



АЛГЕБРА И ТЕОРИЯ ЧИСЕЛ
ALGEBRA AND NUMBER THEORY



UDC 511.42

BENFORD'S LAW AND APPROXIMATION OF LOGARITHMS OF NATURAL
NUMBERS BY RATIONAL NUMBERS

V. I. Bernik, N. I. Kalosha, D. V. Vasilyev

Institute of Mathematics of the National Academy of Sciences of Belarus, Minsk, Belarus
e-mail: bernik.vasili@mail.ru, kalosha@im.bas-net.by, vasilyev@im.bas-net.by

Received: 02.04.2025

Revised: 19.05.2025

Accepted: 23.05.2025

Keywords: diophantine approximation, Benford's law, first digit distribution, powers of integers.

Abstract. The paper is devoted to studying the frequencies at which first digits occur in series formed by powers of integer numbers. A number of generalizations of this problem are considered, and the relation between the distribution of first digits and Diophantine properties of logarithms is discussed. In conclusion of the article, several interesting problems in modern theory of Diophantine approximation are proposed.

ЗАКОН БЕНФОРДА И ПРИБЛИЖЕНИЕ ЛОГАРИФМОВ НАТУРАЛЬНЫХ ЧИСЕЛ
РАЦИОНАЛЬНЫМИ

В. И. Берник, Н. И. Калоша, Д. В. Васильев

Институт математики НАН Беларуси, Минск, Беларусь
e-mail: bernik.vasili@mail.ru, kalosha@im.bas-net.by, vasilyev@im.bas-net.by

Поступила: 02.04.2025

Исправлена: 19.05.2025

Принята: 23.05.2025

Ключевые слова: диофантовы приближения, закон Бенфорда, распределение первых цифр, степени целых чисел.

Аннотация. В статье исследованы частотные свойства первых цифр в последовательности, образованной степенями целых чисел. Рассматривается ряд обобщений этой проблемы, а также обсуждается связь между распределением первых цифр и диофантовыми свойствами логарифмов. Предлагается ряд актуальных проблем в теории диофантовых приближений.

1. Introduction

If we exclude initial zeros in the decimal representation of any non-zero real number, there will be nine possibilities for the leading digit: $a = 1, 2, \dots, 9$. Surprisingly, in most cases these digits occur with different frequencies, and the frequency of a digit a is roughly equal to

$$\lg \frac{a+1}{a}. \quad (1)$$

The first publication describing this phenomenon is due to astronomer and mathematician Simon Newcomb [1]. Frank Benford [2] provided numerous real-world examples that exhibit this distribution of first digits. Newcomb's observation became known as Benford's law, and the distribution (1) as Benford's distribution.

Benford's law holds for many random and deterministic sequences, and the sequence 2^n , $n = 1, 2, \dots$, is a notable example. Recently, Benford's law for power sequences was studied by Hürliemann [3].

We are going to prove that sequences a^n , where $a = 2, 3$, follow Benford's law. We are also going to estimate the residual term and show how our estimate is related to Diophantine properties of logarithms of natural numbers and their combinations. Several generalizations of these facts will be discussed, and computer simulations will be used to evaluate the strength of the obtained theoretical results.

This study originates from a conversation between Vasili Bernik and academician Yuri Prokhorov at a number-theoretic conference in Bielefeld. Not every result presented in the paper is new; some of them are repeated for completeness and ease of understanding.

2. The main results

Let $\nu(A, Q)$ be the frequency with which a positive integer A occurs as leading digits in the first Q values of the sequence 2^n ($1 \leq n \leq Q$). Let $[\beta]$ and $\{\beta\}$ denote respectively the integer and fractional parts of a real number β .

Theorem 2.1. *Let A be an arbitrary positive integer. Then there are infinitely many positive integers $n = n(A)$ such that the decimal representation of A coincides with the leading digits in the decimal representation of 2^n .*

The proof of the theorem is based on two simple lemmas.

Lemma 2.2. *The number $\lg 2$ is irrational.*

Lemma 2.3. *For any irrational number α , the sequence $\{n\alpha\}$, $n = 1, 2, \dots$, is everywhere dense in $[0, 1)$.*

The proof of Lemma 2.2 is commonly known, and Lemma 2.3 can be proved using Dirichlet's pigeonhole principle. In fact, Weyl's criterion [4] implies a much stronger result.

Lemma 2.4. *The sequence $\{n\alpha\}$, $n = 1, 2, \dots$, is uniformly distributed on $[0, 1)$ if and only if α is an irrational number.*

Now it is easy to prove Theorem 2.1. Let A have k decimal digits, $A = \overline{a_1 a_2 \dots a_k}$, $0 \leq a_j \leq 9$, $a_1 \neq 0$. Let us write 2^n in the form

$$2^n = 10^{n \lg 2} = 10^{[n \lg 2]} \cdot 10^{\{n \lg 2\}}.$$

The leading digits of 2^n coincide with A if

$$\lg \frac{A}{10^{k-1}} \leq \{n \lg 2\} < \lg \frac{A+1}{10^{k-1}}$$

or

$$\{n \lg 2\} \in [\lg A, \lg(A+1)) \bmod 1. \quad (2)$$

It follows from Lemmas 2.2–2.4 that the condition (2) holds for infinitely many n .

Let $I \subset [0, 1)$ be an interval or a finite union of intervals, and let $N_I(\alpha, Q)$ be the number of positive integers n , $1 \leq n \leq Q$, such that $\{n\alpha\} \in I$. If $\{n\alpha\}$ is uniformly distributed, we have

$$N_I(Q) = (1 + o(1)) Q |I|.$$

Now taking $\alpha = \lg 2$, $I = [\lg A, \lg(A+1)) \bmod 1$ yields

$$N_I(\lg 2, Q) = (1 + o(1)) Q \lg \frac{A+1}{A} = Q \lg \frac{A+1}{A} + R(Q), \quad (3)$$

$$\lim_{Q \rightarrow \infty} Q^{-1} R(Q) = 0,$$

i. e., the frequency with which the leading digits of 2^n coincide with the digits of A is asymptotically equal to $\nu(A) = \lg \frac{A+1}{A}$. In particular,

$$\nu(1) = \lg 2 = 0.3010\dots, \quad \nu(2) = \lg \frac{3}{2} = 0.1760\dots,$$

$$\nu(8) = \lg \frac{9}{8} = 0.0511\dots, \quad \nu(9) = \lg \frac{10}{9} = 1 - \lg 9 = 0.0457\dots$$

Thus decimal representations of the numbers 2^n start with the digit 1 more than 6 times more frequently compared to the digit 9.

A natural question arises: how accurately is the value $N_I(\lg 2, Q)$ approximated by the number $Q \lg \frac{A+1}{A}$? To answer this question, we must estimate from above the remainder $R(Q)$ in (3). This estimate, in turn, is determined by the measure of irrationality of $\lg 2$, i. e., by how well $\lg 2$ is approximated by rational numbers.

Lemma 2.5. *Let the following inequality hold for an irrational number β and any integers $p, q \in \mathbb{Z} \times \mathbb{N}$:*

$$\left| \beta - \frac{p}{q} \right| > c(\beta) q^{-\lambda}, \quad \lambda \geq 2. \quad (4)$$

Then for any interval $I \subset [0, 1)$ we have

$$N_I(\beta, Q) = |I|Q + O\left(Q^{1-\frac{1}{\lambda-1}} \ln Q\right), \quad (5)$$

where the implicit constant in the Vinogradov symbol O does not exceed $\frac{2^{2\lambda+4}}{c(\beta)} + 1$.

Lemma 2.5 can be proved using Vinogradov's "little glasses" method [5] by decomposing the characteristic function of the interval I into a Fourier series. Then the residual term can be estimated from the inequality

$$\sum_{v=1}^L ||v\beta||^{-1} < \frac{2^{\lambda+1}}{c(\beta)} L^{\lambda-1} \ln L, \quad L \geq 8,$$

where $||x||$ is the distance from the real number x to the nearest integer.

The class of numbers $M(\lambda)$ such that the inequality (4) holds is very broad. All real numbers with bounded partial quotients in their continued fraction representations (for example, all quadratic irrationals) lie in the class $M(2)$. For an arbitrary $\lambda > 2$, all real algebraic numbers (Roth, [6]) and almost all real numbers in the sense of Lebesgue measure (Khinchine, [7]) lie in $M(\lambda)$. Today it is known [8] that $\lg 2 \in M(\lambda)$ for $\lambda = \lambda_0 = 2^{42}$, and thus Lemma 2.5 leads to the following quantitative form of Theorem 2.1.

Theorem 2.6. *Let $B(A, Q)$ be the number of positive integers n , $1 \leq n \leq Q$, such that the leading decimal digits of 2^n coincide with the decimal representation of A . Then for any $\varepsilon > 0$ we have*

$$B(A, Q) = Q \lg \frac{A+1}{A} + O\left(Q^{1-1/(\lambda_0-1)+\varepsilon}\right). \quad (6)$$

If we replace the sequence $2^n, n = 1, 2, \dots$, by the sequence $e^n, 1, 2, \dots$, then from results of Masayoshi Hata and Elena Rukhadze [9; 10] we obtain that the residual term in (6) can be replaced by

$$O(Q^{0.66}).$$

It is easy to see that the number 2 in Theorems 2.1 and 2.6 can be replaced by an arbitrary natural number $b \geq 2, b \neq 10^l, l = 0, 1, 2, \dots$. This yields results similar to Theorem 2.6 with the same residual term as in (5) if for all $p, q \in \mathbb{Z} \times \mathbb{N}$ the inequality

$$\left| \lg b - \frac{p}{q} \right| > c(b) q^{-\lambda(b)} \quad (7)$$

is satisfied [8].

Another natural generalization of the problem can be stated as follows. Take two natural numbers A_1 and A_2 . Let $B(A_1, A_2, Q)$ be the number of integers n , $1 \leq n \leq Q$, for which 2^n starts with A_1 , and 3^n starts with A_2 .

Theorem 2.7. *There exists a real number μ , $0 < \mu < 1$, such that for any $0 < \varepsilon < 1 - \mu$ we have*

$$B(A_1, A_2, Q) = Q \lg \frac{A_1+1}{A_1} \lg \frac{A_2+1}{A_2} + O_\varepsilon(Q^{\mu+\varepsilon})$$

as $Q \rightarrow \infty$.

Theorem 2.7 can be proved similarly to Theorem 2.6 since the sequence of two-dimensional vectors $\vec{a}_n = (\{n \lg 2\}, \{n \lg 3\})$ is uniformly distributed in the square $[0, 1) \times [0, 1)$. Hence, these vectors infinitely often belong to the rectangle

$$[\lg A, \lg(A+1)) \times [\lg B, \lg(B+1)) \bmod 1.$$

This follows from the fact that the numbers $1, \lg 2, \lg 3$ are linearly independent over the field of rational numbers and the multivariate Weyl criterion [4]. Moreover, there exists a constant $\lambda_1 = \lambda_1(2, 3)$ that provides a quantitative characteristic of this linear independence of the form

$$|a_3 \lg 3 + a_2 \lg 2 + a_1| > c(2, 3) H^{-\lambda_1}, \quad (8)$$

where $a_j \in \mathbb{Z}$, $a_2^2 + a_3^2 \neq 0$, $H = \max_{1 \leq j \leq 3} |a_j|$.

The estimates (8) and more general estimates

$$|a_k \lg p_k + a_{k-1} \lg p_{k-1} + \dots + a_2 \lg 2 + a_1| > c(2, \dots, p_k) H^{-\lambda_2}, \quad (9)$$

$a_j \in \mathbb{Z}$, $\sum_{j=2}^k a_j^2 \neq 0$, where p_j is the j th prime number, were first obtained by A. Baker [11]. In the paper [8], Baker wrote these estimates in the form (8), (9). From these estimates, we can obtain quantitative results about the uniform distribution of fractions in a k -dimensional unit cube, which are similar to (5). Having obtained such results, we can directly prove Theorem 2.7 and the following more general theorem.

Theorem 2.8. *Let p_k be the k th prime number, and for arbitrary natural numbers A_1, A_2, \dots, A_k let $B(A_1, \dots, A_k, Q)$ be the number of positive integers n , $1 \leq n \leq Q$, such that the leading digits of 2^n coincide with the digits A_1 , 3^n has the same property with respect to A_2, \dots, p_k^n – to A_k . Then we can specify μ_1 , $0 < \mu_1 < 1$, such that for any ε_1 , $0 < \varepsilon_1 < 1 - \mu_1$, as $Q \rightarrow \infty$ we have*

$$B(A_1, \dots, A_k, Q) = Q \prod_{s=1}^k \lg \frac{A_s + 1}{A_s} + O_{\varepsilon_1}(Q^{\mu_1 + \varepsilon_1}).$$

Note that estimates from below for linear forms of the type (9) can be much more accurate if we take particular combinations of numbers b_1, \dots, b_k .

Of course, it isn't necessary to restrict Theorem 2.8 only to prime numbers. The result holds for composite numbers, as long as the set b_1, \dots, b_k satisfies the requirement that their logarithms, taken together with the number 1, are linearly independent over the field of rational numbers. For example, we can take $b_1 = 4$ and $b_2 = 9$, but can't take $b_1 = 2$, $b_2 = 4$ or $b_1 = 2$, $b_2 = 3$, $b_3 = 6$.

Let us consider another generalization. Let $B_2^{(s)}(Q)$ be the number of positive integers n , $1 \leq n \leq Q$, such that the decimal digits of the number 2^n , starting from the $(s+1)$ -th position, coincide with the number A_1 . From now on, we allow A_1 to have leading zeros, for example, $A_1 = 002$). Further, let

$$v_s(A_1) = \sum_{t=10^{s-1}}^{10^s-1} \lg \frac{10^s t + A_1 + 1}{10^s t + A_1}.$$

Theorem 2.9. *As $Q \rightarrow \infty$, we have*

$$B_2^{(s)}(Q) = Q v_s(A_1) + O_{\varepsilon}(Q^{1-1/(\lambda_0-1)+\varepsilon}). \quad (10)$$

The proof of Theorem 2.9 is similar to the proof of Theorem 2.6, since in the decimal representation of 2^n , the digits of A_1 appear starting from the $(s+1)$ -th position if $\{n \lg 2\}$ lies in the interval $[\lg(10^s t + A), \lg(10^s t + A + 1)) \bmod 1$ for some s -digit number t . Theorem 2.6 holds for any such interval. Now it remains to calculate a sum of the right-hand sides of the expressions of the type (6) over all $9 \cdot 10^{s-1}$ possible values of t .

Clearly, this direct approach leads to a significant increase of the residual term in (10) because of the implicit constant in the Vinogradov symbol. Take, for example, $s = 6$, then the residual term in (10) for the sequence e^n becomes at least 10^6 . Therefore, to obtain meaningful estimates of the remainder term of (10), we need to take Q of the order 10^{21} , which is very large.

The results of numerical experiments (see Section 3) suggest that fluctuations of the residual term for the individual intervals cancel each other out, to a degree, when summation is performed. We were able to formally prove that this type of interference does occur for a union of evenly spaced intervals of equal length.

Lemma 2.10. *Let the inequality (7) hold for an irrational β and all $(p, q) \in \mathbb{Z} \times \mathbb{N}$. Then for all integers $M > 1$, $S \leq M$, real numbers a and b , $0 < b - a < M^{-1}$, and*

$$V = \cup_{j=0}^S [a + jM^{-1}, b + jM^{-1}),$$

the asymptotic equality

$$N_V(\beta, Q) = |V|Q + O_{\varepsilon}(Q^{1-\frac{1}{\lambda-1}} \ln Q),$$

holds, where the implicit constant in the Vinogradov symbol O does not exceed

$$2^{2\lambda+10} \ln M \left(\frac{1}{c(M\beta)M^{\lambda-2}} + \frac{1}{c(\beta)} \right).$$

The result of Lemma 2.10 depends on Diophantine properties of both β and $M\beta$. The relation between $c(\beta)$ and $c(M\beta)$ is not simple, but for most numbers (in the sense of Lebesgue measure) these quantities can be considered to be equal.

Recall that the integer base 2 can be replaced with an arbitrary real base $a > 1$. Take

$$\lg a_1 = \frac{\sqrt{5}-1}{2}, \quad \lg a_2 = \sqrt{2}, \quad \lg a_3 = e, \quad \lg a_4 = \pi$$

and let $B_{a_j}(Q)$, $1 \leq j \leq 4$, be the number of integers n , $1 \leq n \leq Q$, such that the leading digits of the number a_j^n coincide with the digits of a positive integer A . Using well-known results about rational approximation of $\lg a_j$, we can retrace the proof of Theorem 2.6 to obtain the following result.

Theorem 2.11. *As $Q \rightarrow \infty$, we have*

$$B_{a_1}(Q) = Q \lg \frac{A+1}{A} + O(\ln Q); \quad (11)$$

$$B_{a_2}(Q) = Q \lg \frac{A+1}{A} + O(\ln Q); \quad (12)$$

$$B_{a_3}(Q) = Q \lg \frac{A+1}{A} + O(\ln^2 Q); \quad (13)$$

$$B_{a_4}(Q) = Q \lg \frac{A+1}{A} + O\left(Q^{\frac{5}{6}}\right). \quad (14)$$

For almost all a and an arbitrary positive constant $\varepsilon > 0$, we have

$$B_a(A, Q) = Q \lg \frac{A+1}{A} + O(\ln^{2+\varepsilon} Q). \quad (15)$$

There is every reason to expect that the estimate of the residual term in (11) will be the tightest among (11)–(15) since the golden ratio $\frac{\sqrt{5}-1}{2}$ has the continued fraction representation $[1, 1, \dots]$, leading to the estimate (4) with $c(a_1) = (\sqrt{5})^{-1} - \delta$, $\lambda = 2$, for any $\delta > 0$ and $q > q_0(\delta)$.

Let us consider another generalization of the original problem. Take $m_1 = 10^x, \dots, m_k = 10^k$. For these sequences, we can define $B'(A_1, \dots, A_k, Q)$ similarly to $B(A_1, \dots, A_k, Q)$ in Theorem 2.8 by replacing p_j with m_j . Then the well-known metric lower bound on values of polynomials with integer coefficients [12] can be used to obtain the following theorem.

Theorem 2.12. *For almost all x and any $\delta > 0$, we have*

$$B'(A_1, \dots, A_k, Q) = Q \prod_{s=1}^k \lg \frac{A_s+1}{A_s} + O_\delta(\ln^{k+\delta} Q).$$

3. Results of computational experiments

In order to evaluate the accuracy of our theoretical bounds, we have performed a number of computations for large numbers Q , obtaining the following results.

Calculation of $B(A, Q)$ for $b = 2$ and all $A = 1, 2, \dots, 9$ shows that the deviation of $B(A, Q)$ from the asymptotic estimate $Q \lg \frac{A+1}{A}$ does not exceed 7 for $1 \leq Q \leq 10^6$. This suggests that $\lg 2$ is like most real numbers, i. e., that it lies in the class $M(2)$.

Looking at the sequences 2^n and 3^n simultaneously, for the quantity $B(A_1, A_2, Q)$ we have obtained that for all combination of the numbers (A_1, A_2) , $1 \leq A_1 \leq 9$, $1 \leq A_2 \leq 9$, $Q \leq 10^6$, we have

$$\left| B(A_1, A_2, Q) - Q \lg \frac{A_1+1}{A_1} \lg \frac{A_2+1}{A_2} \right| \leq 12.$$

Finally, considering the second, third, fourth, fifth and sixth digits for bases 2 and 3, and for all possible digits $0 \leq A_1 \leq 9$, we obtain that the maximum deviation of $B_2^{(s)}(A_1, Q)$ and $B_3^{(s)}(A_1, Q)$ from $Q \nu_s(A_1)$ for $0 \leq Q \leq 10^6$ does not exceed 22 ($s = 2$), 65 ($s = 3$), 122 ($s = 4$), 405 ($s = 5$), 921 ($s = 6$).

These results are visualized in the Figures 1 and 2 below.

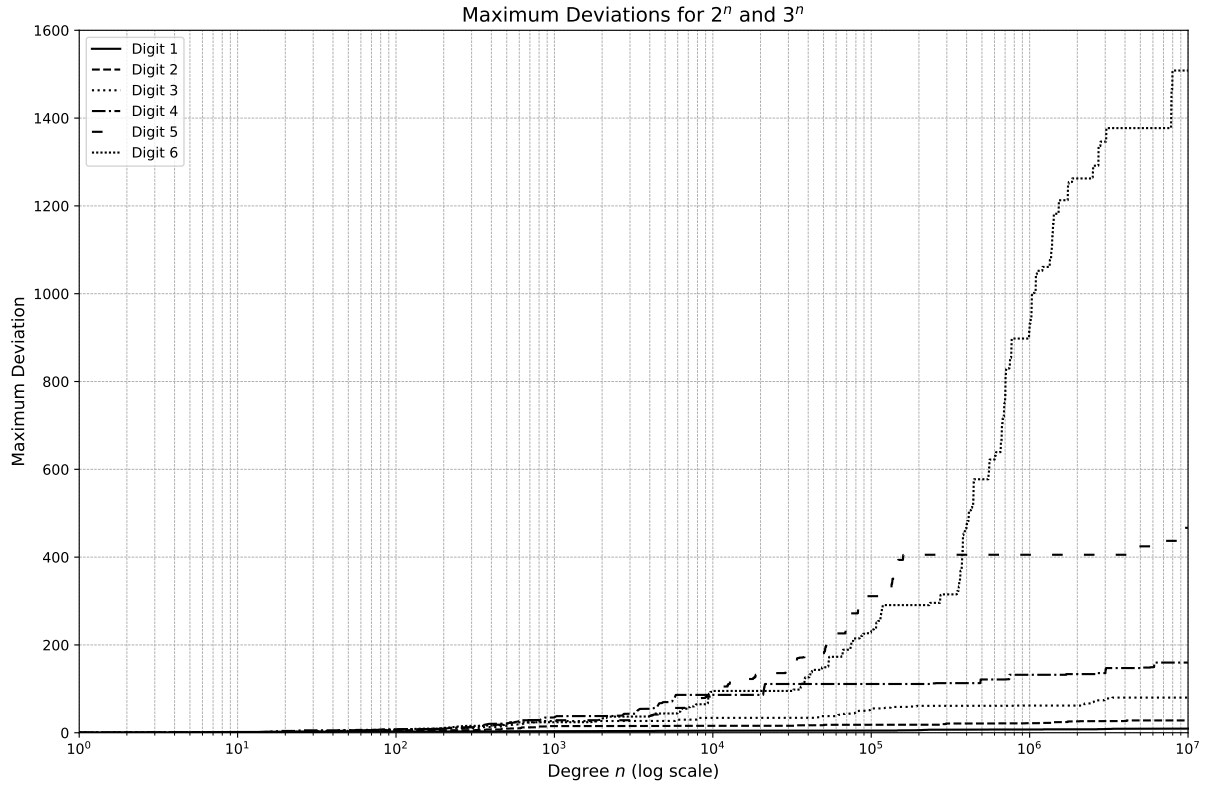


Fig. 1. Maximum deviations of $B_2^{(s)}(A_1, Q)$ and $B_3^{(s)}(A_1, Q)$ from the respective asymptotic values

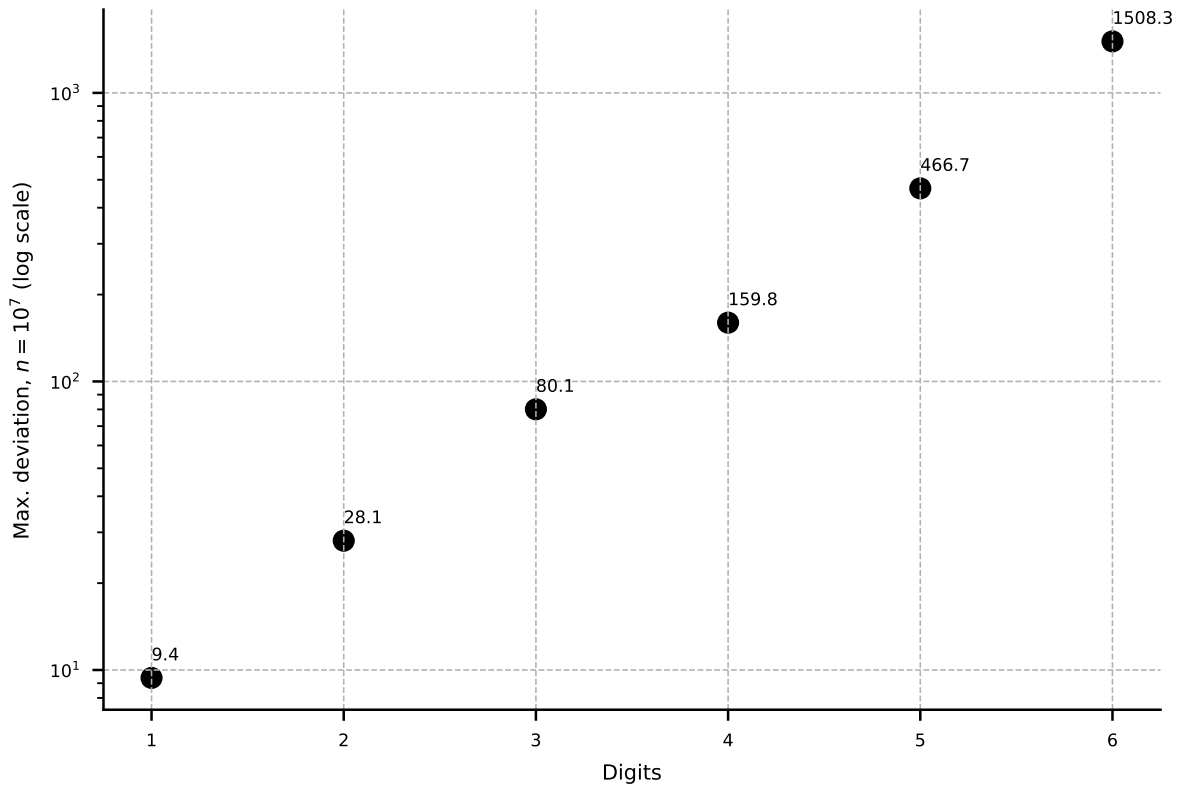


Fig. 2. Maximum deviations of $B_2^{(s)}(A_1, Q)$ and $B_3^{(s)}(A_1, Q)$ for $Q = 10^7$ as s changes from 1 to 6

Fig. 2 suggests that the asymptotic growth of the respective deviation is proportional to e^s and not 10^{s-1} , as expected from considering the 10^{s-1} intervals.

4. Conclusion

The problem of quantitative characterization of statistical properties of the first digits of integer powers has a long history. We have established the connection between this problem and Diophantine properties of logarithms, and obtained estimates for the remainder term in the asymptotic expression for the number of integer powers with specific first digits. We have also proposed numerous generalizations of this problem and provided state of the art solutions. Our results rely on known theorems establishing Diophantine properties of the respective real logarithms. However, these theorems are often very inexact, and improving them often requires solving difficult classical problems of Diophantine approximation. In certain cases, metric approach allows researchers to circumvent this obstacle by proving the desired Diophantine properties for a subset of a box $T \subset \mathbb{R}^n$ of sufficiently large Lebesgue measure [12].

5. Topical problems in the theory of Diophantine approximation

The results presented in the paper show how Diophantine properties of numbers can have surprising consequences in other areas of mathematics. To close out the article, let us formulate several topical problems in Diophantine approximation.

Problem 1. Consider the well-known Dirichlet-type theorem on solutions of the inequality

$$|P(x)| < Q^{-n} \quad (16)$$

in polynomials $P(x)$, $\deg P \leq n$, $H(P) \leq Q$. Obtain bounds on the measure of sets $\sigma(P)$ such that inequality (16) holds for points in these sets.

Problem 2. Study inequality (16) for a) irreducible polynomials and b) reducible polynomials.

Problem 3. Generalize inequality (16) to the fields of a) complex numbers and b) p -adic numbers.

Problem 4. Consider the inequality

$$|D(P)| < 2^{2n-2-2v}, \quad v \geq 0, \quad (17)$$

where $D(P)$ is the discriminant of a polynomial P with integer coefficients of degree n and height $H(P) \leq Q$. Find upper and lower bounds for the number of such polynomials satisfying the inequality (17) in the fields of real and p -adic numbers.

The authors would like to thank Y. V. Prokhorov, Y. V. Nesterenko, A. Dubickas, F. Goetze and V. G. Safonov for a number of useful constructive comments.

This work was supported by the Institute of Mathematics of the National Academy of Sciences of Belarus within the framework of the state programme “Convergence–2020”.

References

1. Newcomb S. On the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 1881, vol. 4, pp. 39–40.
2. Benford F. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 1938, vol. 78, pp. 551–572.
3. Hürlimann W. Generalizing Benford's Law Using Power Laws: Application to Integer Sequences. *International Journal of Mathematics and Mathematical Sciences*, 2009, art. 970284, 10 pp. <https://doi.org/10.1155/2009/970284>.
4. Kuipers L., Niederreiter H. *Uniform Distribution of Sequences*. New York, Wiley, 1974, 390 p. (Pure and Applied Mathematics).
5. Vinogradov I. M. *Selected Papers*. Berlin, Springer-Verlag, 1985, 401 p.
6. Roth K. Rational approximations to algebraic numbers. *Mathematica*, 1955, vol. 2, pp. 1–20.
7. Khinchine A. Zur metrischen Theorie der diophantischen Approximationen. *Mathematische Annalen*, 1924, vol. 92, pp. 115–125.
8. Baker A., Wüstholz G. Linear forms in logarithms of algebraic numbers. *Journal für die reine und angewandte Mathematik*, 1993, vol. 442, pp. 19–62.

9. Hata M. Legendre type polynomials and irrationality measures. *Journal für die reine und angewandte Mathematik*, 1990, vol. 407, pp. 99–125.
10. Rukhadze E. A. A problem on the distribution of first digits. *Moscow University Mathematics Bulletin*, 1987, vol. 42, no. 1, pp. 25–29 (in Russian).
11. Baker A. *Transcendental Number Theory*. Cambridge, Cambridge University Press, 1990, 155 p.
12. Bernik V., Kleinbock D., Margulis G. A. Khinchine-type theorems on manifolds: convergence case for standard and multiplicative versions. *International Mathematics Research Notices*, 2001, vol. 9, pp. 453–486.