The American Mathematical Monthly

ISSN: 0002-9890 (Print) 1930-0972 (Online) Journal homepage: https://www.tandfonline.com/loi/uamm20

The Surprising Accuracy of Benford's Law in Mathematics

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

To cite this article: Zhaodong Cai, Matthew Faust, A. J. Hildebrand, Junxian Li & Yuan Zhang (2020) The Surprising Accuracy of Benford's Law in Mathematics, The American Mathematical Monthly, 127:3, 217-237, DOI: 10.1080/00029890.2020.1690387

To link to this article: https://doi.org/10.1080/00029890.2020.1690387

Published online: 24 Feb 2020.

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. Surprisingly, 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 $10^9 \log_{10} 2 = 301029995.66\ldots$. Similar "perfect hits" can be observed in other instances, such as the digit 1 and 2 counts for the first billion powers of 3. We prove results that explain many, but not all, of these surprising accuracies, and we relate the observed behavior to classical results in Diophantine approximation as well as recent deep conjectures in this area.

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

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

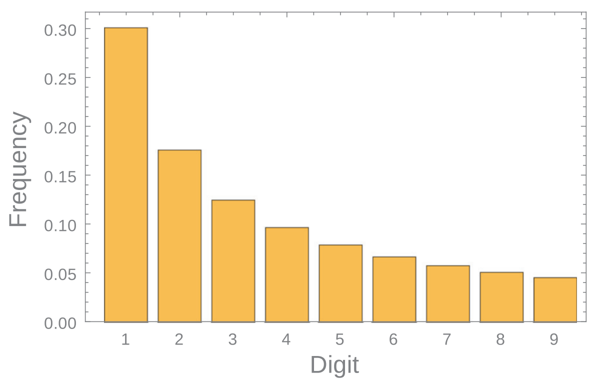

Figure 1. The Benford distribution, $P(d) = \log_{10}(1 + 1/d)$.

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 several MONTHLY articles (e.g., [**15**, **20**, **21**]). It has also long been known (see, e.g., [**8**]) that Benford's law holds for many "natural" mathematical sequences with sufficiently fast rate of growth, such

doi.org/10.1080/00029890.2020.1690387
MSC: Primary 60F99, Secondary 11K31; 11Y55
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uamm.

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 $d \in \{1, 2, \dots, 9\}$, 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), 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 $d \in \{1, 2, \dots, 9\}$. 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 $1, 2, \dots, 9$, 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 square root 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 [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 on 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 appear 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 article, we take the reader along for the ride in this adventure in mathematical research and discovery, and we describe the results that came out of this work.

Outline of the article. The rest of this article is organized as follows. In Sections 2–4, we present the numerical data alluded to above, we formalize several notions of "surprising" accuracy, and we pose three questions suggested by the numerical observations that will serve as guideposts for our investigations. The remainder of the article is devoted to unraveling the mysteries behind the numerical observations and uncovering, to the extent possible, the underlying general phenomena. 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 József 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 spurred 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 nontrivial 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 $N = 10^9$.

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.

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

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 on 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 AND BOUNDED ERRORS.

Motivated by the observations in Table 1, we now formalize several notions of "surprising" 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, $D(\pi) = D(3.141\ldots) = 3$ and $D(1/6) = D(0.166\ldots) = 1$.

We write $\lfloor x \rfloor$ (respectively, $\lceil x \rceil$) for the *floor* (respectively, *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 $d \in \{1, 2, \ldots, 9\}$, we define the associated *leading digit counting function* as

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

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

$$B_d(N) = N P(d) = N \log_{10}\left(1 + \frac{1}{d}\right).$$

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

Definition 1 (Perfect hits and bounded errors). Let $\{a_n\}$ be a sequence of positive real numbers and let $d \in \{1, 2, \ldots, 9\}$.

(i) We call the Benford prediction for the leading digit d in the sequence $\{a_n\}$ a **perfect hit** if it satisfies either

$$S_d(N, \{a_n\}) = \lfloor B_d(N) \rfloor \quad \text{for all } N \in \mathbb{N}, \tag{2}$$

or

$$S_d(N, \{a_n\}) = \lceil B_d(N) \rceil \quad \text{for all } N \in \mathbb{N}, \tag{3}$$

i.e., if the actual leading digit count is *always* equal to the predicted count rounded *down* (respectively, *up*) to the nearest integer. In the first case we call the Benford prediction a **lower perfect hit**, while in the second case we call it an **upper perfect hit**.

(ii) We say that the Benford prediction for the leading digit d in the sequence $\{a_n\}$ has **bounded error** if there exists a constant C such that

$$|S_d(N, \{a_n\}) - B_d(N)| \leq C \quad \text{for all } N \in \mathbb{N}. \tag{4}$$

Remark. Define the *Benford error* as the difference between the actual and predicted leading digit counts:

$$E_d(N, \{a_n\}) = S_d(N, \{a_n\}) - B_d(N). \tag{5}$$

Then the above definitions can be restated in terms of the Benford error as follows.

$$\text{lower perfect hit} \iff -1 < E_d(N, \{a_n\}) \leq 0 \quad \text{for all } N \in \mathbb{N}, \tag{6}$$

$$\text{upper perfect hit} \iff 0 \leq E_d(N, \{a_n\}) < 1 \quad \text{for all } N \in \mathbb{N}, \tag{7}$$

$$\text{bounded error} \iff |E_d(N, \{a_n\})| \leq C \quad \text{for some } C \text{ and all } N \in \mathbb{N}. \tag{8}$$

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* perfect hit in the sense of Definition 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 perfect hit in the sense of Definition 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 article, 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 (5) as N varies. 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.

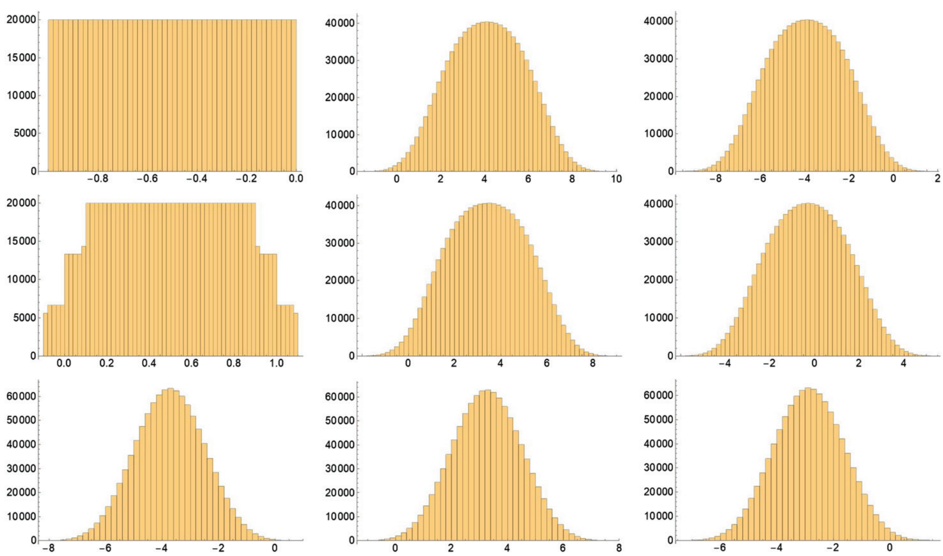

Figure 2. Distribution of the Benford errors for the sequence $\{2^n\}$, based on the first billion terms of this sequence. The three rows of histograms show the distributions of Benford errors for digits 1–3, 4–6, and 7–9, respectively.

The most noticeable, and least surprising, feature in Figure 2 is the approximately 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 article. 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 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, respectively, the Benford error functions, associated with the sequence $\{2^n\}$. We will need a slight generalization of $S_d(N)$, defined by

$$S_I(N) = S_I(N, \{2^n\}) = \#\{n \leq N : D(2^n) \in I\}, \tag{9}$$

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

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

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

Suppose first that $d > 1$ and $2^N < d$. In this case we have $2^n \leq 2^N < d$ and hence $D(2^n) < d$ for all $n \leq N$, 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 (10) holds trivially when $2^N < d$, and we can henceforth assume that

$$2^N \geq d. \tag{11}$$

Let k be the unique integer satisfying

$$d \cdot 10^k \leq 2^N < d \cdot 10^{k+1}. \tag{12}$$

Our assumption (11) ensures that k is a *nonnegative* integer, and rewriting (12) as

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

yields the explicit formula

$$k = \lfloor N \log_{10} 2 - \log_{10} d \rfloor . \tag{13}$$

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

$$[d \cdot 10^i, 2d \cdot 10^i), \quad i = 0, 1, \dots . \tag{14}$$

Since each such interval is of the form $[x, 2x)$, it contains exactly one term 2^n. Hence, since 2^N is one such term, 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 (14) overlaps with the range $[2^1, 2^N]$. By the definition of k (see (12)), 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 (13) for k then yields the desired relation (10). ■

From Lemma 2 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 3 (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*

$$E_1(N) = -\{N\alpha\}, \tag{15}$$

$$E_4(N) = \{N\alpha\} + \{N\alpha - \alpha\} + \{N\alpha + \alpha\} - 1, \tag{16}$$

where $\alpha = \log_{10} 2$. (Recall that $\{x\}$ denotes the fractional part of x.)

Proof. Since $S_1(N) = S_{[1,2)}(N)$, applying Lemma 2 with $d = 1$ gives

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

which proves (15).

The proof of the second formula is slightly more involved. Noting that $S_4(N) = S_{[4,5)}(N)$, we have

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

Applying Lemma 2 to each of the terms on the right of this relation, we obtain

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

Since $E_4(N) = S_4(N) - N \log_{10}(5/4)$, this yields the desired formula (16). ■

As an immediate consequence of the formulas (15) and (16) we obtain the bounds

$$-1 < E_1(N) \le 0 \quad \text{for all } N \in \mathbb{N}, \tag{17}$$

$$-1 \le E_4(N) < 2 \quad \text{for all } N \in \mathbb{N}. \tag{18}$$

In particular, the Benford errors for digits 1 and 4 for the sequence $\{2^n\}$ are bounded, thus providing a partial answer to Question 2. Moreover, the bound (17) is precisely the condition (6) characterizing a (lower) perfect hit, so the Benford prediction for leading digit 1 for the sequence $\{2^n\}$ is indeed a true perfect hit in the sense of Definition 1. Hence, at least one of the nine (potential) "perfect hits" observed in Table 1 turned out to be "for real": the leading digit 1 count for the sequence $\{2^n\}$ is *always* equal to the Benford prediction rounded down to the nearest integer.

What about the other eight entries in this table that represented perfect hits at $N = 10^9$? Are these "for real" as well, or are they mere coincidences? Are there cases where rounding *up* the Benford prediction always gives the exact leading digit count? We will address these questions in Section 9, but we first use the results of Theorem 3 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 the following.

Theorem 4 (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 $s < t$,*

$$\lim_{N\to\infty} \frac{1}{N}\#\{n \le N : s \le E_i(n) < t\} = \int_s^t f_i(x)\,dx \quad (i = 1, 4), \tag{19}$$

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

$$f_1(x) = \begin{cases} 1 & \text{if } -1 \le x \le 0, \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

$$f_4(x) = \begin{cases} 1/3 & \text{if } 3\alpha - 1 \le x \le 0 \text{ or } 1 \le x < 2 - 3\alpha, \\ 2/3 & \text{if } 0 \le x < 1 - 3\alpha \text{ or } 3\alpha \le x < 1, \\ 1 & \text{if } 1 - 3\alpha \le x < 3\alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

where $\alpha = \log_{10} 2 = 0.30103\ldots$.

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 $[1 - 3\alpha, 3\alpha] \approx [0.097, 0.903]$, $[0, 1]$, and $[3\alpha - 1, 2 - 3\alpha] \approx [-0.097, 1.097]$, 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 a superposition of three uniform distributions, given by the density function $f_4(x)$.

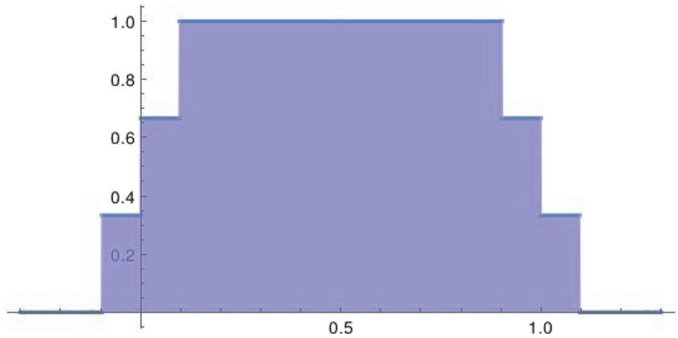

Figure 3. The probability density function $f_4(x)$.

Proof of Theorem 4. By Theorem 3 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

$$X_1 = -U, \tag{22}$$
$$X_4 = U + \{U + \alpha\} + \{U - \alpha\} - 1, \tag{23}$$

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

From (22) 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 (23), one can check that X_4 is a superposition 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

$$\log_{10} a \notin \mathbb{Q}. \tag{24}$$

Condition (24) 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, [**8**]. For our purposes, we need a specific form of this connection that involves the concept of *interval discrepancy* defined as follows.

Definition 5 (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

$$\Delta(N, \alpha, I) = \#\{n \leq N : \{n\alpha\} \in I\} - N|I|, \tag{25}$$

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 6 (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*

$$E_d(N, \{a^n\}) = \Delta(N, \alpha, [\log_{10} d, \log_{10}(d+1)), \tag{26}$$

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

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

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

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

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

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 (26). ∎

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, e.g., [**9**, **17**])

$$D_N(\{n\alpha\}) = \frac{1}{N} \sup_{0 \leq s < t \leq 1} |\Delta(N, \alpha, [s, t))|.$$

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 $\Delta(N, \alpha, I)$. 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 6, the question of whether the Benford error is bounded leads naturally to the following question about the behavior of the interval discrepancy:

Question 4. *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 by the following theorem of Kesten.

Proposition 7 (Bounded interval discrepancy [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*

$$t - s = \{k\alpha\} \quad \textit{for some } k \in \mathbb{Z} \setminus \{0\}. \tag{27}$$

This result has an interesting history going back nearly a century. The sufficiency of condition (27) 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 (27) 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 (Explicit formula for interval discrepancy [19]). *Let α be irrational, $k \in \mathbb{Z} \setminus \{0\}$, and $0 \le s \le 1 - \{k\alpha\}$. Then we have, for any $N \in \mathbb{N}$,*

$$\Delta(N, \alpha, [s, s + \{k\alpha\})) = \begin{cases} -\displaystyle\sum_{h=0}^{k-1} \Big(\{N\alpha - h\alpha - s\} - \{-h\alpha - s\}\Big) & \textit{if } k > 0, \\ \displaystyle\sum_{h=1}^{|k|} \Big(\{N\alpha + h\alpha - s\} - \{h\alpha - s\}\Big) & \textit{if } k < 0. \end{cases} \tag{28}$$

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

$$\{x - t\} - \{x - s\} = \begin{cases} 1 - (t - s) & \text{if } s \le \{x\} < t, \\ -(t - s) & \text{otherwise,} \end{cases} \tag{29}$$

which holds for any real numbers x and t with $0 \le s < t \le 1$. Setting $x = \{n\alpha\}$ in (29) 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 (28). 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 a "perfect hit" in the sense of Definition 1.

Theorem 9 (Perfect hits and bounded Benford errors). *Let a be a positive real number satisfying (24), and let $d \in \{1, 2, \ldots, 9\}$.*

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

$$a^k = \frac{d+1}{d} 10^m \quad \text{for some } k \in \mathbb{Z} \setminus \{0\} \text{ and } m \in \mathbb{Z}. \tag{30}$$

Moreover, if this condition is satisfied, then the limit distribution of the Benford error $E_d(N, \{a^n\})$ (in the sense of (19)) exists and is a weighted average of $|k|$ uniform distributions.

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

- *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*

$$d = 1 \text{ and } a = 2 \cdot 10^m \quad \text{for some } m \in \mathbb{Z}; \tag{31}$$

- *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*

$$d = 9 \text{ and } a = 9 \cdot 10^m \quad \text{for some } m \in \mathbb{Z}. \tag{32}$$

Moreover, if one of the conditions (31) and (32) holds, then the limit distribution of the Benford error $E_d(N, \{a^n\})$ exists and is uniform on $[-1, 0]$ in the first case, and uniform on $[0, 1]$ in the second case.

Theorem 9 can be viewed as a far-reaching generalization of the results (Theorems 3 and 4) 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 (30) of Theorem 9.

Special cases and consequences. If a is an integer ≥ 2 that is not divisible by 10, then condition (30) 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.

Theorem 9, in conjunction with Table 2, allows us to completely settle Questions 1 and 2 on the true nature of the (potential) "perfect hits" observed in Table 1: Of the 27 entries in this table, only the digit-sequence pair $(1, \{2^n\})$ satisfies the perfect hit criterion of Theorem 9 and thus represents a true perfect hit (more precisely, a lower perfect hit). An additional five entries—the pairs $(1, \{5^n\})$, $(4, \{2^n\})$, $(4, \{5^n\})$, and $(9, \{3^n\})$—appear in Table 2 and thus represent cases in which the Benford error is bounded. On the other hand, none of the remaining 21 entries in Table 1 appears in

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. The two entries in boldface, corresponding to the digit-sequence pairs $(1, \{2^n\})$ and $(9, \{9^n\})$, denote cases where the Benford prediction is a perfect hit in the sense of Definition 1.

Digit d	Sequences $\{a^n\}$ with bounded Benford error
1	$\{\mathbf{2}^{\boldsymbol{n}}\}$, $\{5^n\}$
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	$\{3^n\}$, $\{\mathbf{9}^{\boldsymbol{n}}\}$

Table 2, so in all of these cases the Benford error is unbounded and the "perfect hits" observed in Table 1 for some of these cases are mere coincidences.

Proof of Theorem 9. (i) Let a and d be given as in the theorem, and set

$$\alpha = \log_{10} a, \;\; s = \log_{10} d, \;\; t = \log_{10}(d+1). \tag{33}$$

With these notations the condition (30) can be restated as follows.

$$t = s + \{k\alpha\} \quad \text{for some } k \in \mathbb{Z} \setminus \{0\}. \tag{34}$$

We begin by showing that (34) holds if and only if the Benford error is bounded. By Lemma 6 we have

$$E_d(N, \{a^n\}) = \Delta(N, \alpha, [s, t)). \tag{35}$$

By Kesten's theorem (Proposition 7) and since by assumption α is irrational, it follows that $E_d(N, \{a^n\})$ is bounded as a function of N if and only if condition (34) holds. This proves the first assertion of part (i) of the theorem.

Next, we assume that (34) holds and consider the distribution of the Benford error in this case. Combining (35) with Ostrowski's theorem (Proposition 8) gives the explicit formula

$$E_d(N, \{a^n\}) = \begin{cases} -\displaystyle\sum_{h=0}^{k-1} \Big(\{N\alpha - h\alpha - s\} - \{-h\alpha - s\}\Big) & \text{if } k > 0, \\ \displaystyle\sum_{h=1}^{|k|} \Big(\{N\alpha + h\alpha - s\} - \{h\alpha - s\}\Big) & \text{if } k < 0, \end{cases} \tag{36}$$

with $s = \log_{10} d$ and $\alpha = \log_{10} a$ as in (33). Since, by Weyl's theorem (see (22)), 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

$$X_{k,\alpha,s} = \begin{cases} -\sum_{h=0}^{k-1}\Big(\{U - h\alpha - s\} - \{-h\alpha - s\}\Big) & \text{if } k > 0, \\ \sum_{h=1}^{|k|}\Big(\{U + h\alpha - s\} - \{h\alpha - s\}\Big) & \text{if } k < 0, \end{cases} \tag{37}$$

where U is uniformly distributed on $[0, 1]$. The latter distribution is clearly a superposition of $|k|$ uniform distributions, thus proving the last assertion of part (i) of the theorem.

(ii) If (31) or (32) holds, then the bounded error condition (30) is satisfied with $k = 1$ or $k = -1$, so the two expressions for the Benford error $E_d(N, \{a^n\})$ in (36) reduce to a single term, and it is easily checked that this term is equal to $-\{N\alpha\}$ in case (31) holds, and $\{N\alpha\}$ if (32) holds. In the first case, $E_d(N, \{a^n\})$ is contained in the interval $(-1, 0)$ and thus satisfies the lower perfect hit criterion (6), while in the second case $E_d(N, \{a^n\})$ falls into the interval $(0, 1)$ and satisfies the upper perfect hit criterion (7). In either case, Weyl's theorem shows that $E_d(N, \{a^n\})$ is uniformly distributed over the respective interval.

For the converse direction, let a and d be given as in the theorem. and suppose that the Benford prediction for digit d and the sequence $\{a^n\}$ is a (lower or upper) perfect hit. In particular, this implies that the Benford error, $E_d(N, \{a^n\})$, is bounded, so by part (i) $E_d(N, \{a^n\})$ has a limit distribution given by the random variable $X_{k,\alpha,s}$ defined in (37), where k is a nonzero integer and α and s are given by (33).

Suppose first that $k = \pm 1$. Then (37) reduces to

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

Thus, $X_{k,\alpha,s}$ has uniform distribution on the interval $[-1 + \theta, \theta]$, where $\theta = \{-s\}$ if $k = 1$, and $\theta = 1 - \{\alpha - s\}$ if $k = -1$. Since we assumed that the Benford prediction is a perfect hit, we must have either $\theta = 0$ (for a lower perfect hit) or $\theta = 1$ (for an upper perfect hit). In the first case, we have $k = 1$ and $s = 0$, and hence $d = 1$, $t = \log_{10} 2$, and $\{\alpha\} = t - s = \log_{10} 2$. Therefore $a = 10^\alpha = 2 \cdot 10^m$ for some integer m, which is the desired condition (31) for a lower perfect hit. A similar argument shows that in the case $\theta = 1$, the upper perfect hit condition, (32), holds.

Now suppose $|k| \geq 2$. To complete the proof of the necessity of the conditions (31) and (32), it suffices to show that in this case we cannot have a perfect hit. We will do so by showing that the support of the random variable $X_{k,\alpha,s}$ in (37) covers an interval of length greater than 1 and thus, in particular, cannot be equal to one of the intervals $[-1, 0]$ and $[0, 1]$ corresponding to a perfect hit.

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 (37) can be written in the form

$$X_{k,\alpha,s} = U' + \sum_{h=1}^{|k|-1}\{U' - \{h\alpha\}\} + C, \tag{38}$$

where $C = C(k, \alpha, s)$ is a constant. (Note that since $|k| \geq 2$, the sum on the right of (38) contains at least one term.) 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 (38) yields

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

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

$$\max X_{k,\alpha,s} - \min X_{k,\alpha,s} \geq |k|(\lambda_{i+1} - \lambda_i). \tag{39}$$

By the pigeonhole principle, one of the intervals $[\lambda_i, \lambda_{i+1})$, $i = 0, \ldots, |k| - 1$, must have length greater than $1/|k|$ 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 $\{h\alpha\}$, $h = 0, \ldots, |k| - 1$, and α is irrational. It follows that, for some $i \in \{0, \ldots, |k| - 1\}$, the right-hand side of (39) is strictly greater than 1. Hence $X_{k,\alpha,s}$ is supported on an interval of length greater than 1, and the proof of Theorem 9 is complete. ∎

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 (30) of Theorem 9 is not satisfied.

Exhibit B, revisited. For the sequence $\{2^n\}$ the Benford error is unbounded exactly for the digits $d = 2, 3, 5, 6, 7, 8, 9$ (see 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 6), it is natural to consider analogous questions about the limiting distribution of the interval discrepancy. In particular, one can ask:

Question 5. *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 7 and 8), 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 József 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 100 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 [**4**] provides a beautifully written, and exceptionally well-motivated, exposition of this work and the profound ideas that underlie 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**, **3**].

Proposition 10 (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*

$$\lim_{N\to\infty} \frac{1}{N}\#\left\{n \le N : u \le \frac{\Delta(N, \alpha, [0, s]) - C_1 \log N}{C_2\sqrt{\log N}} < v\right\}$$
$$= \frac{1}{\sqrt{2\pi}} \int_u^v e^{-x^2/2}\, dx \quad \text{for all } u < v, \tag{40}$$

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 consistent with the behavior of the Benford error we had observed in Figure 2 for the digits 2, 3, 5, 6, 7, 8, 9.

Can Proposition 10 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 6, 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 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.

Of these two limitations to applying Proposition 10 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 (40) 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 (Proposition 7), would imply bounded interval discrepancy.

Beck's heuristic. The restriction of α to quadratic irrationals in Proposition 10 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:

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}. \tag{41}$$

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

$$\Delta(N, \alpha, I) \approx \epsilon_1 a_1 + \epsilon_2 a_2 + \cdots + \epsilon_k a_k, \tag{42}$$

where the ϵ_i are independent random variables with values ± 1 and $k = k(N)$ is defined by $q_k \leq N < q_{k+1}$, where p_i/q_i denotes the ith partial quotient in the continued fraction expansion of α. By the standard central limit theorem in probability theory (see, e.g., [11, Section VIII.4]), such a sum has an asymptotically normal distribution if it satisfies

$$\lim_{k\to\infty} \frac{a_k^2}{\sum_{i=1}^k a_i^2} = 0. \tag{43}$$

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 (43). On the other hand, Beck [1, p. 38] also notes that in the above situation (43) is essentially necessary for a central limit theorem to hold.

Application to Benford errors. Since, by Lemma 6, $E_d(N, \{a^n\}) = \Delta(N, \alpha, I_d)$, where $\alpha = \log_{10} a$ and $I_d = [\log_{10} d, \log_{10}(d+1))$, Beck's heuristic suggests the following conjecture.

Conjecture 11 (Central limit theorem for Benford errors). *Let $a > 0$ be a real number satisfying (24), and suppose that the continued fraction expansion of $\alpha = \log_{10} a$ satisfies (43). Then, for any digit $d \in \{1, 2, \ldots, 9\}$ that does* not *satisfy the "bounded error" condition (30) of Theorem 9, 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*

$$\lim_{N\to\infty} \frac{1}{N} \# \left\{ n \leq N : u \leq \frac{E_d(N, \{a^n\}) - A_N}{B_N} < v \right\} = \frac{1}{\sqrt{2\pi}} \int_u^v e^{-x^2/2}\, dx \quad \textit{for all } u < v. \tag{44}$$

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 (43). 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 (43). We are similarly ignorant about the nature of the continued fraction expansion of any number of the form

$$\alpha = \log_{10} a, \ a \in \mathbb{N}, \ \log_{10} a \notin \mathbb{Q}. \tag{45}$$

Thus Conjecture 11 does not shed light on the leading digit behavior of the simplest 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 (43) (e.g., numbers a such that $\log_{10} a$ is a quadratic irrational), 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 (45) satisfy (43) *and* assume the truth of Conjecture 11, then we would be able to conclude that the Benford error for sequences $\{a^n\}$ with a as in (45) 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 (43) holds for the numbers of the form (45).

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 (43). 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 (41) of a "random" real number, implies that the set of real numbers $\alpha > 0$ for which (43) 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 (45) behave like "typical" irrational numbers α, we are thus led to the following unexpected conjecture:

Conjecture 12 (Nonnormal distribution of Benford errors for integer sequences $\{a^n\}$). *Let a be any integer greater than or equal to* 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* (44).

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 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 small numbers."

In light of this conjecture, it is natural to ask why the distributions observed in Figure 2 appeared to have a normal shape. We believe there are two phenomena at work. For one, the number of "relevant" continued fraction terms a_i in the approximation (42) of $\Delta(N, \alpha, I)$ can be expected to be around $\log N$ for most α. Thus, even for values N on the order of one billion, the number of terms in the approximating sum of random variables on the right of (42) 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 (43) 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 (43), $\Delta(N, \alpha, I)$ 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 other notions of a "perfect hit," such as situations where rounding the Benford prediction *up or down* always gives the exact count, without insisting on the same type of rounding as we have done in our definition of a (lower or upper) perfect hit. For sequences of the form $\{a^n\}$, these cases can be completely characterized using the methods of this article. For example, the digit-sequence pair $(d, \{a^n\}) = (1, \{5^n\})$, which was one of the potential perfect hits in Table 1, turns out to be a true perfect hit in the above more relaxed sense, but not in the sense of our definition.

One can consider leading digits with respect to more general bases than base 10. The Benford distribution (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) 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 can be extended to more general bases b.

One can ask if similar results hold for more general classes of sequences than the geometric sequences we have considered here. We expect this to be the case for "generic" sequences defined by linear recurrences—including the Fibonacci and Lucas sequences—because solutions to such recurrences can be expressed as linear combinations of geometric sequences, and it seems reasonable to expect that the leading digit behavior of such a linear combination is determined by that of the "dominating" geometric sequence involved.

One can seek to more directly tie the behavior of the Benford error to that of the continued fraction expansion of $\alpha = \log_{10} a$. For example, the heuristic of Beck described in Section 10 (see (42)) suggests that it might be possible to relate the size and behavior of the 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 "surprising" 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. Heuristic arguments, as well as numerical experiments we have carried out, suggest that these indices N are closely tied to the denominators in the continued fraction expansion of $\log_{10} a$.

ACKNOWLEDGMENTS. We are grateful to the referees for their careful reading of the paper and helpful suggestions and comments. This work originated with an undergraduate research project carried out in 2016 at the *Illinois Geometry Lab* (IGL) at the University of Illinois; we thank the IGL for providing this opportunity.

REFERENCES

[1] Beck, J. (2001). Randomness in lattice point problems. *Discrete Math.* 229(1–3): 29–55.

[2] Beck, J. (2010). Randomness of the square root of 2 and the giant leap. Part 1. *Period. Math. Hungar.* 60(2): 137–242.

[3] Beck, J. (2011). Randomness of the square root of 2 and the giant leap. Part 2. *Period. Math. Hungar.* 62(2): 127–246.

[4] Beck, J. (2014). *Probabilistic Diophantine Approximation*. Springer Monographs in Mathematics. Heidelberg: Springer.
[5] Berger, A., Hill, T. P., Rogers, E. (2017). Benford online bibliography. www.benfordonline.net
[6] Berndt, B. C., Kim., S., Zaharescu, C. (2018). The circle problem of Gauss and the divisor problem of Dirichlet—Still unsolved. *Amer. Math. Monthly*. 125(2): 99–114.
[7] Cai, Z., Faust, M., Hildebrand, A. J., Li, J., Zhang, Y. (2019). Leading digits of Mersenne numbers. *Exp. Math.* (to appear). doi.org/10.1080/10586458.2018.1551162
[8] Diaconis, P. (1977). The distribution of leading digits and uniform distribution mod 1. *Ann. Probab*. 5(1): 72–81.
[9] Drmota, M., Tichy, R. (1997). *Sequences, Discrepancies and Applications*. Lecture Notes in Mathematics. Berlin: Springer.
[10] Erdős, P. (1964). Problems and results on diophantine approximations. *Compos. Math.* 16: 52–65.
[11] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed. New York–London–Sydney: John Wiley & Sons, Inc.
[12] Guy, R. (1988). The strong law of small numbers. *Amer. Math. Monthly*. 95(8): 697–712.
[13] Hardy, G. H., Littlewood, J. E. (1922). Some problems of diophantine approximation: The lattice-points of a right-angled triangle. *Proc. Lond. Math. Soc.* 2(1): 15–36.
[14] Hecke, E. (1922). Über analytische Funktionen und die Verteilung von Zahlen mod. eins. *Abh. Math. Sem. Univ. Hamburg*. 1(1): 54–76.
[15] Hill, T. P. (1995). The significant-digit phenomenon. *Amer. Math. Monthly*. 102(4): 322–327.
[16] Kesten, H. (1966). On a conjecture of Erdős and Szüsz related to uniform distribution mod 1. *Acta Arith.* 12(1): 193–212.
[17] Kuipers, L., Niederreiter, H. (1974). *Uniform Distribution of Sequences*. New York–London–Sydney: Wiley-Interscience
[18] Ostrowski, A. (1927). Mathematische Miszellen. IX. Notiz zur Theorie der Diophantischen Approximationen. *Jahresber. Dtsch. Math. Ver*. 36: 178–180.
[19] Ostrowski, A. (1930). Mathematische Miszellen. XVI. Notiz zur Theorie der linearen Diophantischen Approximationen. *Jahresber. Dtsch. Math. Ver*. 39: 34–46.
[20] Raimi, R. A. (1976). The first digit problem. *Amer. Math. Monthly*. 83(7): 521–538.
[21] Ross, K. A. (2011). Benford's law, a growth industry. *Amer. Math. Monthly*. 118(7): 571–583.
[22] Weyl, H. (1916). Über die Gleichverteilung von Zahlen mod. Eins. *Math. Ann.* 77(3): 313–352.

ZHAODONG CAI received his B.S. in mathematics from the University of Illinois in 2017 and is currently a Ph.D. student at the University of Pennsylvania. When not studying mathematics, he likes solving chess puzzles.
Department of Mathematics, University of Pennsylvania, David Rittenhouse Lab, 209 South 33rd St., Philadelphia, PA 19104, USA
zhcai@sas.upenn.edu

MATTHEW FAUST is a Ph.D. student studying mathematics at Texas A&M University. He received B.S. degrees in computer engineering and mathematics in 2018 from the University of Illinois. He is interested in algebraic combinatorics and algebraic geometry. In his free time, he enjoys strategy games, preferably cooperative with Yuan Zhang.
Department of Mathematics, Texas A&M University, Mailstop 3368, College Station, TX 77843, USA
mfaust@math.tamu.edu

A. J. HILDEBRAND received his Ph.D. in mathematics from the University of Freiburg in 1983 and has been at the University of Illinois since 1986, becoming professor emeritus in 2012. Since retiring from the University of Illinois, he has supervised over 100 undergraduates on research projects in pure and applied mathematics and allied areas. The present article grew out of one of these projects.
Department of Mathematics, University of Illinois, 1409 W. Green St., Urbana, IL 61801, USA
ajh@illinois.edu

JUNXIAN LI received her Ph.D. in 2018 from the University of Illinois under the supervision of Alexandru Zaharescu. Her research interests are in number theory. During her Ph.D. studies, she has enjoyed doing research with enthusiastic undergraduates as a graduate mentor in the Illinois Geometry Lab. She is currently a postdoc at the Max Planck Institute for Mathematics.

Max Planck Institute for Mathematics, Vivatsgasse 7, 53111 Bonn, Germany
jli135@mpim-bonn.mpg.de

YUAN ZHANG began his college education at the University of Illinois as a major in natural resources and environment sciences. He soon changed his major to mathematics, receiving his B.S. degree in 2018. He is currently a Ph.D. student studying mathematics at the University of Virginia. He is interested in algebraic topology and algebraic combinatorics. In his free time, he enjoys playing strategy games, especially with his friend Matt Faust.
Department of Mathematics, University of Virginia, 141 Cabell Dr., Kerchof Hall, Charlottesville, VA 22904, USA
yz3yq@virginia.edu

100 Years Ago This Month in *The American Mathematical Monthly*
Edited by Vadim Ponomarenko

Dr. Emmy Noether, daughter of professor Max Noether, and Dr. Schmeidler have been appointed privatdozents at the University of Göttingen.

The Bordin prize (3000 francs) of the Academy of Sciences of the Institute of France is awarded in mathematics every two years for a memoir on a proposed problem. For 1919 the question was: "In the theory of integrals of total differentials of the third kind and of double integrals relative to an algebraic function of two independent variables, there has been proved the existence of certain integers, whose value is difficult to obtain and may depend on the arithmetic nature of the coefficients of the equation of the surface corresponding to the function. The Academy asks for a detailed study of these numbers in some important special cases." This prize was awarded to Professor Solomon Lefschetz, of Kansas University, for a memoir entitled "Sur certains nombres invariants des variétés algébriques avec application aux variétés abéliennes."

—Excerpted from "Notes and News" (1920). 27(3): 141–146.