BENFORD'S LAW AND RANDOM INTEGER DECOMPOSITION WITH CONGRUENCE STOPPING CONDITION

XINYU FANG, STEVEN J. MILLER, MAXWELL SUN, AND AMANDA VERGA

ABSTRACT. Benford's law is a statement about the frequency that each digit arises as the leading digit of numbers in a dataset. It is satisfied by various common integer sequences, such as the Fibonacci numbers, the factorials, and the powers of most integers. In this paper, we prove that integer sequences resulting from a random integral decomposition process (which we model as discrete "stick breaking") subject to a certain congruence stopping condition approaches Benford distribution asymptotically. We also show that our requirement on the number of congruence classes defining the congruence stopping condition is necessary for Benford behavior to occur and is a critical point; deviation from that would result in drastically different behavior.

CONTENTS

The authors are supported by NSF Grant DMS2241623, NSF Grant DMS1947438, Williams College, and University of Michigan. We thank the referee for helpful suggestions.

1. Introduction

1.1. **Background.** Benford's Law, named after the physicist and mathematician Frank Benford who observed it in 1938, describes the non-uniform distribution of first digits in many real-world datasets. According to this law (which we define precisely below), the digit 1 arises as the leading digit approximately 30% of the time, 2 approximately 17% of the time, and so on, with larger digits occurring less frequently. This counterintuitive pattern emerges due to the logarithmic nature of the distribution. It can be observed in a wide range of naturally occurring datasets, such as financial reports, census data, scientific constants, and even seemingly unrelated fields like social media statistics. Today, there are numerous applications of Benford's law including in voting fraud detection [Nig], economics [Töd, V-BFJ], geology [NM1], signal processing [PHA], and the physical sciences [BMP, Eli, MSPZ, NWR, PTTV, SM1, SM2]. See [BH2, Mil1] for more on the general theory and fields where it is observed.

Some commonly seen integer sequences can be proven to satisfy Benford's law exactly, when considering the asymptotic limit as more and more terms of the sequence are included. Some examples include the Fibonacci numbers [Dun, Wa], the factorials [Sar], the powers of 2 [Rai], and the powers of almost any other number [Rai].

Some continuous processes also satisfy Benford's law exactly (in the asymptotic limit as the process continues through time). One is an exponential growth or decay process: If a quantity is exponentially increasing or decreasing in time, then the percentage of time that it has each first digit satisfies Benford's law asymptotically (i.e. increasing accuracy as the process continues through time).

Given its ubiquity and many applications, it is therefore of interest to study which mathematical processes lead to Benford behavior. In general, it is often true that arithmetic operations (such as sums or products) of random variables yield a random variable that is closer to satisfying Benford's Law [Adh, AS, Bha, JKKKM, Lév1, Lév2, MN1, Rob, Sak, Sch1, Sch2, Sch3, ST]. However, this is not always the case (see for example [BH1]). In certain cases, a central limit theorem law is attainable, where Benfordness follows from the convergence of the distribution of mantissas (see Section 2) to the uniform distribution.

In this paper, we focus on the Benfordness of sequences of integers resulting from random integer decomposition processes with stopping conditions defined by congruence relations. Such sequences are naturally of number theoretic interest, as they arise from a process that is defined by congruence conditions.

1.2. **Random Integer Decomposition.** A *random integer decomposition process* is defined as follows. Start with some positive integer L. First, choose a random integer X in $\{1, \ldots, L-1\}$ uniformly and decompose L into two components: X and $(L-X)$. This state is called the first *level*. Then repeat the procedure on both components, resulting in 4 components - this is the second *level*, and so on. Whenever a component reaches 1, it stops decomposing further, and the procedure continues on all existing components that are larger than 1.

Clearly, the above process will eventually result in a sequence of 1's, which is uninteresting. However, the resulting sequence of components becomes more interesting when extra conditions are imposed. For example, Becker et al. [B+] studied the decomposition process in which one of the two components (say the first one X) always stops decomposing further, so it becomes an element in the resulting sequence, while the other component (namely $L-X$) continues to decompose until the remaining alive component reaches 1. In other words, it is a one-sided decomposition process.

The authors showed that for such a process, as $L \to \infty$, the resulting collection of components approaches Benford behavior (Definition 2.3).

It is natural to ask what other interesting conditions we can impose on the decomposition process so that it also yields a sequence that satisfies Benford's law, and we found that stopping conditions defined by congruence classes modulo an integer are good candidates, since the observation by Becker et al. provides strong indication that a process starting with L odd and stopping at even components should also result in Benford behavior. Indeed, that is the case (Theorem 1.1). Before stating the result precisely, let us define some terminology.

A *stopping condition* is a subset $\mathfrak{S} \subset \mathbb{Z}_+$ containing the element 1. A *random integer decomposition process with stopping condition* $\mathfrak{S}$ is defined as a random integer decomposition process in which a component stops decomposing further (and becomes part of the final sequence) as soon as it falls inside the set $\mathfrak{S}$. For example, the random integer decomposition process (without stopping condition) that we described at the beginning can be seen as one with the trivial stopping condition $\mathfrak{S} = \{1\}$. At the end of such a process, the resulting components are all in $\mathfrak{S}$, and we investigate the behavior of this collection of components (for example, in the trivial case, they would just be L copies of 1's; in general, it would be a multi-set with elements in $\mathfrak{S}$).

Theorem 1.1. *Start with an odd integer $L > 0$. Let the stopping set be $\mathfrak{S} = \{1\} \cup \{2m : m \in \mathbb{Z}_+\}$. Then the multi-set of ending components $\{X_i^{(L)}\}_{1 \le i \le m_L}$ resulting from the random integer decomposition process with stopping condition $\mathfrak{S}$ converges to strong Benford behavior (cf. Definition 2.3) as $L \to \infty$.*

In particular, due to Remark 2.4, this means that for any $s \in [1, B)$, the proportion of ending components with significand at most s converges to $\log_B(s)$ in probability as $L \to \infty$.

This is a special case of our Main Theorem below, which asserts the conclusion for stopping conditions defined by residue classes of a general modulus.

Theorem 1.2. *Fix an even modulus $n \ge 2$ and a subset $S \subset \{0, \dots, n-1\}$ of size $n/2$ representing the residue classes. Let the stopping set be*

$$\mathfrak{S} \ := \ \{1\} \cup \{m \in \mathbb{Z}_+ : m = qn + r,\ r \in S, q \in \mathbb{Z}\}. \tag{1.2.1}$$

Consider the collection $\{X_i^{(L,R)}\}_{1 \le i \le m_{L,R}}$ of all *components resulting from R random decomposition processes each starting from $L > 0$ subject to stopping condition $\mathfrak{S}$. This collection converges to strong Benford behavior (cf. Definition 2.3) given that $R > (\log L)^3$ as $L \to \infty$.*

In particular, for all $s \in [1, B)$, and for any choice of $R_L > (\log L)^3$ for each L, we have that the proportion of elements in $\{X_i^{(L,R)}\}_{1 \le i \le m_{L,R}}$ with significand at most s approaches $\log_B(s)$ in probability.

Remark 1.3. *Note that we need to consider the combined collection of many random processes in the case of a general modulus (Theorem 1.2), but did not have to do so in the special case of modulus equal to 2 (Theorem 1.1) because the latter process is guaranteed to continue until the remaining alive component hits 1. When we have a general modulus, the process might stop in one (or very few) steps with positive probability regardless of L, which is clearly non-Benford. Taking the combined collection eliminates this issue.*

1.3. **Stick Breaking Model.** We find it convenient to formulate the problem as the physical process of breaking an integer-length stick at integral points, and studying the final collection of stick lengths. More precisely, the process described earlier corresponds to starting with a stick of length L, breaking it into two pieces of lengths X and $L - X$ (this is called the first *level*), and continuing the process subject to the stopping condition $\mathfrak{S}$. Each round of breaking all existing sticks not already in the stopping set $\mathfrak{S}$ once is called a *level*. For convenience, we say that a stick *dies* or becomes *dead* when its length falls into the stopping set, and *alive* otherwise. Then our main results are formulated equivalently as follows.

Proposition 1.4. *Start with a stick of odd integer length L. Let the stopping set be $\mathfrak{S} = \{1\} \cup \{2m : m \in \mathbb{Z}_+\}$. Namely, a stick dies whenever its length is 1 or even. Then the collection $\{X_i^{(L)}\}_{1 \leq i \leq m_L}$ of lengths of all dead sticks (including those of length 1) converges to strong Benford behavior as $L \to \infty$.*

Proposition 1.5. *Fix an even modulus $n \geq 2$ and a subset $S \subset \{0, \dots, n-1\}$ of size $n/2$ representing the residue classes. Let the stopping set be*

$$\mathfrak{S} \ := \ \{1\} \cup \{m \in \mathbb{Z}_+ : m = qn + r, \ r \in S, q \in \mathbb{Z}\}. \tag{1.3.1}$$

If we start with R identical sticks of positive integer length $L \notin \mathfrak{S}$, then the collection of ending stick lengths $\{X_i^{(L,R)}\}_{1 \leq i \leq m_{L,R}}$ converges to strong Benford behavior given that $R > (\log L)^3$ as $L \to \infty$.

Keeping in mind the obvious correspondence of the two setups, we will from now on adopt the language of the stick breaking process and aim to prove Proposition 1.5 which readily implies our Main Theorem.

1.4. **Plan of Proof.** Our strategy to prove Proposition 1.5 is to consider a continuous approximation of the discrete stick breaking process. The congruence stopping conditions will be approximated by "probabilistic stopping" in the continuous case, where each stick stops breaking with a fixed probability given by the proportion of congruence classes present in the stopping set, namely $|S|/n$. We first prove, using tools from probability theory and analysis, that the continuous analogue converges to Benford behavior; this is done in Section 3. Then, in Section 4, we show that the discrete and continuous processes are sufficiently "close" in a precise sense, which allows us to deduce that the discrete fragmentation also results in Benford behavior. The necessity of the particular type of congruence stopping condition leading to strong Benford behavior is shown in Section 5.

Acknowledgements. The authors are supported by NSF Grant DMS2241623, NSF Grant DMS1947438, Williams College, and University of Michigan. We would like to thank the referee for helpful suggestions.

2. Benford's Law and Strong Benford Behavior

In this section, we give a precise definition of what it means for a sequence of random collections of positive real numbers to *converge to strong Benford behavior*, made rigorous from what is commonly known as Benford's law.

Fix a base $B > 0$. Any $x > 0$ can be written as

$$x \ = \ S_B(x) \cdot B^{k_B(x)} \tag{2.0.1}$$

where $S_B(x) \in [1, B)$ is the *significand* of x base B and $k_B(x) = \lfloor \log_B(x) \rfloor$ is the *exponent*. The *mantissa* of x is defined to be

$$M_B(x) \ = \ \log_B(x) - k_B(x),$$

namely, the fractional part of $\log_B(x)$.

We have the following standard definition (see for example [MN1]).

Definition 2.1 (Benford's law for a deterministic sequence of numbers)**.** *A sequence of positive real numbers* $(a_i)_{i\in\mathbb{N}}$ *is said to be* Benford base B *if*

$$\lim_{N\to\infty} \frac{\#\{i \le N : 1 \le S_B(a_i) \le s\}}{N} \;=\; \log_B s \tag{2.0.2}$$

for all $s \in [1, B]$.

The notion of Benford behavior for a positive real random variable supported on $(0, \infty)$ is defined as follows.

Definition 2.2. *A probability distribution* $\mathcal{D}$*, supported on* $(0, \infty)$*, is said to be* Benford base B *if for* $X \sim \mathcal{D}$*,* $M_B(X)$ *follows the uniform distribution on* $[0, 1]$*. This is equivalent to saying that*

$$\mathbb{P}(1 \le S_B(X) \le s) \;=\; \log_B s \tag{2.0.3}$$

for all $s \in [1, B]$.

This is sometimes referred to as *strong Benford*, as opposed to *weak Benford*, which only concerns the leading digits of a sequence of numbers. Since we are interested in the limiting behavior of a random sequence of finite collections of stick lengths, we give the following precise definition of "convergence to Benford".

For a random process whose realizations are given by a sequence of finite collections of positive real numbers $(\{X_i^{(n)} : 1 \le i \le m_n\})_n$, we may define for each state n the empirical distribution function of the significands

$$P_n(s) := \frac{1}{m_n}\left|\left\{i : S_B\left(X_i^{(n)}\right) \le s\right\}\right|, \text{ for all } s \in [1, B].$$

In other words, P_n is a random function giving the proportion of elements in the multi-set

$$\{X_i^{(n)} : 1 \le i \le m_n\}$$

whose significand is at most s.

Definition 2.3 (Convergence to strong Benford behavior, [B+])**.** *A random sequence of finite collections of positive real numbers* $(\{X_i^{(n)} : 1 \le i \le m_n\})_n$ *is said to* converge to strong Benford behavior (base B) *if* [1]

(1)

$$\lim_{n\to\infty} \mathbb{E}[P_n(s)] \;=\; \log_B(s) \tag{2.0.4}$$

and

(2)

$$\lim_{n\to\infty} \mathrm{Var}[P_n(s)] \;=\; 0. \tag{2.0.5}$$

Remark 2.4. *By Chebyshev's inequality, the above condition implies that* $P_n(s) \to \log_B(s)$ *in probability as* $n \to \infty$.

[1]For each n, m_n is a positive integer-valued random variable, and each $X_i^{(n)}$ is a positive real-valued random variable. Moreover, recall that a collection is a multi-set, so there may be repetitions in $\{X_i^{(n)} : 1 \le i \le m_n\}$.

In our stick breaking model, the multi-set $\{X_i^{(n)} : 1 \le i \le m_n\}$ is the collection of ending stick lengths from a given stick breaking process, with index n being either the total number of levels N (cf. Section 1.3), the starting stick length L, and/or the number of starting sticks R (sometimes the sequence is doubly-indexed by both L and R); these quantities go to infinity in the limit. For simplicity this collection may be abbreviated as $\{X_i\}$, but the limit will be explicitly stated in terms of the indexing parameter.

We recall some notations from [B+] for later use in the proofs. For $s \in [1, B)$, we define the indicator function of "significand at most s" by

$$\varphi_s(x) \;:=\; \begin{cases} 1, & \text{if the significand of } x \text{ is at most } s \\ 0, & \text{otherwise.} \end{cases} \tag{2.0.6}$$

We can then denote the proportion of elements in a set $\{X_i\}$ whose significand is at most s by

$$P(s) \;:=\; \frac{\sum_i \varphi_s(X_i)}{\#\{X_i\}}. \tag{2.0.7}$$

3. Continuous Stick Fragmentation

We now investigate the continuous stick breaking process that serves as an approximation to the discrete process with congruence stopping condition that we are interested in. For simplicity, in the continuous breaking problem we always assume the initial length is 1, since scaling of stick lengths does not affect Benfordness. The continuous process is as follows.

Start from R sticks of length $L > 0$. Fix a positive integer $k \ge 2$. We call a stick *alive* if it continues to break in the next level and *dead* otherwise. All initial sticks are assumed to be alive and each breaks into k pieces in the first level with the $(k-1)$ breaking points being the coordinates of a random variable chosen from some *good* probability distribution $\mathcal{D}$ on $[0,1]^{k-1}$ (this is a continuity condition defined in (3.2.5); roughly, we require $\mathcal{D}$ to be sufficiently continuous (see Example 3.5)). The breaking of each living stick is independent from others. After each level, each new stick obtained continues to be *alive* with probability r and *dead* with probability $1-r$. If $r = 1/2$, this is exactly the continuous analogue of the setting in Proposition 1.5 where we stop at exactly half of the residue classes.

The main result in this section is the following. Throughout, we adopt the convention that $\log x$ stands for the natural logarithm of x, although the base usually does not play a role unless we explicitly state it.

Theorem 3.1. *In the above setting, when $r = 1/k$, the collection of stick lengths after $N \ge \log R$ levels almost surely converges to strong Benford behavior as $R \to \infty$.*

Intuitively, the condition $r = 1/k$ implies that the number of living sticks at each level should stay constant, so the resulting distribution is nice. When r is away from this critical threshold, the limiting distribution becomes non-Benford. See Section 5 for more details in that case.

To present our proof of Theorem 3.1, we will first recall the basic stick breaking model (Section 3.1) that our continuous process generalizes, and the Mellin transform condition (Section 3.2) which is the main analytical tool to prove convergence to Benford in our model. We also prove a new theoretical result (Theorem 3.4) that gives a large family of examples of *good* distributions.

3.1. **The basic model.** We recall the following basic stick decomposition model studied in [B+]. Start with a stick of length L (this stick is indexed by 1) and fix a continuous probability distribution $\mathcal{D}$ with density function supported on $[0,1]$. Choose $p_1 \in [0,1]$ according to $\mathcal{D}$ and break L into p_1L and $(1-p_1)L$. This is the first *level*. Now for each subsequent level, for each existing stick of length ℓ obtained in the previous level indexed by i (namely, the i-th stick obtained in the process), sample independently a new ratio $p_i \in [0,1]$ according to $\mathcal{D}$ and break it into two pieces of length ℓp_i and $\ell(1-p_i)$ respectively. Repeat the same process on every new stick obtained in the previous level, where each breaking involves sampling a new ratio $p_i \in [0,1]$ according to $\mathcal{D}$. Then at the end of the N-th level, we obtain 2^N sticks of lengths

$$\begin{aligned}
X_1 &= Lp_1p_2p_4\dots p_{2^{N-2}}p_{2^{N-1}} \\
X_2 &= Lp_1p_2p_4\dots p_{2^{N-2}}(1-p_{2^{N-1}}) \\
&\vdots \\
X_{2^N-1} &= L(1-p_1)(1-p_3)(1-p_7)\dots(1-p_{2^{N-1}-1})p_{2^N-1} \\
X_{2^N} &= L(1-p_1)(1-p_3)(1-p_7)\dots(1-p_{2^{N-1}-1})(1-p_{2^N-1})
\end{aligned}$$

Note that we choose each the p_i's independently from each other according to $\mathcal{D}$. Becker et al. proved the following.

Theorem 3.2 ([B+]). *The above basic stick decomposition process $\{X_i^{(N)} : 1 \le i \le 2^N\}_N$ converges to strong Benford behavior in the limit $N \to \infty$ if $\mathcal{D}$ satisfies the Mellin transform condition (3.2.2).*

We discuss the Mellin transform condition in more detail next.

3.2. **Mellin transform condition.** For a continuous real-valued function $f : [0,\infty) \to \mathbb{R}$, let $\mathcal{M}f$ denote its *Mellin transform* defined by

$$\mathcal{M}f(s) \;=\; \int_0^\infty f(x)x^{s-1}dx. \tag{3.2.1}$$

Let $\mathcal{F} = \{\mathcal{D}_j\}_{j\in I}$ be a family of probability distributions with associated density functions f_j supported on $[0,\infty)$ and $p : \mathbb{Z}_+ \to I$. We say that $\mathcal{F}$ satisfies the *Mellin transform condition* if the following holds and the convergence is uniform over all choices of p:

$$\lim_{n\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \prod_{m=1}^{n} \mathcal{M}f_{\mathcal{D}_{p(m)}}\left(1 - \frac{2\pi i\ell}{\log B}\right) \;=\; 0. \tag{3.2.2}$$

This condition was proposed by Jang et al. in [JKKKM, Theorem 1.1]. The following corollary of [JKKKM, Theorem 1.1 & Lemma 1.2] relating the Mellin transform property to Benford behavior will be used repeatedly in our proofs of Benfordness results, so we restate it here for ease of reference.

Theorem 3.3 ([JKKKM, Theorem 1.1]). *Let $\mathcal{F} = \{\mathcal{D}_j\}_{j\in I}$ be a family of probability distributions with associated density functions f_j supported on $[0,\infty)$ satisfying the Mellin transform property and $p : \mathbb{Z}_+ \to I$. Let $X_1 \sim \mathcal{D}_{p(1)}$. For all $i \ge 2$, let X_i be a random variable with probability density function given by*

$$\theta^{-1}f_{\mathcal{D}_{p(i)}}(x/\theta) \tag{3.2.3}$$

where θ is the value of the previous random variable X_{i-1}. Then if $Y_n = \log_B X_n$, we have

$$\begin{aligned} &|\mathbb{P}(Y_n \mod 1 \in [a,b]) - (b-a)| \\ &\leq \ (b-a) \cdot \left| \lim_{n\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \prod_{m=1}^{n} \mathcal{M} f_{\mathcal{D}_{p(m)}} \left(1 - \frac{2\pi i \ell}{\log B}\right) \right|. \end{aligned} \tag{3.2.4}$$

In particular, the limiting distribution as $n \to \infty$ of X_n is Benford base B.

The product random variable of i independent random variables, each with distribution given by $\mathcal{D}_{p(m)}$ for $1 \leq m \leq i$, has probability density given by (3.2.3); thus, the sequence of such product random variables, denoted by (X_n), converges to a base B Benford random variable according to Theorem 3.3.

We prove the following new result that gives a sufficient condition on $\mathcal{F}$ for it to satisfy the Mellin transform condition. A weaker version of the result is briefly discussed in [JKKKM]. We include the proof in Appendix A.

Theorem 3.4. *$\mathcal{F}$ satisfies the Mellin transform condition if it is finite and each $f_j \in \mathcal{F}$ are α_j-Hölder continuous (for some $0 < \alpha_j \leq 1$) and supported only on $[0,1]$. In particular, for such an $\mathcal{F}$, a sequence of products of random variables distributed according to some sequence $(f_j) \subset \mathcal{F}$ approaches Benford behavior, and the rate of this convergence is uniform over all such sequences.*

Consider a probability distribution $\mathcal{D}$ on $\mathbb{R}^m$ that is supported on $[0,1]^m$ with cumulative distribution function F. Think of this as giving the m cut points on a stick, not necessarily ordered. For $X \sim \mathcal{D}$, Let $\text{rk}_i(X)$ denote its ith smallest coordinate, where $1 \leq i \leq m$. Let $\text{rk}_0(X) = 0$ and $\text{rk}_{m+1}(X) = 1$. Then, we say that $\mathcal{D}$ is *good* if

$$Y_i \ = \ \text{rk}_{i+1}(X) - \text{rk}_i(X) \tag{3.2.5}$$

has Hölder continuous density for all $0 \leq i \leq m$. In other words, if X represents the cut points of a stick, then we require the distances between adjacent ones to have Hölder continuous densities. This definition is necessary for exploring stick breaking, in which we must choose multiple breaking points of a stick from a distribution and then consider distributions of ratios between the lengths of children and their parents. That is, if such a distribution is *good*, then Theorem 3.4 applies. In fact, many distributions of interest are *good*. For instance, we have the following family of examples.

Example 3.5. *Suppose that $\mathcal{D}$ is the product of m independent 1-dimensional distributions $\mathcal{D}_i$ with densities f_i and cumulative densities F_i. If the f_i are Hölder continuous, then $\mathcal{D}$ is good.*

Proof. Let $Y_i = \text{rk}_{i+1}(X) - \text{rk}_i(X)$ for some $X \sim D$. Assume that $1 \leq i < m$. Then

$$1 - F_{Y_i}(c) \ = \ \sum_{j=1}^{m} \sum_{\substack{S \subseteq [m]\setminus\{j\} \\ |S|=i-1}} \int_0^1 f_j(x) \prod_{l\in S} F_l(x) \prod_{l\notin S,\ l\neq j} (1 - F_l(x+c))\ dx \tag{3.2.6}$$

where j denotes the index of the coordinate of X that corresponds to $\text{rk}_i(X)$ and S denotes the set of indices of the coordinates of X that are less than X_j. By continuity of the f_i's, we can differentiate with respect to c and move the differentiation inside of the integral to obtain

$$f_{Y_i}(c) = \sum_{j=1}^{m} \sum_{\substack{S\subseteq[m]\setminus\{j\} \\ |S|=i-1}} \int_0^1 f_j(x) \prod_{l\in S} F_l(x) \sum_{l'\notin S,\ l'\neq j} (-f_l(x+c)) \prod_{l\notin S,\ l\neq j,l'} (1-F_l(x+c))\, dx. \quad (3.2.7)$$

Now, the F_i are continuously differentiable, so they are also Hölder continuous. We can then take the minimal exponent α among the the f_i to obtain that f_{Y_i} is α-Hölder continuous since sums, products and integrals of α-Hölder continuous functions are α-Hölder continuous. We can similarly show that f_{Y_i} is α-Hölder continuous when $i = 0, m$. □

From now on, we assume all distributions from which breaking points are sampled are *good*.

3.3. **Proof of Theorem 3.1 Assuming Independence.** To prove Theorem 3.1, we first state and prove a version with the extra assumption that the alive/dead status of the sticks are independent.

Theorem 3.6. *When $r = 1/k$ and the alive/dead status of each stick is independent, the process ends in finitely many levels with probability 1, and the collection of ending stick lengths almost surely converges to strong Benford behavior as $R \to \infty$.*

We first show that the process starting from a single stick terminates in finitely many levels.

Proof of finite termination. Let p be the probability that it does not terminate. In such a case, one of the live children of the initial stick initiates a breaking that does not terminate. Thus, we have, if A is the number of live children of the original stick,

$$\begin{aligned} p &= \sum_{a=1}^{k} \mathbb{P}(A=a)\mathbb{P}(\text{at least one of the } a \text{ live children initiates infinite breaking}) \\ &= \sum_{a=1}^{k} \binom{k}{a} \frac{1}{k^a} \left(1-\frac{1}{k}\right)^{k-a} (1-(1-p)^a) \\ &= \sum_{a=1}^{k} \binom{k}{a} \frac{1}{k^a} \left(1-\frac{1}{k}\right)^{k-a} - \sum_{a=1}^{k} \binom{k}{a} \left(\frac{1-p}{k}\right)^{a} \left(1-\frac{1}{k}\right)^{k-a} \\ &= \left[1-\left(1-\frac{1}{k}\right)^{k}\right] - \left[\left(1-\frac{p}{k}\right)^{k} - \left(1-\frac{1}{k}\right)^{k}\right] \\ &= 1-\left(1-\frac{p}{k}\right)^{k}. \end{aligned} \quad (3.3.1)$$

Now, we have that, by Bernoulli's inequality,

$$\left(1-\frac{p}{k}\right)^{k} \geq 1-p \quad (3.3.2)$$

with equality if and only if $p = 0$. But we do have equality, so $p = 0$, as desired. □

Now, consider the process where all R sticks are being broken simultaneously. The above result implies that for any given R, this process also ends in finitely many levels with probability 1. Now we show the second part of Theorem 3.6.

Let n_i be the number of live sticks present at the i^{th} level so that $n_0 = R$. Then, we have the following:

Lemma 3.7. *For $i \geq 0$,*

$$\mathbb{P}\left(|n_j - R| \leq t \ \forall\ 0 \leq j \leq i\right) \ \geq \ 1 - \frac{2i^3R(k-1)}{t^2k} \tag{3.3.3}$$

if $0 < t < R$.

Proof. The result is trivial for $i = 0$. We proceed with induction on i. Assume the result for i; we show it for $i+1$. We have that

$$\begin{aligned}\mathbb{P}\left(|n_j - R| \leq t \ \forall\ 0 \leq j \leq i+1\right) \ &\geq \ \mathbb{P}\left(|n_j - R| \leq \frac{i}{i+1}t \ \forall\ j \leq i, \quad |n_{i+1} - n_i| \leq \frac{1}{i+1}t\right) \\ &\geq \ \mathbb{P}\left(|n_j - R| \leq \frac{i}{i+1}t \ \forall j \leq i\right) \\ &\quad - \mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t, \quad n_i < 2R\right).\end{aligned} \tag{3.3.4}$$

Now, note that the conditional law of n_{i+1} conditioned on $n_i = m$ is a binomial distribution with mk trials with probability $1/k$ of success. Thus, conditioning on n_i, it has expectation n_i and variance $n_i(1-1/k)$ (since $\mathbb{E}[n_{i+1}|n_i = m] = mk(1/k) = m$ and $\mathrm{Var}(n_{i+1}|n_i = m) = m(1-1/k)$). So, by Chebyshev's inequality, for all integers $0 \leq m < 2R$ (such that $n_i = m$ with positive probability),

$$\mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t \ \Big|\ n_i = m\right) \ < \ \frac{m(1-1/k)}{\frac{1}{(i+1)^2}t^2} \ \leq \ \frac{2(i+1)^2R(k-1)}{t^2k}. \tag{3.3.5}$$

It follows that (taking the term in the sum below to be 0 if $n_i = m$ has probability 0)

$$\begin{aligned}\mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t, \ n_i < 2R\right) \ &= \ \sum_{m=0}^{2R-1} \mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t \ \Big|\ n_i = m\right)\mathbb{P}\left(n_i = m\right) \\ &\leq \ \frac{2(i+1)^2R(k-1)}{t^2k} \sum_{m=0}^{2R-1} \mathbb{P}\left(n_i = m\right) \\ &\leq \ \frac{2(i+1)^2R(k-1)}{t^2k}\end{aligned} \tag{3.3.6}$$

and we have that, from (3.3.4) and the inductive hypothesis,

$$\begin{aligned}\mathbb{P}\left(|n_j - R| \leq t \ \forall\ 0 \leq j \leq i+1\right) \ &\geq \ 1 - \frac{2i^3R(k-1)}{\frac{i^2}{(i+1)^2}t^2k} - \frac{2(i+1)^2R(k-1)}{t^2k} \\ &\geq \ 1 - \frac{2(i+1)^3R(k-1)}{t^2k}.\end{aligned} \tag{3.3.7}$$

□

For any R and N, define

$$P_R(s) \ := \ \frac{\sum_i \varphi_s(X_i)}{\#\{X_i\}} \tag{3.3.8}$$

where the sum runs over the set of resulting sticks in a process starting with R sticks (which is finite with probability 1). We show $\mathbb{E}[P_R(s)] \to \log_B(s)$ and $\mathrm{Var}(P_R(s)) \to 0$ as $R \to \infty$.

Proposition 3.8. $\mathbb{E}[P_R(s)] \to \log_B(s)$ *as $R \to \infty$.*

Proof. We first show the existence of a function $h(R) \to \infty$ as $R \to \infty$ such that the expectation of the average of $\varphi_s(X_i)$ over sticks X_i that die within the first $h(R)$ levels goes to $\log_B(s)$ as $R \to \infty$. Later we will argue that the collection of sticks that die after $h(R)$ levels make negligible contribution to $\mathbb{E}[P_R(s)]$, so this is sufficient. Define

$$P_R^n(s) \;:=\; \frac{\sum_{X_i \text{ in first } n \text{ levels}} \varphi_s(X_i)}{\#\{X_i | X_i \text{ in first } n \text{ levels}\}}. \tag{3.3.9}$$

Given a realization A of the alive or dead status of each stick that occurs in the entire decomposition process (without fixing the realization of its length), for any stick X_i belonging to level n coming from a living ancestor in level $n-1$,

$$|\mathbb{E}[\varphi_s(X_i)|A] - \log_B(s)| \;\leq\; f(n) \tag{3.3.10}$$

where f satisfies $f(n) \to 0$ as $n \to \infty$ by Theorem 3.4. We now show that in each of the first $h(R)$ levels, a roughly equal number of sticks become dead. We may take $h(R) = \lfloor R^{1/10} \rfloor$ and $t = R^{2/3}$ and apply Lemma 3.7. Then we obtain that,

$$\mathbb{P}(R - R^{2/3} < n_i < R + R^{2/3} \;\forall\; i \leq h(R)) \;\geq\; 1 - \frac{2R^{3/10}R(k-1)}{R^{4/3}k} \;\geq\; 1 - 2R^{-1/30} \;\to\; 1 \tag{3.3.11}$$

as $R \to \infty$. Let $\overline{n_i}$ be the number of sticks that become dead at level i, so that $\overline{n_i} = kn_{i-1} - n_i$. Then we have, with probability going to 1,

$$(k-1)R - (k+1)R^{2/3} \;<\; \overline{n_i} \;<\; (k-1)R + (k+1)R^{2/3}, \tag{3.3.12}$$

for all $i \leq h(R)$, which implies, when R is sufficiently large,

$$\left(k - \frac{3}{2}\right) R \;<\; \overline{n_i} \;<\; \left(k - \frac{1}{2}\right) R. \tag{3.3.13}$$

Now assume that A is a realization for which the above is true, so that conditioning on A the n_i are constants. Then, letting $P'_R(s) = P_R^{h(R)}(s)$,

$$\begin{aligned}
|\mathbb{E}(P'_R(s)|A) - \log_B(s)| \;&\leq\; \frac{1}{\sum_{i=1}^{h(R)} \overline{n_i}} \sum_{i=1}^{h(R)} f(i)\overline{n_i} \\
&\leq\; \frac{1}{(k-\frac{3}{2})Rh(R)} \sum_{i=1}^{h(R)} f(i)(k - \frac{1}{2})R \\
&\leq\; 3\frac{1}{h(R)} \sum_{i=1}^{h(R)} f(i) =: \delta(R) \;\to\; 0
\end{aligned} \tag{3.3.14}$$

as $R \to \infty$. Let $\mathcal{A}$ be the (countable) collection of realizations for the alive/dead-ness of the sticks that terminate finitely such that (3.3.13) is true for all $i \leq h(R)$. Let $E = \left(\bigcup_{A \in \mathcal{A}} A\right)^c$ and observe that we have shown above that $\mathbb{P}(E) \to 0$ as $R \to \infty$. Then, since the events in $\mathcal{A}$ are pairwise disjoint,

$$\begin{aligned}
|\mathbb{E}(P'_R(s)) - \log_B(s)| \;&\leq\; |\mathbb{E}(P'_R(s)|E) - \log_B(s)|\mathbb{P}(E) + \sum_{A \in \mathcal{A}} |\mathbb{E}(P'_R(s)|A) - \log_B(s)|\mathbb{P}(A) \\
&\leq\; \mathbb{P}(E) + \delta(R) \sum_{A \in \mathcal{A}} \mathbb{P}(A)
\end{aligned} \tag{3.3.15}$$

which approaches 0 as $R \to \infty$. We have used the fact that $P'_R(s)$ is bounded by 0 and 1. This implies $\mathbb{E}[P'_R(s)] \to \log_B(s)$. Now, for the sticks after level $h(R)$, simply note that

$$|\mathbb{E}[\varphi_s(X_i)|A] - \log_B(s)| \;\leq\; f(h(R)) \;\leq\; \sup_{n \geq h(R)} f(n) \tag{3.3.16}$$

which tends to 0 as $R \to \infty$. $P_R(s)$ is a weighted average of these $\varphi_s(X_i)$ and $P'_R(s)$, so $|\mathbb{E}(P_R(s)) - \log_B(s)| \to 0$, as desired. More precisely, let $P''_R(s)$ be the proportion of sticks X_i that die after level $h(R)$ such that $\varphi_s(X_i) = 1$ (or $\log_B(s)$ if no such sticks exist) and note that the above implies that $|\mathbb{E}[P''_R(s)|A] - \log_B(s)| \leq f(h(R))$. The conditioning on A can be removed (via the triangle inequality and law of total expectation) and the result follows after noting that $P_R(s)$ is a linear combination of $P''_R(s)$ and $P'_R(s)$ (with nonnegative coefficients). □

Proposition 3.9. $\mathrm{Var}(P_R(s)) \to 0$ *as* $R \to \infty$.

Proof. Now we analyze the variance. Let A be a choice of dead and alive sticks for all levels in which the process ends after a finite number of levels. Let (X_i, X_j) be a pair of final sticks and suppose both die after at least $\log(R)$ levels and they die at least $\log(\log(R))$ levels apart (i.e., they have enough independence). Let X be their most recent common ancestor so that $X_i = XY_i$ and $X_j = XY_j$ where Y_i and Y_j are each products of at least $\log(\log(R))$ random variables. Then,

$$\begin{aligned} \mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)|A] &= \int_x \mathbb{E}[\varphi_s(xY_i)\varphi_s(xY_j)|X = x, A]d\mu_X(x) \\ &= \int_x \mathbb{E}[\varphi_s(xY_i)|A]\mathbb{E}[\varphi_s(xY_j)|A]d\mu_X(x) \end{aligned} \tag{3.3.17}$$

by independence, so

$$\begin{aligned} \left|\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)|A] - \log_B^2(s)\right| &\leq \int_x \left|\mathbb{E}[\varphi_s(xY_i)|A]\mathbb{E}[\varphi_s(xY_j)|A] - \log_B^2(s)\right| d\mu_X(x) \\ &\leq \int_x |\mathbb{E}[\varphi_s(xY_i)|A] - \log_B(s)|d\mu_X(x) \\ &\quad + \int_x |\mathbb{E}[\varphi_s(xY_j)|A] - \log_B(s)|d\mu_X(x) \\ &\leq 2f(\lceil \log(\log(R)) \rceil) \end{aligned} \tag{3.3.18}$$

where we have used Theorem 3.4 and Theorem 3.3 in the last step. Thus, it suffices to show that the proportion of pairs of the following two types among all pairs of dead sticks goes to 0 as $R \to \infty$, with high probability:

(1) at least one of X_i,X_j dies before $\log(R)$ levels, or

(2) X_i, X_j have a common ancestor less than $\log(\log(R))$ levels before they both die.

Let M be the total number of dead sticks ever.

To show (1) occurs with low probability, we first show that the number of sticks that die within the first $\log(R)$ levels is small compared to M with probability going to 1. Keep our choice of $h(R)$ and t earlier. Therefore when R is sufficiently large, using the upper bound from (3.3.13), we get that as number of sticks that die within the first $\log(R)$ levels is upper bounded by

$$\left(k - \frac{1}{2}\right)\log(R)R \tag{3.3.19}$$

with probability going to 1. Now again using (3.3.13), we can lower bound M by lower bounding the total number of sticks that die within the first $h(R)$ levels. This gives

$$M \;\geq\; h(R)\left(k-\frac{3}{2}\right)R \;=\; \left(k-\frac{3}{2}\right)R^{11/10} \tag{3.3.20}$$

with probability going to 1. Since

$$\frac{(k-\frac{1}{2})\log(R)R}{(k-\frac{3}{2})R^{11/10}} \;\to\; 0 \tag{3.3.21}$$

as $R \to \infty$, we have shown that the proportion of sticks that die in the first $\log(R)$ levels among all goes to 0 as $R \to \infty$ with probability going to 1. This then implies that the number of pairs that involve a stick of this type also takes up a diminishing proportion of all pairs of final sticks as $R \to \infty$.

Now we show (2) is rare, namely, that the number of pairs X_i, X_j having a common ancestor at most $\log(\log R)$ levels before they both die is $o(M^2)$ with high probability. Fix some X_i. Then, the number of sticks, dead or alive, that share the α ancestor of X_i and is $\alpha-\beta$ levels away is at most k^β. Thus, the number of X_j that satisfying (2) when paired with X_i is bounded above by

$$\sum_{\alpha=1}^{\lfloor\log(\log R)\rfloor}\sum_{\beta=0}^{\lfloor\log(\log R)\rfloor} k^\beta \;\leq\; \log(\log R)\frac{k^{\log(\log R)}-1}{k-1} \;\leq\; \log(\log R)(\log R)^{\log k}. \tag{3.3.22}$$

Hence, the number of such pairs is bounded above by $M(\log R)^{1+\log k} = o(M^2)$ by (3.3.20). Let E be the event that all of the above inequalities hold. Then, (3.3.18) implies that

$$|\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)|E] - \log_B^2(s)| \leq 2f(\lceil\log(\log(R))\rceil) \tag{3.3.23}$$

and since $\mathbb{P}(E) \to 1$ as $R \to \infty$ it follows that $\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)] \to \log_B^2(s)$ as well, implying that $\mathrm{Var}[P_R(s)] \to 0$. □

3.4. **Proof of Theorem 3.1.** Now we explain how to build on the proof above to prove Theorem 3.1 without additional assumption. Without assuming independence on the alive/dead status of the sticks, we have the following weaker version of Lemma 3.7.

Lemma 3.10. *For $i \geq 0$,*

$$\mathbb{P}(|n_j - R| \leq t \;\forall\; j \leq i) \;\geq\; 1 - \frac{2i^3Rk^2}{t^2} \tag{3.4.1}$$

if $t < R$.

Proof. As in the proof of Lemma 3.7, we proceed with induction on i, noting that the result is trivial for $i = 0$. By the same calculation, (3.3.4) holds, That is,

$$\mathbb{P}(|n_j - R| \leq t \;\forall\; j \leq i+1) \;\geq\; \mathbb{P}\left(|n_j - R| \leq \frac{i}{i+1}t \;\forall\; j \leq i\right) - \mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t,\; n_i < 2R\right). \tag{3.4.2}$$

We have that n_{i+1} is the sum of n_i independent random variables with mean 1 and variance bounded by k^2. Thus, conditioning on $n_i = m$, it has expectation m and variance at most mk^2. Chebyshev's inequality implies

$$\mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t \;\middle|\; n_i = m\right) \;<\; \frac{mk^2}{\frac{1}{(i+1)^2}t^2} \;\leq\; \frac{2(i+1)^2Rk^2}{t^2}. \tag{3.4.3}$$

for $m < 2R$ such that $\mathbb{P}(n_i = m) > 0$. It follows that

$$\mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t,\ n_i < 2R\right) = \sum_{m=0}^{2R-1} \mathbb{P}\left(|n_{i+1} - n_i| > \frac{1}{i+1}t \;\middle|\; n_i = m\right) \mathbb{P}\left(n_i = m\right)$$
$$\leq \frac{2(i+1)^2 Rk^2}{t^2} \sum_{m=0}^{2R-1} \mathbb{P}\left(n_i = m\right)$$
$$\leq \frac{2(i+1)^2 Rk^2}{t^2}. \tag{3.4.4}$$

We then have that

$$\mathbb{P}(|n_j - R| \leq t\ \forall\ j \leq i+1)\ \geq\ 1 - \frac{2i^3 Rk^2}{\frac{i^2}{(i+1)^2} t^2 k} - \frac{2(i+1)^2 Rk^2}{t^2 k}\ \geq\ 1 - \frac{2(i+1)^3 Rk^2}{t^2} \tag{3.4.5}$$

which completes the induction. □

Theorem 3.1 follows from essentially the same arguments as in proof of Theorem 3.6 using Lemma 3.10. We highlight the necessary changes below.

For any R and N, define

$$P_{R,N}(s)\ :=\ \frac{\sum_i \varphi_s(X_i)}{\#\{X_i\}} \tag{3.4.6}$$

where the sum runs over the set of resulting sticks in the first N levels of a process starting with R sticks. We prove $\mathbb{E}[P_{R,N}(s)] \to \log_B(s)$ and $\mathrm{Var}(P_{R,N}(s)) \to 0$ if $N \geq \log(R)$ and $R \to \infty$. Keep the choices of $h(R) = R^{1/10}$ and $t = R^{2/3}$ in the proof of Theorem 3.6. For sticks that die after $h(R)$ levels, we know that

$$|\mathbb{E}[\varphi_s(X_i)] - \log_B(s)|\ \leq\ f(h(R)) \tag{3.4.7}$$

where the right-hand-side goes to 0 in R. Therefore it again suffices to estimate the errors

$$|\mathbb{E}[\varphi_s(X_i)] - \log_B(s)| \tag{3.4.8}$$

for X_i that dies within the first $h(R)$ levels. Now the exact same argument applies simply after replacing Lemma 3.7 with Lemma 3.10.

For variance, let M now denote the number of resulting sticks after N levels. By the same logic,

$$\mathbb{E}[\varphi_s(X_i)\varphi_s(X_j)]\ \to\ \log_B^2(s) \tag{3.4.9}$$

uniformly given that X_i and X_j have a most recent common ancestor more than $\log(\log R)$ levels away from X_i and both X_i, X_j die after at least $\log(R)$ levels. It therefore suffices to show that such pairs X_i, X_j make up a proportion of all pairs of dead sticks that tends to 1. This is done in the same way as in the proof of Theorem 3.6.

4. Discrete Stick Fragmentation

In this section, we return to the setting of discrete stick fragmentation with congruence stopping condition and present the proofs of Proposition 1.4 and Proposition 1.5. The overall strategy, adopted from that in §3 of [B+], is to approximate the discrete process with an appropriate continuous analogue, and by showing that the two processes are "close" in a precise sense, deduce the desired result from the corresponding continuous result.

4.1. **Proof of Theorem 1.4.** In order to carry out the approximation strategy outlined above, we define a continuous process and a discrete process based on the same sequence of random ratios, so that the latter is the process we are interested in and the former known to be Benford. Our goal is to show that their end results are "close" enough so that the Benfordness of the former implies that of the latter. The two processes are defined as follows. Let $(c_i)_{i\geq 0}$ be a sequence of random numbers chosen from $(0,1)$ with respect to the uniform distribution.

- Let $\mathcal{Q}$ denote the continuous process. In this process, we start with a stick of length $h_0 = L$. For each $i \geq 1$, break off a fragment of length $Y_i = c_{i-1}h_{i-1}$ at the i-th level, which becomes *dead*, namely, stops breaking further. The other stick of length $h_i = h_{i-1} - Y_i = (1-c_{i-1})h_{i-1}$ stays *alive* and continues to break in the next step.
- Let $\mathcal{P}$ denote the discrete process. In this process, we start with a stick of length $\ell_0 = L$. For each $i \geq 1$, break off a fragment of length $X_i = 2\lceil \frac{c_{i-1}(\ell_{i-1}-1)}{2} \rceil$ at the i-th level, which becomes *dead*. Note that by construction, X_i is an even integer taking values in $[2, \ell_{i-1}-1]$.The remaining stick of length $\ell_i = \ell_{i-1} - X_i$ (which is always an odd integer) stays *alive* and continues to break in the next step.
- Moreover, a stick in $\mathcal{P}$ also becomes dead if it has length 1. In that case, the corresponding stick in $\mathcal{Q}$ also dies.

We first derive the following lemma that bounds the length of a stick X_k in $\mathcal{P}$ with the length of the corresponding stick Y_k in $\mathcal{Q}$.

Lemma 4.1. *Given that $\ell_k, h_k > 2$, we have,*

$$Y_k \prod_{i=1}^{k-1}\left(1 - \frac{2}{\ell_i - 2}\right) - 2 \ \leq\ X_k \ \leq\ Y_k \prod_{i=1}^{k-1}\left(1 + \frac{2}{\ell_i - 2}\right) + 2\prod_{i=1}^{k-1}\frac{\ell_i}{\ell_i - 4}. \tag{4.1.1}$$

Proof. Let d_k be the rounded version of c_k used in $\mathcal{P}$, i.e., $d_k = X_{k+1}/\ell_k$. Then, note that

$$(\ell_k - 1)c_k \ \leq\ X_{k+1} \ \leq\ (\ell_k - 1)c_k + 2 \tag{4.1.2}$$

so that

$$\left(1 - \frac{1}{\ell_k}\right)c_k \ \leq\ d_k \ \leq\ \left(1 - \frac{1}{\ell_k}\right)c_k + \frac{2}{\ell_k} \qquad \Longrightarrow \qquad |d_k - c_k| \ \leq\ \frac{2}{\ell_k}. \tag{4.1.3}$$

It follows that

$$\ell_k \ =\ L\prod_{i=0}^{k-1}(1-d_i) \ \leq\ L\prod_{i=0}^{k-1}\left(1 - c_i + \frac{2}{\ell_i}\right) \ \leq\ L\prod_{i=0}^{k-1}(1-c_i)\prod_{i=0}^{k-1}\left(1 + \frac{2}{\ell_i(1-c_i)}\right). \tag{4.1.4}$$

We have that

$$\ell_i(1-c_i) \ \geq\ \ell_i(1-d_i) - 2 \ =\ \ell_{i+1} - 2, \tag{4.1.5}$$

so

$$\ell_k \ \leq\ L\prod_{i=0}^{k-1}(1-c_i)\prod_{i=0}^{k-1}\left(1 + \frac{2}{\ell_{i+1} - 2}\right) \ \leq\ h_k\prod_{i=1}^{k}\left(1 + \frac{2}{\ell_i - 2}\right). \tag{4.1.6}$$

Equation (4.1.5) also implies that $1 - c_i - \frac{2}{\ell_i} \geq 1 - c_i\left(1 - \frac{2}{\ell_{i+1}-2}\right)$, so that

$$\ell_k \ \geq\ L\prod_{i=0}^{k-1}\left(1 - c_i - \frac{2}{\ell_i}\right) \ \geq\ L\prod_{i=0}^{k-1}\left[(1-c_i)\left(1 - \frac{2}{\ell_{i+1}-2}\right)\right] \ \geq\ h_k\prod_{i=1}^{k}\left(1 - \frac{2}{\ell_i - 2}\right). \tag{4.1.7}$$

We can then multiply (4.1.6) by d_k to get

$$X_{k+1} \;\leq\; h_k d_k \prod_{i=1}^{k-1}\left(1+\frac{2}{\ell_{i+1}-2}\right) \;\leq\; \left(Y_{k+1}+\frac{2h_k}{\ell_k}\right)\prod_{i=1}^{k-1}\left(1+\frac{2}{\ell_{i+1}-2}\right) \tag{4.1.8}$$

and then use (4.1.7) to obtain

$$\begin{aligned} X_{k+1} \;&\leq\; Y_{k+1}\prod_{i=1}^{k}\left(1+\frac{2}{\ell_i-2}\right)+2\prod_{i=1}^{k}\left(1+\frac{2}{\ell_i-2}\right)\left(1-\frac{2}{\ell_i-2}\right)^{-1} \\ &\leq\; Y_{k+1}\prod_{i=1}^{k}\left(1+\frac{2}{\ell_i-2}\right)+2\prod_{i=1}^{k}\frac{\ell_i}{\ell_i-4}. \end{aligned} \tag{4.1.9}$$

We can reason similarly by multiplying (4.1.7) with d_k to obtain

$$X_{k+1} \;\geq\; Y_{k+1}\prod_{i=1}^{k}\left(1-\frac{2}{\ell_i-2}\right)-2. \tag{4.1.10}$$

□

Let $g(x)$ be a function that goes to infinity as $x \to \infty$ with $g(x) = o(\sqrt{\log(x)})$. Let $h(x)$ be a function that goes to infinity as $x \to \infty$. The following corollary of Lemma 4.1 essentially says that X_k and Y_k are very close given that k is not too large and ℓ_{k-1}, Y_k are large enough.

Corollary 4.2. *For all $k < g(L)\log L$ such that $\ell_{k-1} > \log^2(L)+2$ and $Y_k > h(L)$, we have*

$$Y_k(1-o(1)) \;\leq\; X_k \;\leq\; Y_k(1+o(1)). \tag{4.1.11}$$

Proof. By Lemma 4.1, we have

$$\begin{aligned} X_k \;&\leq\; Y_k\prod_{i=1}^{k-1}\left(1+\frac{2}{\ell_i-2}\right)+2\prod_{i=1}^{k-1}\frac{\ell_i}{\ell_i-4} \\ &\leq\; Y_k\left(1+\frac{2}{\log^2(L)}\right)^{k-1}+2\left(\frac{\log^2(L)}{\log^2(L)-4}\right)^{k-1} \\ &\leq\; Y_k\left(1+\frac{2}{\log^2(L)}\right)^{g(L)\log L}+2\left(1+\frac{8}{\log^2(L)}\right)^{g(L)\log L} \\ &\leq\; Y_k\exp\left(\frac{2g(L)}{\log L}\right)+2\exp\left(\frac{8g(L)}{\log L}\right). \end{aligned} \tag{4.1.12}$$

As $L \to \infty$, $\frac{g(L)}{\exp(L)} \to 0$, so $\exp\left(\frac{2g(L)}{\log L}\right) \to 1$ and $2\exp\left(\frac{8g(L)}{\log L}\right) = O(1)$. Now by our assumption $Y_k \to \infty$, we get asymptotically that

$$X_k \;\leq\; Y_k(1+o(1)). \tag{4.1.13}$$

For the other inequality, apply Lemma 4.1 again to get

$$\begin{aligned} X_k &\geq Y_k \prod_{i=1}^{k-1} \left(1 - \frac{2}{\ell_i - 2}\right) - 2 \\ &\geq Y_k \left(1 + \frac{2}{\log^2(L)}\right)^{k-1} - 2 \\ &\geq Y_k \left(1 - \frac{2}{\log^2(L)}\right)^{g(L)\log L} - 2 \\ &\geq Y_k (2e)^{-\frac{2g(L)}{\log L}} - 2. \end{aligned} \tag{4.1.14}$$

Again $(2e)^{-\frac{2g(L)}{\log L}} \to 1$ as $L \to \infty$, and since $Y_k \to \infty$, we get asymptotically that

$$X_k \;\geq\; Y_k(1 - o(1)). \tag{4.1.15}$$

. □

The following lemma then helps us translate Benfordness of $\mathcal{Q}$ to that of $\mathcal{P}$ given that they are close enough in the sense above. This is essentially [B+, Lemma 3.3], but we give a different proof here. Let $\{Z_i\}_L = \{Z_1, \dots, Z_{k_L}\}$ denote a finite sequence of random variables whose length k_L depends on L.

Lemma 4.3. *Suppose* $\left\{Y_i^{(L)}\right\}_L = \left\{Y_1^{(L)}, Y_2^{(L)}, \dots, Y_{k_L}^{(L)}\right\}$ *is strong Benford as* $L \to \infty$. *Then if* $\left\{X_i^{(L)}\right\}_L = \left\{X_1^{(L)}, X_2^{(L)}, \dots, X_{k_L}^{(L)}\right\}$ *is such that*

$$Y_i^{(L)}(1 - o_L(1)) \;\leq\; X_i^{(L)} \;\leq\; Y_i^{(L)}(1 + o_L(1)) \tag{4.1.16}$$

as $L \to \infty$, $\left\{X_i^{(L)}\right\}_L$ *is strong Benford as* $L \to \infty$.

Proof. Let

$$P_L^X(s) := \frac{\#\{i : 1 \leq i \leq k_L,\ S_B(X_i^{(L)}) \leq s\}}{k_L} \tag{4.1.17}$$

and similarly for $P_L^Y(s)$. The above definition clearly works for $1 \leq s < B$, but we also let it hold for $s < 1$, simply noting that the quantity is 0, and $s \geq B$, simply noting that the quantity becomes 1. Now, observe that, from the condition, $S_B(Y_i^{(L)})(1 - o_L(1)) \leq X_i^{(L)}/B^{\lfloor \log_B(Y_i^{(L)}) \rfloor} \leq S_B(Y_i^{(L)})(1 + o_L(1))$. For L large enough such that $1 - o_L(1) > 1/B$ and $1 + o_L(1) < B$, we have $1/B < \frac{X_i^{(L)}}{Y_i^{(L)}} < B$. Since significands are bounded by B, it follows that $S_B(Y_i^{(L)}) - c(L)/B < S_B(X_i^{(L)})b < S_B(Y_i^{(L)}) + c(L)/B$ for some $b \in \{1/B, 1, B\}$ and $c(L)$ that approaches 0 as $L \to \infty$. It follows that $S_B(X_i^{(L)}) \leq s$ implies that $S_B(Y_i^{(L)}) < s + c(L)$ or $S_B(Y_i^{(L)}) > B - c(L)$. Also, $S_B(X_i^{(L)}) \leq s$ if $1 + c(L) < S_B(Y_i^{(L)}) \leq s - c(L)$. Thus,

$$P_L^Y(s - c(L)) - P_L^Y(1 + c(L)) \;\leq\; P_L^X(s) \;\leq\; P_L^Y(s + c(L)) + 1 - P_L^Y(B - c(L)) \tag{4.1.18}$$

It is then easy to see that $\mathbb{E}[P_L^X(s)] \to \log_B(s)$ as $L \to \infty$ for $1 \leq s \leq B$ by strong Benfordness of $\{Y_i^{(L)}\}_L$ and monotonicity of $P_L^Y(s)$ (as a function of s). Now, observe that

$$P_L^X(s)^2 \;\geq\; P_L^Y(s - c(L))^2 - 2P_L^Y(s - c(L))P_L^Y(1 + c(L)) \;\geq\; P_L^Y(s - c(L))^2 - 2P_L^Y(1 + c(L)) \tag{4.1.19}$$

and

$$\begin{aligned}P_L^X(s)^2 &\le P_L^Y(s+c(L))^2 + 2P_L^Y(s+c(L))(1-P_L^Y(B-c(L))) + (1-P_L^Y(B-c(L)))^2 \\ &\le P_L^Y(s+c(L))^2 + 3(1-P_L^Y(B-c(L))) \end{aligned} \tag{4.1.20}$$

so that, using the fact that $\mathbb{E}[P_L^Y(s)^2] \to \log_B(s)^2$, we obtain that $\mathbb{E}[P_L^X(s)^2] \to \log_B(s)^2$ as well. □

By [B+, Theorem 1.9], the process $\mathcal{Q}$ is Benford. Given the lemma above, it now suffices to show that the premises of Corollary 4.2 are satisfied for almost all k. The following lemma shows that the process ends within $g(L)\log L$ levels with probability going to 1, so the first condition that k is not too large is almost always true.

Lemma 4.4. *Let F_L be the number of fragments generated by a stick of length L. As $L \to \infty$,*

$$\mathbb{P}[(\log\log L)^2 < F_L < g(L)\log L] \;=\; 1 - o(1). \tag{4.1.21}$$

Proof. We first show the upper bound using Markov's inequality. We prove by induction that

$$\mathbb{E}[F_\ell] \;=\; 1 + 2\sum_{\substack{0<j<\ell \\ j \text{ even}}} \frac{1}{j}. \tag{4.1.22}$$

It is clear that $\mathbb{E}[F_1] = 1$. We have the recurrence

$$\mathbb{E}[F_L] \;=\; \frac{2}{L-1}\sum_{\substack{\ell<L \\ \ell \text{ odd}}} (1 + \mathbb{E}[F_\ell]) \tag{4.1.23}$$

since there is a $\frac{2}{L-1}$ probability of breaking off a piece of length ℓ in the first break for $1 \le \ell \le L-1$ and ℓ odd. By the induction hypothesis, we have

$$\begin{aligned}\mathbb{E}[F_L] &= \frac{2}{L-1}\sum_{\substack{\ell<L \\ \ell \text{ odd}}} \left(1 + \left(1 + 2\sum_{\substack{0<j<\ell \\ j \text{ even}}} \frac{1}{j}\right)\right) \\ &= \frac{2}{L-1}\cdot\frac{L-1}{2} + \frac{2}{L-1}\sum_{\substack{\ell<L \\ \ell \text{ odd}}}\left(1 + 2\sum_{\substack{0<j<\ell \\ j \text{ even}}}\frac{1}{j}\right) \\ &= 1 + \frac{2}{L-1}\left(\frac{L-1}{2} + 2\sum_{\substack{0<j<L-2 \\ j \text{ even}}} \frac{\frac{L-j-1}{2}}{j}\right) \\ &= 1 + \frac{2}{L-1}\left(1 + \sum_{\substack{0<j<L-2 \\ j \text{ even}}}\left(1 + \frac{L-j-1}{j}\right)\right) \\ &= 1 + \frac{2}{L-1} + \frac{2}{L-1}\sum_{\substack{0<j<L-2 \\ j \text{ even}}} \frac{L-1}{j}\end{aligned} \tag{4.1.24}$$

$$= 1 + 2 \sum_{\substack{0<j<L \\ j \text{ even}}} \frac{2}{j},$$

where (4.1.24) follows from the previous step by observing that each $\frac{1}{j}$ is counted

$$\#\{l \text{ odd} : j < \ell < L\} \;=\; \frac{L-j-1}{2} \tag{4.1.25}$$

many times. This completes the induction step, so we have shown (4.1.22). Now since

$$\sum_{\substack{0<j<L \\ j \text{ even}}} \frac{1}{j} \;\sim\; \frac{1}{2}\log(L/2), \tag{4.1.26}$$

we have

$$\mathbb{E}[F_L] \;\sim\; \log L + O(1). \tag{4.1.27}$$

By Markov's inequality,

$$\mathbb{P}(F_L > g(L)\log L) \;\leq\; \frac{\log L + O(1)}{g(L)\log L} \;=\; O\left(\frac{1}{g(L)}\right). \tag{4.1.28}$$

The proof of the lower bound follows the exact same reasoning as the proof of Lemma 3.4 in [B+]. □

Corollary 4.5. *Let k_L be the total number of sticks when the process $\mathcal{P}$ ends. Let*

$$k'_L \;=\; |\{k : \ell_k \geq \log^3(L)\}|. \tag{4.1.29}$$

Then with probability going to 1,

$$\lim_{L\to\infty} \frac{k'_L}{k_L} \;=\; 1. \tag{4.1.30}$$

Moreover, for all k such that $\ell_k \geq \log^3(L)$, we have $Y_{k+1} \to \infty$ as $L \to \infty$ uniformly with probability going to 1.

Proof. The following argument is essentially the same as the one given in the proof of [B+, Corollary 3.5]. We include it here for completeness. Note that $k_L - k'_L$ is the number of sticks generated after ℓ_k first becomes smaller than $\log^3(L)$, and is thus upper bounded by $\log(\log^3(L))g(\log^3(L))$ with probability going to 1 by Lemma 4.4. On the other hand, $k_L > (\log\log L)^2$ with probability going to 1. Therefore as $g(L) = o(\sqrt{\log(L)})$,

$$\lim_{L\to\infty} \frac{k'_L}{k_L} \;=\; 1 - \lim_{L\to\infty} \frac{k_L - k'_L}{k_L} \;>\; 1 - \frac{\log(\log^3(L))g(\log^3(L))}{(\log\log(L))^2} \;=\; 1 \tag{4.1.31}$$

with probability going to 1. To prove the second part of the Corollary, note that, for k such that $\ell_k \geq \log^3(L)$,

$$c_k \geq \frac{1}{g(L)\log^2(L)} \qquad \Longrightarrow \qquad X_{k+1} \geq \frac{\log L}{g(L)} \tag{4.1.32}$$

which approaches infinity. The probability of the former occurring for all such k is

$$\left(1 - \frac{1}{g(L)\log^2(L)}\right)^{g(L)\log L} \;=\; (1 - o(1))e^{-\frac{1}{\log L}} \;\to\; 1. \tag{4.1.33}$$

Thus we immediately deduce the same holds for Y_{k+1} in view of Lemma 4.1. This completes the proof. □

We have verified that all conditions required in Corollary 4.2 are satisfied with probability going to 1, so we are done via the following lemma.

Lemma 4.6. *Given that* $k'_L = (1+o_L(1))k_L$ *with probability* $1-o_L(1)$, $\mathcal{X}_L = \left\{X_1^{(L)}, X_2^{(L)}, \dots, X_{k_L}^{(L)}\right\}$ *approaches strong Benford behavior if and only if* $\mathcal{X}'_L = \left\{(X'_1)^{(L)}, (X'_2)^{(L)}, \dots, (X'_{k'_L})^{(L)}\right\}$ *does.*

Proof. Without loss of generality, assume $k'_L \le k_L$. Let

$$P_L(s) = \frac{\#\{1 \le i \le k_L : S_B(X_i^{(L)}) \le s\}}{k_L},$$

and

$$P'_L(s) = \frac{\#\{1 \le i \le k'_L : S_B((X'_i)^{(L)}) \le s\}}{k'_L}.$$

Then

$$|P_L(s) - P'_L(s)| \le \frac{k_L - k'_L}{k'_L} = o_L(1)$$

with probability $1 - o_L(1)$. So

$$\mathbb{E}[|P_L(s) - P'_L(s)|] \le (1 - o_L(1)) \cdot o_L(1) + o_L(1) \cdot 1 = o_L(1).$$

Therefore

$$\lim_{L\to\infty} \mathbb{E}[P_L(s)] = \log_B(s) \iff \lim_{L\to\infty} \mathbb{E}[P'_L(s)] = \log_B(s).$$

Moreover,

$$\begin{aligned} & P_L(s) - o_L(s) \le P'_L(s) \le P_L(s) + o_L(s) \\ \implies & (P_L(s) - o_L(s))^2 \le P'_L(s)^2 \le (P_L(s) + o_L(s))^2 \\ \implies & |P_L(s)^2 - P'_L(s)^2| = o_L(s). \end{aligned}$$

Therefore

$$\lim_{L\to\infty} \mathrm{Var}[P_L(s)] = 0 \iff \lim_{L\to\infty} \mathrm{Var}[P'_L(s)] = 0.$$

□

4.2. **Proof of Theorem 1.5.** For any integer $\ell > 1$, $r \in \{0, \dots, n-1\}$, let

$$p_r(\ell) \;=\; \frac{|(n\mathbb{Z} + r) \cap [1, \dots, \ell - 1]|}{\ell - 1}. \tag{4.2.1}$$

In other words, $p_r(\ell)$ is the proportion of integers between 1 and $\ell - 1$ falling into the residue class r modulo n. Note that

$$\frac{1}{n} - \frac{1}{\ell - 1} \;\le\; p_r(\ell) \;\le\; \frac{1}{n} + \frac{1}{\ell - 1} \tag{4.2.2}$$

for all r. Define a discrete distribution $\mathcal{D}_\ell$ on $\{0, \dots, n-1\}$ by

$$\mathbb{P}(X_\ell = r) \;=\; p_r(\ell). \tag{4.2.3}$$

Fix starting stick length $L \in \mathbb{Z}_+ \backslash \mathfrak{S}$. We define a discrete process $\mathcal{P}$, and a continuous process $\mathcal{Q}$ that depends on $\mathcal{P}$ as follows.

- In both processes, we start with a stick of the same integer length $L > 1$. Both starting sticks are assumed to be alive. (Since we are defining the process recursively, assume that at the start of each level, every living stick in $\mathcal{Q}$ uniquely corresponds to a living stick in $\mathcal{P}$ and vice versa. This is clearly true in the first level. We will see from our construction that this property is always preserved.)

- At each level, for each living stick in $\mathcal{P}$ of length ℓ, choose a random ratio $p \in (0,1)$ uniformly and a residue class $r \in \{0, \ldots, n-1\}$ with respect to the distribution $\mathcal{D}_\ell$. Suppose $m = |(n\mathbb{Z} + r) \cap [1, \ldots, \ell - 1]|$. Let X be the $(\lfloor mp \rfloor + 1)$-th smallest integer in $[1, \ldots, \ell - 1]$ with residue r modulo n.
- Cut the stick in $\mathcal{P}$ into pieces of lengths X and $\ell - X$, and cut the corresponding stick (of length h) in $\mathcal{Q}$ into pieces of lengths ph and $(1-p)h$.
- Now, in process $\mathcal{P}$, any new stick generated becomes dead if its length is in $\mathfrak{S}$, and in this case the corresponding stick in $\mathcal{Q}$ dies, too.
- Continue to the next level until all sticks die.

By choosing the ratio p and the residue class r of X independently, we ensure that dead/alive status of a new stick in either process is independent of the ratio p used to generate its length. In particular, in the continuous process, the probability that a new stick dies is always close to $1/2$ with an error of at most $\frac{n+4}{2(\ell-1)}$ (sum over $n/2$ residues and then an error of $\frac{2}{\ell-1}$ to account for stopping at length 1).

We want to argue the following:

(1) The continuous process $\mathcal{Q}$ thus constructed is "close" to the process in Theorem 3.1, and thus results in strong Benford behavior.

(2) For almost all pairs of corresponding ending sticks X_k, Y_k in $\mathcal{P}$, $\mathcal{Q}$ respectively, we have

$$Y_k(1 - o(1)) \ \leq \ X_k \ \leq \ Y_k(1 + o(1)) \tag{4.2.4}$$

as $L \to \infty$, so that we can apply Lemma 4.3 to argue that $\mathcal{P}$ is Benford.

4.2.1. *Proof of First Item.*

Lemma 4.7. *Let T_i be the number of living sticks at level i of length at least $\frac{L}{(\log L)^i}$. Given that $L > (n+5)(\log L)^j$, $\log L > 10j$, and $R > 2(\log L)^2$, we have that*

$$\mathbb{P}\left(T_i \geq R\left(1 - \frac{5i}{\log L}\right) \ \forall\ 0 \leq i \leq j\right) \ \geq \ \left(1 - \frac{2(\log L)^2}{R}\right)^j \ \geq \ 1 - \frac{2j(\log L)^2}{R}. \tag{4.2.5}$$

Proof. We proceed with induction on j. The result is clearly true for $j = 0$ since $T_0 = R$. We show the result for j implies that for $j+1$. From now on we condition on the history up to the j-th level. Consider a stick at level j of length at least $\frac{L}{(\log L)^j}$. For each of its children, the probability of being shorter than $\frac{L}{(\log L)^{j+1}}$ is at most $\frac{1}{\log L}$ and the probability of being alive is at least

$$\frac{1}{2} - \frac{n+4}{2\left(\frac{L}{(\log L)^j} - 1\right)} \ \geq \ \frac{1}{2} - \frac{n+4}{2(n+4)\log L} \ = \ \frac{1}{2} - \frac{1}{2\log L}. \tag{4.2.6}$$

So the probability that the child is both of length at least $\frac{L}{(\log L)^{j+1}}$ and alive is bounded below by

$$\mathbb{P}(\text{alive}) - \mathbb{P}\left(\text{length} \leq \frac{L}{(\log L)^{j+1}}\right) \geq \frac{1}{2} - \frac{1}{2\log L} - \frac{1}{\log L} = \frac{1}{2} - \frac{3}{2\log L}. \tag{4.2.7}$$

Let T'_{j+1} be the number of live sticks at level $j+1$ of length at least $\frac{L}{(\log L)^{j+1}}$ whose parent is of length at least $\frac{L}{(\log L)^j}$. Since there are T_j such parents generating $2T_j$ children in total, summing the above probability over each child, we have that

$$\mathbb{E}(T'_{j+1}|T_j) \;\geq\; 2T_j\left(\frac{1}{2}-\frac{3}{2\log L}\right) \;\geq\; T_j\left(1-\frac{3}{\log L}\right). \tag{4.2.8}$$

Moreover, for each parent, the variance of the number of its alive children that are of length at least $\frac{L}{(\log L)^{j+1}}$ is at most $2^2 = 4$ since it has at most 2 children in total. Also, each sub-process starting from one of these parents is independent from another, so the total variance $\mathrm{Var}(T'_{j+1}) \leq 4T_j$. Then, conditioning on the history of the process up to the j-th level, by Chebyshev's inequality,

$$\mathbb{P}\left(T_{j+1} < T_j\left(1-\frac{5}{\log L}\right)\right) \;\leq\; \mathbb{P}\left(|T'_{j+1}-\mathbb{E}(T'_{j+1})| > T_j\frac{2}{\log L}\right) \;<\; \frac{4T_j}{\frac{4T_j^2}{(\log L)^2}} \;<\; \frac{2(\log L)^2}{R}, \tag{4.2.9}$$

where the last inequality is true with probability $\left(1-\frac{2(\log L)^2}{R}\right)^j$ by the induction hypothesis. This implies

$$\mathbb{P}\left(T_{j+1} \;\geq\; T_j\left(1-\frac{5}{\log L}\right)\right) \;\geq\; 1-\frac{2(\log L)^2}{R}. \tag{4.2.10}$$

Notice that

$$\left(1-\frac{5j}{\log L}\right)\left(1-\frac{5}{\log L}\right) \;>\; 1-\frac{5(j+1)}{\log L}, \tag{4.2.11}$$

so that

$$T_{j+1} \;\geq\; T_j\left(1-\frac{5}{\log L}\right) \implies T_{j+1} \;\geq\; R\left(1-\frac{5(j+1)}{\log L}\right) \tag{4.2.12}$$

given that $T_j \geq R(1-\frac{5j}{\log L})$. Thus we have

$$\mathbb{P}\left(T_{j+1} \;\geq\; R\left(1-\frac{5(j+1)}{\log L}\right)\right) \;\geq\; 1-\frac{2(\log L)^2}{R} \tag{4.2.13}$$

given that $T_j \geq R(1-\frac{5j}{\log L})$. This completes the induction step. □

Corollary 4.8. *For sufficiently large L and $R > (\log L)^3$, the number of live sticks ever is bounded below by*

$$R\sqrt{\log L} \tag{4.2.14}$$

with probability at least

$$1-\frac{4(\log L)^{5/2}}{R}. \tag{4.2.15}$$

This is also a lower bound for the number of dead sticks ever, with the same probability.

Proof. Let n_i be the number of alive sticks at level i. First, note that the number of dead sticks generated at level i is $2n_{i-1}-n_i$, and summing this from $i=1$ to infinity yields $2n_0+n_1+n_2+\cdots$ which is bounded below by the total number of live sticks in all levels. Now our lower bound follows from Lemma 4.7 by taking $j = \lfloor 2\sqrt{\log L}\rfloor$ and L large enough so that all assumptions there hold and that

$$\frac{5\cdot\lfloor 2\sqrt{\log L}\rfloor}{\log L} \;<\; \frac{1}{2}. \tag{4.2.16}$$

□

Let M be the total number of dead sticks. We have that, with probability going to 1, $M \geq R\sqrt{\log L}$. Now, we wish to show that $\mathbb{E}[P_{R,L}(s)] \to \log_B(s)$. It suffices to show that $\mathbb{E}[\varphi_s(X_i)] \to \log_B(s)$ uniformly for a proportion of X_i going to 1. We first show that almost all sticks die after $\frac{1}{2}\log\log L$ levels. First, note that there are at most $R2^i$ alive sticks at level i and thus at most $R2^{i-1}\cdot 2 = R2^i$ new dead sticks are generated at level i. Thus, the number of dead sticks generated at or before level j is $\sum_{i=1}^{j} R2^i \leq R2^{j+1}$. Thus, the number of sticks before level $\frac{1}{2}\log\log L$ is at most

$$2R\cdot 2^{(\log\log L)/2} \;=\; 2R(\log L)^{(\log 2)/2} \;=\; o(R\sqrt{\log L}). \tag{4.2.17}$$

That is, the proportion of sticks before level $\frac{1}{2}\log\log L$ goes to 0. Thus, we may assume that X_i dies at a later level. However, it is a product of independent random variables, each chosen from some finite set. Moreover, the length of this product is increasing in L, so by Theorem 3.4, $\mathbb{E}[\varphi_s(X_i)]$ approaches $\log_B(s)$ uniformly, and the conclusion follows.

We now wish to show that $\mathrm{Var}[P_{R,L}(s)] \to 0$. The same strategy as in the proof of Theorem 3.6 works with slight modifications that we highlight below. Recall that the goal is to show that

$$\frac{1}{M^2}\sum_{i,j}\mathbb{E}[\varphi_s(X_iX_j)] \;\to\; \log_B(s)^2, \tag{4.2.18}$$

where X_i, X_j denote a pair of dead sticks. Based on our observation above, we may restrict our attention to the collection of pairs of sticks only involving those that die after at least $\frac{1}{2}\log\log L$ levels. Note that running the exact same argument as in the proof of Theorem 3.6 with $k = 2$, we obtain that the number of pairs with high dependency (as described in (2) in that proof) is bounded above by $M(\log R)^{1+\log 2} = o(M^2)$, so we are done.

4.2.2. *Proof of Second Item.*

Lemma 4.9. *At each level of $\mathcal{P}$, given that a stick of length ℓ breaks into sticks of lengths X and $\ell - X$, with ratio p in process $\mathcal{Q}$, we have*

$$\left|\frac{X}{\ell} - p\right| \;\leq\; \frac{n+1}{\ell}. \tag{4.2.19}$$

This also implies that

$$\left|\frac{\ell - X}{\ell} - (1-p)\right| \;\leq\; \frac{n+1}{\ell}, \tag{4.2.20}$$

so we have the same bound for the error between the corresponding ratios in $\mathcal{P}$ and $\mathcal{Q}$ regardless of which child we look at.

Proof. We prove this for $r \neq 0$. Since $m = \lfloor \frac{\ell-1-r}{n} \rfloor + 1$, we have

$$\frac{\ell-1-r}{n} \;\leq\; m \;\leq\; \frac{\ell-1-r}{n} + 1.$$

Now $X = \lfloor pm \rfloor n + r$ whenever $r \neq 0$. (Note that here if $r = 0$, we have $m = \lfloor \frac{\ell-1}{n} \rfloor$ and $X = \lfloor pm \rfloor n + n$ instead.) So

$$\begin{aligned} p(\ell-1-r) - n + r \;&\leq\; X \;\leq\; p(\ell-1-r) + pn + r \\ \implies p - \frac{p+pr+n-r}{\ell} \;&\leq\; \frac{X}{\ell} \;\leq\; p + \frac{pn+r-p-pr}{\ell}. \end{aligned} \tag{4.2.21}$$

Notice that

$$|p+pr+n-r| \;=\; |n+p-(1-p)r| \;\leq\; n+1 \tag{4.2.22}$$

and

$$|pn+r-p-pr| \;=\; |p(n-1)+(1-p)r| \;\leq\; n, \tag{4.2.23}$$

so we have the desired. One easily verifies the result for $r = 0$ following a similar calculation. □

Corollary 4.10. *Consider a pair of sticks (ℓ_j, h_j) at level $j \geq 1$, where ℓ_j is in process $\mathcal{P}$ and h_j is the corresponding one in process $\mathcal{Q}$. Denote their ancestors as (ℓ_i, h_i) for $0 \leq i \leq j-1$, with $\ell_0 = h_0 = L$. Suppose $h_{i+1} = p_i h_i$ for all $0 \leq i \leq j-1$. Then we have*

$$h_j \prod_{i=0}^{j-1} \left(1 - \frac{n+1}{p_i \ell_i}\right) \;\leq\; \ell_j \;\leq\; h_j \prod_{i=0}^{j-1} \left(1 + \frac{n+1}{p_i \ell_i}\right). \tag{4.2.24}$$

Proof. By Lemma 4.9, we have for all $1 \leq i \leq j$,

$$p_{i-1} \left(1 - \frac{n+1}{p_{i-1}\ell_{i-1}}\right) \;\leq\; \frac{\ell_i}{\ell_{i-1}} \;\leq\; p_{i-1} \left(1 + \frac{n+1}{p_{i-1}\ell_{i-1}}\right), \tag{4.2.25}$$

and the corollary follows by taking the product over all such i. □

Corollary 4.11.

$$h_j \prod_{i=1}^{j} \left(1 - \frac{n+1}{\ell_i - n - 1}\right) \;\leq\; \ell_j \;\leq\; h_j \prod_{i=1}^{j} \left(1 + \frac{n+1}{\ell_i - n - 1}\right). \tag{4.2.26}$$

Proof. This follows from Corollary 4.10 using the lower bound

$$p_{i-1}\ell_{i-1} \;\geq\; \ell_i - n - 1 \tag{4.2.27}$$

which follows from Lemma 4.9. □

Lemma 4.12. *Let $f(L), g(L), h(L)$ be some functions in L that go to infinity as $L \to \infty$ with $g(L) = o(f(L))$. Then for any dead stick ℓ_j with $j < g(L)$, if $\ell_j > f(L)+n+1$ and the corresponding sticks $h_j > h(L)$, we have*

$$h_j(1 - o(1)) \;\leq\; \ell_j \;\leq\; h_j(1 + o(1)). \tag{4.2.28}$$

Proof. From Corollary 4.2, we have

$$\ell_j \;\geq\; h_j \left(1 - \sum_{i=1}^{j} \frac{n+1}{\ell_i - n - 1}\right) \;\geq\; h_j \left(1 - g(L)\frac{n+1}{f(L) - n - 1}\right) \;=\; h_j(1 - o(1)). \tag{4.2.29}$$

For the upper bound, we have that

$$\ell_j \;\leq\; h_j \prod_{i=1}^{j} \left(1 + \frac{n+1}{\ell_i - n - 1}\right) \;\leq\; h_j \left(1 + \frac{n+1}{f(L) - n - 1}\right)^{g(L)}. \tag{4.2.30}$$

As $L \to \infty$, the expression above multiplying h_j approaches

$$\lim_{L\to\infty} \exp\left(g(L)\frac{n+1}{f(L) - n - 1}\right) \;=\; 1 \tag{4.2.31}$$

so $\ell_j \leq h_j(1 + o(1))$, as desired. □

Now the goal is to determine f and g so that

$$\mathbb{P}(\text{A stick dies within } g(L) \text{ levels}) \;=\; 1 - o(1) \tag{4.2.32}$$

and

$$\lim_{L\to\infty} \frac{\#\text{dead sticks with length larger than } f(L)+n+1}{\#\text{all dead sticks}} \;=\; 1. \tag{4.2.33}$$

The intuition is that we want to show most sticks die within the first $g(L)$ levels, and that most sticks that ever occur are long, i.e., larger than $f(L)$.

Lemma 4.13. *Let M be the number of dead sticks ever in a process starting with R sticks of length L. Then*

$$\mathbb{P}(M < R(\log L)\nu(L)) \;\to\; 1 \tag{4.2.34}$$

as $L \to \infty$, where $\nu(L)$ is any function that goes to infinity as $L \to \infty$.

Proof. Let M_L be the number of dead sticks resulting from the process of breaking a single stick of length L. We have that $M_L = 1$ whenever $L \in \mathfrak{S}$. We prove by induction on L that when $L \notin \mathfrak{S}$

$$\mathbb{E}[M_L] \;\le\; 6n^2 \log L. \tag{4.2.35}$$

(Here $\log(x)$ is short-hand for $\log_e(x)$.) When $1 < L \le 3n^2$ this is clear since

$$6n^2 \log L \;\ge\; 3n^2 \cdot 2\log(2) \;\ge\; 3n^2 \;\ge\; L \tag{4.2.36}$$

and $M_L \le L$ for all L.

When $L > 3n^2$ and $L \notin \mathfrak{S}$, we have

$$\begin{aligned}
\mathbb{E}[M_L] \;&=\; \frac{1}{L-1} \sum_{1\le \ell \le L-1} \left(\mathbb{E}[M_\ell] + \mathbb{E}[M_{L-\ell}]\right) \\
&=\; \frac{2}{L-1} \sum_{1\le \ell \le L-1} \mathbb{E}[M_\ell] \\
&\le\; \left(\frac{2}{L-1} \sum_{\substack{1\le \ell \le L-1 \\ \ell \in \mathfrak{S}}} 1 \right) + \left(\frac{2}{L-1} \sum_{\substack{1\le \ell \le L-1 \\ \ell \notin \mathfrak{S}}} 6n^2 \log(\ell) \right) \\
&\le\; \frac{2}{L-1}\left(\frac{L-1}{2} + \frac{n}{2} + 1\right) + \frac{12n^2}{L-1} \log \left(\prod_{\substack{1\le \ell \le L-1 \\ \ell \notin \mathfrak{S}}} \ell \right) \\
&\le\; 1 + \frac{n+2}{L-1} + 6n^2 \log \left(\left(\prod_{\substack{1\le \ell \le L-1 \\ \ell \notin \mathfrak{S}}} \ell \right)^{\frac{2}{L-1}} \right) \\
&\le\; 1 + 6n^2 \log L.
\end{aligned} \tag{4.2.37}$$

Note that we used the fact that

$$|[1, L-1] \cap \mathfrak{S}| \;\le\; \left(\left\lfloor \frac{L-1}{n} \right\rfloor + 1 \right) \cdot \frac{n}{2} + 1 \;\le\; \frac{L-1}{2} + \frac{n}{2} + 1. \tag{4.2.38}$$

To see the last inequality,

$$
\begin{aligned}
1+\frac{n+2}{L-1}+6n^2\log\left(\left(\prod_{\substack{1\le\ell\le L-1\\ \ell\notin\mathfrak{S}}}\ell\right)^{\frac{2}{L-1}}\right) \;&\le\; 1+6n^2\log L \\
\iff \frac{n+2}{6n^2}+\log\left(\left(\prod_{\substack{1\le\ell\le L-1\\ \ell\notin\mathfrak{S}}}\ell\right)^{2}\right) \;&\le\; (L-1)\log L \\
\iff e^{(n+2)/(6n^2)} \;&\le\; \frac{L^{L-1}}{\left(\prod_{\substack{1\le\ell\le L-1\\ \ell\notin\mathfrak{S}}}\ell\right)^2} \\
\Longleftarrow e^{1/(3n)}n^n \;&\le\; \frac{L^{L-1}}{\left(\prod_{\substack{n+1\le\ell\le L-1\\ \ell\notin\mathfrak{S}}}\ell\right)^2}. \qquad (4.2.39)
\end{aligned}
$$

Note that on the RHS, the product has at most $\frac{L-1}{2}$ terms and the first $n/2$ terms are at most $2n$, so we have

$$\frac{L^{L-1}}{\left(\prod_{\substack{n<\ell\le L-1\\ \ell\notin\mathfrak{S}}}\ell\right)^2} \;\ge\; \frac{L^n}{(2n)^n}\cdot\frac{L^{L-n-1}}{\left(\prod_{\substack{2n<\ell\le L-1\\ \ell\notin\mathfrak{S}}}\ell\right)^2} \;\ge\; (3n)^n \;\ge\; e^{1/(3n)}n^n. \qquad (4.2.40)$$

Thus the induction step is complete. By Markov's inequality,

$$\mathbb{P}(M>R(\log L)\nu(L)) \;\le\; \frac{R\mathbb{E}[M_L]}{R(\log L)\nu(L)} \;\le\; \frac{6n^2R\log L}{R(\log L)\nu(L)} \;=\; O\left(\frac{1}{\nu(L)}\right) \;\to\; 0 \qquad (4.2.41)$$

as $L\to\infty$. □

Since at each level before the process ends, the number of sticks increase by at least 1, we have that the total number of levels at most $R(\log L)\nu(L)$ with probability going to 1 as $L\to\infty$. Thus we can take

$$g(L) \;=\; R(\log L)\nu(L). \qquad (4.2.42)$$

Lemma 4.14. *Let $M_{\ell,k}$ denote the number of dead sticks with length smaller than k coming from a process starting with a stick of length ℓ. Let $c=24n^2$. Then for any $k\ge 2n,\ell>1$, we have*

$$\mathbb{E}[M_{\ell,k}] \;\le\; c\log(k). \qquad (4.2.43)$$

In particular,

$$\mathbb{E}[M_{L,\log^2(L)}] \;\le\; 2c\log\log L. \qquad (4.2.44)$$

Proof. When $\ell\le k$, we have trivially

$$\mathbb{E}[M_{\ell,k}] \;=\; \mathbb{E}[M_\ell] \;\le\; \frac{c}{4}\log(\ell) \;\le\; \frac{c}{4}\log(k) \qquad (4.2.45)$$

by (4.2.35). When $\ell>k$ and $\ell\in\mathfrak{S}$, we have

$$\mathbb{E}[M_{\ell,k}] \;=\; 0. \qquad (4.2.46)$$

For a fixed k, we prove the result by induction on ℓ.

$$\begin{aligned}
\mathbb{E}[M_{\ell,k}] &= \frac{1}{\ell-1}\sum_{1\le x\le \ell-1}(\mathbb{E}[M_{x,k}]+\mathbb{E}[M_{\ell-x,k}]) \\
&= \frac{2}{\ell-1}\sum_{1\le x\le \ell-1}\mathbb{E}[M_{x,k}] \\
&= \frac{2}{\ell-1}\left(\sum_{1\le x\le k}\mathbb{E}[M_{x,k}]+\sum_{\substack{x\notin\mathfrak{S}\\ k<x\le \ell-1}}\mathbb{E}[M_{x,k}]\right) \\
&\le \frac{2}{\ell-1}\left(k\cdot\frac{c}{4}\log(k)+\frac{\ell-k-1+n}{2}\cdot c\log(k)\right) \\
&= c\log(k)\frac{\frac{1}{2}k+(\ell-k-1+n)}{\ell-1} \\
&\le c\log(k), \qquad (4.2.47)
\end{aligned}$$

where the last step uses $\frac{k}{2}\ge n$. □

Corollary 4.15. *Let ℓ_i denote a stick occurring at level i in process $\mathcal{P}$.*

$$\frac{\#\{\ell_i>\log^2(L): i\le g(L)\}}{\#\{\ell_i : i\le g(L)\}} \;\to\; 1 \qquad (4.2.48)$$

as $L\to\infty$ with probability going to 1.

Proof. We have from Lemma 4.14 that

$$\mathbb{E}[\#\{\ell_i\le\log^2(L): i\le g(L)\}] \;\le\; R\cdot 2c\log\log L. \qquad (4.2.49)$$

By Markov's inequality,

$$\mathbb{P}(\#\{\ell_i\le\log^2(L): i\le g(L)\} > R(\log L)^{1/3}) \;\le\; \frac{R\cdot 2c\log\log L}{R(\log L)^{1/3}} \;\to\; 0 \qquad (4.2.50)$$

as $L\to\infty$. In other words,

$$\#\{\ell_i>\log^2(L): i\le g(L)\} \;<\; R(\log L)^{1/3} \qquad (4.2.51)$$

with probability going to 1. On the other hand, by Corollary 4.8, we have as long as $R>(\log L)^3$,

$$\#\{\ell_i : i\le g(L)\} \;\ge\; R(\log L)^{1/2} \qquad (4.2.52)$$

with probability going to 1. Since

$$\frac{R(\log L)^{1/3}}{R(\log L)^{1/2}} \;\to\; 0 \qquad (4.2.53)$$

as $L\to\infty$, we have the desired. □

Now to see that item (2) is true, it suffices to show that the premises of Lemma 4.12 are satisfied for most dead sticks. By Lemma 4.13, almost all sticks die within the first $g(L)=R(\log L)\nu(L)$ levels, where $\nu(L)$ is any function that blows up as $L\to\infty$. By Corollary 4.15, almost all dead sticks are at least $f(L)=\log^2(L)$ in length. We can choose ν such that $g(L)=o(f(L))$. This completes the proof of Proposition 1.5. □

5. Non-Benfordness Results

In this section, we prove non-Benfordness of the decomposition processes when the stopping condition deviates from the required ones, namely "$|S| = n/2$" in Theorem 1.2 and "$r = 1/k$" in Theorem 3.1, thus showing that the conditions we impose in those theorems are necessary.

5.1. Continuous Case: When $r \neq 1/k$.

Recall the continuous stick breaking process described in Section 3 and the notation there. We claimed that when $r \neq 1/k$, the limiting distribution of the stick lengths is non-Benford. We now prove precise statements of what happens in those cases.

Theorem 5.1. *When $r > 1/k$, there is positive probability that the process with $R = 1$ does not end in finitely many levels.*

Proof. Let A be some integer that is sufficiently large (we can determine what this means later). There then exists some fixed j such that $n_j > A$ with positive probability p^*. Now, consider $i \geq j$. Conditioning on the event $n_i = m$, we have that n_{i+1} is a random variable with mean mrk and variance $mr(1-r)$. Thus, by Chebyshev's inequality, we have that

$$\mathbb{P}\left(n_{i+1} > n_i\left(1 + \frac{rk-1}{2}\right) \;\Big|\; n_i\right) \;\geq\; 1 - \mathbb{P}\left(|n_{i+1} - n_i rk| \geq n_i \frac{rk-1}{2} \;\Big|\; n_i\right) \;\geq\; 1 - \frac{n_i r(1-r)}{n_i^2\left(\frac{rk-1}{2}\right)^2}. \tag{5.1.1}$$

We can then let $a = \frac{r(1-r)}{A\left(\frac{rk-1}{2}\right)^2}$ and $c = 1 + \frac{rk-1}{2}$. Then the above inequality can be written as

$$\mathbb{P}(n_{i+1} > cn_i \mid n_i) \;\geq\; 1 - \frac{aA}{n_i}. \tag{5.1.2}$$

It follows that

$$\begin{aligned}\mathbb{P}(n_{i+1} > Ac^{i-j+1} \mid n_i > Ac^{i-j}) \;&\geq\; \mathbb{P}(n_{i+1} > cn_i \mid n_i > Ac^{i-j}) \\ &\geq\; \inf_{n_i > Ac^{i-j}}\left(1 - \frac{aA}{n_i}\right) \;\geq\; 1 - ac^{j-i}.\end{aligned} \tag{5.1.3}$$

Hence, the probability that $n_i > Ac^{i-j}$ for all $i \geq j$ given that $n_j > A$ is at least

$$p' \;=\; (1-a)(1-ac^{-1})(1-ac^{-2})\cdots. \tag{5.1.4}$$

Now, since $\lim_{x\to 0}\log(1-x)/x = -1$, we may set A large enough so that a is sufficiently small such that $\log(1-ac^t) > -2ac^t$ for $t \leq 0$. We then have

$$\log(p') \;=\; \sum_{t=0}^{\infty}\log(1-ac^{-t}) \;>\; \sum_{t=0}^{\infty} -2ac^{-t} \;=\; -\frac{2a}{1-\frac{1}{c}}. \tag{5.1.5}$$

In particular, $p' \geq e^{-2a/(1-1/c)} > 0$. Thus the probability that $n_i > Ac^{i-j}$ for all $i \geq j$ is at least p^*p' which is positive. Hence, not only is the process infinite with positive probability, but also the number of alive sticks at each level blows up with positive probability.

□

Theorem 5.2. *When $r < 1/k$ and the alive/dead status of the children of the same stick are possibly dependent on one another, the collection of stick lengths does not converge to strong Benford behavior for sufficiently large bases B.*

Proof.

Lemma 5.3. *We have that*

$$\mathbb{E}[M_R] \;=\; R\frac{k-kr}{1-kr} \tag{5.1.6}$$

Proof. Let p_i be the probability that exactly i of the children of the first stick are alive. Then,

$$\begin{aligned}\mathbb{E}[M_1] \;&=\; \sum_{i=0}^{k} p_i(\mathbb{E}[M_i]+k-i) \;=\; \sum_{i=0}^{k} p_i(i\mathbb{E}[M_1]+k-i)\\ &=\; k+(\mathbb{E}[M_1]-1)\sum_{i=0}^{k} ip_i \;=\; k-kr+kr\mathbb{E}[M_1]\end{aligned} \tag{5.1.7}$$

so that

$$\mathbb{E}[M_1] \;=\; \frac{k-kr}{1-kr}. \tag{5.1.8}$$

Linearity of expectation then implies the result. □

Corollary 5.4. *We have that*

$$M_R \;\le\; 2R\frac{k-kr}{1-kr} \tag{5.1.9}$$

with probability at least $1/2$.

Proof. This follows directly from Lemma 5.3 and Markov's inequality. □

Lemma 5.5. *Let $a>1$ be some real number and let b_a be the expected number of child sticks that are of length at least L/a starting from a stick of length L. With probability at least*

$$1-\frac{4k^2}{b_a^2(1-r)^2R} \tag{5.1.10}$$

the number of dead sticks of length at least L/a in the first level is at least $b_a(1-r)R/2$.

Proof. Denote this quantity by $M^R_{L,a}(1)$. Then, the probability of a child with length at least L/a being dead is $1-r$. Thus,

$$\mathbb{E}[M^R_{L,a}(1)] \;=\; Rb_a(1-r). \tag{5.1.11}$$

Note that $M^R_{L,a}(1)$ is a sum of independent random variables distributed identically to $M^1_{L,a}(1)$, and $\mathrm{Var}[M^1_{L,a}(1)] \le k^2$, so

$$\mathrm{Var}[M^R_{L,a}(1)] \;\le\; Rk^2. \tag{5.1.12}$$

Chebyshev's inequality then implies

$$\mathbb{P}\left(M^R_{L,a}(1) \le \frac{b_a}{2}(1-r)R\right) \;\le\; \frac{Rk^2}{(Rb_a(1-r)/2)^2} \;=\; \frac{4k^2}{b_a^2(1-r)^2R}. \tag{5.1.13}$$

□

By Lemma 5.5 and Corollary 5.4, the proportion of sticks with length at least L/a is at least

$$\frac{b_a}{2}(1-r)R\left(2R\frac{k-kr}{1-kr}\right)^{-1} \;=\; \frac{b_a}{4}(1-r)\frac{1-kr}{k-kr} \;=\; \frac{b_a(1-kr)}{4k} \tag{5.1.14}$$

with probability at least

$$\frac{1}{2}-\frac{4k^2}{b_a^2(1-r)^2R}. \tag{5.1.15}$$

Now, choose $a = k$. Then, we have that, with some probability approaching $1/2$, at least a proportion of $b_k(1-kr)/(4k)$ of the sticks are of length at least L/k. Moreover, $b_k > 0$ so this proportion is positive. Let

$$B \;>\; k^{\frac{4k}{b_k(1-kr)}}. \tag{5.1.16}$$

Then, these sticks occupy an interval of length

$$\log_B(k) \;<\; \frac{b_k(1-kr)}{4k} \;<\; \frac{1}{4} \tag{5.1.17}$$

in the distribution of the normalized mantissas. It follows that the mantissas of the stick lengths do not almost surely approach a uniform distribution as $R \to \infty$. That is, the stick lengths do not approach Benford behavior. More precisely, this contradicts the first condition for strong Benford behavior. □

5.2. **Discrete Case: When $|S| \neq n/2$.** Now we turn to the setting of the discrete stick fragmentation as in our main theorem (Proposition 1.5) and prove a result showing that when $|S| < n/2$, the final stick lengths are non-Benford. Moreover, we state a conjecture on the behavior of the limiting distribution when $|S| \neq n/2$. Simulation results are also presented to support our conjecture.

Theorem 5.6. *If $|S| < n/2$, then as $R \to \infty$ and $L \to \infty$, the collection of mantissas of ending stick lengths does not converge to strong Benford behavior.*

Theorem 5.7. *If $|S| > n/2$, then as $R \to \infty$ and $L \to \infty$ with the condition $R = \omega(L^2)$, the collection of mantissas of ending stick lengths does not converge to the uniform distribution on $[0,1]$ provided that the base B is greater than $3^{6n^3/|S|}$.*

5.2.1. *Proof of Theorem 5.6.*

Lemma 5.8. *Let*

$$M_{L,m} \;:=\; \#\textit{dead sticks generated by a stick of length } L \textit{ that are of length less than } m. \tag{5.2.1}$$

Then for all $L \notin \mathfrak{S}$, there exists constants m, c only depending on k and n such that

$$\mathbb{E}(M_{L,m}) \;\geq\; c\mathbb{E}(M_L) + 1. \tag{5.2.2}$$

Proof. Let

$$c = \frac{1}{2n+1}, \quad m = 2n^2.$$

For $L \leq m$, the result is clear since $\mathbb{E}(M_{L,m}) = \mathbb{E}(M_L) \geq 2$. We now proceed with induction and assume the result is true for positive integers less than $L > m$. We have,

$$\begin{aligned}
\mathbb{E}(M_{L,m}) \;&=\; \frac{2}{L-1}\sum_{\ell=1}^{L-1}\mathbb{E}(M_{\ell,m}) \\
&=\; \frac{2}{L-1}\sum_{\substack{1\leq \ell \leq L-1 \\ \ell \notin \mathfrak{S}}}\mathbb{E}(M_{\ell,m}) \\
&\geq\; \frac{2}{L-1}\sum_{\substack{1\leq \ell \leq L-1 \\ \ell \notin \mathfrak{S}}} c\mathbb{E}(M_\ell) + \frac{2}{L-1}|[L-1] \setminus \mathfrak{S}|.
\end{aligned} \tag{5.2.3}$$

We now show that

$$|[L-1]\setminus\mathfrak{S}| \;\geq\; \frac{L-1}{2}+c|[L-1]\cap\mathfrak{S}|. \tag{5.2.4}$$

Note that

$$|[L-1]\setminus\mathfrak{S}| \;\geq\; (n-|S|)\left(\frac{L-n-1}{n}\right) \;\geq\; \frac{(n+1)(L-n-1)}{2n} \tag{5.2.5}$$

and

$$|[L-1]\cap\mathfrak{S}| \;\leq\; |S|\left(\frac{L+n-1}{n}\right) \;\leq\; \frac{(n-1)(L+n-1)}{2n} \tag{5.2.6}$$

so it suffices to show that

$$\begin{aligned}\frac{(n+1)(L-n-1)}{2n} \;&\geq\; \frac{L-1}{2}+c\frac{(n-1)(L+n-1)}{2n}\\ \iff c \;&\leq\; \frac{(n+1)(L-n-1)-n(L-1)}{(n-1)(L+n-1)}.\end{aligned} \tag{5.2.7}$$

We have

$$\begin{aligned}\frac{(n+1)(L-n-1)-n(L-1)}{(n-1)(L+n-1)} \;&\geq\; \frac{L-1-n^2-n}{(n-1)(L-1+n)}\\ &\geq\; \frac{(2n^2-n^2-n}{(n-1)(2n^2+n-1)}\\ &\geq\; \frac{n}{2n^2+n-1}\\ &\geq\; \frac{1}{2n+1} \;=\; c.\end{aligned} \tag{5.2.8}$$

Hence, (5.2.4) is true, and can be plugged into (5.2.3) to obtain

$$\begin{aligned}\mathbb{E}(M_{L.m}) \;&\geq\; \frac{2}{L-1}\sum_{\substack{1\leq \ell\leq L-1\\ \ell\notin\mathfrak{S}}} c\mathbb{E}(M_\ell)+\frac{2c}{L-1}|[L-1]\cap\mathfrak{S}|+1\\ &\geq\; 1+\frac{2}{L-1}\sum_{\ell=1}^{L-1} c\mathbb{E}(M_\ell)\\ &\geq\; c\mathbb{E}(M_L)+1.\end{aligned} \tag{5.2.9}$$

The induction is complete. ☐

Lemma 5.9. *Let M_L^R be the total number of dead sticks coming from a process of breaking R identical sticks of length L, and $M_{L,m}^R$ be the number of those shorter than m. Then for m and c satisfying the conclusion of Lemma 5.8, we have as $R\to\infty$,*

$$\mathbb{P}\left(\frac{M_{L,m}^R}{M_L^R}\leq\frac{c}{3}\right) \;\to\; 0. \tag{5.2.10}$$

Proof. By Chernoff's inequality, we have

$$\mathbb{P}\left(M_{L,m}^R\leq\frac{1}{2}R\mathbb{E}(M_{L,m})\right) \;\leq\; e^{-R\mathbb{E}(M_{L,m})/8} \tag{5.2.11}$$

and

$$\mathbb{P}\left(M_L^R\;\geq\frac{3}{2}R\mathbb{E}(M_L)\right) \;\leq\; e^{-R\mathbb{E}(M_L)/10}. \tag{5.2.12}$$

So

$$\mathbb{P}\left(M_{L,m}^R \geq \frac{1}{2}R\mathbb{E}(M_{L,m}) \text{ and } M_L^R \leq \frac{3}{2}R\mathbb{E}(M_L)\right) \geq 1 - e^{-R\mathbb{E}(M_{L,m})/8} - e^{-R\mathbb{E}(M_L)/10}. \quad (5.2.13)$$

In that case,

$$\frac{M_{L,m}^R}{M_L^R} \geq \frac{\frac{1}{2}R\mathbb{E}(M_{L,m})}{\frac{3}{2}R\mathbb{E}(M_L)} = \frac{\mathbb{E}(M_{L,m})}{3\mathbb{E}(M_L)} \geq c/3. \quad (5.2.14)$$

Therefore it suffices to show that

$$1 - e^{-R\mathbb{E}(M_{L,m})/8} - e^{-R\mathbb{E}(M_L)/10} \to 0 \quad (5.2.15)$$

as $R \to \infty$. To do this, it again suffices to show that $\mathbb{E}(M_{L,m}) > 0$ and $\mathbb{E}(M_L) > 0$. Clearly, $\mathbb{E}(M_L) \geq 1$, so it follows from Lemma 5.8 that $\mathbb{E}(M_{L,m}) \geq c + 1 > 0$. □

Now to conclude the proof of (2), note that if the process is strong Benford, by Definition 2.2, the collection of mantissas $M_B(X)$ of dead sticks (in some sense) converges to the uniform distribution on $[0, 1]$, which is continuous. More precisely, for any $s \in [1, B)$ and $\epsilon > 0$,

$$\mathbb{P}[M_B(X) = \log_B(s)] \leq \mathbb{E}[\varphi_s(X) - \varphi_{s-\epsilon}(X)] \to \log_B(s) - \log_B(s - \epsilon) \quad (5.2.16)$$

thus $\mathbb{P}[M_B(X) = \log_B(s)] \to 0$ as $L \to \infty$. So we must have that

$$\frac{M_{L,m}^R}{M_L^R} = \frac{\#\{X : M_B(X) \in \{M_B(1), M_B(2), \ldots, M_B(m-1)\}\}}{M_L^R} \to 0 \quad (5.2.17)$$

as $L \to \infty$ with probability going to 1 as $R \to \infty$. This is false by the previous lemma, so the process does not approach strong Benford behavior. In fact, our argument shows that, in some imprecise sense, the collection of mantissas of such a process does not converge to *any* continuous distribution on $[0, 1]$ as $R \to \infty$ and $L \to \infty$.

5.2.2. *Proof of Theorem 5.7.* Let M_L^R be the total number of dead sticks obtained starting from R sticks of length L.

Lemma 5.10. *We have that*

$$\mathbb{E}[M_L^R] \leq 2n^2R. \quad (5.2.18)$$

Proof. We show the result when $R = 1$ via induction on L. Let $M_L^1 = M_L$. The result is clearly true for $L \leq 2n^2$ since $M_L \leq L$, so assume that $L > 2n^2$ and the result holds for all positive integers smaller than L. We have that,

$$\begin{aligned}
\mathbb{E}[M_L] &= \frac{1}{L-1}\sum_{\ell=1}^{L-1}(\mathbb{E}[M_\ell]+\mathbb{E}[M_{L-\ell}]) \\
&= \frac{2}{L-1}\sum_{\ell=1}^{L-1}\mathbb{E}[M_\ell] \\
&\leq 2+\frac{2}{L-1}\sum_{\substack{1\leq \ell\leq L-1 \\ \ell\notin\mathfrak{S}}} 2n^2 \\
&\leq 2+\frac{2}{L-1}\left\lceil\frac{L-1}{n}\right\rceil(n-|S|)\cdot 2n^2 \\
&\leq 2+\frac{2}{L-1}\left(\frac{L+n-1}{n}\right)\left(\frac{n-1}{2}\right)2n^2 \\
&= 2+2n(n-1)\frac{L+n-1}{L-1} \\
&\leq 2+2n(n-1)\frac{2n^2+n}{2n^2} \\
&= 2n^2-n+1 \;\leq\; 2n^2. \qquad (5.2.19)
\end{aligned}$$

The induction is complete. The result for general R follows from linearity of expectation. □

Corollary 5.11. *With probability at least*

$$1-\frac{L^2}{n^4R} \qquad (5.2.20)$$

we have $M_L^R \leq 3n^2R$.

Proof. First, note that, trivially, $\mathrm{Var}[M_L] \leq L^2$ so that $\mathrm{Var}[M_L^R] \leq RL^2$. Then, Chebyshev's inequality implies that

$$\mathbb{P}(M_L^R > 3n^2R) \;\leq\; \mathbb{P}(|M_L^R-\mathbb{E}(M_L^R)| > n^2R) \;\leq\; \frac{RL^2}{(n^2R)^2} \;=\; \frac{L^2}{n^4R}. \qquad (5.2.21)$$

□

Lemma 5.12. *Let $a>2$ be some real number and assume $L>\frac{2an}{a-2}$. With probability at least*

$$1-\frac{16n^2}{|S|^2R} \qquad (5.2.22)$$

the number of dead sticks of length at least L/a in the first level is at least $|S|R/(2n)$.

Proof. Denote this quantity by $M_{L,a}^R(1)$. Then, given a stick of length L, the number of ways the left child can die and be of length at least L/a is bounded below by

$$|S|\left\lfloor\frac{L-L/a}{n}\right\rfloor \;\geq\; |S|\left(\frac{L}{n}\left(1-\frac{1}{a}\right)-1\right) \;\geq\; \frac{|S|L}{2n}. \qquad (5.2.23)$$

Thus, the probability of any arbitrary child being of length at least L/a and dead is at least $|S|/(2n)$. It follows that

$$\mathbb{E}[M_{L,a}^R(1)] \ \geq \ \frac{|S|R}{n}. \tag{5.2.24}$$

Furthermore,

$$\text{Var}[M_{L,a}^R(1)] \ \leq \ 4R \tag{5.2.25}$$

by independence. Thus, we have, by Chebyshev's inequality,

$$\mathbb{P}\left(M_{L,a}^R(1) \leq \frac{|S|R}{2n}\right) \ \leq \ \mathbb{P}\left(\left|M_{L,a}^R(1) - \mathbb{E}[M_{L,a}^R(1)]\right| \geq \frac{|S|R}{2n}\right) \ \leq \ \frac{4R}{\left(\frac{|S|R}{2n}\right)^2} \ = \ \frac{16n^2}{|S|^2R}. \tag{5.2.26}$$

□

Now, set $a = 3$ and let $L > 6n$. Note that by Lemma 5.12 and Corollary 5.11, with probability at least

$$1 - \frac{L^2}{n^4R} - \frac{16n^2}{|S|^2R} \tag{5.2.27}$$

the proportion of dead sticks of length at least $L/3$ is bounded below by

$$\frac{|S|R}{2n}(3n^2R)^{-1} \ = \ \frac{|S|}{6n^3}. \tag{5.2.28}$$

As $L, R \to \infty$ in a manner such that R grows faster than L^2, this probability approaches 1. Now let,

$$B \ > \ 3^{6n^3/|S|}. \tag{5.2.29}$$

We obtain that

$$\log_B(3) \ < \ \frac{|S|}{6n^3} \tag{5.2.30}$$

but at least $\frac{|S|}{6n^3}$ of the dead sticks are in $[L/3, L]$ so that at least the same fraction of normalized mantissas of dead sticks are in $[1 - \log_B(3), 1]$. It follows that the distribution of mantissas of dead sticks cannot approach the uniform distribution as $R \to \infty$ for any $L > 6n$, nor can such be the case as $L \to \infty$.

6. Further Directions

6.1. **General Number of Parts.** Given Theorem 3.1, it seems likely that a similar result would hold for the discrete analogue. Indeed, we make the following conjecture, which is supported by our simulation results (see, for example, Figure 1).

Conjecture 6.1 (General number of parts)**.** *Fix some positive integer $k \geq 2$, and consider the process where we break each stick into k pieces by choosing $k-1$ cut points recursively following the uniform distribution*[2]*. Fix a modulus $n = tk$ for some $t \geq 1$ and a subset $S \subset \{0, \dots, n-1\}$ of size $(t-1)k$ representing the residue classes. Let the stopping set be*

$$\mathfrak{S} \ := \ \{1\} \cup \{m \in \mathbb{Z}_+ : m = qn + r, \ r \in S, q \in \mathbb{Z}\}. \tag{6.1.1}$$

If we start with R identical sticks of positive integer length $L \notin \mathfrak{S}$, then the collection of ending stick lengths converges to strong Benford behavior given that $R > f(L)$ as $L \to \infty$, where $f(L)$ is some function that goes to infinity as $L \to \infty$. Moreover, if the number of residue classes constituting

[2]This is left intentionally vague. One interpretation is as follows: Choose the first cut point according to the uniform distribution as usual, and then choose the next cut point on the second fragment according to the uniform distribution on that fragment, and so on. If at some point the second fragment has length 1, then the breaking stops - so when the stick is short, it is possible that it only breaks into less than k pieces.

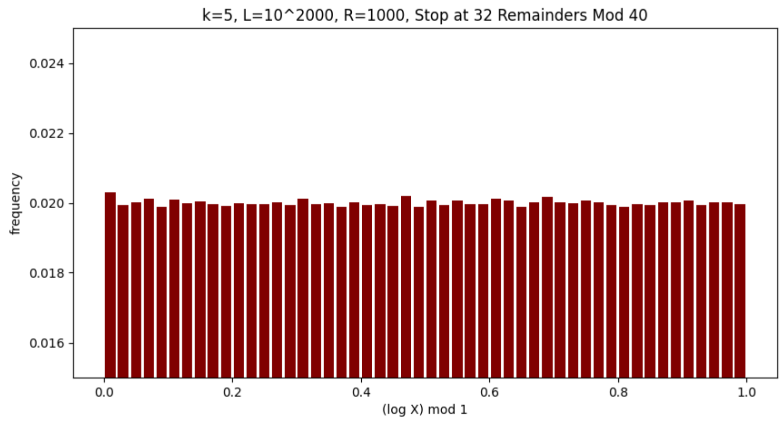

FIGURE 1. Break Into 5 Pieces and Stop at 4/5 of the Residue Classes Modulo 40

the stopping set is not equal to $(t-1)k$*, then the resulting stick lengths do not converge to strong Benford behavior.*

6.2. **General Stopping Condition.** We suspect that the only property of $\mathfrak{S}$ that is fundamentally necessary in Proposition 1.5 is its density in the set of natural numbers. As result, we have the following conjecture:

Conjecture 6.2. *Let* $\mathfrak{S}$ *be such that the limit below exists and let*

$$r = \lim_{n\to\infty} \frac{|\{[1,n]\cap\mathfrak{S}\}|}{n}. \tag{6.2.1}$$

Moreover, assume that $r > 0$*. Then, we have that set of dead stick lengths approaches Benford behavior if and only if* $r = 1/2$*.*

In other words, we believe that $r = 1/2$, the critical threshold, is the only thing required for the process to result in a distribution that approaches Benford behavior; moreover, this threshold is *sharp*, meaning that it gives a necessary and sufficient criterion for Benfordness.

6.3. **Non-Benfordness for General Base.** In the setting of Proposition 1.5, Theorem 5.7 says that in the case $|S| > n/2$, as long as the base B is large enough, the final distribution does not converge to Benford. We conjecture that this is in fact true regardless of the base.

Conjecture 6.3. *The final collection of stick lengths does not converge to strong Benford behavior* for any base B *if the size of* S *is not* $n/2$*. Specifically, if* $|S| > n/2$*, then the limiting distribution depends on the mantissa of* L *base* B*, and the density function of* $\log_B(X/L) \pmod 1$ *is skewed towards 1.*

Note that the proof of Theorem 5.7 would already strongly indicate that the above conjecture is true, although it remains an interesting open question to describe precisely what the distribution looks like.

We have also obtained strong empirical evidence for the conjecture taking $B = 10$ (which is smaller than required in Theorem 5.7). Figure 2 shows a simulation of the process with modulus $n = 40$ and stopping set

$$\mathfrak{S} = \{0, 1, 2, 3, 6, 7, 8, 9, 12, 13, 14, 15, 17, 18, 23, 28, 29, 33, 35, 36, 37\}$$

where $|\mathfrak{S}| = 21$. Note that even though the size of the stopping set is just slightly above the threshold 20, we find that the deviation from uniform distribution is apparent. It is worth noting that even when we vary the stopping set and the significand of L, the pattern persists, as long as $n/2 < |\mathfrak{S}| < n$.

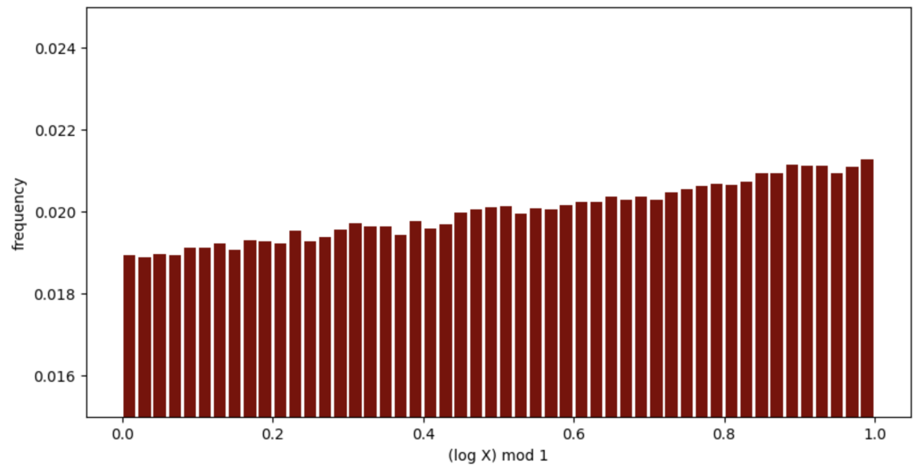

FIGURE 2. Stopping at 21 residues modulo 40 with $L = 10^{10000}$ and $R = 100000$

A heuristic for this is that when $|\mathfrak{S}| = n$, all processes end at the first level, and the resulting stick lengths follow the uniform distribution on $\{1, \ldots, L-1\}$. It is not hard to show that for a random variable X uniformly distributed, $\log(X)$ has smaller mantissa with lower probability and larger mantissa with higher probability.

APPENDIX A. PROOF OF THEOREM 3.4

Note that, for fixed $j \in I$,

$$\begin{aligned} \mathcal{M}f_j\left(1 - \frac{2\pi i\ell}{\log B}\right) &= \int_0^\infty f_j(x)x^{-\frac{2\pi i\ell}{\log B}}dx \\ &= \int_0^\infty f_j(e^{\log x})e^{\log x}e^{-\frac{2\pi i\ell}{\log B}\log x}\frac{dx}{x} \\ &= \int_{-\infty}^\infty g_j(y)e^{-\frac{2\pi i\ell}{\log B}y}dy \\ &= \widehat{g_j}\left(\frac{\ell}{\log B}\right), \end{aligned} \tag{A.0.1}$$

where $g_j(y) = f_j(e^y)e^y$. Moreover, $\|g_j\|_1 = \|f_j\|_1 = 1$, so the Riemann-Lebesgue lemma applies and says that

$$\mathcal{M}f_j\left(1 - \frac{2\pi i\ell}{\log B}\right) \to 0 \tag{A.0.2}$$

as $\ell \to \infty$. Also, for $\ell \neq 0$,

$$\left|\widehat{g_j}\left(\frac{\ell}{\log B}\right)\right| \leq \|g_j\|_1 = 1. \tag{A.0.3}$$

We do not have equality in the above since it follows from triangle inequality and the integrand does not always have the same complex argument (since g_j is continuous). Thus, if we take

$$h(\ell) = \max_j \left|\mathcal{M}f_j\left(1 - \frac{2\pi i\ell}{\log B}\right)\right|, \tag{A.0.4}$$

we have that $h(\ell) < 1$ for $\ell \neq 0$ and also $h(\ell) \to 0$ as $\ell \to \infty$. We now investigate the rate of this convergence. We begin by mimicking the proof of the Riemann-Lebesgue lemma. For any $f : \mathbb{R} \to \mathbb{C}$ continuous and compactly supported, using the substitution $x \mapsto x + \frac{\pi}{\xi}$ for $\xi \neq 0$, we have

$$\hat{f}(\xi) \;=\; \int_{\mathbb{R}} f(x) e^{-ix\xi} dx \;=\; \int_{\mathbb{R}} f\left(x + \frac{\pi}{\xi}\right) e^{-ix\xi} e^{-i\pi} dx \;=\; -\int_{\mathbb{R}} f\left(x + \frac{\pi}{\xi}\right) e^{-ix\xi} dx. \tag{A.0.5}$$

Taking the average, we get

$$|\hat{f}(\xi)| \;\leq\; \frac{1}{2} \int_{\mathbb{R}} \left| f(x) - f\left(x + \frac{\pi}{\xi}\right) \right| dx. \tag{A.0.6}$$

Apply this to $f = \widehat{g_j}$ and $\xi = \frac{\ell}{\log B}$,

$$\begin{aligned} \left| \widehat{g_j}\left(\frac{\ell}{\log B}\right) \right| \;&\leq\; \frac{1}{2} \int_{\mathbb{R}} \left| g_j(x) - g_j\left(x + \frac{\pi \log B}{\ell}\right) \right| dx \\ &\leq\; \frac{1}{2} \int_{\mathbb{R}} \left| f_j(e^x) e^x - f_j(e^{x + \frac{\pi \log B}{\ell}}) e^{x + \frac{\pi \log B}{\ell}} \right| dx \\ &\leq\; \frac{1}{2} \int_0^1 |f_j(u) - c f_j(cu)|\, du \\ &\leq\; \frac{1}{2} \sup_{[0,1]} (|f_j(u) - f_j(cu)| + |f_j(cu) - c f_j(cu)|), \end{aligned} \tag{A.0.7}$$

where $c = e^{\frac{\pi \log B}{\ell}}$ and we used the fact that f is only supported on $[0, 1]$ (this can be easily changed to any compact interval, but for the purpose of this paper all the distributions we consider satisfy this condition). From the assumption that f_j is α_j-Hölder continuous, there exists a constant $\mu \geq 0$ such that

$$|f_j(u) - f_j(cu)| \;\leq\; \mu |(1 - c)u|^{\alpha} \;\leq\; \mu |1 - c|^{\alpha} \tag{A.0.8}$$

and

$$|f_j(cu) - c f_j(cu)| \;\leq\; (1 - c)|f_j(cu)| \;\leq\; |1 - c| M \tag{A.0.9}$$

for all $u \in [0, 1]$, where $M > 0$ is an upper bound for f. Now

$$c \;=\; 1 + \frac{\pi \log B}{\ell} + o(1/\ell), \tag{A.0.10}$$

so

$$|1 - c| \;=\; \frac{\pi \log B}{\ell} + o(1/\ell). \tag{A.0.11}$$

We may assume $0 < \alpha \leq 1$, so that $|1 - c|^{\alpha}$ dominates. There exists some L large enough so that the sum $\sum_{|\ell| \geq L} \ell^{-n\alpha} \to 0$ as $n \to \infty$. By the pigeonhole principle, there exist a $j \in I$ such that $|p^{-1}(\{j\})| = \infty$. Then we have

$$\begin{aligned} \sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{\infty} \prod_{m=1}^{n} \mathcal{M} f_{\mathcal{D}_{p(m)}} \left(1 - \frac{2\pi i \ell}{\log B}\right) \;&\leq\; \sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{\infty} \prod_{\substack{m=1 \\ p(m)=j}}^{n} \mathcal{M} f_j \left(1 - \frac{2\pi i \ell}{\log B}\right) \\ &\leq\; \sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{\infty} \mathcal{M} f_j \left(1 - \frac{2\pi i \ell}{\log B}\right)^n \;\to\; 0 \end{aligned} \tag{A.0.12}$$

as $n \to \infty$. This proves (3.2.2). To see that the convergence is uniform over all p, note that by the pigeonhole principle, for any choice of p and any positive integer N there exists a $j \in I$ such that

$|p^{-1}(\{j\}) \cap \{1, \dots, N\}| \ \geq \ N/|I|$. Therefore for any $\epsilon > 0$, it suffices to take the maximum among all the N's needed for each $j \in I$ so that

$$\sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{\infty} \mathcal{M} f_j \left(1 - \frac{2\pi i \ell}{\log B}\right)^{N/|I|} \ < \ \epsilon. \tag{A.0.13}$$

References

[AS] M. Abromovich, I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, tenth printing, National Bureau of Standards, Applied Mathematics Series **55**, (1972).

[Adh] A. K. Adhikari, *Some results on the distribution of the most significant digit*, Sankhyā: The Indian Journal of Statistics, Series B **31** (1969), 413–420.

[AF] R. L. Adler and L. Flatto, *Uniform distribution of Kakutani's interval splitting procedure*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **38** (1977), no. 4, 253–259.

[B+] T. Becker, D. Burt, David T. C. Corcoran, A. Greaves-Tunnell, J. R. Iafrate, J. Jing, S. J. Miller, J. D. Porfilio, R. Ronan, J. Samranvedhya, F. W. Strauch, and B. Talbut, *Benford's law and continuous dependent random variables*, Annals of Physics **338** (2018), 350-381.

[Ben] F. Benford, *The Law of Anomalous Numbers*, Proceedings of the American Philosophical Society **78** (1938), 551-572.

[BH1] A. Berger and T. P. Hill, *A Basic Theory of Benford's Law*, Probab. Surv. **8** (2011), 1–126.

[BH2] A. Berger, T. P. Hill *An Introduction to Benford's Law*, Princeton University Press, (2015).

[BH3] A. Berger, and T. P. Hill, *Benford Online Bibliography*, http://www.benfordonline.net.

[Ber] A. C. Berry, *The Accuracy of the Gaussian Approximation to the Sum of Independent Variates*, Transactions of the American Mathematical Society **49** (1941), 122-136.

[Bert] J. Bertoin, *Random fragmentation and coagulation processes*, Cambridge Studies in Advanced Mathematics **102**, Cambridge University Press, Cambridge, (2006).

[BDMMM] L. Betti, I. Durmic, Z. Mcdonald, J. B. Miller and S. J. Miller, *Benfordness of measurements resulting from box fragmentation*, (2023), preprint. https://arxiv.org/abs/2304.08335

[Bha] R. N. Bhattacharya, *Speed of convergence of the n-fold convolution of a probability measure on a compact group*, Z. Wahrscheinlichkeitstheorie verw. Geb. **25** (1972), 1–10.

[BMP] B. Buck, A. Merchant and S. Perez, *An Illustration of Benford's First Digit Law Using Alpha Decay Half-Lives*, European Journal of Physics **14** (1993), 59–63.

[Car] I. Carbone, *Discrepancy of LS-sequences of partitions and points*, Ann. Mat. Pura Appl. **191** (2012), 819–844.

[CV] I. Carbone and A. Volčič, *Kakutani's splitting procedure in higher dimension*, Rend. Istit. Mat. Univ. Trieste **39** (2007), 119–126.

[CLM] V. Cuff, A. Lewis and S. J. Miller, *The Weibull distribution and Benford's law*, Involve, a Journal of Mathematics 8-5 (2015), 859–874. DOI 10.2140/involve.2015.8.859.

[Dia] P. Diaconis, *The distribution of leading digits and uniform distribution mod 1*, Ann. Probab. **5** (1979), 72–81.

[DM] I. Durmic and S. J. Miller, *Benford Behaviour of a Higher Dimensional Fragmentation Process*, preprint, (2023), https://librarysearch.williams.edu/permalink/01WIL_INST/1faevhg/alma991013795585602786.

[Dun] R. L. Duncan . *An Application of Uniform Distribution to the Fibonacci Numbers*, The Fibonacci Quarterly. 5: (1967), 137-–140.

[Eli] I. Eliazar, *Harmonic statistics*, Annals of Physics **380** (2017), 168–187.

[Ess] C. Esseen, *On the Liapounoff limit of error in the theory of probability.*, Arkiv För Matematik, Astronomi Och Fysik, (1942).

[Hil1] T. P. Hill, *A Statistical Derivation of the Significant-Digit Law*, Statistical Science **10** (1995), no. 4, 354-363.

[Hil2] T. P. Hill, *The first-digit phenomenon*, American Scientists **86** (1996), 358–363.

[IMS] J. Iagrate, S. J. Miller and F. W. Strauch, *Equipartitions and a distribution for numbers: A statistical model for Benford's law*, Physical Review E **91** (2015), no. 6, 062138 (6 pages).

[IV] M. Infusino and A. Volčič, *Uniform distribution on fractals*, Unif. Distrib. Theory **4** (2009), no. 2, 47–58.

[JKKKM] D. Jang, J. U. Kang, A. Kruckman, J. Kudo and S. J. Miller, *Chains of distributions, hierarchical Bayesian models and Benford's Law*, Journal of Algebra, Number Theory: Advances and Applications, volume 1, number 1 (2009), 37–60.

[Kak] S. Kakutani, *A problem of equidistribution on the unit interval* $[0,1]$, Measure theory (Proc. Conf., Oberwolfach, 1975), Lecture Notes in Math., Vol. 541, Springer, Berlin, (1976), pp. 369–375.

[Khi] A. Y. Khinchin, *Continued Fractions*, Third Edition, The University of Chicago Press, Chicago (1964).

[Kol] A. N. Kolmogoroff, *Über das logarithmisch normale Verteilungsgesetz der Dimensionen der Teilchen bei Zerstückelung*, C. R. (Doklady) Acad. Sci. URSS (N. S.) **31** (1941), 99–101.

[KM] A. Kontorovich and S. J. Miller, *Benford's Law, values of L-functions and the* $3x+1$ *problem*, Acta Arithmetica **120** (2005), no. 3, 269–297.

[Kos] A. E. Kossovsky, *Towards a Better Understanding of the Leading Digits Phenomena*, preprint, (2006), http://arxiv.org/abs/math/0612627.

[LSE] L. M. Leemis, B. W. Schmeiser and D. L. Evans, *Survival Distributions Satisfying Benford's Law*, The American Statistician **54** (2000), no. 3.

[Lem] D. S. Lemons, *On the Numbers of Things and the Distribution of First Digits*, American Journal of Physics (1986), 816–817.

[Lév1] P. Lévy, *Théorie de l'Addition des Variables Aléatoires*, Gauthier-Villars, Paris, (1937).

[Lév2] P. Lévy, *L'addition des variables aléatoires définies sur une circonférence*, Bull. de la S. M. F. **67** (1939), 1–41.

[Loo] J. C. Lootgieter, *Sur la répartition des suites de Kakutani. I*, Ann. Inst. H. Poincare Sect. B (N.S.) **13** (1977), no. 4, 385–410.

[Mil1] S. J. Miller, *Benford's Law: Theory and Applications*, Princeton University Press, Princeton, NJ, (2015).

[Mil2] S. J. Miller, *The Probability Lifesaver*, Princeton University Press, Princeton, NJ, (2017). https://doi.org/10.1515/9781400885381.

[MN1] S. J. Miller and M. Nigrini, *The Modulo* 1 *Central Limit Theorem and Benford's Law for Products*, International Journal of Algebra **2** (2008), no. 3, 119–130.

[MN2] S. J. Miller and M. J. Nigrini, *Order Statistics and Benford's Law*, International Journal of Mathematics and Mathematical Sciences, (2008), 1-13.

[MT-B] S. J. Miller and R. Takloo-Bighash, *An Invitation to Modern Number Theory*, Princeton University Press, Princeton, NJ, (2006).

[MSPZ] M. Moret, V. de Senna, M. Pereira and G. Zebende, *Newcomb-Benford Law in Astrophysical Sources*, International Journal of Modern Physics C **17** (2006), no. 11, 1597–1604.

[New] S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881), 39-40.

[NWR] D. Ni, L. Wei and Z. Ren, *Benford's Law and* β*-Decay Half-Lives*, Communications in Theoretical Physics **51** (2009), 713–716.

[Nig] M. J. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, Auditing: A Journal of Practice and Theory **28** (2009), no. 2, 305–324.

[NM1] M. Nigrini and S. J. Miller, *Benford's Law applied to hydrology data – results and relevance to other geophysical data*, Mathematical Geology **39** (2007), no. 5, 469–490.

[NM2] M. J. Nigrini and S. J. Miller, *Data diagnostics using second order tests of Benford's Law*, John Wiley & Sons, Inc., Hoboken, New Jersey, (2012).

[Oll] J. Olli, *Division point measures resulting from triangle subdivisions*, Geom. Dedicata **158** (2012), 69–86.

[PHA] F. Perez-Gonzalez, G. L. Heileman, and C.T. Abdallah, *Benford's Law in Image Processing*, Proceedings of the IEEE International Conference on Image Processing (ICIP), San Antonio, TX, Volume 1, (2007), pp. 405–408.

[PTTV] L. Pietronero, E. Tosatti, V. Tosatti and A. Vespignani, *Explaining the Uneven Distribution of Numbers in Nature: the Laws of Benford and Zipf*, Physica A: Statistical Mechanics and its Applications, (2001).

[PvZ] R. Pyke and W. R. van Zwet, *Weak convergence results for the Kakutani interval splitting procedure*, Ann. Probab. **32** (2004), no. 1A, 380–423.

[Rai] R. A. Raimi, *The First Digit Problem*, The American Mathematical Monthly, **83:7** (1976), no. 7, 521-538.

[Rob] H. Robbins, *On the equidistribution of sums of independent random variables*, Proc. Amer. Math. Soc. **4** (1953), 786–799.

[RSZ] Z. Rudnick, P. Sarnak, and A. Zaharescu, *The Distribution of Spacings Between the Fractional Parts of* $n^2\alpha$, Invent. Math. **145** (2001), no. 1, 37–57.

[Sak] H. Sakamoto, *On the distributions of the product and the quotient of the independent and uniformly distributed random variables*, Tôhoku Math. J. **49** (1943), 243–260.

[Sar] Sarkar, P. B. (1973). *An Observation on the Significant Digits of Binomial Coefficients and Factorials.* Sankhya B. 35: 363–364.

[Sch1] P. Schatte, *On sums modulo* 2π *of independent random variables*, Math. Nachr. **110** (1983), 243–261.

[Sch2] P. Schatte, *On the asymptotic uniform distribution of sums reduced mod 1*, Math. Nachr. **115** (1984), 275–281.

[Sch3] P. Schatte, *On the asymptotic logarithmic distribution of the floating-point mantissas of sums*, Math. Nachr. **127** (1986),7–20.

[SM1] L. Shao and B. Ma, *First Digit Distribution of Hadron Full Width*, Modern Physics Letters A **24** (2009), no. 40, 3275–3282.

[SM2] L. Shao and B. Ma, *Empirical Mantissa Distributions of Pulsars*, Astroparticle Physics **33** (2010), 255–262.

[Slud] E. Slud, *Entropy and maximal spacings for random partitions*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **41** (1977/78), no. 4, 341–352.

[ST] M. D. Springer and W. E. Thompson, *The distribution of products of independent random variables*, SIAM J. Appl. Math. **14** (1966), 511–526.

[Sta] E. W. Stacy, *A Generalization of the Gamma Distribution*, The Annals of Mathematical Statistics **33** (1962), no. 3, 1187-1192.

[Töd] K.-H. Tödter, *Benford's Law as an Indicator of Fraud in Economics*, German Economic Review **10** (2009), 339–351.

[vZ] W. R. van Zwet, *A proof of Kakutania's conjecture on random subdivision of longest intervals*, Ann. Probability **6** (1978), no. 1, 133–137.

[V-BFJ] S. B. Villas-Boas, Q. Fu and G. Judge, *Benford's law and the FSD distribution of economic behavioral micro data*, Physica A: Statistical Mechanics and its Applications (2017), 711–719.

[Wa] L. C. Washington, *Benford's Law for Fibonacci and Lucas Numbers*, The Fibonacci Quarterly, 19 (2): (1981) 175–177.

Email address: fangxinyu@u.nus.edu

Department of Mathematics, National University of Singapore, Singapore 119077

Email address: sjm1@williams.edu

Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267

Email address: mrsun@mit.edu

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

Email address: amanda.verga@trincoll.edu

Department of Mathematics, Trinity College, Hartford, CT 06106