The surprising accuracy of Benford's law in mathematics

Zhaodong Cai, Matthew Faust, A.J. Hildebrand, Junxian Li, and Yuan Zhang

Abstract. Benford's Law is an empirical "law" governing the frequency of leading digits in numerical data sets. While for real-world data Benford's Law typically represents a relatively crude approximation to the actual frequencies, for mathematical sequences the predictions derived from it can be uncannily accurate. For example, among the first billion powers of 2, exactly 301029995 begin with digit 1, while the Benford prediction for this count is $10^6 \log_{10} 2 = 301029995.66\ldots$. If we ignore the fractional part of the predicted value, this represents a perfect hit. The same "perfect hit" can be observed in the digit 1 counts for the first billion powers of 3 and the first billion powers of 5, and the digit 2 counts among the first billion powers of 3. Are these observations mere coincidences or part of some deeper phenomenon? In this paper we seek to answer this and related questions.

1. INTRODUCTION. *Benford's Law* is the empirical observation that leading digits in many real-world data sets tend to follow the *Benford distribution*, depicted in Figure 1 and given by

$$P(\text{first digit is } d) = P(d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, 2, \ldots, 9. \tag{1.1}$$

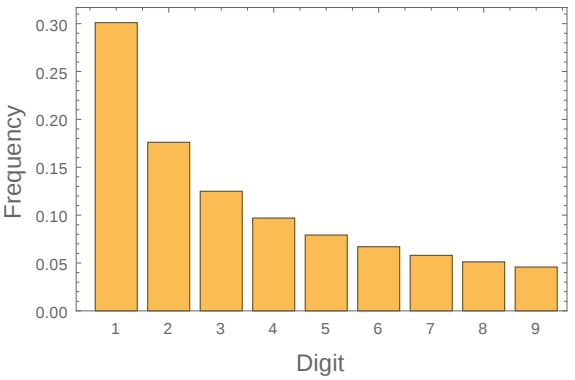

Figure 1. The Benford distribution, $P(d) = \log_{10}(1 + 1/d)$.

Thus, in a data set following Benford's Law, approximately $\log_{10} 2 \approx 30.1\%$ of the numbers begin with digit 1, approximately $\log_{10}(3/2) \approx 17.6\%$ begin with digit 2, while only around $\log_{10}(10/9) \approx 4.6\%$ begin with digit 9.

Benford's Law has been found to be a good match for a wide range of real world data, from populations of cities to accounting data, and it has been the subject of nearly one thousand articles (see the online bibliography [**5**]), including the *Monthly* articles

by Raimi [**20**], Hill [**15**], and Ross [**21**]. It has also long been known (see, e.g., Diaconis [**8**]) that Benford's Law holds for many "natural" mathematical sequences with sufficiently fast rate of growth, such as the Fibonacci numbers, the powers of 2, and the sequence of factorials. In this context, saying that Benford's Law holds is usually understood to mean that, for each digit $d \in \{1, 2, \ldots, 9\}$, the proportion of terms beginning with digit d among the first N terms of the sequence converges to the Benford frequency $P(d)$ given by (1.1), as $N \to \infty$.

How accurate is Benford's Law? Given a sequence such as the powers of 2, Benford's Law predicts that, among the first N terms of the sequence, approximately $N \log_{10}(1 + 1/d)$ begin with digit d, for each $d \in \{1, 2, \ldots, 9\}$. How good are these approximations? A natural benchmark is a random model: Imagine the sequence of leading digits were generated randomly by repeated throws of a 9-sided die with faces marked $1, 2, \ldots, 9$, weighted such that face d comes up with the Benford probability $P(d) = \log_{10}(1 + 1/d)$. Under these assumptions, by the Central Limit Theorem the difference between the actual and predicted digit counts among the first N terms will be roughly of order $\sqrt{N}$. Thus, in a data set consisting of a billion terms (i.e., with $N = 10^9$) it would be reasonable to expect errors on the order of $10,000$.

Random models of the above type form the basis of numerous conjectures in number theory, most notably the Riemann Hypothesis. However, there also exist problems in which, due to additional structure inherent in the problem, it is reasonable to expect smaller errors than the squareroot type errors that are typical for random situations. Two classic examples of this type are the Circle Problem of Gauss and the Divisor Problem of Dirichlet, which have been the subject of a recent *Monthly* article by Berndt, Kim, and Zaharescu [**6**]. In both of these problems the "correct" order of the error terms is believed to be $N^{1/4}$. For $N = 10^9$, this would suggest errors in the order of 100.

Finally, there are examples in number theory in which the approximation error, while still exhibiting "random" behavior, grows at a logarithmic rate. One such case is a problem investigated by Hardy and Littlewood [**13**] concerning the number of lattice points in a right triangle.

How good are the predictions provided by Benford's Law when compared to such benchmarks? The surprising answer is that, in many cases, these predictions seem to be uncannily accurate—more accurate than any of the above benchmarks, and more accurate than even the most optimistic conjectures would lead one to expect. In fact, when we first observed some remarkable coincidences in data we had compiled for a different project[**7**], we thought of them as mere flukes. Later we revisited the problem, approaching it in a systematic manner, expecting to either confirm the "fluke" nature of these coincidences, or to come up with a simple explanation for them.

What we found instead was something far more complex, and more interesting, than any of us had anticipated. Our attempt at getting to the bottom of some seemingly insignificant numerical coincidences turned into a research adventure full of surprises and unexpected twists that required unearthing little known classical results in Diophantine approximation as well as drawing on some of the deepest recent work in the area. In this paper we take the reader along the ride in this adventure in mathematical research and discovery, and we describe the results that came out of this work.

Outline of the paper. The rest of this paper is organized as follows. In Sections 2–4 we present the surprising numerical data alluded to above, we formalize several notions of "unreasonable" accuracy, and we pose three questions suggested by the numerical observations that will serve as guideposts for our investigations. The remainder of

the paper is devoted to unraveling the mysteries behind the numerical observations and uncovering, to the extent possible, the underlying general phenomenon. We proceed in three stages, corresponding to three different levels of sophistication in terms of the mathematical tools used. The three stages are largely independent of each other, and they can be read independently.

In the first stage, consisting of Sections 5 and 6, we use an entirely elementary approach to settle the mystery in a particularly interesting special case. In the second stage, presented in Sections 7–9, we draw on results by Ostrowski and Kesten from the mid 20th century to obtain a general solution to the mystery in the "bounded Benford error" case. In the third stage, contained in Section 10, we bring recent groundbreaking and deep work of Jozsef Beck to bear on the remaining—and most difficult—case, that of an "unbounded Benford error," and we present the surprising denouement of the mystery in this case.

The final section, Section 11, contains some concluding remarks on extensions and generalizations of these results and related results.

2. NUMERICAL EVIDENCE: EXHIBIT A. We begin by presenting some of the numerical data that had spurned our initial investigations. Our data consisted of leading digit counts for the first billion terms of a variety of "natural" mathematical sequences. Carrying out such large scale computations is a highly non-trivial task that, among other things, required the use of specialized C++ libraries for arbitrary precision real number arithmetic. The technical details are described in [**7**].

Table 1 shows the actual leading digit counts for the sequences $\{2^n\}$, $\{3^n\}$, and $\{5^n\}$, along with the predictions provided by Benford's Law, i.e., $N \log_{10}(1 + 1/d)$, where $N = 10^9$.

Digit	Benford Prediction	$\{2^n\}$	$\{3^n\}$	$\{5^n\}$
1	301029995.66	**301029995**	**301029995**	**301029995**
2	176091259.06	176091267	**176091259**	176091252
3	124938736.61	124938729	**124938737**	124938744
4	96910013.01	**96910014**	96910012	**96910013**
5	79181246.05	79181253	**79181247**	79181239
6	66946789.63	66946788	66946787	66946793
7	57991946.98	57991941	57991952	57991951
8	51152522.45	51152528	51152520	51152519
9	45757490.56	45757485	**45757491**	45757494

Table 1. Predicted versus actual counts of leading digits among the first billion terms of the sequences $\{2^n\}$, $\{3^n\}$, $\{5^n\}$. Entries in **boldface** fall within ± 1 of the predicted counts.

Remarkably, nine out of the 27 entries in this table fall within ± 1 of the Benford predictions and are equal to the floor or the ceiling of the predicted values. This is an amazingly good "hit rate" for numbers that are in the order of 10^8. Of the remaining 18 entries, all are within a single digit error of the predicted value.

As remarkable as these observed coincidences seem to be, one has be careful before jumping to conclusions. For example, a "perfect hit" observed at $N = 10^9$ might just be a coincidence that does not persist at other values of N. Such coincidences would not be particularly unusual in case the errors in the Benford approximations have a slow (e.g., logarithmic) rate of growth.

One must also keep in mind Guy's "Strong Law of Small Numbers" [**12**], which refers to situations in which the "true" behavior is very different from the behavior that can be observed within the computable range. Such situations are not uncommon in number theory; Guy's paper includes several examples. Could it be that the uncanny accuracy of Benford's Law observed in Table 1 is just a manifestation of Guy's "Strong Law of Small Numbers", and thus a complete mirage?

3. PERFECT HITS, ALMOST PERFECT HITS, AND BOUNDED ERRORS. Motivated by the observations in Table 1, we now formalize several notions of "unreasonable" accuracy of Benford's Law.

We begin by introducing some basic notations. We denote by $D(x)$ the *leading* (i.e., *most significant*) digit of a positive number x, expressed in its standard decimal expansion and ignoring leading 0's; for example, $D(\pi) = D(3.141\dots) = 3$ and $D(1/6) = D(0.166\dots) = 1$.

We write $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) for the *floor* (resp. *ceiling*) of a real number x, and $\{x\} = x - \lfloor x \rfloor$ for its fractional part.

Given a sequence $\{a_n\}$ of positive real numbers and a digit $d \in \{1, 2, \dots, 9\}$, we define the associated *leading digit counting function* as

$$S_d(N, \{a_n\}) = \#\{n \le N : D(a_n) = d\}, \tag{3.1}$$

where, here and in the sequel, N denotes a positive integer and the notation "$n \le N$" means that n runs over the integers $n = 1, 2, \dots, N$. We denote the *Benford approximation*, or *Benford prediction*, for the counting function $S_d(N, \{a_n\})$ by

$$B_d(N) = NP(d) = N \log_{10}\left(1 + \frac{1}{d}\right), \tag{3.2}$$

and we define the *Benford error* as the difference between the actual and predicted leading digit counts:

$$E_d(N, \{a_n\}) = S_d(N, \{a_n\}) - B_d(N). \tag{3.3}$$

In terms of these notations, the entries in the second column of Table 1 are $B_d(10^9)$, $d = 1, 2, \dots, 9$, while those in the three right-most columns are $S_d(10^9, \{a^n\})$, $d = 1, 2, \dots, 9$, for $a = 2$, $a = 3$, and $a = 5$.

Definition 3.1 (Perfect Hits, Almost Perfect Hits, and Bounded Errors). *Let $\{a_n\}$ be a sequence of positive real numbers and let $d \in \{1, 2, \dots, 9\}$. We call the Benford prediction for leading digit d in the sequence $\{a_n\}$*

- *a **lower perfect hit** if*

$$S_d(N, \{a_n\}) = \lfloor B_d(N) \rfloor \quad \textit{for all } N \in \mathbb{N}, \tag{3.4}$$

  i.e., if the actual leading count is always *equal to the predicted count rounded* down *to the nearest integer;*

- *an **upper perfect hit** if*

$$S_d(N, \{a_n\}) = \lceil B_d(N) \rceil \quad \textit{for all } N \in \mathbb{N}, \tag{3.5}$$

  i.e., if the actual leading count is always *equal to the predicted count rounded* up *to the nearest integer;*

- *an **almost perfect hit** if there exists $\theta \in [0,1]$ such that*

$$S_d(N,\{a_n\}) = \lfloor B_d(N) + \theta \rfloor \quad \textit{for all } N \in \mathbb{N}. \tag{3.6}$$

*Moreover, we say that the Benford prediction for leading digit d in the sequence $\{a_n\}$ has **bounded error** if there exists a constant C such that*

$$|E_d(N,\{a_n\})| \le C \quad \textit{for all } N \in \mathbb{N}. \tag{3.7}$$

Remarks. (1) Note that, while the actual leading digit count, $S_d(N,\{a_n\})$, is necessarily a nonnegative integer, the Benford prediction for this count, $B_d(N) = N\log_{10}(1+1/d)$, represents an irrational number whenever $N \in \mathbb{N}$. Thus, the best we can hope for is that the actual count is equal to the Benford prediction rounded up or down to an integer. In this sense, the "perfect hit" and "almost perfect hit" cases defined above represent best-possible scenarios. In the "almost perfect hit" case both "up" and "down" rounding may be required, while in the "lower perfect hit" and "upper perfect hit" cases the same type of rounding (either "down" or "up") always gives the exact count.

(2) A lower perfect hit corresponds to the case $\theta = 0$ in the definition (3.6) of an almost perfect hit. Similarly, in view of the identity $\lceil x \rceil = \lfloor x+1 \rfloor$ for $x \in \mathbb{R} \setminus \mathbb{Z}$, an upper perfect hit corresponds to the case of $\theta = 1$ in (3.6). In this sense, the notion of an almost perfect hit can be viewed as a natural extension of the concepts of lower and upper perfect hits.

(3) Writing $S_d(N,\{a_n\}) = B_d(N) + E_d(N,\{a_n\})$, the definitions of lower, upper, and almost perfect hits can be restated in terms of the Benford error $E_d(N,\{a_n\})$:

$$\text{lower perfect hit} \iff -1 < E_d(N,\{a_n\}) < 0 \quad \text{for all } N \in \mathbb{N}, \tag{3.8}$$

$$\text{upper perfect hit} \iff 0 < E_d(N,\{a_n\}) < 1 \quad \text{for all } N \in \mathbb{N}, \tag{3.9}$$

$$\text{almost perfect hit} \iff -1+\theta < E_d(N,\{a_n\}) < \theta \quad \text{for all } N \in \mathbb{N} \text{ and some } \theta \in [0,1]. \tag{3.10}$$

As observed above, of the 27 entries in Table 1 nine are equal to the Benford prediction rounded up or down to an integer. Hence, each of these cases represents a *potential* (lower or upper) perfect hit or almost perfect hit in the sense of Definition 3.1. This suggests the following questions:

Question 1 (Perfect Hits). *Which, if any, of the nine* observed *"perfect hits" in Table 1 are "for real", i.e., are instances of a true lower or upper perfect hit, or an almost perfect hit, in the sense of Definition 3.1?*

Question 2 (Bounded Errors). *Which, if any, of the 27 entries in Table 1 represent cases in which the Benford prediction has bounded error?*

In this paper we will provide a complete answer to these questions, not only for the sequences shown in Table 1, but for arbitrary sequences of the form $\{a^n\}$. We encourage the reader to guess the answers to these questions before reading on. Suffice it to say that our own initial guesses turned out to be way off!

4. NUMERICAL EVIDENCE: EXHIBIT B. For further insight into the behavior of the Benford approximations, it is natural to consider the *distribution* of the Benford errors defined in (3.3). Focusing on the sequence $\{2^n\}$, we have computed, for each

digit $d \in \{1, 2, \ldots, 9\}$, the quantities $E_d(N; \{2^n\})$, $N = 1, 2, \ldots, 10^9$, and plotted a histogram of the distribution of these 10^9 terms. The results, shown in Figure 2, turned out to be quite unexpected.

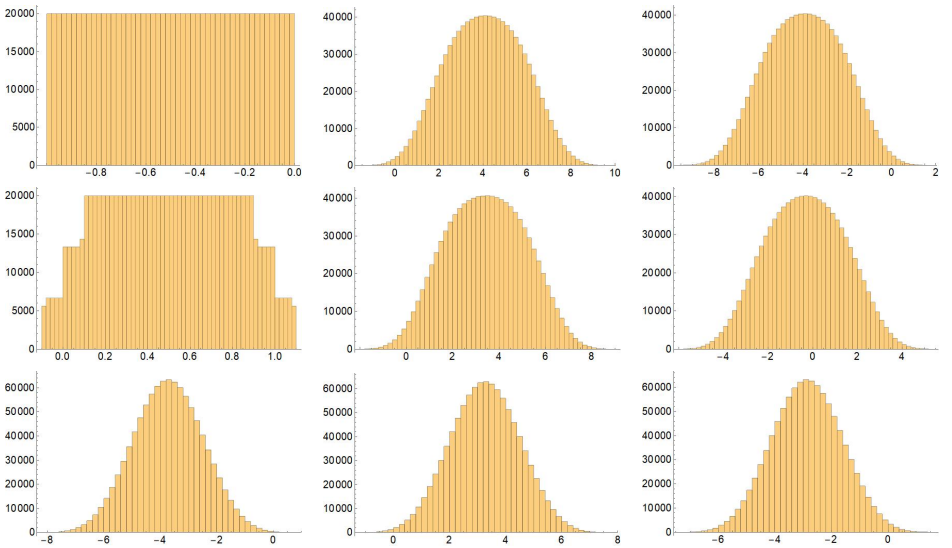

Figure 2. Distribution of the Benford errors for the sequence $\{2^n\}$, based on the first billion terms of this sequence. The three rows of histograms show the distributions of Benford errors for digits 1–3, 4–6, and 7–9, respectively.

The most noticeable, and least surprising, feature in Figure 2 is the distinctive normal shape of seven of the nine distributions shown. This suggests that the corresponding Benford errors are asymptotically normally distributed. The means and standard deviations of these distributions are in the order of single digits, indicating a logarithmic, or even sublogarithmic, growth rate.

The error distribution for digit 1 (shown in the top left histogram) also has an easily recognizable shape: It appears to be a uniform distribution supported on the interval $[-1, 0]$.

By contrast, the error distribution for digit 4 (shown in the middle left histogram) does not resemble any familiar distribution and seems to be a complete mystery. Unraveling this mystery, and discovering the underlying general mechanism, has been a key motivation and driving force in our research; we will describe the results later in this paper. In the meantime, the reader may ponder the following question, keeping in mind the possibility of Guy's "Strong Law of Small Numbers" being in action.

Question 3 (Distribution of Benford Errors). *Which, if any, of the distributions observed in Figure 2 are "for real" in the sense that they represent the true asymptotic behavior of the Benford errors?*

5. UNRAVELING THE DIGIT 1 AND 4 MYSTERIES, I. In this section we focus on the sequence $\{2^n\}$. Using entirely elementary arguments, we seek to unravel some of the mysteries surrounding the leading digit behavior of this sequence we have observed above.

We write

$$S_d(N) = S_d(N, \{2^n\}), \quad E_d(N) = E_d(N, \{2^n\})$$

for the leading digit counting functions, resp. the Benford error functions, associated with the sequence $\{2^n\}$. We will need a slight generalization of $S_d(N)$, defined by

$$S_I(N) = S_I(N, \{2^n\}) = \#\{n \le N : D(2^n) \in I\}, \tag{5.1}$$

where I is an interval in $[1, 10)$.

The key to unlocking the digit 1 and 4 mysteries for the sequence $\{2^n\}$ is contained in the following lemma which provides an explicit formula for $S_I(N)$ for certain intervals I.

Lemma 5.1. *Let $N \in \mathbb{N}$ and $d \in \{1, 2, \ldots, 5\}$. Then*

$$S_{[d,2d)}(N) = \begin{cases} \lfloor N \log_{10} 2 \rfloor & \text{if } d = 1, \\ \lfloor N \log_{10} 2 + \log_{10}(10/d) \rfloor & \text{if } 2 \le d \le 5. \end{cases} \tag{5.2}$$

Proof. Let $N \in \mathbb{N}$ and $d \in \{1, 2, \ldots, 5\}$ be given.

Suppose first that $2^N < d$. In this case we have $2^n \le 2^N < d$ and hence $D(2^n) < d$ for all $n \le N$, and thus $S_{[d,2d)}(N) = 0$. On the other hand, in view of the inequalities

$$0 < N \log_{10} 2 + \log_{10}(10/d) = \log_{10}(2^N/d) + 1 < 1,$$

we have $\lfloor N \log_{10} 2 + \log_{10}(10/d) \rfloor = 0$. Therefore (5.2) holds trivially when $2^N < d$, and we can henceforth assume that

$$2^N \ge d. \tag{5.3}$$

Let k be the unique integer satisfying

$$d \cdot 10^k \le 2^N < d \cdot 10^{k+1}. \tag{5.4}$$

Our assumption (5.3) ensures that k is a *nonnegative* integer, and rewriting (5.4) as

$$\log_{10} d + k \le N \log_{10} 2 < \log_{10} d + k + 1$$

yields the explicit formula

$$k = \lfloor N \log_{10} 2 - \log_{10} d \rfloor . \tag{5.5}$$

Now observe that $d \le D(2^n) < 2d$ holds if and only if 2^n falls into one of the intervals

$$[d \cdot 10^i, 2d \cdot 10^i), \quad i = 0, 1, \ldots \tag{5.6}$$

Since each such interval is of the form $[x, 2x)$, it contains exactly one term 2^n. Hence the number of integers $n \le N$ counted in $S_{[d,2d)}(N)$ is equal to the number of integers i for which the interval (5.6) overlaps with the range $[2^1, 2^N]$. By the definition of k (see (5.4)), this holds if and only if $1 \le i \le k$ in the case $d = 1$, and if and only if $0 \le i \le k$ in the case $d \ge 2$. Thus, $S_{[d,2d)}(N)$ is equal to k in the first case, and $k + 1$ in the second case. Substituting the explicit formula (5.5) for k then yields the desired relation (5.2). ∎

From Lemma 5.1 we derive our first main result, an explicit formula for the Benford errors $E_1(N)$ and $E_4(N)$ associated with the sequence $\{2^n\}$.

Theorem 5.2 (Digit 1 and 4 Benford Errors for $\{2^n\}$). *Let N be a positive integer. Then the Benford errors $E_d(N) = E_d(N, \{2^n\})$ satisfy*

$$E_1(N) = -\{N\alpha\}, \tag{5.7}$$

$$E_4(N) = \{N\alpha\} + \{N\alpha - \alpha\} + \{N\alpha + \alpha\} - 1, \tag{5.8}$$

where $\alpha = \log_{10} 2$.

Proof. For the first formula, note that $D(2^n) = 1$ holds if and only if $1 \le D(2^n) < 2$. Thus we have $S_1(N) = S_{[1,2)}(N)$, and applying Lemma 5.1 with $d = 1$ gives

$$E_1(N) = S_1(N) - B_1(N) = \lfloor N \log_{10} 2 \rfloor - N \log_{10}\left(1 + \frac{1}{1}\right) = -\{N \log_{10} 2\},$$

which proves (5.7).

The proof of the second formula is more involved. The key lies in the relation

$$D(2^n) \neq 4 \iff 1 \le D(2^n) < 2 \text{ or } 2 \le D(2^n) < 4 \text{ or } 5 \le D(2^n) < 10. \tag{5.9}$$

Note that the three conditions on the right of (5.9) are mutually exclusive, and that each of these conditions is of the form appearing in the definition of the quantities $S_{[d,2d)}$. Counting the number of integers $n \le N$ for which the condition on the left (resp. right) side of (5.9) is satisfied therefore yields the following relation between the functions $S_d(N)$ and $S_I(N)$:

$$N - S_4(N) = S_{[1,2)}(N) + S_{[2,4)}(N) + S_{[5,10)}(N).$$

Applying Lemma 5.1 to each of the terms on the right of this relation, we obtain

$$\begin{aligned} S_4(N) &= N - \lfloor N \log_{10} 2 \rfloor - \left\lfloor N \log_{10} 2 + \log_{10} \frac{10}{2} \right\rfloor - \left\lfloor N \log_{10} 2 + \log_{10} \frac{10}{5} \right\rfloor \\ &= N\left(1 - 3\log_{10} 2\right) - 1 + \{N\alpha\} - \{N\alpha + 1 - \alpha\} - \{N\alpha + \alpha\} \\ &= N \log_{10} \frac{5}{4} - 1 + \{N\alpha\} - \{N\alpha - \alpha\} - \{N\alpha + \alpha\}, \end{aligned}$$

where $\alpha = \log_{10} 2$. Since $E_4(N) = S_4(N) - N \log_{10}(5/4)$, this yields the desired formula (5.8). ∎

As an immediate consequence of the formulas (5.7) and (5.8) we obtain the bounds

$$-1 < E_1(N) \le 0 \quad \text{for all } N \in \mathbb{N}, \tag{5.10}$$

$$-1 \le E_4(N) < 2 \quad \text{for all } N \in \mathbb{N}. \tag{5.11}$$

In particular, the Benford errors for digits 1 and 4 for the sequence $\{2^n\}$ are bounded, thus providing a partial answer to Question 3. Moreover, the case $d = 1$ of Lemma 5.1 yields

$$S_1(N) = \lfloor N \log_{10} 2 \rfloor = \lfloor B_1(N) \rfloor \quad \text{for all } N \in \mathbb{N}. \tag{5.12}$$

This shows that the Benford prediction for leading digit 1 for the sequence $\{2^n\}$ is indeed a true perfect hit in the sense of Definition 3.1. Hence, at least one of the five "perfect hits" observed in Table 1 turned out to be "for real".

What about the other four entries in this table that represented perfect hits at $N = 10^9$, i.e., the cases of digits 1 and 2 in the sequence $\{3^n\}$, and digits 1 and 4 in the sequence $\{5^n\}$? Are these "for real" as well, or are they mere coincidences? We will address this question in Section 9 below, but we first use the results of Theorem 5.2 to settle another numerical mystery, namely the distribution of the Benford errors for digits 1 and 4 in Figure 2.

6. UNRAVELING THE DIGIT 1 AND 4 MYSTERIES, II. Continuing our focus on the sequence $\{2^n\}$, we now turn to the *distribution* of the Benford errors $E_1(N)$ and $E_4(N)$ for this sequence and we seek to explain the peculiar shapes of these distributions that we had observed in Figure 2. We will prove:

Theorem 6.1 (Distribution of Digit 1 and 4 Benford Errors for $\{2^n\}$). *The sequences $\{E_1(n)\}$ and $\{E_4(n)\}$ satisfy, for any real numbers $s < t$,*

$$\lim_{N\to\infty} \frac{1}{N}\#\{n \le N : s \le E_i(n) < t\} = \int_s^t f_i(x)\,dx \quad (i = 1, 4), \tag{6.1}$$

where $f_1(x)$ and $f_4(x)$ are defined by

$$f_1(x) = \begin{cases} 1 & \text{if } -1 \le x \le 0, \\ 0 & \text{otherwise,} \end{cases} \tag{6.2}$$

$$f_4(x) = \begin{cases} 1/3 & \text{if } 3\alpha - 1 \le x \le 0 \text{ or } 1 \le x < 2 - 3\alpha, \\ 2/3 & \text{if } 0 \le x < 1 - 3\alpha \text{ or } 3\alpha \le x < 1, \\ 1 & \text{if } 1 - 3\alpha \le x < 3\alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{6.3}$$

where $\alpha = \log_{10} 2$.

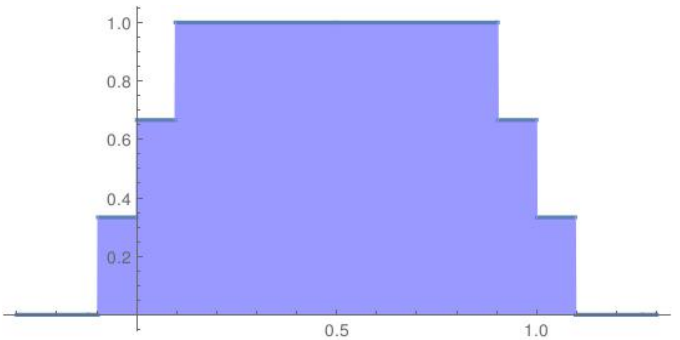

Figure 3. The probability density function $f_4(x)$.

The function $f_1(x)$ is the probability density of a uniform distribution on the interval $[-1, 0]$. The function $f_4(x)$, shown in Figure 3, is a weighted average of three uniform densities supported on the intervals $[1 - 3\alpha, 1]$, $[0, 3\alpha]$, and $[3\alpha - 1, 2 - 3\alpha]$, respectively.

The theorem shows that the error distributions for digits 1 and 4 we had observed in Figure 2 are "for real": The digit 1 error is indeed uniformly distributed over the interval $[-1, 0]$, while the "mystery distribution" of the digit 4 error turns out to be that of the random variable X_4 defined in (6.5).

Proof of Theorem 6.1. By Theorem 5.2 we have

$$E_1(n) = -\{n\alpha\},$$
$$E_4(n) = \{n\alpha\} + \{n\alpha + \alpha\} + \{n\alpha - \alpha\} - 1.$$

The distribution of the numbers $\{n\alpha\}$ in these formulas is well-understood: Indeed, since $\alpha = \log_{10} 2$ is irrational, by *Weyl's Theorem* (see, e.g., [**22**]), these numbers behave like a uniform random variable on the interval $[0, 1]$, in the sense that for any real numbers s, t with $0 \le s < t \le 1$,

$$\lim_{N\to\infty} \frac{1}{N} \#\{n \le N : s \le \{n\alpha\} < t\} = t - s.$$

It follows that the limit distributions of $E_1(n)$ and $E_4(n)$ exist and are those of the random variables

$$X_1 = -U, \tag{6.4}$$
$$X_4 = U + \{U + \alpha\} + \{U - \alpha\} - 1, \tag{6.5}$$

where U is a uniform random variable on $[0, 1]$.

From (6.4) we immediately obtain that X_1 is a uniform random variable on $[-1, 0]$ and hence has density given by the function $f_1(x)$ defined above. Moreover, by considering separately the ranges $0 \le U < \alpha$, $\alpha \le U < 1 - \alpha$, and $1 - \alpha \le U \le 1$ in (6.5), one can check that X_4 is a mixture of three uniform distributions corresponding to these three ranges, and that the density of X_4 is given by the function $f_4(x)$ defined above; we omit the details of this routine, but somewhat tedious, calculation. ■

7. BENFORD ERRORS AND INTERVAL DISCREPANCY. We now consider the case of a general geometric sequence $\{a^n\}$, where a is a positive real number (not necessarily an integer), subject only to the condition

$$\log_{10} a \notin \mathbb{Q}. \tag{7.1}$$

Condition (7.1) serves to exclude sequences such as $\{\sqrt{10}^{\,n}\}$ for which the leading digits behave in a trivial manner.

To make further progress, we exploit the connection between the distribution of leading digits of a sequence and the theory of uniform distribution modulo 1. This connection is well-known, and it has been used to rigorously establish Benford's Law for various classes of mathematical sequences; see, for example, Diaconis [**8**]. For our purposes, we need a specific form of this connection that involves the concept of *interval discrepancy* defined as follows:

Definition 7.1 (Interval Discrepancy). *Let α be a real number, and let I be an interval in $[0, 1]$. For any $N \in \mathbb{N}$, we define the **interval discrepancy** of the sequence $\{n\alpha\}$ with respect to the interval I by*

$$\Delta(N, \alpha, I) = \#\{n \le N : \{n\alpha\} \in I\} - N|I|, \tag{7.2}$$

where $|I|$ denotes the length of I.

The point of this definition is that it allows us to express the Benford error $E_d(N, \{a^n\})$ directly in the form $\Delta(N, \alpha, I)$ with suitable choices of α and I:

Lemma 7.2 (Benford Errors and Interval Discrepancy). *Let a be a positive real number, $N \in \mathbb{N}$, and $d \in \{1, 2, \ldots, 9\}$. Then we have*

$$E_d(N, \{a^n\}) = \Delta(N, \alpha, [\log_{10} d, \log_{10}(d+1)), \tag{7.3}$$

where $\alpha = \log_{10} a$.

Proof. Note that, for any $n \in \mathbb{N}$,

$$\begin{aligned} D(a^n) = d &\iff d \cdot 10^i \le a^n < (d+1) \cdot 10^{i+1} \quad \text{for some } i \in \mathbb{Z} \\ &\iff \log_{10} d + i \le n \log_{10} a < \log_{10}(d+1) + i + 1 \quad \text{for some } i \in \mathbb{Z} \\ &\iff \{n\alpha\} \in [\log_{10} d, \log_{10}(d+1)), \end{aligned}$$

since $\log_{10} a^n = n \log_{10} a = n\alpha$. It follows that

$$S_d(N, \{a^n\}) = \#\{n \le N : \{n\alpha\} \in [\log_{10} d, \log_{10}(d+1))\},$$

and subtracting $B_d(N) = N \log_{10}(1 + 1/d) = N(\log_{10}(d+1) - \log_{10} d)$ on each side yields the desired relation (7.3). ∎

We remark that the interval discrepancy defined above is different from the usual notion of discrepancy of a sequence in the theory of uniform distribution modulo 1, defined as (see, for example, [**17**] and [**9**])

$$D_N(\{n\alpha\}) = \max_{0 \le s < t \le 1} |\Delta(N, \alpha, [s, t))|. \tag{7.4}$$

While there exists a large body of work on the asymptotic behavior of the ordinary discrepancy function D_N, much less is known about the interval discrepancy $\Delta(N, \alpha, I)$. Indeed, the asymptotic behavior of the latter quantity is, in some respects, more subtle and more mysterious than that of the usual discrepancy function. In the following sections we describe some of the key results on interval discrepancies, and we apply these results to Benford errors.

8. INTERVAL DISCREPANCY: RESULTS OF OSTROWSKI AND KESTEN.

In view of Lemma 7.2, the question of whether the Benford error is bounded leads naturally to the following question about the behavior of the interval discrepancy:

Question. *Under what conditions on α and I is the interval discrepancy $\Delta(N, \alpha, I)$ bounded as $N \to \infty$?*

It turns out that this question has a simple and elegant answer, given as follows:

Proposition 8.1 (Bounded Interval Discrepancy (Kesten [16])). *Let α be irrational, and let $I = [s, t)$, where $0 \le s < t \le 1$. Then $\Delta(N, \alpha, I)$ is bounded as $N \to \infty$ if and only if*

$$t - s = \{k\alpha\} \quad \textit{for some } k \in \mathbb{Z} \setminus \{0\}. \tag{8.1}$$

This result has an interesting history going back nearly a century. The sufficiency of condition (8.1) was established by Hecke [**14**] in 1922 for the special case $s = 0$ and by Ostrowski [**18**] in 1927 for general s. The necessity of the condition had been conjectured by Erdős and Szüsz [**10**] and was proved by Kesten [**16**] in 1966.

For the case when condition (8.1) is satisfied, we have the following more precise result that gives an explicit formula for the interval discrepancy. This result is implicit in Ostrowski's paper [**19**] (see formulas (6) and (6') in [**19**]), but since the original paper is not easily accessible, we will provide a proof here.

Proposition 8.2 (Explicit Formula for Interval Discrepancy (Ostrowski [19])). *Let α be irrational, $k \in \mathbb{Z} \setminus \{0\}$, and $0 \le s \le 1 - \{k\alpha\}$. Then we have, for any $N \in \mathbb{N}$,*

$$\Delta(N, \alpha, [s, s + \{k\alpha\}) = \begin{cases} -\displaystyle\sum_{h=0}^{k-1} \Big(\{N\alpha - h\alpha - s\} - \{-h\alpha - s\} \Big) & \text{if } k > 0, \\ \displaystyle\sum_{h=1}^{|k|} \Big(\{N\alpha + h\alpha - s\} - \{h\alpha - s\} \Big) & \text{if } k < 0. \end{cases} \tag{8.2}$$

Proof. Let α, k, and s be given as in the proposition. We start with the elementary identity

$$\{x - t\} - \{x - s\} = \begin{cases} 1 - (t - s) & \text{if } s \le \{x\} < t, \\ -(t - s) & \text{otherwise.} \end{cases} \tag{8.3}$$

which holds for any real numbers x and t with $0 \le s < t \le 1$. Setting $x = \{n\alpha\}$ in (8.3) and summing over $n \le N$, we obtain

$$\begin{aligned} \sum_{n=1}^{N} \Big(\{\{n\alpha\} - t\} - \{\{n\alpha\} - s\} \Big) &= \#\{n \le N : \{n\alpha\} \in [s, t)\} - N(t - s) \\ &= \Delta(N, \alpha, [s, t)). \end{aligned}$$

Specializing t to $t = s + \{k\alpha\}$, the latter sum turns into a telescoping sum in which all except the first and last $|k|$ terms cancel out. More precisely, if $k > 0$, then

$$\begin{aligned} \Delta(N, \alpha, [s, s + \{k\alpha\})) &= \sum_{n=1}^{N} \Big(\{\{n\alpha\} - s - \{k\alpha\}\} - \{\{n\alpha\} - s\} \Big) \\ &= \sum_{n=1}^{N} \Big(\{(n - k)\alpha - s\} - \{n\alpha - s\} \Big) \\ &= -\sum_{h=0}^{k-1} \Big(\{N\alpha - h\alpha - s\} - \{-h\alpha - s\} \Big), \end{aligned}$$

which proves the first case of (8.2). The second case follows by an analogous argument. ∎

9. PERFECT HITS AND BOUNDED ERRORS: THE GENERAL CASE. With the theorems of Kesten and Ostrowski at our disposal, we are finally in a position to settle Questions 1 and 2 and provide a partial answer to Question 3. Our main result is the following theorem, which gives a complete description of all (nontrivial) geometric sequences $\{a^n\}$ and digits $d \in \{1, 2, \ldots, 9\}$ for which the Benford prediction has bounded error or represents one of the "perfect hit" types in Definition 3.1.

Theorem 9.1 (Perfect Hits and Bounded Benford Errors). *Let a be a positive real number satisfying* (7.1)*, and let $d \in \{1, 2, \ldots, 9\}$.*

(i) ***Characterization of bounded Benford errors.*** *The Benford prediction for leading digit d in $\{a^n\}$ has* ***bounded error*** *if and only if*

$$a^k = \frac{d+1}{d} 10^m \quad \text{for some } k \in \mathbb{Z} \setminus \{0\} \text{ and } m \in \mathbb{Z}. \tag{9.1}$$

(ii) ***Characterizations of perfect hits.*** *The Benford prediction for leading digit d in $\{a^n\}$ is*

- *an* ***almost perfect hit*** *(i.e., satisfies $S_d(N, \{a^n\}) = \lfloor B_d(N) + \theta \rfloor$ for all $N \in \mathbb{N}$ and some fixed $\theta \in [0, 1]$) if and only if*

$$a = \frac{d+1}{d} 10^m \text{ or } a = \frac{d}{d+1} 10^m \quad \text{for some } m \in \mathbb{Z}. \tag{9.2}$$

  Moreover, if this condition is satisfied, the parameter θ in the definition of an almost perfect hit is given by

$$\theta = \begin{cases} \{-\log_{10} d\} & \text{if the first case of (9.2) holds,} \\ \log_{10}(d+1) & \text{if the second case of (9.2) holds.} \end{cases}$$

- *a* ***lower perfect hit*** *(i.e., satisfies $S_d(N, \{a^n\}) = \lfloor B_d(N) \rfloor$ for all $N \in \mathbb{N}$) if and only if*

$$d = 1 \text{ and } a = 2 \cdot 10^m \quad \text{for some } m \in \mathbb{Z}; \tag{9.3}$$

- *an* ***upper perfect hit*** *(i.e., satisfies $S_d(N, \{a^n\}) = \lceil B_d(N) \rceil$ for all $N \in \mathbb{N}$) if and only if*

$$d = 9 \text{ and } a = 9 \cdot 10^m \quad \text{for some } m \in \mathbb{Z}; \tag{9.4}$$

(iii) ***Distribution of Benford errors.*** *Suppose one of the above conditions* (9.1)–(9.4)— *holds. Then the distribution of the Benford error $E_d(N, \{a^n\})$ (in the sense of* (6.1)*) is a finite mixture of finite uniform distributions under* (9.1)*, uniform on $[\theta - 1, \theta]$ under* (9.2)*, uniform on $[-1, 0]$ under condition* (9.3)*, and uniform on $[0, 1]$ under* (9.4)*.*

Special cases and consequences. Before proving Theorem 9.1, we present some special cases and consequences of this result.

(1) Up to multiplication of a by a power of 10, there is exactly one case in which the actual leading count is *always* given by the *floor* of the Benford prediction, namely that of digit 1 and the sequence $\{2^n\}$. Similarly, up to multiplication of a by a power of 10, there is exactly one case in which the actual leading count is *always* given by the *ceiling* of the Benford prediction, namely that of digit 9 and the sequence $\{9^n\}$.

(2) For each digit $d \in \{1, 2, \ldots, 9\}$, there are exactly two sequences $\{a^n\}$ with $1 < a < 10$ for which the Benford prediction is an almost perfect hit, corresponding to $a = (d+1)/d$ and $a = 10d/(d+1)$.

(3) If $a > 0$ is *irrational* and not a rational power of 10, then the Benford error for the sequence $\{a^n\}$ is unbounded for all digits $d \in \{1, 2, \ldots, 9\}$.

(4) If a is an integer ≥ 2 that is not divisible by 10, then condition (9.1) reduces to a simple Diophantine equation for the number a and the digit d. This equation has only finitely many solutions, which can be found by considering the prime factorizations of the numbers a, d, and $d+1$. Table 2 gives a complete list of these solutions.

Digit d	Sequences $\{a^n\}$ with bounded Benford error
1	$\{2^n\}$, $\{5^n\}$
2	$\{15^n\}$
3	$\{75^n\}$
4	$\{2^n\}$, $\{5^n\}$, $\{8^n\}$, $\{125^n\}$
5	$\{12^n\}$
6	
7	$\{875^n\}$
8	$\{1125^n\}$
9	$\{3^n\}$, $\{9^n\}$

Table 2. Complete list of digits d and sequences $\{a^n\}$, where $a \geq 2$ is an integer not divisible by 10, for which the Benford prediction has bounded error.

(5) Theorem 9.1 allows us to completely settle Questions 1 and 2 on the true nature of the nine "perfect hits" observed in Table 1. As Table 3 shows, of these nine entries only two are "for real", in the sense of being one of the three types of perfect hits defined in Definition 3.1. An additional three are cases in which the Benford error is bounded, while the remaining four entries are cases in which the Benford error is unbounded.

Sequence	Digit	Observed status at $N = 10^9$	True status
$\{2^n\}$	1	Lower perfect hit	Lower perfect hit
$\{2^n\}$	4	Upper perfect hit	Bounded error
$\{3^n\}$	1	Lower perfect hit	Unbounded error
$\{3^n\}$	2	Lower perfect hit	Unbounded error
$\{3^n\}$	3	Upper perfect hit	Unbounded error
$\{3^n\}$	5	Upper perfect hit	Unbounded error
$\{3^n\}$	9	Upper perfect hit	Bounded error
$\{5^n\}$	1	Lower perfect hit	Almost perfect hit
$\{5^n\}$	4	Lower perfect hit	Bounded error

Table 3. Observed versus true nature of the nine entries in Table 1 for which the actual leading digit count was within ± 1 of the Benford prediction at $N = 10^9$.

(6) Theorem 9.1 can be viewed as a far-reaching generalization of the results (Theorems 5.2 and 6.1) we had obtained above, using a much more elementary approach, for the cases of leading digits 1 and 4 in the sequence $\{2^n\}$. In particular,

the theorem shows that the "brick-shaped" error distributions we had observed in these particular cases are "for real", and that error distributions of this type arise whenever the sequence $\{a^n\}$ and the digit d satisfy the boundedness criterion (9.1) of Theorem 9.1.

Proof of Theorem 9.1. Let a and d be given as in the theorem, and set

$$\alpha = \log_{10} a,\ s = \log_{10} d,\ t = \log_{10}(d+1). \tag{9.5}$$

With these notations the four conditions in the theorem, (9.1)—(9.4), can be restated as follows:

$$t = s + \{k\alpha\} \quad \text{for some } k \in \mathbb{Z} \setminus \{0\}. \tag{9.1'}$$

$$t = s + \{\alpha\} \text{ or } t = s + \{-\alpha\}, \tag{9.2'}$$

$$s = 0 \text{ and } t = \alpha = \log_{10} 2, \tag{9.3'}$$

$$t = 1 \text{ and } s = \alpha = \log_{10} 9, \tag{9.4'}$$

To establish parts (i) and (ii) of the theorem, we need to show these four conditions are, respectively, equivalent to the four cases "bounded Benford error", "almost perfect hit", "lower perfect hit", and "upper perfect hit".

We begin by showing that (9.1)' holds if and only if the Benford error is bounded. By Lemma 7.2 we have

$$E_d(N, \{a^n\}) = \Delta(N, \alpha, [s,t)). \tag{9.6}$$

By Kesten's theorem (Prop. 8.1) it follows that $E_d(N, \{a^n\})$ is bounded as a function of N if and only if condition (9.1)' holds. This proves part (i) of the theorem.

Next, we assume that (9.1)' holds and consider the distribution of the Benford error in this case. Combining (9.6) with Ostrowski's theorem (Prop. 8.2) gives the explicit formula

$$E_d(N, \{a^n\}) = \begin{cases} -\sum_{h=0}^{k-1} \Big(\{N\alpha - h\alpha - s\} - \{-h\alpha - s\} \Big) & \text{if } k > 0, \\ \sum_{h=1}^{|k|} \Big(\{N\alpha + h\alpha - s\} - \{h\alpha - s\} \Big) & \text{if } k < 0, \end{cases} \tag{9.7}$$

with $s = \log_{10} d$ and $\alpha = \log_{10} a$ as in (9.5). Since, by Weyl's Theorem (see (6.4)), the sequence $\{N\alpha\}$ is uniformly distributed modulo 1, it follows that the Benford errors $E_d(N, \{a^n\})$ have a distribution equal to that of the random variable

$$X_{k,\alpha,s} = \begin{cases} -\sum_{h=0}^{k-1} \Big(\{U - h\alpha - s\} - \{-h\alpha - s\} \Big) & \text{if } k > 0, \\ \sum_{h=1}^{|k|} \Big(\{U + h\alpha - s\} - \{h\alpha - s\} \Big) & \text{if } k < 0, \end{cases} \tag{9.8}$$

where U is uniformly distributed on $[0,1]$. The latter distribution is clearly a finite mixture of uniform distributions, thus proving the first assertion of part (iii) of the theorem.

We now turn to the proof of part (ii). Suppose first that the "almost perfect hit" criterion, (9.2)', is satisfied. Then (9.1)' holds with $k = 1$ or $k = -1$, so by (9.7) we have

$$E_d(N, \{a^n\}) = \begin{cases} -\{N\alpha - s\} + \{-s\} & \text{if } k = 1, \\ \{N\alpha + \alpha - s\} - \{\alpha - s\} & \text{if } k = -1. \end{cases} \tag{9.9}$$

Moreover, by (9.8) the distribution of $E_d(N, \{a^n\})$ is that of the random variable

$$X_{1,\alpha,s} = \begin{cases} -\{U - s\} + \{-s\} & \text{if } k = 1, \\ \{U + \alpha - s\} - \{\alpha - s\} & \text{if } k = -1. \end{cases} \tag{9.10}$$

In particular, it follows that $E_d(N, \{a^n\})$ is contained in an interval of the form $(-1 + \theta, \theta)$ with $\theta = \{-s\}$ if $k = 1$, and $\theta = 1 - \{\alpha - s\}$ if $k = -1$. In either case, $E_d(N, \{a^n\})$ satisfies the criterion (3.10) for almost perfect hits. Thus we have established the sufficiency of condition (9.2)' for an almost perfect hit.

Moreover, if (9.2)' holds, then (9.10) shows that the Benford error is uniformly distributed on an interval $[\theta - 1, \theta]$, where $\theta = \{-s\} = \{-\log_{10} d\}$ if $k = 1$ (i.e., in the first case of (9.2)'), and $\theta = 1 - \{\alpha - s\} = 1 - \{-t\} = \{t\} = \log_{10}(d + 1)$ if $k = -1$ (i.e., in the first case of (9.2)'). This proves the "almost perfect hit" case of part (iii) of the theorem.

Now suppose that the Benford prediction is an almost perfect hit. Then (3.10) holds, so the Benford error $E_d(N, \{a^n\})$ is contained in an interval of the form $(-1 + \theta, \theta)$ and, in particular, is bounded. By part (i) of the theorem, it follows that $E_d(N, \{a^n\})$ is given by the explicit formula (9.7) and has the same distribution as the random variable $X_{k,\alpha,s}$ defined in (9.8). To prove (9.2)', it remains to show that we must have $k = 1$ or $k = -1$ in (9.7).

We argue by contradiction. Suppose that $|k| \geq 2$ in (9.7). We will show below that then the support of the random variable $X_{k,\alpha,s}$ covers an interval of length greater than 1. Since $X_{k,\alpha,s}$ is the distribution of the Benford error $E_d(N, \{a^n\}$, it follows that the Benford error cannot be contained in an interval of the form $(-1 + \theta, \theta)$, so we have obtained a contradiction. Therefore we must have $k = 1$ or $k = -1$ as desired.

To prove our claim on the support of $X_{k,\alpha,s}$, suppose $|k| \geq 2$. If we set $U' = \{U - s\}$ if $k > 0$ and $U' = \{U + |k|\alpha - s\}$ if $k < 0$, then U' is uniformly distributed on $[0, 1]$, and (9.8) can be written in the form

$$X_{k,\alpha,s} = U' + \sum_{h=1}^{|k|-1} \{U' - \{h\alpha\}\} + C, \tag{9.11}$$

where $C = C(k, \alpha, s)$ is a constant. Now let $0 < \lambda_1 < \cdots < \lambda_{|k|-1} < 1$ denote the numbers $\{h\alpha\}$, $h = 1, \ldots, |k| - 1$, arranged in increasing order, and set $\lambda_0 = 0$ and $\lambda_{|k|} = 1$. Then (9.11) yields

$$X_{k,\alpha,s} = |k|U' + C_i \quad \text{if } \lambda_i \leq U' < \lambda_{i+1} \tag{9.12}$$

for each $i \in \{0, 1, \ldots, |k| - 1\}$, where $C_i = C_i(k, \alpha, s)$ is a constant. In particular, for each such i the support of $X_{k,\alpha,s}$ covers an interval of length $|k|(\lambda_{i+1} - \lambda_i)$. so we have

$$\max X_{k,\alpha,s} - \min X_{k,\alpha,s} \geq |k|(\lambda_{i+1} - \lambda_i). \tag{9.13}$$

By the pigeonhole principle, one of the intervals $[\lambda_i, \lambda_{i+1})$, $i = 0, \ldots, |k| - 1$, must have length $> 1/|k|$ except in the case when $\lambda_i = i/|k|$ for $i = 0, 1, \ldots, |k|$. But this case is impossible since the numbers λ_i are a permutation of numbers of the form $\{h\alpha\}$ and α is irrational. It follows that, for some $i \in \{0, \ldots, k-1\}$, the right-hand side of (9.13) is strictly greater than 1. Hence $X_{k,\alpha,s}$ is supported on an interval of length greater than 1, thus proving the claim.

To prove the assertions on lower and upper perfect hits, note first that the desired conditions (9.3)' and (9.4)' are the special cases $s = 0$ and $k = 1$, resp. $s = \alpha$ and $k = -1$, of (9.2)'. The latter conditions are equivalent to having $\theta = 0$, resp. $\theta = 1$, in the distribution interval $[\theta - 1, \theta]$ for the Benford error. Thus, (9.3)' holds if and only if the Benford error is contained in $[-1, 0]$, and (9.4)' holds if and only if the Benford error is contained in $[0, 1]$. But by (3.8) and (3.9) the latter two conditions are equivalent to the cases of a lower, resp. upper, perfect hit. This completes the proof of the theorem. ∎

10. THE FINAL FRONTIER: THE CASE OF UNBOUNDED ERRORS. Having characterized the cases when the Benford error is bounded and completely described the behavior of the Benford error for those cases, we now turn to the final—and deepest—piece of the puzzle, the behavior of the Benford error in cases where it is unbounded, i.e., when the boundedness criterion (9.1) of Theorem 9.1 is not satisfied.

Exhibit B, Revisited. For the sequence $\{2^n\}$ the Benford error is unbounded exactly for the digits $d = 2, 3, 5, 6, 7, 8, 9$ (cf. Table 2). Remarkably, those are precisely the digits for which the distribution of the Benford error in Figure 2 has the distinctive shape of a normal distribution. Is this observed behavior for the sequence $\{2^n\}$ "for real", in the sense that the Benford error satisfies an appropriate Central Limit Theorem for these seven digits? Is this behavior "typical" for cases of sequences $\{a^n\}$ and digits d in which the Benford error is unbounded? Could it be that a Central Limit Theorem holds in *all* cases in which the Benford error is unbounded? In other words, is it possible that the distribution of the Benford error for sequences $\{a^n\}$ is either asymptotically normal, or a mixture of uniform distributions?

These are all natural questions suggested by numerical data, and it is not clear where the truth lies. Indeed, we do not *know* the answer, but we will provide heuristics *suggesting* what the truth is and formulate conjectures based on such heuristics.

Interval Discrepancy, Revisited: The Limiting Distribution of $\Delta(N, \alpha, I)$. In view of the connection between Benford errors and the interval discrepancy $\Delta(N, \alpha, I)$ (see Lemma 7.2), it is natural to consider analogous questions about the limiting distribution of the interval discrepancy. In particular, one can ask:

Question. *Under what conditions on α and I does the interval discrepancy $\Delta(N, \alpha, I)$ satisfy a Central Limit Theorem?*

In contrast to the question about bounded interval discrepancy, which had been completely answered more than 50 years ago by Ostrowski and Kesten (see Propositions 8.1 and 8.2), the behavior of $\Delta(N, \alpha, I)$ in the case of unbounded interval discrepancy turns out to be much deeper, and despite some spectacular progress in recent years, a complete understanding remains elusive.

The recent progress on this question is largely due to Jozsef Beck, who over the past three decades engaged in a systematic, and still ongoing, effort to attack questions of this type, for which Beck coined the term "Probabilistic Diophantine Approximation." Beck's work is groundbreaking and extraordinarily deep. The proofs of the results

cited below take up well over one hundred pages and draw on methods from multiple fields, including algebraic and analytic number theory, probability theory, Fourier analysis, and the theory of Markov chains. Beck's recent book "Probabilistic Diophantine Approximation" [**4**] provides a beautifully written, and exceptionally well motivated, exposition of this work, and the profound ideas that underly it. We highly recommend this book to the reader interested in learning more about this fascinating new field at the intersection of number theory and probability theory.

Beck's main result on the behavior of $\Delta(N, \alpha, I)$ is the following theorem. Detailed proofs can be found in his book [**4**], as well as in his earlier papers [**2**] and [**3**].

Proposition 10.1 (Central Limit Theorem for Interval Discrepancy (Beck [4, Theorem 1.1])). *Let α be a quadratic irrational and let $I = [0, s]$, where s is a rational number in $[0, 1]$. Then $\Delta(N, \alpha, [0, s))$ satisfies the Central Limit Theorem*

$$\lim_{N\to\infty} \frac{1}{N} \# \left\{ n \le N : s \le \frac{\Delta(N, \alpha, I) - C_1 \log N}{C_2 \sqrt{\log N}} < t \right\} = \frac{1}{\sqrt{2\pi}} \int_s^t e^{-x^2/2}\, dx \quad \textit{for all } s < t, \tag{10.1}$$

where $C_1 = C_1(\alpha, s)$ and $C_2 = C_2(\alpha, s)$ are constants depending on α and s

This result shows that, under appropriate conditions on α and I, the interval discrepancy, $\Delta(N, \alpha, I)$, is approximately normally distributed with mean and variance growing at a logarithmic rate. This is exactly the type of behavior of the Benford error we had observed in Figure 2 for the digits $2, 3, 5, 6, 7, 8, 9$. Indeed, even the logarithmic rate of growth of the mean and variances in (10.1) is consistent with that observed in Figure 2, and with the numerical size of the errors in Table 1.

Can Proposition 10.1 explain, and rigorously justify, these observations? Unfortunately, the assumptions on α and I in the proposition are too restrictive to be applicable in situations corresponding to Benford errors. Indeed, by Lemma 7.2, the Benford error, $E_d(N, \{a^n\})$, is equal to the interval discrepancy $\Delta(N, \alpha, I_d)$ with $\alpha = \log_{10} a$ and $I_d = [\log_{10} d, \log_{10}(d + 1))$. However, Proposition 10.1 applies only to intervals with *rational endpoints* and thus does not cover intervals of the form I_d. Moreover, in the cases of greatest interest such as the sequence $\{2^n\}$, the number $\alpha = \log_{10} 2$ is not a quadratic irrational and hence not covered by Proposition 10.1.

Of these two limitations to applying Proposition 10.1 to Benford errors, the restriction on the type of interval I seems surmountable. Indeed, Beck [**1**, p. 38] proved a Central Limit Theorem similar to (10.1) for "random" intervals I. Hence, it is at least plausible that the result remains valid for intervals of the type I_d provided $|I_d|$ is not of the form $\{k\alpha\}$ for some $k \in \mathbb{Z}$, which, by Kesten's theorem (Prop. 8.1), would imply bounded interval discrepancy.

Beck's Heuristic. The restriction of α to quadratic irrationals in Proposition 10.1 is due to the fact that quadratic irrationals have a periodic continued fraction expansion, which simplifies the argument. Beck remarks that this restriction can be significantly relaxed, and he provides a heuristic for the class of numbers α for which a Central Limit Theorem should hold, which we now describe.

Consider the continued fraction expansion of α:

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \dots}}}. \tag{10.2}$$

Then, according to Beck's heuristic (see [**1**, p. 38]), the interval discrepancy $\Delta(N, \alpha, I)$ behaves roughly like

$$\Delta(N, \alpha, I) \approx \epsilon_1 a_1 + \epsilon_2 a_2 + \cdots + \epsilon_s a_s, \tag{10.3}$$

where the ϵ_i are independent random variables with values ± 1 and $s = s(N)$ is a suitably chosen cutoff index. By the standard Central Limit Theorem in Probability Theory (see, e.g., Feller [**11**, Section VIII.4]), such a sum has an asymptotically normal distribution if it satisfies the *Lindeberg condition*,

$$\lim_{k\to\infty} \frac{a_k^2}{\sum_{i=1}^{k} a_i^2} = 0. \tag{10.4}$$

Beck [**4**, p. 247] concludes that a Central Limit Theorem for $\Delta(N, \alpha, I)$ can be expected to hold whenever α is an irrational number whose continued fraction expansion satisfies (10.4), while for numbers α that do not satisfy (10.4), a Central Limit Theorem cannot be expected to hold.

Application to Benford Errors. Since, by Lemma 7.2, $E_d(N, \{a^n\}) = \Delta(N, \alpha, I_d)$, where $\alpha = \log_{10} a$ and $I_d = [\log_{10} d, \log_{10}(d+1))$, Beck's heuristic suggests the following conjecture.

Conjecture 10.2 (Central Limit Theorem for Benford Errors). *Let $a > 0$ be a real number satisfying* (7.1)*, and suppose that the continued fraction expansion of $\alpha = \log_{10} a$ satisfies* (10.4)*. Then, for any digit $d \in \{1, 2, \dots, 9\}$ that does* not *satisfy the "bounded error" condition* (9.1) *of Theorem 9.1, the Benford error $E_d(N, \{a^n\})$ is asymptotically normally distributed in the sense that there exist sequences $\{A_N\}$ and $\{B_N\}$ such that*

$$\lim_{N\to\infty} \frac{1}{N} \# \left\{ n \le N : s \le \frac{E_d(N, \{a^n\}) - A_N}{B_N} < t \right\}$$
$$= \frac{1}{\sqrt{2\pi}} \int_s^t e^{-x^2/2}\, dx \quad \textit{for all } s < t. \tag{10.5}$$

This conjecture would explain the normal shape of the distributions of the Benford errors observed in Figure 2 *if* the number $\alpha = \log_{10} 2$ has a continued fraction expansion satisfying (10.4). Unfortunately, we know virtually nothing about the continued fraction expansion of $\log_{10} 2$ and thus are in no position to determine whether or not $\log_{10} 2$ satisfies (10.4). We are similarly ignorant about the nature of the continued fraction expansion of any number of the form

$$\alpha = \log_{10} a,\ a \in \mathbb{N},\ \log_{10} a \notin \mathbb{Q}. \tag{10.6}$$

Thus Conjecture 10.2 does not shed light on the leading digit behavior of the simplest and most interesting class of sequences $\{a^n\}$, namely those where a a positive integer that is not a power of 10.

We can certainly construct numbers a for which $\alpha = \log_{10} a$ satisfies (10.4) (for example, $a = 10^{\sqrt{2}}$), but those constructions are rather artificial, and they do not cover natural families of numbers a such as positive integers or rationals.

If we are willing to believe that all numbers of the form (10.6) satisfy (10.4) *and* assume the truth of Conjecture 10.2, then we would be able to conclude that the Benford error for sequences $\{a^n\}$ with a as in (10.6) satisfies the dichotomy mentioned above: The error is either bounded with a limit distribution that is a finite mixture of uniform distributions, or unbounded with a normal limit distribution. This would be a satisfactory conclusion to our original quest, but it depends on a crucial assumption, namely that (10.4) holds for the numbers of the form (10.6).

How realistic is such an assumption? Alas, it turns out that this assumption is not at all realistic, in the sense that "most" real numbers α do not satisfy (10.4). Indeed, Beck [**1**, p. 39] (see also [**4**, p. 244]) showed that the Gauss-Kusmin theorem, a classical result on the distribution of the terms $a_i = a_i(\alpha)$ in the continued fraction (10.2) of a "random" real number, implies that the set of real numbers $\alpha > 0$ for which (10.4) holds has Lebesgue measure 0. Thus, the condition fails for a "typical" α. Hence, as Beck observes, for a "typical" α, the interval discrepancy $\Delta(N, \alpha, I)$ does *not* satisfy a Central Limit Theorem.

Assuming the numbers $\log_{10} a$ in (10.6) behave like "typical" irrational numbers α, we are thus led to the following unexpected conjecture:

Conjecture 10.3 (Non-normal Distribution of Benford Errors for Integer Sequences $\{a^n\}$). *Let a be any integer ≥ 2 that is not a power of 10, and let $d \in \{1, 2, \ldots, 9\}$. Then the Benford error $E_d(N, \{a^n\})$ does* ***not*** *satisfy a Central Limit Theorem in the sense of* (10.5)*.*

This conjecture, which is based on sound heuristics and thus seems highly plausible, represents a stunning turn-around in our quest to unravel the mysteries behind Figure 2. If true, the conjecture would imply that in *none* of the cases shown in Figure 2 is the distribution asymptotically normal. In particular, the seven distributions in Figure 2 that seemed tantalizingly close to a normal distribution and which appeared to be the most likely candidates for a "real" phenomenon are now being revealed as the (likely) "fakes": The observed normal shapes are (likely) mirages and manifestations of Guy's "Strong Law of Large Numbers".

In light of this conjecture, it is natural to ask why the distributions observed in Figure 2 had such a distinctive normal shape. We believe there are two phenomena at work. For one, the number of "relevant" continued fraction terms a_i in the approximation (10.3) of $\Delta(N, \alpha, I)$ can be expected to be around $\log N$ for most α. Thus, even for values N in the order of one billion, the number of terms in the approximating sum of random variables on the right of (10.3) may be too small to reliably represent the long-term behavior of these sums. Furthermore, while, for a "typical" α, the ratio $a_k^2/(a_1^2 + \cdots + a_k^2)$ appearing in (10.4) is bounded away from 0 for infinitely many values of k, these values of k form a very sparse set of integers, while for "most" k, the above ratio remains small. This would suggest that, even for numbers α that do not satisfy (10.4), $\Delta(N, \alpha, I)$ can be expected to be approximately normal "most of the time".

11. CONCLUDING REMARKS. While our original goal of getting to the bottom of the numerical mysteries in Table 1 and Figure 2 and understanding the underlying

general phenomenon has been largely accomplished, the story does not end here. The results and conjectures obtained suggest a variety of generalizations, extensions, and related questions.

One can consider leading digits with respect to more general bases than base 10. The Benford distribution (1.1) has an obvious generalization for leading digits with respect to an arbitrary integer base $b \geq 3$: simply replace the probabilities $P(d) = \log_{10}(1 + 1/d)$, $d = 1, \ldots, 9$, in (1.1) by the probabilities $P_b(d) = \log_b(1 + 1/d)$, $d = 1, \ldots, b - 1$. We have focused here on the base 10 case for the sake of exposition, but we expect that all of our results and conjectures extend, in a straightforward manner, to this more general setting.

One can ask if similar results hold for more general classes of sequences than the geometric sequences we have considered here. We do expect the results to extend to sequences such as the Fibonacci numbers that are sufficiently to close to geometric sequences, but not for sequences with significantly different rates of growth.

One can seek to more directly tie the behavior of the Benford error to that of the continued fraction expansion of $\alpha = \log_{10} 10$. For example, the heuristic of Beck described in Section 10 suggests that it might be possible to relate the size and behavior Benford error $E_d(N, \{a^n\})$ over a *specific* range for the numbers N to the size and behavior of the continued fraction terms a_k for a corresponding range of indices k.

Finally, one can investigate other measures of "unreasonable" accuracy of the Benford prediction. A particularly interesting one is provided by "record hits" of the Benford prediction, defined as cases where the Benford error at index N is smaller in absolute value than at any previous index. We expect that these indices N are closely tied to the denominators in the continued fraction expansion of $\log_{10} a$.

ACKNOWLEDGMENTS. This work originated with an undergraduate research project carried out in 2016 at the *Illinois Geometry Lab* at the University of Illinois. The experimental results in this paper were generated using the *Illinois Campus Computing Cluster*, a high performance computing platform at the University of Illinois.

REFERENCES

1. József Beck, *Randomness in lattice point problems*, Discrete Math. **229** (2001), no. 1-3, 29–55, Combinatorics, graph theory, algorithms and applications.
2. ———, *Randomness of the square root of 2 and the giant leap, Part 1*, Period. Math. Hungar. **60** (2010), no. 2, 137–242.
3. ———, *Randomness of the square root of 2 and the giant leap, Part 2*, Period. Math. Hungar. **62** (2011), no. 2, 127–246.
4. ———, *Probabilistic Diophantine approximation*, Springer Monographs in Mathematics, Springer, Cham, 2014, Randomness in lattice point counting.
5. Arno Berger, Theodore P. Hill, and E. Rogers, *Benford online bibliography*, `http://www.benfordonline.net`, Last accessed 11.04.2017.
6. Bruce C. Berndt, Sun Kim, and Alexandru Zaharescu, *The circle problem of Gauss and the divisor problem of Dirichlet—still unsolved*, Amer. Math. Monthly **125** (2018), no. 2, 99–114.
7. Zhaodong Cai, Matthew Faust, A. J. Hildebrand, Junxian Li, and Yuan Zhang, *Leading digits of Mersenne numbers*, Preprint (2017).
8. Persi Diaconis, *The distribution of leading digits and uniform distribution* mod 1, Ann. Probability **5** (1977), no. 1, 72–81.
9. Michael Drmota and Robert F. Tichy, *Sequences, discrepancies and applications*, Lecture Notes in Mathematics, vol. 1651, Springer-Verlag, Berlin, 1997.
10. P. Erdős, *Problems and results on diophantine approximations*, Compositio Math. **16** (1964), 52–65 (1964).
11. William Feller, *An introduction to probability theory and its applications. Vol. II*, John Wiley & Sons, Inc., New York-London-Sydney, 1966.
12. Richard K. Guy, *The strong law of small numbers*, Amer. Math. Monthly **95** (1988), no. 8, 697–712.

13. G. H. Hardy and J. E. Littlewood, *Some Problems of Diophantine Approximation: The Lattice-Points of a Right-Angled Triangle*, Proc. London Math. Soc. (2) **20** (1921), no. 1, 15–36.
14. E. Hecke, *über analytische Funktionen und die Verteilung von Zahlen mod. eins*, Abh. Math. Sem. Univ. Hamburg **1** (1922), no. 1, 54–76.
15. Theodore P. Hill, *The significant-digit phenomenon*, Amer. Math. Monthly **102** (1995), no. 4, 322–327.
16. Harry Kesten, *On a conjecture of Erdös and Szüsz related to uniform distribution* mod 1, Acta Arith. **12** (1966/1967), 193–212.
17. Lauwerens Kuipers and Harald Niederreiter, *Uniform distribution of sequences*, Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1974, Pure and Applied Mathematics.
18. Alexander Ostrowski, *Mathematische miszellen. ix. notiz zur theorie der diophantischen approximationen.*, Jahresbericht der Deutschen Mathematiker-Vereinigung **36** (1927), 178–180.
19. ———, *Mathematische miszellen. xvi. zur theorie der linearen diophantischen approximationen.*, Jahresbericht der Deutschen Mathematiker-Vereinigung **39** (1930), 34–46.
20. Ralph A. Raimi, *The first digit problem*, Amer. Math. Monthly **83** (1976), no. 7, 521–538.
21. Kenneth A. Ross, *Benford's law, a growth industry*, Amer. Math. Monthly **118** (2011), no. 7, 571–583.
22. Hermann Weyl, *Über die Gleichverteilung von Zahlen mod. Eins*, Math. Ann. **77** (1916), no. 3, 313–352.

ZHAODONG CAI
University of Pennsylvania, Philadelphia, PA 19104
zhcai@sas.upenn.edu

MATTHEW FAUST
University of Illinois, Urbana, IL 61801
mhfaust2@illinois.edu

A.J. HILDEBRAND (corresponding author)
University of Illinois, Urbana, IL 61801
ajh@illinois.edu

JUNXIAN LI
University of Illinois, Urbana, IL 61801
jli135@illinois.edu

YUAN ZHANG
University of Illinois, Urbana, IL 61801
yzhng195@illinois.edu