BENFORDNESS OF MEASUREMENTS RESULTING FROM BOX FRAGMENTATION

LIVIA BETTI, IRFAN DURMIĆ, ZOE MCDONALD, JACK B. MILLER, AND STEVEN J. MILLER

ABSTRACT. We make progress on a conjecture made by [DM], which states that the d-dimensional frames of m-dimensional boxes resulting from a fragmentation process satisfy Benford's law for all $1 \leq d \leq m$. We provide a sufficient condition for Benford's law to be satisfied, namely that the maximum product of d sides is itself a Benford random variable. Motivated to produce an example of such a fragmentation process, we show that processes constructed from log-uniform proportion cuts satisfy the maximum criterion for $d = 1$.

CONTENTS

1. Introduction 1
1.1. Historical Background 1
1.2. Previous Work on Fragmentation 2
1.3. Results 3
2. Reduction to the Maximum-Volume Face 5
3. A Family of Distributions whose Maximum Side-Lengths are Benford 8
4. Future Work 10
4.1. Restating the Mellin Condition in Terms of Characteristic Functions 11
References 12

1. INTRODUCTION

1.1. Historical Background.

At the dawn of the 20^{th} century, the astronomer and mathematician Simon Newcomb observed that the logarithmic books at his workplace showed a lot of wear and tear at the early pages, but the more he progressed through the book, the less usage could be observed. Newcomb deduced that his colleagues had a "bias" towards numbers starting with the digit 1. In particular, the digit 1 shows up as the first digit roughly 30% of the time, the digit 2 about 17% of the time, and so on. While he did come up with a mathematical model for this interesting relationship, his work stayed mostly unnoticed.

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

Definition 1.1. [Ben, Page 554] We say that data exhibits *(weak) Benford behavior* if the frequency F_d of leading digit d satisfies

$$F_d \ = \ \log_{10} \frac{d \ + \ 1}{d}. \tag{1.1}$$

Nowadays, Benford's Law is used in detecting many different forms of fraud, and its prevalence in the world fascinates not only mathematicians, but many other scientists as well (to learn more about Benford's Law and its many applications, we recommend [BeHi, Nig, Mil1] to name a few).

In 1986, Lemons [Lemons] proposed using Benford's law to analyze the partitioning of a conserved quantity. Since then, driven by the potential application to nuclear fragmentation, mathematicians and

Date: April 14, 2023.
2010 *Mathematics Subject Classification.* 60A10, 11K06 (primary), (secondary) 60E10.
Key words and phrases. Benford's Law, Digit Bias, Fragmentation Process.
This work was supported by NSF grant DMS1947438, and Williams College.

physicists have taken an interest in the Benfordness of various fragmentation processes. Among these processes of interest is *stick fragmentation*. In the unrestricted stick fragmentation model, one begins with a stick of length L. Draw p_1 from a probability distribution on $(0,1)$. This fragments the stick into two sub-sticks of lengths p_1L and $(1-p_1)L$. For each sub-stick, draw another independent probability (p_2 and p_3, respectively) from the same distribution. Repeat this process N times. Of particular interest is whether this fragmentation process follows Benford's law.

1.2. **Previous Work on Fragmentation.** An important definition when studying a more precise statistical version of Benford's law is the notion of the significand of a real number, i.e., its leading digits in scientific notation.

Definition 1.2 (Significand)**.** Given a positive real number x, we say that its *significand base* $B>1$, denoted $S_B(x)$, is the unique real number $S_B(x)\in[1,B)$ such that $k=\log_B(x)-\log_B(S_B(x))$ is an integer. One can then write $x=S_B(x)\cdot B^k$.

As is common practice with these techniques involving proofs of Benford's law, we define a stricter version of Benford behavior.

Definition 1.3 (Strong Benford's Law)**.** We say that a sequence of random variables $X^{(n)}$ *converges to strong Benford behavior* in the base B if

$$\mathbb{P}(S_B(X^{(n)})\leq D)\to\log_B(D), \tag{1.2}$$

for all $D\in[1,B]$. Notice by compactness that this implies uniform convergence of (1.2).

We may now state the previous results on box fragmentation. Becker, et al. [B–] proved a theorem regarding unrestricted stick fragmentation (compare with their Theorem 1.5) which was later generalized by [DM] in the form of the following theorem.

Theorem 1.4 (Benfordness of the m-Volumes of a Branching-Fragmentation Process)**.** *Fix a continuous probability density $f:(0,1)\to\mathbb{R}$ such that its Mellin transform*[1] $\mathcal{M}[f_u]$ *satisfies*

$$\lim_{n\to\infty}\sum_{\substack{\ell=-\infty\\ \ell\neq0}}^{\infty}\left|\prod_{u=1}^{nm}\mathcal{M}[f_u]\left(1-\frac{2\pi i\ell}{\log 10}\right)\right| \;=\; 0, \tag{1.3}$$

where each $f_u(t)$ is either $f(t)$ or $f(1-t)$ (the density of $1-P$ if P has density f). Given an m-dimensional box of m-dimensional volume V, we independently choose density cuts $p_1,p_2,\dots,p_{nm-1},p_{nm}$ from the unit interval stemming from the probability density function f and the associated random variable P. After N iterations we have

$$\begin{aligned} V_1 &= Vp_1p_2p_4\cdots p_{2^{nm-2}}p_{2^{nm-1}}, \quad V_2 \;=\; Vp_1p_2p_4\cdots p_{2^{nm-2}}(1-p_{2^{nm-1}}), \quad \dots, \\ V_{(2^m)^n} &= V(1-p_1)(1-p_3)(1-p_7)\cdots(1-p_{2^{nm-1}-1})(1-p_{2^{nm}-1}). \end{aligned} \tag{1.4}$$

Let φ_s denote the significand indicator function

$$\varphi_s(x) \;:=\; \begin{cases}1 & s_{10}(x)\;\leq\;s\\ 0 & \textit{otherwise}\end{cases}. \tag{1.5}$$

Let $\rho_n(s)$ denote the fraction of volumes $V_1,\dots,V_{(2^m)^n}$ with significand at most s, i.e.,

$$\rho_m^{(n)}(s) \;:=\; \frac{\sum_{i=1}^{(2^m)^n}\varphi_s(V_i)}{(2^m)^n}. \tag{1.6}$$

We have that the following two conditions hold.

(1) $\lim_{n\to\infty}\mathbb{E}[\rho_m^{(n)}(s)]=\log_{10}(s)$,
(2) $\lim_{n\to\infty}\mathrm{Var}\left(\rho_m^{(n)}(s)\right)=0$.

[1]The Mellin transform is related to the Fourier transform by a logarithmic change of variables, which we will discuss further in Section 4. Often, the Mellin and Fourier transforms are useful tool for stating regularity conditions.

Thus, in the limit, the m-dimensional volumes resulting from such a branching-fragmentation process exhibit Benford behavior with high probability.

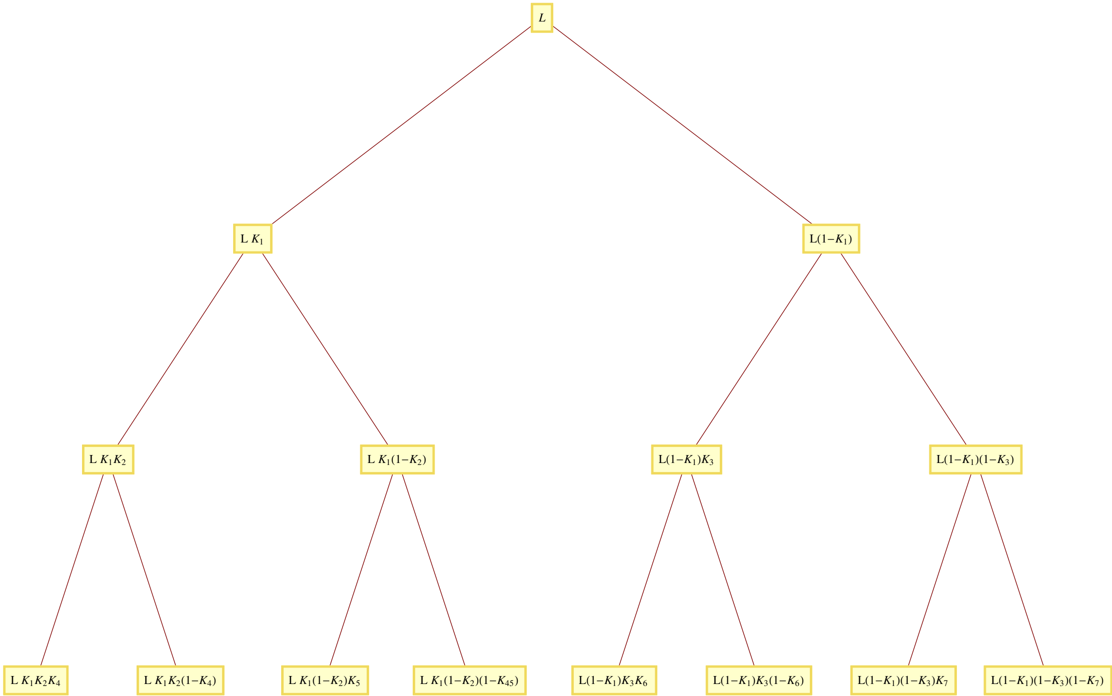

FIGURE 1. The side lengths of a one-dimensional branching-fragmentation process for $n = 3$.

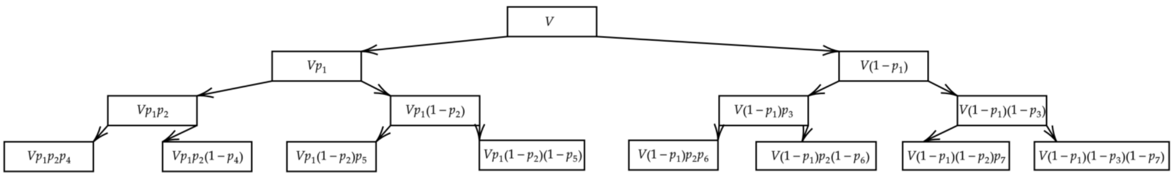

FIGURE 2. The volumes of a three-dimensional branching-fragmentation process for $n = 1$.

Remark 1.5. The exact fragmentation process used in Theorem 1.4 features $2^{m \cdot n}$ boxes at time step n, all of which are concurrent sub-boxes of the original box. We say that this is a *branching-fragmentation process*, as there are exponentially many boxes which naturally are the leaves of a height n binary tree of all the boxes at all the time steps up to n. Theorem 1.4 proves strong concentration, i.e., that the variance goes to zero; morally this is because early decisions in the tree about where to cut have little effect on future boxes that are far apart leaves on the tree.

The proof of Theorem 1.4 suggests that one might observe Benford behavior in the perimeter, area, and other generalized volumes of lower-dimensional faces of boxes resulting from fragmentation.

1.3. **Results.** We prove results about linear-fragmentation processes, which we define as follows.

Definition 1.6 (Box)**.** We say a set $\mathfrak{B} \subset \mathbb{R}^m$ is an *m-dimensional box* if it is a set of the form $[a_1, b_1] \times \cdots \times [a_m, b_m] \subset \mathbb{R}^m$, where $a_i < b_i$ are finite numbers.

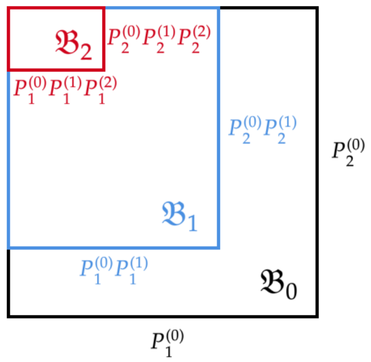

FIGURE 3. A linear fragmentation process for $n = 2$ on a two-dimensional box.

Definition 1.7 (Linear-Fragmentation Process)**.** A *linear-fragmentation process* is a sequence of random variables $\mathfrak{B}_0, \mathfrak{B}_1, \mathfrak{B}_2, \dots$ such that the following hold.

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

The statistics we are interested in studying are the volumes of the frame random variables in a linear-fragmentation process.

Definition 1.8 (d-Volume)**.** Given an m-dimensional box $\mathfrak{B}$ and a positive integer $d \leq m$, we say the *d-volume* of $\mathfrak{B} = \prod_i [a_i, b_i]$ is the sum of the d-dimensional volumes of the d-dimensional faces of $\mathfrak{B}$. More precisely, we define

$$\mathrm{Vol}_d(\mathfrak{B}) \;:=\; 2^{m-d} \sum_{|I| \,=\, d} \; \prod_{i \,\in\, I} (b_i - a_i), \tag{1.7}$$

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

In Section 2, we prove the following theorem.

Theorem 1.9 (Maximum Criterion)**.** *Let $\mathfrak{B} = \mathfrak{B}_0$ be a fixed m-dimensional box. Let $\mathfrak{B}_0 \supset \mathfrak{B}_1 \supset \cdots$ be a a linear-fragmentation process whose proportion cuts P_i have probability density functions $f_i : (0,1) \to (0, \infty)$. Let*

$$V_d^{(n)} \;:=\; \mathrm{Vol}_d(\mathfrak{B}_n) \tag{1.8}$$

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

Remark 1.10. Condition (5) for a linear-fragmentation process can be dropped with more work. The idea is that, by the law of large numbers, one expects the significand of our volumes to be largely influenced by the sides whose proportion cuts have the largest mean; therefore we have a reduction to the case of equal means. Having the same mean and different variances, there is little quantitative difference in our analysis, but for sake of notation it is much clearer to assume that all variances are the same.

When $d = m$ there is only one choice of product, and therefore the maximum criterion is automatically satisfied by a large class of continuous proportion distributions, namely all such distributions P_i

for which repeated independent multiplications by $X = P_1 \cdots P_m$ converges to strong Benford behavior. Note that this gives us a result analogous to those of [B–] and [DM] for the linear-fragmentation process. Therefore, Theorem 1.9 implies the following corollary.

Corollary 1.11 (Benfordness of the m-Volumes of a Linear-Fragmentation Process)**.** *Let $\mathfrak{B} = \mathfrak{B}_0$ be a fixed m-dimensional box. Let $\mathfrak{B}_0 \supset \mathfrak{B}_1 \supset \cdots$ be a a linear-fragmentation process. Then the sequences of box volumes $\mathrm{Vol}_m(\mathfrak{B}_n)$ converges to strong Benford behavior.*

Indeed, one may appeal to the Central Limit Theorem in order to immediately see that the independent products of such P_i satisfy the strong version of Benford's law.

In Section 3, we produce an example family of distributions which satisfy the maximum criterion for $d = 1$, namely those for which $\log_B P_i$ are uniformly distributed. We prove the following theorem.

Theorem 1.12 (Example of the Maximum Criterion being Satisfied)**.** *Let $P_i^{(j)}$ be IID log-uniform distributions. In the case of $d = 1$, i.e., perimeter, the maximum side-lengths*

$$\mathfrak{m}_1^{(n)} := \max_{1 \leq i \leq m} P_i^{(1)} \cdots P_i^{(n)} \tag{1.9}$$

converge to Strong Benford behavior as $n \to \infty$.

In view of Theorem 1.9, this gives an example of Benford behavior for lower dimensional volumes of a box fragmentation process.

Corollary 1.13. *Let $\mathfrak{B} = \mathfrak{B}_0$ be a fixed m-dimensional box. Let $\mathfrak{B}_0 \supset \mathfrak{B}_1 \supset \ldots$ be a linear-fragmentation process whose proportion cuts P_i are identically log-uniform. Then the sequence of frame perimeters $\mathrm{Vol}_1(\mathfrak{B}_n)$ converges to Strong Benford behavior as $n \to \infty$.*

2. Reduction to the Maximum-Volume Face

In this section, the following notation is fixed. We work under the assumptions of Definition 1.7.

- B: a fixed base in $[1, \infty)$.
- m: the dimension of the boxes $\mathfrak{B} = \mathfrak{B}_0 \supset \mathfrak{B}_1 \supset \cdots$.
- d: the dimension of the frames we are considering.
- $P_1^{(0)}, \ldots, P_m^{(0)}$: the initial side lengths (i.e., $b_i - a_i$) of $\mathfrak{B}_0$.
- $P_1^{(n)}, \ldots, P_m^{(n)}$, $n \geq 1$: the m proportions drawn at the nth iteration.
- $S_i^{(n)} := \prod_{t=0}^{n} P_i^{(t)}$: the side lengths of $\mathfrak{B}_n$.
- I, J: dummy indexing sets ranging over subsets of $\{1, \ldots, m\}$ with cardinality d.
- $P_I^{(t)} := \prod_{i \in I} P_i^{(t)}$.
- $v_d^{(n)} := 2^{d-m} V_d^{(n)} = \sum_I \prod_{i \in I} S_i^{(n)} = \sum_I \mathfrak{p}_I^{(n)}$: the d-volume without the constant 2^{m-d}.
- $\mathfrak{p}_I^{(n)} := \prod_{i \in I} S_i^{(n)} = \prod_{t=0}^{n} P_I^{(t)}$: the product of the sides in I.
- $\mathfrak{m}_d^{(n)} := \max_I \mathfrak{p}_I^{(n)}$: the maximum product of d sides.

It suffices to show that the random variables $v_d^{(n)}$ converge to strong Benford behavior, because $v_d^{(n)}$ and $V_d^{(n)}$ only differ by a fixed multiplicative constant of 2^{m-d}. Indeed, if X is Benford, so is cX for any fixed $c > 0$. Moreover, what we like is to control such a sum of products $\sum_I \mathfrak{p}_I^{(n)}$ over $|I| = d$ by using the observation that the maximum product $\mathfrak{m}_d^{(n)}$ should typically be many orders of magnitude larger than the other products. We quantify this statement in the form of Lemma 2.1, which is the tool that allows us to control the strong Benford behavior of our sum of random variables, allowing us to if one ascertains that the strong Benfordness of the maximum is suitable. In rare instances, such as $\mathfrak{m}_d^{(n)} = (B - \varepsilon) \cdot B^k$ where $\varepsilon > 0$ is small, the Benfordness of $\mathfrak{m}_d^{(n)} = \mathfrak{p}_{\mathcal{I}}^{(n)}$ for some $|\mathcal{I}| = d$ does not translate well to the Benfordness of $\mathfrak{m}_d^{(n)} + \sum_{J \neq \mathcal{I}} \mathfrak{p}_{\mathcal{I}}^{(n)}$, since there is an overflow of the digits base B which tampers with the distribution of the significand greatly. We handle these events, showing they almost always never occur (i.e., with probability tending towards 0) in a standard way (cf. §9.3.2 of [MT-B]).

We first require a lemma.

Lemma 2.1 (Wafer Lemma)**.** *Let $0 < \delta_n < 1$ be a decreasing sequence. Then the probability that $v_d^{(n)}$ is at most $(1+\delta_n)$ times $\mathfrak{m}_d^{(n)}$ is*

$$\mathbb{P}\left(\mathfrak{m}_d^{(n)} \le v_d^{(n)} \le (1+\delta_n)\mathfrak{m}_d^{(n)}\right) \;=\; 1 - O\left(\frac{-\log \delta_n}{\sqrt{n}}\right), \tag{2.1}$$

where the implied constant depends on the distribution of $Y_j^{(t)}$ and m. We say that such an event at time n is a δ_n-Wafer.

Proof. Our goal is to show that as $n \to \infty$, it is with probability tending to 1 that there exists a product $\mathfrak{p}_I^{(n)}$ which is significantly greater in magnitude than the other products $\mathfrak{p}_J^{(n)}$ for $J \neq I$. That is, it is with probability tending to 1 that there exists an indexing set $\mathcal{I}$ which has the largest product and is large in the sense that $\log \mathfrak{p}_{\mathcal{I}}^{(n)} - \log \mathfrak{p}_J^{(n)} \ge \alpha_n$ for all $J \neq \mathcal{I}$, where α_n slowly tends towards infinity. We first write for every I

$$\log \mathfrak{p}_I^{(n)} \;=\; \sum_{i \in I} \log S_i^{(n)} \;=\; \sum_{i \in I} \sum_{t \le n} \log P_i^{(t)}. \tag{2.2}$$

Notice that, due to the inequality below, we may reduce to the $d = 1$ dimensional case, since showing that it tends to 1 will squeeze all other probabilities. Indeed,

$$\mathbb{P}\left(\bigcup_{|\mathcal{I}|=d} \bigcap_{J \neq \mathcal{I}} \{\log \mathfrak{p}_{\mathcal{I}}^{(n)} - \log \mathfrak{p}_J^{(n)} \ge \alpha_n\}\right) \;\ge\; \mathbb{P}\left(\bigcup_{i=1}^{m} \bigcap_{j \neq i} \{\log S_i^{(n)} - \log S_j^{(n)} \ge \alpha_n\}\right). \tag{2.3}$$

This can be seen by using the middle expression for $\log \mathfrak{p}_I^{(n)}$ in (2.2). Notice that for $\alpha_n > 0$, the union of events over i is disjoint, therefore we calculate

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{m} \bigcap_{j \neq i} \{\log S_i^{(n)} - \log S_j^{(n)} \ge \alpha_n\}\right) \;&=\; \sum_{i=1}^{m} \mathbb{P}\left(\bigcap_{j \neq i} \{\log S_i^{(n)} - \log S_j^{(n)} \ge \alpha_n\}\right) \\ &=\; \sum_{i=1}^{m} \int_{-\infty}^{\infty} f_i^{(n)}(s) \prod_{j \neq i} F_j^{(n)}(s - \alpha_n)\, ds, \end{aligned} \tag{2.4}$$

where we have used the integral version of the law of total probability with respect to the values that the maximum value $s = \log S_i^{(n)}$ may take, as well as independence of the S_j's. The functions $f_j^{(n)}, F_j^{(n)}$ denote the PDF and CDF of $\log S_j^{(n)}$ respectively. One version of the Berry–Esseen theorem (cf. [Berry] and [Esseen]) gives us, in consideration of (2.2) for each $\log S_j^{(n)}$,

$$F_j^{(n)}(x) \;=\; \Phi\left(\frac{x - n \cdot \mu_P}{\sqrt{n} \cdot \sigma_P}\right) + O_P\left(\frac{1}{\sqrt{n}}\right) \tag{2.5}$$

where Φ is the PDF of the standard normal $\mathcal{N}(0,1)$, and the implied constant for $O_P(1/\sqrt{n})$ is uniform over $x \in \mathbb{R}$. By our convention in Definition 1.7, $\mu_P = \mathbb{E}[\log_B P_j^{(1)}]$ and $\sigma_P = \mathrm{Var}[\log_B P_j^{(1)}]$ are uniform over $1 \le j \le m$. Applying (2.5) to (2.4) yields, for $1 \ll \alpha_n \ll \sqrt{n}$,

$$\begin{aligned}&\sum_{i=1}^{m} \int_{-\infty}^{\infty} f_i^{(n)}(s) \left(\Phi\left(\frac{s - n \cdot \mu_P}{\sqrt{n} \cdot \sigma_P}\right) + O_P\left(\frac{\alpha_n}{\sqrt{n}}\right)\right)^{m-1} ds \\ &\qquad =\; \left(\sum_{i=1}^{m} \int_{-\infty}^{\infty} f_i^{(n)}(s) \Phi\left(\frac{s - n \cdot \mu_P}{\sqrt{n} \cdot \sigma_P}\right)^{m-1} ds\right) \;+\; O_{P,m}\left(\frac{\alpha_n}{\sqrt{n}}\right). \end{aligned} \tag{2.6}$$

Integrating by parts, applying (2.5) to $F_i^{(n)}$ and absorbing error, we obtain

$$O_{P,m}\left(\frac{\alpha_n}{\sqrt{n}}\right) + \sum_{j=1}^{m} \left(1 - \int_{-\infty}^{\infty} \frac{m-1}{\sqrt{n} \cdot \sigma_P} \cdot \Phi\left(\frac{s - n \cdot \mu_P}{\sqrt{n} \cdot \sigma_P}\right)^{m-1} \Phi'\left(\frac{s - n \cdot \mu_P}{\sqrt{n} \cdot \sigma_P}\right) ds\right). \tag{2.7}$$

One may recognize that the above integrand has primitive $(1-\frac{1}{m})\Phi(\frac{s-n\cdot\mu_P}{\sqrt{n}\cdot\sigma_P})^m$, and so each integral contributes $1-\frac{1}{m}$, leaving us with

$$1 - O_{P,m}\left(\frac{\alpha_n}{\sqrt{n}}\right). \tag{2.8}$$

Taking $\alpha_n = -\log\big(\delta_n/\binom{m}{d}\big)$, we have by considering subevents

$$\begin{aligned}\mathbb{P}\left(\mathfrak{m}_d^{(n)} \le v_d^{(n)} \le (1+\delta_n)\mathfrak{m}_d^{(n)}\right) \;&\geq\; \mathbb{P}\left(\bigcup_{|\mathcal{I}|=d}\bigcap_{J\neq\mathcal{I}}\{\log\mathfrak{p}_{\mathcal{I}}^{(n)} - \log\mathfrak{p}_J^{(n)} \ge -\log\big(\delta_n/\tbinom{m}{d}\big)\}\right)\\ &\geq\; \mathbb{P}\left(\bigcup_{i=1}^{m}\bigcap_{j\neq i}\{\log S_i^{(n)} - \log S_j^{(n)} \ge -\log\big(\delta_n/\tbinom{m}{d}\big)\}\right)\\ &=\; 1 - O_{P,m}\left(\frac{-\log\delta_n}{\sqrt{n}}\right). \end{aligned}\tag{2.9}$$

This finishes our proof. □

We claim that Lemma 2.1 reduces the question of strong Benford behavior of $v_d^{(n)}$ to $\mathfrak{m}_d^{(n)}$. That is, the Wafer lemma implies

Lemma 2.2 (Reduction to Max)**.** *Assume $\mathfrak{m}_d^{(n)}$ converges to strong Benford behavior. Then $v_d^{(n)}$ does as well.*

Proof. Let E_n be the event that $\mathfrak{m}_d^{(n)}$ and $v_d^{(n)}$ are a δ_n-Wafer and $(1+\delta_n)S_B(\mathfrak{m}_d^{(n)}) < B$. We condition on this event to prevent an overflow of the order of magnitude. Then

$$S_B(\mathfrak{m}_d^{(n)}) \;\le\; S_B(v_d^{(n)}) \;\le\; (1+\delta_n)S_B(\mathfrak{m}_d^{(n)}). \tag{2.10}$$

Moreover, the conditional probabilities are

$$\mathbb{P}(S_B(\mathfrak{m}_d^{(n)}) \le D/(1+\delta_n) \mid E_n) \;\le\; \mathbb{P}(S_B(v_d^{(n)}) \le D \mid E_n) \;\le\; \mathbb{P}(S_B(\mathfrak{m}_d^{(n)}) \le D \mid E_n). \tag{2.11}$$

Making basic estimates such as inclusion-exclusion, we estimate the unconditional probability as

$$\mathbb{P}\Big(S_B(\mathfrak{m}_d^{(n)}) \le D/(1+\delta_n)\Big) + \mathbb{P}(E_n) - 1 \;\le\; \mathbb{P}(S_B(v_d^{(n)}) \le D) \;\le\; \mathbb{P}\Big(S_B(\mathfrak{m}_d^{(n)}) \le D\Big) + 1 - \mathbb{P}(E_n). \tag{2.12}$$

We show that $v_d^{(n)}$ converges to strong Benford behavior by taking $\delta_n \to 0$ at a slow enough rate.

Because $\delta_n \to 0$, one has

$$\mathbb{P}\Big((1+\delta_n)S_B(\mathfrak{m}_d^{(n)}) < B\Big) \to 1. \tag{2.13}$$

Also, by assuming that δ_n slowly goes to zero in the sense that $\log\Big(\frac{1}{\delta_n}\Big) = o(\sqrt{n})$, we have by Lemma 2.1 that

$$\mathbb{P}(\delta_n\text{-wafer}) \;=\; 1 - O\left(\frac{-\log\delta_n}{\sqrt{n}}\right) \to 1. \tag{2.14}$$

Since E_n is the intersection of these two events, we see that $\mathbb{P}(E_n) \to 1$ because of an inclusion-exclusion bound that we also used to obtain (2.12).

$$\mathbb{P}(A\cap B) \ge \mathbb{P}(A) + \mathbb{P}(B) - 1. \tag{2.15}$$

By our assumption that $\mathfrak{m}_d^{(n)}$ converges to strong Benford behavior, we have that

$$\mathbb{P}\Big(S_B(\mathfrak{m}_d^{(n)}) \le D\Big) \to \log_B(D). \tag{2.16}$$

Because of uniform convergence (due to compactness), we also have that

$$\mathbb{P}\Big(S_B(\mathfrak{m}_d^{(n)}) \le D/(1+\delta_n)\Big) \to \log_B(D). \tag{2.17}$$

Therefore by the squeeze theorem, we deduce that

$$\mathbb{P}\Big(S_B(v_d^{(n)}) \leq D\Big) \to \log_B(D), \tag{2.18}$$

provided that $\mathfrak{m}_d^{(n)}$ converges to strong Benford behavior. □

This proves Theorem 1.9, because $v_d^{(n)}$ and $V_d^{(n)}$ differ by only a constant multiplicative factor of 2^{m-d}.

3. A Family of Distributions whose Maximum Side-Lengths are Benford

For the sake of clean and transparent analysis, we select as our example family identically log-uniform distributions: $\log_B P_i \sim \mathrm{Uniform}(a, b)$ where $a < b \leq 0$. By shifting and scaling each logarithm of a proportion by a constant, we realize that we may "work" with the normalized distribution $\mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$, which has mean zero and variance one. Of course, this means that we are no longer strictly considering a physically realistic linear fragmentation process, because the boxes $\mathfrak{B}_n$ no longer form a descending chain, however for the sake of purely analyzing the Benfordness of our system, this statistical normalization clearly generalizes, and we lose nothing by assuming it. For $1 \leq i \leq m$, we let

$$Z_i^{(n)} \;:=\; \frac{\log_B(P_i^{(1)} \cdots P_i^{(n)}) - n\mu_P}{\sqrt{n} \cdot \sigma_P} \;=\; \frac{\log_B(P_i^{(1)} \cdots P_i^{(n)})}{\sqrt{n}}. \tag{3.1}$$

If $f_n(x)$ denotes the probability density function of any one of the random variables above, then its characteristic function is

$$\widehat{f_n}(k) \;:=\; \int_{-\infty}^{\infty} f_n(x) e^{ikx}\, dx \;=\; \left(\mathrm{sinc}\left(\frac{k\sqrt{3}}{\sqrt{n}}\right)\right)^n. \tag{3.2}$$

Note that the function $\mathrm{sinc} : \mathbb{R} \to \mathbb{R}$ is defined by

$$\mathrm{sinc}(u) \;:=\; \begin{cases} \frac{\sin(u)}{u} & u \neq 0 \\ 1 & u = 0 \end{cases}. \tag{3.3}$$

The above formula for $\widehat{f_n}(k)$ follows by writing $f_n(x)$ as the n-fold convolution of the PDF of P_i, and then normalizing the PDF by subtracting mean and dividing by variance. We then use the fact that the characteristic function (or Fourier transform) turns convolutions into products.

Our goal is to produce an estimate of the closeness of the PDF $f_n(x)$ and the Gaussian function

$$\varphi(x) \;:=\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{3.4}$$

As we will see, this closeness will allow us to correctly estimate the probability of events involving the significand.

Lemma 3.1. *The following estimate is satisfied by the random variables $Z_i^{(n)}$.*

$$f_n(x) \;=\; \varphi(x) + O(n^{-1+4\varepsilon}). \tag{3.5}$$

Proof. We begin by writing, for $|k|\leq n^{\varepsilon}$,

$$\begin{aligned}
\widehat{f_n}(k) &= \left(1-\frac{k^2}{2n}+O\left(\frac{k^4}{n^2}\right)\right)^n \\
&= \exp\left[n\log\left(1-\frac{k^2}{2n}+O\left(\frac{k^4}{n^2}\right)\right)\right] \\
&= \exp\left[n\left(\log\left(1-\frac{k^2}{2n}\right)+\frac{1}{1-\frac{k^2}{2n}}O\left(\frac{k^4}{n^2}\right)\right)\right] \\
&= \exp\left[n\left(-\frac{k^2}{2n}+O\left(\frac{k^4}{n^2}\right)+O(1)O\left(\frac{k^4}{n^2}\right)\right)\right] \\
&= \exp\left[-\frac{k^2}{2n}+O(n^{-1+4\varepsilon})\right] \\
&= e^{-k^2/2}\cdot\left(1+O(n^{-1+4\varepsilon})\right). \qquad (3.6)
\end{aligned}$$

In terms of Fourier inversion, this allows us to manage the bulk part of our sum, namely we obtain

$$\frac{1}{2\pi}\int_{|k|\leq n^{\varepsilon}}\widehat{f_n}(k)e^{-ikx}\,dk \;=\; \varphi(x)+O(n^{-1+4\varepsilon}). \qquad (3.7)$$

It therefore suffices to show that we have adequate bandwidth $|k|\leq n^{\varepsilon}$ for recovering $f_n(x)$ from $\widehat{f_n}(k)$ as $n\to\infty$, i.e., to bound the strength of higher frequencies. Using the bound $|\mathrm{sinc}(u)|\leq |u|^{-1}$, we obtain for all k that

$$|\widehat{f_n}(k)|\leq\left(\frac{|k|\sqrt{3}}{\sqrt{n}}\right)^{-n}. \qquad (3.8)$$

This immediately shows that we may discard the set of frequencies $|k|\geq n^{1/2}$, since our bound yields

$$\int_{|k|\geq n^{1/2}}|\widehat{f_n}(k)|\,dk \;\leq\; \frac{2}{n-1}\cdot 3^{-n/2} \;=\; O\left(\frac{3^{-n/2}}{n}\right). \qquad (3.9)$$

We want to show that the Fourier inversion over the middle range of frequencies $n^{\varepsilon}\leq|k|\leq n^{1/2}$ is also a small error term. This is because for $n^{\varepsilon}\leq|k|\leq n^{1/2}$, we estimate, using $\mathrm{sinc}(\sqrt{3}u)\leq 1-\frac{u^2}{2.1}$ for small u, as well as $(1-\frac{1}{N})^N\leq e^{-1}$ for N large,

$$0\;\leq\;\widehat{f_n}(k)\;\leq\;\widehat{f_n}(n^{-\varepsilon})\;\leq\;\left(1-\frac{n^{-1+2\varepsilon}}{2.1}\right)^n\;\leq\;e^{-n^{2\varepsilon}/2.1}. \qquad (3.10)$$

This allows us to estimate

$$\int_{n^{\varepsilon}\leq|k|\;\leq\;n^{1/2}}|\widehat{f_n}(k)|\,dk\leq 2n^{1/2}e^{-n^{2\varepsilon}/2.1}\;=\;O(n^{1/2}e^{-n^{2\varepsilon}/2.1}). \qquad (3.11)$$

We therefore have proven, combining all of our estimates, that

$$f_n(x)\;=\;\frac{1}{2\pi}\int_{-\infty}^{\infty}\widehat{f_n}(k)e^{-ikx}\,dk=\varphi(x)+O(n^{-1+4\varepsilon})+O(n^{1/2}e^{-n^{2\varepsilon}/2.1})+O(3^{-n/2}/n). \qquad (3.12)$$

This yields the desired estimate.

□

Using estimate (3.5), we are able to prove Theorem 1.12.

Remark 3.2. We crucially rely on the fact that the $Z_i^{(n)}$ are independent. For $d>1$, this is no longer true, as $\mathfrak{p}_I^{(n)},\mathfrak{p}_J^{(n)}$ share proportion cuts even if $I\neq J$. This obstruction should be able to be removed with further work.

Corollary 1.13 follows from applying Theorem 1.9 and Theorem 1.12. Let us now prove Theorem 1.12.

Proof. Because the side-lengths $S_i^{(n)} = P_i^{(1)} \cdots P_i^{(n)}$ are independent and identically distributed, the probability density function $g_n(x)$ for the normalized random variable $(\log_B \mathfrak{m}_1^{(n)} - n\mu_P)/\sqrt{n}\sigma_P$ is given by

$$g_n(x) \ := \ mF_n(x)^{m-1} f_n(x). \tag{3.13}$$

This is a basic fact about order statistics (see [Mil2]). Notice that the support of $Z_i^{(n)}$ is $[-\sqrt{3n}, \sqrt{3n}]$. Using Lemma 3.1, we have

$$F_n(x) \ = \ \int_{-\sqrt{3n}}^{x} \varphi(x) + O(n^{-1+4\varepsilon})\, dx \ = \ \Phi(x) + O(n^{-1/2+4\varepsilon}). \tag{3.14}$$

This allows us to say that the maximum of approximate Gaussian random variables is approximately the maximum of Gaussian random variables, i.e., expanding (3.13) using (3.14) and (3.5), one derives

$$g_n(x) \ = \ m\Phi(x)^{m-1}\varphi(x) + O(m^2 n^{-1/2+4\varepsilon}\varphi(x)) + O(m^2 n^{-3/2+8\varepsilon}) + O(mn^{-1+4\varepsilon}). \tag{3.15}$$

Our last step is to compute the probability that $\log_B \mathfrak{m}_1^{(n)} \in (a,b) + \mathbb{Z}$ where $(a,b) \subset (0,1)$. This is given by integrating $g_n(x)$ over the set $E_n = (a/\sqrt{n}, b/\sqrt{n}) + \mathbb{Z}/\sqrt{n}$. Thus the probability is

$$\int_{E_n} g_n(x)\, dx \ = \ \int_{E_n} m\Phi(x)^{m-1}\varphi(x)\, dx + O(m^2 n^{-1/2+4\varepsilon}) + O(m^2 n^{-1+8\varepsilon}) + O(mn^{-1/2+4\varepsilon}). \tag{3.16}$$

The integral on the right hand side represents the probability that the max of m Gaussian random variables lies in the set E_n, and the probability approaches $(b-a)$. Indeed, one way to see this is that we are performing an improper Riemann sum of width $1/\sqrt{n}$ on the fixed Riemann-integrable function $m\Phi(x)^{m-1}\varphi(x)$, and that the set E_n simply is a "dense" subset of the rectangles. Therefore the integral over E_n in the limit approaches the "probability" that a chosen rectangle intersects E_n, which is $(b-a)$, times the limit of the improper Riemann sums of $m\Phi(x)^{m-1}\varphi(x)$, which is simply 1. Therefore choosing $\varepsilon < 1/8$, we have

$$\lim_{n\to\infty} \int_{E_n} g_n(x)\, dx = (b-a). \tag{3.17}$$

From this we deduce that the maximum perimeter sequence $\mathfrak{m}_1^{(n)}$ exhibits Strong Benford behavior. □

4. Future Work

We conjecture that the maximum criterion, i.e., the assumption in Theorem 1.9, holds for a large family of proportion cut distributions. More precisely, we conjecture the following.

Conjecture 4.1. *Every linear-fragmentation process (that is continuous, with finite mean, variance, and third moment) satisfies the maximum criterion in all dimensions* $1 \le d \le m$.

From our work, this conjecture implies the following corollary.

Corollary 4.2 (Strong Benfordness)**.** *Assume that Conjecture 4.1 holds. Then every linear-fragmentation process satisfies the strong form of Benford's law for all dimensions* $1 \le d \le m$.

While the tools we have employed thus far in our work with linear-fragmentation processes are distinct from the methods used previously in working with branching-fragmentation processes, the only substantial difference between linear-fragmentation and branching-fragmentation is the presence of a binary tree of weakly correlated events. Applying linearity of expectation, one sees that the expectation values of the leaves of the tree are the expectation value of the end of a linear-fragmentation process with the same height. Therefore we obtain the following corollary.

Corollary 4.3. *Assume that Conjecture 4.1 holds. Then every branching-fragmentation process (with proportion cuts* P_i *as in Definition 1.7) satisfies*

$$\lim_{n\to\infty} \mathbb{E}[\rho_d^{(n)}(s)] \ = \ \log_B(s), \tag{4.1}$$

where

$$\rho_d^{(n)}(s) \;=\; \frac{1}{2^{mn}} \sum_{i=1}^{2^{mn}} \varphi_s(\mathrm{Vol}_d(\mathfrak{B}_i)). \tag{4.2}$$

4.1. **Restating the Mellin Condition in Terms of Characteristic Functions.** Our last remark section for this paper concerns how to interpret the Mellin condition that is presented in the works of [B–] and [DM] in terms of characteristic functions for the logarithm of proportion cuts.

If $f(t)$ is the probability density of $P \in (0,1)$ where P is a proportion cut, then we want to state the Mellin condition

$$\lim_{n\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \left|\mathcal{M}[f]\left(1 + \frac{i\ell}{\log B}\right)\right|^n \;=\; 0, \tag{4.3}$$

in terms of a characteristic function condition for $f_n(x)$, the probability density of $Z^{(n)}$, defined as

$$Z^{(n)} \;:=\; \frac{\log_B(P^{(1)} \cdots P^{(n)}) - n\mu_P}{\sqrt{n}\cdot\sigma_P}, \tag{4.4}$$

where μ_P, σ_P are constants which denote the mean and variance of $\log_B P$.

Remark 4.4. We have slightly modified the Mellin condition originally specified in equation 1.3. The reason for this change is that we want (i) an arbitrary base B, (ii) an answer in terms of characteristic functions rather than Fourier transforms (dropping 2π), and (iii) we may without loss of generality always take $f_u = f$ (see Remark 1.7 of [DM]).

Now take $t = P$, $s = \log_B P$, and define

$$g(s) = f(B^s)\cdot B^s \ln(B). \tag{4.5}$$

A change of variables from t to s gives

$$\mathcal{M}[f]\left(1 + \frac{i\ell}{\log B}\right) \;=\; \int_0^\infty f(t)t^{i\ell/\ln(B)}\,dt \;=\; \int_{-\infty}^\infty g(s)e^{is\ell}\,ds \;=\; \widehat{g}(\ell) \tag{4.6}$$

We have by definition of the random variable $Z^{(n)}$ that $f_n(x) = \sqrt{n}\cdot\sigma_P\cdot(g^{*n})(n\mu_P + \sqrt{n}\cdot\sigma_P x)$. Without loss of generality assume that $\mu_P = 0$ and $\sigma_P = 1$. Then one obtains by applying the characteristic transform

$$\widehat{f_n}(k) = (\widehat{g})^n(k/\sqrt{n}). \tag{4.7}$$

Therefore we see that the Mellin condition is equivalent to the statement that

$$\lim_{n\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} |\widehat{f_n}(\ell\sqrt{n})| = \;0. \tag{4.8}$$

Thus we observe that this is a very mild regularity condition, since we expect $\widehat{f_n}(k) \approx e^{-k^2/2}$ for nice P. We have seen this condition concretely hold for the family of proportion distributions in Section 3.

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.

[B–] Becker, Thealexa and Burt, David and Corcoran, Taylor C. and Greaves-Tunnell, Alec and Iafrate, Joseph R. and Jing, Joy and Miller, Steven J. and Porfilio, Jaclyn D. and Ronan, Ryan and Samranvedhya, Jirapat and et al., *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.

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

[BH2] A. Berger, and T. P. Hill, *Benford Online Bibliography*, http://www.benfordonline.net.

[Berry] 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.

[CLM12] 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. Durmić and S. J. Miller, *Benford Behaviour of a Higher Dimensional Fragmentation Process*, preprint 2023. https://librarysearch.williams.edu/permalink/01WIL_INST/1faevhg/alma991013795585602786.

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

[Hi1] T. P. Hill, *A Statistical Derivation of the Significant-Digit Law*, Statistical Science **10** (1995), no. 4, 354-363.

[Hi2] T. P. Hill, *The first-digit phenomenon*, American Scientists **86** (1996), 358–363.

[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 (March 2009), 37–60.

[Jing] Joy Jing, *Benford's Law and Stick Decomposition*.

[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.

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

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

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

[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.

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

[MiNi1] 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.

[MiNi2] S. J. Miller and M. J. Nigrini, *Order Statistics and Benford's Law*, International Journal of Mathematics and Mathematical Sciences, (2008), 1-13.

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

[Nig] 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

[NiMi] 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.

[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.

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

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

Email address: lbetti@u.rochester.edu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ROCHESTER, ROCHESTER, NY 14627

Email address: idurmic@student.jyu.fi

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF JYVÄSKYLÄ, JYVÄSKYLÄ, FI, 40740

Email address: zmcd@bu.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, BOSTON UNIVERSITY, BOSTON, MA, 02215

Email address: jack.miller.jbm82@yale.edu

DEPARTMENT OF MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, CT, 06511

Email address: Steven.J.Miller@williams.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267