ALEA, Lat. Am. J. Probab. Math. Stat. **20**, 713–730 (2023)
DOI: 10.30757/ALEA.v20-27

A generalized Kubilius-Barban-Vinogradov bound for prime multiplicities

Louis H.Y. Chen, Arturo Jaramillo and Xiaochuan Yang

Department of Mathematics, National University of Singapore, Block S17, 10 Lower Kent Ridge Road, Singapore 119076.
E-mail address: matchyl@nus.edu.sg
URL: https://blog.nus.edu.sg/louischen/

Centro de Investigación en Matemáticas, Jalisco S/N, Col. Valenciana 36023 Guanajuato, Gto.
E-mail address: jagil@cimat.mx
URL: https://www.cimat.mx/ jagil

Department of Mathematics, Brunel University London, Uxbridge UB83PN, United Kingdom
E-mail address: xiaochuan.yang@brunel.ac.uk
URL: http://www.xiaochuanyang.com/

Abstract. We present an assessment of the distance in total variation of *arbitrary* collections of prime factor multiplicities of a random number in $[n] = \{1, \ldots, n\}$ and a collection of independent geometric random variables. More precisely, we impose mild conditions on the probability law of the random sample and the aforementioned collection of prime multiplicities, for which a fast decaying bound on the distance towards a tuple of geometric variables holds. Our results generalize and complement those from Kubilius (1964) and Barban and Vinogradov (1964) which consider the particular case of uniform samples in $[n]$ and collection of "small primes". As applications, we show a generalized version of the celebrated Erdös Kac theorem for not necessarily uniform samples of numbers.

1. Introduction

1.1. *Overview.* Let $\mathcal{P}$ denote the set of prime numbers. For a given $n \in \mathbb{N} = \{1, 2, ...\}$, we consider a random variable J_n supported on $[n]$. The goal of this paper is to study asymptotic properties (as n tends to infinity) of the p-adic valuations of J_n, denoted by $\{v_p^n \ ; \ p \in \mathcal{P}\}$ and characterized

Received by the editors January 19th, 2022; accepted October 21st, 2022.
2010 *Mathematics Subject Classification.* 60B12, 11K65.
Key words and phrases. Limit theorems, Probabilistic number theory.
First Author has been supported by some agency. Second Author and Third Author have been supported by some other grant.
This research was supported by FNR Grant R-AGR-3410-12-Z (MISSILe) from the University of Luxembourg and partially supported by Grant R-146-000-230-114 from the National University of Singapore.

by the identity

$$J_n = \prod_{p \in \mathcal{P}} p^{v_p^n}.$$

More precisely, we will determine general conditions over J_n that will allow us to approximate the law of a random vector of the form $\boldsymbol{v}^n = (v_p^n \ ; \ p \in \Gamma_n)$, where Γ_n is a finite subset of $\mathcal{P}$, whose cardinality satisfies suitable growth conditions. As applications of our results, we will prove a version of Erdös-Kac theorem, valid for non-necessarily uniform samples, as well as a Poisson point process approximation for the configuration of non-trivial multiplicities of small primes (see Section 3 for details). Our approach relies on very simple probabilistic and combinatorial arguments, which makes the proof remarkably accessible; although the the applications will typically make use of elementary results from number theory, such as Mertens' formula. It is worth mentioning that one of the three main steps in our proof (see Lemma 2.3) relies on the celebrated Bonferroni inequalities, which are a tool that is close in spirit to classical number theoretical arguments from sieve method. The rest of the proof uses ideas from large deviation estimates, such as the Chernoff bounds.

Thoughout the paper, $[x]$ denotes the set $\{1, ..., \lfloor x \rfloor\}$ for any $x > 0$ and $\{a|b\}$ denotes the event that a divides b for $a, b \in \mathbb{N}$. We write $v_p = v_p^n$ for simplicity.

Motivations and the uniform distribution case

Since the influential manuscript Kubilius (1964), the use of the random vector $\boldsymbol{v}^n$ as a tool for studying divisibility properties of J_n has gained particular traction in probabilistic number theory, as these objects naturally emerge in the study of arithmetic additive functions of uniform samples. Our manuscript takes Kubilius (1964) as starting point. There it was proved that in the particular case where J_n has uniform distribution over $[n]$ and $\Gamma_n = \mathcal{P} \cap [n^{1/\beta_n}]$ for a sequence $\beta_n > 0$ converging to infinity, there exist constants $C, \delta > 0$ such that

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \leq C e^{-\delta \beta_n}, \tag{1.1}$$

where $\boldsymbol{g}^n = (g_p \ ; \ p \in \Gamma_n)$ is a random vector whose entries are independent geometric random variables with

$$\mathbb{P}[g_p = k] = p^{-k}(1 - p^{-1}),$$

for $k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$. The aforementioned bound for $d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$, combined with elementary probabilistic tools, leads to very powerful results in probabilistic number theory. To exemplify this, we would like to mention that (1.1) can be used to obtain a quantitative assessment of the rate of convergence in the celebrated Erdös-Kac theorem, which establishes the asymptotic normality for the number of prime divisors of a uniform sample of $[n]$. The precise statement of Erdös-Kac theorem, as well as its link to the inequality (1.1) and some further generalizations will be presented in Section 3.

Since their publication, the results from Kubilius (1964) have been extended and generalized in several directions. Next we briefly mention some of them. In Barban and Vinogradov (1964) (see also Elliott (1980, Chapter 3)), the bound (1.1) was improved to

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \leq C(e^{-\frac{1}{8}\beta_n \log(\beta_n)} + n^{-\frac{1}{15}}), \tag{1.2}$$

for some (possibly different) constant $C > 0$. In the subsequent papers Elliott (1980) and Arratia et al. (1999) it was proved that if the sequence $\{\beta_n\}_{n \geq 1}$ is bounded, then the left hand side of (1.2) can not go to zero. We would also like to refer the reader to Arratia and Tavaré (1992) and Arratia and Stark (1999) where an analysis of the case of the case where $\beta_n \equiv u$ for all $n \in \mathbb{N}$ was carried out. In such instance it was shown that

$$\lim_n d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) = H(u) \in (0, \infty), \tag{1.3}$$

where H is defined by

$$H(u) := \frac{1}{2} \int_{\mathbb{R}} |\mathcal{B}(v) - e^{-\gamma}| \varrho(u - v) dv + \frac{1}{2} \varrho(u),$$

and $\mathcal{B}, \varrho, \gamma$ denote the Buschstav's function, Dickman's function and Euler's constant respectively. Finally, we would like to mention the paper Tenenbaum (1999), where it was proved that for every $\varepsilon > 0$, there exists a constant $C_\varepsilon > 0$ such that

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \leq C_\varepsilon(\varrho(\beta_n) 2^{(1+\varepsilon)\beta_n} + n^{-1+\varepsilon}).$$

It was proved as well in Tenenbaum (1999) that in the particular case where $\exp\{(\log\log(n))^{5/3} + \varepsilon\} \leq \beta_n \leq n$, the exact rate

$$|d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) - H(\beta_n)| = o(H(\beta_n)), \tag{1.4}$$

holds, and for every $\varepsilon > 0$,

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \leq C_\varepsilon(e^{-\beta_n \log(\beta_n)} + n^{-1+\varepsilon}).$$

The above results give a very complete picture of the behavior of $d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$, for the case where J_n is uniform in $[n]$ and the primes under consideration are all small, in the sense that $\Gamma_n = \mathcal{P} \cap [n^{1/\beta_n}]$. There has however been only little investigation into the problem of giving explicit rates of $d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$ for more general choices for Γ_n and J_n. In this paper we give a partial answer to this, as we address the following questions

- Up to what extent, the above results remain valid for different choices of Γ_n?
- How much flexibility do we have for choosing the distribution of J_n?

We would like to emphasize that the cases where Γ_n contains large primes are of special interest, as the complementary case is nearly completely described by the relation (1.4). We would like to bring the reader's attention to the manuscripts Arratia (2002) and Chen et al. (2022), where the study of divisibility properties for non-uniform samples J_n arise quite naturally in the investigation of probabilistic number theory problems. In both the papers Arratia (2002) and Chen et al. (2022), the sequence J_n is approximated with high degree of accuracy by a sequence of variables Y_n with simpler distributional laws, where the "simplicity in the law" is described by its embedding into a Poisson point process functional in Arratia (2002) and by regarding it as a realization from a sequence of independent variables conditioned over an adequate event in Chen et al. (2022). The approach taken in the current manuscript balances differently the simplicity-accuracy tradeoff, as we compare the multiplicities of the sample with fully independent variables with geometric distribution. The price we pay for obtaining such a strong simplification is an increase on the discrepancy between variables emerging from divisibility properties of J_n and those arising from independent geometric variables, which can be modulated (depending on the problem) by exploiting the flexibility on the choice of Γ_n.

Some heuristic considerations

In order to give an exploratory description of the nature of bounds of the type (1.1), we introduce some suitable notation. For any $D \subset \Gamma_n$ and $m = (m_p \;;\; p \in D) \in \mathbb{N}^D$, set $|m| := \sum_{p \in D} m_p$, $|D| := \sharp D$, $m + 1 := (m_p + 1 \;;\; p \in D)$ and

$$p_D = \prod_{p \in D} p, \qquad\qquad p_D^m = \prod_{p \in D} p^{m_p}$$

with the convention that $p_D = p_D^m = 1$ if $D = \emptyset$. Define as well the events

$$A(D,m) = \{v_p = m_p \text{ for all } p \in D \text{ and } v_q = 0 \text{ for all } q \in \Gamma_n \setminus D\},$$
$$\tilde{A}(D,m) = \{g_p = m_p \text{ for all } p \in D \text{ and } g_q = 0 \text{ for all } q \in \Gamma_n \setminus D\}.$$

This way, we can write

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) = \frac{1}{2} \sum_{D \subset \Gamma_n} \sum_{m \in \mathbb{N}^D} |\mathbb{P}[A(D,m)] - \mathbb{P}[\tilde{A}(D,m)]|. \tag{1.5}$$

Formula (1.5) reduces the problem to estimating $\mathbb{P}[A(D,m)]$ and $\mathbb{P}[\tilde{A}(D,m)]$. One way of doing this, is decomposing the events $A(D,m)$ and $\tilde{A}(D,m)$ as unions and intersections of "elemental events", having the property that their probabilities are easy to approximate. Throughout this paper, the elemental events that will serve for approximating $\mathbb{P}[A(D,m)]$ will consist of the elements of the π-system

$$\begin{aligned}\mathfrak{E} &:= \{\{v_{p_1} \geq m_1, \dots, v_{p_r} \geq m_r\};\ p_1, \dots, p_r \in \mathcal{P} \text{ and } m_1, \dots, m_r \in \mathbb{N}\} \\ &= \{\{p_1^{m_1} \cdots p_r^{m_r} | J_n\};\ p_1, \dots, p_r \in \mathcal{P} \text{ and } m_1, \dots, m_r \in \mathbb{N}\},\end{aligned}$$

while those used to approximate $\mathbb{P}[\tilde{A}(D,m)]$ will consist on the elements of the π-system

$$\tilde{\mathfrak{E}} := \{\{g_{p_1} \geq m_1, \dots, g_{p_r} \geq m_r\};\ p_1, \dots, p_r \in \mathcal{P} \text{ and } m_1, \dots, m_r \in \mathbb{N}\}.$$

This choice is justified by the fact that for every $p_1, \dots, p_r \in \mathcal{P}$ satsifying $p_i \neq p_j$ and $m_1, \dots, m_r \in \mathbb{N}$,

$$\mathbb{P}[g_{p_1} \geq m_1, \dots, g_{p_r} \geq m_r] = \frac{1}{d}, \tag{1.6}$$

where $d := \prod_{i=1}^r p_i^{m_i}$. In order for our heuristic to be accurate, we are required to impose a condition that guarantees that when the variables g_p appearing in (1.6) are replaced by v_p, the associated probability remains approximately equal to $\frac{1}{d}$. Motivated by this, we will assume that J_n and $\{g_p\ ;\ p \in \mathcal{P}\}$ are defined in a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the following hypothesis holds:

($\mathbf{H}_t$) There exist finite constants $\kappa \geq 1, t > 0$ independent of n, such that J_n is supported in $[0,n]$ and for every $a \in \mathbb{N}$,

$$|\mathbb{P}[a \mid J_n] - \frac{1}{a}| \leq \frac{\kappa}{n^t} \quad \text{and} \quad \mathbb{P}[a \mid J_n] \leq \frac{1+\kappa}{a}. \tag{1.7}$$

Despite the fact that condition ($\mathbf{H}_t$) imposes a mildly rigid condition over the law of J_n, it does include the following rich family of probability laws as particular instances.

Example 1.1. Consider the truncated "Pareto type" distribution

$$\pi_{n,s}(k) = \frac{1}{Z_{n,s}} k^{-s} \quad k \in [n], \text{ and } s \in [0,1),$$

where $Z_{n,s} = \sum_{k \in [n]} k^{-s}$. We claim that $J_n \sim \pi_{n,s}$ satisfies ($\mathbf{H}_{1-s}$) with $\kappa = 3$, for all n large. Indeed, we have

$$|\mathbb{P}[a|J_n] - a^{-1}| = \frac{1}{Z_{n,s}} \Big| \sum_{k=1}^{\lfloor n/a \rfloor} (ak)^{-s} - a^{-1} \sum_{k=1}^{n} k^{-s} \Big| \tag{1.8}$$

Notice that for any $a \in [n]$,

$$\sum_{k=1}^{\lfloor n/a \rfloor} (ak)^{-s} - a^{-1} \sum_{k=1}^{n} k^{-s} \leq \int_0^{n/a} (ax)^{-s} dx - a^{-1} \int_1^{n+1} x^{-s} dx \leq a^{-1} \int_0^1 x^{-s} dx \leq (1-s)^{-1}.$$

Similarly, for any $a \in [n]$,

$$a^{-1}\sum_{k=1}^{n} k^{-s} - \sum_{k=1}^{\lfloor n/a \rfloor} (ak)^{-s} \leq a^{-1}\int_0^{n-1} x^{-s} ds - \int_1^{n/a} (ax)^{-s} dx \leq a^{-1}\int_0^{a} x^{-s} dx \leq 1 + (1-s)^{-1}.$$

It remains to bound from below $Z_{n,s}$. We have

$$Z_{n,s} \geq \int_1^{n+1} x^{-s} ds = (1-s)^{-1}(n^{1-s} - 1)$$

Plugging these estimates in (1.8) proves that the first bound of $(\mathbf{H}_{1-s})$ holds. The other bound follows from analogous integral approximation arguments.

Remark 1.2. Distributions considered in the previous example are not asymptotically uniform in the sense that their total variation distance as $n \to \infty$ does not converge to 0. To verify this, we simply compute the sum

$$\sum_{k\in[n/M]} \Big|\frac{k^{-s}}{Z_{n,s}} - \frac{1}{n}\Big|, \quad M > 1.$$

For any $s \in [0,1)$, by choosing M large enough, we see that the summands are bounded from below by c/n, where $c > 0$ depends only on s and M. This gives a lower bound $c/M > 0$ for the total variation distance between the truncated Pareto distributions and the uniform distribution over $[n]$. Therefore, the conclusions of our main result cannot be derived from those in the literature for a uniform distribution in $[n]$.

Example 1.3. Let $\upsilon : [0,\infty) \to \mathbb{R}_+$ be a continuously differentiable function that is monotone over an interval of the form $[M,\infty)$, for some $M \in \mathbb{N}$, with

$$\sup_{n>M} n^t \int_0^1 |\upsilon(sn) - L| ds < \infty,$$

for some $L > 0$ and $t > 0$. If the law of J_n is supported in $\{1,\dots,n\}$ and satisfies

$$\mathbb{P}[J_n = k] := c_n \upsilon(k),$$

for some $k = 1,\dots,n$ and $c_n > 0$, then the variables J_n satisfy condition $(\mathbf{H}_t)$ for n sufficiently large. To verify this, observe that for every natural number $a \in \mathbb{N}$,

$$\mathbb{P}[a \mid J_n] = n^t c_n \left(\frac{1}{n^t}\sum_{j=1}^{\lfloor n/a \rfloor} \upsilon(aj)\right). \tag{1.9}$$

By elementary algebraic manipulations,

$$\frac{1}{n^t}\sum_{j=1}^{\lfloor n/a \rfloor} \upsilon(aj) = O(1/n^t) + \frac{1}{n^t}\sum_{j=M}^{\lfloor n/a \rfloor} \upsilon(aj) = O(1/n^t) + \frac{1}{n}\int_M^{n/a} \upsilon(as) ds = O(1/n^t) + L/a,$$

where $O(1/n^t)$ denote error functions bounded in absolute value by a constant multiple of $1/n^t$, independent of a. By choosing $a = 1$, we get that $n^t c_n = L + O(1/n^t)$. Condition $(\mathbf{H}_t)$ then follows from (1.9). Some particular instances in which the above conditions hold are the case where J_n has uniform distribution over $\{1,\dots,n\}$ and more generally, the case where there exist a non-increasing function $\theta : \mathbb{R}_+ \to \mathbb{R}_+$, with

$$\sup_{n\geq 1} k^t \theta(k) < \infty,$$

as well as constants $\varepsilon > 0$ and $c_n > 0$ such that $\mathbb{P}[J_n = k] = c_{\alpha,n}(\varepsilon + \theta(k))$ for all $k = 1,\dots,n$ and $\mathbb{P}[J_n = k] = 0$ otherwise.

Example 1.4. A generalization of the above two examples gives rise the following large family of probability distributions satisfying $(\mathbf{H}_t)$. Let $\upsilon : [0, \infty) \to \mathbb{R}_+$ be a continuously differentiable function that is monotone over an interval of the form $[M, \infty)$, for some $M \in \mathbb{N}$, with

$$\sup_{n>M} \frac{1}{c_n} \int_0^1 |\upsilon(sn)| ds < \infty,$$

for some $L > 0$ and $t \in (0, 1)$, where

$$c_n := \left(\sum_{k=1}^{n} \upsilon(k) \right)^{-1}.$$

If the law of J_n is supported in $\{1, \dots, n\}$, c_n has a decay of the order n^{-t} and

$$\mathbb{P}[J_n = k] := c_n \upsilon(k),$$

for some $k = 1, \dots, n$ and $c_n > 0$, then the variables J_n satisfy condition $(\mathbf{H}_t)$ for n sufficiently large. The proof of this claim is identical to that of Example 1.3, with the exception that n^t should be replaced by c_n^{-1} and L should be replaced by zero.

Notice that

$$\begin{aligned} A(D, m) &= \{v_p \geq m_p \ ; \ p \in D\} \backslash \bigcup_{p \in \mathcal{P}} \{v_p \geq m_p + 1 \ ; \ p \in D\} \cup \{v_q \geq 1 \ ; \ q \in D^c\} \\ \tilde{A}(D, m) &= \{g_p \geq m_p \ ; \ p \in D\} \backslash \bigcup_{p \in \mathcal{P}} \{g_p \geq m_p + 1 \ ; \ p \in D\} \cup \{g_q \geq 1 \ ; \ q \in D^c\}, \end{aligned}$$

where D^c denotes the complement relative to Γ_n, namely, $D^c := \Gamma_n \backslash D$. Consequently, by an elementary application of the inclusion-exclusion principle,

$$\begin{aligned} \mathbb{P}[A(D, m)] &= \sum_{I \subset \Gamma_n} (-1)^{|I|} \mathbb{P}[\{p_D^m | J_n\} \cap \{p_{D \cap I}^{m+1} p_{D^c \cap I} | J_n\}] \\ &= \sum_{I \subset \Gamma_n} (-1)^{|I|} \mathbb{P}[p_{D \cap I^c}^m p_{D \cap I}^{m+1} p_{D^c \cap I} | J_n] = \sum_{I \subset \Gamma_n} (-1)^{|I|} \mathbb{P}[p_D^m p_I | J_n] \end{aligned}$$

for any $D \subset \Gamma_n$ and $m \in \mathbb{N}^D$. Similarly,

$$\mathbb{P}[\tilde{A}(D, m)] = \sum_{I \subset \Gamma_n} \frac{(-1)^{|I|}}{p_D^m p_I}.$$

Thus, by using the hypothesis $(\mathbf{H}_t)$, we obtain the bound

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \leq \frac{1}{2} \sum_{D \subset \Gamma_n} \sum_{m \in \mathbb{N}_+^D} \sum_{I \subset \Gamma_n} (\mathbb{1}_{\{p_D^m p_I \leq n\}} \frac{1}{n} + \mathbb{1}_{\{p_D^m p_I > n\}} \frac{1}{p_D^m p_I}).$$

By using the geometric sum formula for the second term and the fact that there are at most $\frac{1}{2}\log(n)$ values for m for which $p_D^m p_I \leq n$, we deduce that there exists a universal constant $C > 0$, independent of Γ_n or n, such that

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \leq C \sum_{D \subset \Gamma_n} \sum_{I \subset \Gamma_n} \frac{\log(n)}{n} \leq C 4^{|\Gamma_n|} \frac{\log(n)}{n}. \tag{1.10}$$

In particular, when $\sup_n |\Gamma_n| < \infty$, one deduces that $d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$ is of the order $\frac{\log(n)}{n}$. Naturally, one wonders if in the case where $|\Gamma_n|$ converges to infinity, the above argument leads to a "good" bound for $d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$. Unfortunately, the right hand side of (1.10) diverges if $|\Gamma_n|$ is asymptotically larger than $\frac{\log(n)}{\log(4)}$, which restricts significantly the possible choices of Γ_n.

In this paper, we make suitable adjustments to the argument above, so that we obtain a bound that converges to zero even when $|\Gamma_n|$ is polynomial of any degree in $\log(n)$, see Remark 1.8 for more details. The main idea consists of replacing the use of the inclusion-exclusion principle by Bonferonni-type bounds. This, combined with a careful combinatorial analysis of the resulting terms, leads to a near to optimal bound for $d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$.

1.2. *Statement of the main result.* Recall that J_n is assumed to satisfy the condition $(\mathbf{H}_t)$. To state the main theorem, we introduce the quantities

$$\tau_n := \sum_{p\in\Gamma_n} \frac{1}{p} \qquad \text{and} \qquad \rho_n := \frac{\log n}{\log|\Gamma_n|}.$$

Theorem 1.5. *Suppose that* $\lim_{n\to\infty} \rho_n = \infty$ *and that there exists* $\varepsilon \in (0,1)$*, such that* $|\Gamma_n| \le n^{\tau_n^{-1-\varepsilon}}$ *for all* $n \in \mathbb{N}$*. Suppose that* $(\mathbf{H}_t)$ *holds with* $t > 0$*. Then for* n *sufficiently large, we have*

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \le (4+2\kappa)\exp\Big(-c\min\big(\rho_n\log\rho_n, \log(n)\big)\Big) \tag{1.11}$$

with $c = \frac{\varepsilon\wedge(t/4)}{2(1+\varepsilon)}$.

Remark 1.6. In the case where $\Gamma_n = \mathcal{P}\cap[n^{\frac{1}{\beta_n}}]$ for $\beta_n > 1$ converging to infinity, we have that $\lim_n n^{-\frac{1}{\beta_n}}\log(n^{\frac{1}{\beta_n}})|\Gamma_n| = 1$ due to the prime number theorem. From here it follows that $\lim_n \frac{\rho_n}{\beta_n} = 1$. Consequently, up to a change in the constant c, Theorem 1.5 improves Barban's bound (1.2) when ε is small enough and

$$\beta_n\log(\beta_n) \le \log(n),$$

for n sufficiently large.

Remark 1.7. If ρ_n is uniformly bounded over n, then $|\Gamma_n| \ge n^a$ for some fixed $a\in(0,1)$ independent of n. In this case, by (1.3), in order for a bound for the left hand side of (1.11) to converge to zero at the time that it generalizes the bounds presented by Kubilius and Barban bounds, we have to restrict ourselves to the regime $\lim_{n\to\infty}\rho_n = \infty$. This justifies the appearance of the condition $\lim_{n\to\infty}\rho_n = \infty$ as a hypothesis in Theorem 1.5. However, a bound for case $\sup_n \rho_n < \infty$ can be obtained by means of Equation (1.10).

Remark 1.8. By Mertens' formula (see Tenenbaum (1999, page 14)), we have that for n sufficiently large, $\tau_n \le \log\log(n) + 1$. Consequently, for any $K > 0$, we have that

$$n^{\tau_n^{-1-\varepsilon}} \ge e^{\frac{\log(n)}{(1+\log\log(n))^{1+\varepsilon}}} \ge [\log(n)]^K,$$

provided that n is sufficiently large. Thus, our condition on the cardinality of Γ_n is mild, as one only requires the condition $|\Gamma_n| \le n^{\frac{1}{(1+\log\log(n))^{1+\varepsilon}}}$ to be satisfied.

Remark 1.9. By considering Γ_n of small cardinality, we observe that one cannot surpass the barrier $\log(n)$ in the exponent. For instance, if

$$|\Gamma_n| \le \log\log(n),$$

then we have $\rho_n\log(\rho_n) \ge \log(n)\log\log(n)$ for n sufficiently large, and any estimate of the type $c_1\exp(-c_2\rho_n\log(\rho_n))$ decays faster than $c_1 e^{-c_2\log(n)\log\log(n)}$. The former quantity can't bound $d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$ since for every $p\in\mathcal{P}$,

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) = \frac{1}{2}\sum_{m\in\mathbb{N}_0^{\Gamma_n}} |\mathbb{P}[\boldsymbol{v}^n = m] - \mathbb{P}[\boldsymbol{g}^n = m]| \ge \frac{1}{2}\mathbb{P}\Big[g_p \ge \frac{\log(n)}{\log(p)}\Big],$$

so that

$$d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) \geq p^{-1-(\log(n)/\log(p))} = \frac{1}{pn}.$$

2. Proof of theorem 1.5

The main idea consists on decomposing the right hand side of (1.5) into three pieces, which heuristically correspond to the following instances

- The set $D \subset \Gamma_n$ of prime divisors has large cardinality, see Section 2.1.
- The set $D \subset \Gamma_n$ is small, but the associated multiplicities $m \in \mathbb{N}^D$ are large, see Section 2.2.
- Both the set $D \subset \Gamma_n$ and the multiplicities $m \in \mathbb{N}^D$ are small, see Section 2.3.

A suitable control for each of these instances will be considered in Lemmas 2.1,2.2, 2.3. Once we finish the proofs of such lemmas, they will be combined at the end of the section, to yield a complete proof of Theorem 1.5.

2.1. *Many distinct prime divisors.* First we consider D with large cardinality. We define the threshold α_n of the form

$$\alpha_n := \delta \rho_n, \tag{2.1}$$

where $\delta > 0$ is a positive constant independent of n. The constants α_n also depends on δ, but we omit it from the notation.

Lemma 2.1. *Suppose that the second bound of* $(\mathbf{H}_t)$ *holds. For any* $\delta > 0$ *and* $n \in \mathbb{N}$, *we have*

$$\begin{aligned}
&\sum_{\substack{D \subset \Gamma_n \\ |D| \geq \alpha_n}} \sum_{m \in \mathbb{N}^D} |\mathbb{P}[A(D,m)] - \mathbb{P}[\tilde{A}(D,m)]| \\
&\leq (2+\kappa) \exp\Big(-\frac{\delta\varepsilon}{1+\varepsilon} \rho_n \log(\rho_n) + \delta(1-\log(\delta))\rho_n \Big).
\end{aligned} \tag{2.2}$$

In particular, the condition $\rho_n \to \infty$ *implies that there exists a constant* $N \in \mathbb{N}$ *depending on* ε, δ, *such that for* $n \geq N$,

$$\frac{\delta\varepsilon}{1+\varepsilon} \rho_n \log(\rho_n) \geq 2|\delta(1-\log(\delta))|\rho_n,$$

so that inequality (2.2) *implies that for such* n,

$$\sum_{\substack{D \subset \Gamma_n \\ |D| \geq \alpha_n}} \sum_{m \in \mathbb{N}^D} |\mathbb{P}[A(D,m)] - \mathbb{P}[\tilde{A}(D,m)]| \leq (2+\kappa) \exp\Big(-\frac{\delta\varepsilon}{2(1+\varepsilon)} \rho_n \log(\rho_n) \Big).$$

Proof: The sum on the left-hand side is empty if $\alpha_n > |\Gamma_n|$ in which case the claim is trivial. Suppose that $\alpha_n \leq |\Gamma_n|$. We simply bound

$$|\mathbb{P}[\tilde{A}(D,m)] - \mathbb{P}[A(D,m)]| \leq \mathbb{P}[\tilde{A}(D,m)] + \mathbb{P}[A(D,m)]$$

and estimate the sum of probabilities separately. Notice that

$$\begin{aligned}
\sum_{\substack{D \subset \Gamma_n \\ |D| \geq \alpha_n}} \sum_{m \in \mathbb{N}^D} \mathbb{P}[A(D,m)] &= \sum_{\substack{D \subset \Gamma_n \\ |D| \geq \alpha_n}} \mathbb{P}[v_p \geq 1, \forall p \in D, v_q = 0, \forall q \in \Gamma_n \setminus D] \\
&\leq \sum_{\substack{D \subset \Gamma_n \\ |D| \geq \alpha_n}} \mathbb{P}[v_p \geq 1, \forall p \in D] = \sum_{\substack{D \subset \Gamma_n \\ |D| \geq \alpha_n}} \mathbb{P}[p_D | J_n] \leq \sum_{\substack{D \subset \Gamma_n \\ |D| \geq \alpha_n}} \frac{1+\kappa}{p_D},
\end{aligned}$$

where the last inequality follows from hypothesis $(\mathbf{H}_t)$. Observe that the condition $|\Gamma_n| \leq n^{\tau_n^{-(1+\varepsilon)}}$ implies that $\log(|\Gamma_n|) \leq \frac{\log(n)}{\tau_n^{1+\varepsilon}}$, which leads to the bound

$$\rho_n \geq \tau_n^{1+\varepsilon}. \tag{2.3}$$

Thus, by using Chernoff's bound (see Lemma A.2),

$$\sum_{\substack{D\subset\Gamma_n\\|D|\geq\alpha_n}} \frac{1}{p_D} = \sum_{j=\alpha_n}^{|\Gamma_n|} \frac{1}{j!} \sum_{(p_1,\dots,p_j)\in(\Gamma_n)^j_{\neq}} \frac{1}{p_1\cdots p_j} \leq \sum_{j=\alpha_n}^{\infty} \frac{\tau_n^j}{j!} \leq \Big(\frac{e\tau_n}{\alpha_n}\Big)^{\alpha_n}, \tag{2.4}$$

where $(\Gamma_n)^j_{\neq}$ is the set of j-tuples of elements in Γ_n with distinct coordinates. We observe by (2.3) that

$$\begin{aligned}\log\Big(\Big(\frac{e\tau_n}{\alpha_n}\Big)^{\alpha_n}\Big) &\leq \delta\rho_n\Big(1+\frac{1}{1+\varepsilon}\log(\rho_n)-\log(\delta)-\log(\rho_n)\Big)\\ &\leq -\frac{\delta\varepsilon}{1+\varepsilon}\rho_n\log(\rho_n)+\delta(1-\log(\delta))\rho_n,\end{aligned}$$

yielding the desired estimate for the sum of $\mathbb{P}[A(D,m)]$. The same argument gives an analogous bound for the sum of $\mathbb{P}[\tilde{A}(D,m)]$, ending the proof. $\square$

2.2. *Fewer prime divisors with overall high multiplicities.* Now we consider the case of fewer prime divisors $|D| \leq \alpha_n$. Recall that $|m| \geq |D|$ since $m \in \mathbb{N}^D$. We are interested in the situation where the prime divisors D have overall high multiplicities in the sense that $|m| \geq \beta_n$ with β_n much larger than α_n.

Lemma 2.2. *Suppose that the second bound of* $(\mathbf{H}_t)$ *holds. Let* α_n *be given by* (2.1) *and define*

$$\beta_n := \frac{2\delta}{(1+\varepsilon)\log(1.5)}\rho_n\log(\rho_n). \tag{2.5}$$

Then we have for all $n \in \mathbb{N}$,

$$\begin{aligned}&\sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}} |\mathbb{P}[\tilde{A}(D,m)]-\mathbb{P}[A(D,m)]|\\ &\leq (2+\kappa)\exp\Big(-\frac{\delta}{1+\varepsilon}\rho_n\log(\rho_n)+5\log(1.5)\delta\rho_n+\log(\delta\rho_n)\Big).\end{aligned} \tag{2.6}$$

In particular, the condition $\rho_n \to \infty$ *guarantees the existence of a constant* $N \in \mathbb{N}$ *depending on* ε, δ, *such that that for* $n \geq N$,

$$\frac{\delta}{1+\varepsilon}\rho_n\log(\rho_n) \geq 2|5\log(1.5)\delta\rho_n+\log(\delta\rho_n)|,$$

so that inequality (2.6) *implies that for such* n,

$$\sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}} |\mathbb{P}[\tilde{A}(D,m)]-\mathbb{P}[A(D,m)]| \leq (2+\kappa)\exp\Big(-\frac{\delta}{2(1+\varepsilon)}\rho_n\log(\rho_n)\Big).$$

Proof: As in the proof of Lemma 2.2, we simply estimate the difference of probabilities by their sum. We start by handling the terms involving the quantities $\mathbb{P}[\tilde{A}(D,m)]$. Notice that

$$\begin{aligned}\sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}}\mathbb{P}[\tilde{A}(D,m)] &\leq \sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\mathbb{P}\Big[\sum_{p\in D}g_p\geq\beta_n\,|\,g_p\geq 1,p\in D\Big]\mathbb{P}\Big[g_p\geq 1,p\in D\Big]\\ &= \sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\mathbb{P}\Big[\sum_{p\in D}\hat{g}_p\geq\beta_n\Big]\frac{1}{p_D},\end{aligned}$$

where $\mathcal{L}(\hat{g}_p,p\in D)=\mathcal{L}(g_p,p\in D\,|\,g_p\geq 1,p\in D)$. One readily checks that $(\hat{g}_p,p\in D)$ is a family of independent $\mathbb{N}$-valued geometric random variables with

$$\mathbb{P}[\hat{g}_p=k]=p^{-(k-1)}(1-p^{-1}),\quad k\in\mathbb{N},p\in D. \tag{2.7}$$

For any $p\in D$, a direct computation leads to the uniform bound $\mathbb{E}[(1.5)^{\hat{g}_p}]\leq 3$ and the following concentration bound

$$\mathbb{P}\Big[\sum_{p\in D}\hat{g}_p\geq\beta_n\Big]\leq(1.5)^{-\beta_n}3^{|D|}\leq(1.5)^{-(\beta_n-3\alpha_n)}$$

holds by Markov's inequality. This leads to

$$\begin{aligned}\sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}}\mathbb{P}[\tilde{A}(D,m)] &\leq (1.5)^{-(\beta_n-3\alpha_n)}\sum_{j=0}^{\alpha_n\wedge|\Gamma_n|}\frac{1}{j!}\sum_{(p_1,\dots,p_j)\in(\Gamma_n)^j_{\neq}}\frac{1}{p_1\cdots p_j}\\ &\leq (1.5)^{-(\beta_n-3\alpha_n)}\sum_{j=0}^{\alpha_n\wedge|\Gamma_n|}\frac{\tau_n^j}{j!}\leq\alpha_n(1.5)^{-(\beta_n-3\alpha_n)}(\tau_n)^{\alpha_n}. \end{aligned}\tag{2.8}$$

Now we move to the analysis of the sums involving $\mathbb{P}[A(D,m)]$. As before, we write

$$\begin{aligned}\sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}}\mathbb{P}[A(D,m)] &\leq \sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}}\mathbb{P}[v_p\geq m_p,\forall p\in D]\\ &\leq \sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}}\frac{1+\kappa}{p_D^m}\\ &= \sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\mathbb{P}\Big[\sum_{p\in D}\hat{g}_p\geq\beta_n\Big]\frac{1+\kappa}{p_D}\prod_{p\in D}\frac{1}{1-p^{-1}}. \end{aligned}\tag{2.9}$$

where (2.9) follows from the definition of $\hat{g}_p$ at (2.7). Bounding the product in (2.9) by $2^{|D|}\leq(1.5)^{2\alpha_n}$ and then using the argument leading to (2.8), we obtain

$$\sum_{\substack{D\subset\Gamma_n\\|D|\leq\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\|m|\geq\beta_n}}\mathbb{P}[A(D,m)]\leq(1+\kappa)\alpha_n(1.5)^{-(\beta_n-5\alpha_n)}(\tau_n)^{\alpha_n}.$$

To summarize, both sums have analogous upper bounds. It follows from (2.1)-(2.5) that

$$\begin{aligned}\log\Big(\alpha_n(1.5)^{-(\beta_n-5\alpha_n)}(\tau_n)^{\alpha_n}\Big) &\leq \log(\delta\rho_n)-(\beta_n-5\delta\rho_n)\log(1.5)+\frac{\delta}{1+\varepsilon}\rho_n\log(\rho_n)\\ &\leq -\frac{\delta}{1+\varepsilon}\rho_n\log(\rho_n)+5\log(1.5)\delta\rho_n+\log(\delta\rho_n).\end{aligned}$$

The conclusion follows immediately. $\square$

2.3. *Fewer prime divisors with moderate multiplicities.* Now we handle the remaining case where $|D| \leq \alpha_n$ and $|m| \leq \beta_n$. This is the only part where we seek cancellations between the probability mass functions of independent (i.e. $\boldsymbol{g}^n$) and dependent (i.e. $\boldsymbol{v}^n$) vectors.

Lemma 2.3. *Let α_n and β_n be given by (2.1) and (2.5) with $\delta \in (0, t/3)$. Suppose $(\mathbf{H}_t)$ holds with $t > 0$. For n sufficiently large, we have*

$$\sum_{\substack{D\subset\Gamma_n \\ |D|\leq\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D \\ |m|\leq\beta_n}} |\mathbb{P}[A(D,m)] - \mathbb{P}[\tilde{A}(D,m)] \leq (3+2\kappa)\exp(-c\min[\log(n), \rho_n\log(\rho_n)]) \tag{2.10}$$

with $c = \min(t - 3\delta, \frac{\delta\varepsilon}{2(1+\varepsilon)})$. In particular, the condition $\varepsilon \in (0,1)$ implies that $\frac{\delta\varepsilon}{2(1+\varepsilon)} \leq \frac{\delta}{4}$, so that $t - 3\delta \geq \frac{\delta\varepsilon}{2(1+\varepsilon)}$ for $\delta \leq 0.3t$, thus yielding the inequality

$$\sum_{\substack{D\subset\Gamma_n \\ |D|\leq\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D \\ |m|\leq\beta_n}} |\mathbb{P}[A(D,m)] - \mathbb{P}[\tilde{A}(D,m)] \leq (3+2\kappa)\exp(-\frac{\delta\varepsilon}{2(1+\varepsilon)}\min[\log(n), \rho_n\log(\rho_n)])$$

Proof: We claim that for any positive odd integer γ, the following estimate holds:

$$\begin{aligned}
&\sum_{\substack{D\subset\Gamma_n \\ |D|\leq\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D \\ |m|\leq\beta_n}} |\mathbb{P}[A(D,m)] - \mathbb{P}[\tilde{A}(D,m)]| \\
&\leq \frac{3}{\sqrt{\gamma}}\left(\frac{e\tau_n}{\gamma}\right)^{\gamma}\alpha_n(2e)^{2\tau_n} + \begin{cases} \frac{2(1+\kappa)|\Gamma_n|^{\gamma+1}}{n}\alpha_n e^{2\alpha_n} & \text{if } \sqrt{\beta_n|\Gamma_n|} \leq \alpha_n, \\ \frac{2(1+\kappa)|\Gamma_n|^{\gamma+1}}{n}\alpha_n\left(\frac{e^2\beta_n|\Gamma_n|}{\alpha_n^2}\right)^{\alpha_n} & \text{else.} \end{cases}
\end{aligned} \tag{2.11}$$

In order to show this, we apply Bonferonni-type estimates in Lemma A.1 for any positive odd integer $1 \leq \gamma \leq |\Gamma_n|$, as well as the condition $(\mathbf{H}_t)$, yielding

$$\begin{aligned}
|\mathbb{P}[A(D,m)] - \mathbb{P}[\tilde{A}(D,m)]| &\leq \sum_{\substack{I\subset\Gamma_n \\ |I|\leq\gamma}} \left|\mathbb{P}[p_D^m p_I | J_n] - \frac{1}{p_D^m p_I}\right| + 4\kappa \sum_{\substack{I\subset\Gamma_n \\ |I|=\gamma+1}} \frac{1}{p_D^m p_I} \\
&\leq \frac{2|\Gamma_n|^{\gamma+1}}{n^t} + \frac{4\kappa}{p_D^m}\frac{\tau_n^\gamma}{\gamma!} \\
&\leq \frac{2|\Gamma_n|^{\gamma+1}}{n^t} + \frac{6\kappa}{p_D^m}\frac{1}{\sqrt{2\pi\gamma}}\left(\frac{e\tau_n}{\gamma}\right)^{\gamma},
\end{aligned}$$

where we used Stirling's estimate $n! \geq \sqrt{2\pi n}(n/e)^n$ in the last inequality. We first show

$$\sum_{\substack{D\subset\Gamma_n \\ |D|\leq\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D \\ |m|\leq\beta_n}} \frac{2|\Gamma_n|^{\gamma+1}}{n^t} \leq \begin{cases} \frac{4|\Gamma_n|^{\gamma+1}}{n^t}\alpha_n e^{2\alpha_n} & \text{if } \sqrt{\beta_n|\Gamma_n|} \leq \alpha_n, \\ \frac{4|\Gamma_n|^{\gamma+1}}{n^t}\alpha_n\left(\frac{e^2\beta_n|\Gamma_n|}{\alpha_n^2}\right)^{\alpha_n} & \text{else.} \end{cases} \tag{2.12}$$

Notice that

$$|\{m \in \mathbb{N}^D : |m| \leq \beta_n\}| = \sum_{a=|D|}^{\beta_n} \binom{a-1}{|D|-1} \leq \frac{(\beta_n)^{|D|}}{|D|!}.$$

Thus, using Stirling's formula in the last inequality, we have

$$\sum_{\substack{D\subset\Gamma_n \\ |D|\leq\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D \\ |m|\leq\beta_n}} 1 \leq 1 + \sum_{j=1}^{\alpha_n} \binom{\Gamma_n}{j}\frac{\beta_n^j}{j!} \leq 1 + \sum_{j=1}^{\alpha_n} \frac{(\beta_n|\Gamma_n|)^j}{(j!)^2} \leq 1 + \sum_{j=1}^{\alpha_n} \left(\frac{e^2\beta_n|\Gamma_n|}{j^2}\right)^j.$$

Observe that the summand of the last sum as a function of j increases then decreases as j grows from 1 to ∞, and it attains the maximum at $j = \sqrt{\beta_n|\Gamma_n|}$. If $\sqrt{\beta_n|\Gamma_n|} \le \alpha_n$, we bound the summands from above by $e^{2\sqrt{\beta_n|\Gamma_n|}} \le e^{2\alpha_n}$, otherwise, the summands are bounded by $(\frac{e^2\beta_n|\Gamma_n|}{\alpha_n^2})^{\alpha_n}$, thus yielding (2.12).

Next we show

$$\sum_{\substack{D\subset\Gamma_n\\|D|\le\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D\\|m|\le\beta_n}} \frac{6\kappa}{p_D^m}\frac{1}{\sqrt{2\pi\gamma}}\Big(\frac{e\tau_n}{\gamma}\Big)^\gamma \le \frac{6\kappa}{\sqrt{\gamma}}\Big(\frac{e\tau_n}{\gamma}\Big)^\gamma \alpha_n(2e)^{2\tau_n}. \tag{2.13}$$

Manipulating the sum over m as in (2.9) gives

$$\sum_{\substack{m\in\mathbb{N}^D\\|m|\le\beta_n}} \frac{1}{p_D^m} = \prod_{p\in D}\frac{1}{1-p^{-1}}\mathbb{P}\Big[\sum_{p\in D} g_p \le \beta_n, g_p \ge 1, p\in D\Big] \le 2^{|D|}\mathbb{P}[g_p\ge 1, p\in D] = \frac{2^{|D|}}{p_D}.$$

Hence, handling the sum of $(p_D)^{-1}$ as in (2.4) yields

$$\sum_{\substack{D\subset\Gamma_n\\|D|\le\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D\\|m|\le\beta_n}} \frac{6\kappa}{p_D^m} \le 6\kappa\sum_{j=0}^{\alpha_n}2^j\sum_{\substack{D\subset\Gamma_n\\|D|=j}}\frac{1}{p_D} \le 6\kappa\sum_{j=0}^{\alpha_n}\frac{(2\tau_n)^j}{j!} \le 6\kappa\sqrt{2\pi}\sum_{j=0}^{\alpha_n}\Big(\frac{2e\tau_n}{j}\Big)^j,$$

where we used again Stirling's estimate the last inequality. Observe that the summands, regarded as a function of j, attains its maximum at $j = \lfloor 2\tau_n\rfloor$. Hence,

$$\sum_{\substack{D\subset\Gamma_n\\|D|\le\alpha_n}} \sum_{\substack{m\in\mathbb{N}^D\\|m|\le\beta_n}} \frac{6\kappa}{p_D^m} \le 6\kappa\sqrt{2\pi}\alpha_n(2e)^{2\tau_n},$$

leading to (2.13). Combining (2.12) and (2.13) gives (2.11).

It remains to prove (2.10). To this end, we apply (2.11) with $\gamma = \gamma_n$ given by

$$\gamma_n = \max\{k \le \alpha_n : k \text{ is an odd integer}\}.$$

We distinguish i) $\sqrt{\beta_n|\Gamma_n|} \le \alpha_n$ and ii) $\sqrt{\beta_n|\Gamma_n|} > \alpha_n$. In case i),

$$\begin{aligned}\log\Big(\frac{|\Gamma_n|^{\gamma_n+1}}{n^t}\alpha_n e^{2\alpha_n}\Big) &= (\delta\rho_n+1)\frac{\log(n)}{\rho_n} + \log(\delta\rho_n) + 2\delta\rho_n - t\log(n)\\ &\le -\big[t-\delta+(\rho_n)^{-1} - 2\delta[\log(|\Gamma_n|)]^{-1} - \frac{\log(\delta\log(n))}{\log(n)}\big]\log(n)\\ &\le -(t-3\delta)\log(n)\end{aligned}$$

for all n sufficiently large, where we used the condition $|\Gamma_n| \ge e^2$. On the other hand, by $\rho_n \ge \tau_n^{1+\varepsilon}$, we have

$$\begin{aligned}\log\Big(\big(\frac{e\tau_n}{\gamma_n}\big)^{\gamma_n}\alpha_n(2e)^{2\tau_n}\Big) &= \delta\rho_n\Big(1-\log(\delta) - \frac{\varepsilon}{1+\varepsilon}\log(\rho_n)\Big) + \log(\delta\rho_n) + 4\rho_n^{(1+\varepsilon)^{-1}}\\ &\le -\frac{\delta\varepsilon}{1+\varepsilon}\rho_n\log(\rho_n) + \Big(1-\log(\delta)+4\delta^{-1}\rho_n^{-\frac{\varepsilon}{1+\varepsilon}}\Big)\delta\rho_n + \log(\delta\rho_n)\\ &\le -\frac{\delta\varepsilon}{2(1+\varepsilon)}\rho_n\log(\rho_n)\end{aligned} \tag{2.14}$$

for n sufficiently large, ending the proof of (2.10) for case i). In case ii), we have

$$\begin{aligned}
&\log\Big(\frac{|\Gamma_n|^{\gamma_n+1}}{n^t}\alpha_n\Big(\frac{e^2\beta_n|\Gamma_n|}{\alpha_n^2}\Big)^{\alpha_n}\Big)\\
&= (2\delta\rho_n+1)\frac{\log(n)}{\rho_n}+\log(\delta\rho_n)\\
&\qquad+\delta\rho_n\Big(2+\log\Big[\frac{2\delta}{(1+\varepsilon)\log(1.5)}\Big]+\log(\rho_n\log(\rho_n))-2\log(\delta\rho_n)\Big)-t\log(n)\\
&\le -[t-2\delta+(\rho_n)^{-1}]\log(n)-\delta\rho_n\log(\rho_n)+\delta\rho_n\log\log(\rho_n)\\
&\qquad+\Big(2+\log\Big[\frac{2\delta}{(1+\varepsilon)\log(1.5)}\Big]-2\log(\delta)\Big)\delta\rho_n+\log(\delta\rho_n)\\
&\le -(t-3\delta)\log(n)
\end{aligned}$$

for n sufficiently large. This, together with (2.14), ends the proof of (2.10) for case ii), thereby ending the proof of the lemma. □

Proof of Theorem 1.5: Let $\delta\in(0,0.3t)$ with t given at (1.7). Recall the definitions of α_n,β_n at (2.1) and (2.5), which depend on δ and ε in the statement of Theorem 1.5. Using the expression (1.5) for the total variance distance, there exists a constant $N\in\mathbb{N}$ depending on ε,δ, such that

$$\begin{aligned}
d_{\mathrm{TV}}(\boldsymbol{v}^n,\boldsymbol{g}^n) &= \frac12\sum_{\substack{D\subset\Gamma_n\\ |D|>\alpha_n}}\sum_{m\in\mathbb{N}^D}|\mathbb{P}[A(D,m)]-\mathbb{P}[\tilde A(D,m)]|\\
&+\frac12\sum_{\substack{D\subset\Gamma_n\\ |D|\le\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\ |m|\ge\beta_n}}|\mathbb{P}[A(D,m)]-\mathbb{P}[\tilde A(D,m)]|\\
&+\frac12\sum_{\substack{D\subset\Gamma_n\\ |D|\le\alpha_n}}\sum_{\substack{m\in\mathbb{N}^D\\ |m|\le\beta_n}}|\mathbb{P}[A(D,m)]-\mathbb{P}[\tilde A(D,m)]|.
\end{aligned}$$

Since $\rho_n\to\infty$, by choosing $\delta=(0.3t)\wedge\varepsilon$, we derive by using Lemmas 2.1-2.3 that

$$d_{\mathrm{TV}}(\boldsymbol{v}^n,\boldsymbol{g}^n)\le(4+2\kappa)\exp\Big(-\frac{\varepsilon\wedge(t/4)}{2(1+\varepsilon)}\min(\log n,\rho_n\log\rho_n)\Big).$$

This completes the proof. □

3. Applications

The total variation bound obtained in the paper allows to transfer distributional approximation results valid in the realm of independent random variables to the dependent prime multiplicities of a random sample J_n, at the price of an extra error term that appears in Theorem 1.5.

In this section, we mention two such possibilities: a generalized Erdös-Kac theorem, and a Poisson process approximation result. The first result (Theorem 3.1) generalise considerably the classical Erdös-Kac theorem for additive functions of a uniform random variable in $[n]$. The second result (Theorem 3.2) seems to be new even when J_n is uniform. We stress that these results are manifestations of the non-trivial fact that a large subset of prime multiplicities of a rather general random sample in $[n]$ are very close to be independent random variables. This non-trivial fact is the main message of this paper, proved in Theorem 1.5. Once the fact is established, either normal approximation (in Theorem 3.1) or Poisson approximation (in Theorem 3.2) for sum of weakly dependent random variables is not hard to prove.

3.1. *Generalized Erdös-Kac theorem.* Consider the prime factor counting function

$$\omega(k) = \sum_{p \in \mathcal{P}} \mathbb{1}_{\{p|k\}},$$

and the law of $w(J_n)$ with J_n satisfies $(\mathbf{H}_t)$ with $t > 0$. In the special case where J_n is uniformly distributed in $[n]$ (which satisfies $(\mathbf{H}_t)$ with $t = 1$ by Example 1.1), the celebrated Erdös-Kac theorem (see Erdös and Kac (1940)) states that $(\omega(J_n) - \log\log(n))/\sqrt{\log\log(n)}$ converges in distribution to a standard Gaussian random variable. To assess the rate of convergence of the Erdös-Kac theorem, one can use the Wasserstein distance given by

$$d_W(X, Y) = \sup_{g \in Lip_1} |\mathbb{E}[g(X)] - \mathbb{E}[g(Y)]|,$$

where X, Y are real-valued random variables and Lip_1 denotes the set of 1-Lipchitz functions.

There are several probabilistic proofs for obtaining assessments of the distance between $\omega(J_n)$ subject to normalization and a standard Gaussian random variable. We refer to Harper (2009) for two proofs relying on Stein's method in the Kolmogorov metric. We would like to highlight the intention of Harper (2009): it aims to present arguments which are highly accessible, in the sense that they do not rely on sophisticated machinery from number theory. Theorem 3.1 below has a similar feature, as it only makes use of Theorem 1.5, which is based simply on some Bonferroni-type estimates and elementary large deviation bounds. The inequality (1.11) provides a rather simple scheme for achieving the rate (3.1) which is the same as Harper (2009), not only for the case in which J_m is uniform, but for a wide range of variables satisfying $(\mathbf{H_t})$. Obtaining an optimal rate of convergence in the Erdos Kac theorem is a highly non-trivial task that escapes the reach of our techniques. However, the Charlier expansions for the law of $\omega(J_n)$ presented in Barbour et al. (2014), imply that the optimal rate is of the order $\log\log(n)^{-1/2}$, which differs from our bound just by a term of the order $\log\log\log(n)$.

Theorem 3.1. *Suppose that the law of J_n is such that there is a constant $\kappa > 0$ independent of n, such that $\{J_n\}_{n\geq 1}$ satisfies the condition $(\mathbf{H_t})$. Then there exists a constant $C > 0$, such that*

$$d_W\left(\frac{\omega(J_n) - \log\log(n)}{\sqrt{\log\log(n)}}, N\right) \leq C\frac{\log\log\log(n)}{\sqrt{\log\log(n)}}, \tag{3.1}$$

where N is a standard Gaussian random variable denotes the Wasserstein distance.

Proof: We observe that

$$\omega(J_n) = \sum_{p \in \mathcal{P} \cap [1,n]} \mathbb{1}_{\{v_p \geq 1\}}.$$

Define the arithmetic function

$$\tilde{\omega}_n(k) := \sum_{p \in \mathcal{P} \cap [1, n^{\frac{1}{3\log\log(n)^2}}]} \mathbb{1}_{\{p \text{ divides } k\}}.$$

By Merten's formula (see Tenenbaum (1999, Theorem 9)), we have that

$$|\sum_{p \in \mathcal{P} \cap [1,n]} \frac{1}{p} - \log\log(n)| \leq 1,$$

for n large. Combining this inequality with $(\mathbf{H_t})$, we deduce the existence of a constant $C > 0$ such that

$$\mathbb{E}[\omega(J_n) - \omega_n(J_n)] + \mathbb{V}\mathrm{ar}[\omega(J_n) - \tilde{\omega}_n(J_n)] \leq C\log\log\log(n). \tag{3.2}$$

and

$$|\mathbb{E}[\omega(J_n)] - \log\log(n)| + |\mathbb{V}\mathrm{ar}[\omega(J_n)] - \log\log(n)| \le C \log\log\log(n). \tag{3.3}$$

Observe that

$$\begin{aligned}
&d_W(\omega(J_n) - \mathbb{E}[\omega(J_n)], M_n) \\
&\qquad \le d_W(\omega(J_n) - \mathbb{E}[\omega(J_n)], \tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)]) \\
&\qquad + d_W(\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)], \tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)]\mathbb{1}_{\{|\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)]| \le \log\log(n)^2\}}) \\
&\qquad + d_W((\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)])\mathbb{1}_{|\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)]| \le \log\log(n)^2}, \sum_{p \in \mathcal{P}} \mathbb{1}_{\{|p| \le n^{\frac{1}{3\log\log(n)^2}}\}} g_p) \\
&\qquad + d_W(\sum_{p \in \mathcal{P}} \mathbb{1}_{\{|p| \le n^{\frac{1}{3\log\log(n)^2}}\}} g_p, M_n),
\end{aligned}$$

where the g_p are independent random variables with geometric law of parameter $1/p$ and M_n denotes a centered Gaussian random variable with variance $\log\log(n)$. Using the Berry Esseen estimations, as well as (3.2) and (3.3), we thus get

$$\begin{aligned}
&d_W(\omega(J_n) - \mathbb{E}[\omega(J_n)], M_n) \\
&\le C\log\log\log(n) + d_W((\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)])\mathbb{1}_{|\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)]| \le \log\log(n)^2}, \sum_{p \in \mathcal{P}} \mathbb{1}_{\{|p| \le n^{\frac{1}{3\log\log(n)^2}}\}} g_p).
\end{aligned} \tag{3.4}$$

Define

$$Y_n := (\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)])\mathbb{1}_{|\tilde{\omega}_n(J_n) - \mathbb{E}[\tilde{\omega}_n(J_n)]| \le \log\log(n)^2}.$$

and

$$\tilde{Y}_n := \sum_{p \in \mathcal{P}} \mathbb{1}_{\{|p| \le n^{\frac{1}{3\log\log(n)^2}}\}} g_p.$$

Observe that Y_n takes values in the set $\{-\log\log(n)^2, \dots, \log\log(n)^2\}$ and consequently, from a straighforward examination of the Wasserstein distance,

$$\begin{aligned}
d_W(Y_n, \tilde{Y}_n) &\le \mathbb{E}[|\tilde{Y}_n| \mathbb{1}_{\{\log\log(n)^2 \le |\tilde{Y}_n|\}}] + \sum_{k=-\log\log(n)^2}^{\log\log(n)^2} |k| |\mathbb{P}[\tilde{Y}_n = k] - \mathbb{P}[Y_n = k]| \\
&\le C + C\log\log(n)^4 d_{\mathrm{TV}}(Y_n, \tilde{Y}_n).
\end{aligned}$$

Observe that by the prime number theorem, $\log(|\mathcal{P} \cap [1, n^{\frac{1}{3\log\log(n)^2}}]|) = O(\frac{\log(n)}{3\log\log(n)^2})$. Therefore, by Theorem 1.5

$$\begin{aligned}
d_W(Y_n, \tilde{Y}_n) &\le \mathbb{P}[\tilde{Y}_n \ge \log\log(n)^2] + \sum_{k=-\log\log(n)^2}^{\log\log(n)^2} |k| |\mathbb{P}[\tilde{Y}_n = k] - \mathbb{P}[Y_n = k]| \\
&\le C + C\log\log(n)^4 e^{-\log\log(n)^4} \le C'.
\end{aligned}$$

Combining the above inequality with (3.4), we get

$$d_W(\omega(J_n) - \mathbb{E}[\omega(J_n)], M_n) \le C\log\log\log(n).$$

The above inequality leads to the desired conclusion. □

3.2. *Poisson approximation.* Now we show how our general bound is relevant in the case of Poisson approximation. Consider the counting process

$$X_n(t) = \sum_{p \in \mathcal{P} \cap [a_n, a_n^{e^t}]} \mathbb{1}_{\{v_p \geq 1\}}, \quad t \in [0,1]$$

where a_n diverges as $n \to \infty$ and there exists $\varepsilon > 0$ such that

$$a_n \leq e^{\frac{\log(n)/e}{[\log\log(n)]^{1+\varepsilon}}}.$$

Hence the cardinality of $\Gamma_n := \mathcal{P} \cap [a_n, a_n^{e^t}]$ satisfies the condition of Theorem 1.5. Denote by $\{\widetilde{X}_n(t), t \in [0,1]\}$ the process obtained by replacing all the v_p's by g_p's. By abuse of language, we continue to use X_n to denote the counting measure induced by X_n and make the same convention for $\tilde{X}_n$.

Recall that the total variation distance between random measures η and η' is defined as $d_{\mathrm{TV}}(\eta, \eta') = \inf \mathbb{P}[\eta_1 \neq \eta_2]$ where the infimum is taken over all pair (η_1, η_2) of random measures such that its first marginal is equal in law to η and its second marginal equal to η'.

Theorem 3.2. *Let η_n be a Poisson point process with the same intensity as that of $\tilde{X}_n$. Then*

$$d_{\mathrm{TV}}(X_n, \eta_n) \leq d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n) + \frac{2}{a_n}.$$

Proof: Let $(\boldsymbol{v}^n, \boldsymbol{g}^n)$ be the optimal coupling such that $\mathbb{P}[\boldsymbol{v}^n \neq \boldsymbol{g}^n] = d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n)$. Then

$$d_{\mathrm{TV}}(X_n, \tilde{X}_n) \leq \mathbb{P}[X_n \neq \tilde{X}_n] = d_{\mathrm{TV}}(\boldsymbol{v}^n, \boldsymbol{g}^n).$$

On the other hand, we can obtain an error estimate for marginal distributions of $\tilde{X}_n$ and that of η_n. For any fixed $t \in [0,1]$, it follows from classical Poisson limit theorem Durrett (2019, Section 3.6) that

$$\begin{aligned} d_{\mathrm{TV}}(\widetilde{X}_n(t), \eta_n([0,t])) &\leq \sum_{p \in \mathcal{P} \cap [a_n, a_n^{e^t}]} \mathbb{P}[g_p \geq 1]^2 \\ &\leq \max\left\{p^{-1} : p \in \mathcal{P} \cap [a_n, a_n^{e^t}]\right\} \sum_{p \in \mathcal{P} \cap [a_n, a_n^{e^t}]} \frac{1}{p} \leq \frac{2}{a_n}, \end{aligned}$$

where we used Mertens' formula in the last step. □

Appendix A. Some elementary probabilistic estimates

We prove a Bonferonni-type estimates.

Lemma A.1. *For any positive odd integer $\gamma \leq |\Gamma_n|$,*

$$\sum_{\substack{I \subset \Gamma_n \\ |I| \leq \gamma}} (-1)^{|I|} \mathbb{P}[p_D^m p_I | J_n] \leq \mathbb{P}[A(D,m)] \leq \sum_{\substack{I \subset \Gamma_n \\ |I| \leq \gamma+1}} (-1)^{|I|} \mathbb{P}[p_D^m p_I | J_n], \tag{A.1}$$

$$\sum_{\substack{I \subset \Gamma_n \\ |I| \leq \gamma}} (-1)^{|I|} \frac{1}{p_D^m p_I} \leq \mathbb{P}[\tilde{A}(D,m)] \leq \sum_{\substack{I \subset \Gamma_n \\ |I| \leq \gamma+1}} (-1)^{|I|} \frac{1}{p_D^m p_I}. \tag{A.2}$$

Proof: Notice that

$$A(D,m) = \{p_D^m | J_n\} \cap \Big(\big(\bigcup_{q \in \Gamma_n \setminus D} \{q | J_n\}\big) \cup \big(\bigcup_{q \in D} \{q^{m_q+1} | J_n\}\big)\Big)^c.$$

Hence,

$$\begin{aligned}\mathbb{P}[A(D,m)] &= \mathbb{P}[p_D^m|J_n] - \mathbb{P}\big[\big(\bigcup_{q\in\Gamma_n\setminus D}\{qp_D^m|J_n\}\big)\cup\big(\bigcup_{q\in D}\{qp_D^m|J_n\}\big)\big]\\ &= \mathbb{P}[p_D^m|J_n] - \mathbb{P}\big[\bigcup_{q\in\Gamma_n}\{qp_D^m|J_n\}\big].\end{aligned}$$

By Bonferonni inequalities Durrett (2019, Exercise 1.6.10), we have for any positive odd integer $\gamma < |\Gamma_n|$ that

$$\sum_{\substack{\emptyset\neq I\subset\Gamma_n\\|I|\le\gamma+1}}(-1)^{|I|+1}\mathbb{P}\big[\bigcap_{q\in I}\{qp_D^m|J_n\}\big] \le \mathbb{P}\big[\bigcup_{q\in\Gamma_n}\{qp_D^m|J_n\}\big] \le \sum_{\substack{\emptyset\neq I\subset\Gamma_n\\|I|\le\gamma}}(-1)^{|I|+1}\mathbb{P}\big[\bigcap_{q\in I}\{qp_D^m|J_n\}\big].$$

Observe that $\cap_{q\in I}\{qp_D^m|J_n\} = \{p_D^m p_I|J_n\}$. Thus,

$$\begin{aligned}\mathbb{P}[A(D,m)] &\le \mathbb{P}[p_D^m|J_n] - \sum_{\substack{\emptyset\neq I\subset\Gamma_n\\|I|\le\gamma+1}}(-1)^{|I|+1}\mathbb{P}\big[\bigcap_{q\in I}\{qp_D^m|J_n\}\big] = \sum_{\substack{I\subset\Gamma_n\\|I|\le\gamma+1}}(-1)^{|I|}\mathbb{P}[p_D^m p_I|J_n],\\ \mathbb{P}[A(D,m)] &\ge \mathbb{P}[p_D^m|J_n] - \sum_{\substack{\emptyset\neq I\subset\Gamma_n\\|I|\le\gamma}}(-1)^{|I|+1}\mathbb{P}\big[\bigcap_{q\in I}\{qp_D^m|J_n\}\big] = \sum_{\substack{I\subset\Gamma_n\\|I|\le\gamma}}(-1)^{|I|}\mathbb{P}[p_D^m p_I|J_n],\end{aligned}$$

ending the proof of (A.1). The same argument, with independence and the exact law of g_p, yields (A.2). We leave the details to the interested reader. □

We record here the Chernoff bound for Poisson random variables.

Lemma A.2. *If $0 < \lambda < x$ and M is a Poisson random variable with parameter $\lambda > 0$, then*

$$\mathbb{P}[M \ge x] \le e^{-\lambda}(e\lambda)^x x^{-x}.$$

Equivalently,

$$\sum_{k\ge x}\frac{\lambda^k}{k!} \le \Big(\frac{e\lambda}{x}\Big)^x.$$

Acknowledgements

We would like to thank an anonymous referee for carefully reading the first version of this manuscript and making very valuable comments.

References

Arratia, R. On the amount of dependence in the prime factorization of a uniform random integer. In *Contemporary combinatorics*, volume 10 of *Bolyai Soc. Math. Stud.*, pp. 29–91. János Bolyai Math. Soc., Budapest (2002). MR1919568.

Arratia, R., Barbour, A. D., and Tavaré, S. The Poisson-Dirichlet distribution and the scale-invariant Poisson process. *Combin. Probab. Comput.*, **8** (5), 407–416 (1999). MR1731976.

Arratia, R. and Stark, D. A total variation distance invariance principle for primes, permutations and Poisson-Dirichlet (1999). Preprint.

Arratia, R. and Tavaré, S. The cycle structure of random permutations. *Ann. Probab.*, **20** (3), 1567–1591 (1992). MR1175278.

Barban, M. B. and Vinogradov, A. I. On the number-theoretic basis of probabilistic number theory. *Dokl. Akad. Nauk SSSR*, **154**, 495–496 (1964). MR0163895.

Barbour, A. D., Kowalski, E., and Nikeghbali, A. Mod-discrete expansions. *Probab. Theory Related Fields*, **158** (3-4), 859–893 (2014). MR3176367.

Chen, L. H. Y., Jaramillo, A., and Yang, X. A probabilistic approach to the Erdös-Kac theorem for additive functions. *ArXiv Mathematics e-prints* (2022). arXiv: 2102.05094.

Durrett, R. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge (2019). ISBN 978-1-108-47368-2. MR3930614.

Elliott, P. D. T. A. *Probabilistic number theory II: Central limit theorems*, volume 240 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin-New York (1980). ISBN 0-387-90438-7. MR560507.

Erdös, P. and Kac, M. The Gaussian law of errors in the theory of additive number theoretic functions. *Amer. J. Math.*, **62**, 738–742 (1940). MR2374.

Harper, A. J. Two new proofs of the Erdős-Kac theorem, with bound on the rate of convergence, by Stein's method for distributional approximations. *Math. Proc. Cambridge Philos. Soc.*, **147** (1), 95–114 (2009). MR2507311.

Kubilius, J. *Probabilistic methods in the theory of numbers*. Translations of Mathematical Monographs, Vol. 11. American Mathematical Society, Providence, R.I. (1964). MR0160745.

Tenenbaum, G. Crible d'Ératosthène et modèle de Kubilius. In *Number theory in progress, Vol. 2 (Zakopane-Kościelisko, 1997)*, pp. 1099–1129. de Gruyter, Berlin (1999). MR1689563.