n=1}^{N}\mathbb{E}[X_n^4]\Big)^{\frac14}.$$

Further, applying the Cauchy-Schwarz inequality we find

$$\mathbb{E}\Big[\min\Big(1,\Big|\sum_{n=1}^{N}X_n^2-1\Big|\Big)\Big] \le \Big(\mathbb{E}\Big[\Big(\sum_{n=1}^{N}X_n^2-1\Big)^2\Big]\Big)^{\frac12}.$$

Thus the stated result follows immediately from Proposition 2.2. □

So far we have treated martingale difference sequences of real valued random variables. Let us now turn to a martingale difference sequence $Z_1, \ldots, Z_N$ of complex valued random variables, where we aim to show that the partial sums $S_N = \sum_{n=1}^{N} Z_n$ have a standard complex normal distribution. To achieve this we shall compute the Fourier transform $\mathbb{E}[e^{it_1\mathrm{Re}(S_N)+it_2\mathrm{Im}(S_N)}]$ and show that this approximates $e^{-(t_1^2+t_2^2)/4}$ (which is the Fourier transform of a standard complex Gaussian).

Theorem 2.4. *Let $Z_1, \ldots, Z_n$ 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 real numbers t_1 and t_2 we have, with $t^2 = (t_1^2+t_2^2)/2$,*

$$\mathbb{E}[e^{it_1Re(S_N)+it_2Im(S_N)}] = e^{-t^2/2} + O\Big(e^{t^2}\Big(\sum_{n=1}^{N}\mathbb{E}[|Z_n|^4]\Big)^{\frac14}\Big) + O\Big(e^{t^2}\Big(\mathbb{E}\Big[\Big(\sum_{n=1}^{N}|Z_n|^2-1\Big)^2\Big]\Big)^{\frac12}\Big)$$
$$+ O\Big(e^{t^2}\max_{\phi}\Big(\mathbb{E}\Big[\Big(\sum_{n=1}^{N}(e^{-i\phi}Z_n^2+e^{i\phi}\overline{Z_n}^2)\Big)^2\Big]\Big)^{\frac12}\Big).$$

Proof. Write

$$\frac{t_1+it_2}{2}=\frac{te^{i\theta}}{\sqrt{2}} \qquad \text{with } t=\frac{\sqrt{t_1^2+t_2^2}}{\sqrt{2}},$$

so that

$$t_1\mathrm{Re}(S_N)+t_2\mathrm{Im}(S_N)=t\Big(\frac{e^{-i\theta}S_N+e^{i\theta}\overline{S_N}}{\sqrt{2}}\Big)=t\sum_{n=1}^{N}\frac{e^{-i\theta}Z_n+e^{i\theta}\overline{Z_n}}{\sqrt{2}}.$$

Now $X_n=(e^{-i\theta}Z_n+e^{i\theta}\overline{Z_n})/\sqrt{2}$ forms a real valued martingale difference sequence, and we may apply Theorem 2.3. It follows that

$$\mathbb{E}[e^{i(t_1\mathrm{Re}(S_N)+t_2\mathrm{Im}(S_N))}]=e^{-t^2/2}+O\Big(e^{t^2}\Big(\sum_{n=1}^{N}\mathbb{E}[X_n^4]\Big)^{\frac14}\Big)+O\Big(e^{t^2}\Big(\mathbb{E}\Big[\Big(\sum_{n=1}^{N}X_n^2-1\Big)^2\Big]\Big)^{\frac12}\Big).$$

Now note that $\mathbb{E}[X_n^4]\ll\mathbb{E}[|Z_n|^4]$ and so the first error term above is

$$O\Big(e^{t^2}\Big(\sum_{n=1}^{N}\mathbb{E}[|Z_n|^4]\Big)^{\frac14}\Big).$$

Regarding the second error term, note that

$$\begin{aligned}\Big(\sum_{n=1}^{N}X_n^2-1\Big)^2&=\Big(\sum_{n=1}^{N}\Big(\frac{e^{-2i\theta}Z_n^2+e^{2i\theta}\overline{Z_n}^2}{2}+|Z_n|^2\Big)-1\Big)^2\\&\ll\Big(\sum_{n=1}^{N}|Z_n|^2-1\Big)^2+\Big(\sum_{n=1}^{N}(e^{-2i\theta}Z_n^2+e^{2i\theta}\overline{Z_n}^2)\Big)^2.\end{aligned}$$

The theorem follows readily. □

3. Application to random multiplicative functions

We now apply the work in Section 2 to the study of random multiplicative functions. From now on, we shall denote by $P(n)$ the largest prime factor of the integer n.

Theorem 3.1. *Let f denote a random Steinhaus multiplicative function, and let a_n denote a sequence of complex numbers. Put*

$$V=\sum_{n\le N}|a_n|^2,$$

and define the complex valued random variable

$$Z:=\frac{1}{\sqrt{V}}\sum_{n\le N}a_nf(n).$$

Suppose $\mathcal{S}$ is a subset of $[2,N]$ such that for some $1\ge\epsilon>0$ the following three conditions hold:

(1). We have

$$\sum_{\substack{n\le N\\ n\notin\mathcal{S}}} |a_n|^2 \le \epsilon^2 V.$$

(2). We have

$$\Big|\sum_{\substack{m_1,m_2,n_1,n_2\in\mathcal{S}\\ m_1m_2=n_1n_2\\ m_1\neq n_1, m_2\neq n_2\\ P(m_1)=P(n_1)\\ P(m_2)=P(n_2)}} a_{m_1}a_{m_2}\overline{a_{n_1}a_{n_2}}\Big| \le \epsilon^2 V^2.$$

(3). We have

$$\Big|\sum_{\substack{m_1,m_2,n_1,n_2\in\mathcal{S}\\ m_1m_2=n_1n_2\\ P(m_1)=P(n_1)=P(m_2)=P(n_2)}} a_{m_1}a_{m_2}\overline{a_{n_1}a_{n_2}}\Big| \le \epsilon^4 V^2.$$

Then for any real numbers t_1 and t_2 we have, with $t^2=(t_1^2+t_2^2)/2$

$$\mathbb{E}[e^{it_1 Re(Z)+it_2 Im(Z)}] = e^{-t^2/2}+O(e^{t^2}\epsilon).$$

Proof. Put

$$\widetilde{Z}=\frac{1}{\sqrt{V}}\sum_{n\in\mathcal{S}} a_n f(n).$$

Note that, using the Cauchy–Schwarz inequality and assumption (1),

$$\begin{aligned}\Big|\mathbb{E}[e^{it_1\mathrm{Re}(Z)+it_2\mathrm{Im}(Z)}]-\mathbb{E}[e^{it_1\mathrm{Re}(\widetilde{Z})+it_2\mathrm{Im}(\widetilde{Z})}]\Big| &\ll \mathbb{E}[|t_1\mathrm{Re}(Z-\widetilde{Z})+t_2\mathrm{Im}(Z-\widetilde{Z})|]\\ &\ll \Big(t^2\mathbb{E}[|Z-\widetilde{Z}|^2]\Big)^{\frac12}=\Big(\frac{t^2}{V}\sum_{\substack{n\le N\\ n\notin\mathcal{S}}}|a_n|^2\Big)^{\frac12}=O(\epsilon e^{t^2}).\end{aligned}$$

Thus it is enough to compute $\mathbb{E}[e^{it_1\mathrm{Re}(\widetilde{Z})+it_2\mathrm{Im}(\widetilde{Z})}]$, and we approach this using our work in Section 2.

For each prime $p\le N$ define

$$\widetilde{Z}_p=\frac{1}{\sqrt{V}}\sum_{\substack{n\in\mathcal{S}\\ P(n)=p}} a_n f(n),$$

so that $\widetilde{Z}=\sum_{p\le N}\widetilde{Z}_p$. Notice that each term in the sum defining $\widetilde{Z}_p$ involves $f(p)$, which is independent of all $f(\ell)$ with ℓ being a prime $<p$. Thus $\widetilde{Z}_p$ forms a martingale difference sequence as p varies over all the primes at most N. Therefore, we may apply Theorem 2.4 to evaluate $\mathbb{E}[e^{it_1\mathrm{Re}(\widetilde{Z})+it_2\mathrm{Im}(\widetilde{Z})}]$. The martingale decomposition given above was pioneered by Harper [7], motivated by work of Blei and Janson [2], and related decompositions have appeared for instance in [14].

Now observe that

$$\sum_{p\le N}\mathbb{E}[|\widetilde{Z}_p|^4] = \frac{1}{V^2}\sum_{p\le N}\sum_{\substack{m_1,m_2,n_1,n_2\in\mathcal{S}\\ P(m_1)=P(m_2)=P(n_1)=P(n_2)=p}} a_{m_1}a_{m_2}\overline{a_{n_1}a_{n_2}}\mathbb{E}[f(m_1m_2)\overline{f(n_1n_2)}]$$
$$= \frac{1}{V^2}\sum_{\substack{m_1,m_2,n_1,n_2\in\mathcal{S}\\ P(m_1)=P(m_2)=P(n_1)=P(n_2)\\ m_1m_2=n_1n_2}} a_{m_1}a_{m_2}\overline{a_{n_1}a_{n_2}},$$

which is bounded (in magnitude) by ϵ^4 by our assumption (3). Thus the first error term in applying Theorem 2.4 is $O(e^{t^2}\epsilon)$.

Next observe that

$$\mathbb{E}\Big[\Big(\sum_{p\le N}|\widetilde{Z}_p|^2-1\Big)^2\Big] = \sum_{p,q\le N}\mathbb{E}[|\widetilde{Z}_p|^2|\widetilde{Z}_q|^2] - 2\sum_{p\le N}\mathbb{E}[|\widetilde{Z}_p|^2]+1.$$

Now

$$\sum_{p,q\le N}\mathbb{E}[|\widetilde{Z}_p|^2|\widetilde{Z}_q|^2] = \frac{1}{V^2}\sum_{p,q\le N}\sum_{\substack{m_1,n_1\in\mathcal{S}\\ P(n_1)=P(m_1)=p}}\sum_{\substack{m_2,n_2\in\mathcal{S}\\ P(m_2)=P(n_2)=q}} a_{m_1}a_{m_2}\overline{a_{n_1}a_{n_2}}\mathbb{E}[f(m_1m_2)\overline{f(n_1n_2)}]$$
$$= \frac{1}{V^2}\sum_{\substack{m_1,n_1,m_2,n_2\in\mathcal{S}\\ m_1m_2=n_1n_2\\ P(m_1)=P(n_1)\\ P(m_2)=P(n_2)}} a_{m_1}a_{m_2}\overline{a_{n_1}a_{n_2}} = \frac{1}{V^2}\Big(\sum_{n\in\mathcal{S}}|a_n|^2\Big)^2 + O(\epsilon^2),$$

upon isolating the terms $m_1=n_1$ and $m_2=n_2$, and then using assumption (2) to bound the remaining terms. Further

$$\sum_{p\le N}\mathbb{E}[|\widetilde{Z}_p|^2] = \frac{1}{V}\sum_{p\le N}\sum_{\substack{m,n\in\mathcal{S}\\ P(m)=P(n)=p}} a_m\overline{a_n}\mathbb{E}[f(m)\overline{f(n)}] = \frac{1}{V}\sum_{n\in\mathcal{S}}|a_n|^2,$$

so that

$$\mathbb{E}\Big[\Big(\sum_{p\le N}|\widetilde{Z}_p|^2-1\Big)^2\Big] = \Big(\frac{1}{V}\sum_{n\in\mathcal{S}}|a_n|^2-1\Big)^2 + O(\epsilon^2) = O(\epsilon^2),$$

upon using assumption (1) in the last step. Therefore the second error term while using Theorem 2.4 may also be bounded by $O(e^{t^2}\epsilon)$.

Finally consider the third error term in Theorem 2.4, which involves the maximum over ϕ of

$$\mathbb{E}\Big[\Big(\sum_{p\le N}(e^{-i\phi}\widetilde{Z}_p^2+e^{i\phi}\overline{\widetilde{Z}}_p{}^2)\Big)^2\Big] = \sum_{p,q\le N}\mathbb{E}\Big[(e^{-i\phi}\widetilde{Z}_p^2+e^{i\phi}\overline{\widetilde{Z}}_p{}^2)(e^{-i\phi}\widetilde{Z}_q^2+e^{i\phi}\overline{\widetilde{Z}}_q{}^2)\Big].$$

If $p\neq q$ then

$$\mathbb{E}[\widetilde{Z}_p^2\widetilde{Z}_q^2] = \mathbb{E}[\widetilde{Z}_p^2\overline{\widetilde{Z}}_q{}^2] = \mathbb{E}[\overline{\widetilde{Z}}_p{}^2\widetilde{Z}_q^2] = \mathbb{E}[\overline{\widetilde{Z}}_p{}^2\overline{\widetilde{Z}}_q{}^2] = 0,$$

as may be seen by expanding these terms and noting that no diagonal terms arise. We are left with the terms $p = q$ which contribute

$$\ll \sum_{p \le N} \mathbb{E}\Big[|\widetilde{Z}_p|^4\Big] \ll \epsilon^4,$$

from our work on the first error term. Thus the third error term appearing in Theorem 2.4 is $O(e^{t^2}\epsilon^2)$, and the proof of the theorem is complete. □

Let us now record a simplified version of Theorem 3.1 when a_n is the indicator function of a set.

Corollary 3.2. *Let $\mathcal{A}$ be a non-empty subset of natural numbers and let f denote a random Steinhaus multiplicative function. Define the complex valued random variable*

$$Z = \frac{1}{\sqrt{|\mathcal{A}|}} \sum_{n \in \mathcal{A}} f(n).$$

Let $\mathcal{S}$ be a subset of $\mathcal{A}$, with all elements in $\mathcal{S}$ being at least 2. Suppose that $1 \ge \epsilon \ge 0$ is such that the following three conditions are met:

(1). $|\mathcal{A} \backslash \mathcal{S}| \le \epsilon^2 |\mathcal{A}|$.

(2). The number of solutions to $m_1 m_2 = n_1 n_2$ with m_1, m_2, n_1 and $n_2 \in \mathcal{S}$, and $m_1 \ne n_1$, $m_1 \ne n_2$ is bounded by $\epsilon^4 |\mathcal{A}|^2$.

(3). For each prime p

$$\#\{s \in \mathcal{S} : \quad P(s) = p\} \le \epsilon^4 |\mathcal{A}|.$$

Then for any real numbers t_1 and t_2 we have, with $t^2 = (t_1^2 + t_2^2)/2$

$$\mathbb{E}[e^{it_1 Re(Z) + it_2 Im(Z)}] = e^{-t^2/2} + O(e^{t^2}\epsilon).$$

Proof. We apply Theorem 3.1 with a_n denoting the indicator function of the set $\mathcal{A}$. Condition (1) of Theorem 3.1 holds by assumption (1) here.

The sum in condition (2) of Theorem 3.1 counts non-diagonal solutions of $m_1 m_2 = n_1 n_2$ with $m_i, n_i \in \mathcal{S}$ together with special diagonal solutions $m_1 = n_2$ and $m_2 = n_1$ with $P(m_1) = P(m_2) = P(n_1) = P(n_2)$. By our assumption (2) the non-diagonal solutions are bounded by $\epsilon^4 |\mathcal{A}|^2$. As for the special diagonal solutions, these are bounded by

$$\sum_p (\#\{s \in \mathcal{S} : P(s) = p\})^2 \le \epsilon^4 |\mathcal{A}| \sum_p \#\{s \in \mathcal{S} : P(s) = p\} \le \epsilon^4 |\mathcal{A}|^2,$$

upon using our assumption (3). Thus condition (2) of Theorem 3.1 holds, with $2\epsilon^4$ in place of ϵ^2 there.

Finally the sum in condition (3) of Theorem 3.1 is bounded above by the count of non-diagonal solutions to $m_1 m_2 = n_1 n_2$, together with two copies of the special diagonal solutions bounded above. Thus this sum is bounded by $3\epsilon^4 |\mathcal{A}|^2$.

We have checked that the conditions in Theorem 3.1 are met (with a slightly larger value of ϵ there), and the stated result follows. □

The key condition in Corollary 3.2 is the assumption (2) on non-diagonal solutions. The third assumption is often harmless, and our next lemma shows that it holds automatically for large subsets of intervals.

Lemma 3.3. *For all primes p, and all $x \geq y \geq 3$*

$$\#\{x < n \leq x+y : \quad P(n) = p\} \ll y \exp(-\tfrac{1}{2}\sqrt{\log y \log\log y}).$$

Proof. If $p > \exp(\frac{1}{2}\sqrt{\log y \log\log y})$ then

$$\#\{x < n \leq x+y : \quad P(n) = p\} \leq \frac{y}{p} + 1 \ll y \exp(-\tfrac{1}{2}\sqrt{\log y \log\log y}).$$

If $p < \exp(\frac{1}{2}\sqrt{\log y \log\log y}) =: z$ (say) then

$$\#\{x < n \leq x+y : \quad P(n) = p\} \leq \Psi(x+y, z) - \Psi(x, z) \leq \Psi(y, z),$$

where $\Psi(x, z)$ denotes the number of integers below x all of whose prime factors are below z, and the inequality used above is due to Hildebrand [10]. The bound of the lemma now follows from the familiar estimate

$$\Psi(y, z) = yu^{-(1+o(1))u}$$

with $u = \log y/\log z = 2\sqrt{\log y/\log\log y}$. □

Proof of Theorem 1.1. Since the convergence of characteristic functions implies the convergence in distribution (see, for example, [6]), the theorem follows immediately from Corollary 3.2 and Lemma 3.3. □

Several times above we have encountered the relation $m_1 m_2 = n_1 n_2$, and we now record a simple parametrization of these solutions, setting up notation that we shall use later.

Lemma 3.4. *The solutions to $m_1 m_2 = n_1 n_2$ may be parameterized as*

$$m_1 = ga, \quad m_2 = hb, \quad n_1 = gb, \quad n_2 = ha,$$

where $(a, b) = 1$. Diagonal solutions correspond to solutions with $g = h$ or $a = b$ (in which case $a = b = 1$).

Proof. All we have done is to write $g = (m_1, n_1)$ and $h = (m_2, n_2)$. □

4. Random multiplicative functions in short intervals: Proof of Corollary 1.2

We now use our work in Section 3 to study partial sums of random Steinhaus multiplicative functions over short intervals $\mathcal{A} = [x, x+y]$. Throughout we think of y as large, and $x \geq y$. Our goal is to show Corollary 1.2, which states that in the range $y \leq x/(\log x)^{2\log 2 - 1 + \epsilon}$ the limiting distribution of $\sum_{x \leq n \leq x+y} f(n)$ is Gaussian (improving upon the earlier result in [3]). Below we use the standard notation $\Omega(n)$ to denote the number of prime factors of n counted with multiplicity.

Proposition 4.1. *Let y be large with $y \le x$. If $y \le x/(\log x)^2$ take $\mathcal{S}$ to be all of $\mathcal{A} = [x, x+y]$, and in the range $y > x/(\log x)^2$ define $\mathcal{S}$ to be subset of elements in $\mathcal{A}$ satisfying $\Omega(n) \le (1+\epsilon)\log\log x$. Then, the number of integers in $[x, x+y]$ that are not in $\mathcal{S}$ is $o(y)$. Further, in the range $y \le x(\log x)^{1-2\log 2-\epsilon}$, the number of off-diagonal solutions to*

$$m_1 m_2 = n_1 n_2 \qquad \text{with } m_1, m_2, n_1, n_2 \in \mathcal{S},$$

is $o(y^2)$.

Deduction of Corollary 1.2. We apply Corollary 3.2 with $\mathcal{A} = [x, x+y]$ and $\mathcal{S}$ as defined above. The first and the second conditions in Corollary 3.2 follow from Proposition 4.1, and the third condition follows from Lemma 3.3. Thus Corollary 3.2 applies and shows that the Fourier transform of $\sum_{x\le n\le x+y} f(n)$ matches that of a complex Gaussian, giving the desired result. □

The proof of Proposition 4.1 relies on the following result of Shiu [21].

Lemma 4.2 (Shiu)**.** *Let $f(n)$ be a non-negative multiplicative function such that*
(i) $f(p^\ell) \le A_1^\ell$ for some positive constant A_1, and
(ii) for any $\epsilon > 0$, $f(n) \le A_2 n^\epsilon$ for some $A_2 = A_2(\epsilon)$.
Then, for all $\sqrt{x} \le y \le x$ we have

$$\sum_{x\le n\le x+y} f(n) \ll \frac{y}{\log x} \exp\Big(\sum_{p\le x} \frac{f(p)}{p}\Big).$$

Proof of Proposition 4.1. We first show that $|\mathcal{A}\backslash\mathcal{S}| = o(y)$. In the range $y \le x/(\log x)^2$ we have $\mathcal{S} = \mathcal{A}$ and so this holds trivially. Suppose then that $x/(\log x)^2 \le y \le x$, where the claim is simply that most integers in $[x, x+y]$ have the expected number of prime factors, namely $\sim \log\log x$. We may deduce this quickly from Lemma 4.2, taking there f to be the completely multiplicative function $f(n) = \exp(\epsilon'\Omega(n))$ with $\epsilon' = 1/\log\log\log x$. Then, an application of Lemma 4.2 gives

$$\begin{aligned}|\mathcal{A}\backslash\mathcal{S}| &\le \exp(-\epsilon'(1+\epsilon)\log\log x) \sum_{x<n\le x+y} f(n)\\ &\ll \exp(-\epsilon'(1+\epsilon)\log\log x)\frac{y}{\log x}\exp\Big(\sum_{p\le x}\frac{f(p)}{p}\Big) \ll y\exp\Big(-\frac{\epsilon'\epsilon}{2}\log\log x\Big),\end{aligned}$$

proving our claim (with lots of room to spare).

We now focus on the main thrust of the proposition, which is to estimate the number of non-diagonal solutions to $m_1 m_2 = n_1 n_2$ with all variables being in $\mathcal{S}$. We use the parameterization in Lemma 3.4, and write below $\delta = y/x$. Thus our goal is to bound

$$\sum_{\substack{a\ne b\\(a,b)=1}} \sum_{\substack{g\ne h\\ ga,gb,ha,hb\in\mathcal{S}}} 1.$$

Since m_1, m_2, n_1, and n_2 must all lie in the interval $[x, x+y]$, we must have

$$\frac{g}{h} = \frac{m_1}{n_2} \in [1/(1+\delta), (1+\delta)], \qquad \text{and} \qquad \frac{a}{b} = \frac{m_1}{n_1} \in [1/(1+\delta), (1+\delta)]. \tag{4.1}$$

Since we are only interested in off-diagonal solutions (with $a \neq b$ and $g \neq h$), we must have a, b, g, and h all being $\gg 1/\delta$. In particular, there are no off-diagonal solutions if $y \leq c\sqrt{x}$ for a suitable constant c, since then ga (for instance) would be $\gg 1/\delta^2 > x$. Assume henceforth that $y \gg \sqrt{x}$. Ignoring the condition $(a,b) = 1$, our goal now is to bound

$$\sum_{\substack{g,a,b,h \gg 1/\delta \\ ga, gb, ha, hb \in \mathcal{S}}} 1 \ll \sum_{\substack{g,h \\ 1/\delta \ll g,h \ll \sqrt{x}}} \sum_{\substack{a,b \\ ga,gb,ha,hb \in \mathcal{S}}} 1. \tag{4.2}$$

In the last estimate above, we used that g/h and a/b are both in $[(1+\delta)^{-1}, (1+\delta)]$, and since $ga \leq x$ we must have either a and b being $\ll \sqrt{x}$ or g and h being $\ll \sqrt{x}$; by symmetry we restricted attention to the latter case.

If g is given, then by (4.1) there are $\ll \delta g$ choices for h. If g and h are fixed, then there are $\ll y/g$ choices each for a and b. Therefore, the number of solutions in (4.2) may be bounded by

$$\ll \sum_{1/\delta \ll g \ll \sqrt{x}} (\delta g)(y/g)^2 \ll \delta y^2 (\log x). \tag{4.3}$$

This is $o(y^2)$ when $y = o(x/\log x)$, and establishes the proposition in that range.

Now suppose that $x/(\log x)^2 \leq y \leq x$. In this range, we exploit that $\mathcal{S}$ contains only those integers $n \in [x, x+y]$ with $\Omega(n) \leq K := (1+\epsilon)\log\log x$. If ga, gb, ha and hb are all in $\mathcal{S}$ then we must have $2K - \Omega(g) - \Omega(a) - \Omega(b) - \Omega(h) \geq 0$, so that we may bound the quantity in (4.2) by

$$\ll \sum_{\substack{g,h \\ 1/\delta \ll g,h \ll \sqrt{x}}} \sum_{\substack{a,b \\ ga,gb,ha,hb \in [x,x+y]}} 2^{2K - \Omega(g) - \Omega(a) - \Omega(b) - \Omega(h)}. \tag{4.4}$$

Here the weight $2^{2K-\Omega(gabh)}$ was chosen with the benefit of hindsight, starting with $\lambda^{2K-\Omega(gabh)}$ for $\lambda \geq 1$ and optimizing the value of λ.

Noting that $(1+\delta)^{-1} \le g/h \le (1+\delta)$, and invoking Lemma 4.2 we may bound the quantity in (4.4) by

$$\begin{aligned}
&\ll 2^{2K} \sum_{\substack{1/\delta \ll g,h \ll \sqrt{x} \\ (1+\delta)^{-1} \le g/h \le (1+\delta)}} 2^{-\Omega(g)-\Omega(h)} \Big(\sum_{a \in [x/g,(x+y)/g]} 2^{-\Omega(a)} \Big)^2 \\
&\ll 2^{2K} \sum_{\substack{1/\delta \ll g,h \ll \sqrt{x} \\ (1+\delta)^{-1} \le g/h \le (1+\delta)}} 2^{-\Omega(g)-\Omega(h)} \Big(\frac{y}{g \log x} \exp\Big(\sum_{p \le x} \frac{1}{2p} \Big) \Big)^2 \\
&\ll \frac{2^{2K} y^2}{\log x} \sum_{\substack{1/\delta \ll g,h \ll \sqrt{x} \\ (1+\delta)^{-1} \le g/h \le (1+\delta)}} \frac{2^{-\Omega(g)-\Omega(h)}}{g^2}. \qquad (4.5)
\end{aligned}$$

It remains to bound the sums over g and h in (4.5). To this end, we split the sum over g into dyadic blocks $G \le g \le 2G$ with $1/\delta \ll G \ll \sqrt{x}$. In the range $1/\delta \ll G \le 1/\delta^2$ note that

$$\sum_{\substack{G \le g \le 2G \\ (1+\delta)^{-1} \le g/h \le (1+\delta)}} \frac{2^{-\Omega(g)-\Omega(h)}}{g^2} \ll \frac{1}{G^2} \sum_{G \le g \le 2G} \sum_{(1+\delta)^{-1} g \le h \le (1+\delta) g} 1 \ll \frac{1}{G^2} G(G\delta) \ll \delta.$$

In the range $1/\delta^2 < G \ll \sqrt{x}$, using Lemma 4.2 twice we obtain the bound

$$\sum_{\substack{G \le g \le 2G \\ (1+\delta)^{-1} \le g/h \le (1+\delta)}} \frac{2^{-\Omega(g)-\Omega(h)}}{g^2} \ll \frac{1}{G^2} \sum_{G \le g \le 2G} 2^{-\Omega(g)} \frac{\delta G}{(\log G)^{\frac{1}{2}}} \ll \frac{\delta}{\log G}.$$

Splitting the interval $1/\delta \ll g \ll \sqrt{x}$ into dyadic blocks, and using the above two estimates, we conclude that the sums over g and h in (4.5) contribute $\ll \delta \log(1/\delta) + \delta \log\log x \ll \delta \log\log x$. Inserting this in (4.5), we conclude that the number of off-diagonal solutions is

$$\ll \frac{2^{2K} y^2}{\log x} \delta \log\log x \qquad (4.6)$$

which is $o(y^2)$ in the range $y \le x/(\log x)^{2\log 2 - 1 + 2\epsilon}$, upon recalling that $K = (1+\epsilon)\log\log x$. □

Remark 4.3. Proposition 4.1 also answers the following question: what is the largest y such that the product set of $\mathcal{A} = [x, x+y]$ has size $|\mathcal{A} \cdot \mathcal{A}| \gg |\mathcal{A}|^2$? Since the Cauchy-Schwarz inequality gives $|\mathcal{A} \cdot \mathcal{A}| \ge |\mathcal{S} \cdot \mathcal{S}| \ge |\mathcal{S}|^4 / E_\times(\mathcal{S})$, from Proposition 4.1 we find that the product set of $[x, x+y]$ has its maximal size $\sim y^2/2$ in the range $y \le x/(\log x)^{2\log 2 - 1 + \epsilon}$. On the other hand, if y is larger than $x/(\log x)^{2\log 2 - 1 - \epsilon}$ then apart from $o(y^2)$ exceptions, an element in the product set $[x, x+y] \cdot [x, x+y]$ would be in $[x^2, x^2 + 2xy + y^2]$ and have about $2\log\log x$ prime factors, and an application of Selberg's work [20] shows that there are at most $o(y^2)$ such elements. Thus the largest y in this problem is of size $x/(\log x)^{2\log 2 - 1 + o(1)}$. There are other closely related problems where the same threshold arises; for instance see [19] for work

on product sets of dense subsets of the first N integers (and a random version is studied in [15]), and see [22] for a study of product sets of arithmetic progressions.

Remark 4.4. The recent paper [18] studies high moments of random multiplicative functions over short intervals $[x, x+y]$, and produces a range of y where the high moments match the Gaussian moments (establishing a central limit theorem). The valid range for y there is weaker than what we establish in Corollary 1.2, and only when x/y is larger than an arbitrarily large power of $\log x$ does the method of moments yield a central limit theorem.

The flexibility of restricting to a dense subset $\mathcal{S}$ of $[x, x+y]$ can facilitate the computation of some higher moments. Indeed the key point in our argument is that, when restricted to integers with a typical number of prime factors, the fourth moment matches that of a Gaussian so long as $y \le x/(\log x)^{2\log 2-1+\epsilon}$. The argument in Remark 4.3 shows that the fourth moment blows up if $y \ge x/(\log x)^{2\log 2-1-\epsilon}$. Even when restricted to integers with a typical number of prime factors, higher moments will still blow up, so that Corollary 1.2 is not accessible by the method of moments.

To illustrate briefly, consider the range of y for which the sixth moment blows up. Let $\mathcal{S}$ be any dense subset of $[x, x+y]$, and let $\mathcal{S}_0$ denote the elements in $\mathcal{S}$ with $\Omega(n) = (1+o(1))\log\log x$ so that $|\mathcal{S}_0|$ is also $\sim y$. Then, an application of the Cauchy-Schwarz inequality gives,

$$\mathbb{E}\Big[\Big|\sum_{n\in\mathcal{S}} f(n)\Big|^6\Big] \ge \mathbb{E}\Big[\Big|\sum_{n\in\mathcal{S}_0} f(n)\Big|^6\Big] \ge \frac{|\mathcal{S}_0|^6}{|\mathcal{S}_0\cdot\mathcal{S}_0\cdot\mathcal{S}_0|}.$$

Now the triple product set $\mathcal{S}_0\cdot\mathcal{S}_0\cdot\mathcal{S}_0$ is a subset of the integers in $[x^3,(x+y)^3]$ having $(3+o(1))\log\log x$ prime factors, and this set has size $yx^2(\log x)^{2-3\log 3+o(1)}$. Therefore

$$\mathbb{E}\Big[\Big|\sum_{n\in\mathcal{S}} f(n)\Big|^6\Big] \ge \frac{y^5}{x^2}(\log x)^{3\log 3-2+o(1)},$$

and this is much bigger than y^3 if $y \ge x/(\log x)^{\frac{3}{2}\log 3-1-\epsilon}$. Note that $\frac{3}{2}\log 3 - 1 = 0.6479\ldots$, while $2\log 2 - 1 = 0.3862\ldots$.

5. Proof of Corollary 1.3

Let $\mathcal{A}$ denote the set of integers in $[x, x+y]$ that are the sum of two squares, where we assume that x and y are large with $x^{\frac{1}{3}} \le y = o(x)$. The lower bound on y ensures, by work of Hooley [11], that $|\mathcal{A}| \asymp y/\sqrt{\log x}$. As with the proof of Corollary 1.2, we shall apply Corollary 3.2 with a suitable choice of $\mathcal{S} \subset \mathcal{A}$. Note that the third condition required in Corollary 3.2 follows from Lemma 3.3, and it remains to specify $\mathcal{S}$ and verify the first two conditions there. As in Section 4, we write $y = \delta x$.

Consider first the range $x^{\frac{1}{3}} \le y \le x/(\log x)^3$, where we shall simply take $\mathcal{S} = \mathcal{A}$. Thus the first condition in Corollary 3.2 is immediate, and it remains to bound the number of non-diagonal solutions to $m_1m_2 = n_1n_2$ with m_1, m_2, n_1, n_2 in $\mathcal{A}$. The argument leading up to (4.3) shows that the number of non-diagonal solutions with m_1, m_2, n_1, n_2 in $[x, x+y]$

(ignoring that they are sums of two squares) is $\ll \delta y^2 \log x$. Since $|\mathcal{A}| \gg y/\sqrt{\log x}$, in the range $\delta \le (\log x)^{-3}$ we see that this bound is $\ll |\mathcal{A}|^2/\log x$, which verifies condition 2.

Therefore we may assume that y is in the range $x/(\log x)^3 \le y = o(x)$. In this range we require a more careful choice of the set $\mathcal{S}$. Let $a(n)$ denote the indicator function of the set of integers that are sums of two squares. Recall that $a(n)$ is a multiplicative function with $a(p^k) = 1$ if p is 2 or $p \equiv 1 \bmod 4$, and for $p \equiv 3 \bmod 4$ given by $a(p^{2k}) = 1$ and $a(p^{2k+1}) = 0$. A typical integer n of size x that is a sum of two squares will have about $\frac{1}{2}\log\log x$ prime factors, and indeed such an integer will have about $\frac{1}{2}k$ prime factors below e^{e^k}. Our set $\mathcal{S}$ will consist of such typical sums of two squares.

More precisely, let $\epsilon > 0$ be small, and let $\mathcal{S}$ be the subset of integers $n \in \mathcal{A}$ satisfying $\Omega(n; e^{e^k}) \le (\frac{1}{2} + \epsilon)k$ for each natural number k in the range $1/\delta \le e^{e^k} \le x$. Here $\Omega(n;t)$ counts the number of prime powers p^a dividing n with $p \le t$. We begin by showing that $|\mathcal{A}\backslash\mathcal{S}|$ is small for small δ, which would verify condition 1 of Corollary 3.2.

Let k be a given integer in the range $\log\log(1/\delta) \le k \le \log\log x$; note that since $\delta = o(1)$, we know that k is large (tending to infinity with x). We first bound the number of integers n in $\mathcal{A}$ that have $\Omega(n; e^{e^k}) \ge (\frac{1}{2} + \epsilon)k$. We apply Shiu's result Lemma 4.2 taking there $f(n) = \exp(k^{-\frac{1}{2}}\Omega(n; e^{e^k}))a(n)$ to obtain

$$\#\{n \in \mathcal{A} : \quad \Omega(n; e^{e^k}) \ge (\tfrac{1}{2} + \epsilon)k\} \le e^{-(\frac{1}{2}+\epsilon)\sqrt{k}} \sum_{x \le n \le x+y} f(n) \ll |\mathcal{A}| \exp(-\epsilon\sqrt{k}).$$

Summing this over all k in the range $\log\log(1/\delta) \le k \le \log\log x$, it follows that $|\mathcal{A}\backslash\mathcal{S}| = o(x)$, as desired.

Having verified condition 1 of Corollary 3.2, it remains lastly to check condition 2; namely to check that there are few non-diagonal solutions to $m_1m_2 = n_1n_2$ with m_1, m_2, n_1, $n_2 \in \mathcal{S}$. We parametrize solutions as in Lemma 3.4 writing $m_1 = ga$, $m_2 = hb$, $n_1 = gb$, and $n_2 = ha$, where $(a,b) = 1$. We may assume that $g \ne h$ and $a \ne b$ (since we only want non-diagonal solutions). As in (4.1) we must have g/h and a/b lying in the interval $[(1+\delta)^{-1}, 1+\delta]$, so that we may assume that g, h, a, b are all $\gg 1/\delta$. Since a and b are coprime, and ga and gb are both sums of two squares, it follows that g, a, and b must all be sums of two squares; similarly h must also be a sum of two squares. Thus, we may suppose that g, h, a, and b are all $\gg 1/\delta$, are all sums of two squares, and as in (4.2) our task is to bound

$$\sum_{\substack{g,h \\ 1/\delta \ll g,h \le \sqrt{x}}} \sum_{\substack{a,b \\ ga,gb,ha,hb \in \mathcal{S}}} a(g)a(h)a(a)a(b). \tag{5.1}$$

Above we omitted the condition $(a,b) = 1$; noted that $\max(g,h)$ or $\max(a,b)$ must be $\le \sqrt{x}$, and assumed that the former condition holds by symmetry.

Break the sum over g in (5.1) into dyadic blocks $G < g \le 2G$ where $1/\delta \ll G \le \sqrt{x}$. We wish to estimate the contribution to (5.1) arising from such a dyadic block. Select k to be the least integer with $e^{e^k} \ge \max(1/\delta, G)$. Since ga, gb, ha, hb are all in $\mathcal{S}$ we must have $\Omega(g; e^{e^k}) + \Omega(h; e^{e^k}) + \Omega(a; e^{e^k}) + \Omega(b; e^{e^k}) \le (1+2\epsilon)k$. Therefore the contribution of this

dyadic block is

$$\ll 2^{(1+2\epsilon)k} \sum_{\substack{G<g\le 2G \\ h\in(g/(1+\delta),g(1+\delta))}} a(g)a(h)2^{-\Omega(g;e^{e^k})-\Omega(h;e^{e^k})} \sum_{a,b\in(x/g,(x+y)/g)} a(a)a(b)2^{-\Omega(a;e^{e^k})-\Omega(b;e^{e^k})}.$$

Now using Lemma 4.2, the sum over a above may be bounded by

$$\ll \frac{y}{g\log x}\exp\Big(\sum_{\substack{p\le e^{e^k}\\ p\equiv 1 \bmod 4}}\frac{1}{2p}+\sum_{\substack{e^{e^k}<p\le x\\ p\equiv 1 \bmod 4}}\frac{1}{p}\Big)\ll \frac{y}{g\sqrt{\log x}}e^{-\frac{k}{4}}.$$

Naturally the same bound holds for the sum over b, and we conclude that the contribution of the dyadic block $G\le g\le 2G$ is

$$\ll e^{k((1+2\epsilon)\log 2-1/2)}\frac{y^2}{G^2\log x}\sum_{\substack{G<g\le 2G\\ h\in(g/(1+\delta),g(1+\delta))}} a(g)a(h)2^{-\Omega(g;e^{e^k})-\Omega(h;e^{e^k})}. \tag{5.2}$$

Consider first the case when $1/\delta \ll G\le 1/\delta^2$. Here we bound $a(g)a(h)2^{-\Omega(g;e^{e^k})-\Omega(h;e^{e^k})}$ by 1, and note that given g there are $\ll \delta g$ choices for h. Noting also that e^k is of size $\log G$, we conclude that in this range of G, the quantity in (5.2) may be bounded by

$$\ll \delta(\log G)^{\frac{1}{4}}\frac{y^2}{\log x}. \tag{5.3}$$

Now consider the range $1/\delta^2\le G\le\sqrt{x}$. Here we may use Lemma 4.2 twice to bound the quantity in (5.2) by

$$\ll (\log G)^{(1+2\epsilon)\log 2-1/2}\frac{y^2}{G^2\log x}\sum_{G\le g\le 2G}a(g)2^{-\Omega(g;e^{e^k})}\frac{\delta g}{\log G}(\log G)^{\frac{1}{4}}$$

$$\ll \frac{y^2}{\log x}\frac{\delta}{(\log G)^{2-(1+2\epsilon)\log 2}}. \tag{5.4}$$

We now return to the problem of bounding (5.1). Using (5.3) the contribution of the dyadic blocks with $1/\delta\ll G\le 1/\delta^2$ may be bounded by $\ll \delta(\log 1/\delta)^{\frac{5}{4}}y^2/\log x$ (since there are $\ll \log(1/\delta)$ such dyadic blocks). Using (5.4), the contribution of all the dyadic blocks with $1/\delta^2\le G\le\sqrt{x}$ is $\ll \delta y^2/\log x$ — the key fact here is that $2-(1+2\epsilon)\log 2>1$ (for suitably small ϵ) so that when $(\log G)^{-(2-(1+2\epsilon)\log 2)}$ is summed over the powers of 2 in this range, the resulting sum is $\ll 1$. We conclude that (5.1), which bounds the non-diagonal solutions to $m_1m_2=n_1n_2$, may be bounded by

$$\ll \delta(\log 1/\delta)^{\frac{5}{4}}\frac{y^2}{\log x}=o(|\mathcal{A}|^2).$$

This completes our verification of condition 2 in Corollary 3.2, and thus our proof of Corollary 1.3.

We remark that the proof goes through for more general sifted sets $\mathcal{A}$. For instance, suppose $\mathcal{A}$ is the set of integers composed of primes lying in subset of the primes with relative density ρ. Then so long as $\rho \le 1/\log 4 - \epsilon$, and $y = o(x)$ is such that $[x, x+y]$ contains the expected number of elements of $\mathcal{A}$ (which is about $y/(\log x)^{1-\rho}$), then one can obtain a suitable central limit theorem.

6. Shifted primes: Proof of Corollary 1.4

To deduce Corollary 1.4 from Theorem 1.1, we need only show that the number of non-diagonal solutions to $(p+k)(q+k) = (r+k)(s+k)$ (with p, q, r and s being primes below N) is $o(\pi(N)^2)$. We use the parametrization of Lemma 3.4 to write $p+k = ga$, $q+k = hb$, $r+k = gb$ and $s+k = ha$, with $a \neq b$ and $g \neq h$ (since we are only interested in bounding non-diagonal solutions). Thus, with $\mathbb{1}_{\mathcal{P}}$ denoting the indicator function of primes, we must bound

$$\sum_{a\neq b} \sum_{\substack{g\neq h \\ ga,gb,ha,hb\le N+k}} \mathbb{1}_{\mathcal{P}}(ga-k)\mathbb{1}_{\mathcal{P}}(gb-k)\mathbb{1}_{\mathcal{P}}(ha-k)\mathbb{1}_{\mathcal{P}}(hb-k).$$

Note that either $\max(a,b)$ or $\max(g,h)$ must be $\le \sqrt{N+k}$. By symmetry, we may restrict attention to the former case, and also assume that $a < b$. Thus the non-diagonal solutions are bounded by

$$\ll \sum_{a<b\le \sqrt{N+k}} \sum_{g,h\le (N+k)/b} \mathbb{1}_{\mathcal{P}}(ga-k)\mathbb{1}_{\mathcal{P}}(gb-k)\mathbb{1}_{\mathcal{P}}(ha-k)\mathbb{1}_{\mathcal{P}}(hb-k).$$

For each small prime p with $p \nmid kab(b-a)$, g must avoid two distinct residue classes mod p (namely the residue classes k/a and k/b mod p) in order for $ga-k$ and $gb-k$ to be prime. For the primes p dividing $kab(b-a)$ ignore any constraints that g must satisfy mod p. Then, a straight-forward upper bound sieve gives

$$\sum_{g\le (N+k)/b} \mathbb{1}_{\mathcal{P}}(ga-k)\mathbb{1}_{\mathcal{P}}(gb-k) \ll \frac{N}{b(\log N)^2}\Big(\frac{|k|ab(b-a)}{\phi(|k|ab(b-a))}\Big)^2$$

and of course the same holds for the sum over h. We conclude that the off-diagonal solutions are bounded by

$$\begin{aligned}
&\ll \sum_{a<b\le\sqrt{N+k}} \frac{N^2}{b^2(\log N)^4}\Big(\frac{|k|ab(b-a)}{\phi(|k|ab(b-a))}\Big)^4 \\
&\ll \frac{N^2}{(\log N)^4}\Big(\frac{|k|}{\phi(|k|)}\Big)^4 \sum_{a<b\le\sqrt{N+k}} \frac{1}{b^2}\Big(\Big(\frac{a}{\phi(a)}\Big)^{12} + \Big(\frac{b}{\phi(b)}\Big)^{12} + \Big(\frac{b-a}{\phi(b-a)}\Big)^{12}\Big),
\end{aligned}$$

upon using the AM-GM inequality. Since $\sum_{n\le x}(n/\phi(n))^{12} \ll x$, it readily follows that the above is

$$\ll \frac{N^2}{(\log N)^3}\Big(\frac{|k|}{\phi(|k|)}\Big)^4,$$

which suffices since k is a fixed non-zero integer.

7. Proof of Corollary 1.5

This section records the largest subset $\mathcal{A} \subseteq [1, N]$ that we know for which $\sum_{n\in\mathcal{A}} f(n)$ has a Gaussian distribution. The construction is essentially due to Ford [5], who showed that there is a subset $\mathcal{B} \subseteq [1, N]$ with $|\mathcal{B}| \geq N(\log N)^{-\theta}(\log\log N)^{-\frac{3}{2}}$ such that $E_\times(\mathcal{B}) \ll |\mathcal{B}|^2(\log\log N)^4$ (see [5, Lemma 3.1, 3.2]).

Let $\mathcal{A}$ range uniformly over all subsets of $\mathcal{B}$ with $\lfloor\rho|\mathcal{B}|\rfloor$ elements, where $\rho = (\log\log N)^{-5}$. Let us compute the average number of non-diagonal solutions to $m_1m_2 = n_1n_2$ with m_1, m_2, n_1, $n_2 \in \mathcal{A}$. Note that any such non-diagonal solution must arise from a non-diagonal solution to $m_1m_2 = n_1n_2$ with m_1, m_2, n_1, $n_2 \in \mathcal{B}$. Since at least three of m_1, m_2, n_1, n_2 must be distinct, such a non-diagonal solution would count as a non-diagonal solution in $\mathcal{A}$ with "probability" $\ll \rho^3$ (three elements of $\mathcal{A}$ are specified, and there are $\binom{|\mathcal{B}|-3}{|\mathcal{A}|-3}$ ways of choosing the remaining elements). It follows that the average number of non-diagonal solutions in $\mathcal{A}$ is $\ll \rho^3 E_\times(\mathcal{B}) \ll |\mathcal{A}|^2(\log\log N)^{-1}$.

We deduce that there exists a subset $\mathcal{A}$ of $[1, N]$ with $|\mathcal{A}| \geq N(\log N)^{-\theta}(\log\log N)^{-7}$ such that the number of non-diagonal solutions to $m_1m_2 = n_1n_2$ with m_1, m_2, n_1, $n_2 \in \mathcal{A}$ being at most $|\mathcal{A}|^2(\log\log N)^{-1}$. Corollary 1.5 now follows from Theorem 1.1.

8. Proof of Theorem 1.6

In this section we study the distribution of random multiplicative functions twisted by $e(n\alpha)$. A key input in understanding such sums is the following result of Montgomery and Vaughan.

Lemma 8.1 (Montgomery-Vaughan [17]). *Let g be a multiplicative function with $|g(n)| \leq 1$ for all n. Let x be large, and let α be a real number. Suppose α has a rational approximation u/v such that $|\alpha - u/v| \leq 1/v^2$, where $(u, v) = 1$ and v lies in the interval $R \leq v \leq x/R$ for some parameter $R \geq 2$. Then*

$$\sum_{n\leq x} g(n)e(n\alpha) \ll \frac{x}{\log x} + \frac{x}{\sqrt{R}}(\log R)^{3/2}.$$

To prove Theorem 1.6 we shall apply Theorem 3.1, taking $\mathcal{A} = \mathcal{S}$ be the set of all positive integers up to x and $a_n = e(n\theta)$. The variance V equals $\lfloor x\rfloor$, and we must check the three criteria given in Theorem 3.1. The first condition holds automatically since $\mathcal{S} = \mathcal{A}$. We now check the third condition, and then consider the second condition (which requires the most work). The third condition in Theorem 3.1 requires a good bound for

$$\Big|\sum_{\substack{m_1,m_2,n_1,n_2\leq x\\ m_1m_2=n_1n_2\\ P(m_1)=P(n_1)=P(m_2)=P(n_2)}} e(\theta(m_1+m_2-n_1-n_2))\Big| \leq \sum_{\substack{m_1,m_2\leq x\\ P(m_1)=P(m_2)}} d(m_1m_2). \tag{8.1}$$

Since $d(m_1m_2) \le d(m_1)d(m_2) \le \frac{1}{2}(d(m_1)^2 + d(m_2)^2)$, we may bound the above quantity by

$$\le \sum_{m_1 \le x} d(m_1)^2 \sum_{\substack{m_2 \le x \\ P(m_2)=P(m_1)}} 1.$$

Given m_1, arguing as in Lemma 3.3 we may bound the inner sum over m_2 by $\ll x\exp(-\frac{1}{2}\sqrt{\log x \log\log x}) \ll x/(\log x)^{10}$, so that the quantity in (8.1) may be bounded by

$$\ll \frac{x}{(\log x)^{10}} \sum_{m_1 \le x} d(m_1)^2 \ll \frac{x^2}{(\log x)^7},$$

which is more than we need.

It remains to verify the second condition, which requires us to bound

$$\sum_{\substack{m_1,m_2,n_1,n_2 \le x \\ m_1m_2=n_1n_2 \\ P(m_1)=P(n_1) \\ P(m_2)=P(n_2) \\ m_1\neq n_1, m_2 \neq n_2}} e((m_1+m_2-n_1-n_2)\theta).$$

We use the parametrization in Lemma 3.4, to write $m_1 = ga$, $m_2 = hb$, $n_1 = gb$ and $n_2 = ha$. The constraints on m_1, m_2, n_1, n_2 then become $(a,b) = 1$ with $a \neq b$, $P(ab) \le \min(P(g), P(h))$, and $\max(a,b) \times \max(g,h) \le x$. Thus the sum we wish to bound becomes

$$\sum_{\substack{\max(a,b)\times\max(g,h)\le x \\ a\neq b,\ (a,b)=1 \\ P(ab)\le \min(P(g),P(h))}} e((g-h)(a-b)\theta). \tag{8.2}$$

Since $\max(a,b) \times \max(g,h) \le x$, we may break the sum above into the cases (1) when $\max(g,h) \le \sqrt{x}$, (2) when $\max(a,b) \le \sqrt{x}$, taking care to subtract the terms satisfying (3) $\max(a,b)$ and $\max(g,h)$ both below $\sqrt{x}$.

Before turning to these cases, we record a preliminary lemma which will be useful in our analysis.

Lemma 8.2. *Let θ be an irrational number satisfying the Diophantine condition (1.4). Let $\mathcal{L} = \mathcal{L}(x)$ denote the set of all integers ℓ with $|\ell| \le \sqrt{x}$ such that for some $v \le (\log x)^5$ one has $\|v\ell\theta\| \le x^{-\frac{1}{3}}$. Then 0 is in $\mathcal{L}$, and for any two distinct elements ℓ_1, $\ell_2 \in \mathcal{L}$ we have $|\ell_1 - \ell_2| \gg (\log x)^5$.*

Proof. Evidently 0 is in $\mathcal{L}$, and the main point is the spacing condition satisfied by elements of $\mathcal{L}$. If ℓ_1 and ℓ_2 are distinct elements of $\mathcal{L}$ then there exist v_1, $v_2 \le (\log x)^5$ with $\|v_1\ell_1\theta\| \le x^{-\frac{1}{3}}$ and $\|v_2\ell_2\theta\| \le x^{-\frac{1}{3}}$. It follows that $\|v_1v_2(\ell_1-\ell_2)\theta\| \le 2(\log x)^5 x^{-\frac{1}{3}}$. The desired bound on $|\ell_1 - \ell_2|$ now follows from the Diophantine property that we required of θ, namely that $\|q\theta\| \gg \exp(-q^{\frac{1}{50}})$. □

8.1. **Case 1:** $\max(g,h) \le \sqrt{x}$**.** Suppose that g and h are given with g and h below $\sqrt{x}$, and consider the sum over a and b in (8.2). We distinguish two sub-cases, depending on whether $g-h$ lies in $\mathcal{L}$ or not. Consider first the situation when $g-h \notin \mathcal{L}$. Using Möbius inversion to detect the condition that $(a,b)=1$, the sums over a and b may be expressed as (the $O(1)$ error term accounts for the term $a=b=1$ which must be omitted)

$$\sum_{\substack{k\le x/\max(g,h)\\ P(k)\le \min(P(g),P(h))}} \mu(k) \sum_{\substack{r,s\le x/(k\max(g,h))\\ P(r),P(s)\le \min(P(g),P(h))}} e(k(g-h)(r-s)\theta) + O(1)$$

$$= \sum_{\substack{k\le x/\max(g,h)\\ P(k)\le \min(P(g),P(h))}} \mu(k) \Big| \sum_{\substack{r\le x/(k\max(g,h))\\ P(r)\le \min(P(g),P(h))}} e(k(g-h)r\theta)\Big|^2 + O(1). \tag{8.3}$$

If $k > (\log x)^2$ then we bound the sum over r above by $x/(k\max(g,h))$, and so these terms contribute to (8.3) an amount

$$\ll \sum_{k>(\log x)^2} \frac{x^2}{k^2\max(g,h)^2} \ll \frac{x^2}{(\log x)^2\max(g,h)^2}.$$

Now consider $k \le (\log x)^2$, and find (using Dirichlet's theorem) a rational approximation u/v to $k(g-h)\theta$ with $|k(g-h)\theta - u/v| \le 1/(vx^{\frac{1}{3}})$ and $v \le x^{\frac{1}{3}}$. Since $g-h \notin \mathcal{L}$ by assumption, it follows that $v \ge (\log x)^3$, and therefore an application of Lemma 8.1 shows that the sum over r in (8.3) is $\ll x/(k\max(g,h)\log x)$. Thus the terms $k \le (\log x)^2$ contribute to (8.3) an amount bounded by

$$\sum_{k\le(\log x)^2} \frac{x^2}{k^2\max(g,h)^2(\log x)^2} \ll \frac{x^2}{(\log x)^2\max(g,h)^2}.$$

Summing this over all $g,\ h \le \sqrt{x}$, we conclude that the contribution of terms with $\max(g,h)\le\sqrt{x}$ and $g-h\notin\mathcal{L}$ to (8.2) is

$$\ll \sum_{g,h\le\sqrt{x}} \frac{x^2}{(\log x)^2\max(g,h)^2} \ll \frac{x^2}{\log x}.$$

Now consider the contribution of the terms $\max(g,h)\le\sqrt{x}$ where $g-h$ lies in $\mathcal{L}$. Note that in (8.2) we allow for the possibility that $g=h$; we begin by estimating these terms (which could also be handled as in our argument for the terms in (8.1)). The terms $g=h\le\sqrt{x}$ give

$$\le \sum_{g\le\sqrt{x}} \Big(\sum_{\substack{a\le x/g\\ P(a)\le P(g)}} 1\Big)^2 \le \sum_{(\log x)^2\le g\le\sqrt{x}} \frac{x^2}{g^2} + \sum_{g\le(\log x)^2} \Psi(x/g,(\log x)^2) \ll \frac{x^2}{\log x}.$$

Now consider the terms with $g-h \in \mathcal{L}$ with $g-h \neq 0$. Bounding the sum over a and b trivially by $\leq (x/\max(g,h))^2$, we see that the contribution of these terms is

$$\ll \sum_{\substack{g\neq h\leq \sqrt{x}\\ g-h\in\mathcal{L}}} \frac{x^2}{\max(g,h)^2} \ll \sum_{g\leq\sqrt{x}} \frac{x^2}{g^2} \sum_{\substack{h<g\\ g-h\in\mathcal{L}}} 1 \ll \sum_{g\leq\sqrt{x}} \frac{x^2}{g^2}\frac{g}{(\log x)^5} \ll \frac{x^2}{(\log x)^4},$$

where we used Lemma 8.2 to bound the sum over h.

We conclude that the contribution of terms with $\max(g,h)\leq\sqrt{x}$ to (8.2) is $\ll x^2/\log x$, completing our discussion of this case.

8.2. **Case 2:** $\max\{a,b\}\leq\sqrt{x}$**.** Here we must bound

$$\sum_{\substack{a\neq b\leq\sqrt{x}\\ (a,b)=1}} \Big| \sum_{\substack{g\leq x/\max(a,b)\\ P(ab)\leq P(g)}} e(g(a-b)\theta)\Big|^2.$$

Again we distinguish the cases when $a-b\in\mathcal{L}$, and when $a-b\notin\mathcal{L}$. In the first case, we bound the sum over g above trivially by $\leq x/\max(a,b)$, and thus these terms contribute (using Lemma 8.2)

$$\ll \sum_{a\leq\sqrt{x}} \frac{x^2}{a^2} \sum_{\substack{b<a\\ a-b\in\mathcal{L}}} 1 \ll \sum_{a\leq\sqrt{x}} \frac{x^2}{a^2}\frac{a}{(\log x)^5} \ll \frac{x^2}{(\log x)^4}.$$

Now consider the case when $a-b\notin\mathcal{L}$. Using Dirichlet's theorem we may find a rational approximation u/v to $(a-b)\theta$ such that $|(a-b)\theta-u/v|\leq 1/(vx^{\frac{1}{3}})$ and $v\leq x^{\frac{1}{3}}$. Since $(a-b)\notin\mathcal{L}$, it follows that $v\geq(\log x)^5$. Therefore, two applications of Lemma 8.1 give

$$\sum_{\substack{g\leq x/\max(a,b)\\ P(g)\geq P(ab)}} e(g(a-b)\theta) = \sum_{g\leq x/\max(a,b)} e(g(a-b)\theta) - \sum_{\substack{g\leq x/\max(a,b)\\ P(g)<P(ab)}} e(g(a-b)\theta) \ll \frac{x}{\max(a,b)\log x}.$$

Thus the contribution of the terms with $a-b\notin\mathcal{L}$ is

$$\ll \sum_{a,b\leq\sqrt{x}} \frac{x^2}{(\log x)^2\max(a,b)^2} \ll \frac{x^2}{\log x}.$$

Thus the contribution to (8.2) from the Case 2 terms is $\ll x^2/\log x$.

8.3. **Case 3:** $\max(a,b)$ **and** $\max(g,h)\leq\sqrt{x}$**.** Here we must bound

$$\sum_{\substack{a\neq b\leq\sqrt{x}\\ (a,b)=1}} \Big| \sum_{\substack{g\leq\sqrt{x}\\ P(ab)\leq P(g)}} e(g(a-b)\theta)\Big|^2,$$

and our argument in Case 2 above furnishes the bound $\ll x^2/\log x$.

Combining our work in the three cases, we conclude that the quantity in (8.2) is $\ll x^2/\log x$. This verifies the second condition needed to apply Theorem 3.1 and completes our proof of Theorem 1.6.

9. Rademacher random multiplicative functions

In this section we briefly indicate the analogues of our results in the Rademacher model of random multiplicative functions, where $f(p) = \pm 1$ with equal probability (and chosen independently for different primes), and $f(n)$ is taken to be 0 if n has a square factor. In our work above, a key role was played by the fourth moment, which in the Steinhaus case led to solutions to $m_1m_2 = n_1n_2$ and to the notion of the multiplicative energy. If we consider the corresponding fourth moment in the Rademacher case, we are led to (with $\mathcal{A}$ denoting a set of square-free integers)

$$\mathbb{E}\Big[\Big(\sum_{n\in\mathcal{A}} f(n)\Big)^4\Big] = \sum_{\substack{n_1,n_2,n_3,n_4\in\mathcal{A}\\ n_1n_2n_3n_4=\square}} 1.$$

Thus the analogue of the multiplicative energy here is what may be termed the *square energy of* $\mathcal{A}$ namely:

$$E_\square(\mathcal{A}) := \#\{(n_1,n_2,n_3,n_4)\in\mathcal{A}^4 : n_1n_2n_3n_4 = \square\}.$$

Note that there are $\sim 3|\mathcal{A}|^2$ diagonal solutions, given by the pairings $n_1 = n_2$ and $n_3 = n_4$; $n_1 = n_3$ and $n_2 = n_4$; or $n_1 = n_4$ and $n_2 = n_3$. Taking this difference into account, and arguing as in our proof of Theorem 3.1 and the simplified Theorem 1.1 we can establish the following result (whose proof we omit).

Theorem 9.1. *Let $\mathcal{A}\subset[1,N]$ be a set of square-free integers with*

$$|\mathcal{A}| \geq N\exp(-\tfrac{1}{3}\sqrt{\log N\log\log N}).$$

Suppose that there exists a subset $\mathcal{S}\subset\mathcal{A}$ with $|\mathcal{S}| = (1+o(1))|\mathcal{A}|$ and satisfying

$$E_\square(\mathcal{S}) = (3+o(1))|\mathcal{S}|^2.$$

Then as f ranges over Rademacher random multiplicative functions, the quantity

$$\frac{1}{\sqrt{|\mathcal{A}|}}\sum_{n\in\mathcal{A}} f(n)$$

is distributed like a standard (real) normal random variable with mean 0 and variance 1.

With suitable modifications to the proofs, the results that we have enunciated for the Steinhaus model would extend to the Rademacher case. We sketch one example, treating the distribution of Rademacher random multiplicative functions in short intervals, which is an analogue of Corollary 1.2 and improves upon the earlier work in [3].

Corollary 9.2. *Let x and y be large, with $x^{1/5}\log x \ll y \leq x/(\log x)^{\alpha-\epsilon}$ where $\alpha = 2\log 2 - 1$. Let $\mathcal{T}$ denote the set of all square-free integers in $[x, x+y]$. Then, for a random Rademacher multiplicative function f, the quantity*

$$\frac{1}{\sqrt{|\mathcal{T}|}}\sum_{x\leq n\leq x+y} f(n),$$

is distributed like a standard normal random variable with mean 0 and variance 1.

In [3] such a result was established in the range $x^{\frac{1}{5}}\log x \ll y = o(x/\log x)$. The lower bound on y is to ensure that the interval $[x, x+y]$ contains $\gg y$ square-free integers (which follows from [4]). Thus we restrict attention to the range $x/(\log x)^2 \le y \le x/(\log x)^{\alpha-\epsilon}$. In this range we choose $\mathcal{S}$ to be the set of integers $n \in \mathcal{T}$ with $\Omega(n) \le (1+\epsilon)\log\log x$. Then, as in the proof of Proposition 4.1, we have $|\mathcal{T}\backslash\mathcal{S}| = o(y)$, and we need only show that the number of non-diagonal solutions to $n_1n_2n_3n_4 = \square$ (with $n_i \in \mathcal{S}$) is $o(y^2)$.

Write, as earlier, $y = \delta x$. Let r denote the gcd of n_1 and n_2, and let s denote the gcd of n_3 and n_4. Write $n_1 = ru_1$, $n_2 = ru_2$, and $n_3 = sv_1$ and $n_4 = sv_2$. Now u_1u_2 and v_1v_2 are square-free integers (since $(u_1, u_2) = (v_1, v_2) = 1$), and their product is a square, which means that $u_1u_2 = v_1v_2$. Using our parametrization in Lemma 3.4, write $u_1 = ga$, $u_2 = hb$, $v_1 = gb$, $v_2 = ha$, where now we also know that $(g, h) = (a, b) = 1$. Thus we have parametrized our solutions to $n_1n_2n_3n_4 = \square$ as $n_1 = rga$, $n_2 = rhb$, $n_3 = sgb$, $n_4 = sha$. If two out of the three possibilities $r = s$, $g = h$, or $a = b$ occur, then we obtain diagonal solutions; therefore we may assume that at most one of the equalities $r = s$, $g = h$, or $a = b$ can occur.

Since the variables n_i must all lie in $[x, x+y]$, it follows that

$$\frac{n_1n_2}{n_3n_4} = \frac{r^2}{s^2} \in [(1+\delta)^{-2}, (1+\delta)^2], \qquad \text{or} \qquad \frac{r}{s} \in [(1+\delta)^{-1}, (1+\delta)].$$

In particular, either $r = s$, or both r and s are $\gg 1/\delta$. Similarly we also find that g/h and a/b must lie in $[(1+\delta)^{-1}, (1+\delta)]$, so that either $g = h = 1$ or g, $h \gg 1/\delta$, and either $a = b = 1$ or a, $b \gg 1/\delta$.

Case 1: $g = h = 1$, or $a = b = 1$. If $g = h = 1$, then we must bound the number of non-diagonal solutions to $n_1n_3 = n_2n_4$, and our work in Proposition 4.1 shows that this is $o(y^2)$. Similarly if $a = b = 1$ then we have non-diagonal solutions to $n_1n_4 = n_2n_3$, and these again are $o(y^2)$.

Case 2: $r = s$. This is a little different from Case 1, since we may have $r = s$ without both being necessarily 1. Suppose first that $r = s \le \sqrt{x}$. Here note that we are counting off-diagonal solutions to $u_1u_2 = v_1v_2$, where u_1, u_2, v_1, v_2 are in $[x/r, (x+y)/r]$, and $\Omega(u_1)$, $\Omega(u_2)$, $\Omega(v_1)$, $\Omega(v_2)$ are all below $(1+\epsilon)\log\log x$ which is $\le (1+2\epsilon)\log\log(x/r)$. Therefore, Proposition 4.1 applies here to show that the number of non-diagonal choices for u_1, u_2, v_1, v_2 is $o(y^2/r^2)$, and summing over $r \le \sqrt{x}$ produces a bound of $o(y^2)$ for this count.

Now suppose that $r = s > \sqrt{x}$. Here we have $\le (y/r+1)$ choices for u_1 and u_2, and then v_1 and v_2 are fixed in $O(d(u_1u_2)) = O(x^\epsilon)$ ways. Therefore these terms contribute

$$\ll \sum_{\sqrt{x} \le r = s \le x} \Big(\frac{y^2}{r^2} + 1\Big)x^\epsilon = o(y^2).$$

Case 3: $r \ne s$, $g \ne h$, $a \ne b$. We now ignore the conditions that $(g, h) = (a, b) = 1$, so that the pairs r, s; g, h; and a, b are now all on an equal footing. Since $\Omega(rga)$, $\Omega(rhb)$, $\Omega(sgb)$, $\Omega(sha)$ are all assumed to be $\le K := (1+\epsilon)\log\log x$, it follows that $\Omega(r) + \Omega(s) + \Omega(g) + \Omega(h) + \Omega(a) + \Omega(b) \le 2K$. Without loss of generality we may assume that $\max(r, s) \ll$

$\max(g,h) \ll \max(a,b)$, so that $\max(r,s) \ll x^{\frac{1}{3}}$, and $\max(g,h) \ll (x/\max(r,s))^{\frac{1}{2}}$. Therefore it is enough to bound

$$\sum_{\substack{1/\delta \ll r,s \ll x^{\frac{1}{3}} \\ r/s \in [(1+\delta)^{-1},(1+\delta)]}} \sum_{\substack{1/\delta \ll g,h \ll (x/\max(r,s))^{\frac{1}{2}} \\ g/h \in [(1+\delta)^{-1},(1+\delta)]}} \sum_{\substack{a \in [x/(rg),(x+y)/(rg)] \\ b \in [x/(sh),(x+y)/(sh)]}} 2^{2K-\Omega(r)-\Omega(s)-\Omega(g)-\Omega(h)-\Omega(a)-\Omega(b)}.$$

Using Shiu's Lemma 4.2 to bound the sums over a and b, we see that the above may be bounded by

$$\ll \frac{2^{2K}y^2}{\log x} \sum_{\substack{1/\delta \ll r,s \ll x^{\frac{1}{3}} \\ r/s \in [(1+\delta)^{-1},(1+\delta)]}} \frac{2^{-\Omega(r)-\Omega(s)}}{r^2} \sum_{\substack{1/\delta \ll g,h \ll x^{\frac{1}{2}} \\ g/h \in [(1+\delta)^{-1},(1+\delta)]}} \frac{2^{-\Omega(g)-\Omega(h)}}{g^2}.$$

The sums over r, s, and g, h above are nearly identical to the sums over g and h appearing in (4.5), and thus may be bounded by $\ll (\delta \log\log x)$ (using our work leading up to (4.6)). Thus, we conclude that the number of non-diagonal solutions counted in this case is

$$\ll \frac{2^{2K}y^2}{\log x}(\delta \log\log x)^2 = o(y^2).$$

This completes our proof of Corollary 9.2.

References

[1] J. Benatar, A. Nishry, and B. Rodgers. Moments of polynomials with random multiplicative coefficients. *Mathematika*, 68(1):191–216, 2022.

[2] R. Blei and S. Janson. Rademacher chaos: tail estimates versus limit theorems. *Ark. Mat.*, 42(1):13–29, 2004.

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

[4] M. Filaseta and O. Trifonov. On gaps between squarefree numbers. II. *J. London Math. Soc. (2)*, 45(2):215–221, 1992.

[5] K. Ford. Extremal properties of product sets. *Proc. Steklov Inst. Math.*, 303(1):220–226, 2018. Published in Russian in Tr. Mat. Inst. Steklova **3**03 (2018), 239–245.

[6] A. Gut. *Probability: a graduate course*. Springer Texts in Statistics. Springer, New York, second edition, 2013.

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

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

[9] H. Helson. Hankel forms. *Studia Math.*, 198(1):79–84, 2010.

[10] A. Hildebrand. Integers free of large prime divisors in short intervals. *Quart. J. Math. Oxford Ser. (2)*, 36(141):57–69, 1985.

[11] C. Hooley. On the intervals between numbers that are sums of two squares. III. *J. Reine Angew. Math.*, 267:207–218, 1974.

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

[13] O. Klurman, I. D. Shkredov, and M. W. Xu. On the random Chowla conjecture. *Geom. Funct. Anal.*, 33(3):749–777, 2023.
[14] Y.-K. Lau, G. Tenenbaum, and J. Wu. On mean values of random multiplicative functions. *Proc. Amer. Math. Soc.*, 141(2):409–420, 2013.
[15] D. Mastrostefano. On maximal product sets of random sets. *J. Number Theory*, 224:13–40, 2021.
[16] D. L. McLeish. Dependent central limit theorems and invariance principles. *Ann. Probability*, 2:620–628, 1974.
[17] H. L. Montgomery and R. C. Vaughan. Exponential sums with multiplicative coefficients. *Invent. Math.*, 43(1):69–82, 1977.
[18] M. Pandey, V. Y. Wang, and M. W. Xu. Partial sums of typical multiplicative functions over short moving intervals. *Algebra & Number Theory*, to appear. arXiv:2207.11758.
[19] C. Pomerance and A. Sárközy. On products of sequences of integers. In *Number theory, Vol. I (Budapest, 1987)*, volume 51 of *Colloq. Math. Soc. János Bolyai*, pages 447–463. North-Holland, Amsterdam, 1990.
[20] A. Selberg. Note on a paper by L. G. Sathe. *J. Indian Math. Soc. (N.S.)*, 18:83–87, 1954.
[21] P. Shiu. A Brun-Titchmarsh theorem for multiplicative functions. *J. Reine Angew. Math.*, 313:161–170, 1980.
[22] M. W. Xu and Y. Zhou. On product sets of arithmetic progressions. *Discrete Anal.*, pages Paper No. 10, 31, 2023.

Department of Mathematics, Stanford University, Stanford, CA, USA
Email address: ksound@stanford.edu

Department of Mathematics, Stanford University, Stanford, CA, USA
Email address: maxxu@stanford.edu