Tony Haddad & Dimitris Koukoulopoulos

On Arratia's coupling and the Dirichlet law for the factors of a random integer

ON ARRATIA'S COUPLING AND THE DIRICHLET LAW FOR THE FACTORS OF A RANDOM INTEGER

by Tony Haddad & Dimitris Koukoulopoulos

Abstract. — Let $x \geqslant 2$, let N_x be an integer chosen uniformly at random from the set $\mathbb{Z} \cap [1, x]$, and let $(V_1, V_2, \dots)$ be a Poisson–Dirichlet process of parameter 1. We prove that there exists a coupling of these two random objects such that

$$\mathbb{E} \sum_{i \geqslant 1} \left| \log P_i - V_i \log x \right| \asymp 1,$$

where the implied constants are absolute and $N_x = P_1 P_2 \cdots$ is the unique factorization of N_x into primes or ones with the P_i's being non-increasing. This establishes a 2002 conjecture of Arratia, who constructed a coupling for which the left-hand side in the above estimate is $\ll \log\log x$, and who also proved that the left-hand side is $\geqslant 1 - o(1)$ for all couplings. In addition, we use our refined coupling to give a probabilistic proof of the Dirichlet law for the average distribution of the integer factorization into k parts proved in 2023 by Leung and we improve on its error term.

Résumé (Sur le couplage d'Arratia et la loi de Dirichlet pour les facteurs d'un entier aléatoire)
Soit $x \geqslant 2$, soit N_x un entier choisi uniformément au hasard dans l'ensemble $\mathbb{Z} \cap [1, x]$, et soit $(V_1, V_2, \dots)$ un processus de Poisson-Dirichlet de paramètre 1. Nous montrons l'existence d'un couplage de ces deux objets aléatoires satisfaisant à

$$\mathbb{E} \sum_{i \geqslant 1} \left| \log P_i - V_i \log x \right| \asymp 1,$$

où les constantes implicites sont absolues et où $(P_i)_{i \geqslant 1}$ est l'unique suite décroissante de nombres premiers ou de uns qui satisfait à $N_x = P_1 P_2 \cdots$. Ce résultat établit une conjecture d'Arratia (2002), qui avait auparavant construit un couplage dont l'espérance ci-dessus satisfaisait à $\ll \log\log x$, et montré que cette espérance est toujours $\geqslant 1 - o(1)$ pour tout couplage. De plus, nous utilisons le couplage pour fournir une preuve probabiliste de la loi de Dirichlet pour la distribution moyenne des factorisations en k parties d'un entier, résultat initialement établi en 2023 par Leung, et nous en améliorons le terme d'erreur.

Contents

1. Introduction

Let N_x be an integer chosen uniformly at random from the set $\mathbb{Z} \cap [1, x]$. We may then factor it uniquely as $N_x = P_1 P_2 \cdots$ with the P_i's forming a non-increasing sequence of primes or ones. In 1972, Billingsley [5] showed that, for any fixed positive integer r, the joint distribution of the random vector

$$\left(\frac{\log P_1}{\log x}, \ldots, \frac{\log P_r}{\log x} \right)$$

converges in distribution as $x \to \infty$ to the first r components of the Poisson–Dirichlet distribution (of parameter 1).

There are many ways to define the Poisson–Dirichlet distribution. One of the most intuitive ones involves a "stick-breaking" process that we will use throughout the paper. We start by sampling a sequence of i.i.d. random variables $(U_i)_{i \geqslant 1}$ that are all uniformly distributed in $[0, 1]$. We then define the sequence $(L_i)_{i \geqslant 1}$ in the following way:

$$L_1 := U_1 \quad \text{and} \quad L_j := U_j \prod_{i=1}^{j-1} (1 - U_i) \quad \text{for } j \geqslant 2.$$

The distribution of the process $\boldsymbol{L} = (L_1, L_2, \ldots)$ is called the *GEM distribution* (of parameter 1), named after Griffiths, Engen and McCloskey (see [10, Ch. 41] for a discussion on the history of this distribution). Lastly, we sort the components of $\boldsymbol{L}$ in non-increasing order to create $\boldsymbol{V} = (V_1, V_2, \ldots)$. The distribution of this process

is the *Poisson–Dirichlet distribution* (of parameter 1).[1] We note that both $\sum_{i \geqslant 1} L_i$ and $\sum_{i \geqslant 1} V_i$ are equal to 1 almost surely.

In 2000, Tenenbaum [19] studied the rate of convergence in Billingsley's theorem by providing an asymptotic series for the difference between the cumulative distribution functions of $\left(\frac{\log P_1}{\log x}, \ldots, \frac{\log P_r}{\log x} \right)$ and of $(V_1, \ldots, V_r)$.

Another way to give a quantitative version of Billingsley's result is by constructing a *coupling* of N_x and $\boldsymbol{V}$, i.e., a single probability space over which lives copies of N_x and $\boldsymbol{V}$, such that the expectation

$$\mathbb{E} \sum_{i \geqslant 1} \left| \log P_i - V_i \log x \right| \tag{1.1}$$

is bounded by a positive monotone function that is $o(\log x)$ as $x \to \infty$. The random variables N_x and $\boldsymbol{V}$ must be strongly correlated in this new probability space to achieve this. Indeed, if, for instance, P_1 and V_1 were independent, then we would have that $P_1 \leqslant x^{1/3}$ and $V_1 > 2/3$ can happen at the same time with positive probability, which implies $\left| \log P_1 - V_1 \log x \right| \geqslant (\log x)/3$ with positive probability. Hence, a coupling with N_x and $\boldsymbol{V}$ being independent (also called a *trivial coupling*) makes (1.1) $\asymp \log x$.

In 2002, Richard Arratia [2] constructed a coupling satisfying

$$\mathbb{E} \sum_{i \geqslant 1} \left| \log P_i - V_i \log x \right| \ll \log\log x \tag{1.2}$$

for all $x \geqslant 3$. Moreover, he conjectured that there is a coupling for N_x and $\boldsymbol{V}$ with the expectation above being $O(1)$. The main goal of this paper is to prove this conjecture:

Theorem 1. — *There is a coupling of N_x and $\boldsymbol{V}$ satisfying*

$$\mathbb{E} \sum_{i \geqslant 1} \left| \log P_i - V_i \log x \right| \asymp 1$$

for all $x \geqslant 1$.

Remarks

(a) Theorem 1 is optimal in the sense that no coupling of N_x and $\boldsymbol{V}$ can make this expectation tend to 0 as $x \to \infty$. This follows directly from the triangle inequality:

$$\mathbb{E} \sum_{i \geqslant 1} \left| \log P_i - V_i \log x \right| \geqslant \mathbb{E} \left| \sum_{i \geqslant 1} \left(\log P_i - V_i \log x \right) \right| = \mathbb{E} \left[\log(x/N_x) \right] = 1 - o(1).$$

(b) If X and Y are two random variables taking values in a separable metric space (S, d), their *Wasserstein distance* equals

$$\mathrm{d}_W(X, Y) := \inf \mathbb{E} \left[\mathrm{d}(X, Y) \right], \tag{1.3}$$

where the infimum is taken over all possible couplings of X and Y. Taking S to be the ℓ^1 space of real valued sequences with its associated metric, and using Remark (a)

---

[1] The GEM and Poisson–Dirichlet distributions have more general definitions involving typically a parameter θ. In the rest of the paper, we will not be mentioning the parameter since we will always work with $\theta = 1$.

above, we find that Theorem 1 is equivalent to the statement that $\mathrm{d_W}(\boldsymbol{P}_x, \boldsymbol{V}) \asymp 1/\log x$, where $\boldsymbol{P}_x := \left(\frac{\log P_1}{\log x}, \frac{\log P_2}{\log x}, \dots\right)$.

(c) The coupling we construct contains a copy of the random process $(N_x)_{x \geqslant 1}$. This fact might be useful in applications of Theorem 1.

(d) Let σ be a random permutation uniformly distributed in the permutation group S_n. It is well known that the factorization into primes of N_x and the decomposition into disjoint cycles of σ share similar statistics when $n \approx \log x$. In 2006, Arratia, Barbour and Tavaré [3] have proved that there exists a coupling between σ and $\boldsymbol{V}$ such that

$$(1.4) \qquad \mathbb{E} \sum_{i \geqslant 1} |C_i - nV_i| \sim \frac{\log n}{4},$$

with C_i being the number of cycles of length i in σ. They showed that (1.4) was optimal by using the inequality $|C_i - nV_i| \geqslant \|nV_i\|$ where $\|\cdot\|$ is the distance to the closest integer, and computing $\mathbb{E}\sum_{i \geqslant 1} \|nV_i\|$. This breaks the analogy between primes and permutations since Theorem 1 and (1.4) are not of the same order of magnitude when n is replaced by $\log x$. The main reason why it is possible to get a better result in Theorem 1 is because the set $\{\log p : p \text{ primes}\}$ have much shorter gaps around $\log x$ than the gaps of $\mathbb{Z}$ around n.

1.1. Application to the distribution of factorizations of random integers

We now give some applications of Theorem 1 to the theory of divisors of integers. Let Δ^∞ be the set of sequences $(v_i)_{i \geqslant 1}$ of non-negative reals satisfying $\sum_{i \geqslant 1} v_i = 1$, and let $\psi \colon \Delta^\infty \to \mathbb{R}$ be a K-Lipschitz function with respect to the ℓ^1-metric on Δ^∞. Since the diameter of Δ^∞ is 2, the image of ψ is contained in an interval of length $2K$. Consequently, ψ is bounded, and we have $\mathbb{E}[\mathbb{1}_{N_x=1} \cdot \psi(\boldsymbol{V})] \ll_\psi 1/x$. By Theorem 1, it follows that

$$(1.5) \qquad \mathbb{E}\left[\mathbb{1}_{N_x \geqslant 2} \cdot \psi\left(\tfrac{\log P_1}{\log N_x}, \tfrac{\log P_2}{\log N_x}, \dots\right)\right] = \mathbb{E}\left[\psi(V_1, V_2, \dots)\right] + O_\psi(1/\log x).$$

Using this simple fact, we obtain:

Corollary 1.1. — *Let $\rho(n) := \min\{d | n : d \geqslant \sqrt{n}\}$. Then there is a constant $c \in (1/2, 1)$ such that*

$$\sum_{n \leqslant x} \log \rho(n) = cx \log x + O(x) \qquad (x \geqslant 1).$$

Proof. — Let

$$\psi(v_1, v_2, \dots) := \inf\left(\mathcal{D}(v_1, v_2, \dots) \cap [\tfrac{1}{2}, 1]\right)$$

with $\mathcal{D}(v_1, v_2, \dots) := \{\sum_{i \in B} v_i : B \subseteq \mathbb{N}\}$. It is easy to check that this is a 1-Lipschitz function. Note that the left-hand side of (1.5) with this definition for ψ is exactly equal to

$$\frac{1}{\lfloor x \rfloor} \sum_{2 \leqslant n \leqslant x} \frac{\log \rho(n)}{\log n}.$$

With (1.5) and partial summation, we prove the corollary with $c = \mathbb{E}[\psi(\boldsymbol{V})]$. $\qquad \square$

For the next application, we use the coupling again to extract information about integer factorizations. However, to obtain the desired result, we require much more than just the expectation in Theorem 1.

Let Δ^{k-1} be the set of k-tuples $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}_{\geqslant 0}^k$ satisfying $\alpha_1 + \cdots + \alpha_k = 1$. We also need to define a special class of functions:

DEFINITION 1.2 (The class of functions $\mathcal{F}_k(\boldsymbol{\alpha})$). — Given $k \in \mathbb{Z}_{\geqslant 2}$ and $\boldsymbol{\alpha} \in \Delta^{k-1}$, let $\mathcal{F}_k(\boldsymbol{\alpha})$ be the set of functions $f \colon \mathbb{N}^k \to \mathbb{R}_{\geqslant 0}$ satisfying the following three properties:

(a) We have $\sum_{d_1 \cdots d_k = n} f(d_1, \ldots, d_k) = 1$ for all $n \in \mathbb{N}$. Thus, for each fixed $n \in \mathbb{N}$, the function $f(\boldsymbol{d})$ defines a probability mass function on the set of vectors $\boldsymbol{d} \in \mathbb{N}^k$ satisfying $d_1 \cdots d_k = n$.

(b) Whenever $\boldsymbol{d}$ satisfies

$$d_i := \begin{cases} p & \text{if } i = j, \\ 1 & \text{if } i \neq j. \end{cases}$$

for some $1 \leqslant j \leqslant k$ and prime p, then $f(\boldsymbol{d}) = \alpha_j$.

(c) The function f is *multiplicative*, i.e., for any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{N}^k$ with $a_1 \cdots a_k$ and $b_1 \cdots b_k$ being co-prime, we have the property

$$f(a_1 b_1, \ldots, a_k b_k) = f(a_1, \ldots, a_k) \cdot f(b_1, \ldots, b_k).$$

REMARK. — Let $\omega(n)$ denote the number of distinct prime factors of n. If $d_1 \cdots d_k$ is a square-free number and $f \in \mathcal{F}_k(\boldsymbol{\alpha})$, then properties (b) and (c) of the definition above imply that

$$(1.6) \qquad\qquad f(d_1, \ldots, d_k) = \prod_{j=1}^k \alpha_j^{\omega(d_j)}.$$

We will use the class of functions $\mathcal{F}_k(\boldsymbol{\alpha})$ to define certain "random factorizations" into k parts of a random integer. Specifically, let us fix $f \in \mathcal{F}_k(\boldsymbol{\alpha})$ and $x \geqslant 1$. We then define a *random k-factorization corresponding to f* to be a random vector $\boldsymbol{D}_{f,x} = (D_{f,x,1}, \ldots, D_{f,x,k})$ taking values on $\mathbb{N}^k$ and satisfying the formula[2]

$$(1.7) \qquad\qquad \mathbb{P}\Big[D_{f,x,i} = d_i \ \forall i \leqslant k \,\Big|\, N_x = n\Big] = f(d_1, \ldots, d_k)$$

for all $n \in \mathbb{N}$ and all k-tuples $(d_1, \ldots, d_k) \in \mathbb{N}^k$ with $d_1 \cdots d_k = n$.

Here are three examples of such random factorizations.

EXAMPLE 1 (Uniform sampling). — Let $f_{\mathrm{U}(k)}(d_1, \ldots, d_k) := \tau_k(d_1 \cdots d_k)^{-1}$ with $\tau_k(n)$ being the number of k-factorizations of n. Then $f_{\mathrm{U}(k)} \in \mathcal{F}_k(\frac{1}{k}, \ldots, \frac{1}{k})$. If $f_{\mathrm{U}(k)}$ is seen as a probability mass function as in (1.7), then we are sampling $D_{f_{\mathrm{U}(k)},x,1} \cdots D_{f_{\mathrm{U}(k)},x,k}$ uniformly among all k-factorizations of N_x.

---

[2]Strictly speaking, we only need property (a) of Definition 1.2 to define $\boldsymbol{D}_{f,x}$. But we will also need the other two properties when proving Theorem 2 below.

Example 2 (Recursive sampling). — Let $f_{\mathrm{R}(k)}(d_1,\ldots,d_k) \coloneqq \prod_{j=1}^{k-1} \tau(d_j \cdots d_k)^{-1}$ with $\tau(n)$ being the number of divisors of n. Then $f_{\mathrm{R}(k)} \in \mathcal{F}_k(\frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^{k-1}}, \frac{1}{2^{k-1}})$. One way to realize this random k-factorization is by first sampling uniformly a divisor $D_{f_{\mathrm{R}(k)},x,1}$ of N_x. Then, for all $j < k$, we recursively sample $D_{f_{\mathrm{R}(k)},x,j}$ uniformly among the divisors of $N_x/D_{f_{\mathrm{R}(k)},x,1} \cdots D_{f_{\mathrm{R}(k)},x,j-1}$.

Example 3 (Multinomial sampling). — For any fixed $\boldsymbol{\alpha} \in \Delta^{k-1}$, we define the function $f_{\mathrm{M}(\boldsymbol{\alpha})}(d_1,\ldots,d_k) \coloneqq \prod_{i=1}^{k} \alpha_i^{\Omega(d_i)} \cdot \prod_{p|n} \binom{\nu_p(d_1 \cdots d_k)}{\nu_p(d_1),\ldots,\nu_p(d_k)}$ with $\nu_p(d)$ being the p-valuation of d and $\Omega(d)$ being the number of prime factors of d counted with multiplicity. The function $f_{\mathrm{M}(\boldsymbol{\alpha})}$ is in $\mathcal{F}_k(\boldsymbol{\alpha})$. This sampling can be understood as considering a sequence of i.i.d. random variables $(B_i)_{i\geqslant 1}$ satisfying $\mathbb{P}[B_i = j] = \alpha_j$ and constructing the k-factorization $D_{f_{\mathrm{M}(\boldsymbol{\alpha})},x,1} \cdots D_{f_{\mathrm{M}(\boldsymbol{\alpha})},x,k} = N_x$ as

$$D_{f_{\mathrm{M}(\boldsymbol{\alpha})},x,j} \coloneqq \prod_{i\geqslant 1:\ B_i = j} P_i,$$

where $P_1 P_2 \cdots$ is the prime factorization of N_x as before. With this definition, the vectors $(\nu_p(D_{f_{\mathrm{M}(\boldsymbol{\alpha})},x,1}),\ldots,\nu_p(D_{f_{\mathrm{M}(\boldsymbol{\alpha})},x,k}))$ all follow multinomial distributions for every prime p.

Starting with the pioneering work of Deshouillers, Dress, and Tenenbaum [7], many different people have studied the distribution of $\boldsymbol{D}_{f,x}$ for various choices of $f \in \mathcal{F}_k(\boldsymbol{\alpha})$, and they proved that it converges on a logarithmic scale to a certain *Dirichlet distribution*. Recall that if $\boldsymbol{\alpha} \in \Delta^{k-1}$ satisfies $\alpha_i > 0$ for all i, we say that a Δ^{k-1}-valued random vector $\boldsymbol{Z}$ follows the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ if

$$\mathbb{P}\big[Z_i \leqslant u_i\ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\boldsymbol{u}) \coloneqq \prod_{i=1}^{k} \Gamma(\alpha_i)^{-1} \underset{\substack{0 \leqslant t_i \leqslant u_i\ \forall i < k \\ t_1 + \cdots + t_{k-1} \leqslant 1}}{\int \cdots \int} \prod_{i=1}^{k} t_i^{\alpha_i - 1}\, \mathrm{d}t_1 \cdots \mathrm{d}t_{k-1}.$$

With this notation, most known results on $\boldsymbol{D}_{f,x}$ can be stated in the following form: uniformly for $x \geqslant 2$ and $\boldsymbol{u} \in [0,1]^{k-1}$, we have

$$(1.8) \qquad \mathbb{P}\big[D_{f,x,i} \leqslant N_x^{u_i}\ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\boldsymbol{u}) + O_{\boldsymbol{\alpha}}\big((\log x)^{-\min\{\alpha_1,\ldots,\alpha_k\}}\big).$$

As we indicated above, the first result of this type was obtained in 1979 by Deshouillers, Dress, and Tenenbaum [7], who proved (1.8) for $f = f_{\mathrm{U}(2)}$ (which is also equal to $f_{\mathrm{R}(2)}$; both cases involve sampling a divisor of N_x uniformly). For higher values of k, Nyandwi and Smati [18] extended this result in 2013 by proving (1.8) for $f = f_{\mathrm{U}(3)}$, and La Bretèche and Tenenbaum [6] further proved it for $f = f_{\mathrm{R}(3)}$ in 2016. In 2023, Leung [15] demonstrated that (1.8) applies to $f = f_{\mathrm{U}(k)}$ for all $k \geqslant 2$. Leung also claimed that his proof could be adapted to any f in a more general class of k-dimensional multiplicative functions than $\mathcal{F}_k(\boldsymbol{\alpha})$ (for a proof, see [16]). This more general class allows the quantities $f(1,\ldots,1,p,1,\ldots,1)$, with p being at the jth coordinate, to equal α_j *on average* rather than pointwise.

Returning to 2007, Bareikis and Manstavičius [4] generalized and improved the result of Deshouillers, Dress, and Tenenbaum. They showed that if f belongs to the

more general class of two-dimensional multiplicative functions containing $\mathcal{F}_2(\alpha_1, \alpha_2)$, as described in Leung's claim, then (1.8) holds with the improved error term

$$\mathbb{P}\big[D_{f,x,1} \leqslant N_x^u\big] = F_{(\alpha_1,\alpha_2)}(u) + O_{\alpha_1,\alpha_2}\Big((1 + u\log x)^{-\alpha_1}(1 + (1-u)\log x)^{-\alpha_2}\Big).$$

All results mentioned above rely on techniques from Fourier analysis, such as the Landau–Selberg–Delange method, as their main ingredients to get to their results.

We have a new approach to this problem: since the size of the divisors on a logarithmic scale of any integer is entirely determined by the size of its prime factors, one might expect that the distribution of $\boldsymbol{D}_{f,x}$ can be accessed via a quantitative form of Billingsley's theorem. Indeed, using the coupling from Theorem 1, we establish the following result, which proves (1.8) in full generality with a Bareikis-Manstavičius type error-term.

Theorem 2 (Dirichlet law for the factorization into k parts). — *Let $k \geqslant 2$, let $\boldsymbol{\alpha} \in \Delta^{k-1}$ be fixed with $\alpha_i > 0$ for all i, and let $f \in \mathcal{F}_k(\boldsymbol{\alpha})$. In addition, let $x > 1$ and let $\boldsymbol{D}_{f,x}$ be the random k-factorization corresponding to f.*

For any $\boldsymbol{u} \in [0,1]^{k-1}$ with at least one $i \in \{1, \ldots, k-1\}$ with $u_i \neq 1$, we have

$$\mathbb{P}\big[D_{f,x,i} \leqslant N_x^{u_i} \ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\boldsymbol{u})$$
$$+ O\bigg(\sum_{\substack{1 \leqslant i < k \\ u_i \neq 1}} \frac{1}{(1 + u_i\log x)^{1-\alpha_i}(1 + (1-u_i)\log x)^{\alpha_i}}\bigg);$$

the implied constant in the big-Oh is completely uniform in all parameters.

Remarks

(a) When $u \in [0, \frac{1}{2}]$, we have

$$(1 + u_i\log x)^{1-\alpha_i}(1 + (1-u_i)\log x)^{\alpha_i} \geqslant (0.5\log x)^{\alpha_i}.$$

Similarly, when $u \in [\frac{1}{2}, 1]$, we have

$$(1 + u_i\log x)^{1-\alpha_i}(1 + (1-u_i)\log x)^{\alpha_i} \geqslant (0.5\log x)^{1-\alpha_i}.$$

Therefore, we find that the expression in the big-Oh in Theorem 2 is

$$\leqslant (0.5\log x)^{-\min\{\alpha_1,\ldots,\alpha_k, 1-\alpha_1,\ldots,1-\alpha_k\}}$$

uniformly in $\boldsymbol{u} \in [0,1]^{k-1}$. Since $1 - \alpha_j \geqslant \alpha_i$ for all $i \neq j$ by our assumption that $\alpha_1 + \cdots + \alpha_k = 1$, we conclude that the error term in Theorem 2 is $\ll_{\boldsymbol{\alpha}} (\log x)^{-\min\{\alpha_1,\ldots,\alpha_k\}}$, thus recovering Leung's estimate (1.8) when f lies in the class $\mathcal{F}_k(\boldsymbol{\alpha})$.

(b) If we relied solely on Arratia's original coupling from [2], which leads to (1.2), our approach would yield the same Theorem 2 with each instance of $\log x$ in the error term being replaced by $\log x/\log_2 x$. This adjustment would still offer an improvement over Leung's result if all u_i values are sufficiently far from 0 or 1. We provide further explanations about this point in the remarks following Lemma 9.4.

The proof of Theorem 2 is based on a 1987 result of Donnelly and Tavaré [8], who proved the following probabilistic version of Leung's theorem: If $\boldsymbol{V} = (V_1, V_2, \dots)$ is a Poisson–Dirichlet process, and $(C_i)_{i \geqslant 1}$ is a sequence of i.i.d. random variables that is independent of $\boldsymbol{V}$, supported on $\{1, \dots, k\}$ and satisfying $\mathbb{P}[C_i = j] = \alpha_j$ for all j, then

$$(1.9) \qquad \Big(\sum_{i:\ C_i = 1} V_i, \dots, \sum_{i:\ C_i = k} V_i \Big)$$

follows exactly the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$. In 1998, Arratia [1] used this result to show that $\mathbb{P}\big[D_{f_{\mathrm{U}(2)},x,1} \leqslant N_x^u\big]$ is $F_{(\frac{1}{2},\frac{1}{2})}(u) + o(1)$ as $x \to \infty$ with probabilistic methods. We use the coupling to bridge between the distribution of (1.9) and Theorem 2 and get an explicit error term.

Remark. — Here is a brief heuristic about the shape of the error term we obtain in Theorem 2. Suppose that $N_x \neq 1$ and let $\boldsymbol{\delta}_{f,x} := \big(\frac{\log D_{f,x,1}}{\log N_x}, \dots, \frac{\log D_{f,x,k}}{\log N_x}\big)$. There exists a coupling between $\boldsymbol{\delta}_{f,x}$ and the random vector (1.9) such that their distance is of typical size $\asymp 1/\log x$. For each j, the marginal distribution of the jth component of $\mathrm{Dir}(\boldsymbol{\alpha})$ is $\mathrm{Beta}(\alpha_j, 1 - \alpha_j)$, which is why we get the error term of Theorem 2.

1.2. Structure of the paper. — We have organized the paper in two main parts.

Part 1 contains the proof of Theorem 1. It is divided as follows:

  – In Section 2, we present the coupling implicit in Theorem 1 and present a proof of the latter as a corollary of four key results (Lemmas 2.1–2.3 and Proposition 2.4). Lemma 2.2 is simple and proved right away.

  – In Section 3, we prove Lemma 2.1.

  – In Section 4, we explain another way to realize a GEM process that was presented by Arratia in [2], and we use it to prove Lemma 2.3. This alternative way of describing a GEM process will be also key in proving Proposition 2.4.

  – Sections 5, 6 and 7 are reserved to prove Proposition 2.4.

Finally, Part 2 contains the proof of Theorem 2 and it is organized in the following way:

  – In Section 8, we present the argument due to Donnelly–Tavaré showing that random k-partitions of the components of a Poisson–Dirichlet process are distributed according to a Dirichlet law.

  – In Section 9, we use the coupling of Section 2 to construct a coupling of the random k-factorization $\boldsymbol{D}_{f,x}$ and an analogous k-partition of the components of the Poisson–Dirichlet process. We then use this coupling to reduce Theorem 2 to estimating two boundary events, one involving number-theoretic objects and the other one using pure probabilistic objects.

  – In Section 10, we prove the necessary estimate for the number-theoretic boundary event, and in Section 11 we show the analogous probabilistic estimate.

Remark. — The readers interested only in Theorem 2 do not need to go beyond Section 2 in Part 1.

1.3. Notation. — We let $\log_j$ denote the j-iteration of the natural logarithm, meaning that $\log_1 = \log$ and $\log_j = \log \circ \log_{j-1}$ for $j \geqslant 2$.

Throughout the paper, the letter p is reserved for prime numbers and the letter n is reserved for natural numbers, unless stated otherwise. Given such p and n, we write $\nu_p(n)$ for the p-adic valuation of n, that is to say the largest integer $v \geqslant 0$ such that $p^v | n$. In addition, we write $\omega(n)$ for the number of distinct prime factors of n.

Moreover, for $n \in \mathbb{N}$, we let $s(n)$ denote its largest square-full divisor. Also, we let $n^\flat := n/s(n)$ and we note that $n^\flat$ is always square-free and co-prime to $s(n)$.

We write $\pi(x)$ for the number of primes $\leqslant x$. We shall also use heavily Chebyshev's function $\theta(x) := \sum_{p \leqslant x} \log p$.

To describe various estimates, we use Vinogradov's notation $f(x) \ll g(x)$ or Landau's notation $f(x) = O(g(x))$ to mean that $|f(x)| \leqslant C \cdot g(x)$ for a positive constant C. If C depends on a parameter α, we write $f(x) \ll_\alpha g(x)$ or $f(x) = O_\alpha(g(x))$. If two positive functions f, g have the same order of magnitude in the sense that $f(x) \ll g(x) \ll f(x)$, then we write $f(x) \asymp g(x)$.

If P is some proposition, then the indicator function $\mathbb{1}_P$ will be equal to 1 if P is true and 0 if P is false.

Part 1. Sharpening Arratia's coupling

2. The coupling

In this section, we describe the coupling behind Theorem 1. To construct it, we begin with an ambient probability space Ω containing the following objects:

– a GEM process $\boldsymbol{L} = (L_1, L_2, \dots)$;
– three mutually independent random variables U_1', U_2' and U_3' that are also independent from $\boldsymbol{L}$, and which are uniformly distributed in the open interval $(0, 1)$.

We shall extract an integer N_x and a Poisson–Dirichlet process $\boldsymbol{V}$ as deterministic functions of these random objects. Extracting $\boldsymbol{V}$ is done by sorting the components of $\boldsymbol{L}$ in non-increasing order. The extraction of N_x is more complicated, and we need to introduce additional notation to describe it.

Let $(\lambda_j)_{j \geqslant 0}$ be the increasing sequence of positive real numbers defined by $\lambda_0 := e^{-\gamma}$ and

$$\lambda_j := \exp\left(-\gamma + \sum_{i \leqslant j} \frac{1}{v_i q_i}\right) \quad \text{for } j \geqslant 1,$$

with γ being the Euler-Mascheroni constant and $q_j = p_j^{v_j}$ being the jth smallest prime power, i.e., $(q_j)_{j \geqslant 1}$ is the sequence $2, 3, 2^2, 5, 7, 2^3, 3^2, \dots$ Note that

$$(2.1) \qquad\qquad \lambda_j = \log q_j + O(1/(\log q_j)^2)$$

by a stronger version of Mertens's estimate (Proposition A.2).

Moreover, we have

$$q_{j+1} = q_j + O(q_j/(\log q_j)^3). \tag{2.2}$$

Indeed, the prime number theorem (Proposition A.1) implies that for some sufficiently large constant C, we have $\theta(q_j + Cq_j/(\log q_j)^3) - \theta(q_j) > 0$, whence $q_{j+1} \leqslant q_j + Cq_j/(\log q_j)^3$, as required.

Next, we define the step-function $h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ by

$$h(t) := \sum_{j \geqslant 1} (\log q_j) \cdot \mathbb{1}_{\lambda_{j-1} < t \leqslant \lambda_j}. \tag{2.3}$$

In particular, (2.1) and (2.2) imply that, if $r(t) := |h(t) - t|$, then

$$r(t) \ll \min\{t, t^{-2}\} \quad \text{for all } t > 0. \tag{2.4}$$

Using the above notation, here is how to extract N_x from $\boldsymbol{L}, U_1', U_2'$ and U_3':

(1) Construct the sequence of random prime powers or ones $(Q_i)_{i \geqslant 1}$ by letting $Q_i := e^{h(L_i \log x)}$. Note that we have $Q_i = 1$ whenever $L_i \leqslant e^{-\gamma}/\log x$ so there are only finitely many Q_i's that are prime powers.

(2) Define the random integer $J_x := \prod_{j \geqslant 2} Q_j$.

(3) Define the *extra prime* $P_{\mathtt{extra}}$ as the smallest element in the set $\{1\} \cup \{\text{primes}\}$ that would satisfy $\theta(P_{\mathtt{extra}}) \geqslant U_1' \theta(x/J_x)$, where $\theta(y) = \sum_{p \leqslant y} \log p$ is Chebyshev's function. (In particular, we have $P_{\mathtt{extra}} = 1$ when $J_x > x/2$; otherwise, $P_{\mathtt{extra}}$ is a prime $\leqslant x/J_x$.)

(4) Let μ_x be the probability measure induced by the random variable $M_x := J_x P_{\mathtt{extra}}$, and let ν_x be the uniform counting measure on $\mathbb{Z} \cap [1, x]$. Then, by Lemma B.2 and our assumption that U_2' and U_3' exist on Ω, there exists a random variable N_x on Ω such that:

– N_x is uniformly distributed on $\mathbb{Z} \cap [1, x]$,
– $\mathbb{P}\big[M_x \neq N_x\big] = \mathrm{d}_{\mathrm{TV}}(\mu_x, \nu_x)$

with d_{TV} being the total variation distance defined in (B.1).

This completes the definition of our coupling, since the space Ω contains a Poisson–Dirichlet process $\boldsymbol{V}$ and also a random variable N_x with distribution ν_x.

In Section 2.1, we show how to use the coupling to prove Theorem 1. Lastly, in Section 2.2, we make some technical remarks on the coupling.

2.1. Reducing Theorem 1 to three lemmas and a proposition. — With the following four key results, we directly get Theorem 1. Recall that $s(n)$ denotes the largest square-full divisor of the integer n. In addition, let

$$\Theta_x := \sum_{i \geqslant 1} r(V_i \log x). \tag{2.5}$$

Lemma 2.1 (The ℓ^1 distance within the coupling). — *When $M_x = N_x$, we have the inequality*

$$\sum_{i \geqslant 1} |\log P_i - V_i \log x| \leqslant \log(x/N_x) + 2 \cdot \log s(N_x) + 2 \cdot \Theta_x.$$

Lemma 2.2 (Properties of N_x). — *Fix $\alpha \in [0,1)$ and $\beta \in [0,1/2)$. Uniformly over $x \geqslant 1$, we have*

$$\mathbb{E}\big[(x/N_x)^\alpha s(N_x)^\beta\big] \ll_{\alpha,\beta} 1.$$

Proof. — We must show that

$$S := \sum_{n \leqslant x} (x/n)^\alpha s(n)^\beta \ll_{\alpha,\beta} x.$$

Indeed, if we let $b = s(n)$ and $a = n^\flat = n/b$, then

$$S \leqslant \sum_{\substack{b \leqslant x \\ b \text{ square-full}}} x^\alpha b^{\beta-\alpha} \sum_{a \leqslant x/b} a^{-\alpha} \ll_\alpha \sum_{\substack{b \leqslant x \\ b \text{ square-full}}} x^\alpha b^{\beta-\alpha} (x/b)^{1-\alpha} \leqslant x \sum_{b \text{ square-full}} b^{\beta-1},$$

where we used our assumption that $\alpha < 1$. Since we have assumed that $\beta < 1/2$, the last sum over b converges, thus completing the proof. $\qquad\square$

Lemma 2.3 (Properties of Θ_x). — *Fix $\alpha \geqslant 0$. Uniformly over $x \geqslant 1$, we have*

$$\mathbb{E}\big[e^{\alpha\Theta_x}\big] \ll_\alpha 1.$$

Proposition 2.4 (Total variation distance between M_x and N_x). — *For $x \geqslant 2$, we have*

$$\mathbb{P}\big[M_x \neq N_x\big] \ll \frac{1}{\log x}.$$

The proof of Lemma 2.1 is given in Section 3, and the proof of Lemma 2.3 is given in Section 4. The proof of Proposition 2.4 is the longest part. We set it up in Sections 5 and 6 to eventually give it in Section 7. Here is how we get Theorem 1 with these results:

Proof of Theorem 1. — Let $S := \sum_{i \geqslant 1} |\log P_i - V_i \log x|$. The proof that $\mathbb{E}[S] \gg 1$ is explained in the remark after the statement of the theorem in the introduction. For the upper bound, we first note that we always have the trivial bound $S \leqslant 2 \log x$. This bound and Lemma 2.1 gives us

$$S \leqslant \mathbb{1}_{M_x \neq N_x} \cdot (2 \log x) + \mathbb{1}_{M_x = N_x} \cdot \big(\log(x/N_x) + 2 \cdot \log s(N_x) + 2 \cdot \Theta_x \big).$$

Taking expectations on both sides, we get $\mathbb{E}[S] \ll 1$ with Lemmas 2.2–2.3 and with Proposition 2.4. $\qquad\square$

In fact, if we condition on the event $M_x = N_x$, we can obtain a much stronger bound:

Proposition 2.5. — *Fix $\alpha \in [0, 1/4)$. For $x \geqslant 2$, we have*

$$\mathbb{E}\Big[\exp\big(\alpha \sum_{i \geqslant 1} |\log P_i - V_i \log x|\big) \,\big|\, M_x = N_x \Big] \ll_\alpha 1.$$

Proof. — This follows readily by Hölder's inequality and by Lemmas 2.1–2.3. $\qquad\square$

2.2. Remarks on the coupling

(a) As discussed previously, we have $\lambda_{j-1} \approx \lambda_j \approx \log q_j$, and thus $(\log x)L_i \approx \log Q_i$ as long as L_i is not too small. In particular, we expect that $\sum_{i \geqslant 1} \log Q_i$ would be too close to $\log x$, and thus $\prod_{i \geqslant 1} Q_i$ cannot serve as a proxy of N_x. This is the reason we have to delete Q_1 from the factors of J_x, and we insert instead an extra random prime $P_{\mathtt{extra}}$ conveniently chosen so that $J_x P_{\mathtt{extra}}$ has a distribution close to ν_x.

As we already remarked, we have $P_{\mathtt{extra}} = 1$ if, and only if, $J_x > x/2$ (which happens rarely); otherwise, $P_{\mathtt{extra}}$ is a prime $\leqslant x/J_x$. As a matter of fact, for all $j \in \mathbb{Z} \cap [1, x/2]$, we have

$$(2.6) \qquad \mathbb{P}\big[P_{\mathtt{extra}} = p \,|\, J_x = j\big] = \frac{\mathbb{1}_{p \leqslant x/j} \cdot \log p}{\theta(x/j)}.$$

This is the crucial property that will allow us to show that $M_x = J_x P_{\mathtt{extra}}$ is close to being uniformly distributed.

(b) The coupling we defined above is a modification of Arratia's coupling in [2]. Some of the differences in our definition are purely aesthetic. The one major difference is within step (3), which is the whole reason why we obtain a stronger bound than (1.2). The construction of Arratia's extra prime $P_{\mathtt{Arratia}}$ had a different distribution which satisfied

$$(2.7) \qquad \mathbb{P}\big[P_{\mathtt{Arratia}} = p \,|\, J_x = j\big] = \frac{1}{1 + \pi(x/j)}$$

for all $j \leqslant x$ and all $p \in \{1\} \cup \{\text{primes} \leqslant x/j\}$. It is possible to get the inequality in Lemma 2.1 with Arratia's original coupling. However, it would be impossible to get a version of Proposition 2.4 with a bound better than $\log_2 x/\log x$.

3. The ℓ^1 distance within the coupling

In this section we establish Lemma 2.1. We need the following rearrangement inequality.

Lemma 3.1 (Rearrangement inequality). — *For any two non-increasing sequences of real numbers $(x_i)_{i \geqslant 1}$ and $(y_i)_{i \geqslant 1}$, and for any two permutations $\sigma, \rho \colon \mathbb{N} \to \mathbb{N}$, we have*

$$\sum_{i \geqslant 1} |x_i - y_i| \leqslant \sum_{i \geqslant 1} |x_{\sigma(i)} - y_{\rho(i)}|.$$

Proof. — See [3, Lem. 3.2]. □

Proof of Lemma 2.1. — Recall the definitions of the sequence of prime powers $(Q_i)_{i \geqslant 1}$, the extra prime $P_{\mathtt{extra}}$ and M_x. We create another sequence of primes or ones $(\widetilde{P}_i)_{i \geqslant 1}$ in the following way:

  – We set $\widetilde{P}_1 := P_{\mathtt{extra}}$.
  – If $i \geqslant 2$ with $Q_i = 1$, then we set $\widetilde{P}_i := 1$.
  – If $i \geqslant 2$ with $Q_i > 1$, we set $\widetilde{P}_i$ to be the only prime dividing Q_i.

We let $(\widehat{P}_i)_{i \geqslant 1}$ be the sequence $(\widetilde{P}_i)_{i \geqslant 1}$ in non-increasing order, and we set

$$\widehat{M}_x := \prod_{i \geqslant 1} \widetilde{P}_i = \prod_{i \geqslant 1} \widehat{P}_i.$$

Since $M_x = N_x$, we have $P_i \geqslant \widehat{P}_i$ for all i, because $(\widehat{P}_i)_{i \geqslant 1}$ is a subsequence of the non-increasing sequence $(P_i)_{i \geqslant 1}$. Therefore,

$$\sum_{i \geqslant 1} |\log P_i - V_i \log x| \leqslant \sum_{i \geqslant 1} |\log \widehat{P}_i - V_i \log x| + \sum_{i \geqslant 1} \log(P_i/\widehat{P}_i).$$

We have that $\prod_{i \geqslant 1} P_i/\widehat{P}_i = M_x/\widehat{M}_x$. Furthermore, the integer $M_x/\widehat{M}_x$ only contains prime factors whose square divides N_x. Thus, $M_x/\widehat{M}_x$ divides $s(N_x)$, and

(3.1) $$\sum_{i \geqslant 1} \log(P_i/\widehat{P}_i) = \log(M_x/\widehat{M}_x) \leqslant \log s(N_x),$$

and thus

(3.2) $$\sum_{i \geqslant 1} |\log P_i - V_i \log x| \leqslant \sum_{i \geqslant 1} |\log \widehat{P}_i - V_i \log x| + \log s(N_x).$$

Next, we use the rearrangement inequality (Lemma 3.1) to find

$$\sum_{i \geqslant 1} |\log \widehat{P}_i - V_i \log x| \leqslant |\log P_{\texttt{extra}} - L_1 \log x| + \sum_{i \geqslant 2} |\log \widetilde{P}_i - L_i \log x|$$

$$\leqslant |\log P_{\texttt{extra}} - L_1 \log x| + \sum_{i \geqslant 2} |\log Q_i - L_i \log x|$$

$$+ \sum_{i \geqslant 2} \log(Q_i/\widetilde{P}_i),$$

where we used that $\widetilde{P}_i \leqslant Q_i$ for each i. Moreover, we have $\prod_{i \geqslant 2} Q_i/\widetilde{P}_i = M_x/\widehat{M}_x$, so (3.1) implies

$$\sum_{i \geqslant 1} |\log \widehat{P}_i - V_i \log x| \leqslant |\log P_{\texttt{extra}} - L_1 \log x| + \sum_{i \geqslant 2} |\log Q_i - L_i \log x| + \log s(N_x).$$

Finally, we also note that $\log P_{\texttt{extra}} = \log M_x - \sum_{i \geqslant 2} \log Q_i$ and $L_1 = 1 - \sum_{i \geqslant 2} L_i$. Since we have assumed that $M_x = N_x$, we have

$$|\log P_{\texttt{extra}} - L_1 \log x| \leqslant \log(x/N_x) + \sum_{i \geqslant 2} |\log Q_i - L_i \log x|.$$

Combining the two above displayed inequalities with (3.2), we conclude that

$$\sum_{i \geqslant 1} |\log P_i - V_i \log x| \leqslant \log(x/N_x) + 2 \sum_{i \geqslant 2} |\log Q_i - L_i \log x| + 2 \log s(N_x).$$

To complete the proof, recall that $\log Q_i = h(L_i \log x)$ and $r(t) = |h(t) - t|$, whence

$$\sum_{i \geqslant 2} |\log Q_i - L_i \log x| = \sum_{i \geqslant 2} r(L_i \log x) \leqslant \Theta_x.$$

This proves Lemma 2.1. □

4. Another realization of the GEM distribution

For the proof of Lemma 2.3 and Proposition 2.4, we will need to be more precise as to how the GEM process $\boldsymbol{L}$ is sampled in the coupling. The construction we present below is also the one used by Arratia [2].

Definition 4.1 (The Poisson process $\mathscr{R}$)

(a) We denote by $\mathscr{R}$ the Poisson process on $\mathbb{R}^2_{>0}$ that has intensity measure $e^{-wy}\,\mathrm{d}w\,\mathrm{d}y$. We may assume that the w-coordinates of the points of $\mathscr{R}$ are all distinct since this happens almost surely.

(b) We index the points of $\mathscr{R} = \big\{(W_i, Y_i) : i \in \mathbb{Z}\big\}$ according to the following rules:
   - $W_i < W_{i+1}$ for all $i \in \mathbb{Z}$;
   - if we let $S_i \coloneqq \sum_{\ell \geqslant i} Y_\ell$ for all $i \in \mathbb{Z}$, then we have $S_1 \leqslant \log x < S_0$.

Remark. — By the mapping theorem (Proposition B.3), projecting $\mathscr{R}$ on the w-axis yields a Poisson Process with intensity $\mathrm{d}w/w$ and, similarly, projecting $\mathscr{R}$ on its y-axis yields a Poisson Process with intensity $\mathrm{d}y/y$. Therefore, the w-coordinates of the points in $\mathscr{R}$ have almost surely exactly one limit point at 0 and they are almost surely unbounded. Hence, the indexing (W_i, Y_i) in part (b) of the above definition is well-defined.

The following lemma describes the distribution of the point process S_i.

Lemma 4.2 (Scale-invariant spacing lemma). — *The point process* $\{S_i : i \in \mathbb{Z}\}$ *is a Poisson process on the positive real line with intensity measure* $\mathrm{d}s/s$.

Proof. — See [3, Lem. 7.1]. □

Using this lemma, we have the following description of the GEM distribution.

Proposition 4.3 (Arratia, [2]). — *The process* $\big(1 - \frac{S_1}{\log x}, \frac{Y_1}{\log x}, \frac{Y_2}{\log x}, \dots\big)$ *follows a GEM distribution.*

Proof. — Applying the map $T_x(s) \coloneqq \log_2 x - \log s$ to the points of $\{S_i : i \in \mathbb{Z}\}$ yields a homogeneous Poisson process on the real line with constant rate 1, by the mapping theorem (Proposition B.3) and Lemma 4.2. Furthermore, we have $T_x(S_0) < 0 \leqslant T_x(S_1)$ and $(T_x(S_i))_{i \in \mathbb{Z}}$ increasing. Therefore, $T_x(S_1)$ and $T_x(S_{i+1}) - T_x(S_i)$ for all $i \geqslant 1$ are independent exponential random variables of parameter 1. If X is a standard exponential random variable, then $1 - e^{-X}$ is a uniform random variable in $[0, 1]$. We conclude that $1 - S_1/\log x, Y_1/S_1, Y_2/S_2, \dots$ are independent uniform random variables in $[0, 1]$. The proposition follows by the characterization of the GEM distribution described in the introduction. □

For the next sections and the proof below, we will assume that the GEM process $\boldsymbol{L} = (L_1, L_2, \dots)$ sampled for our coupling was determined by $\mathscr{R}$ by defining

$L_1 := 1 - S_1/\log x$ and $L_j := Y_{j-1}/\log x$ for $j \geqslant 2$. In this setting, we now have

$$(4.1) \qquad J_x = \prod_{i \geqslant 1} e^{h(Y_i)}.$$

4.1. **Proof of Lemma 2.3.** — We conclude this section by using the realization of the GEM described here to prove Lemma 2.3. Note that

$$(4.2) \qquad \Theta_x = r(L_1 \log x) + \sum_{i \geqslant 1} r(Y_i) \leqslant \Theta_\infty + O(1)$$

with $\Theta_\infty := \sum_{i \in \mathbb{Z}} r(Y_i)$. With Campbell's theorem (Proposition B.5), we directly compute that

$$(4.3) \qquad \mathbb{E}[e^{\alpha \Theta_\infty}] = \exp\left(\int_0^\infty \frac{e^{\alpha r(y)} - 1}{y}\, \mathrm{d}y\right).$$

This integral is convergent for all fixed $\alpha > 0$ because $e^{\alpha r(y)} - 1 \ll_\alpha r(y) \ll \min\{y, y^{-2}\}$. Combining this fact with (4.2) proves that $\mathbb{E}[e^{\alpha \Theta_x}] \ll_\alpha 1$. We have thus established Lemma 2.3.

5. An integer-friendly version of J_x

Let Λ be the von Mangoldt function, that is to say

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \text{ for some prime power } p^k, \\ 0 & \text{otherwise.} \end{cases}$$

Recall the Poisson Process $\mathscr{R}$ given in Definition 4.1. We then define

$$\mathscr{R}^* := \left\{\left(WY/h(Y), e^{h(Y)}\right) : (W, Y) \in \mathscr{R},\ Y > e^{-\gamma}\right\}.$$

Without loss of generality, we may assume that the points $WY/h(Y)$ with $(W, Y) \in \mathscr{R}$ and $Y > e^{-\gamma}$ are all distinct. By the mapping theorem (Proposition B.3) applied to the map $(w, y) \mapsto (wy/h(y), e^{h(y)})$ on the Poisson process $\mathscr{R}$ restricted to $\mathbb{R}_{>0} \times \mathbb{R}_{>e^{-\gamma}}$, the random set $\mathscr{R}^*$ is a Poisson process on the space $\mathbb{R}_{>0} \times \{\text{prime powers}\}$ with mean measure μ^* satisfying

$$\mu^*(B \times \{q\}) = \int_B \frac{\Lambda(q)}{q^{1+t}}\, \mathrm{d}t$$

for any $B \subseteq \mathbb{R}_{>0}$ and any $q \in \mathbb{N}$. Note that the w-coordinates of the points in $\mathscr{R}$ restricted to $\mathbb{R}_{>0} \times \mathbb{R}_{>e^{-\gamma}}$ are now bounded with probability one. For each such realization of the points of $\mathscr{R}^*$, there is a unique labeling of them as $\{(T_i^*, Q_i^*) : i \in \mathbb{Z}_{\leqslant K}\}$ in such a way that the following properties hold:

- $T_{i-1}^* < T_i^*$ for all $i \in \mathbb{Z}_{\leqslant K}$;
- $\prod_{i=1}^K Q_i^* \leqslant x < \prod_{i=0}^K Q_i^*$.

(Note that K is a random variable and is not fixed.) We then define the random integer

$$J_x^* := \prod_{i=1}^K Q_i^*.$$

One advantage of introducing J_x^* is that the computation for the distribution of J_x^* can be done more easily and precisely than the distribution of J_x. We perform this calculation in this section. We will then see in the next section that the probability that J_x is different from J_x^* is small enough to be negligible in the calculation of the total variation distance in Section 7.

This random integer J_x^* was used by Arratia in [2], and he estimated its distribution in his Lemma 2. For the sake of completeness, we give a proof for this estimation.

Lemma 5.1 (Arratia [2]). — *For $x \geqslant 2$ and $1 \leqslant j \leqslant x$, we have*

$$\mathbb{P}\big[J_x^* = j\big] = \frac{1}{j \log x}\big(1 + O(1/\log x)\big).$$

Proof. — Let $t > 0$ and let q be a prime power, and consider the random variable $N_q(t)$ that counts the number of points (T_i^*, Q_i^*) with $T_i^* > t$ and $Q_i^* = q$. Note that $N_q(t)$ has Poisson distribution with parameter $\Lambda(q)/q^{1+t} \log q$. Moreover, let $I_t^* := \prod q^{N_q(t)}$ with the product being over all prime powers q. On one hand, when t is fixed, the expectation $\mathbb{E}[(I_t^*)^{-s}]$ can be expanded as a Dirichlet series whose i^{th} coefficient equals $\mathbb{P}[I_t^* = i]$. This Dirichlet series converges absolutely for $\text{Re}(s) \geqslant 0$. On the other hand, we have

$$\mathbb{E}[(I_t^*)^{-s}] = \prod_q \mathbb{E}\big[q^{-sN_q(t)}\big] = \prod_q \exp\Big(\tfrac{\Lambda(q)}{q^{1+t} \log q}(q^{-s} - 1)\Big) = \frac{\zeta(1+t+s)}{\zeta(1+t)}.$$

It follows that $\mathbb{P}[I_t^* = i] = i^{-1-t}/\zeta(1+t)$ for all $i \geqslant 1$. All points of $\mathscr{R}^*$ have distinct T^*-coordinates with probability one. In this situation, we have that $J_x^* = j$ if, and only if, there exists exactly one point $(T^*, Q^*) \in \mathscr{R}^*$ such that $I_{T^*}^* = j$ and $Q^* > x/j$. Thus, we have

$$\mathbb{1}_{J_x^* = j} = \sum_{(T^*, Q^*) \in \mathscr{R}^*} \mathbb{1}_{I_{T^*}^* = j \text{ and } Q^* > x/j}$$

almost surely. Taking expectations on both sides and using the Mecke equation (Proposition B.6), we obtain the distribution of J_x^*:

$$(5.1) \qquad \mathbb{P}\big[J_x^* = j\big] = \int_0^\infty \bigg(\sum_{q > x/j} \frac{\Lambda(q)}{q^{1+t}} \bigg) \cdot \frac{j^{-1-t}}{\zeta(1+t)}\, \mathrm{d}t.$$

We want to estimate this integral. Let $S(u) := \sum_{q \leqslant u} \Lambda(q)/q$. We have $S(u) = \log u + O(1)$ for $u \geqslant 1$ by Mertens's estimate [13, Th. 3.4(a)]. We use partial summation to get

$$\sum_{q > x/j} \frac{\Lambda(q)}{q^{1+t}} = \int_{x/j}^\infty u^{-t}\, \mathrm{d}S(u) = \frac{(x/j)^{-t}}{t}\,(1 + O(t)).$$

for all $t > 0$. By putting this estimate in (5.1), we have

$$\mathbb{P}\big[J_x^* = j\big] = \frac{1}{j} \int_0^\infty \frac{x^{-t}(1 + O(t))}{t \cdot \zeta(1+t)}\, \mathrm{d}t.$$

Since $\zeta(1+t) \geqslant 1$ for all $t > 0$, the portion of the integral over $t \geqslant 1$ is $\ll 1/(x \log x)$. On the other hand, if $t \in (0, 1]$, we have $1/\zeta(1+t) = t + O(t^2)$. We conclude that

$$\mathbb{P}[J_x^* = j] = \frac{1}{j} \int_0^1 x^{-t}(1 + O(t)) \, dt + O(1/jx \log x) = \frac{1}{j \log x}\big(1 + O(1/\log x)\big).$$

This concludes the proof. $\qquad\qquad\square$

6. When J_x and J_x^* are different

To prove Proposition 2.4, we must get a hold of the distribution of J_x. Since we have a good approximation for the distribution of J_x^*, it will be enough, for our purposes, to show that the event $\{J_x \neq J_x^*\}$ occurs with low probability.

For any $t > 0$, consider the random variable

$$I_t := \sum_{(W,Y) \in \mathscr{R}} Y \cdot \mathbb{1}_{W>t}.$$

We compute below the distribution of I_t.

Proposition 6.1 (Arratia, [2]). — *For any fixed $t > 0$, the random variable I_t follows an exponential distribution of parameter t, i.e., $\mathbb{P}[I_t > y] = e^{-ty}$ for all $y > 0$.*

Proof. — A direct application of Campbell's theorem (Proposition B.5) implies that $\mathbb{E}[e^{sI_t}] = t/(t-s)$ for $\mathrm{Re}(s) < t$, which agrees with the moment generating function of an exponential distribution of parameter t. This completes the proof. $\qquad\square$

Let η be the smallest positive constant satisfying

$$\frac{a}{h(a)} \cdot \frac{h(b)}{b} \leqslant 1 + \frac{\eta}{\min\{a,b\}^2}.$$

for all $a, b > e^{-\gamma}$. Such a constant must exist by (2.4). In addition, let

$$r_0 := \sup_{y>0} r(y).$$

Let us then define the following events:

- $\mathcal{E}_1 := \big\{S_1 < \log x - R_x - r_0, \ S_0 > \log x + R_x + 2r_0\big\}$, where $R_x := \sum_{i \geqslant 2} r(Y_i)$.
- $\mathcal{E}_2$ is the event where $W_i/W_0 > 1 + \eta/\min\{Y_0, Y_i\}^2$ for all $i \geqslant 1$.
- $\mathcal{E}_3$ is the event where $W_0/W_i > 1 + \eta/\min\{Y_0, Y_i\}^2$ for all $i \leqslant -1$.

The variable R_x depends on the value of x since the labeling of points in $\mathscr{R}$ change as x grows, even if $\mathscr{R}$ stays fixed.

Lemma 6.2. — *For $x > 1$, we have $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \subseteq \{J_x = J_x^*\}$. In particular,*

$$\mathbb{P}\big[J_x \neq J_x^*\big] \leqslant \mathbb{P}[\mathcal{E}_1^c] + \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] + \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c].$$

Proof. — Recall that $J_x = \prod_{i \geqslant 1} e^{h(Y_i)}$ and assume that $\mathcal{E}_1$ occurs. Then

$$(6.1) \qquad \log J_x = \sum_{i=1}^{\infty} h(Y_i) \leqslant \sum_{i \geqslant 1} \big(Y_i + r(Y_i)\big) \leqslant S_1 + R_x + r_0 < \log x,$$

since $r(Y_1) \leqslant r_0$. Similarly, we have

$$\log\big(J_x \cdot e^{h(Y_0)}\big) = \sum_{i=0}^{\infty} h(Y_i) \geqslant \sum_{i \geqslant 0} \big(Y_i - r(Y_i)\big) \geqslant S_0 - R_x - 2r_0 > \log x.$$

Therefore, $\mathcal{E}_1$ implies the inequalities

(6.2) $J_x < x < J_x e^{h(Y_0)}.$

In particular, we must have $Y_0 > e^{-\gamma}$.

Assume now further that $\mathcal{E}_2$ and $\mathcal{E}_3$ also occur. We claim that this implies $J_x = J_x^*$. Let B_+ be the set of integers $i \geqslant 1$ such that $Y_i > e^{-\gamma}$, and let B_- be the set of integers $i \leqslant -1$ such that $Y_i > e^{-\gamma}$. By our assumption that $\mathcal{E}_2 \cap \mathcal{E}_3$ occurs and by the definition of η, we have

(6.3) $\dfrac{W_i Y_i}{h(Y_i)} > \dfrac{W_0 Y_0}{h(Y_0)}$ if $i \in B_+$

and, similarly,

(6.4) $\dfrac{W_i Y_i}{h(Y_i)} < \dfrac{W_0 Y_0}{h(Y_0)}$ if $i \in B_-$.

Now, let $K \geqslant 0$ and (T_j^*, Q_j^*) with $j \in \mathbb{Z}_{\leqslant K}$ be the indexing of points of $\mathscr{R}^*$ given in Section 5, that is to say we have $T_{j-1}^* < T_j^*$ for all $j \in \mathbb{Z}_{\leqslant K}$, and such that $\prod_{j=0}^{K} Q_j^* > x \geqslant \prod_{j=1}^{K} Q_j^*$. Let $j_0 \in \mathbb{Z}_{\leqslant K}$ be such that $T_{j_0}^* = W_0 Y_0 / h(Y_0)$ (which exists because $Y_0 > e^{-\gamma}$). Using relations (6.3) and (6.4), we find

$$\prod_{j_0 < j \leqslant K} Q_j^* = \prod_{i \in B_+} e^{h(Y_i)} = \prod_{i \geqslant 1} e^{h(Y_i)} = J_x.$$

On the other hand, we have $\prod_{j_0 \leqslant j \leqslant K} Q_j^* = e^{h(Y_0)} J_x$. Hence, using (6.2), we find that $\prod_{j_0 < j \leqslant K} Q_j^* < x < \prod_{j_0 \leqslant j \leqslant K} Q_j^*$. In particular, $j_0 = 0$ and thus $J_x^* = \prod_{j_0 < j \leqslant K} Q_j^* = J_x$, as claimed.   □

The following lemma requires a lot more care than in Arratia's paper [2] since he only needed $\mathbb{P}[J \neq J^*] \ll \log_2 x / \log x$ to be true for his coupling.

Lemma 6.3. — *We have* $\mathbb{P}[\mathcal{E}_1^c] \ll 1/\log x$ *for* $x \geqslant 2$.

Proof. — We may assume that x is large enough. We have $\mathbb{E}[e^{R_x}] \leqslant \mathbb{E}[e^{\Theta_\infty}] \ll 1$ by (4.3). Therefore, the event $R_x > \log_2 x$ occurs with probability $\ll 1/\log x$. Using Lemma 4.2, we find that the probability that there exists $i \in \mathbb{Z}$ such that $|S_i - \log x| \leqslant 4r_0$ is $O(1/\log x)$. In conclusion,

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{k \in \{0,1\}} \mathbb{P}\Big[4r_0 < |S_k - \log x| \leqslant R_x + 2r_0, \ R_x \leqslant \log_2 x\Big] + O\big(1/\log x\big).$$

Note that if there exists $k \in \mathbb{Z}$ with $W_k > (\log_2 x)^{-10}$ and $|S_k - \log x| \leqslant \log_2 x + 2r_0$,

then $I_{(\log_2 x)^{-10}} \geqslant \log x - \log_2 x - 2r_0$. So, using Proposition 6.1, we find

$$\mathbb{P}\Big[\exists k : |S_k - \log x| \leqslant \log_2 x + 2r_0, \ W_k > (\log_2 x)^{-10}\Big]$$
$$\leqslant \mathbb{P}\big[I_{(\log_2 x)^{-10}} > \log x - \log_2 x - 2r_0\big] = \exp\Big(- \frac{\log x - \log_2 x - 2r_0}{(\log_2 x)^{10}}\Big)$$

for x large enough. Consequently,

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{k\in\{0,1\}} \mathbb{P}\Big[4r_0 < |S_k - \log x| \leqslant R_x + 2r_0, \ R_x \leqslant \log_2 x, \ W_k \leqslant (\log_2 x)^{-10}\Big]$$
$$+ O\big(1/\log x\big).$$

Let m_0 be the largest integer such that $2^{m_0-1}r_0 \leqslant \log_2 x$. Hence, if $2r_0 < R_x \leqslant \log_2 x$, then there exists a unique integer $m \in [2, m_0]$ such that $2^{m-1}r_0 < R_x \leqslant 2^m r_0$, in which case the condition $|S_k - \log x| \leqslant R_x + 2r_0$ implies that $|S_k - \log x| \leqslant 2^m r_0 + 2r_0 \leqslant 2^{m+1}r_0$. We conclude that

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{\substack{k\in\{0,1\}\\2\leqslant m\leqslant m_0}} \mathbb{P}\Big[R > 2^{m-1}r_0, \ |S_k - \log x| \leqslant 2^{m+1}r_0, \ W_k \leqslant (\log_2 x)^{-10}\Big]$$
$$+ O\big(1/\log x\big).$$

Note that $\mathbb{1}_{R_x>2^{m-1}r_0} \leqslant (4^{1-m}/r_0^2)\mathbb{1}_{R_x>2r_0}R_x^2$. In addition, if $R_x > 2r_0$, then we have $\sum_{i\geqslant 2} r(Y_i)^2 \leqslant r_0 R_x < R_x^2/2$, whence $R_x^2 \leqslant 4\sum_{i>j\geqslant 2} r(Y_i)r(Y_j)$. We conclude that

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{\substack{k\in\{0,1\}\\2\leqslant m\leqslant m_0}} \frac{4^{2-m}}{r_0^2}\mathbb{E}\Big[\sum_{i,j:\ i>j>k} r(Y_i)r(Y_j)\cdot \mathbb{1}_{|S_k-\log x|\leqslant 2^{m+1}r_0}\cdot \mathbb{1}_{W_k\leqslant(\log_2 x)^{-10}}\Big]$$
$$+ O\big(1/\log x\big),$$

Therefore, to complete the proof of the lemma, it is enough to show that

$$(6.5)\quad E(z) := \mathbb{E}\Big[\sum_{i,j,k\in\mathbb{Z}:\ i>j>k} r(Y_i)r(Y_j)\cdot \mathbb{1}_{|S_k-\log x|\leqslant z}\cdot \mathbb{1}_{W_k\leqslant(\log_2 x)^{-10}}\Big] \ll \frac{z}{\log x}$$

uniformly for $z \in [0, 4\log_2 x]$.

For the rest of the proof, we fix $z \in [0, 4\log_2 x]$. Given $t'' > t' > t > 0$, let

$$I_{t,t',t''} := \sum_{\substack{(W,Y)\in\mathscr{R}\\W\in\mathbb{R}_{>t}\setminus\{t',t''\}}} Y.$$

Since the W_i's are almost surely distinct, we have

$$E(z) \leqslant \mathbb{E}\Big[\sum\sum\sum r(Y')r(Y'')\cdot \mathbb{1}_{|Y+Y'+Y''+I_{W,W',W''}-\log x|\leqslant z}\cdot \mathbb{1}_{W\leqslant(\log_2 x)^{-10}}\Big].$$

with the triple sum being over points $(W,Y),(W',Y'),(W'',Y'') \in \mathscr{R}$ satisfying $W'' > W' > W$. Hence, using the Mecke equation (Proposition B.6) and the fact

that $I_{t,t',t''}$ has the same distribution as I_t, we find

$$E(z) \leqslant \int \cdots \int_{\substack{0 < w < w' < w'' \\ w \leqslant (\log_2 x)^{-10} \\ y,y',y'' \geqslant 0 \\ y+y'+y'' \leqslant \log x + z}} r(y')r(y'')\mathbb{P}\Big[\big|I_w + y + y' + y'' - \log x\big| \leqslant z\Big]\frac{\mathrm{d}w \cdots \mathrm{d}y''}{e^{wy+w'y'+w''y''}},$$

where the integral is sixfold with variables w, w', w'', y, y', y''. Proposition 6.1 implies

$$\mathbb{P}\Big[\big|I_w + y + y' + y'' - \log x\big| \leqslant z\Big] \leqslant e^{-w(\log x - y - y' - y'')}(e^{wz} - e^{-wz})$$
$$\ll wze^{-w(\log x - y - y' - y'')},$$

since $w \leqslant (\log_2 x)^{-10}$ and $z \leqslant 4\log_2 x$. Consequently,

$$E(z) \ll \int \cdots \int_{\substack{0 < w < w' < w'' \\ y,y',y'' \geqslant 0 \\ y+y'+y'' \leqslant \log x + z}} r(y')r(y'')zwe^{-w\log x - (w'-w)(y'+y'') - (w''-w')y''} \,\mathrm{d}w \cdots \mathrm{d}y''$$

$$= \frac{z}{(\log x)^2} \iiiint_{\substack{t,y,y',y'' \geqslant 0 \\ y+y'+y'' \leqslant \log x + z}} \frac{r(y')r(y'')}{y''(y'+y'')}te^{-t} \,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}y'\,\mathrm{d}y'',$$

where we made the change of variables $t = w\log x$. Since $\int_0^\infty r(u)/u\,\mathrm{d}u \ll 1$ and $\log x + z \ll \log x$, relation (6.5) follows. This completes the proof of the lemma. $\quad\square$

Lemma 6.4. — *We have $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] + \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c] \ll 1/\log x$ for $x \geqslant 2$.*

Proof. — Recall that $\mathcal{E}_2$ failing means that there exists $i \geqslant 1$ such that $W_i/W_0 \leqslant 1 + \eta/\min\{Y_0, Y_i\}^2$. In addition, recall that the event $\mathcal{E}_1$ implies that $Y_0 > e^{-\gamma}$ (this was explained in the beginning of the proof of Lemma 6.2). Hence, we have

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] \leqslant \mathbb{E}\Bigg[\sum_{\substack{(W,Y) \in \mathcal{R} \\ (W',Y') \in \mathcal{R}}} \mathbb{1}_{1 < \frac{W'}{W} \leqslant 1 + \eta/\min\{Y,Y'\}^2} \cdot \mathbb{1}_{Y',Y > e^{-\gamma}} \cdot \mathbb{1}_{I_{W,W'} + Y' \in (\log x - Y, \log x]}\Bigg]$$

with $I_{t,t'} := \sum_{(W,Y) \in \mathcal{R},\, W \in \mathbb{R}_{>t} \smallsetminus \{t'\}} Y$. We use Mecke's equation as in the proof of (6.5) to get that the expectation above is equal to

$$\iiiint_{\substack{0 < w < w' < w(1+\eta/\min\{y,y'\}^2) \\ y,y' > e^{-\gamma},\, y' \leqslant \log x}} \mathbb{P}\big[\log x - y < I_w + y' \leqslant \log x\big] \cdot e^{-wy-w'y'} \,\mathrm{d}w\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y'.$$

We have $e^{-w'y'} \leqslant e^{-wy'}$, thus

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] \ll \iiint_{\substack{w>0,\, y,y' > e^{-\gamma} \\ y' \leqslant \log x}} \mathbb{P}\big[\log x - y < I_w + y' \leqslant \log x\big] \cdot \frac{we^{-w(y+y')}}{\min\{y,y'\}^2} \,\mathrm{d}w\,\mathrm{d}y\,\mathrm{d}y'.$$

By Proposition 6.1, we have

$$\mathbb{P}\big[\log x - y < I_w + y' \leqslant \log x\big] \leqslant \begin{cases} e^{-w(\log x - y - y')} & \text{if } y \leqslant \log x, \\ 1 & \text{if } y > \log x. \end{cases}$$

Therefore,

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] \ll \iiint_{\substack{w>0,\ y,y'>e^{-\gamma} \\ y,y' \leqslant \log x}} \frac{w e^{-w \log x}}{\min\{y,y'\}^2}\, \mathrm{d}w\, \mathrm{d}y\, \mathrm{d}y' + \iiint_{\substack{w>0,\ y,y'>e^{-\gamma} \\ y' \leqslant \log x < y}} \frac{w e^{-w(y+y')}}{(y')^2}\, \mathrm{d}w\, \mathrm{d}y\, \mathrm{d}y'$$

$$\leqslant \iint_{e^{-\gamma}<y,y' \leqslant \log x} \frac{1}{\min\{y,y'\}^2 (\log x)^2}\, \mathrm{d}y\, \mathrm{d}y' + \iint_{\substack{y,y'>e^{-\gamma} \\ y' \leqslant \log x < y}} \frac{1}{(yy')^2}\, \mathrm{d}y\, \mathrm{d}y'$$

$$\ll 1/\log x.$$

This completes the proof of the claimed bound on $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c]$.

Finally, we bound $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c]$ using a very similar argument. We have

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c]$$

$$\leqslant \mathbb{E}\left[\sum\sum_{(W,Y),(W',Y') \in \mathscr{R}} \mathbb{1}_{1 < \frac{W}{W'} \leqslant 1 + \frac{\eta}{\min\{Y,Y'\}^2}} \cdot \mathbb{1}_{Y',Y>e^{-\gamma}} \cdot \mathbb{1}_{I_W \in (\log x - Y, \log x]} \right]$$

$$= \iiiint_{\substack{0<w'<w<w'(1+\eta/\min\{y,y'\}^2) \\ y,y'>e^{-\gamma}}} \mathbb{P}\big[\log x - y < I_w \leqslant \log x\big] \cdot e^{-wy-w'y'}\, \mathrm{d}w\, \mathrm{d}w'\, \mathrm{d}y\, \mathrm{d}y'$$

$$\leqslant J_1 + J_2,$$

where

$$J_1 := \iiiint_{\substack{0<w'<w<w'(1+\eta/\min\{y,y'\}^2) \\ y,y'>e^{-\gamma},\ y \leqslant \log x}} e^{-w \log x - w'y'}\, \mathrm{d}w\, \mathrm{d}w'\, \mathrm{d}y\, \mathrm{d}y'$$

and

$$J_2 := \iiiint_{\substack{0<w'<w<w'(1+\eta/\min\{y,y'\}^2) \\ y > \log x,\ y'>e^{-\gamma}}} e^{-wy-w'y'}\, \mathrm{d}w\, \mathrm{d}w'\, \mathrm{d}y\, \mathrm{d}y'.$$

Using $e^{-w \log x} \leqslant e^{-w' \log x}$, we find

$$J_1 \ll \iiint_{\substack{w'>0,\ y,y'>e^{-\gamma} \\ y \leqslant \log x}} \frac{w' e^{-w'(y' + \log x)}}{\min\{y,y'\}^2}\, \mathrm{d}w'\, \mathrm{d}y\, \mathrm{d}y'$$

$$= \iint_{\substack{y,y'>e^{-\gamma} \\ y \leqslant \log x}} \frac{1}{\min\{y,y'\}^2 (y' + \log x)^2}\, \mathrm{d}y\, \mathrm{d}y'$$

$$\ll 1/\log x.$$

Similarly, we have

$$J_2 \ll \underset{\substack{w'>0,\ y'>e^{-\gamma} \\ y>\log x}}{\iiint} \frac{w'e^{-w'(y+y')}}{\min\{y,y'\}^2}\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y'$$

$$= \underset{\substack{y'>e^{-\gamma} \\ y>\log x}}{\iint} \frac{1}{\min\{y,y'\}^2(y+y')^2}\,\mathrm{d}y\,\mathrm{d}y'$$

$$\ll 1/\log x.$$

This implies that $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c] \ll 1/\log x$, thus completing the proof of the lemma. $\quad\square$

As an immediate corollary of Lemmas 6.2, 6.3 and 6.4, we have:

PROPOSITION 6.5. — *For $x \geqslant 2$, we have $\mathbb{P}\big[J_x \neq J_x^*\big] \ll 1/\log x$.*

7. Proof of Proposition 2.4

With a good estimation of the distribution of J_x^* and with the fact that $J_x = J_x^*$ with a quantifiably high probability, we are able to give an upper bound on the total variation distance between the distribution of M_x and N_x.

Proof of Proposition 2.4. — Recall that we constructed N_x in the coupling such that $\mathbb{P}[M_x \neq N_x] = \mathrm{d}_{\mathrm{TV}}(\mu_x, \nu_x)$ with μ_x and ν_x being the distribution of M_x and N_x respectively. Note that $M_x > x$ must imply that $J_x > x$, which means that $\mathcal{E}_1$ cannot happen in that situation by (6.1). Hence, using Lemma 6.3, we see that it is enough to show that

$$\mathbb{P}[M_x \in A] = \frac{\#A}{\lfloor x \rfloor} + O\big(1/\log x\big)$$

uniformly over all $A \subseteq \mathbb{Z} \cap [1, x]$.

For each $j \leqslant x$, we define the sets

$$A_j := \bigcup_{p:\ pj \in A} \Big(\frac{\theta(p-1)}{\theta(x/j)}, \frac{\theta(p)}{\theta(x/j)} \Big].$$

Note that $M_x \in A$ and $J_x \leqslant x/2$ if, and only if, $J_x = j$ for some $j \leqslant x/2$ and $U_1' \in A_j$. With this in mind, we define the following events:

- $\mathcal{B}_1 := \{J_x = J_x^* \leqslant x/2\}$;
- $\mathcal{B}_2 := \{M_x \in A\}$;
- $\mathcal{B}_3 := \{J_x^* = j \text{ and } U_1' \in A_j \text{ for some } j \leqslant x/2\}$.

Since $\mathcal{B}_1 \cap \mathcal{B}_2 = \mathcal{B}_1 \cap \mathcal{B}_3$, we have

$$\Big| \mathbb{P}[\mathcal{B}_2] - \mathbb{P}[\mathcal{B}_3] \Big| \leqslant \mathbb{P}\big[\mathcal{B}_1^c\big] \leqslant \mathbb{P}\big[J_x \neq J_x^*\big] + \mathbb{P}\big[x/2 < J_x^* \leqslant x\big] \ll \frac{1}{\log x}$$

by Lemma 5.1 and by Proposition 6.5. Therefore, we have

$$\mathbb{P}\big[M_x \in A\big] = \sum_{j \leqslant x/2} \mathbb{P}\big[J_x^* = j, \ U_1' \in A_j\big] + O(1/\log x)$$

(7.1)

$$= \sum_{p,j:\ pj\in A} \frac{\log p}{\theta(x/j)} \cdot \mathbb{P}[J_x^* = j] + O(1/\log x).$$

Since $\theta(t)/t = 1 + O\big((\log t)^{-2}\big)$ for all $t \geqslant 2$ by the prime number theorem [13, Th. 8.1], we have

(7.2) $$\frac{\mathbb{P}[J_x^* = j]}{\theta(x/j)} = \frac{1}{\lfloor x \rfloor \log x} \left(1 + O\left(\frac{1}{(\log(x/j))^2} + \frac{1}{\log x}\right)\right)$$

for $j \leqslant x/2$. In addition, note that

(7.3) $$\sum_{p,j:\ pj\in A} \log p \cdot \left(\frac{1}{(\log(x/j))^2} + \frac{1}{\log x}\right) \ll \sum_{j \leqslant x/2} \left(\frac{x}{j(\log(x/j))^2} + \frac{x}{j \log x}\right) \ll x.$$

By combining (7.1), (7.2) and (7.3), we reduce the problem to showing the following:

(7.4) $$\sum_{p,j:\ pj\in A} \frac{\log p}{\log x} = \#A + O\big(x/\log x\big).$$

If we set $L(a) := \sum_{p|a} \log p$, then we have

(7.5) $$\sum_{p,j:\ pj\in A} \frac{\log p}{\log x} = \sum_{a\in A} \frac{L(a)}{\log x} = \#A - \sum_{a\in A} \frac{\log(x/a) + (\log a - L(a))}{\log x}.$$

The quantity $\log(x/a)$ is non-negative whenever $a \leqslant x$. Furthermore, the integer $a/(\prod_{p|a} p)$ is always a divisor of $s(a)$, thus $0 \leqslant \log a - L(a) \leqslant \log s(a)$. Therefore,

$$\sum_{a\in A} \frac{\log(x/a) + (\log a - L(a))}{\log x} \ll \sum_{a \leqslant x} \frac{\log(x/a) + \log s(a)}{\log x}$$

$$= \frac{\lfloor x \rfloor}{\log x} \cdot \mathbb{E}\big[\log(x/N_x) + \log s(N_x)\big]$$

$$\ll \frac{x}{\log x}$$

by Lemma 2.2. With this inequality combined with (7.5), we establish (7.4), and hence it completes the proof of Proposition 2.4.   □

Part 2. Factorization into k parts

8. Theorem 2 for the probabilistic model

With the following probabilistic version of Theorem 2 given by Donnelly and Tavaré in 1987 [8, §3], we see how a Poisson–Dirichlet process relates to the Dirichlet distribution. We repeat their proof since they use a Poisson process that will come in handy later.

Proposition 8.1 (Donnelly–Tavaré [8]). — *Let $\boldsymbol{V} = (V_1, V_2, \dots)$ be a Poisson–Dirichlet process, let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k) \in \Delta^{k-1}$ such that $\alpha_i > 0$ for all $1 \leqslant i \leqslant k$, and let $(C_i)_{i \geqslant 1}$ be a sequence of i.i.d. random variables (also independent of $\boldsymbol{V}$) with $\mathbb{P}[C_i = j] = \alpha_j$ for all $i \geqslant 1$. Then the random vector*

$$\Big(\sum_{i:\, C_i = 1} V_i, \dots, \sum_{i:\, C_i = k} V_i \Big)$$

is distributed according to $\mathrm{Dir}(\boldsymbol{\alpha})$.

Proof. — Let $X_1 > X_2 > \cdots$ be the points of a Poisson process on $\mathbb{R}_{>0}$ with intensity $(e^{-x}/x)\,\mathrm{d}x$, and let $S := \sum_{i \geqslant 1} X_i$. For each $i \geqslant 1$, Let $V_i := X_i/S$. We then know that $(V_1, V_2, \dots)$ follows the Poisson–Dirichlet distribution [9, Th. 2.2]. With the coloring theorem (Proposition B.4), the point processes

$$\Pi_m = \{X_i : C_i = m\}$$

form independent Poisson processes of intensity $\alpha_m \cdot (e^{-x}/x)\,\mathrm{d}x$ for all $m = 1, \dots, k$. We then let

$$S_m := \sum_{X \in \Pi_m} X$$

for $m = 1, 2, \dots, k$. With Campbell's theorem (Proposition B.5), we compute the moment generating function

$$\log \mathbb{E}[e^{sS_m}] = \alpha_m \cdot \int_0^\infty \frac{e^{sx} - 1}{x} \cdot e^{-x}\,\mathrm{d}x = -\alpha_m \log(1 - s)$$

for $\mathrm{Re}(s) < 1$. This coincides with the moment generating function of $\mathrm{Gamma}(\alpha_m, 1)$ distribution. We thus deduce that the vector

$$\Big(\sum_{i:\, C_i = 1} V_i, \dots, \sum_{i:\, C_i = k} V_i \Big) = \Big(\frac{S_1}{\sum_{m=1}^k S_m}, \dots, \frac{S_k}{\sum_{m=1}^k S_m} \Big)$$

follows the distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ (see [12, Ch. 49, pp. 485-487] for a proof). $\qquad\square$

9. A coupling for $\boldsymbol{D}_{f,x}$

Fix $k \geqslant 2$, $\boldsymbol{\alpha} \in \Delta^{k-1}$ with $\alpha_i > 0$ for all i, and $f \in \mathcal{F}_k(\boldsymbol{\alpha})$. In addition, let N_x be a random integer uniformly distributed in $\mathbb{Z} \cap [1, x]$, let $\boldsymbol{D}_{f,x}$ be a random vector satisfying (1.7), and let $\boldsymbol{Z}$ be a Δ^{k-1}-valued random variable distributed according to $\mathrm{Dir}(\boldsymbol{\alpha})$. We also consider the random vector $\boldsymbol{\delta}_{f,x} \in \Delta^{k-1}$ defined by

$$(9.1) \qquad \boldsymbol{\delta}_{f,x} := \Big(\frac{\log D_{f,x,1}}{\log N_x}, \dots, \frac{\log D_{f,x,k}}{\log N_x} \Big)$$

when $N_x \geqslant 2$, and $\boldsymbol{\delta}_{f,x} = (1, 0, \dots, 0)$ when $N_x = 1$. We note that Theorem 2 is about comparing the cumulative distribution function of $\boldsymbol{\delta}_{f,x}$ with the one for $\boldsymbol{Z}$. To achieve this, we construct an appropriate coupling between N_x, $\boldsymbol{D}_{f,x}$ and $\boldsymbol{Z}$. This will build on the coupling given in Section 2.

Lemma 9.1. — *Let $x \geqslant 2$ and assume the above notation. There exists a coupling of N_x, $\boldsymbol{V}$, $\boldsymbol{D}_{f,x}$ and $\boldsymbol{Z}$ such that $\mathbb{P}[\mathcal{E}^c] \ll 1/\log x$ with*

$$\mathcal{E} := \left\{ \|\boldsymbol{\delta}_{f,x} - \boldsymbol{Z}\|_\infty \leqslant \frac{2 \cdot \log(x/N_x) + 3 \cdot \log s(N_x) + 2 \cdot \Theta_x}{\log x} \right\},$$

with $\boldsymbol{\delta}_{f,x}$ defined as in (9.1), Θ_x defined as in (2.5), and $\|\cdot\|_\infty$ being the supremum norm in $\mathbb{R}^k$.

Proof. — Our starting point is the coupling of N_x and $\boldsymbol{V}$ described in Section 2, and it remains to construct $\boldsymbol{D}_{f,x}$ and $\boldsymbol{Z}$. To do this, we need to equip this space with the following additional random variables that are all independent of each other and of N_x and $\boldsymbol{V}$, and whose role will become apparent later:

– a sequence $(C'_i)_{i \geqslant 1}$ of i.i.d. random variables such that $\mathbb{P}[C'_i = \ell] = \alpha_\ell$ for all $i \geqslant 1$ and all $\ell = 1, 2, \ldots, k$;

– for each natural number n and each prime p, we have independent random vectors $\boldsymbol{X}_p(n) := \big(X_{p,1}(n), \ldots, X_{p,k}(n)\big)$ with

$$\mathbb{P}\big[X_{p,i}(n) = e_i \ \forall i \leqslant k\big] = f(p^{e_1}, \ldots, p^{e_k}),$$

for every non-negative integer solutions to $e_1 + \cdots + e_k = \nu_p(n)$.

To construct $\boldsymbol{D}_{f,x}$, we may take

$$D_{f,x,i} := \prod_p p^{X_{p,i}(N_x)} \quad \text{for } i = 1, 2, \ldots, k.$$

Since $f \in \mathcal{F}_k(\boldsymbol{\alpha})$ satisfies property (c) of Definition 1.2, we may easily check that the distribution of $\boldsymbol{D}_{f,x} = (D_{f,x,1}, \ldots, D_{f,x,k})$ is indeed in accordance with (1.7).

Next, we construct $\boldsymbol{Z}$. Firstly, we need to construct a new sequence of random variables $(C_i)_{i \geqslant 1}$, whose definition depends deterministically on the previously introduced random variables. The following events must occur:

– We have $N_x = n$ for some choice of $n \in \mathbb{Z} \cap [1, x]$.

– For all $j \in \{1, \ldots, k\}$ and all $p \mid n^\flat$, we have $X_{p,j}(n) = \mathbb{1}_{j = \ell(p)}$ for some choice of sequence $(\ell(p))_{p \mid n^\flat}$ taking values in $\{1, \ldots, k\}$.

– For all $i \geqslant 1$, we have that $C'_i = \ell'_i$ for some choice sequence $(\ell'_i)_{i \geqslant 1}$ taking values in $\{1, \ldots, k\}$.

Recall the factorization $N_x = P_1 P_2 \cdots$ with the P_i's forming a non-increasing sequence of primes or ones. Then we construct $(C_i)_{i \geqslant 1}$ in the following way:

– If $i \geqslant 1$ is such that $P_i > 1$ and $P_i \mid n^\flat$, then we let $C_i := \ell(P_i)$.

– Otherwise, we let $C_i := \ell'_i$.

Since $f \in \mathcal{F}_k(\boldsymbol{\alpha})$ satisfies property (b) of Definition 1.2, a straightforward computation reveals that for any $n \in \mathbb{Z} \cap [1, x]$, any $(\ell_i)_{i \in I}$ taking values in $\{1, \ldots, k\}$ and any finite set $I \subset \mathbb{N}$, we have

$$\mathbb{P}\big[C_i = \ell_i \ \forall i \in I \mid N_x = n\big] = \prod_{j=1}^k \alpha_j^{\#\{i \in I : \ell_i = j\}},$$

implying that the random variables $(C_i)_{i \geqslant 1}$ are independent of each other, of N_x and of $\boldsymbol{V}$, and they satisfy

$$\mathbb{P}[C_i = \ell] = \alpha_\ell \quad \text{for } i = 1, 2, \dots \text{ and } \ell = 1, 2, \dots, k.$$

Finally, equipped with these random variables and motivated by the proof of Donnelly and Tavaré in Section 8, we take

$$\boldsymbol{Z} := \Big(\sum_{i : C_i = 1} V_1, \dots, \sum_{i : C_i = k} V_k \Big),$$

which is distributed according to $\mathrm{Dir}(\boldsymbol{\alpha})$ by the discussion in Section 8.

Recall the definition of $\boldsymbol{\delta}_{f,x}$ in (9.1). We introduce the auxiliary random variables

(9.2)
$$\boldsymbol{\delta}_{f,x}^* := \Big(\frac{\log D_{f,x,1}}{\log x}, \dots, \frac{\log D_{f,x,k}}{\log x} \Big),$$
$$\boldsymbol{\rho}_{f,x} := \Big(\sum_{i : C_i = 1} \frac{\log P_i}{\log x}, \dots, \sum_{i : C_i = k} \frac{\log P_i}{\log x} \Big).$$

We claim that

(9.3)
$$\big\| \boldsymbol{\delta}_{f,x} - \boldsymbol{\delta}_{f,x}^* \big\|_\infty \leqslant \frac{\log(x/N_x)}{\log x}.$$

Indeed, if $N_x = 1$, both the left-hand and right-hand sides of the inequality are always equal to 1. For $N_x \geqslant 2$, we note that

$$\big\| \boldsymbol{\delta}_{f,x} - \boldsymbol{\delta}_{f,x}^* \big\|_\infty = \frac{\log(x/N_x)}{\log x} \cdot \max_{1 \leqslant i \leqslant k} \frac{\log D_{f,x,i}}{\log N_x},$$

from which (9.3) follows.

Furthermore, we also assert that

(9.4)
$$\big\| \boldsymbol{\delta}_{f,x}^* - \boldsymbol{\rho}_{f,x} \big\|_\infty \leqslant \frac{\log s(N_x)}{\log x}.$$

To verify this, suppose $N_x = n$. Then we have

$$\begin{aligned}
\log D_{x,\ell} &= \sum_{p | n^\flat} X_{p,\ell}(n) \log p + \sum_{p | s(n)} X_{p,\ell}(n) \log p \\
&= \sum_{i \geqslant 1 : P_i | n^\flat} X_{P_i,\ell}(n) \log P_i + \sum_{p | s(n)} X_{p,\ell}(n) \log p \\
&= \sum_{i \geqslant 1 : P_i | n^\flat} \mathbb{1}_{C_i = \ell} \log P_i + \sum_{p | s(n)} X_{p,\ell}(n) \log p,
\end{aligned}$$

whence (9.4) follows readily.

Now if the event $\mathcal{E}$ does not happen, then

$$\big\| \boldsymbol{\delta}_{f,x} - \boldsymbol{Z} \big\|_\infty > \frac{2 \cdot \log(x/N_x) + 3 \cdot \log s(N_x) + 2 \cdot \Theta_x}{\log x}.$$

Together with (9.3) and (9.4), this implies that

$$\big\| \boldsymbol{\rho}_{f,x} - \boldsymbol{Z} \big\|_\infty > \frac{\log(x/N_x) + 2 \cdot \log s(N_x) + 2 \cdot \Theta_x}{\log x}.$$

By Lemma 2.1, we must then have that $M_x \neq N_x$. In conclusion, we have proved that $\mathcal{E}^c \subseteq \{M_x \neq N_x\}$. Thus, Proposition 2.4 shows that $\mathbb{P}[\mathcal{E}^c] \ll 1/\log x$, as required. $\square$

In order to make use of Lemma 9.1, we need to introduce some further notation. Let
$$R_{\mathrm{NT}} := 3 \cdot \frac{\log(x/N_x^\flat)}{\log x} \quad \text{and} \quad R_{\mathrm{PD}} := 2 \cdot \frac{\Theta_x}{\log x}.$$
We write $B(\boldsymbol{x}, r)$ for the closed ball of radius $r \geqslant 0$ centered at $\boldsymbol{x}$ in the normed vector space $(\mathbb{R}^k, \|\cdot\|_\infty)$. We recall that Δ^{k-1} is the standard $(k-1)$-dimensional simplex. For any subset $A \subseteq \Delta^{k-1}$, we write ∂A for the boundary of A in the relative topology of Δ^{k-1}.

In addition, we define the following events:

- $\mathcal{E}_{\mathrm{NT}}(A) := \{B(\boldsymbol{\delta}_{f,x}^*, R_{\mathrm{NT}}) \cap \partial A = \varnothing\}$;
- $\mathcal{E}_{\mathrm{PD}}(A) := \{B(\boldsymbol{Z}, R_{\mathrm{PD}}) \cap \partial A = \varnothing\}$.

LEMMA 9.2. — *For any measurable set $A \subseteq \Delta^{k-1}$, we have*
$$\mathbb{P}[\boldsymbol{\delta}_{f,x} \in A] = \mathbb{P}[\boldsymbol{Z} \in A] + O\Big(\mathbb{P}[\mathcal{E}_{\mathrm{NT}}(A)^c] + \mathbb{P}[\mathcal{E}_{\mathrm{PD}}(A)^c] + \frac{1}{\log x}\Big).$$

Proof. — Let $\mathcal{E}$ be as in Lemma 9.1. Hence, it suffices to show that

$$(9.5) \qquad \Big|\mathbb{P}[\mathcal{E} \cap \{\boldsymbol{\delta}_{f,x} \in A\}] - \mathbb{P}[\mathcal{E} \cap \{\boldsymbol{Z} \in A\}]\Big| \leqslant \mathbb{P}[\mathcal{E}_{\mathrm{NT}}(A)^c] + \mathbb{P}[\mathcal{E}_{\mathrm{PD}}(A)^c].$$

We shall prove the following stronger statement: if $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, then $\boldsymbol{\delta}_{f,x} \in A$ if, and only if, $\boldsymbol{Z} \in A$.

Indeed, assume that $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, and we also have $\boldsymbol{\delta}_{f,x} \in A$. Aiming for a contradiction, we also assume that $\boldsymbol{Z} \notin A$.

Let
$$R_{\mathrm{NT}}' := (\log x)^{-1}(2 \cdot \log(x/N_x) + 3 \cdot \log s(N_x)).$$
Note that $B(\boldsymbol{\delta}_{f,x}, R_{\mathrm{NT}}') \subseteq B(\boldsymbol{\delta}_{f,x}^*, R_{\mathrm{NT}})$ by (9.3). By assumption of $\mathcal{E}_{\mathrm{NT}}(A)$, we know that $B(\boldsymbol{\delta}_{f,x}, R_{\mathrm{NT}}') \cap \partial A = \varnothing$. Since $\boldsymbol{\delta}_{f,x} \in A$ and $B(\boldsymbol{\delta}_{f,x}, R_{\mathrm{NT}}') \cap \Delta^{k-1}$ is connected (being the intersection of two convex sets), we have that
$$B(\boldsymbol{\delta}_{f,x}, R_{\mathrm{NT}}') \cap \Delta^{k-1} \subseteq A.$$
And by assumption of $\mathcal{E}_{\mathrm{PD}}(A)$, we have $\boldsymbol{Z} \in \Delta^{k-1} \smallsetminus A$ and $B(\boldsymbol{Z}, R_{\mathrm{PD}}) \cap \partial A = \varnothing$, which implies
$$B(\boldsymbol{Z}, R_{\mathrm{PD}}) \cap \Delta^{k-1} \subseteq \Delta^{k-1} \smallsetminus A.$$
In particular, we find that $B(\boldsymbol{\delta}_{f,x}, R_{\mathrm{NT}}') \cap B(\boldsymbol{Z}, R_{\mathrm{PD}}) \cap \Delta^{k-1} = \varnothing$. On the other hand, we know that $\|\boldsymbol{\delta}_{f,x} - \boldsymbol{Z}\|_\infty \leqslant R_{\mathrm{NT}}' + R_{\mathrm{PD}}$ when $\mathcal{E}$ occurs by Lemma 9.1. In particular, there exists a point $\boldsymbol{y}$ on the line segment connecting $\boldsymbol{\delta}_{f,x}$ and $\boldsymbol{Z}$ such that $\|\boldsymbol{\delta}_{f,x} - \boldsymbol{y}\|_\infty \leqslant R_{\mathrm{NT}}'$ and $\|\boldsymbol{y} - \boldsymbol{Z}\|_\infty \leqslant R_{\mathrm{PD}}$. Since the sets Δ^{k-1}, $B(\boldsymbol{\delta}_{f,x}, R_{\mathrm{NT}}')$ and $B(\boldsymbol{Z}, R_{\mathrm{PD}})$ are convex, we conclude that $\boldsymbol{y}$ lies in their intersection. But we had seen before that this intersection is the empty set. We have thus arrived at a contradiction. This completes the proof that if $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, and we also know that $\boldsymbol{\delta}_{f,x} \in A$, then we must also have that $\boldsymbol{Z} \in A$.

Conversely, we may show by a simple variation of the above argument that if $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, and we also know that $\boldsymbol{Z} \in A$, then we must also have that $\boldsymbol{\delta}_{f,x} \in A$. This completes the proof of the lemma. $\qquad\square$

To prove Theorem 2, we need to bound $\mathbb{P}[\mathcal{E}_{\mathrm{NT}}(A)^c]$ and $\mathbb{P}[\mathcal{E}_{\mathrm{PD}}(A)^c]$ when A equals the set
$$\Delta_{\mathbf{u}}^{k-1} := \{\boldsymbol{x} \in \Delta^{k-1} : x_i \leqslant u_i \ \forall i < k\}.$$
with $\boldsymbol{u} \in (0,1]^{k-1}$. Note that
$$\partial\big(\Delta_{\mathbf{u}}^{k-1}\big) \subseteq \{\boldsymbol{x} \in \Delta^{k-1} : \exists i < k \text{ such that } x_i = u_i < 1\}$$
(here it is important that the boundary of $\Delta_{\mathbf{u}}^{k-1}$ is defined with respect to the topology of Δ^{k-1}). Therefore,
$$\mathcal{E}_{\mathrm{NT}}(\Delta_{\mathbf{u}}^{k-1})^c \subseteq \bigcup_{\substack{i<k\\u_i\neq 1}} \mathcal{B}_{\mathrm{NT}}^{(i)}(u_i), \quad \text{where} \quad \mathcal{B}_{\mathrm{NT}}^{(i)}(u) := \Big\{\big|\log(D_{f,x,i}/x^u)\big| \leqslant 3\log(x/N_x^\flat)\Big\},$$
as well as
$$\mathcal{E}_{\mathrm{PD}}(\Delta_{\mathbf{u}}^{k-1})^c \subseteq \bigcup_{\substack{i<k\\u_i\neq 1}} \mathcal{B}_{\mathrm{PD}}^{(i)}(u_i), \quad \text{where} \quad \mathcal{B}_{\mathrm{PD}}^{(i)}(u) := \Big\{\big|Z_i - u\big| \leqslant 2\Theta_x/\log x\Big\},$$

We then have the following two crucial estimates:

LEMMA 9.3. — *For $i \in \{1,2,\ldots,k-1\}$ and $u \in (0,1)$, we have the uniform estimate*
$$\mathbb{P}\big[\mathcal{B}_{\mathrm{NT}}^{(i)}(u)\big] \ll \frac{1}{(1+u\log x)^{1-\alpha_i}(1+(1-u)\log x)^{\alpha_i}}.$$

LEMMA 9.4. — *For $i \in \{1,2,\ldots,k-1\}$ and $u \in (0,1)$, we have*
$$\mathbb{P}\big[\mathcal{B}_{\mathrm{PD}}^{(i)}(u)\big] \ll \frac{1}{(1+u\log x)^{1-\alpha_i}(1+(1-u)\log x)^{\alpha_i}}.$$

Using Lemmas 9.2–9.4 together with Proposition 8.1 yields immediately Theorem 2. Thus, it remains to prove Lemmas 9.3 and 9.4, which we do in the next two sections.

REMARKS

(a) The proofs of Lemmas 9.3 and 9.4 are rather involved. However, it is possible to obtain slightly weaker versions of them in a rather easy manner, which will then lead to a correspondingly weaker version of Theorem 2.

First, we prove a variation of Lemma 9.2. Let $\mathcal{E}^*(A)$ be the event $B(\boldsymbol{Z}, R^*) \cap \partial A = \varnothing$ with
$$R^* := \frac{2\cdot\log(x/N_x) + 3\cdot\log s(N_x) + 2\cdot\Theta_x}{\log x}.$$
A straightforward modification of the proof of Lemma 9.2 implies
$$(9.6) \qquad \mathbb{P}[\boldsymbol{D}_{f,x} \in A] = \mathbb{P}[\boldsymbol{Z} \in A] + O\big(\mathbb{P}[\mathcal{E}^*(A)^c] + 1/\log x\big)$$

for any measurable set A. With Chernoff's bound and Lemmas 2.2–2.3, the probability that $R^* > \log_2 x/\log x$ is $\ll 1/\log x$. Moreover, for any $\delta \in (0, 1/4]$, we can show by a direct computation with the Dirichlet distribution the uniform bound

$$\mathbb{P}\Big[\exists j < k \text{ such that } |Z_j - u_j| \leqslant \delta\Big] \ll \sum_{1 \leqslant j < k} \frac{\delta}{(u_j + \delta)^{1-\alpha_j}(1 - u_j + \delta)^{\alpha_j}}$$

for all $\boldsymbol{u} \in [0,1]^{k-1}$ (see Lemma 11.1 below for a proof of this claim). Taking $\delta = \log_2 x/\log x$ proves that

$$\mathbb{P}\big[D_{f,x,i} \leqslant N_x^{u_i} \ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\boldsymbol{u})$$
$$+ O\bigg(\sum_{\substack{1 \leqslant i < k \\ u_i \neq 1}} \frac{1}{(1 + u_i \log x/\log_2 x)^{1-\alpha_i}(1 + (1 - u_i)\log x/\log_2 x)^{\alpha_i}} \bigg).$$

(b) As a matter of fact, the above proof could have worked with Arratia's original coupling from [2]. Under this coupling, we have $\mathbb{P}[M_x \neq N_x] \ll \log_2 x/\log x$ (see Remark (b) at the end of Section 2), so that Lemma 9.1 would hold with part (a) replaced by the weaker bound $\mathbb{P}[\mathcal{E}^c] \ll \log_2 x/\log x$, and thus (9.6) would hold with $\log_2 x/\log x$ in place of $1/\log x$.

10. Proof of Lemma 9.3 using elementary number-theoretic techniques

Note that

$$(10.1) \qquad \mathbb{P}\big[N_x^\flat \leqslant x/(\log x)^3\big] \leqslant \frac{1}{\log x}\mathbb{E}\big[(x/N_x^\flat)^{1/3}\big] \ll \frac{1}{\log x}$$

by Lemma 2.2 applied with $\alpha = \beta = 1/3$.

Fix now $i \in \{1, \ldots, k-1\}$ and $u \in (0, 1)$. For simplicity, let us write

$$\alpha := \alpha_i$$

for the remainder of this section. Given z, let us define

$$S(z) := \sum_{\substack{n \leqslant x \\ n^\flat \in (x/z, 2x/z]}} \sum_{\substack{d_1 \cdots d_k = n \\ d_i \in [x^u z^{-3}, x^u z^3]}} f(d_1, \ldots, d_k).$$

Using (10.1) and (1.7), we find that

$$(10.2) \qquad \mathbb{P}\big[\mathcal{B}_{\mathrm{NT}}^{(i)}(u)\big] \leqslant \frac{1}{\lfloor x \rfloor} \sum_{1 \leqslant m \leqslant 5 \log_2 x} S(2^m) + O\big(1/\log x\big).$$

Hence Lemma 9.3 follows readily from the estimate below:

Lemma 10.1. — *There exists a universal constant $x_0 \geqslant 2$ such that if $x \geqslant x_0$ and $z \in [2, (\log x)^4]$, then we have the uniform estimate*

$$S(z) \ll \frac{\log z}{\sqrt{z}} \cdot \frac{x}{(1 + u \log x)^{1-\alpha}(1 + (1 - u)\log x)^{\alpha}}.$$

Proof. — If $d_1 \cdots d_k = n$, then we may uniquely write $d_j = d_j' d_j''$ with $d_j' | n^\flat$ and $d_j'' | s(n)$. In particular, $d_j'' \leqslant s(n) \leqslant x/n^\flat \leqslant z$ for all j. Hence, if $d_i \in [x^u z^{-3}, x^u z^3]$, we must also have that $d_i' \in [x^u z^{-4}, x^u z^3]$. Hence, using property (c) of Definition 1.2, we deduce that

$$\sum_{\substack{d_1 \cdots d_k = n \\ d_i \in [x^u z^{-3}, x^u z^3]}} f(d_1, \ldots, d_k) \leqslant \sum_{\substack{d_1' \cdots d_k' = n^\flat \\ d_i' \in [x^u z^{-4}, x^u z^3]}} f(d_1', \ldots, d_k') \sum_{d_1'' \cdots d_k'' = s(n)} f(d_1'', \ldots, d_k'').$$

Relation (1.6) implies that $f(d_1', \ldots, d_k') = \prod_{j=1}^{k} \alpha_j^{\omega(d_j')}$. Moreover, using property (a) of Definition 1.2, we find that the sum over the d_j'''s equals 1. In conclusion,

$$\sum_{\substack{d_1 \cdots d_k = n \\ d_i \in [x^u z^{-3}, x^u z^3]}} f(d_1, \ldots, d_k) \leqslant \sum_{\substack{d_1' \cdots d_k' = n^\flat \\ d_i' \in [x^u z^{-4}, x^u z^3]}} \prod_{j=1}^{k} \alpha_j^{\omega(d_j')}$$

$$= \sum_{\substack{d | n^\flat \\ d \in [x^u z^{-4}, x^u z^3]}} \alpha^{\omega(d)} (1-\alpha)^{\omega(n^\flat/d)}.$$

Let $n^\flat = dm$ and $s(n) = b$. If $n^\flat \in [x/z, 2x/z]$ and $d \in [x^u z^{-4}, x^u z^3]$, then $m \in [x^{1-u} z^{-4}, 2x^{1-u} z^3]$. Hence, we find that

$$\begin{aligned}
S(z) &\leqslant \sum_{\substack{dm \in [x/z, 2x/z] \\ d \in [x^u z^{-4}, x^u z^3] \\ m \in [x^{1-u} z^{-4}, 2x^{1-u} z^3]}} \alpha^{\omega(d)} (1-\alpha)^{\omega(m)} \sum_{\substack{b \leqslant z \\ b \text{ square-full}}} 1 \\
&\ll \sqrt{z} \sum_{\substack{dm \in [x/z, 2x/z] \\ d \in [x^u z^{-4}, x^u z^3] \\ m \in [x^{1-u} z^{-4}, 2x^{1-u} z^3]}} \alpha^{\omega(d)} (1-\alpha)^{\omega(m)} = \sqrt{z} \cdot (S_1 + S_2),
\end{aligned}$$
(10.3)

where S_1 denotes the double sum over d and m with the additional constraint $d \leqslant \sqrt{2x/z}$, and S_2 is the corresponding sum over pairs (d, m) with $d > \sqrt{2x/z}$ and $m \leqslant \sqrt{2x/z}$.

First, we bound S_1. Note that for the conditions on d to be compatible we must have that $u \leqslant 2/3$, provided that x is large enough. Assuming this is the case, we apply twice Proposition A.3 to find that

$$\begin{aligned}
S_1 &\leqslant \sum_{\substack{d \leqslant \sqrt{2x/z} \\ d \in [x^u z^{-4}, x^u z^3]}} \alpha^{\omega(d)} \sum_{m \leqslant 2x/dz} (1-\alpha)^{\omega(m)} \\
&\ll \sum_{\substack{d \leqslant \sqrt{2x/z} \\ d \in [x^u z^{-4}, x^u z^3]}} \alpha^{\omega(d)} \cdot \frac{x (\log x)^{-\alpha}}{dz} \\
&\ll \frac{x (\log z)(1 + u \log x)^{\alpha-1} (\log x)^{-\alpha}}{z}.
\end{aligned}$$
(10.4)

We bound S_2 in a similar manner. For this sum to be non-empty, we have $u \geqslant 1/3$. If this is the case, then we have

$$S_2 \leqslant \sum_{\substack{m \leqslant \sqrt{2x/z} \\ m \in [x^{1-u}z^{-4}, 2x^{1-u}z^3]}} (1-\alpha)^{\omega(m)} \sum_{d \leqslant 2x/mz} \alpha^{\omega(d)}$$

(10.5)
$$\ll \sum_{\substack{m \leqslant \sqrt{2x/z} \\ m \in [x^{1-u}z^{-4}, 2x^{1-u}z^3]}} (1-\alpha)^{\omega(m)} \cdot \frac{x(\log x)^{\alpha-1}}{mz}$$

$$\ll \frac{x(\log z)(1 + (1-u)\log x)^{-\alpha}(\log x)^{\alpha-1}}{z}.$$

Combining (10.3), (10.4) and (10.5), while keeping in mind that $S_1 = 0$ if $u > 2/3$ and that $S_2 = 0$ if $u < 1/3$, completes the proof of the lemma, $\square$

11. Proof of Lemma 9.4 using probabilistic techniques

Since $\boldsymbol{Z}$ follows the distribution $\mathrm{Dir}(\boldsymbol{\alpha})$, each component Z_i follows a $\mathrm{Beta}(\alpha_i, 1-\alpha_i)$ distribution, meaning that if $[a,b] \subseteq [0,1]$, then

(11.1)
$$\begin{aligned}
\mathbb{P}\big[Z_i \in [a,b]\big] &= \frac{1}{\Gamma(\alpha_i)\Gamma(1-\alpha_i)} \int_a^b \frac{\mathrm{d}t}{t^{1-\alpha_i}(1-t)^{\alpha_i}} \\
&= \frac{\sin(\pi\alpha_i)}{\pi} \int_a^b \frac{\mathrm{d}t}{t^{1-\alpha_i}(1-t)^{\alpha_i}},
\end{aligned}$$

by Euler's reflection formula. For this reason, we need the following preliminary estimate.

LEMMA 11.1. — *Uniformly for $u \in [0,1]$, $\alpha \in (0,1)$ and $\delta > 0$, we have*

$$\sin(\pi\alpha) \int_{[u-\delta, u+\delta] \cap [0,1]} \frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^{\alpha}} \ll \frac{\delta}{(u+\delta)^{1-\alpha}(1-u+\delta)^{\alpha}}.$$

In particular, if $\delta \geqslant 1/\log x$, then

$$\sin(\pi\alpha) \int_{[u-\delta, u+\delta] \cap [0,1]} \frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^{\alpha}} \ll \frac{\delta \log x}{(1+u\log x)^{1-\alpha}(1+(1-u)\log x)^{\alpha}}.$$

Proof. — Note that both sides of the claimed inequality are invariant under the change of variables $(\alpha, u) \mapsto (1-\alpha, 1-u)$. Hence, we may assume without loss of generality that $u \in [0, 1/2]$. In addition, observe that

$$\sin(\pi\alpha) \int_0^1 \frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^{\alpha}} = \frac{\sin(\pi\alpha)}{\Gamma(\alpha)\Gamma(1-\alpha)} = \pi.$$

Hence the lemma is trivially true if $\delta > 1/4$.

We have thus reduced the proof to the case when $\alpha \in (0,1)$, $u \in [0, 1/2]$ and $\delta \in [0, 1/4]$. In particular, $u + \delta \leqslant 3/4$, so that $1 - t \in [1/4, 1]$ for all $t \in [0, u+\delta]$.

It thus suffices to prove that

$$(11.2) \qquad \sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{dt}{t^{1-\alpha}} \ll \frac{\delta}{(u+\delta)^{1-\alpha}}.$$

Assume first that $\delta \leqslant u/2$. We then have $t \geqslant u/2$ whenever $t \in [u-\delta, u+\delta]$. Using also the trivial bound $\sin(\pi\alpha) \leqslant 1$, we conclude that

$$\sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{dt}{t^{1-\alpha}} \leqslant \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{dt}{(u/2)^{1-\alpha}} \ll \frac{\delta}{(u+\delta)^{1-\alpha}}.$$

This proves the lemma in this case.

Finally, assume that $\delta \geqslant u/2$. We then use that

$$\sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{dt}{t^{1-\alpha}} \leqslant \sin(\pi\alpha) \int_0^{3\delta} \frac{dt}{t^{1-\alpha}} = \frac{\sin(\pi\alpha)}{\alpha} \cdot (3\delta)^\alpha$$

$$\ll \delta^\alpha \asymp \frac{\delta}{(u+\delta)^{1-\alpha}}.$$

This completes the proof of the lemma in all cases. $\qquad\qquad \square$

Let us now show Lemma 9.4. For simplicity of notation, let us fix $i \in \{1, 2, \ldots, k\}$ and $u \in (0,1)$, and let us set $\alpha := \alpha_i$ and

$$\Delta := \frac{1}{(1 + u\log x)^{1-\alpha}(1 + (1-u)\log x)^\alpha}.$$

Thus, our goal is to show that

$$(11.3) \qquad \mathbb{P}\Big[\big|Z_i - u\big| \leqslant 2\Theta_x/\log x\Big] \ll \Delta.$$

Recall that

$$\Theta_x = \sum_{j\geqslant 1} r(V_j \log x),$$

and that there exists an absolute constant $c \geqslant 1$ such that $r(y) \leqslant c\min\{y, y^{-2}\}$ for all $y > 0$ (see (2.4)). Since there are at most 10 indices j such that $V_j \geqslant 0.1$, we find

$$\Theta_x \leqslant \Theta_x' + 10c, \quad \text{where} \quad \Theta_x' := \sum_{j\geqslant 1:\ V_j < 0.1} r(V_j \log x).$$

Now, using Lemma 11.1 and relation (11.1), we have

$$\mathbb{P}\Big[\big|Z_i - u\big| \leqslant 200c/\log x\Big] \ll \Delta.$$

On the other hand, if $\frac{200c}{\log x} < |Z_i - u| \leqslant \frac{2\Theta_x}{\log x} \leqslant \frac{20c+2\Theta_x'}{\log x}$, then we must have $\Theta_x' > 90c$. In particular, there exists $m \in \mathbb{Z}$ such that $2^m > 40c$ and $\Theta_x' \in (2^m, 2^{m+1}]$, whence $|Z_i - u| \leqslant \frac{20c + 2^{m+2}}{\log x}$. We thus conclude that

$$\mathbb{P}\Big[\big|Z_i - u\big| \leqslant \frac{20c + 2\Theta_x'}{\log x}\Big] \leqslant \sum_{m\in\mathbb{Z}:\ 2^m > 40c} \mathbb{P}\Big[\big|Z_i - u_i\big| \leqslant \frac{20c + 2^{m+2}}{\log x},\ \Theta_x > 2^m\Big] + O(\Delta).$$

Therefore Lemma 9.4 will follow if we can prove that

$$(11.4) \qquad \mathbb{P}\Big[\big|Z_i - u\big| \leqslant \frac{20c + 4\kappa}{\log x},\ \Theta_x' > \kappa\Big] \ll \frac{\Delta}{\kappa} \qquad \text{for all } \kappa \geqslant 40c.$$

We shall make a further reduction. If we set

$$G(\lambda) := \#\big\{ j \geqslant 1 : V_j \log x \in (\lambda, 2\lambda] \big\},$$

then we have

$$\Theta_x' = \sum_{j \geqslant 1:\, V_j < 0.1} r(V_j \log x) \leqslant c \sum_{j \geqslant 1:\, V_j < 0.1} \min \big\{ V_i \log x, (V_i \log x)^{-2} \big\}$$

$$\leqslant c \sum_{\substack{m \geqslant 0 \\ 2^m \leqslant 0.1 \log x}} \frac{G(2^m)}{4^m} + c \sum_{m < 0} 2^{m+1} G(2^m)$$

$$\leqslant 5c \max_{\substack{m \geqslant 0 \\ 2^m \leqslant 0.1 \log x}} \Big(\frac{G(2^m)}{2^{3m/2}} \Big) + 5c \max_{m < 0} \big(2^{m/2} G(2^m) \big),$$

since $\sum_{m \geqslant 0} 2^{-m/2} < 5$ and $\sum_{m < 0} 2^{1+m/2} < 5$. We thus find

$$\mathbb{P}\Big[|Z_i - u| \leqslant \frac{20c + 4\kappa}{\log x},\ \Theta_x' > \kappa \Big]$$

$$\leqslant \sum_{\substack{m \geqslant 0 \\ 2^m \leqslant 0.1 \log x}} \mathbb{P}\Big[|Z_i - u| \leqslant \frac{20c + 4\kappa}{\log x},\ G(2^m) > \frac{2^{3m/2}\kappa}{10c} \Big]$$

$$+ \sum_{m < 0} \mathbb{P}\Big[|Z_i - u| \leqslant \frac{20c + 4\kappa}{\log x},\ G(2^m) > \frac{\kappa 2^{-m/2}}{10c} \Big].$$

Note that in both sums, we have $G(2^m) \geqslant 4$, since $\kappa \geqslant 40c$. Hence, Markov's inequality implies

$$\mathbb{P}\Big[|Z_i - u| \leqslant \frac{20c + 4\kappa}{\log x},\ \Theta_x' > \kappa \Big]$$

$$\leqslant \sum_{\substack{m \geqslant 0 \\ 2^m \leqslant 0.1 \log x}} \frac{100c^2}{\kappa^2 2^{3m}} \mathbb{E}\Big[\mathbb{1}_{|Z_i - u| \leqslant \frac{20c+4\kappa}{\log x}} \cdot \mathbb{1}_{G(2^m) \geqslant 4} \cdot G(2^m)^2 \Big]$$

$$+ \sum_{m < 0} \frac{100c^2 2^m}{\kappa^2} \mathbb{E}\Big[\mathbb{1}_{|Z_i - u| \leqslant \frac{20c+4\kappa}{\log x}} \cdot \mathbb{1}_{G(2^m) \geqslant 4} \cdot G(2^m)^2 \Big].$$

This reduces (11.4), and thus Lemma 9.4, to proving the following estimate:

Lemma 11.2. — *Uniformly for $\mu \geqslant 1$ and $\lambda \in (0, 0.1 \log x]$, we have*

$$\mathbb{E}\Big[\mathbb{1}_{|Z_i - u| \leqslant \mu/\log x} \cdot \mathbb{1}_{G(\lambda) \geqslant 4} \cdot G(\lambda)^2 \Big] \ll (\lambda + \mu) \cdot \Delta.$$

Proof. — Let us call E the quantity we seek to bound from above. Consider two independent Poisson processes $(A_j)_{j \geqslant 1}$ with intensity $\alpha (e^{-x}/x)\,\mathrm{d}x$ and $(B_j)_{j \geqslant 1}$ with intensity $(1-\alpha)(e^{-x}/x)\,\mathrm{d}x$. Hence, the union of the two processes is a new Poisson process of intensity $(e^{-x}/x)\,\mathrm{d}x$. The sum $S_A := \sum_{j \geqslant 1} A_j$ has distribution $\mathrm{Gamma}(\alpha, 1)$ and

the sum $S_B := \sum_{j \geqslant 1} B_j$ has distribution Gamma$(1-\alpha, 1)$. If we also set $S := S_A + S_B$, then, as we saw in the proof of Proposition 8.1, we have

$$E = \mathbb{E}\Big[\mathbb{1}_{|S_A/S - u| \leqslant \mu/\log x} \cdot \mathbb{1}_{G_A(\lambda) + G_B(\lambda) \geqslant 4} \cdot \big(G_A(\lambda) + G_B(\lambda)\big)^2\Big],$$

where $G_A(\lambda) = \sum_{j \geqslant 1} \mathbb{1}_{A_j \log x \in (\lambda S, 2\lambda S]}$ and $G_B(\lambda) = \sum_{j \geqslant 1} \mathbb{1}_{B_j \log x \in (\lambda S, 2\lambda S]}$. We know $G_A(\lambda) + G_B(\lambda) \geqslant 4$. Hence, if $G_A(\lambda) \geqslant G_B(\lambda)$, then we must have $G_A(\lambda) \geqslant 2$ and $G_A(\lambda) \leqslant G_A(\lambda)^2/2$, whence $G_A(\lambda) \leqslant 2 \sum \sum_{k > j \geqslant 1} \mathbb{1}_{A_k \log x, \, A_j \log x \in (\lambda S, 2\lambda S]}$. So, we find

$$\big(G_A(\lambda) + G_B(\lambda)\big)^2 \leqslant 4G_A(\lambda)^2 = 4G_A(\lambda) + 8\sum_{k > j \geqslant 1}\sum \mathbb{1}_{A_k \log x, \, A_j \log x \in (\lambda S, 2\lambda S]}$$

$$\leqslant 16 \sum_{k > j \geqslant 1}\sum \mathbb{1}_{A_k \log x, \, A_j \log x \in (\lambda S, 2\lambda S]}.$$

The analogous inequality also holds when $G_B(\lambda) \geqslant G_A(\lambda)$, with the roles of A and B reversed. We conclude that

$$E \leqslant 16(E_A + E_B),$$

where

$$E_A := \mathbb{E}\Big[\mathbb{1}_{|S_A/S - u| \leqslant \mu/\log x} \sum \sum_{k > j \geqslant 1} \mathbb{1}_{A_k \log x, \, A_j \log x \in (\lambda S, 2\lambda S]}\Big]$$

and E_B is defined analogously. Using Mecke's equation (cf. Proposition B.6), we find

$$E_A = \alpha^2 \iint_{2a_1 > a_2 > a_1 > 0} \mathbb{E}\Big[\mathbb{1}_{|\frac{a_1 + a_2 + S_A}{a_1 + a_2 + S} - u| \leqslant \frac{\mu}{\log x}} \prod_{j=1}^2 \mathbb{1}_{a_j \log x \in (\lambda(a_1 + a_2 + S), 2\lambda(a_1 + a_2 + S))]}\Big]$$
$$\times \frac{e^{-a_1 - a_2}}{a_1 a_2} \, da_1 \, da_2$$

$$= \frac{\alpha^2}{\Gamma(\alpha)\Gamma(1-\alpha)} \iiiint_{\substack{2a_1 > a_2 > a_1 > 0, \ s_1, s_2 > 0 \\ |\frac{a_1 + a_2 + s_1}{a_1 + a_2 + s_1 + s_2} - u| \leqslant \frac{\mu}{\log x} \\ a_j \log x \in (\lambda(a_1 + a_2 + s_1 + s_2), 2\lambda(a_1 + a_2 + s_1 + s_2))] \ (j=1,2)}} \frac{e^{-a_1 - a_2 - s_1 - s_2}}{a_1 a_2 s_1^{1-\alpha} s_2^\alpha} \, da_1 \, da_2 \, ds_1 \, ds_2,$$

where we used that the a_j's lie in the same dyadic interval to deduce that $a_2 < 2a_1$. We make the change of variables $t = s_1/(s_1 + s_2)$ and $s = s_1 + s_2$. Since $\lambda/\log x \leqslant 0.1$ and $a_1 < a_2 < 2a_1$, the conditions $a_j \log x \in (\lambda(a_1 + a_2 + s), 2\lambda(a_1 + a_2 + s)]$ for $j = 1, 2$ imply that $a_j \log x \in (\lambda s, 5\lambda s]$. Knowing also that $|\frac{a_1 + a_2 + s_1}{a_1 + a_2 + s} - u| \leqslant \frac{\mu}{\log x}$, we find $|t - u| = |\frac{s_1}{s_1 + s_2} - u| \leqslant \frac{10\lambda + \mu}{\log x}$. Finally, we use Euler's reflection formula to write $\Gamma(\alpha)\Gamma(1-\alpha) = \pi/\sin(\pi\alpha)$. We conclude that

$$E_A \leqslant \frac{\sin(\pi\alpha)}{\pi} \iiiint_{\substack{a_2 > a_1 > 0, \ s > 0, \ t \in [0,1] \\ |t - u| \leqslant (10\lambda + \mu)/\log x \\ a_j \log x \in (\lambda s, 5\lambda s] \ (j=1,2)}} \frac{e^{-a_1 - a_2 - s}}{a_1 a_2 t^{1-\alpha}(1-t)^\alpha} \, da_1 \, da_2 \, ds \, dt.$$

For every fixed value of s, the integral over a_1 and a_2 is $\leqslant (\log 5)^2$, since $\int_w^{5w} \frac{e^{-a}}{a}\, \mathrm{d}a \leqslant \log 5$ for any $w > 0$. We also have $\int_0^\infty e^{-s}\, \mathrm{d}s = 1$. We thus conclude that

$$E_A \leqslant \frac{(\log 5)^2 \sin(\pi\alpha)}{\pi} \int\limits_{\substack{t \in [0,1] \\ |t-u| \leqslant (10\lambda+\mu)/\log x}} \frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^\alpha}.$$

Using Lemma 11.1 shows that $E_A \ll (\lambda + \mu)\Delta$. The same estimate holds for E_B too, thus completing the proof of the lemma. □

Part 3. Appendices

Because this paper lies in the intersection of number theory and probability theory, readers coming from one of these fields might not be familiar with standard results of the other one. For this reason, we gather here some key results from both fields. We present these results in the following two sections.

Appendix A. Tools from Number Theory

Let $\theta(x) := \sum_{p \leqslant x} \log p$ and let $\psi(x) := \sum_{n \leqslant x} \Lambda(n)$, where

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \text{ for some prime power } p^k, \\ 0 & \text{otherwise.} \end{cases}$$

Understanding these functions gives information about the distribution of primes. In 1896, Charles de la Vallée Poussin and Jacques Hadamard proved the prime number theorem, thus establishing that $\psi(x) \sim \theta(x) \sim x$ as $x \to \infty$. De la Vallée Poussin proved a more precise version of these estimates, with the error term $O\big(xe^{-c\sqrt{\log x}}\big)$ for some positive absolute constant c (see, for example, [13, Ch. 8]). For the purposes of the present paper, however, the following weaker version of his result suffices:

Proposition A.1 (The prime number theorem). — *For $x \geqslant 2$, we have*

$$\theta(x) = x + O(x/(\log x)^3)$$

and

$$\psi(x) = x + O(x/(\log x)^3).$$

We shall also need a stronger version of Mertens's estimate, which in its classical form, yields an estimate for the product $\prod_{p \leqslant x}(1 - 1/p)$.

Proposition A.2 (Strong Mertens's estimate). — *For $x \geqslant 2$, we have*

$$\sum_{p^k \leqslant x} \frac{1}{kp^k} = \log_2 x + \gamma + O(1/(\log x)^3).$$

Proof. — Using the prime number theorem (Proposition A.1), there exists a constant c_0 satisfying

$$\sum_{p^k \leqslant x} \frac{1}{kp^k} = \int_{2^-}^x \frac{\mathrm{d}\psi(t)}{t \log t} = \log_2 x + c_0 + O(1/(\log x)^3).$$

for $x \geqslant 2$. For the proof of $\gamma = c_0$, see the proof of [17, Th. 2.7].   □

In the proof of Theorem 2, we need the following estimate:

PROPOSITION A.3. — *Uniformly for $\alpha \in [0,1]$ and $x, y \geqslant 2$, we have*

$$\sum_{n \leqslant x} \alpha^{\omega(n)} \ll x(\log x)^{\alpha-1} \quad and \quad \sum_{x/y < n \leqslant xy} \frac{\alpha^{\omega(n)}}{n} \ll (\log y)(\log(xy))^{\alpha-1}.$$

Proof. — The first bound follows readily with [13, Th. 14.2]. Let us now prove the second one.

When $y \geqslant \sqrt{x}$, we have

$$\sum_{x/y < n \leqslant xy} \frac{\alpha^{\omega(n)}}{n} \leqslant \sum_{\substack{p|n \Rightarrow p \leqslant y^3}} \frac{\alpha^{\omega(n)}}{n} = \prod_{p \leqslant y^3} \left(1 + \frac{\alpha}{p-1}\right) \ll (\log y)^\alpha$$

by the inequality $1 + t \leqslant e^t$ and Mertens's theorem (Proposition A.2). This completes the proof in this case. Finally, assume that $y \leqslant \sqrt{x}$. We then have

$$\sum_{x/y < n \leqslant xy} \frac{\alpha^{\omega(n)}}{n} \leqslant \sum_{\substack{m \in \mathbb{Z} \\ e^m \in [x/y, exy]}} \sum_{n \in (e^{m-1}, e^m]} \frac{\alpha^{\omega(n)}}{n} \leqslant \sum_{\substack{m \in \mathbb{Z} \\ e^m \in [x/y, exy]}} e^{1-m} \sum_{n \leqslant e^m} \alpha^{\omega(n)}.$$

For every m as above, the innermost sum is $\ll e^m m^{\alpha-1} \asymp e^m (\log(xy))^{\alpha-1}$ by the first part of the lemma and our assumption that $y \leqslant \sqrt{x}$. Since there are $\leqslant 1 + \log y \ll \log y$ choices for m, the needed estimate follows in this last case too.   □

APPENDIX B. TOOLS FROM PROBABILITY THEORY

B.1. THE TOTAL VARIATION DISTANCE. — The *total variation distance* is a metric between two probability distributions. Let μ and ν be two probability measures on $\mathbb{C}$. Then

(B.1) $$\mathrm{d}_{\mathrm{TV}}(\mu, \nu) := \sup |\mu(A) - \nu(A)|,$$

where the supremum is taken over all Lebesgue-measurable subsets of $\mathbb{C}$. For any real number a, let

$$a^+ := \max\{a, 0\} \quad \text{and} \quad a^- := \max\{-a, 0\}.$$

When μ and ν are supported on $\mathbb{Z}_{\geqslant 1}$, here are some alternative definitions of the total variation distance:

LEMMA B.1. — *Let μ and ν be two probability measure supported on $\mathbb{N}$. Then*

$$\mathrm{d}_{\mathrm{TV}}(\mu, \nu) = \sum_{i \geqslant 1} (\mu(i) - \nu(i))^+ = \sum_{i \geqslant 1} (\mu(i) - \nu(i))^-.$$

Proof. — Let $E := \{i \in \mathbb{N} : \mu(i) > \nu(i)\}$ and let $\xi := \mu - \nu$. Note that for any $B \subseteq \mathbb{N}$, we have $\xi(B \cap E) \geqslant 0 \geqslant \xi(B \cap E^c)$. Therefore, for any $A \subseteq \mathbb{N}$, we have

$$\xi(A) = \xi(E) + \xi(A \cap E^c) - \xi(A^c \cap E) \leqslant \xi(E)$$

and

$$\xi(A) = \xi(E^c) + \xi(A \cap E) - \xi(A^c \cap E^c) \geqslant \xi(E^c).$$

Therefore, $\mathrm{d}_{\mathrm{TV}}(\mu, \nu) = \sup_{A \subseteq \mathbb{N}} |\xi(A)| = \max\{\xi(E), -\xi(E^c)\}$. Note that $\xi(E) = \sum_{i \geqslant 1}(\mu(i) - \nu(i))^+$, that $-\xi(E^c) = \sum_{i \geqslant 1}(\mu(i) - \nu(i))^-$ and that

$$\sum_{i \geqslant 1}(\mu(i) - \nu(i))^+ = \sum_{i \geqslant 1}\left[(\mu(i) - \nu(i)) + (\mu(i) - \nu(i))^-\right] = \sum_{i \geqslant 1}(\mu(i) - \nu(i))^-.$$

The lemma follows. $\qquad\square$

The total variation distance will be especially useful in this paper because of the following proposition. In [2, §3.8], Arratia proved that for any two random variables X and Y returning positive integers, we can always construct X' and Y' within the same probability space such that:

  – X' and Y' have the same distribution as X and Y, respectively;

  – Y' is a function of X', U, V where (U, V) is a point uniformly chosen in the unit square independent of X';

  – $\mathbb{P}[X' \neq Y'] = \mathrm{d}_{\mathrm{TV}}(X, Y)$.

We repeat his proof here. Let μ and ν be two probability measures supported on $\mathbb{N}$. Let $z_j := \sum_{i \leqslant j} \frac{(\mu(i) - \nu(i))^-}{\mathrm{d}_{\mathrm{TV}}(\mu, \nu)}$ (with $z_0 := 0$). We consider the function $f_{\mu,\nu} : \mathbb{N} \times (0,1)^2 \to \mathbb{N}$ defined as

$$f_{\mu,\nu}(m; a, b) := \begin{cases} m & \text{if } a \cdot \mu(m) \leqslant \nu(m), \\ \sum_{i \geqslant 1} i \cdot \mathbb{1}_{z_{i-1} < b \leqslant z_i} & \text{otherwise.} \end{cases}$$

This is the function used in Section 2 for the extraction of N_x. Note that if $a, b \in (0, 1)$ and $m \in \mathbb{N}$, then we have the equivalency

(B.2) $$f_{\mu,\nu}(m; a, b) \neq m \quad \Longleftrightarrow \quad a \cdot \mu(m) > \nu(m).$$

Indeed, the direction "$\Rightarrow$" is obvious. To see the converse direction, note that if $a \cdot \mu(m) > \nu(m)$, then $(\mu(m) - \nu(m))^- = 0$, and hence the interval $(z_{m-1}, z_m]$ is empty.

LEMMA B.2 (Arratia, [2]). — *Let μ and ν be two probability measures supported on $\mathbb{N}$, let X be a random variable with law μ, and let U and U' be two uniform random variables in $(0, 1)$ such that X, U, U' are independent. Let $Y := f_{\mu,\nu}(X; U, U')$ with $f_{\mu,\nu}$ defined as above. Then $\mathbb{P}[X \neq Y] = \mathrm{d}_{\mathrm{TV}}(\mu, \nu)$ and $\mathbb{P}[Y \in A] = \nu(A)$.*

Proof. — Using (B.2), we find that

(B.3) $$\{X \neq Y\} = \{U \cdot \mu(X) > \nu(X)\}.$$

Furthermore, we directly compute that

(B.4) $$\mathbb{P}\left[U \cdot \mu(m) > \nu(m), X = m\right] = (\mu(m) - \nu(m))^+.$$

Therefore,

$$\text{(B.5)} \qquad \mathbb{P}[X \neq Y] = \mathbb{P}[U \cdot \mu(X) > \nu(X)] = \sum_{m \geqslant 1} (\mu(m) - \nu(m))^+ = \mathrm{d}_{\mathrm{TV}}(\mu, \nu)$$

with Lemma B.1.

Next, we prove that $\mathbb{P}[Y = n] = \nu(n)$ for any $n \in \mathbb{N}$. Note that $Y = n$ if, and only if, one of two disjoint events happen: either we have $U \cdot \mu(n) \leqslant \nu(n)$ with $X = n$, or we have $U \cdot \mu(X) > \nu(X)$ with $z_{n-1} < U' \leqslant z_n$. Therefore, with (B.4) and (B.5), we have

$$\mathbb{P}[Y = n] = \mathbb{P}\big[U \cdot \mu(n) \leqslant \nu(n), X = n\big] + \mathbb{P}\big[U \cdot \mu(X) > \nu(X)\big] \cdot \mathbb{P}\big[z_{n-1} < U' \leqslant z_n\big]$$
$$= \Big(\mu(n) - \big(\mu(n) - \nu(n)\big)^+\Big) + \mathrm{d}_{\mathrm{TV}}(\mu, \nu) \cdot \frac{\big(\mu(n) - \nu(n)\big)^-}{\mathrm{d}_{\mathrm{TV}}(\mu, \nu)} = \nu(n),$$

where we used (B.3) and (B.5) to show that $\mathbb{P}\big[U \cdot \mu(X) > \nu(X)\big] = \mathrm{d}_{\mathrm{TV}}(\mu, \nu)$. This concludes the proof of the lemma. □

B.2. Poisson Processes. — The following definition and propositions are borrowed from Kingman's book on Poisson processes [11]. Let $(S, \mathcal{S})$ be a measurable space with S being a subset of $\mathbb{R}^d$ for some $d \geqslant 1$. A *Poisson process* on a state space S with mean measure μ is a random countable subset $\Pi \subseteq S$ such that:

– for any disjoint measurable subsets $A_1, \ldots, A_n$ of S, the random variables $\#(\Pi \cap A_1), \ldots, \#(\Pi \cap A_n)$ are independent;
– the random variable $\#(\Pi \cap A)$ is a Poisson random variable of parameter $\mu(A)$ for any $A \subset S$ measurable.

We can see Π as an element of the measurable space $(\Omega_S, \mathcal{F}_S)$ where Ω_S is the set of countable subsets of S and $\mathcal{F}_S$ is the smallest σ-algebra for which the map $\Pi \mapsto \#\{\Pi \cap B\}$ is measurable for all $B \in \mathcal{S}$. If μ has no atoms, meaning no singleton with positive probability, and is σ-finite, meaning that S is a countable union of measurable sets with finite measure, then a Poisson process with mean measure μ always exists (see [11], the existence theorem in section 2.5 for a proof). If μ is absolutely continuous with respect to the Lebesgue measure, then the function $\lambda\colon S \to \mathbb{R}_{\geqslant 0}$ such that $\mu(A) = \int_A \lambda(x)\,\mathrm{d}x$ for all measurable subsets $A \subseteq S$ is called the *intensity* of the Poisson process. Here are a few important propositions about Poisson processes that we will use in the paper and proved in [11]:

Proposition B.3 (Mapping theorem, [11, §2.3]). — *If Π is a Poisson process with mean measure μ on S, and $f : S \to T$ is a measurable function such that $\mu^*(B) := \mu(f^{-1}(B))$ has no atoms, then $f(\Pi)$ is a Poisson process on T with measure μ^*.*

Proposition B.4 (Coloring theorem, [11, §5.1]). — *If Π is a Poisson process with mean measure μ on S, and the points are randomly colored with k colors such that the probability of a point receiving the color i is p_i, and such that the color of a point is independent of different points and of the position of the point. Let Π_i be the*

subset of Π with color i. Then all the Π_i are independent Poisson processes with mean measures $\mu_i = p_i \mu$.

PROPOSITION B.5 (Campbell's theorem, [11, §3.2]). — *Let Π be any Poisson process on S with mean measure μ. Let $f : S \to \mathbb{R}$ be a measurable function. Then*

$$\Sigma = \sum_{X \in \Pi} f(X)$$

is absolutely convergent almost surely if, and only if,

$$\int_S \min\{f, 1\} \, d\mu < \infty.$$

If this condition holds, then

$$\mathbb{E}[e^{s\Sigma}] = \exp\left(\int_S (e^{sf} - 1) \, d\mu\right)$$

for any complex s for which the integral converges. Moreover,

$$\mathbb{E}[\Sigma] = \int_S f \, d\mu$$

if the integral converges. In the case where it converges, we also have

$$\mathrm{Var}[\Sigma] = \int_S f^2 \, d\mu.$$

Many probabilities or expectations that involve Poisson processes in this paper can be reformulated as

$$\mathbb{E} \sum_{X \in \Pi} f(\Pi \smallsetminus \{X\}, X).$$

In these cases, there is a generalization of the formula for $\mathbb{E}[\Sigma]$ in Campbell's theorem, called the Mecke equation, allowing us to compute these objects:

PROPOSITION B.6 (Mecke equation, [14, Th. 4.5]). — *Let Π be a Poisson process on S with a σ-finite mean measure μ, and let $f : \Omega_S \times S^k \to [0, \infty)$ be measurable. Then we have*

$$\mathbb{E} \sum_{\substack{X_1, \ldots, X_k \\ \text{all distinct}}} f(\Pi \smallsetminus \{X_1, \ldots, X_k\}; X_1, \ldots, X_k)$$

$$= \int_S \cdots \int_S \mathbb{E}[f(\Pi; x_1, \ldots, x_k)] \, d\mu(x_1) \cdots d\mu(x_k).$$

REFERENCES

[1] R. ARRATIA – "On the central role of scale invariant Poisson processes on $(0, \infty)$", in *Microsurveys in discrete probability (Princeton, NJ, 1997)*, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 41, American Mathematical Society, Providence, RI, 1998, p. 21–41.

[2] ———, "On the amount of dependence in the prime factorization of a uniform random integer", in *Contemporary combinatorics*, Bolyai Soc. Math. Stud., vol. 10, János Bolyai Math. Soc., Budapest, 2002, p. 29–91.

[3] R. Arratia, A. D. Barbour & S. Tavaré – "A tale of three couplings: Poisson-Dirichlet and GEM approximations for random permutations", *Combin. Probab. Comput.* **15** (2006), no. 1-2, p. 31–62.

[4] G. Bareikis & E. Manstavičius – "On the DDT theorem", *Acta Arith.* **126** (2007), no. 2, p. 155–168.

[5] P. Billingsley – "On the distribution of large prime divisors", *Period. Math. Hungar.* **2** (1972), p. 283–289.

[6] R. de la Bretèche & G. Tenenbaum – "Sur les processus arithmétiques liés aux diviseurs", *Adv. in Appl. Probab.* **48** (2016), no. A, p. 63–76.

[7] J.-M. Deshouillers, F. Dress & G. Tenenbaum – "Lois de répartition des diviseurs. I", *Acta Arith.* **34** (1979), no. 4, p. 273–285 (loose errata).

[8] P. Donnelly & S. Tavaré – "The population genealogy of the infinitely-many neutral alleles model", *J. Math. Biol.* **25** (1987), no. 4, p. 381–391.

[9] S. Feng – *The Poisson-Dirichlet distribution and related topics. Models and asymptotic behaviors*, Probability and its Appl., Springer, Heidelberg, 2010.

[10] N. L. Johnson, S. Kotz & N. Balakrishnan – *Discrete multivariate distributions*, Wiley Series in Probab. and Stat.: Applied Probab. and Stat., John Wiley & Sons, Inc., New York, 1997.

[11] J. F. C. Kingman – *Poisson processes*, Oxford Studies in Probab., vol. 3, The Clarendon Press, Oxford University Press, New York, 1993.

[12] S. Kotz, N. Balakrishnan & N. L. Johnson – *Continuous multivariate distributions. Vol. 1. Models and applications*, second ed., Wiley Series in Probab. and Stat.: Applied Probab. and Stat., Wiley-Interscience, New York, 2000.

[13] D. Koukoulopoulos – *The distribution of prime numbers*, Graduate Studies in Math., vol. 203, American Mathematical Society, Providence, RI, 2019.

[14] G. Last & M. Penrose – *Lectures on the Poisson process*, Institute of Math. Stat. Textbooks, vol. 7, Cambridge University Press, Cambridge, 2018.

[15] S.-K. Leung – "Dirichlet law for factorisation of integers, polynomials and permutations", *Math. Proc. Cambridge Philos. Soc.* **175** (2023), no. 3, p. 649–676.

[16] ———, "Dirichlet law with arbitrary parameters for factorizations of integers", Retrieved on 24 August 2025 from `https://sites.google.com/view/sunkaileung/miscellaneous`, 2025.

[17] H. L. Montgomery & R. C. Vaughan – *Multiplicative number theory. I. Classical theory*, Cambridge Studies in Advanced Math., vol. 97, Cambridge University Press, Cambridge, 2007.

[18] S. Nyandwi & A. Smati – "Distribution laws of pairs of divisors", *Integers* **13** (2013), article no. A13 (13 pages).

[19] G. Tenenbaum – "A rate estimate in Billingsley's theorem for the size distribution of large prime factors", *Q. J. Math.* **51** (2000), no. 3, p. 385–403.

Tony Haddad, Département de mathématiques et de statistique, Université de Montréal,
CP 6128 succ. Centre-Ville, Montréal, QC H3C 3J7, Canada
E-mail : `tony.haddad@umontreal.ca`
Url : `https://haddadt.github.io/`

Dimitris Koukoulopoulos, Département de mathématiques et de statistique, Université de Montréal,
CP 6128 succ. Centre-Ville, Montréal, QC H3C 3J7, Canada
E-mail : `dimitris.koukoulopoulos@umontreal.ca`
Url : `https://dms.umontreal.ca/~koukoulo/`