Benford Behavior Resulting From Stick and Box Fragmentation Processes

Bruce Fang[a,*], Steven J. Miller[a]

[a] *Department of Mathemtics, Williams College, Williamstown, 01267, MA, USA*

Abstract

Benford's law is the statement that in many real-world data set, the probability of having digit d in base B, where $1 \leq d \leq B$, as the first digit is $\log_B((d+1)/d)$. We sometimes refer to this as weak Benford behavior, and we say that a data set exhibits strong Benford behavior in base B if the probability of having significand at most s, where $s \in [1, B)$, is $\log_B(s)$. We examine Benford behaviors in two different probabilistic model: stick and box fragmentation. Building on the joint work of Becker et al. [1] on the single proportion stick fragmentation model, we employ combinatorial identities on multinomial coefficients to reduce the multi-proportion stick fragmentation model to the single proportion model. We provide a necessary and sufficient condition for the lengths of the stick fragments to converge to strong Benford behavior along with a quantification of the discrepancy from uniform distribution on $[0, 1]$ in terms of irrationality exponent. Then we answer a conjecture posed by Betti et al. [5] on the high dimensional box fragmentation model. Using tools from order statistics, we prove that under some conditions, faces of any arbitrary dimension of the box have total volume converging to strong Benford's behavior.

Keywords: Benford's law, high-dimensional fragmentation, equidistribution mod 1, irrationality exponent, multinomial identities, order statistics
2008 MSC: 60A10, 11K06 (primary), 60E10 (secondary)

Contents

*Corresponding author.
Email addresses: fangbaojun2002@gmail.com (Bruce Fang), sjm1@williams.edu or steven.miller.mc96@aya.yale.edu (Steven J. Miller)

1. Introduction

At the beginning of the 20$^{\text{th}}$ century, astronomer and mathematician Simon Newcomb [14] noticed an unusual pattern in the logarithmic tables he used at work. The early pages of the books were far more worn than the later ones, suggesting that numbers starting with smaller digits were consulted more often. From this, Newcomb inferred that people had a tendency to encounter numbers beginning with 1 more frequently. Specifically, he found that 1 appeared as the leading digit about 30% of the time, 2 about 17%, with the frequency decreasing for larger digits. Although he formulated a mathematical explanation for this curious phenomenon, his discovery initially went largely overlooked.

It took another 57 years after Newcomb's discovery for physicist Frank Benford to make the exact same observation as Newcomb: the first pages of logarithmic tables were used far more than others. He formulated this law as follows.

Definition 1. *[4, Page 554] Data exhibits* ***(weak) Benford behavior base*** *B if the frequency F_d of leading digit d is*

$$F_d \;=\; \log_B\left(\frac{d+1}{d}\right). \tag{1}$$

Nowadays, Benford's Law is used in detecting many different forms of fraud, and its prevalence in the world fascinates not only mathematicians, but many other scientists as well (to learn more about Benford's law and its many applications, we recommend [3, 13, 15] to name a few).

In 1986, Lemons [12] proposed using Benford's law to analyze the partitioning of a conserved quantity. Since then, driven by the potential application to nuclear fragmentation, mathematicians and physicists have taken an interest in the Benford behavior of various

fragmentation processes. Among these processes of interest is *stick fragmentation*. In a 1-dimensional stick fragmentation model, one begins with a stick of length L. Draw p_1 from a probability distribution on $(0, 1)$. This fragments the stick into two sub-sticks of lengths $p_1 L$ and $(1 - p_1)L$. For each sub-stick, draw two probabilities p_2 and p_3 from the same distribution. We perform this fragmentation process for a total of N stages, where at each stage we fragment all the sub-sticks created during the previous stage according to probabilities. Hence, by the end of N stages, there are 2^N sticks. Of particular interest is whether or not this fragmentation process converges to Benford's behavior.

1.1. Previous Work on Fragmentation

An important definition when studying a more precise statistical version of Benford behavior is the notion of the significand of a real number, i.e., its leading digits in scientific notation.

Definition 2. *Given a positive real number x, we say that its* ***significand base*** *$B > 1$, denoted $S_B(x)$, is the unique real number $S_B(x) \in [1, B)$ such that $x = S_B(x) \cdot B^k$, where k is an integer.*

As is common practice with these techniques involving proofs of Benford behavior, we define a more general version of Benford behavior that allows for processes that do not necessarily exhibit Benford behavior at first but converge to it in the limit.

Definition 3. *We say that a sequence of random variables $X^{(n)}$* ***converges to strong Benford behavior*** *base B if*

$$\mathbb{P}(S_B(X^{(n)}) \leq s) \to \log_B(s) \tag{2}$$

for all $s \in [1, B]$. Notice by compactness that this implies uniform convergence of (2).

An equivalent formulation to the above is the **Uniform Distribution Characterization**, which is especially suited for investigation of products of random variables; see [6] for a proof.

Proposition 1 (Uniform Distribution Characterization)**.** *[6] A sequence of random variables $X^{(n)}$ converges to strong Benford behavior in base B if and only if*

$$\mathbb{P}\left(\log_B(X^{(n)}) \bmod 1 \leq t\right) \to t, \tag{3}$$

for all $t \in [0, 1]$. If (1) is satisfied, then we say that $\log_B(X^{(n)})$ ***converges to being equidistributed mod 1****.*

We may now state some previous results on stick fragmentation. Becker et al. [1] proved a theorem regarding a fixed proportion 1-dimensional stick fragmentation process (compare with their Theorem 1.11), which provides a necessary and sufficient condition for the process to converge to strong Benford behavior and quantifies the discrepancy from uniform distribution mod 1 in terms of irrationality exponent. We first introduce some necessary definitions.

Definition 4. *[11] Suppose that x is a real number. The* ***irrationality exponent*** *μ of x is the supremum of the set of μ such that $0 < |x - p/q| < 1/q^{\mu}$ is satisfied by an infinite number of coprime integer pairs (p, q) with $q > 0$. If such a set does not exist, then we say x has irrationality exponent ∞.*

Definition 5. *[11] For a finite sequence $\{x_i\}_{i=1}^{n}$, denote by $A([a, b), n)$ the number of x_i's such that $x_i \bmod 1 \in [a, b)$ for $0 \leq a < b \leq 1$. Then we call the number*

$$D_n := \sup_{0 \leq a < b \leq 1} \left| \frac{A([a, b), n)}{n} - (b - a) \right| \tag{4}$$

the ***discrepancy*** *of the sequence. It measures how far a sequence is from being uniform mod 1.*

Now we are ready to state the theorem from [1].

Theorem 1. *[1] Choose any $p \in (0, 1)$. In Stage 1, cut a given stick at proportion p to create two pieces of length p and $1 - p$. In Stage 2, cut each resulting piece into two pieces at the same proportion p. Perform this process for a total of N stages, generating 2^N sticks with $N + 1$ distinct lengths (assuming $p \neq 1/2$) given by*

$$\begin{aligned} x_1 &= Lp^N \\ x_2 &= Lp^{N-1}(1 - p) \\ x_3 &= Lp^{N-2}(1 - p)^2 \\ &\vdots \\ x_N &= Lp(1 - p)^{N-1} \\ x_{N+1} &= L(1 - p)^N, \end{aligned} \tag{5}$$

where the frequency of x_n is $\binom{N}{n}/2^N$. Choose y so that $B^y = (1 - p)/p$, which is the ratio of adjacent lengths (i.e., x_{i+1}/x_i). The decomposition process results in stick lengths that converge to strong Benford's behavior base B if and only if $y \notin \mathbb{Q}$. If y has finite irrationality exponent, the discrepancy of the sequence $\{\log_B(x_i)\}_{i=1}^{N}$ can be quantified in terms of that exponent, and there is a power savings.

Theorem 1 suggests the possibility of examining another fixed proportion 1-dimensional stick fragmentation model, where we cut a piece into an arbitrary $m \geq 2$ number of pieces according $m-1$ fixed proportions at each stage. Using new techniques involving multinomial coefficients to reduce the problem into the original fixed proportion 1-dimensional stick fragmentation model in [1], this paper establishes the following result (see Section 2).

Theorem 2. *For any integer $m > 2$, choose $p_1, p_2, \ldots, p_{m-1} \in (0, 1)$ such that $p_1 + p_2 + \cdots + p_{m-1} < 1$. Set $p_m := 1 - (p_1 + p_2 + \cdots + p_{m-1})$. At each stage, we cut all sticks according to proportions $p_1, p_2, \ldots, p_{m-1}$ to create m pieces. After stage N, we have m^N sticks in total, of lengths*

$$A^{(N)}_{k_1, k_2, \ldots, k_m} = p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}, \tag{6}$$

for $0 \leq k_1, k_2, \ldots, k_m \leq N$ such that $k_1 + k_2 + \cdots + k_m = N$. Let $y_i = \log_B(p_i/p_{i+1})$ for $1 \leq i \leq m-1$. Then the decomposition process results in stick lengths that converge to strong Benford's behavior base B if and only if $y_i \notin \mathbb{Q}$ for some $1 \leq i \leq m-1$. Let κ_0 be the least irrationality exponent among all the irrational y_i's. Then the discrepancy of the sequence $\left\{\log_B(A^{(N)}_{k_1,k_2,\ldots,k_m})\right\}_{\substack{0 \leq k_1,k_2,\ldots,k_m \leq N \\ k_1+k_2+\cdots+k_m=N}}$ is $O(N^{\delta(-1/(\kappa_0-1)+\epsilon')})$ for some $\delta > 0$ and for every $\epsilon' > 0$.

Another possibility is to consider fragmentation processes in higher dimensions. This was considered by Durmić and Miller in [8] and Benford behavior was established in the case of volume of an arbitrary m-dimensional box under mild assumptions. This was generalized by Betti et al. in [5] to any d-dimensional volume of d-dimensional faces of an arbitrary m-dimensional box, where $d \leq m$. To state their results, we start with some definitions.

Definition 6. *We say that a set $\mathfrak{B} \subset \mathbb{R}^m$ is an m-**dimensional box** if it is a set of the form $[a_1, b_1] \times \cdots \times [a_m, b_m] \subset \mathbb{R}^m$, where $a_i < b_i$ are finite numbers.*

Definition 7. *A **linear-fragmentation process** is a sequence of random variables $\mathfrak{B}_0, \mathfrak{B}_1, \mathfrak{B}_2, \ldots$ such that the following holds.*

1. *The random variables $\mathfrak{B}_i$ are m-dimensional boxes.*
2. *The random variables $\mathfrak{B}_i$ form a descending chain $\mathfrak{B}_0 \supset \mathfrak{B}_1 \supset \mathfrak{B}_2 \supset \ldots$.*
3. *The distribution $\mathfrak{B}_{n+1}$ conditioned on $\mathfrak{B}_n$ is some fixed distribution of independent proportion cuts $P_1, \ldots, P_m$ along each Cartesian axis. These P_i are fixed over all $n \geq 0$.*
4. *The proportion cuts P_i are continuous random variables with finite mean, variance, and third moment.*
5. *$\mathbb{E}[\log_B P_i] = \mu_P \in \mathbb{R}$ and $\mathrm{Var}[\log_B P_i] = \sigma_P^2 > 0$ for all $1 \leq i \leq m$.*

Definition 8. *Given an m-dimensional box $\mathfrak{B}$ and a positive integer $d \leq m$, we say the d-**volume** of $\mathfrak{B} = \prod_i [a_i, b_i]$ is the sum of the d-dimensional volumes of the d-dimensional faces of $\mathfrak{B}$:*

$$\mathrm{Vol}_d(\mathfrak{B}) \;:=\; 2^{m-d} \sum_{|I|=d} \prod_{i \in I} (b_i - a_i), \tag{7}$$

where we are summing over all subsets $I \subset \{1, \ldots, m\}$ with cardinality d.

[5] established a sufficient condition for strong Benford behavior to emerge (see their Theorem 1.9), which involves the maximum d-dimensional volume of a d-dimensional face.

Theorem 3 (Maximum Criterion)**.** *[5] Let $\mathfrak{B} := \mathfrak{B}_0$ be a fixed m-dimensional box and $\mathfrak{B}_0 \supset \mathfrak{B}_1 \supset \cdots$ be a linear-fragmentation process whose proportion cuts P_i have probability density functions $f_i : (0,1) \to (0, \infty)$. Let*

$$V_d^{(N)} \;:=\; \mathrm{Vol}_d(\mathfrak{B}_N) \tag{8}$$

be the sequence of volumes obtained from this process and $\mathfrak{m}_d^{(N)}$ denote the maximum product of d sides of $\mathfrak{B}_N$. If $\mathfrak{m}_d^{(N)}$ converges to strong Benford behavior base B as $N \to \infty$, then so too does $V_d^{(N)}$ converge to strong Benford behavior base B as $N \to \infty$.

In the same paper, the Maximum Criterion was verified for $d = m$ for any linear fragmentation process, and for $d = 1$ in a special case.

Theorem 4. *Let $P_i^{(j)}$ be i.i.d. log-uniform distributions, i.e., $\log_B P_i^{(j)}$'s have uniform distributions. In the case of $d = 1$, i.e., perimeter, the maximum side length of each box*

$$\mathfrak{m}_1^{(N)} \;=\; \max_{1\le i\le m} P_i^{(1)} \cdots P_i^{(N)} \tag{9}$$

converges to strong Benford behavior base B as $N \to \infty$.

Hence, it was conjectured in the same paper (see Conjecture 4.1) that for any linear fragmentation process, $m_d^{(N)}$ converges to strong Benford behavior.

Conjecture 1. *[5] Every linear fragmentation process satisfies the Maximum Criterion in all dimensions $1 \le d \le m$.*

We prove the conjecture for any $m \ge 2$ and $d \le m$ under some mild conditions on the linear fragmentation process (see Section 3).

Theorem 5. *Suppose that $\mathfrak{B}_0 \supset \mathfrak{B}_1 \supset \mathfrak{B}_2 \supset \cdots$ is a linear fragmentation process on an m-dimensional box for an arbitrary $m \ge 2$, such that the proportion cut P_i at every stage satisfies the following.*

1. *We have $\mathbb{E}[\log_B(P_i)] = 0$ and $\mathrm{Var}[\log_B(P_i)] = 1$.*
2. *We have $\log_B(P_i)$ is supported on $[-C, C]$, where $C > 0$ is a constant.*
3. *Suppose that $P_i^{(k)}$ is a sequence of i.i.d. random variables $\sim P_i$. Define the random variable*

$$Z_i^{(N)} \;:=\; \frac{\log_B(P_i^{(1)} \cdots P_i^{(N)})}{\sqrt{N}}. \tag{10}$$

Then $Z_i^{(N)}$ has probability density function $f_{Z_i^{(N)}}(z_i) = \varphi(z_i) + A(z_i)$ and cumulative density function $F_{Z_i^{(N)}}(z_i) = \Phi(z_i) + B(z_i)$, where $\varphi(x)$ and $\Phi(x)$ are the probability density function and cumulative density function of the standard normal $N(0, 1)$, $A(x) = O(N^{-1/2-\delta})$ and $B(x) = O(N^{-\delta})$ for some $\delta > 0$.

Then this linear fragmentation process satisfies the condition of the Maximum Criterion for any $d \le m$, i.e., $\mathfrak{m}_d^{(N)}$ converges to strong Benford behavior base B as $N \to \infty$ for any $d \le m$, thus by Theorem 3, $V_d^{(N)}$ converges to strong Benford behavior base B as $N \to \infty$ for any $d \le m$.

2. Proof of Theorem 2

2.1. Preliminary

In [10], Fang, Irons, Lippelman and Miller proved that if $y_i \in \mathbb{Q}$ for all $1 \le i \le m - 1$, then the distribution of the stick lengths does not follow Benford behavior as $N \to \infty$. To be exact, they were able to show that the distribution of the logarithms of the stick lengths converges to a discrete distribution, contrary to a uniform distribution on $[0, 1]$ which is an

equivalent condition for convergence to strong Benford behavior by Uniform Distribution Characterization 1. Hence, having $y_i \notin \mathbb{Q}$ for some $1 \le i \le m-1$ is a necessary condition for the distribution of the stick lengths to follow Benford behavior.

For the remainder of this section, we show that having $y_i \notin \mathbb{Q}$ for some $1 \le i \le m-1$ is also a sufficient condition for the distribution of the stick lengths to converge to Benford's behavior. Since we can always reorder the $p_i^{k_i}$'s in the stick length $A^{(N)}_{k_1,k_2,\ldots,k_m} := p_1^{k_1}p_2^{k_2}\cdots p_m^{k_m}$ and thus the order of y_i's in the factorization, then it suffices to show that the first exponent $y_1 \notin \mathbb{Q}$ is a sufficient condition for the distribution of the stick lengths to converge to Benford's behavior.

We know that after stage N, the number of sticks of length $A^{(N)}_{k_1,k_2,\ldots,k_m}$ is equal to the multinomial coefficient

$$\binom{k_1+k_2+\cdots+k_m}{k_1,k_2,\ldots,k_m} \;=\; \binom{N}{k_1,k_2,\ldots,k_m}. \tag{11}$$

From the proof of Theorem , we know that the stick length $A^{(N)}_{k_1,k_2,\ldots,k_m}$ has the factorization

$$A^{(N)}_{k_1,k_2,\ldots,k_m} \;=\; p_1^{k_1}p_2^{k_2}\cdots p_m^{k_m} \;=\; \left(\frac{p_1}{p_2}\right)^{k_1}\left(\frac{p_2}{p_3}\right)^{k_1+k_2}\cdots\left(\frac{p_{m-1}}{p_m}\right)^{\sum_{j=1}^{m-1}k_j}(p_m)^N, \tag{12}$$

which motivates us to define the exponents $y_i = \log_B(p_i/p_{i+1})$ for $1 \le i \le m-1$. Now, fix an interval $(a,b) \subset (0,1)$. Let

$$\chi_{(a,b)}(x) \;:=\; \mathbb{1}\Big(\log_B(x) \bmod 1 \in (a,b)\Big) \tag{13}$$

be the indicator function for $\log_B(x) \bmod 1 \in (a,b)$. Then after stage N, the probability that a stick length mod 1 is in (a,b) equals

$$F_N(a,b) \;:=\; \frac{1}{m^N}\sum_{\substack{0\le k_1,k_2\ldots,k_m\le N\\ k_1+k_2+\cdots+k_m=N}}\binom{N}{k_1,k_2,\ldots,k_m}\chi_{(a,b)}\left(\prod_{i=1}^{m}p_i^{k_i}\right). \tag{14}$$

Our goal is to prove that as $N\to\infty$,

$$F(a,b) \;:=\; \lim_{N\to\infty}F_N(a,b) \;=\; b-a, \tag{15}$$

which is a sufficient condition for Benford behavior.

2.2. Combinatorial identities

The key steps to our proof are two combinatorial identities related to multinomial coefficients, which we present and prove below. We first review the multinomial coefficients.

Definition 9. *Suppose that $n, k_1, k_2, \ldots, k_m \ge 0$ and $k_1+\cdots+k_m = n$, then we define the multinomial coefficients by*

$$\binom{n}{k_1,k_2,\ldots,k_m} \;:=\; \frac{n!}{k_1!k_2!\cdots k_m!}. \tag{16}$$

The following identity on the factorization of multinomial coefficients allows us to express multinomial coefficients in terms of product of binomial coefficients. As we shall see, this forms the key ingredient that reduces our multi-proportion problem into the original single-proportion problem in [1].

Lemma 1. *[16] For $m \geq 1$,*

$$\binom{k_1 + k_2 + \cdots + k_m}{k_1, k_2, \ldots, k_m} = \prod_{i=1}^{m} \binom{k_1 + \cdots + k_i}{k_i}. \tag{17}$$

The proof for the identity above is a straightforward induction through the expansion of the multinomial and binomial coefficients. Below is another identity that is especially useful later for evaluating sums of multinomial coefficients.

Lemma 2. *[16] For $N \in \mathbb{Z}_{\geq 0}$ and $m \in \mathbb{Z}_{>0}$,*

$$m^N = \sum_{\substack{0 \leq k_1, k_2, \ldots, k_m \leq N \\ k_1 + k_2 + \cdots + k_m = N}} \binom{N}{k_1, k_2, \ldots, k_m} = \sum_{\substack{0 \leq k_1, \ldots, k_{m-1} \leq N \\ k_1 + k_2 + \cdots + k_{m-1} = N}} 2^{k_1} \binom{N}{k_1, k_2, \cdots, k_{m-1}}. \tag{18}$$

The most straightforward proof of the identity above relies on two different ways of representing and expanding m^N: one as $(\underbrace{1 + \cdots + 1}_{N})^N$ and the other as $(2 + \underbrace{1 + \cdots + 1}_{N-2})^N$. See [16] for more.

2.3. Truncation

We prove that for any $\epsilon \in (0, 1)$, the contribution to $F_N(a, b)$ of multinomial coefficients with $k_1 + k_2 < N^\epsilon$ is negligible. This allows us to only consider $k_1 + k_2$ that are sufficiently large for these stick lengths $A_{k_1, k_2, \ldots, k_m}$.

Proposition 2. *For any $\epsilon \in (0, 1)$,*

$$\sum_{\substack{0 \leq k_1, k_2 \ldots, k_m \leq N \\ k_1 + k_2 + \cdots + k_m = N \\ k_1 + k_2 < N^\epsilon}} \binom{N}{k_1, k_2, \ldots, k_m} \leq (m-2)^N N^{N^\epsilon + 2}. \tag{19}$$

Proof. We know that if $k_1 + k_2 + \cdots + k_m = N$, then

$$\begin{aligned} \binom{N}{k_1, k_2, \ldots, k_m} &= \frac{N(N-1)\cdots(N - (k_1 + k_2) + 1)}{k_1! k_2!} \cdot \frac{(N - (k_1 + k_2))!}{k_3! k_4! \cdots k_m!} \\ &\leq N^{k_1 + k_2} \cdot \binom{N - (k_1 + k_2)}{k_3, k_4, \ldots, k_m}. \end{aligned} \tag{20}$$

Hence,

$$\begin{aligned}
&\sum_{\substack{0\le k_1,k_2,\ldots,k_m\le N\\ k_1+k_2+\cdots+k_m=N\\ k_1+k_2<N^\epsilon}} \binom{N}{k_1,k_2,\ldots,k_m}\\
&= \sum_{\substack{0\le k_1,k_2<N^\epsilon\\ k_1+k_2<N^\epsilon}} N^{k_1+k_2} \sum_{\substack{0\le k_3,k_4,\ldots,k_m\le N-(k_1+k_2)\\ k_3+k_4+\cdots+k_m=N-(k_1+k_2)}} \binom{N-(k_1+k_2)}{k_3,k_4,\ldots,k_m}\\
&\le (m-2)^N \sum_{\substack{0\le k_1,k_2<N^\epsilon\\ k_1+k_2<N^\epsilon}} N^{k_1+k_2}\\
&\le (m-2)^N (N^\epsilon)^2 N^{N^\epsilon} \le (m-2)^N N^{N^\epsilon+2}. \qquad (21)
\end{aligned}$$

□

We know the probability that a stick length $p_1^{k_1}p_2^{k_2}\cdots p_m^{k_m}$ satisfies $k_1+k_2<N^\epsilon$ is

$$\frac{1}{m^N} \sum_{\substack{0\le k_1,k_2\ldots,k_m\le N\\ k_1+k_2+\cdots+k_m=N\\ k_1+k_2<N^\epsilon}} \binom{N}{k_1,k_2,\ldots,k_m}. \qquad (22)$$

According to Proposition 2, we have that (22) is bounded above by $((m-2)/m)^N N^{N^\epsilon+2}$. Hence, the logarithm of the probability is bounded above by

$$\log\left(\left(\frac{m-2}{m}\right)^N N^{N^\epsilon+2}\right) = N\log\left(\frac{m-2}{m}\right) + (N^\epsilon+2)\log(N). \qquad (23)$$

It is clear that (23) goes to $-\infty$ as $N\to\infty$. Thus the probability goes to 0 as $N\to\infty$. So it suffices to consider the case where $k_1+k_2\ge N^\epsilon$. Now, fix $k(N)\ge N^\epsilon$ and let $0\le k_1,k_2\le k(N)$ with $k_1+k_2=k(N)$ and $0\le k_3,k_4,\ldots,k_m\le N-k(N)$ with $k_3+k_4+\cdots+k_m=N-k(N)$. By [1], we know that the frequency k_1 of stick length $A^{(N)}_{k_1,k_2,\ldots,k_m}=p_1^{k_1}p_2^{k_2}\ldots p_m^{k_m}$ follows a binomial distribution with mean $k(N)/2$ and standard deviation $\sqrt{k(N)}/2$. Pick some $\delta\in(0,\epsilon/10)$. We see that it suffices to consider cases where $|k_1-k(N)/2|<(\lceil N^\delta\rceil\sqrt{k(N)})/2$, because the probability that k_1 is outside this range is asymptotically small by Chebyshev's inequality:

$$\mathbb{P}\left(\left|k_1-\frac{k(N)}{2}\right| \ge \frac{\lceil N^\delta\rceil\sqrt{k(N)}}{2}\right) \le \frac{1}{\lceil N^\delta\rceil^2}. \qquad (24)$$

2.4. Near uniform probability within small intervals

We keep the same notation and definition and fix ϵ, δ, $k(N):=k_1+k_2$, $k_3,k_4,\ldots,k_m$ as before. Our goal is to prove that the stick length $A^{(N)}_{k_1,k_2,\ldots,k_m}$ is roughly uniformly distributed over small intervals of k_1. Since $\lceil N^\delta\rceil = o(\lceil N^\delta\rceil\lceil\sqrt{k(N)}/2\rceil)$, we can evenly divide the range

of k_1 between $k(N)/2 \pm \lceil N^\delta \rceil \lceil \sqrt{k(N)}/2 \rceil$ into intervals of $\lceil N^\delta \rceil$ values of k_1, where the ℓ^{th} interval starts with $k_{1,\ell} = k_{1,\ell,0}$ and ranges over $k_{1,\ell,i}$ for $0 \le i \le \lceil N^\delta \rceil - 1$, as defined below:

$$\begin{aligned} k_{1,\ell} &:= f_\ell\left(\frac{k(N)}{2}\right) + \ell\lceil N^\delta \rceil, \\ k_{1,\ell,i} &:= f_\ell\left(\frac{k(N)}{2}\right) + \ell\lceil N^\delta \rceil + i, \quad 0 \le i \le N^\delta - 1, \end{aligned} \tag{25}$$

where $f_\ell(\cdot) := \lceil \cdot \rceil$ when $\ell \ge 0$ and $f_\ell(\cdot) := \lfloor \cdot \rfloor$ when $\ell < 0$. We see that ℓ ranges from $-\lceil \sqrt{k(N)}/2 \rceil$ to $\lceil \sqrt{k(N)}/2 \rceil$. Note that we are using the floor and the ceiling functions to ensure that $f_{1,\ell,i}$'s have integer values. For convenience, we will make a slight abuse of notation from now on to drop the floor and the ceiling signs, as they have negligible effect on our calculation. We want to show that the difference $\left|\binom{k(N)}{k_{1,\ell,i}} - \binom{k(N)}{k_{1,\ell,j}}\right|$ is asymptotically smaller than $\binom{k(N)}{k_{1,\ell}}$ uniformly for all $0 \le i < j \le N^\delta - 1$, which would imply that the stick length is roughly uniformly distributed over small intervals of k_1. Since the binomial distribution is symmetric around its mean, it suffices to look at $\ell \ge 0$. Moreover, the probability density function is monotonically decreasing to the right of the mean, the difference is uniformly bounded within each interval and

$$\left|\binom{k(N)}{k_{1,\ell,i}} - \binom{k(N)}{k_{1,\ell,j}}\right| \le \left|\binom{k(N)}{k_{1,\ell}} - \binom{k(N)}{k_{1,\ell+1}}\right|. \tag{26}$$

We follow Section (5.52) of [1] to obtain a bound for the difference.

Proposition 3. *For $\ell \le \sqrt{k(N)}/2$,*

$$\left|\binom{k(N)}{k_{1,\ell}} - \binom{k(N)}{k_{1,\ell+1}}\right| \le O\left(\binom{k(N)}{k_{1,\ell}} \cdot N^{-\frac{3\epsilon}{10}}\right). \tag{27}$$

Since the proof follows similar argument to Section (5.52) of [1], we leave the proof details to Appendix A.

2.5. Equidistribution within small intervals

In this subsection, we want to show that for fixed $k(N), k_3, \ldots, k_m$, logarithm base B of the stick length $\log_B(A^{(N)}_{k_{1,\ell,i},k_2,\ldots,k_m})$ for $0 \le i \le \lceil N^\delta \rceil - 1$ converges to being equidistributed mod 1, in the sense that logarithm of the stick length is thought of as a random variable that takes value in $\log_B(A^{(N)}_{k_{1,\ell,i},k_2,\ldots,k_m})$ for $0 \le i \le \lceil N^\delta \rceil - 1$ with equal probabilities. We first state the following theorem which provides an easy criterion for checking equidistribution mod 1.

Lemma 3 (Weyl's Criterion)**.** *[17] A sequence $\{a_n\}_{n=1}^\infty$ is equidistributed mod 1 if and only if for all nonzero integers ℓ,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} e^{2\pi i \ell a_j}. \tag{28}$$

Now, notice for fixed $k(N)$ and ℓ, $\log_B(A^{(N)}_{k_1,\ell,i},k_2,\ldots,k_m)$ forms an arithmetic progression over the ℓ^{th} interval of k_1 with common difference $\log_B(p_1/p_2) \in \mathbb{Q}$. Then equidistribution mod 1 is a direct consequence of the following corollary of Lemma 3.

Corollary 1. *[11] For any $a \in \mathbb{R}$, the arithmetic progression $\{a+nd\}_{n=1}^{\infty}$ is equidistributed mod 1 if and only if $d \notin \mathbb{Q}$.*

Corollary 1 suffices to establish convergence to equidistribution mod 1 of $\log_B(A^{(N)}_{k_1,\ell,i},k_2,\ldots,k_m)$. However, it does not provide any quantitative measure of the discrepancy from uniform distribution on $[0,1]$. It turns out that the key ingredient behind this quantification is irrationality exponent. We restate the definition here.

Definition 10. *[11] Suppose that x is a real number. The **irrationality exponent** μ of x is the supremum of the set of μ such that $0 < |x - p/q| < 1/q^{\mu}$ is satisfied by an infinite number of coprime integer pairs (p,q) with $q > 0$. If such a set does not exist, then we say x has irrationality exponent ∞.*

Since $\mathbb{Q}$ is dense in $\mathbb{R}$, then every real number can be approximated by rational numbers. However, what differentiates them is how well they can be approximated. Irrationality exponent measures exactly how well a real number can be approximated by rational numbers. The bigger the irrationality exponent, the finer the rational approximation can be. The following are some well-known facts about irrationality exponent, which comes handy later in this section.

Proposition 4. *[11] The irrationality exponent of any rational number is 1. As a consequence of Dirichlet's approximation theorem, which states that for any irrational number x,*

$$\left| x - \frac{p}{q} \right| \;<\; \frac{1}{q^2} \tag{29}$$

for infinitely (p,q) where p and q are coprime, the irrationality exponent of any irrational number is at least 2. Further more, any algebraic irrational (irrational numbers that are zeros of polynomials over $\mathbb{Q}$) has irrationality exponent exactly 2.

We now state a theorem that quantifies discrepancy of the logarithm of the stick length in terms of irrationality exponent.

Theorem 6. *[11, Theorem 3.2] Let κ be the irrationality exponent of $\log_B(p_1/p_2)$ and $J_\ell(a,b) \subset \{0,1,\ldots,N^\delta - 1\}$ be the set of indices i such that $\log_B(A^{(N)}_{k_1,\ell,i},k_2,\ldots,k_m)$ mod 1 is in (a,b). If $\kappa < \infty$, then*

$$|J_\ell(a,b)| \;=\; (b-a)N^\delta + O\left(N^{\delta\left(1-\frac{1}{\kappa-1}+\epsilon'\right)}\right), \tag{30}$$

for every $\epsilon' > 0$ and there is a power saving with the error term. The error term is optimal in the sense that

$$\left| \frac{J_\ell(a,b)}{N^\delta} - (b-a) \right| \;=\; \Omega\left(N^{\delta(-\frac{1}{\kappa-1})-\epsilon'}\right) \tag{31}$$

for every $\epsilon' > 0$. If $\kappa = \infty$, then

$$|J_\ell(a,b)| \;=\; (b-a)N^\delta + o(N^\delta). \tag{32}$$

Note that [11] has a different convention for irrationality exponent and if the irrationality exponent of x is μ in our definition, then it is $\mu-1$ in their definition. This is the reason for the change made to the presentation of Theorem [11], where there is a $-1/(\kappa-1)$ instead of a $1/\kappa$ in the exponent.

2.6. Evaluation of sums within and across intervals

We first proceed to our calculation with the case where $\log_B(p_1/p_2)$ has a finite irrationality exponent $\kappa < \infty$. For fixed $k(N) := k_1 + k_2, k_3, k_4, \dots k_m$, and $\ell \leq \sqrt{k(N)}/2$, we first count the number of lengths $A^{(N)}_{k_{1,\ell,i},k_2,\dots,k_m}$ such that $\log_B(A^{(N)}_{k_{1,\ell,i},k_2,\dots,k_m}) \bmod 1$ is in (a,b) within the ℓ^{th} interval. By definition, this is given by

$$\sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}}. \tag{33}$$

By Proposition 3 and Section 2.5, we have

$$\begin{aligned}
\sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} &= \sum_{i\in J_\ell(a,b)} \left\{ \binom{k(N)}{k_{1,\ell}} + O\left(\binom{k(N)}{k_{1,\ell}} N^{-\frac{3\epsilon}{10}} \right) \right\} \\
&= \left\{ \binom{k(N)}{k_{1,\ell}} \sum_{i\in J_\ell(a,b)} 1 \right\} + O\left(\binom{k(N)}{k_{1,\ell}} N^{-\frac{3\epsilon}{10}} \sum_{i\in J_\ell(a,b)} 1 \right) \\
&= (b-a)N^{\delta} \binom{k(N)}{k_{1,\ell}} + O\left(\binom{k(N)}{k_{1,\ell}} N^{\delta\left(1-\frac{1}{\kappa-1}+\epsilon'\right)} \right).
\end{aligned} \tag{34}$$

We justify the error term in the last line of (34). Since κ is the irrationality exponent of an irrational number, then $\kappa \geq 1$ (see [2]). Also, $\epsilon' > 0$, so $-1/2 < -1/(\kappa-1)+\epsilon'$. Moreover, for some proper choice of ϵ', we have $-1/(\kappa-1)+\epsilon' < 0$. Combining these with the fact that $\delta \in (0, \epsilon/10)$ gives the dominant error term in the last line of (34).

Now, we count the number of stick lengths $A^{(N)}_{k_{1,\ell,i},k_2,\dots,k_m}$ such that $\log_B(A^{(N)}_{k_{1,\ell,i},k_2,\dots,k_m}) \bmod 1$ is in (a,b) over all the intervals. By the truncation in Section 2.3 the main term comes from the sum over ℓ from $-\sqrt{k(N)}/2$ to $\sqrt{k(N)}/2$. The number is given by

$$\sum_{\ell=-k(N)/(2N^\delta)}^{k(N)/(2N^\delta)} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} = \sum_{\ell=-\sqrt{k(N)}/2}^{\sqrt{k(N)}/2} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} + \sum_{|\ell|>\sqrt{k(N)}/2} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}}. \tag{35}$$

Using (24), we can provide an upper bound on the second sum on the RHS of (35):

$$\begin{aligned}
\sum_{|\ell|>\sqrt{k(N)}/2} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} &\leq \mathbb{P}\left(\left| k_1 - \frac{k(N)}{2} \right| \geq \frac{N^\delta \sqrt{k(N)}}{2} \right) \cdot \sum_{0\leq k_1 \leq N} \binom{k(N)}{k_1} \\
&= \mathbb{P}\left(\left| k_1 - \frac{k(N)}{2} \right| \geq \frac{N^\delta \sqrt{k(N)}}{2} \right) \cdot 2^{k(N)} = O\left(\frac{2^{k(N)}}{N^{2\delta}} \right).
\end{aligned} \tag{36}$$

On the other hand, we apply (34) to the first sum on the RHS of (35):

$$\sum_{\ell=-\sqrt{k(N)}/2}^{\sqrt{k(N)}/2} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} = \sum_{\ell=-\sqrt{k(N)}/2}^{\sqrt{k(N)}/2} (b-a)N^\delta \binom{k(N)}{k_{1,\ell}} + O\left(\sum_{\ell=-\sqrt{k(N)}/2}^{\sqrt{k(N)}/2} \binom{k(N)}{k_{1,\ell}} N^{\delta\left(1-\frac{1}{\kappa-1}+\epsilon'\right)} \right). \tag{37}$$

By (36) and (5.17) and (5.18) of [1], by choosing $q = N^\delta$, we have

$$\sum_{\ell=-\sqrt{k(N)}/2}^{\sqrt{k(N)}/2} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} = (b-a)2^{k(N)} + O\left(2^{k(N)} N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)}\right). \tag{38}$$

Substituting (36) and (38) into (35) yields

$$\sum_{\ell=-k(N)/(2N^\delta)}^{k(N)/(2N^\delta)} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} = (b-a)2^{k(N)} + O\left(2^{k(N)} N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)}\right). \tag{39}$$

Finally, we sum over all $k(N), k_3, k_4, \ldots, k_m$, which gives us the total number of lengths $A^{(N)}_{k_1,k_2,\ldots,k_m}$ such that $\log_B(A^{(N)}_{k_1,k_2,\ldots,k_m}) \bmod 1$ is in (a,b). We break the sum based on whether $k(N) \geq N^\epsilon$ or not. For $k(N) \geq N^\epsilon$, we apply (39):

$$\begin{aligned}
&\sum_{\substack{0\leq k(N),k_3,k_4,\ldots,k_m\leq N \\ k(N)+k_3+k_4+\cdots+k_m=N \\ k(N)\geq N^\epsilon}} \binom{N}{k(N),k_3,k_4,\ldots,k_m} \sum_{\ell=-k(N)/(2N^\delta)}^{k(N)/(2N^\delta)} \sum_{i\in J_\ell(a,b)} \binom{k(N)}{k_{1,\ell,i}} \\
&= (b-a) \sum_{\substack{0\leq k(N),k_3,k_4,\ldots,k_m\leq N \\ k(N)+k_3+k_4+\cdots+k_m=N \\ k(N)\geq N^\epsilon}} 2^{k(N)} \binom{N}{k(N),k_3,k_4,\ldots,k_m} \\
&\quad + O\left(N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)} \sum_{\substack{0\leq k(N),k_3,k_4,\ldots,k_m\leq N \\ k(N)+k_3+k_4+\cdots+k_m=N \\ k(N)\geq N^\epsilon}} 2^{k(N)} \binom{N}{k(N),k_3,k_4,\ldots,k_m} \right).
\end{aligned} \tag{40}$$

Note that by Lemma 2

$$\begin{aligned}
&\sum_{\substack{0\le k(N),k_3,k_4,\ldots,k_m\le N\\ k(N)+k_3+k_4+\cdots+k_m=N\\ k(N)\ge N^\epsilon}} 2^{k(N)}\binom{N}{k(N),k_3,k_4,\ldots,k_m}\\
&= m^N - \sum_{\substack{0\le k_1,k_2,\ldots,k_m\le N\\ k_1+k_2+\cdots+k_m=N\\ k_1+k_2<N^\epsilon}}\sum_{k_1=0}^{k_1+k_2}\binom{N}{k_1,k_2,\ldots,k_m}\\
&= m^N + O((m-2)^N N^{N^\epsilon+2}), \qquad (41)
\end{aligned}$$

where in the last line, we apply Proposition 2. Substituting (41) into (40) gives

$$\begin{aligned}
&\sum_{\substack{0\le k(N),k_3,k_4,\ldots,k_m\le N\\ k(N)+k_3+k_4+\cdots+k_m=N\\ k(N)\ge N^\epsilon}}\binom{N}{k(N),k_3,k_4,\ldots,k_m}\sum_{\ell=-k(N)/(2N^\delta)}^{k(N)/(2N^\delta)}\sum_{i\in J_\ell(a,b)}\binom{k(N)}{k_{1,\ell,i}}\\
&= (b-a)m^N + O\left((m-2)^N N^{N^\epsilon+2} + m^N N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)} + (m-2)^N N^{N^\epsilon+2} N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)}\right)\\
&= (b-a)m^N + O\left(m^N N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)}\right). \qquad (42)
\end{aligned}$$

For $k(N) < N^\epsilon$, by Lemma 17 and Proposition 2 we have

$$\begin{aligned}
&\sum_{\substack{0\le k(N),k_3,k_4,\ldots,k_m\le N\\ k(N)+k_3+k_4+\cdots+k_m=N\\ k(N)<N^\epsilon}}\binom{N}{k(N),k_3,k_4,\ldots,k_m}\sum_{\ell=-k(N)/(2N^\delta)}^{k(N)/(2N^\delta)}\sum_{i\in J_\ell(a,b)}\binom{k(N)}{k_{1,\ell,i}}\\
&\le \sum_{\substack{0\le k(N),k_3,k_4,\ldots,k_m\le N\\ k(N)+k_3+k_4+\cdots+k_m=N\\ k(N)<N^\epsilon}}\binom{N}{k(N),k_3,k_4,\ldots,k_m}\sum_{k_1=0}^{k(N)}\binom{k(N)}{k_1}\\
&= \sum_{\substack{0\le k_1,k_2,\ldots,k_m\le N\\ k_1+k_2+\cdots+k_m=N\\ k_1+k_2<N^\epsilon}}\binom{N}{k_1,k_2,\ldots,k_m} = O\left((m-2)^N N^{N^\epsilon+2}\right). \qquad (43)
\end{aligned}$$

Combining (42) and (43), we have

$$\begin{aligned}
&\sum_{\substack{0\le k(N),k_3,k_4,\ldots,k_m\le N\\ k(N)+k_3+k_4+\cdots+k_m=N}}\binom{N}{k(N),k_3,k_4,\ldots,k_m}\sum_{\ell=-k(N)/(2N^\delta)}^{k(N)/(2N^\delta)}\sum_{i\in J_\ell(a,b)}\binom{k(N)}{k_{1,\ell,i}}\\
&= (b-a)m^N + O\left(N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)}m^N + N^{N^\epsilon+2}(m-2)^N\right)\\
&= (b-a)m^N + O\left(N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)}m^N\right), \qquad (44)
\end{aligned}$$

After dividing by m^N, the total number of sticks after stage N, we arrive at

$$F_N(a,b) \;=\; b-a+O\left(N^{\delta\left(-\frac{1}{\kappa-1}+\epsilon'\right)}\right). \tag{45}$$

We have shown that if $\log_B(p_1/p_2)$ has a finite irrationality exponent κ, then the logarithm of the stick lengths is equidistributed mod 1, and thus the stick lengths converge to the Benford distribution. Moreover, the discrepancy between the distribution of the logarithm of the stick length and uniform distribution mod 1 is quantified in terms of the irrationality exponent of $\log_B(p_1/p_2)$. Equation (45) tells us that the smaller the irrationality exponent, the smaller the discrepancy is (thanks to the guarantee that the error term is optimal), and the closer the distribution of the stick length is to the Benford distribution. If $\log(p_1/p_2)$ has infinite irrationality exponent, the calculation is exactly the same, except that we start with the error term $o(N^\delta)$ instead of $O(N^{\delta\left(1-\frac{1}{\kappa-1}+\epsilon'\right)})$ in Section 2.5, and so we end up with the $F_N(a,b) \;=\; b-a+o(1)$.

Thus, this establishes the strong Benford behavior of fixed multi-proportion 1-dimensional stick fragmentation process when $y_i \notin \mathbb{Q}$ for some $1 \le i \le m-1$. On a final note, recall from Section 2.1 our choices of the order of $p_1, p_2, \ldots, p_m$ and thus factorization are arbitrary. Hence, to obtain an optimal error term, we simply need to choose an order of $p_1, p_2, \ldots, p_m$ such that $\log_B(p_1/p_2)$ is irrational and its irrationality exponent is minimized. Let κ_0 be the irrationality exponent of $\log_B(p_1/p_2)$. Then we have

$$F_N(a,b) \;=\; b-a+O\left(N^{\delta\left(-\frac{1}{\kappa_0-1}+\epsilon'\right)}\right). \tag{46}$$

□

3. Proof of Theorem 5

3.1. Preliminary

In this section, we prove Theorem 5, i.e., resolve Conjecture 4.1 of [5] that we mentioned in Section 1 under some mild conditions. Let us first recall the conjecture.

Conjecture 2. *Every linear fragmentation process satisfies the maximum criterion in all dimensions* $1 \le d \le m$.

[5] has verified the conjecture in the case of $d = 1$ and arbitrary m for i.i.d. log-uniform distributions, i.e., the proportion at which each dimension is cut is P_i, where $\log P_i \sim \text{Uniform}(a,b)$ and $a < b \le 0$. They decide to normalize P_i to $\text{Uniform}(-\sqrt{3},\sqrt{3})$ with mean 0 and variance 1, which simplifies the calculation later on and does not affect the nature or the solution to the problem. After stage N, the side length of the i^{th} dimension of the box, after taking logarithm and normalizing, is

$$Z_i^{(N)} \;:=\; \frac{\log_B(P_i^{(1)}\cdots P_i^{(N)}) - N\mu_P}{\sqrt{N}\sigma_P} \;=\; \frac{\log_B(P_i^{(1)}\cdots P_i^{(N)})}{\sqrt{N}}. \tag{47}$$

By the CLT, $Z_i^{(N)}$ converges to $N(0,1)$. Moreover, Lemma 3.1 states that if $f_{Z_i^{(N)}}(z)$ is the probability density function of $Z_i^{(N)}$ and $\varphi(z)$ is the probability density function of $N(0,1)$, then for all $z \in [-\sqrt{3}N, \sqrt{3}N]$ and any $\epsilon > 0$, we have

$$f_{Z_i^{(N)}}(z_i) \;=\; \varphi(z_i) + O(N^{-1+4\epsilon}), \tag{48}$$

where

$$\varphi(x) \;:=\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{49}$$

Due to the presence of the error term, to establish the strong Benford behavior of the process, they need to rely on the fact that the distributions for the cuts are finitely supported with uniformly bounded supports. They also need to rely on the fact that the exponent of the error term is less than $-1/2$ for sufficiently small $\epsilon > 0$, which allows the error term from integrating $f_{Z_i^{(N)}}(z)$ over the support to be asymptotically 0.

For the rest of this paper, we prove Conjecture 4.1 of [5] for any $m \geq 2$ and $d \leq m$. Physically, this means that we consider any arbitrary dimensional volume of faces of that dimension of the m-dimensional box. We assume that the $\log(P_i)$ are i.i.d., have mean 0 and variance 1, are finitely supported with support $[-C, C]$, and that the probability density function (PDF) of $Z_i^{(N)}$ is $f_{Z_i^{(N)}}(z_i) = \varphi(z_i) + A(z_i)$, where $A(z_i) = O(N^{-1/2-\delta})$ for some $\delta > 0$. We also assume that the cumulative density function (CDF) of $Z_i^{(N)}$ is given by $F_{Z_i^{(N)}}(z_i) = \Phi(z_i) + B(z_i)$, where Φ is the CDF of $N(0,1)$ and $B(z_i) = O(N^{-\delta})$ (for an example of a distribution satisfying these condition, see [5]). We prove that the maximum d-dimensional volume $\mathfrak{m}_d^{(N)}$ of a d-dimensional face of the m-dimensional box converges to strong Benford behavior. Once we prove this statement, then by the maximum criterion established in [5], we have that the d-dimensional volume $V_d^{(N)}$ of the d-dimensional faces of the m-dimensional box also converges to strong Benford behavior.

The key idea behind our proof is the use of the Mean Value Theorem and a recursive method of change of variables and consistent cancellation that provide upper bounds on the error terms in terms of Gaussian densities. We shall explore this more in Section 3.6.

To begin, we use knowledge from order statistics to find the PDF of $\mathfrak{m}_d^{(N)}$.

Definition 11. *Suppose that $X_1, X_2, \ldots, X_k$ are random variables, and $[]_i$ returns the i^{th} largest number among a list $(a_1, a_2, \ldots, a_k)$ of real numbers. For each outcome ω in the sample space Ω, we define*

$$X_{(k-i+1)}(w) \;:=\; [(X_1(w), X_2(w), \ldots, X_k(w))]_{k-i+1}. \tag{50}$$

*We say that the random variable $X_{(k-i+1)}$ is **the i^{th} order statistics**, or the i^{th} largest random variable among $X_1, X_2, \ldots, X_k$. Hence, as random variables, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(k)}$.*

Let $S_i^{(N)}$ be the side length of the i^{th} dimension of the m-dimensional box after stage N. Then the maximum volume of a d-dimensional face is the product of the longest d sides of

the box, i.e., $\mathfrak{m}_d^{(N)} = \prod_{i=1}^{d} S_{(m+1-i)}^{(N)}$. We define

$$Y^{(N)} \;:=\; \frac{\log_B(\mathfrak{m}_d^{(N)})}{\sqrt{N}} \;=\; \sum_{i=1}^{d} \frac{\log_B(S_{(m+1-i)}^{(N)})}{\sqrt{N}} \;=\; \sum_{i=1}^{d} Z_{(m+1-i)}^{(N)}, \tag{51}$$

To find the PDF of $Y^{(N)}$, we need the joint PDF of $Z_{(m+1-i)}^{(N)}$ for all $1 \le i \le d$. This is a standard question in order statistics.

Proposition 5. *[7] For i.i.d. random variables $X_1, \ldots, X_k$ with PDF $f(x)$ and CDF $F(x)$, the joint PDF of $X_{(i)}, \ldots, X_{(k)}$ where $1 \le i \le k$ is given by*

$$f_{X_{(i)},\ldots,X_{(k)}}(x_i, \ldots, x_k) \;=\; C_k^i f(x_i) \cdots f(x_k)(F(x_i))^{i-1}, \tag{52}$$

where $C_k^i := k!/(i-1)!$.

Thus, in our case, the joint PDF of $Z_{(m+1-i)}^{(N)}$ for all $1 \le i \le d$ is

$$\begin{aligned} &f_{Z_{(m+1-d)},\ldots,Z_{(m)}}(z_{m+1-d}, \ldots, z_m) \\ &\;=\; C_m^{m+1-d} \left(\Phi(z_{m+1-d}) + B(z_{m+1-d})\right)^{m-d} \prod_{i=1}^{d} \left(\varphi(z_{m+i-d}) + A(z_{m+i-d})\right). \end{aligned} \tag{53}$$

By binomial expansion,

$$(\Phi(z_{m+1-d}) + B(z_{m+1-d}))^{m-d} \;=\; \sum_{j=0}^{m-d} \binom{m-d}{j} (\Phi(z_{m+1-d}))^{m-d-j} (B(z_{m+1-d}))^{j}. \tag{54}$$

Note that $B(z_{m+1-d}) = O(N^{-\delta})$ and $0 \le \Phi(x) \le 1$, which we shall constantly use without reference for the remainder of the paper. Then any term where $j > 1$ is at most of the same magnitude as $B(z_{m+1-d})$. We have

$$(\Phi(z_{m+1-d}) + B(z_{m+1-d}))^{m-d} \;=\; (\Phi(z_{m+1-d}))^{m-d} + O(B(z_{m+1-d})). \tag{55}$$

We make a slight abuse of notation here to write the error term above as $B(z_{m+1-d})$, which does not change the order of magnitude of the error term. Hence,

$$\begin{aligned} &f_{Z_{(m+1-d)},\ldots,Z_{(m)}}(z_{m+1-d}, \ldots, z_m) \\ &\;=\; C_m^{m+1-d} \left(\Phi(z_{m+1-d})^{m-d} + B(z_{m+1-d})\right) \prod_{i=1}^{d} \left(\varphi(z_{m+i-d}) + A(z_{m+i-d})\right). \end{aligned} \tag{56}$$

3.2. PDF of $Y^{(N)}$

In this section, we find the PDF of $Y^{(N)}$ as defined in (53). The CDF $F_{Y^{(N)}}(y)$ of $Y^{(N)}$ is given by integrating over the appropriate region in $\mathbb{R}^d$, that is, over all the values $(z_{m+1-d}, \ldots, z_m)$ of $(Z_{(m+1-d)}^{(N)}, \ldots, Z_{(m)}^{(N)})$ that sums to y. We know that $Z_{(m+1-d)}^{(N)}, \ldots Z_{(m)}^{(N)}$ are not independent, i.e., they have to satisfy $Z_{(m+1-d)}^{(N)} \le \cdots \le Z_{(m)}^{(N)}$. In general, we know

that any $1 \le j \le d$, z_{m+j-d} can take value from $z_{m+(j-1)-d}$ to $(y - \sum_{i=1}^{j-1} z_{m+i-d})/(d-j+1)$. The only exception is z_{m+1-d}, whose upper bound is already correctly stated but the lower bound is $-C\sqrt{N}$. Hence, we have

$$\begin{aligned} & F_{Y^{(N)}}(y) \\ &= \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} f_{Z_{(m+1-d)},\ldots,Z_{(m)}}(z_{m+1-d},\ldots,z_m) dz_m \cdots dz_{m+1-d}, \end{aligned} \tag{57}$$

where we use the product of integrals to denote

$$\prod_{j=k}^{d} \int_{z_j}^{y_j} \; := \; \int_{z_k}^{y_k} \cdots \int_{z_j}^{y_j} \cdots \int_{z_d}^{y_d}. \tag{58}$$

We separate out the main term $C_m^{m+1-d} \Phi(z_{m+1-d})^{m-d} \prod_{i=1}^{d} \varphi(z_{m+i-d})$ of the joint PDF (56) and denote it by

$$m_{Y^{(N)}}(z_{m+1-d},\ldots,z_m) \; := \; C_m^{m+1-d} \Phi(z_{m+1-d})^{m-d} \prod_{i=1}^{d} \varphi(z_{m+i-d}). \tag{59}$$

We also denote the error term of the joint PDF (56) by $e_{Y^{(N)}}(z_{m+1-d},\ldots,z_m)$. Let $\mathcal{M}_{Y^{(N)}}(y)$ and $\mathcal{E}_{Y^{(N)}}(y)$ be the main term and the error term of $F_{Y^{(N)}}(y)$. Then

$$\begin{aligned} F_{Y^{(N)}}(y) &= \mathcal{M}_{Y^{(N)}}(y) + \mathcal{E}_{Y^{(N)}}(y) \\ \mathcal{M}_{Y^{(N)}}(y) &= \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} m_{Y^{(N)}}(z_{m+1-d},\ldots,z_m) dz_m \cdots dz_{m+1-d} \\ \mathcal{E}_{Y^{(N)}}(y) &= \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} e_{Y^{(N)}}(z_{m+1-d},\ldots,z_m) dz_m \cdots dz_{m+1-d}. \end{aligned} \tag{60}$$

We are now ready to find the PDF $f_{Y^{(N)}}$ of $Y^{(N)}$. We know that

$$f_{Y^{(N)}}(y) \; = \; \frac{d}{dy} F_{Y^{(N)}}(y) \; = \; \frac{d}{dy} \mathcal{M}_{Y^{(N)}}(y) + \frac{d}{dy} \mathcal{E}_{Y^{(N)}}(y). \tag{61}$$

To differentiate $\mathcal{M}_{Y^{(N)}}(y)$ and $\mathcal{E}_{Y^{(N)}}(y)$, we apply the **Leibniz integral rule** for differentiate under the integral sign.

Lemma 4. *[9] Suppose that $f(x,t)$ is a function such that both $f(x,t)$ and $df(x,t)/dx$ are continuous in t and x in the xt-plane, and $a(x)$ and $b(x)$ are also continuously differentiable. Then*

$$\frac{d}{dx} \int_{a(x)}^{b(x)} f(x,t) dt \; = \; f(x,b(x)) \cdot \frac{d}{dx} b(x) - f(x,a(x)) \cdot \frac{d}{dx} a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t) dt. \tag{62}$$

For the rest of this paper, one can easily check the regularity conditions on $f(x,t)$, $a(x)$, and $b(x)$ each time we apply the Leibniz integral rule, so there will not be explicit mention of the regularity conditions again. We first make some observations on the effect of d/dy on

$$\int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} g(z_{m+1-d},\ldots,z_m)dz_m\cdots dz_{m+1-d} \tag{63}$$

for any Riemann integrable function $g(z_{m+1-d},\ldots,z_m)$. One can check that each application of the Leibniz integral rule to (63) gives two integrals, one of which has the lower bound of the integration coinciding with the upper bound of the integration and thus equals to 0. As a result, we are always left with one integral that interchanges the order of the differentiation and the integration signs. Eventually, (63) becomes

$$\begin{aligned}
&\int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \frac{\partial}{\partial y}\left(\int_{z_{m-1}}^{y-\sum_{i=1}^{d-1} z_{m+i-d}} g(z_{m+1-d},\ldots,z_m)dz_m\right) dz_{m-1}\cdots dz_{m+1-d} \\
&= \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} g\left(z_{m+1-d},\ldots,z_{m-1},y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1}\cdots dz_{m+1-d}.
\end{aligned} \tag{64}$$

Applying (64) to $d\mathcal{M}_{Y^{(N)}}(y)/dy$ and $d\mathcal{E}_{Y^{(N)}}(y)/dy$, where $\mathcal{M}_{Y^{(N)}}(y)$ and $\mathcal{E}_{Y^{(N)}}(y)$ are as defined in (60), we have

$$\begin{aligned}
\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y) &= \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} m_{Y^{(N)}}\left(z_{m+1-d},\ldots,z_{m-1},y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \\
&\quad \cdots dz_{m+1-d} \\
\frac{d}{dy}\mathcal{E}_{Y^{(N)}}(y) &= \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} e_{Y^{(N)}}\left(z_{m+1-d},\ldots,z_{m-1},y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \\
&\quad \cdots dz_{m+1-d}.
\end{aligned} \tag{65}$$

3.3. Outline of problem

We are ready to formulate our problem. We want to show that $\mathfrak{m}_d^{(N)}$ converges to strong Benford behavior, which is equivalent to showing that $\log_B(\mathfrak{m}_d^{(N)})$ converges to being equidistributed mod 1. Hence, in terms of the probability density function $f_{Y^{(N)}}(y)$ of $Y^{(N)} := \log_B(\mathfrak{m}_d^{(N)})/\sqrt{N}$, we can formulate our problem as to showing

$$F_N(a,b) := \sum_{n=-dC\cdot N}^{dC\cdot N-1} \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} f_{Y^{(N)}}(y)dy \approx b-a, \tag{66}$$

for all $(a,b) \subset (0,1)$ for sufficiently large N. Let the main term and the error term of $F_N(a,b)$ be defined respectively as

$$\mathcal{M}_N(a,b) \;:=\; \sum_{n=-dC\cdot N}^{dC\cdot N-1} \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y)dy; \tag{67}$$

$$\mathcal{E}_N(a,b) \;:=\; \sum_{n=-dC\cdot N}^{dC\cdot N-1} \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \frac{d}{dy}\mathcal{E}_{Y^{(N)}}(y)dy. \tag{68}$$

Our goal is to show

$$\mathcal{M}_N(a,b) \;\approx\; b-a, \quad \mathcal{E}_N(a,b) \;\approx\; 0. \tag{69}$$

3.4. Upper bound on error term $\mathcal{E}_N(a,b)$

We want to show that the error term $\mathcal{E}_N(a,b)$ is negligible. We start by giving some bounds on $d\mathcal{E}_{Y^{(N)}}(y)/dy$ from (65). Recall that $e_{Y^{(N)}}(z_{m+1-d},\dots,z_m)$ is the error term of the joint density function (56). Let $[\ell]$ be short hand for $\{1,\dots,\ell\}$. Then

$$\begin{aligned}
&\frac{1}{C_m^{m+1-d}} e_{Y^{(N)}}(z_{m+1-d},\dots,z_m) \\
&\quad= \Phi(z_{m+1-d})^{m-d}\left(\prod_{i=1}^{d}(\varphi(z_{m+i-d})+A(z_{m+i-d})) - \prod_{i=1}^{d}\varphi(z_{m+i-d})\right) \\
&\qquad + B(z_{m+1-d})\prod_{i=1}^{d}(\varphi(z_{m+i-d})+A(z_{m+i-d})) \\
&\quad= \left(\Phi(z_{m+1-d})^{m-d}+B(z_{m+1-d})\right)\left(\sum_{S\subsetneq[d]}\left(\prod_{i\in S}\varphi(z_{m+i-d})\right)\left(\prod_{i\in[d]\setminus S}A(z_{m+i-d})\right)\right) \\
&\qquad + B(z_{m+1-d})\prod_{i=1}^{d}\varphi(z_{m+i-d}) \\
&\ll \sum_{S\subsetneq[d]}\left(\prod_{i\in S}\varphi(z_{m+i-d})\right)\left(\prod_{i\in[d]\setminus S}A(z_{m+i-d})\right) + N^{-\delta}\prod_{i=1}^{d}\varphi(z_{m+i-d}).
\end{aligned} \tag{70}$$

We know from (65) that to bound $d\mathcal{E}_{Y^{(N)}}(y)/dy$, it suffices to bound

$$\begin{aligned}
D_1(y) \;:=\; \sum_{\substack{S\subset[d]\\ S\neq[d]}} \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1}z_{m+i-d}}{d-j+1}} \left(\prod_{i\in S}\varphi(z_{m+i-d})\right) \\
\cdot \left(\prod_{i\in[d]\setminus S}A(z_{m+i-d})\right)\Bigg|_{z_m=y-\sum_{i=1}^{d-1}z_{m+i-d}} dz_{m-1}\cdots dz_{m+1-d}
\end{aligned} \tag{71}$$

as well as

$$D_2(y) \;:=\; N^{-\delta} \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \prod_{i=1}^{d} \varphi(z_{m+i-d}) \Bigg|_{z_m = y - \sum_{i=1}^{d-1} z_{m+i-d}} dz_{m-1} \cdots dz_{m+1-d}, \tag{72}$$

because then we have $d\mathcal{E}_{Y^{(N)}} dy \ll D_1(y) + D_2(y)$. We first look at $D_1(y)$. Suppose that k_S is the largest index such that $k \in [d] \setminus S$. Such an index k exists, since $S \neq [d]$. First, assume that we have S's where $k_S = d$. Then the integrand of (71) becomes

$$\begin{aligned}
&\left(\prod_{i \in S} \varphi(z_{m+i-d}) \right) \left(\prod_{i \in [d] \setminus S} A(z_{m+i-d}) \right) \Bigg|_{z_m = y - \sum_{i=1}^{d-1} z_{m+i-d}} \\
&\quad = \left(\prod_{i \in S} \varphi(z_{m+i-d}) \right) \left(\prod_{i \in [d-1] \setminus S} A(z_{m+i-d}) \right) A\left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) \\
&\quad \ll N^{-\frac{1}{2}-\delta} \left(\prod_{i \in S} \varphi(z_{m+i-d}) \right) \left(\prod_{i \in [d-1] \setminus S} A(z_{m+i-d}) \right),
\end{aligned} \tag{73}$$

where in the last line we use the fact that $A(x) = O(N^{-1/2-\delta})$. Note that the lower and upper bounds on the interval of integration for each integral in (71) are $O(\sqrt{N})$, since $y, z_{m+1-d}, \ldots, z_{m-1}$ are all $O(\sqrt{N})$. Choose a constant C' such that $C'\sqrt{N}$ is an upper bound on the upper bounds on the interval of integration for all the integrals in (71) and $-C'\sqrt{N}$ is a lower bound on the interval of integration for all the integrals in (71). Hence,

$$\begin{aligned}
&\int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \left(\prod_{i \in S} \varphi(z_{m+i-d}) \right) \left(\prod_{i \in [d] \setminus S} A(z_{m+i-d}) \right) \Bigg|_{z_m = y - \sum_{i=1}^{d-1} z_{m+i-d}} \\
&dz_{m-1} \cdots dz_{m+1-d} \\
&\quad \ll N^{-1/2-\delta} \underbrace{\int_{-C'\sqrt{N}}^{C'\sqrt{N}} \cdots \int_{-C'\sqrt{N}}^{C'\sqrt{N}}}_{d-1} \left(\prod_{i \in S} \varphi(z_{m+i-d}) \right) \left(\prod_{i \in [d-1] \setminus S} N^{-1/2-\delta} \right) dz_{m-1} \cdots dz_{m+1-d} \\
&\quad \ll N^{-1/2-\delta-|[d-1]\setminus S|\delta} \prod_{i \in S} \left(\int_{-C'\sqrt{N}}^{C'\sqrt{N}} \varphi(z_{m+i-d}) dz_{m+i-d} \right) \\
&\quad \ll N^{-1/2-(1+|[d-1]\setminus S|)\delta} \;\ll\; N^{-1/2-\delta},
\end{aligned} \tag{74}$$

where in the last line we use the fact that

$$\int_{-C'\sqrt{N}}^{C'\sqrt{N}} \varphi(x) dx \;\ll\; \int_{-\infty}^{\infty} \varphi(x) dx \;=\; 1, \tag{75}$$

which we shall use frequently without reference from now on. Since the number of subsets S of $[d]$ is finite, then (71) gives us

$$D_1(y) \ll N^{-1/2-\delta}. \tag{76}$$

The case when $k_S < d$ is similar, though more care is needed to identify the $O(N^{-1/2-\delta})$ decay, which essentially comes from the $A(z_{m+k_S-d})$ term. With this in mind, using the same method for the case when $k_S = d$ as well as the convolution formula for Gaussian PDF, we can again show that $D_1(y) \ll N^{-1/2-\delta}$. We leave the details to Appendix B.

We now turn to $D_2(y)$ as defined in (72), which is relatively more straightforward to bound. We have

$$D_2(y) \ll N^{-\delta} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{d-1} \prod_{i=1}^{d-1} \varphi(z_{m+i-d}) \varphi\left(y - \sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \cdots dz_{m+1-d}. \tag{77}$$

We recognize that the integral is the convolution of d standard Gaussian density functions. We know that the convolution is the probability density function of sum of d i.i.d. standard Gaussian random variables, which itself is also with mean 0 and variance d. Hence,

$$D_2(y) \ll N^{-\delta} \frac{1}{\sqrt{2\pi d}} e^{-y^2/(2d)}. \tag{78}$$

Thus,

$$\frac{d}{dy}\mathcal{E}_{Y^{(N)}}(y) \ll D_1(y) + D_2(y) \ll N^{-\delta} \frac{1}{\sqrt{2\pi d}} e^{-y^2/(2d)} + N^{-1/2-\delta}. \tag{79}$$

We are now ready to bound the error term $\mathcal{E}_N(a,b)$ from (68):

$$\begin{aligned} \mathcal{E}_N(a,b) &= \sum_{n=-dC\cdot N}^{dC\cdot N-1} \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \frac{d}{dy}\mathcal{E}_{Y^{(N)}}(y) dy \\ &\ll N^{-\delta} \int_{-dC\sqrt{N}}^{dC\sqrt{N}} \frac{1}{\sqrt{2\pi d}} e^{-y^2/(2d)} dy + \int_{-dC\cdot\sqrt{N}}^{dC\cdot\sqrt{N}} N^{-1/2-\delta} dy \ll N^{-\delta}. \end{aligned} \tag{80}$$

3.5. Strategy for remainder of proof

Our next task is to show that the main term $\mathcal{M}_N(a,b)$ defined in (67) satisfies

$$\mathcal{M}_N(a,b) := \sum_{-dC\cdot N}^{dC\cdot N-1} \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y) dy \approx b - a. \tag{81}$$

Our strategy is the following. First, we want to write the integral of $d\mathcal{M}_{Y^{(N)}}(y)/dy$ over each interval $[(a+n)/\sqrt{N}, (b+n)/\sqrt{N}]$ as the sum of a main term $m_n(a,b)$ that is constant with respect to y within the interval and an error term $e_n(a,b)$ for each n. Essentially, this

shows that $d\mathcal{M}_{Y^{(N)}}(y)/dy$ is almost constant within a small interval, incurring an negligible cost that, as we shall prove, does not accumulate, i.e.,

$$\mathcal{M}_{\text{err},N}(a,b) \;=\; \sum_{n=-dC\cdot N}^{dC\cdot N-1} e_n(a,b) \;\approx\; 0. \tag{82}$$

Finally, we prove that the sum of the main term

$$\mathcal{M}_{\text{main},N}(a,b) \;=\; \sum_{n=-dC\cdot N}^{dC\cdot N-1} m_n(a,b) \tag{83}$$

is a Riemann sum that converges to $b-a$. It then follows that

$$\mathcal{M}_N(a,b) \;=\; \mathcal{M}_{\text{main},N}(a,b) + \mathcal{M}_{\text{err},N}(a,b) \;\approx\; b-a. \tag{84}$$

Thus, based on the estimate on the error term in (68),

$$F_N(a,b) \;=\; \mathcal{M}_N(a,b) + \mathcal{E}_N(a,b) \;\approx\; b-a, \tag{85}$$

which establishes the equidistribution result.

3.6. Equidistribution within small interval

In this section, we show that $d\mathcal{M}_{Y^{(N)}}(y)/dy$ is equidistributed within each small interval $[(a+n)/\sqrt{N},(b+n)/\sqrt{N}]$. For each n, we write the integral over $[(a+n)/\sqrt{N},(b+n)/\sqrt{N}]$ in terms of a main term $m_n(a,b)$ that is constant with respect to y and an error term $e_n(a,b)$, i.e.,

$$\int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y)dy \;=\; m_n(a,b) + e_n(a,b), \tag{86}$$

where

$$\begin{aligned} m_n(a,b) \;&:=\; \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \frac{d}{dy}\mathcal{M}_{Y^{(N)}}\left(\frac{n}{\sqrt{N}}\right)dy, \\ e_n(a,b) \;&:=\; \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y) - \frac{d}{dy}\mathcal{M}_{Y^{(N)}}\left(\frac{n}{\sqrt{N}}\right)dy. \end{aligned} \tag{87}$$

Since $d\mathcal{M}_{Y^{(N)}}(n/\sqrt{N})/dy$ is constant, then the main term $m_n(a,b)$ is

$$m_n(a,b) \;=\; \frac{b-a}{\sqrt{N}}\frac{d}{dy}\mathcal{M}_{Y^{(N)}}\left(\frac{n}{\sqrt{N}}\right). \tag{88}$$

We now quantify the error term. We start by providing an upper bound on

$$\left|\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y) - \frac{d}{dy}\mathcal{M}_{Y^{(N)}}\left(\frac{n}{\sqrt{N}}\right)\right|, \tag{89}$$

for $y \in [n/\sqrt{N}, (n+1)/\sqrt{N}] \supset [(a+n)/\sqrt{N}, (b+n)/\sqrt{N}]$. As we shall see, $d\mathcal{M}_{Y^{(N)}}(y)/dy$ is a continuous function over $[n/\sqrt{N}, (n+1)/\sqrt{N}]$ and a differentiable function over $(n/\sqrt{N}, (n+1)/\sqrt{N})$. By the Mean Value Theorem, for any $y_1, y_2 \in [n/\sqrt{N}, (n+1)/\sqrt{N}]$ with $y_1 < y_2$, we have

$$\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y_1) - \frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y_2) \;=\; \frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(c_n)(y_1 - y_2) \tag{90}$$

for some $c_n \in (y_1, y_2)$. Hence

$$\left|\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y_1) - \frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y_2)\right| \;\le\; \left|\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(c_n)\right| |y_1 - y_2| \;\le\; \frac{1}{\sqrt{N}}\left|\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(c_n)\right|. \tag{91}$$

We want to provide an upper on $|d^2\mathcal{M}_{Y^{(N)}}(y)/dy^2|$ over the interval $[n/\sqrt{N}, (n+1)/\sqrt{N}]$. Recall from (65) that

$$\begin{aligned}
\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y) \;&=\; \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} m_{Y^{(N)}}\left(z_{m+1-d}, \ldots, z_{m-1}, y - \sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \\
&\quad \cdots dz_{m+1-d} \\
&=\; \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} (\Phi(z_{m+1-d}))^{m-d} \prod_{i=1}^{d-1} \varphi(z_{m+i-d}) \\
&\quad \cdot \varphi\left(y - \sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \cdots dz_{m+1-d}.
\end{aligned} \tag{92}$$

Following the same procedure as shown in Section 3.2, we can again exchange the order of differentiation and integration. Since the product in the integrand in (92) goes from $j = 2$ to $d - 1$, we need to discuss the cases when $d = 2$ and $d > 2$ separately. If $d = 2$,

$$\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y) \;=\; \int_{-C\sqrt{N}}^{\frac{y}{2}} m_{Y^{(N)}}\left(z_{m-1}, y - z_{m-1}\right) dz_{m-1}. \tag{93}$$

By the Leibniz integral rule and (59),

$$\begin{aligned}
&\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \\
&=\; \frac{1}{2} m_{Y^{(N)}}\left(\frac{y}{2}, \frac{y}{2}\right) + \int_{-C\sqrt{N}}^{\frac{y}{2}} \frac{\partial}{\partial y}\left(m_{Y^{(N)}}(z_{m-1}, y - z_{m-1})\right) dz_{m-1} \\
&\ll\; \varphi\left(\frac{y}{2}\right) + \int_{-C\sqrt{N}}^{\frac{y}{2}} (\Phi(z_{m-1}))^{m-2} \varphi(z_{m-1}) \cdot (-1) \cdot (y - z_{m-1}) \varphi(y - z_{m-1}) dz_{m-1}.
\end{aligned} \tag{94}$$

We know that y is an upper bound on the value of $z_{m-1} + z_m$, then $y \ge -2C\sqrt{N}$, and $y/2 \ge -C\sqrt{N}$. We now break into cases when $y \le 0$ and $y \ge 0$. When $y \le 0$, $\varphi(z_{m-1})$ is at

most $\varphi(y/2)$ on $[-C\sqrt{N}, y/2]$. So (94) becomes

$$\begin{aligned}\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) &\ll \varphi\left(\frac{y}{2}\right) + \varphi\left(\frac{y}{2}\right)\int_{-\infty}^{\infty}|y - z_{m-1}|\varphi(y - z_{m-1})dz_{m-1} \\ &= \varphi\left(\frac{y}{2}\right) + \varphi\left(\frac{y}{2}\right)\int_{-\infty}^{\infty}|z_{m-1}|\varphi(z_{m-1})dz_{m-1} \ll \varphi\left(\frac{y}{2}\right), \end{aligned} \tag{95}$$

where in the last line we use the fact that the integral $\int_{-\infty}^{\infty}|x|\varphi(x)dx$ is the expected value of the absolute value of a standard normal random variable, which is finite. When $y \geq 0$, we have

$$y - z_{m-1} \geq y - \frac{y}{2} = \frac{y}{2} \geq 0. \tag{96}$$

Hence, (94) becomes

$$\begin{aligned}\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) &\ll \varphi\left(\frac{y}{2}\right) + \int_{-C\sqrt{N}}^{\frac{y}{2}}(y - z_{m-1})\varphi(y - z_{m-1})dz_{m-1} \\ &= \varphi\left(\frac{y}{2}\right) + \varphi(y - z_{m-1})\Big|_{z_{m-1}=-C\sqrt{N}}^{z_{m-1}=\frac{y}{2}} \ll \varphi\left(\frac{y}{2}\right) + \varphi\left(y + C\sqrt{N}\right), \end{aligned} \tag{97}$$

where in the second line we employ the fact that $\int(y-x)\varphi(y-x)dx = \varphi(y-x)$. Thus, if $d = 2$, regardless of whether $y \leq 0$ or $y \geq 0$, $d^2\mathcal{M}_{Y^{(N)}}(y)/dy^2$ is on the order of $\varphi(y/2)+\varphi(y+C\sqrt{N})$. This concludes the estimate of $d^2\mathcal{M}_{Y^{(N)}}(y)/dy^2$ for $d = 2$.

If $d > 2$, we repeatedly apply Leibniz integral rule as we have demonstrated in Subsection 3.4 to obtain

$$\begin{aligned}&\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \\ &= \int_{-C\sqrt{N}}^{\frac{y}{d}}\prod_{j=2}^{d-2}\int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1}z_{m+i-d}}{d-j+1}}\frac{\partial}{\partial y}\Bigg(\int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2}z_{m+i-d}}{2}} m_{Y^{(N)}}\left(z_{m+1-d}, \ldots, z_{m-1}, y - \sum_{i=1}^{d-1}z_{m+i-d}\right) \\ &\quad dz_{m-1}\Bigg)dz_{m-2}\cdots dz_{m+1-d} \\ &\ll \int_{-C\sqrt{N}}^{\frac{y}{d}}\prod_{j=2}^{d-2}\int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1}z_{m+i-d}}{d-j+1}}(\Phi(z_{m+1-d}))^{m-d}\left(\prod_{i=1}^{d-2}\varphi(z_{m+i-d})\right)\left(\varphi\left(\frac{y-\sum_{i=1}^{d-2}z_{m+i-d}}{2}\right)\right)^2 \\ &\quad + \Bigg(\int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2}z_{m+i-d}}{2}}(\Phi(z_{m+1-d}))^{m-d}\left(\prod_{i=1}^{d-1}\varphi(z_{m+i-d})\right)(-1)\left(y - \sum_{i=1}^{d-1}z_{m+i-d}\right) \\ &\quad \cdot\varphi\left(y - \sum_{i=1}^{d-1}z_{m+i-d}\right)dz_{m-1}\Bigg)dz_{m-2}\cdots dz_{m+1-d}. \end{aligned} \tag{98}$$

Now, similar to the case when $d = 2$, we estimate the second integrand in (98) by breaking down into cases when $(y - \sum_{i=1}^{d-2}z_{m+i-d})/2 \leq 0$ and $(y - \sum_{i=1}^{d-2}z_{m+i-d}) \geq 0$. We leave the

details to Appendix C. Hence, we have $d^2\mathcal{M}_{Y^{(N)}}(y)/dy^2$ in (98) is bounded above by

$$\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-2}\left(\prod_{i=1}^{d-2}\varphi(z_{m+i-d})\right)\varphi\left(\frac{y-\sum_{i=1}^{d-2}z_{m+i-d}}{2}\right)dz_{m-2}\cdots dz_{m+1-d}$$
$$+\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-2}\left(\prod_{i=1}^{d-2}\varphi(z_{m+i-d})\right)\varphi\left(y-\left(\sum_{i=1}^{d-2}z_{m+i-d}\right)-z_{m-2}\right)dz_{m-2}\cdots dz_{m+1-d}. \tag{99}$$

We first obtain some estimate on the first integral of (99). The idea behind estimating this integral is recursive uses of change of variables that lead to consistent cancellation. We first do the change of variable $z_{m-2}\mapsto z_{m-2}+(y-\sum_{i=1}^{d-3}z_{m+i-d})/5$. We have

$$\begin{aligned}&\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-2}\left(\prod_{i=1}^{d-2}\varphi(z_{m+i-d})\right)\varphi\left(\frac{y-\sum_{i=1}^{d-2}z_{m+i-d}}{2}\right)dz_{m-2}\cdots dz_{m+1-d}\\ &\ll\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-2}e^{-\frac{\sum_{i=1}^{d-3}z_{m+i-d}^2}{2}}\cdot e^{-\frac{\left(z_{m-2}+\left(y-\sum_{i=1}^{d-3}z_{m+i-d}\right)/5\right)^2}{2}}\\ &\quad\cdot e^{-\frac{\frac{1}{4}\left(-z_{m-2}+4\left(y-\sum_{i=1}^{d-3}z_{m+i-d}\right)/5\right)^2}{2}}dz_{m-2}\cdots dz_{m+1-d}.\end{aligned} \tag{100}$$

Notice that in the expansion of the exponents, we get a cancellation of the term $2z_{m-2}((y-\sum_{i=1}^{d-3}z_{m+i-d})/5)$. Hence, (100) is bounded above by

$$\begin{aligned}&\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-3}e^{-\frac{\sum_{i=1}^{d-3}z_{m+i-d}^2}{2}}\cdot e^{-\frac{\frac{1}{5}\left(y-\sum_{i=1}^{d-3}z_{m+i-d}\right)^2}{2}}\left(\int_{-\infty}^{\infty}e^{-\frac{\frac{5}{4}z_{m-2}^2}{2}}dz_{m-2}\right)dz_{m-3}\cdots dz_{m+1-d}\\ &\ll\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-3}e^{-\frac{\sum_{i=1}^{d-3}z_{m+i-d}^2}{2}}\cdot e^{-\frac{\frac{1}{5}\left(y-\sum_{i=1}^{d-3}z_{m+i-d}\right)^2}{2}}dz_{m-3}\cdots dz_{m+1-d}\end{aligned} \tag{101}$$

We see that the estimate in (101) has a similar structure, which allows us to use induction to estimate the interval. For an arbitrary k such that $1\le k\le d-2$, we want to look at

$$\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k}e^{-\frac{\sum_{i=1}^{k}z_{m+i-d}^2}{2}}\cdot e^{-\frac{A_k(y-\sum_{i=1}^{k}z_{m+i-d})^2}{2}}dz_{m+k-d}\cdots dz_{m+1-d}, \tag{102}$$

where $A_k>0$. Using the change of variable $z_{m+k-d}\mapsto z_{m+k-d}+\frac{A_k}{1+A_k}(y-\sum_{i=1}^{k-1}z_{m+i-d})$,

(102) is bounded above by

$$\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k} e^{-\frac{\sum_{i=1}^{k-1} z_{m+i-d}^2}{2}} \cdot e^{-\frac{\left(z_{m+k-d}+\frac{A_k}{1+A_k}\left(y-\sum_{i=1}^{k-1} z_{m+i-d}\right)\right)^2}{2}}$$

$$\cdot e^{-\frac{A_k\left(-z_{m+k-d}+\frac{1}{1+A_k}\left(y-\sum_{i=1}^{k-1} z_{m+i-d}\right)\right)^2}{2}} dz_{m+k-d}\cdots dz_{m+1-d}. \tag{103}$$

Expanding the exponents also gives a cancellation of the term

$$\frac{2A_k}{1+A_k} z_{m+k-d}\left(y-\sum_{i=1}^{k-1} z_{m+i-d}\right). \tag{104}$$

Hence, (102) is bounded above by

$$\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k} e^{-\frac{\sum_{i=1}^{k-1} z_{m+i-d}^2}{2}} \cdot e^{-\frac{A_{k-1}\left(y-\sum_{i=1}^{k-1} z_{m+i-d}\right)^2}{2}} e^{-\frac{(1+A_k) z_{m+k-d}^2}{2}} dz_{m+k-d}\cdots dz_{m+1-d}, \tag{105}$$

where

$$A_{k-1} \;:=\; \frac{A_k^2}{1+A_k^2}+\frac{A_k}{(1+A_k)^2}. \tag{106}$$

Note that since $A_k > 0$, then $A_{k+1} > 0$. Hence, (102) becomes bounded above by

$$\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k-1} e^{-\frac{\sum_{i=1}^{k-1} z_{m+i-d}^2}{2}} \cdot e^{-\frac{A_{k-1}\left(y-\sum_{i=1}^{k-1} z_{m+i-d}\right)^2}{2}} \left(\int_{-\infty}^{\infty} e^{-\frac{(1+A_k) z_{m+k-d}^2}{2}} dz_{m+k-d}\right)$$

$$dz_{m+(k-1)-d}\cdots dz_{m+1-d}$$

$$\ll \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k-1} e^{-\frac{\sum_{i=1}^{k-1} z_{m+i-d}^2}{2}} \cdot e^{-\frac{A_{k-1}\left(y-\sum_{i=1}^{k-1} z_{m+i-d}\right)^2}{2}} dz_{m+(k-1)-d}\cdots dz_{m+1-d}, \tag{107}$$

Thus, by induction, (100) becomes bounded above by

$$e^{-(A_0 y^2)/2}, \tag{108}$$

where $A_0 > 0$ is recursively defined by (106) with $A_{d-2} = 1/4$. Similarly, we can show by induction that the second integral in (99) has the estimate

$$\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-2} \left(\prod_{i=1}^{d-2}\varphi(z_{m+i-d})\right)\left(\varphi\left(y-\left(\sum_{i=1}^{d-2} z_{m+i-d}\right)-z_{m-2}\right)\right) dz_{m-2}\cdots dz_{m+1-d}$$

$$\ll e^{-(B_0 y^2)/2} \tag{109}$$

for some $B_0 > 0$. Hence, if $d > 2$, (99) becomes

$$\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \;\ll\; e^{-(A_0y^2)/2} + e^{-(B_0y^2)/2} \;\ll\; e^{-(C_0y^2)/2}, \tag{110}$$

where $C_0 := \min\{A_0, B_0\}$. Recall from (97) that when $d = 2$,

$$\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \;\ll\; \varphi\left(\frac{y}{2}\right) + \varphi\left(y + C\sqrt{N}\right). \tag{111}$$

Thus, regardless of whether $d = 2$ or $d > 2$,

$$\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \;\ll\; e^{-\frac{D_0y^2}{2}} + e^{-\frac{D_0\left(y+C\sqrt{N}\right)^2}{2}}, \tag{112}$$

where $D_0 := \min\{C_0, 1/4\}$. Note that $e^{-D_0y^2/2}$ and $\varphi(\sqrt{D_0}(y + C\sqrt{N}))$ each has only one global extreme, at $y = 0$ and $y = -C\sqrt{N}$ respectively. Hence, locally on $[n/\sqrt{N}, (n+1)/\sqrt{N}]$ the functions are monotonic and can be bounded above by the sum of its values at the two end points. Hence, for all n,

$$\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \;\ll\; e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}}, \tag{113}$$

for all $y \in [n/\sqrt{N}, (n+1)/\sqrt{N}]$. Thus, returning to (91), for all $y_1, y_2 \in [n/\sqrt{N}, (n+1)/\sqrt{N}]$,

$$\begin{aligned}&\left|\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y_1) - \frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y_2)\right| \\ &\quad\ll\; \frac{1}{\sqrt{N}}\left(e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}}\right).\end{aligned} \tag{114}$$

Thus, the error term $e_n(a, b)$ as defined in (87) is bounded by

$$\begin{aligned}e_n(a,b) \;&\le\; \int_{\frac{a+n}{\sqrt{N}}}^{\frac{b+n}{\sqrt{N}}} \left|\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y) - \frac{d}{dy}\mathcal{M}_{Y^{(N)}}\left(\frac{n}{\sqrt{N}}\right)\right| dy \\ &\ll\; \left(\frac{1}{\sqrt{N}}\right)^2\left(e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}} + e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}}\right).\end{aligned} \tag{115}$$

3.7. Upper bound on error term $\mathcal{M}_{\mathrm{err},N}(a, b)$

In this section, we want to prove that the error term of the main term $\mathcal{M}_N(a, b)$ is small, i.e.,

$$\mathcal{M}_{\mathrm{err,N}}(a,b) \;=\; \sum_{n=-dC\cdot N}^{dC\cdot N-1} e_n(a,b) \;\approx\; 0. \tag{116}$$

Based on the estimate $e_n(a,b)$ in (115), we can pull out one of the $1/\sqrt{N}$ factors in the bracket and obtain

$$\begin{aligned}\mathcal{M}_{\mathrm{err},N}(a,b) &= \sum_{n=-dC\cdot N}^{dC\cdot N-1} e_n(a,b)\\ &\ll \frac{1}{\sqrt{N}}\cdot\left(\frac{1}{\sqrt{N}}\sum_{n=-dC\cdot N}^{dC\cdot N-1}\left(e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}\right)^2}{2}}+e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}\right)^2}{2}}+e^{-\frac{D_0\left(\frac{n}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}}+e^{-\frac{D_0\left(\frac{n+1}{\sqrt{N}}+C\sqrt{N}\right)^2}{2}}\right)\right).\end{aligned} \tag{117}$$

We observe that the above in the parentheses are Riemann sums for $e^{-D_0x^2/2}$ and $e^{-D_0(x+C\sqrt{N})^2/2}$ on $(-\infty,\infty)$, which are both finite. Hence

$$\mathcal{M}_{\mathrm{err},N}(a,b) \ll \frac{1}{\sqrt{N}}. \tag{118}$$

We have thus established that the error term $\mathcal{M}_{\mathrm{err},N}(a,b)$ of the main term $\mathcal{M}_N(a,b)$ is negligible.

3.8. Evaluation of main term $\mathcal{M}_{\mathrm{main},N}(a,b)$

Finally, we establish the main term $\mathcal{M}_{\mathrm{main},N}(a,b)$ of the main term $\mathcal{M}_N(a,b)$. From (83) and (88), we see that

$$\begin{aligned}\mathcal{M}_{\mathrm{main},N}(a,b) &= \sum_{n=-dC\cdot N}^{dC\cdot N-1} m_n(a,b)\\ &= (b-a)\cdot\left(\frac{1}{\sqrt{N}}\sum_{n=-dC\cdot N}^{dC\cdot N-1}\frac{d}{dy}\mathcal{M}_{Y^{(N)}}\left(\frac{n}{\sqrt{N}}\right)\right).\end{aligned} \tag{119}$$

Since the term in the parentheses above is a Riemann sum for $\int_{-\infty}^{\infty}\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y)dy$, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{n=-2C\cdot N}^{2C\cdot N-1}\frac{d}{dy}\mathcal{M}_{Y^{(N)}}\left(\frac{n}{\sqrt{N}}\right) &= \int_{-\infty}^{\infty}\frac{d}{dy}\mathcal{M}_{Y^{(N)}}(y)dy+o(1)\\ &= \mathcal{M}_{Y^{(N)}}(\infty)-\mathcal{M}_{Y^{(N)}}(-\infty)+o(1).\end{aligned} \tag{120}$$

Hence the main term $\mathcal{M}_{\mathrm{main},N}(a,b)$ becomes

$$\mathcal{M}_{\mathrm{main},N}(a,b) = (b-a)(\mathcal{M}_{Y^{(N)}}(\infty)-\mathcal{M}_{Y^{(N)}}(-\infty))+o(1). \tag{121}$$

Recall that

$$\mathcal{M}_{Y^{(N)}}(y) = \int_{-C\sqrt{N}}^{\frac{y}{d}}\prod_{j=2}^{d}\int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1}z_{m+i-d}}{d-j+1}} m_{Y^{(N)}}(z_{m+1-d},\dots,z_m)dz_m\cdots dz_{m+1-d}. \tag{122}$$

To simplify our calculation, we want to extend the interval of integration of the outer integral to $-\infty$. We do so by showing that the tail of the integral is small, i.e.,

$$T_N \;:=\; \int_{-\infty}^{-C\sqrt{N}} \prod_{j=2}^{d} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} m_{Y^{(N)}}(z_{m+1-d},\ldots,z_m)dz_m\cdots dz_{m+1-d} \;\approx\; 0. \tag{123}$$

We have

$$\begin{aligned} T_N \;&\ll\; \int_{-\infty}^{-C\sqrt{N}} \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-1} (\Phi(z_{m+1-d}))^{m-d} \prod_{i=1}^{d} \varphi(z_{m+i-d})dz_m\cdots dz_{m+1-d} \\ &\ll\; \int_{-\infty}^{-C\sqrt{N}} \varphi(z_{m+1-d})dz_{m+1-d}, \end{aligned} \tag{124}$$

Hence, we have reduced the problem to showing the smallness of the Gaussian tail. We first make a definition.

Definition 12. *For functions $f(x)$ and $g(x)$, we say that $f(x)=\Theta(g(x))$ if $f(x)=O(g(x))$ and $g(x)=O(f(x))$.*

The following result is a straightforward calculation that provides an upper bound on the Gaussian tail.

Proposition 1. *[13] Suppose that $g:\mathbb{R}\to\mathbb{R}_{\geq 0}$ is a function such that $g(N)=\Theta(N^{\epsilon'})$ for some $\epsilon'>0$. Then*

$$\int_{-\infty}^{-g(N)} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}dx \;=\; \int_{g(N)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}dx \;\ll\; e^{-g(N)/2}. \tag{125}$$

Back to (124), by Proposition 1

$$T_N \;\ll\; e^{-C\sqrt{N}}. \tag{126}$$

Thus, (122) becomes

$$\mathcal{M}_{Y^{(N)}}(y) \;=\; \int_{-\infty}^{\frac{y}{d}} \prod_{j=2}^{d} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} m_{Y^{(N)}}(z_{m+1-d},\ldots,z_m)dz_m\cdots dz_{m+1-d} + O\left(e^{-C\sqrt{N}}\right). \tag{127}$$

Now, by Proposition 5,

$$m_{Y^{(N)}}(z_{m+1-d},\ldots,z_m) \;:=\; C_m^{m+1-d} \left(\Phi(z_{m+1-d})\right)^{m-d} \prod_{i=1}^{d} \varphi(z_{m+i-d}) \tag{128}$$

is the joint PDF of $W_{(m+1-d)}, \ldots, W_{(m)}$, where $W_1, \ldots, W_m$ are i.i.d. $\sim N(0,1)$. Thus, following our derivation in Subsection 3.2,

$$H(y) \;:=\; \int_{-\infty}^{\frac{y}{d}} \prod_{j=2}^{d} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} m_{Y^{(N)}}(z_{m+1-d}, \ldots, z_m) dz_m \cdots dz_{m+1-d} \tag{129}$$

is the CDF of $\sum_{i=1}^{d} W_{(m+i-d)}$. We know that for a cumulative density function $H(y)$, $H(-\infty) = 0$ and $H(\infty) = 1$. Hence,

$$\begin{aligned} \mathcal{M}_{Y^{(N)}}(-\infty) \;&=\; H(-\infty) + O(e^{-C\sqrt{N}}) \;=\; O\left(e^{-C\sqrt{N}}\right) \\ \mathcal{M}_{Y^{(N)}}(\infty) \;&=\; H(\infty) + O(e^{-C\sqrt{N}}) \;=\; 1 + O\left(e^{-C\sqrt{N}}\right). \end{aligned} \tag{130}$$

Thus, substituting these two estimates into (121) yields

$$\mathcal{M}_{\text{main},N}(a,b) \;=\; (b-a)\left(1 + O\left(e^{-C\sqrt{N}}\right)\right) + o(1)) \;=\; (b-a) + o(1). \tag{131}$$

Returning to (85) and (84), combined with the estimate for the error term $\mathcal{E}_N(a,b)$ in (80) and the error term $\mathcal{M}_{\text{err},N}(a,b)$ of $\mathcal{M}_N(a,b)$ in (118), we have

$$\begin{aligned} F_N(a,b) \;&=\; \mathcal{M}_N(a,b) + \mathcal{E}_N(a,b) \\ &=\; \mathcal{M}_{\text{main},N}(a,b) + \mathcal{M}_{\text{err},N}(a,b) + \mathcal{E}_N(a,b) \\ &=\; (b-a) + o(1) + O\left(\frac{1}{\sqrt{N}}\right) + O(N^{-\delta}) \\ &=\; (b-a) + o(1). \end{aligned} \tag{132}$$

We conclude that $\log_B(\mathfrak{m}_d^{(N)})$ converges to being equidistributed mod 1, and therefore by Uniform Distribution Characterization $\mathfrak{m}_d^{(N)}$ converges to strong Benford behavior. □

4. Funding

This work was partially supported by Williams College Summer Science Program Research Fellowship, the Finnerty Fund, and NSF Grant DMS2241623.

Appendix A. Proof of Proposition 3

In this section, we prove Proposition 3 which provides a quantitative bound on the difference among probabilities within an interval. Let us first recall the statement of Proposition 3.

Proposition 3. *For $\ell \le \sqrt{k(N)}/2$,*

$$\left| \binom{k(N)}{k_{1,\ell}} - \binom{k(N)}{k_{1,\ell+1}} \right| \;\le\; O\left(\binom{k(N)}{k_{1,\ell}} \cdot N^{-\frac{3\epsilon}{10}} \right). \tag{A.1}$$

Proof. We first factor out $\binom{k(N)}{k_{1,\ell}}$ from the difference:

$$\begin{aligned}
&\binom{k(N)}{k_{1,\ell}} - \binom{k(N)}{k_{1,\ell+1}} \;=\; \binom{k(N)}{\frac{k(N)}{2}+\ell N^\delta} - \binom{k(N)}{\frac{k(N)}{2}+(\ell+1)N^\delta} \\
&= \frac{k(N)!}{\left(\frac{k(N)}{2}+\ell N^\delta\right)!\left(\frac{k(N)}{2}-\ell N^\delta\right)!} - \frac{k(N)!}{\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)!\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)} \\
&= \frac{k(N)!\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)!\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)! - k(N)!\left(\frac{k(N)}{2}+\ell N^\delta\right)!\left(\frac{k(N)}{2}-\ell N^\delta\right)!}{\left(\frac{k(N)}{2}+\ell N^\delta\right)!\left(\frac{k(N)}{2}-\ell N^\delta\right)!\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)!\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)!} \\
&= \frac{k(N)!}{\left(\frac{k(N)}{2}+\ell N^\delta\right)!\left(\frac{k(N)}{2}-\ell N^\delta\right)!} \\
&\cdot \frac{\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)! - \left(\frac{k(N)}{2}+\ell N^\delta\right)!\left(\frac{k(N)}{2}-\ell N^\delta\right)!}{\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)!\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)!} \\
&= \binom{k(N)}{k_{1,\ell}}\left(1 - \frac{\left(\frac{k(N)}{2}+\ell N^\delta\right)!\left(\frac{k(N)}{2}-\ell N^\delta\right)!}{\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)!\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)!}\right). \qquad \text{(A.2)}
\end{aligned}$$

Now, we analyze the term

$$\alpha_{\ell,N} \;:=\; \frac{\left(\frac{k(N)}{2}+\ell N^\delta\right)!\left(\frac{k(N)}{2}-\ell N^\delta\right)!}{\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)!\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)!}. \qquad \text{(A.3)}$$

We want to show that $\alpha_{\ell,N} \to 1$ as $N \to \infty$, so that the difference in (A.2) is asymptotically much smaller than the main term $\binom{k(N)}{k_{1,\ell}}$. We have

$$\begin{aligned}
\frac{\left(\frac{k(N)}{2}-(\ell+1)N^\delta\right)^{N^\delta}}{\left(\frac{k(N)}{2}+(\ell+1)N^\delta\right)^{N^\delta}} \;&\le\; \alpha_{\ell,N} \;\le\; \frac{\left(\frac{k(N)}{2}-\ell N^\delta\right)^{N^\delta}}{\left(\frac{k(N)}{2}+\ell N^\delta\right)^{N^\delta}} \\
\left(1 - \frac{4(\ell+1)N^\delta}{k(N)+2(\ell+1)N^\delta}\right)^{N^\delta} \;&\le\; \alpha_{\ell,N} \;\le \left(1 - \frac{4\ell N^\delta}{k(N)+2\ell N^\delta}\right)^{N^\delta}. \qquad \text{(A.4)}
\end{aligned}$$

Since $\ell \le \sqrt{k(N)}/2$, then

$$0 \;\le\; \frac{4\ell N^\delta}{k(N)+2\ell N^\delta} \;\le\; \frac{4\ell N^\delta}{k(N)} \;\le\; \frac{2N^\delta}{\sqrt{k(N)}}. \qquad \text{(A.5)}$$

Since $k(N) \ge N^\epsilon$, and $\delta \in (0, \epsilon/10)$, then

$$0 \;\le\; \frac{2N^{2\delta}}{\sqrt{k(N)}} \;\le\; \frac{2N^{\epsilon/10}}{N^{\epsilon/2}} \;=\; 2N^{-2\epsilon/5} \;=\; O(N^{-2\epsilon/5}). \qquad \text{(A.6)}$$

Similarly, we also have

$$\frac{4(\ell+1)N^\delta}{k(N)+2(\ell+1)N^\delta} \;=\; O(N^{-2\epsilon/5}). \tag{A.7}$$

Hence, for sufficiently large N,

$$\begin{aligned} 0 \;&<\; 1-\frac{4(\ell+1)N^\delta}{k(N)+2(\ell+1)N^\delta} \;<\; 1 \\ 0 \;&<\; 1-\frac{4\ell N^\delta}{k(N)+2\ell N^\delta} \;<\; 1. \end{aligned} \tag{A.8}$$

Returning to (A.4), given (A.8), we have

$$\left(1-\frac{4(\ell+1)N^\delta}{k(N)+2(\ell+1)N^\delta}\right)^{N^\delta} \;\le\; \alpha_{\ell,N} \;\le\; \left(1-\frac{4\ell N^\delta}{k(N)+2\ell N^\delta}\right)^{N^\delta}. \tag{A.9}$$

Using binomial expansion,

$$\sum_{j=1}^{N^\delta}\binom{N^\delta}{j}(-1)^j\left(\frac{4(\ell+1)N^\delta}{k(N)+2(\ell+1)N^\delta}\right)^j \;\le\; \alpha_{\ell,N}-1 \;\le\; \sum_{j=1}^{N^\delta}\binom{N^\delta}{j}(-1)^j\left(\frac{4\ell N^\delta}{k(N)+2\ell N^\delta}\right)^j. \tag{A.10}$$

We first bound the right sum in (A.4). Using the assumption that $\ell \le \sqrt{k(N)}/2$, we have

$$\begin{aligned} \left|\sum_{j=1}^{N^\delta}\binom{N^\delta}{j}(-1)^j\left(\frac{4\ell N^\delta}{k(N)+2\ell N^\delta}\right)^j\right| \;&\le\; \sum_{j=1}^{N^\delta} N^{j\delta}\left(\frac{4\ell N^\delta}{k(N)}\right)^j \\ &\le\; \sum_{j=1}^{N^\delta} N^{j\delta}\left(\frac{2\sqrt{k(N)}N^\delta}{k(N)}\right)^j \\ &=\; \sum_{j=1}^{N^\delta}\left(\frac{2N^{2\delta}}{\sqrt{k(N)}}\right)^j \\ &=\; \frac{2N^{2\delta}}{\sqrt{k(N)}}\cdot\frac{1-\left(\frac{2N^{2\delta}}{\sqrt{k(N)}}\right)^{N^\delta}}{1-\frac{2N^{2\delta}}{\sqrt{k(N)}}}, \end{aligned} \tag{A.11}$$

where on the last line we use the geometric series formula. Since $k(N) \ge N^\epsilon$, and $\delta \in (0, \epsilon/10)$, then

$$0 \;\le\; \frac{2N^{2\delta}}{\sqrt{k(N)}} \;\le\; \frac{2N^{\epsilon/5}}{N^{\epsilon/2}} \;=\; 2N^{-3\epsilon/10}. \tag{A.12}$$

Hence

$$\left|\sum_{j=1}^{N^\delta}\binom{N^\delta}{j}(-1)^j\left(\frac{4\ell N^\delta}{k(N)+2\ell N^\delta}\right)^j\right| \;\le\; 2N^{-3\epsilon/10}\cdot\frac{1}{1-2N^{-3\epsilon/10}}$$
$$=\; O\left(N^{-3\epsilon/10}\right). \tag{A.13}$$

Similarly, for the left sum in (A.4), we also have

$$\left|\sum_{j=1}^{N^\delta}\binom{N^\delta}{j}(-1)^j\left(\frac{4(\ell+1)N^\delta}{k(N)+2(\ell+1)N^\delta}\right)^j\right| \;=\; O\left(N^{-3\epsilon/10}\right). \tag{A.14}$$

Applying (A.13) and (A.14) to (A.4), we get

$$|\alpha_{\ell,N}-1| \;=\; O\left(N^{-3\epsilon/10}\right). \tag{A.15}$$

Thus, substituting the estimate (A.15) back to (A.2) gives us

$$\left|\binom{k(N)}{k_{1,\ell}}-\binom{k(N)}{k_{1,\ell+1}}\right| \;\le\; O\left(\binom{k(N)}{k_{1,\ell}}\cdot N^{-3\epsilon/10}\right). \tag{A.16}$$

□

Appendix B. Case for $k_S < d$

In this appendix, we want to show that $D_1(y) \ll N^{-1/2-\delta}$ when $k_S < d$. Recall that S is a proper subset of $[d]$ and k_S is the largest index such that $k \in [d]\setminus S$, and that $D_1(y)$ is defined in (71) to be

$$D_1(y) \;:=\; \sum_{\substack{S\subset[d]\\ S\neq[d]}}\int_{-C\sqrt{N}}^{\frac{y}{d}}\prod_{j=2}^{d-1}\int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1}z_{m+i-d}}{d-j+1}}\left(\prod_{i\in S}\varphi(z_{m+i-d})\right)$$
$$\cdot\left(\prod_{i\in[d]\setminus S}A(z_{m+i-d})\right)\Bigg|_{z_m=y-\sum_{i=1}^{d-1}z_{m+i-d}} dz_{m-1}\cdots dz_{m+1-d}. \tag{B.1}$$

Let $[\ell_1;\ell_2]$ denote $\{\ell_1,\dots,\ell_2\}$ if ℓ_1,ℓ_2 are integers such that $\ell_1\le\ell_2$, and let it be $\emptyset$ otherwise. Then the integrand of (71) becomes

$$\left(\prod_{i\in S}\varphi(z_{m+i-d})\right)\left(\prod_{i\in[d]\setminus S}A(z_{m+i-d})\right)\Bigg|_{z_m=y-\sum_{i=1}^{d-1}z_{m+i-d}}$$
$$=\left(\prod_{i\in S\setminus[k_S+1;d]}\varphi(z_{m+i-d})\right)\left(\prod_{i\in[d]\setminus(S\cup\{k_S\})}A(z_{m+i-d})\right)\cdot A(z_{m+k_S-d})$$
$$\cdot\left(\prod_{i\in[k_S+1;d-1]}\varphi(z_{m+i-d})\right)\varphi\left(y-\sum_{i=1}^{d-1}z_{m+i-d}\right). \tag{B.2}$$

Hence,

$$\int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \left(\prod_{i\in S} \varphi(z_{m+i-d})\right) \left(\prod_{i\in[d]\setminus S} A(z_{m+i-d})\right) \Bigg|_{z_m = y-\sum_{i=1}^{d-1} z_{m+i-d}}$$
$$dz_{m-1}\cdots dz_{m+1-d}$$
$$= \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \left(\prod_{i\in S\setminus[k_S+1;d]} \varphi(z_{m+i-d})\right) \left(\prod_{i\in[d]\setminus(S\cup\{k_S\})} A(z_{m+i-d})\right)$$
$$\cdot A(z_{m+k_S-d}) \cdot \left(\prod_{i=k_S+1}^{d-1} \varphi(z_{m+i-d})\right) \cdot \varphi\left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1}\cdots dz_{m+1-d}. \quad \text{(B.3)}$$

To bound (B.3), we first give some estimate on the following

$$\prod_{j=k_S+1}^{d-1} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \left(\prod_{i=k_S+1}^{d-1} \varphi(z_{m+i-d})\right) \cdot \varphi\left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1}\cdots dz_{m+(k_S+1)-d}$$
$$\ll \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-k_S-1} \left(\prod_{i=k_S+1}^{d-1} \varphi(z_{m+i-d})\right) \cdot \varphi\left(\left(y-\sum_{i=1}^{k_S} z_{m+i-d}\right) - \sum_{i=k_S+1}^{d-1} z_{m+i-d}\right) dz_{m-1}$$
$$\cdots dz_{m+(k_S+1)-d}$$
$$\ll \frac{1}{\sqrt{2\pi(d-k_S)}} e^{-\frac{\left(y-\sum_{i=1}^{k_S} z_{m+i-d}\right)^2}{2(d-k_S)}}, \quad \text{(B.4)}$$

where in the last line we use the fact that the second last line is exactly the convolution of $d-k_S$ standard Gaussian density function evaluated at $y-\sum_{i=1}^{k_S} z_{m+i-d}$, which is exactly the probability density function of sum of $d-k_S$ independent standard Gaussian random variables and thus is itself also Gaussian with mean 0 and variance $d-k_S$, evaluated at

$y - \sum_{i=1}^{k_S} z_{m+i-d}$. Hence, (B.3) is bounded above by

$$\begin{aligned}
&\int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{k_S-1} \int_{z_m+(j-1)-d}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \left(\prod_{i\in S\setminus[k_S+1;d]} \varphi(z_{m+i-d})\right)\left(\prod_{i\in[d]\setminus(S\cup\{k_S\})} A(z_{m+i-d})\right) \\
&\quad \int_{z_m+(k_S-1)-d}^{\frac{y-\sum_{i=1}^{k_S-1} z_{m+i-d}}{d-k_S+1}} A(z_{m+k_S-d}) \cdot \frac{1}{\sqrt{2\pi(d-k_S)}} e^{-\frac{\left(y-\sum_{i=1}^{k_S} z_{m+i-d}\right)^2}{2(d-k_S)}} dz_{m+k_S-d}\cdots dz_{m+1-d} \\
&\ll N^{-\frac{1}{2}-\delta} \underbrace{\int_{-C'\sqrt{N}}^{C'\sqrt{N}} \cdots \int_{-C'\sqrt{N}}^{C'\sqrt{N}}}_{k_S-1} \left(\prod_{i\in S\setminus[k_S+1;d]} \varphi(z_{m+i-d})\right)\left(\prod_{i\in[d]\setminus(S\cup\{k_S\})} A(z_{m+i-d})\right) \\
&\quad dz_{m+(k_S-1)-d}\cdots dz_{m+1-d} \\
&\ll N^{-\frac{1}{2}-\delta} \prod_{i\in S\setminus[k_S+1;d]} \left(\int_{-C'\sqrt{N}}^{C'\sqrt{N}} \varphi(z_{m+i-d}) dz_{m+i-d}\right) \\
&\quad \cdot \prod_{i\in[d]\setminus(S\cup\{k_S\})} \left(\int_{-C'\sqrt{N}}^{C'\sqrt{N}} A(z_{m+i-d}) dz_{m+i-d}\right) \\
&\ll N^{-1/2-(1+|[d]\setminus(S\cup\{k_S\})|)\delta} \ll N^{-1/2-\delta},
\end{aligned} \tag{B.5}$$

where in the second line, we use the change of variable $z_{m+k_S-d} \mapsto z_{m+k_S-d} - (\sum_{i=1}^{k_S-1} z_{m+i-d} + y)$ and the fact that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(d-k_S)}} e^{-\frac{(z_{m+k_S-d})^2}{2(d-k_S)}} dz_{m+k_S-d} \;=\; 1. \tag{B.6}$$

Since the number of proper subsets S of $[d]$ is finite, then when $k_S < d$, we have that $D_1(y) \ll N^{-1/2-\delta}$ by definition of $D_1(y)$ in (B.1).

Appendix C. Case for $d > 2$

In this section, we want to obtain the following estimate on $d^2\mathcal{M}_{Y^{(N)}}(y)/dy^2$ when $d > 2$

$$\begin{aligned}
&\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \\
&\quad \ll \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{d-2} \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right) dz_{m-2}\cdots dz_{m+1-d} \\
&\quad + \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{d-2} \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \varphi\left(y - \left(\sum_{i=1}^{d-2} z_{m+i-d}\right) - z_{m-2}\right) dz_{m-2}\cdots dz_{m+1-d}.
\end{aligned} \tag{C.1}$$

First, recall from (98) that

$$\begin{aligned}
&\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y) \\
&\ll \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-2} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} (\Phi(z_{m+1-d}))^{m-d} \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \left(\varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right)\right)^2 \\
&\quad + \left(\int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}} (\Phi(z_{m+1-d}))^{m-d} \left(\prod_{i=1}^{d-1} \varphi(z_{m+i-d})\right) (-1) \left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) \right. \\
&\quad \left. \cdot \varphi\left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \right) dz_{m-2} \cdots dz_{m+1-d}. \qquad \text{(C.2)}
\end{aligned}$$

Since y is an upper bound on $\sum_{i=1}^{d} z_{m+i-d}$ and $z_{m-2} \le z_{m-1}, z_m$, then we have $z_{m-2} \le (z_{m-1}+z_m)/2 \le (y-\sum_{i=1}^{d-2} z_{m+i-d})/2$. When $(y-\sum_{i=1}^{d-2} z_{m+i-d})/2 \le 0$, $\varphi(z_{m-1})$ is at most $\varphi((y-\sum_{i=1}^{d-2} z_{m+i-d})/2)$. Hence,

$$\begin{aligned}
&\int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}} (\Phi(z_{m+1-d}))^{m-d} \left(\prod_{i=1}^{d-1} \varphi(z_{m+i-d})\right) (-1) \left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) \\
&\cdot \varphi\left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \\
&\quad \ll \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right) \int_{-\infty}^{\infty} \left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) \\
&\qquad \cdot \varphi\left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \\
&\quad \ll \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right) \int_{-\infty}^{\infty} |z_{m-1}| \varphi(z_{m-1}) dz_{m-1}, \qquad \text{(C.3)}
\end{aligned}$$

where in the last line we use the change of variable $z_{m-1} \mapsto -z_{m-1} + y - \sum_{i=1}^{d-2} z_{m+i-d}$. The integral $\int_{-\infty}^{\infty} |x|\varphi(x)dx$ is the expected value of the absolute value of a standard normal random variable, which is finite. Hence, when $(y-\sum_{i=1}^{d-2} z_{m+i-d})/2 \le 0$,

$$\begin{aligned}
&\int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}} (\Phi(z_{m+1-d}))^{m-d} \left(\prod_{i=1}^{d-1} \varphi(z_{m+i-d})\right) (-1) \left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) \\
&\cdot \varphi\left(y-\sum_{i=1}^{d-1} z_{m+i-d}\right) dz_{m-1} \\
&\quad \ll \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right). \qquad \text{(C.4)}
\end{aligned}$$

On the other hand, when $(y - \sum_{i=1}^{d-2} z_{m+i-d})/2 \geq 0$,

$$y - \sum_{i=1}^{d-1} z_{m+i-d} \;\geq\; y - \sum_{i=1}^{d-2} z_{m+i-d} - \left(\frac{y - \sum_{i=1}^{d-2} z_{m+i-d}}{2} \right) \;=\; \frac{y - \sum_{i=1}^{d-2} z_{m+i-d}}{2} \;\geq\; 0. \tag{C.5}$$

Hence, we find

$$\begin{aligned}
&\int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}} (\Phi(z_{m+1-d}))^{m-d} \left(\prod_{i=1}^{d-1} \varphi(z_{m+i-d}) \right) (-1) \left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) \\
&\cdot \varphi \left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) dz_{m-1} \\
&\quad \ll \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d}) \right) \int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}} \left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) \varphi \left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) dz_{m-1} \\
&\quad = \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d}) \right) \varphi \left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) \Bigg|_{z_{m-1}=z_{m-2}}^{z_{m-1}=\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}} \\
&\quad = \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d}) \right) \left(\varphi \left(\frac{y - \sum_{i=1}^{d-2} z_{m+i-d}}{2} \right) - \varphi \left(y - \left(\sum_{i=1}^{d-2} z_{m+i-d} \right) - z_{m-2} \right) \right).
\end{aligned} \tag{C.6}$$

Thus, if $d > 2$, regardless of whether $(y - \sum_{i=1}^{d-2} z_{m+i-d})/2 \leq 0$ or $(y - \sum_{i=1}^{d-2} z_{m+i-d})/2 \geq 0$, we have by combining (C.4) and (C.6) that

$$\begin{aligned}
&\int_{z_{m-2}}^{\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}} (\Phi(z_{m+1-d}))^{m-d} \left(\prod_{i=1}^{d-1} \varphi(z_{m+i-d}) \right) (-1) \left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) \\
&\quad \cdot \varphi \left(y - \sum_{i=1}^{d-1} z_{m+i-d} \right) dz_{m-1} \\
&\quad \ll \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d}) \right) \left(\varphi \left(\frac{y - \sum_{i=1}^{d-2} z_{m+i-d}}{2} \right) + \varphi \left(y - \left(\sum_{i=1}^{d-2} z_{m+i-d} \right) - z_{m-2} \right) \right).
\end{aligned} \tag{C.7}$$

Hence, (98) becomes

$$\begin{aligned}
&\frac{d^2}{dy^2}\mathcal{M}_{Y^{(N)}}(y)\\
&\ll \int_{-C\sqrt{N}}^{\frac{y}{d}} \prod_{j=2}^{d-2} \int_{z_{m+(j-1)-d}}^{\frac{y-\sum_{i=1}^{j-1} z_{m+i-d}}{d-j+1}} \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \left(\varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right)\right)^2\\
&\quad + \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \left(\varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right) + \varphi\left(y - \left(\sum_{i=1}^{d-2} z_{m+i-d}\right) - z_{m-2}\right)\right)\\
&\quad dz_{m-2}\cdots dz_{m+1-d}\\
&\ll \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-2} \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \varphi\left(\frac{y-\sum_{i=1}^{d-2} z_{m+i-d}}{2}\right) dz_{m-2}\cdots dz_{m+1-d}\\
&\quad + \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{d-2} \left(\prod_{i=1}^{d-2} \varphi(z_{m+i-d})\right) \varphi\left(y - \left(\sum_{i=1}^{d-2} z_{m+i-d}\right) - z_{m-2}\right) dz_{m-2}\cdots dz_{m+1-d},
\end{aligned} \tag{C.8}$$

which is exactly the estimate we need.

References

[1] Becker, T.; Burt, D.; Corcoran, C.; Greaves-Tunnell, A.; Iafrate, J. R.; Jing, J.; Miller, S. J.; Porfilio, J. D.; Ronan, R.; Samranvedhya, J.; Talbut, B.; Strauch, F. W. Benford's Law and Continuous Dependent Random Variables. *Annals of Physics* **2018**, *388*, 350–381.

[2] Becher V.; Bugeaud Y.; Slaman T. A. The Irrationality Exponents of Computable Numbers. *Proc. Amer. Math. Soc.* **2016**, *144*, 1509–1521.

[3] Berger, A.; Hill, T. P. *An Introduction to Benford's Law*, 1st ed.; Princeton University Press: Princeton, NJ, USA, **2015**; ISBN 978-0-691-16306-2.

[4] Benford, F. The Law of Anomalous Numbers. *Proc. Amer. Philos. Soc.* **1938**, *78*, 551–572.

[5] Betti L.; Durmić I.; McDonald Z.; Miller J. B.; Miller S. J. Benfordness of Measurements Resulting from Box Fragmentation. *J. Stat. Theory Pract.* **2025**, *19*, 59. `https://doi.org/10.1007/s42519-025-00476-w`.

[6] Diaconis P. The Distribution of Leading Digits and Uniform Distribution mod 1. *Ann. Probab.* **1979**, *5*, 72-81.

[7] Casella G.; Berger R. *Statistical Inference*, 2nd ed.; Duxbury Press: Pacific Grove, CA, USA, **2002**; ISBN 978-8-131-50394-2.

[8] Durmić I.; Miller S. J. Benford Behaviour of a Higher Dimensional Fragmentation Process. Undergraduate thesis, available at: https://librarysearch.williams.edu/permalink/01WIL_INST/1faevhg/alma991013795585602786.

[9] Folland G. B. *Real Analysis: modern techniques and their applications*, John Wiley & Sons: Hoboken, NJ, USA, **1999**; ISBN 978-0-471-31716-6.

[10] Fang B.; Irons A.; Lippelman E.; Miller S. J. Benford Behavior in Stick Fragmentation Problems. *Stats* **2025**, *8*, 91. https://doi.org/10.3390/stats8040091.

[11] Kuipers, L.; Niederreiter, H. *Uniform Distribution of Sequences*, 1st ed.; John Wiley & Sons: Hoboken, NJ, USA, **1974**; ISBN 978-0-471-51045-1.

[12] Lemons, D. S. On the Number of Things and the Distribution of First Digits. *Am. J. Phys.* **1986**, *56*, 816–817.

[13] Miller, S. J. *Benford's Law: Theory and Applications*, 1st ed.; Princeton University Press: Princeton, NJ, USA, **2015**; ISBN 978-0-691-14761-1.

[14] Newcomb, S. Note on the Frequency of Use of the Different Digits in Natural Numbers. *Am. J. Math.* **1881**, *4*, 39–40.

[15] Nigrini, M. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, 1st ed.; John Wiley & Sons: Hoboken, NJ, USA, **2012**; ISBN 978-1-119-20309-4.

[16] Stanley R. *Enumerative Combinatorics, vol 1*, 2nd ed; Cambridge University Press, Cambridge, UK, **2012**; ISBN 978-0-521-66351-9.

[17] Weyl, H. Über die Gleichverteilung von Zahlen mod. Eins (On the distribution of numbers modulo one). *Math. Ann.* **1916**, *77*, 313–352.