On the discrepancy of powers of random variables

Nicolas Chenavier*, Dominique Schneider †

Abstract

Let (d_n) be a sequence of positive numbers and let (X_n) be a sequence of positive independent random variables. We provide an upper bound for the deviation between the distribution of the mantissaes of $(X_n^{d_n})$ and the Benford's law. If d_n goes to infinity at a rate at most polynomial, this deviation converges a.s. to 0 as N goes to infinity.

Keywords: Benford's law; discrepancy; mantissa.
AMS 2010 Subject Classifications: 60B10 . 11K38

1 Introduction

A sequence of positive numbers (x_n) is said to satisfy the first digit phenomenon if

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{F(x_n)=k} = \log_{10}\left(1+\frac{1}{k}\right), k = 1,\ldots,9,$$

where $F(x_n)$ is the first digit of x_n, and where $\mathbb{1}_A$ denotes the indicator function of any subset A. Such a phenomenon was observed by Benford and Newcomb on real life numbers [1, 13]. It is extensively used in various domains, such as fraud detection [14], computer design [8] and image processing [17]. As an extension of the first digit phenomenon, the notion of Benford sequence is introduced as follows. Let μ_{10} be the measure on the interval $[1, 10)$ defined by

$$\mu_{10}([1, a)) = \log_{10} a, (1 \leq a < 10),$$

where $\log_{10} a$ denotes the logarithm in base 10 of a. Let $\mathcal{M}_{10}(x)$ be the mantissa in base 10 of a positive number x, i.e. $\mathcal{M}_{10}(x)$ is the unique number in $[1, 10)$ such that there exists an integer k satisfying $x = \mathcal{M}_{10}(x)10^k$. A set of numbers (x_n) is referred to as a Benford sequence if for any $1 \leq a < 10$, we have

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\mathcal{M}_{10}(x_n)\in[1,a)} = \mu_{10}([1, a)).$$

In particular, each Benford sequence satisfies the first digit phenomenon since $F(x) = k$ if and only if $\mathcal{M}_{10}(x) \in [k, k + 1)$, with $x > 0$, $k = 1,\ldots, 9$. For instance, the sequences (2^n), $(n!)$ and (n^n) are Benford. For various examples of sequences of positive numbers whose mantissae are (or approach to be) distributed with respect to μ_{10}, see e.g. [5, 6]. More recently, several authors have provided examples of sequences of random variables whose mantissa distribution

*Université Littoral Côte d'Opale, EA 2797, LMPA, 50 rue Ferdinand Buisson, F-62228 Calais, France. E-mail: nicolas.chenavier@univ-littoral.fr, *corresponding author*

†Université Littoral Côte d'Opale, EA 2797, LMPA, 50 rue Ferdinand Buisson, F-62228 Calais, France. E-mail: dominique.schneider@univ-littoral.fr

converges to μ_{10} [3, 10, 16] or whose the sequence of mantissae is almost surely distributed with respect to μ_{10}. For a wide panorama on Benford sequences, see the reference books [2, 12].

It is well known that a sequence (x_n) of positive numbers is Benford in base 10 if and only if the sequence of its fractional parts $(\{\log_{10} x_n\})$ is uniformly distributed in $[0,1)$. According to the Weyl's criterion (see e.g. [9], p7), the sequence (x_n) is Benford if and only if, for any $h \in \mathbf{Z}^*$, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2i\pi h \log_{10} x_n} = 0.$$

To define a deviation between a sequence and the Benford's law, the notion of discrepancy is introduced as follows. Let $u = (u_n)$ be a sequence of real numbers. The discrepancy modulo 1 of order N of u, associated with the natural density, is defined as

$$D_N(u) = \sup_{0\le a<b<1} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{[a,b)} \left(\{u_n\}\right) - (b-a) \right|.$$

For more details on the discrepancy, see e.g. [9], p100–131. For a sequence $x = (x_n)$, if we set $x_n = 10^{u_n}$, we write $\widetilde{D}_N(x) = D_N(u)$. The quantity $\widetilde{D}_N(x)$ deals with the deviation between μ_{10} and the distribution of the first N terms of $(\mathcal{M}_{10}(x_n))$ since $\{\log_{10} x_n\} = \log_{10}(\mathcal{M}_{10}(x_n))$. Hence

$$\widetilde{D}_N(x) = \sup_{1\le s<t<10} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{[s,t)} \left(\mathcal{M}_{10}(x_n)\right) - \mu_{10}([s,t)) \right|.$$

In particular, $x = (x_n)$ is Benford if and only if $\widetilde{D}_N(x)$ converges to 0 as N goes to infinity. Through misuse of language, we also say that $\widetilde{D}_N(x)$ is the discrepancy of $x = (x_n)$.

In this paper, we consider the following problem. Let (X_n) be a sequence of positive independent random variables. We say that (X_n) is a.s. Benford if $\omega - \mathbb{P}\,a.s.$ the sequence $(X_n(\omega))$ is Benford. As observed in [7], several deterministic sequences at a power d tend to be Benford when the power d is large enough. The aim of our paper is to provide general conditions on the distribution of the random sequence $X = (X_n)$ to ensure that $X^{(d)} = (X_n^{d_n})$ is a.s. Benford for any sequence of positive numbers (d_n) such that d_n converges to infinity at a rate at most polynomial.

First, we give some notation. In what follows, the function log denotes the natural logarithm. For any functions f, g, we write $g(x) \underset{x\to\infty}{\sim} f(x)$ if and only if $\frac{g(x)}{f(x)} \underset{x\to\infty}{\longrightarrow} 1$. Moreover, we write $g(x) = O(f(x))$ if and only if there exists a positive number M and a real number x_0 such that $|g(x)| \le M|f(x)|$ for any $x \ge x_0$.

We are now prepared to state our first theorem, which provides an upper bound for the discrepancy.

Theorem 1. *Let (d_n) be a (deterministic) sequence of positive numbers such that $d_n = O\left(n^\theta\right)$ for some $\theta \ge 0$. Let $X = (X_n)$ be a sequence of positive independent random variables satisfying the following two conditions:*

(i) there exists $\alpha > 0$ such that $\sum_{n=1}^{\infty} \mathbb{P}\left(|\log X_n| > n^\alpha\right) < \infty$;

(ii) there exists a sequence of nonnegative numbers (r_n), with $r_n = O(n^{-\beta})$ for some $\beta > 0$, and their exist four constants $c_1, c_2, \gamma, \delta > 0$, such that for n large enough and for each $h \in \mathbf{N}^*$, we have

$$\left| \mathbb{E}\left[e^{2i\pi h \log X_n} \right] \right| \le c_1 h^{-\gamma} + c_2 h^{\delta} r_n. \tag{1}$$

Then there exist an integrable random variable C_0 and a constant c_0 such that, for any $N \geq 1$, we have $\omega - \mathbb{P}$ a.s.

$$\widetilde{D}_N(X^{(d)}(\omega)) \leq C_0(\omega) \cdot (\log N)^2 \cdot N^{-\frac{1}{2}} + c_0 \left(\frac{1}{N} \sum_{n=1}^{N} (d_n)^{-\gamma} + (\log N)^{\frac{1}{\delta+1}} \cdot N^{-\frac{\min\{\beta-\delta\theta,1\}}{\delta+1}} \right),$$

where $X^{(d)}(\omega) = (X_n^{d_n}(\omega))$.

The above theorem is obvious if the upper bound does not converge to 0. However, if $\delta\theta < \beta$, it provides a non-trivial estimate for the discrepancy when d_n goes to infinity at a rate at most polynomial. As a consequence, we obtain the following result.

Corollary 2. *Let (d_n) be such that $d_n = O\left(n^\theta\right)$ for some $\theta > 0$ and $d_n \underset{n\to\infty}{\longrightarrow} \infty$. Assume that $X = (X_n)$ satisfies the assumptions (i) and (ii) for some $\alpha, \beta, \gamma, \delta > 0$, with $\delta\theta < \beta$. Then $\widetilde{D}_N(X^d(\omega))$ converges $\omega - \mathbb{P}$ a.s. to 0, at a rate of convergence provided in Theorem 1. In particular, the sequence $(X_n^d(\omega))$ is a.s. Benford.*

In particular, if $X = (X_n)$ and (d_n) satisfy the assumptions of Corollary 2, with the more restrictive condition $d_n = O(n^\sigma)$ for each $\sigma > 0$, then the discrepancy of $X^{(d)}(\omega)$ can be bounded as follows:

$$\sup_{1\leq s<t<10} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{[s,t)} \left(\mathcal{M}_{10}(X_n^{d_n}(\omega)) \right) - \mu_{10}([s,t)) \right| \leq C(\omega) \cdot \frac{1}{N} \sum_{n=1}^{N} d_n^{-\gamma}.$$

It is rather surprising that $X^{(d)}(\omega)$ is a.s. Benford for a sequence $d = (d_n)$ which converges arbitrarily slowly to infinity. On the opposite, it appears that for several classes of (deterministic) sequences (x_n), the sequence $(x_n^{d_n})$ is Benford, when (d_n) converges to infinity at a rate *at less* polynomial (see e.g. Theorem 2 in [11]). As a second consequence of Theorem 1, the following corollary deals with the case where the sequence (d_n) is constant.

Corollary 3. *Let $d_n = d$ for each $n \geq 1$ and let $X = (X_n)$ be such that the assumptions (i) and (ii) hold for some $\alpha, \beta, \gamma, \delta > 0$. Then there exist an integrable random variable $C_0(\omega)$ and a constant c_0 such that, for any $N \geq 1$, we have $\omega - \mathbb{P}$ a.s.*

$$\widetilde{D}_N(X^d(\omega)) \leq C_0(\omega) \cdot (\log N)^2 \cdot N^{-\frac{1}{2}} + c_0 \left(d^{-\gamma} + (\log N)^{\frac{1}{\delta+1}} \cdot N^{-\frac{\min\{\beta,1\}}{\delta+1}} \right),$$

where $X^d(\omega) = (X_n^d(\omega))$.

In particular, as d goes to infinity, the sequence $X^d = (X_n^d)$ tends to be a.s. Benford in the sense that its discrepancy converges to 0 as $d, N \to \infty$. In a different context, such a convergence was already observed in Theorem 1 in [7], in which it is stated that two (deterministic) sequences at a large power tend to be Benford.

The assumption (i) of Theorem 1 is few restrictive. Indeed, thanks to the Markov's inequality, such a condition is satisfied when $\mathbb{E}[X_n]$ and $\mathbb{E}[X_n^{-1}]$ are negligible compared to $n^{-1-\epsilon}e^{n^\alpha}$ for some $\alpha, \epsilon > 0$. The assumption (ii) of Theorem 1 is in a way classical and is discussed in Remark 1.

Our paper is organized as follows. In Section 2, we prove Theorem 1. This result is illustrated through several examples of standard distributions in Section 3. These examples deal with discrete and continuous random variables respectively. In the rest of the paper, we denote by c a generic constant which is independent of ω, N and (d_n), but which may depend on other quantities.

2 Proof of Theorem 1

To prove Theorem 1, we apply two well-known inequalities. The first one deals with the discrepancy and is referred to as the Erdös-Turán inequality (see e.g. [**?**]).

Theorem 4. *(Erdös-Turán inequality) Let $x = (x_n)$ be a sequence of real numbers and let $N \geq 1$. Then, for every integer $H \geq 1$, we have*

$$\widetilde{D}_N(x) \leq \frac{1}{H+1} + \sum_{h=1}^{H} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2i\pi h \log_{10} x_n} \right|.$$

The second inequality which we apply gives a deviation beween a sum of unit random complex numbers and the expectation of this sum. Such a result is due to Cohen and Cuny (Theorem 4.10 in [4]) and is re-written in our context.

Theorem 5. *(Cohen & Cuny, 2006) Let (Y_n) be a sequence of independent random variables, with values in $\mathbf{R}$. Assume that there exists $\eta > 0$, such that $\sum_{n=1}^{\infty} \mathbb{P}(|Y_n| > n^{\eta}) < \infty$. Let (a_n) be a sequence of complex numbers. Then there exist universal constants $\epsilon > 0$ and $C > 0$, such that*

$$\mathbb{E}\left[\sup_{N>K\geq 1} \sup_{T\geq 1} \exp\left(\epsilon \cdot \frac{\max_{|t|\leq T} \left| \sum_{n=K+1}^{N} a_n \left(e^{2i\pi t Y_n} - \mathbb{E}\left[e^{2i\pi t Y_n} \right] \right) \right|^2}{\log(1+T)\log(1+N^{\eta}) \sum_{n=K+1}^{N} |a_n|^2} \right) \right] \leq C.$$

In the rest of the paper, with a slight abuse of notation, we omit the dependence in ω, e.g. we write $\widetilde{D}_N(X^{(d)})$ instead of $\widetilde{D}_N(X^{(d)}(\omega))$. We are now prepared to prove our first theorem.

Proof of Theorem 1. According to the Erdös-Turán inequality, we have for any $H \geq 1$,

$$\widetilde{D}_N(X^{(d)}) \leq \frac{1}{H+1} + \sum_{h=1}^{H} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2i\pi h \log_{10} X_n^{d_n}} \right|.$$

Hence,

$$\begin{aligned} \widetilde{D}_N(X^{(d)}) \leq \frac{1}{H+1} &+ \sum_{h=1}^{H} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[e^{2i\pi h \log_{10} X_n^{d_n}} \right] \right| \\ &+ \sum_{h=1}^{H} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} \left(e^{2i\pi h \log_{10} X_n^{d_n}} - \mathbb{E}\left[e^{2i\pi h \log_{10} X_n^{d_n}} \right] \right) \right|. \quad (2) \end{aligned}$$

First, we provide an upper bound for the term on the bottom. To do it, we take $a_n = 1$, $Y_n = \log_{10} X_n^{d_n}$ and $K = 1$. Since $d_n = O(n^{\theta})$, we obtain for n large enough that $\mathbb{P}(|Y_n| > n^{\eta}) \leq \mathbb{P}(|\log X_n| > n^{\alpha})$ with $\eta > \alpha + \theta$. Hence, according to the assumption (i), we have $\sum_{n=1}^{\infty} \mathbb{P}(|Y_n| > n^{\eta}) < \infty$. It follows from Theorem 5 that

$$\mathbb{E}\left[\sup_{N>1} \sup_{T\geq 1} \max_{|t|\leq T} \frac{\left| \sum_{n=2}^{N} \left(e^{2i\pi t \log_{10} X_n^{d_n}} - \mathbb{E}\left[e^{2i\pi t \log_{10} X_n^{d_n}} \right] \right) \right|^2}{\log(1+T)\log(1+N^{\eta})(N-1)} \right] \leq C.$$

In particular, there exists an integrable random variable $c(\omega)$ such that, for any $N \geq 2$, $T \geq 1$, $|t| \leq T$ we have $\omega - \mathbb{P}\, a.s.$

$$\left| \frac{1}{N} \sum_{n=1}^{N} \left(e^{2i\pi t \log_{10} X_n^{d_n}} - \mathbb{E}\left[e^{2i\pi t \log_{10} X_n^{d_n}} \right] \right) \right| \leq c(\omega) \cdot \sqrt{\log(1+T)} \cdot \sqrt{\frac{\log(1+N^{\eta})}{N}}.$$

Notice that we have considered a sum over $n = 1, \dots, N$ and not over $n = 2, \dots, N$ in the above equation because $\left| e^{2i\pi t \log_{10} X_1} - \mathbb{E}\left[e^{2i\pi t \log_{10} X_1} \right] \right| \leq 2$. By taking $T = H$ and $t = h$, we obtain for any $N \geq 1, H \geq 1$ that

$$\begin{aligned} \sum_{h=1}^{H} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} \left(e^{2i\pi h \log_{10} X_n^{d_n}} - \mathbb{E}\left[e^{2i\pi h \log_{10} X_n^{d_n}} \right] \right) \right| & \\ \leq c(\omega) \sum_{h=1}^{H} \frac{1}{h} \sqrt{\log(1+H)} \cdot \sqrt{\frac{\log(1+N^{\eta})}{N}} & \\ \leq c'(\omega) \log H \sqrt{\log(1+H)} \cdot \sqrt{\frac{\log(1+N^{\eta})}{N}}. & \end{aligned} \tag{3}$$

Secondly, we provide an upper bound for the second term in the right-hand side in (2). To do it, let N_0 be such that the inequality (1) holds for each $N \geq N_0$. Then

$$\left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[e^{2i\pi h \log_{10} X_n^{d_n}} \right] \right| \leq \frac{1}{N} \sum_{n=1}^{N_0} \left| \mathbb{E}\left[e^{2i\pi h d_n \log_{10} X_n} \right] \right| + \frac{1}{N} \sum_{n=N_0+1}^{N} \left| \mathbb{E}\left[e^{2i\pi h d_n \log_{10} X_n} \right] \right|.$$

Bounding $\left| \mathbb{E}\left[e^{2i\pi h d_n \log_{10} X_n} \right] \right|$ by 1 in the first sum and applying the inequality (1) in the second sum for the right-hand side, we get

$$\left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[e^{2i\pi h \log_{10} X_n^{d_n}} \right] \right| \leq \frac{N_0}{N} + c_1 \cdot \frac{1}{N} \sum_{n=1}^{N} \left(\frac{h d_n}{\log(10)} \right)^{-\gamma} + c_2 \cdot \frac{1}{N} \sum_{n=1}^{N} \left(\frac{h d_n}{\log(10)} \right)^{\delta} r_n.$$

Besides, $\sum_{h=1}^{H} \frac{1}{h} \leq c \log H$, $\sum_{h=1}^{H} \frac{1}{h^{1+\gamma}} \leq c$ and $\sum_{h=1}^{H} \frac{1}{h^{1-\delta}} \leq c H^{\delta}$. This implies that

$$\sum_{h=1}^{H} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[e^{2i\pi h \log_{10} X_n^{d_n}} \right] \right| \leq c \cdot \left(\frac{\log H}{N} + \frac{1}{N} \sum_{n=1}^{N} (d_n)^{-\gamma} + \frac{1}{N} \sum_{n=1}^{N} (d_n)^{\delta} r_n \cdot H^{\delta} \right).$$

Since $d_n = O\left(n^{\theta}\right)$ and $r_n = O\left(n^{-\beta}\right)$, we have $\frac{1}{N} \sum_{n=1}^{N} (d_n)^{\delta} r_n \leq c \cdot \log N \cdot N^{-1}$ if $\beta - \delta\theta = 1$ and $\frac{1}{N} \sum_{n=1}^{N} (d_n)^{\delta} r_n \leq c \cdot N^{-\min\{\beta - \delta\theta, 1\}}$ otherwise. This together with (2) and (3) implies that

$$\begin{aligned} \widetilde{D}_N(X^{(d)}) \leq \frac{1}{H+1} + c''(\omega) \cdot \log H \cdot \sqrt{\log(1+H)} \cdot \sqrt{\frac{\log(1+N^{\eta})}{N}} & \\ + c \cdot \left(\frac{1}{N} \sum_{n=1}^{N} (d_n)^{-\gamma} + \log N \cdot N^{-\min\{\beta - \delta\theta, 1\}} \cdot H^{\delta} \right). & \end{aligned}$$

Optimizing the right-hand side over $H \geq 1$, we conclude the proof of Theorem 1 by taking

$$H = \left\lfloor (\log N)^{-\frac{1}{\delta+1}} \cdot N^{\frac{\min\{\beta - \delta\theta, 1\}}{\delta+1}} \right\rfloor + 1.$$

□

Remark 1. The assumption given in Equation (1) has been chosen in such a way that it holds when X_n follows the (discrete) uniform distribution on $\{1, \dots, n\}$. Indeed, in this case, we have

$$\left| \mathbb{E}\left[e^{2i\pi h \log X_n} \right] \right| = \left| \frac{1}{n} \sum_{k=1}^{n} e^{2i\pi h \log k} \right| \leq \frac{1}{\sqrt{n}} + \frac{1}{n} \left| \sum_{k=\lfloor \sqrt{n} \rfloor + 1}^{n} e^{2i\pi h \log k} \right|,$$

According to the Van der Corput's theorem (see e.g. [9], p17), this shows that

$$\left|\mathbb{E}\left[e^{2i\pi h \log X_n}\right]\right| \leq \frac{8}{\sqrt{h}} + \frac{1+4\sqrt{h}}{\sqrt{n}} + \frac{6}{n} + \frac{3h}{n\sqrt{n}}.$$

In particular, this satisfies Equation (1) with $\gamma = \frac{1}{2}$, $\delta = 1$ and $r_n = \frac{1}{\sqrt{n}}$. However, our assumption (ii) and our assumption on the independence of the random variables X_n remain restrictive. We hope, in a future paper, to extent Theorem 1 with more general conditions.

Remark 2. The main tool to derive the rate of the discrepancy is contained in Theorem 5. Besides, as a consequence of Corollary 3, we deduce that $\omega - \mathbb{P}\ a.s.$

$$\lim_{d\to\infty} \limsup_{N\to\infty} \widetilde{D}_N(X^d) = 0. \tag{4}$$

In particular, when d is large, the sequence $X^d = (X_n^d)$ tends to be a Benford sequence. However, Theorem 5 is not necessary to derive Equation (4) because the latter can be proved directly by standard arguments. Indeed, it follows from the law of large numbers (for independent non-stationary random variables) and the Erdös-Turán inequality that for all fixed $H \geq 1$,

$$\limsup_{N\to\infty} \widetilde{D}_N(X^d) \leq \frac{1}{H+1} + \sum_{h=1}^{H} \frac{1}{h} \limsup_{N\to\infty} \frac{1}{N} \left|\sum_{n=1}^{N} \mathbb{E}\left[e^{2i\pi h d \log_{10} X_n}\right]\right|.$$

Besides, according to (1), we know that

$$\lim_{d\to\infty} \limsup_{N\to\infty} \frac{1}{N} \left|\sum_{n=1}^{N} \mathbb{E}\left[e^{2i\pi h d \log_{10} X_n}\right]\right| = 0.$$

Hence, by taking $H \to \infty$, this proves that $\lim_{d\to\infty} \limsup_{N\to\infty} \widetilde{D}_N(X^d) = 0$. However, the main contribution of our paper is to provide an explicit rate of convergence for the discrepancy of X^d as d goes to infinity.

3 Examples

In this section, we give several examples of sequences of random variables satisfying the assumptions (i) and (ii) of Theorem 1. Our examples deal with discrete and continuous random variables respectively.

3.1 Discrete random variables

The following proposition provides sufficient conditions for discrete random variables to ensure that the assumption (ii) of Theorem 1 is satisfied for $\gamma = \delta = 1$.

Proposition 6. *Let (X_n) be a sequence of random variables with finite expectation and such that $X_n \geq 1$ a.s.. Assume that there exists a sequence of modes (m_n) such that the sequences $(\mathbb{P}(X_n = k))_{k\leq m_n}$ and $(\mathbb{P}(X_n = k))_{k>m_n}$ are non-decreasing and non-increasing respectively. Moreover, assume that for some $\beta > 0$ one of the two following cases is satisfied:*

- ***Case 1:*** *$m_n \cdot n^{-\beta} \underset{n\to\infty}{\longrightarrow} \infty$ and $\sup_{n\geq 1} m_n \mathbb{P}(X_n = m_n) < \infty$;*
- ***Case 2:*** *$\sup_{n\geq 1} m_n < \infty$, $\mathbb{P}(X_n = m_n) = O\left(n^{-\beta}\right)$ and $\mathbb{E}\left[\frac{1}{X_n}\right] = O\left(n^{-\beta}\right)$.*

Then for n large enough and for each $h \geq 1$, we have:

$$\left|\mathbb{E}\left[\, e^{2i\pi h \log X_n}\,\right]\right| \leq c_1 h^{-1} + c_2 h n^{-\beta}$$

where c_1, c_2 are two constants.

Proof of Proposition 6. First, we provide a generic upper bound for $\mathbb{E}\left[\, e^{2i\pi h \log X_n}\,\right]$ which is independent of the two above cases. Then we deduce a specific upper bound for this expectation which depends this time on the case which is considered.

To do it, we write $\mathbb{E}\left[\, e^{2i\pi h \log X_n}\,\right] = \lim_{N\to\infty} \sum_{k=1}^{N} e^{2i\pi h \log k}\, \mathbb{P}\left(X_n = k\right)$. Let $N \geq 1$ be fixed. It follows from the Abel transformation that

$$\begin{aligned}\sum_{k=1}^{N} e^{2i\pi h \log k}\, \mathbb{P}\left(X_n = k\right) = \mathbb{P}\left(X_n = N+1\right) \sum_{j=1}^{N} e^{2i\pi h \log j} \\ - \sum_{k=1}^{N}\sum_{j=1}^{k} e^{2i\pi h \log j}\left(\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right).\end{aligned}$$

Since $\left|\mathbb{P}\left(X_n = N+1\right)\sum_{j=1}^{N} e^{2i\pi h \log j}\right| \leq N\,\mathbb{P}\left(X_n = N+1\right)$ converges to 0 as N goes to infinity (because $\mathbb{E}\left[\, X_n\,\right] < \infty$), it is enough prove that

$$\left|\sum_{k=1}^{N}\sum_{j=1}^{k} e^{2i\pi h \log j}\left(\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right)\right| \leq \frac{c_1}{h} + h c_2 n^{-\beta},$$

for some constants c_1, c_2. To do it, we apply the following lemma.

Lemma 7. *For each $h \geq 1$, $k \geq 1$, we have*

$$\left|\sum_{j=1}^{k} e^{2i\pi h \log j}\right| \leq \frac{k}{2\pi h} + 1 + \pi h \log k.$$

Proof of Lemma 7. First, we notice that

$$\sum_{j=1}^{k} e^{2i\pi h \log j} = k^{2i\pi h+1} R_k(f),$$

where $R_k(f) := \sum_{j=0}^{k-1} \int_{\frac{j}{k}}^{\frac{j+1}{k}} f\left(\frac{j+1}{k}\right) \mathrm{d}t$ is the Riemann sum of the function $f : t \mapsto t^{2i\pi h}$ on $[0,1]$ with n regular steps of length n^{-1}. Hence

$$\begin{aligned}\left|\sum_{j=1}^{k} e^{2i\pi h \log j}\right| &\leq k\left|\int_0^1 f(t)\mathrm{d}t\right| + k\left|R_k(f) - \int_0^1 f(t)\mathrm{d}t\right| \\ &\leq \frac{k}{2\pi h} + k\left|R_k(f) - \int_0^1 f(t)\mathrm{d}t\right|,\end{aligned}$$

where the second inequality comes from the fact that $\int_0^1 f(t)\mathrm{d}t = \frac{1}{2i\pi h+1}$. Besides,

$$\begin{aligned}\left|R_k(f) - \int_0^1 f(t)\mathrm{d}t\right| &= \left|\sum_{j=0}^{k-1}\int_{\frac{j}{k}}^{\frac{j+1}{k}} \left(f\left(\frac{j+1}{k}\right) - f(t)\right)\mathrm{d}t\right| \\ &\leq \left|\int_0^{\frac{1}{k}} \left(f\left(\frac{1}{k}\right) - f(t)\right)\mathrm{d}t\right| + \sum_{j=1}^{k-1}\int_{\frac{j}{k}}^{\frac{j+1}{k}} \left(\frac{j+1}{k} - t\right)\cdot\frac{2\pi h k}{j}\mathrm{d}t,\end{aligned}$$

where the last line is a consequence of the mean value inequality. Integrating the right-hand side over t, we get

$$\left|R_k(f) - \int_0^1 f(t)\mathrm{d}t\right| \leq \frac{1}{k} + 2\pi h \sum_{j=1}^{k-1} \frac{1}{2jk} \leq \frac{1}{k} + \pi h \cdot \frac{\log k}{k}.$$

This concludes the proof of Lemma 7. □

According to Lemma 7, we have

$$\left|\sum_{k=1}^{N}\sum_{j=1}^{k} e^{2i\pi h \log j}\left(\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right)\right| \leq \sum_{k=1}^{N}\left(\frac{k}{2\pi h} + 1 + \pi h \log k\right)\left|\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right|.$$

Since the sequences $\left(\mathbb{P}\left(X_n = k\right)\right)_{k\leq m_n}$ and $\left(\mathbb{P}\left(X_n = k\right)\right)_{k\geq m_n}$ are non-decreasing and non-increasing respectively, we get

$$\begin{aligned}
&\sum_{k=1}^{N}\left(\frac{k}{2\pi h} + 1 + \pi h \log k\right)\left|\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right| \\
&= \sum_{k=1}^{m_n-1}\left(\frac{k}{2\pi h} + \pi h \log k\right)\left(\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right) \\
&+ \sum_{k=m_n}^{N}\left(\frac{k}{2\pi h} + \pi h \log k\right)\left(\mathbb{P}\left(X_n = k\right) - \mathbb{P}\left(X_n = k+1\right)\right) \\
&\qquad + 2\,\mathbb{P}\left(X_n = m_n\right) - \mathbb{P}\left(X_n = N\right) - \mathbb{P}\left(X_n = 1\right).
\end{aligned}$$

With standard computations, we get:

$$\sum_{k=1}^{m_n-1} k\left(\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right) \leq m_n\,\mathbb{P}\left(X_n = m_n\right),$$

$$\sum_{k=1}^{m_n-1} \log k\left(\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right) \leq \log m_n\,\mathbb{P}\left(X_n = m_n\right),$$

$$\sum_{k=m_n}^{N} k\left(\mathbb{P}\left(X_n = k\right) - \mathbb{P}\left(X_n = k+1\right)\right) \leq m_n\,\mathbb{P}\left(X_n = m_n\right) + 1,$$

$$\begin{aligned}
\sum_{k=m_n}^{N} \log k\left(\mathbb{P}\left(X_n = k\right) - \mathbb{P}\left(X_n = k+1\right)\right) &\leq \sum_{k=m_n}^{N-2} \log\left(1 + \frac{1}{k}\right)\mathbb{P}\left(X_n = k+1\right) \\
&\quad + \log m_n\,\mathbb{P}\left(X_n = m_n\right).
\end{aligned}$$

Using the fact that $\log\left(1+\frac{1}{k}\right)\mathbb{P}\left(X_n = k+1\right) \leq \frac{1}{k}\,\mathbb{P}\left(X_n = k\right)$ for each $k \geq m_n$, we deduce that

$$\sum_{k=1}^{N}\left(\frac{k}{2\pi h} + 1 + \pi h \log k\right)\left|\mathbb{P}\left(X_n = k+1\right) - \mathbb{P}\left(X_n = k\right)\right| \leq \frac{c_1}{h} + \pi h s_n, \tag{6}$$

where

$$c_1 = \frac{1}{2\pi}\left(2\sup_{n\geq 1} m_n\,\mathbb{P}\left(X_n = m_n\right) + 1\right)$$

and

$$s_n = 2\log m_n\,\mathbb{P}\left(X_n = m_n\right) + \sum_{k=m_n}^{N-1}\frac{1}{k}\,\mathbb{P}\left(X_n = k\right) + 2\,\mathbb{P}\left(X_n = m_n\right).$$

The inequality (6) is independent of the two cases considered in the assumptions of Proposition 6. Now, we deal with the terms c_1 and s_n by discussing these two cases.

- Case 1: if $m_n \cdot n^{-\beta} \underset{n\to\infty}{\longrightarrow} \infty$ for some $\beta > 0$ and $\sup_{n\geq 1} m_n\,\mathbb{P}\left(X_n = m_n\right) < \infty$, we obtain that $c_1 < \infty$. Moreover, $s_n = O\left(n^{-\beta}\right)$ since $\log m_n = O(m_n)$ and

$$\sum_{k=m_n}^{\infty}\frac{1}{k}\,\mathbb{P}\left(X_n = k\right) \leq \sum_{k=m_n}^{\infty}\frac{1}{k}\,\mathbb{P}\left(X_n = m_n\right) \underset{n\to\infty}{\sim} \log m_n \cdot \mathbb{P}\left(X_n = m_n\right).$$

- Case 2: if $\sup_{n\geq 1} m_n < \infty$, $\mathbb{P}\left(X_n = m_n\right) = O\left(n^{-\beta}\right)$ and $\mathbb{E}\left[\frac{1}{X_n}\right] = O\left(n^{-\beta}\right)$ for some $\beta > 0$, we also obtain that $c_1 < \infty$ and $s_n = O\left(n^{-\beta}\right)$.

This concludes the proof of Proposition 6. □

We give below three examples of sequences of random variables $X = (X_n)$ by checking the assumption (i) of Theorem 1 and one of the two cases of Proposition 6. According to Theorem 1 and Proposition 6, the discrepancy for each example can be bounded as follows:

$$\widetilde{D}_N(X^{(d)}) \leq C_0(\omega)\cdot(\log N)^2 \cdot N^{-\frac{1}{2}} + c_0\left(\frac{1}{N}\sum_{n=1}^{N}(d_n)^{-1} + (\log N)^{\frac{1}{2}}\cdot N^{-\frac{1}{2}\cdot\min\{\beta-\theta,1\}}\right).$$

In particular, if $(d_n) \to \infty$ with $d_n = O\left(n^{\theta}\right)$ and $\theta > \beta$, the sequence $X^{(d)} = (X_n^{d_n})$ is a.s. Benford.

Example 1. Assume that X_n has a geometric distribution with parameter $p_n = O\left(n^{-\beta}\right)$. Here $m_n = 1$, so that $\mathbb{P}\left(X_n = 1\right) = p_n = O\left(n^{-\beta}\right)$. We also obtain the same order for $\mathbb{E}\left[\frac{1}{X_n}\right] = -\frac{p_n}{1-p_n}\cdot\log(1-p_n)$. In particular, the third conditions of Case 2 are satisfied. Besides, if $p_n e^{n^{\alpha}} n^{-\alpha'} \underset{n\to\infty}{\longrightarrow} \infty$ for some $\alpha > 0, \alpha' > 1$, the assumption (i) holds since

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|\log X_n| > n^{\alpha}\right) \leq \sum_{n=1}^{\infty}\frac{1}{p_n e^{n^{\alpha}}} < \infty$$

according to the Markov's inequality.

Example 2. Let X_n be a random variable with distribution $\mathbb{P}\left(X_n = k\right) = \frac{\alpha_n}{(n+k)^{1+\epsilon}}$, where α_n is the normalizing constant and $\epsilon > 0$. In particular, we have

$$\epsilon n^{\epsilon} \leq \alpha_n \leq \epsilon(n+1)^{\epsilon} \tag{7}$$

since

$$\frac{1}{\int_n^\infty x^{-(1+\epsilon)}\mathrm{d}x} \leq \alpha_n := \frac{1}{\sum_{k=1}^\infty (n+k)^{-(1+\epsilon)}} \leq \frac{1}{\int_{n+1}^\infty x^{-(1+\epsilon)}\mathrm{d}x}.$$

Here $m_n = 1$ and the third conditions of Case 2 are satisfied. Indeed, the first one is trivial and for the second one we have $\mathbb{P}(X_n = 1) = O\left(n^{-(1+\epsilon)}\right)$. For the third condition, let $\beta < 1$. According to (7), we have $\frac{1}{k} \cdot \frac{\alpha_n \cdot n^\beta}{(n+k)^{1+\epsilon}} \leq \frac{\epsilon}{k(k+1)^{1-\beta}}$. It follows from the dominated convergence theorem that

$$\lim_{n\to\infty} n^\beta \cdot \mathbb{E}\left[\frac{1}{X_n}\right] = \sum_{k=1}^\infty \lim_{n\to\infty} \frac{1}{k} \cdot \frac{\alpha_n \cdot n^\beta}{(n+k)^{1+\epsilon}} = 0.$$

This checks the third condition of Case 2 for each $\beta < 1$. Besides, the assumption (i) holds since for each $n \geq 1$ and for each $\alpha > 0$, we have

$$\mathbb{P}(|\log X_n| > n^\alpha) = \sum_{k=\lfloor e^{n^\alpha}+1\rfloor}^\infty \frac{\alpha_n}{(n+k)^{1+\epsilon}} \leq \sum_{k=\lfloor e^{n^\alpha}+1\rfloor}^\infty \frac{\epsilon(n+1)^\epsilon}{(n+k)^{1+\epsilon}} \underset{n\to\infty}{\sim} \frac{n^\epsilon}{e^{\epsilon n^\alpha}}.$$

Example 3. Assume that X_n has a (discrete) uniform distribution in $\{a_n, \ldots, b_n\}$, with $a_n < b_n$, $b_n \cdot n^{-\beta} \to \infty$ for some $\beta > 0$, and $\limsup \frac{a_n}{b_n} < 1$. Here we take $m_n = b_n$. The two conditions of Case 1 are satisfied. Indeed, the first one holds because $b_n \cdot n^{-\beta} \to \infty$. The second one comes from the fact that $\limsup \frac{a_n}{b_n} < 1$ and $m_n\, \mathbb{P}(X_n = m_n) = \frac{b_n}{b_n - a_n + 1}$. Besides, a sufficient and few restrictive assumption on b_n to ensure that the assumption (i) holds is: $b_n = O(e^{n^\alpha})$ for some $\alpha > 0$. Notice that if $\frac{a_n}{b_n}$ converges to 1, the random variables X_n are asymptotically deterministic. It is not surprising that the property (b) cannot hold in this context since there exist deterministic sequences such that, at any power d, the sequences are not Benford.

3.2 Continuous random variables

Let $X = (X_n)$ be a sequence of random variables. We first state three properties which imply the assumption (ii) of Theorem 1 when they are simultaneously satisfied.

(a) For any $n \geq 1$, the density f_n of X_n exists and is a piecewise absolutely continuous function. In what follows, we denote by k_n the number of sub-domains of f_n and by $I_{n,j} := [a_{n,j}, b_{n,j}]$ the j-th sub-domain, with $a_{n,j} \leq b_{n,j} \leq a_{n,j+1}$ for each $1 \leq j \leq k_n - 1$. The k_n-th interval is of the form $I_{n,k_n} = [a_{n,k_n}, +\infty)$. In particular, f_n is a.e. differentiable on $\bigcup_{j=1}^{k_n} I_{n,j}$ and $f_n = 0$ on the complement.

(b) $\limsup_{N\to\infty} \sum_{j=1}^{k_N} \sup_{x\in I_{N,j}} |x f_N(x)| < \infty$.

(c) $\limsup_{N\to\infty} \sum_{j=1}^{k_N} \int_{I_{N,j}} |x f_N'(x)| \mathrm{d}x < \infty$.

Under the above assumptions, the following proposition ensures that the assumption (ii) of Theorem 1 holds, with $\gamma = 1$ and $a_n = 0$ for each $n \geq 1$.

Proposition 8. *If the properties hold (a), (b) and (c) hold simultaneously, then for n large enough and for each $h \in \mathbf{N}^*$, we have* $\left|\mathbb{E}\left[e^{2i\pi h \log X_n}\right]\right| \leq c_1 h^{-1}$.

Proof of Proposition 8. It is enough to prove the following inequality:

$$\limsup_{N\to\infty} \sup_{h\in\mathbf{N}^*} h\left|\mathbb{E}\left[e^{2i\pi h\log X_N}\right]\right| < \infty.$$

To do it, we assume without loss of generality that $k_n = 1$ for each n, with $I_{n,j} =: I_n = [a_n, b_n]$. In particular, the density f_n is absolutely continuous on $[a_n, b_n]$ and equals 0 on the complement. This gives for any $N \geq 1, h \geq 1$

$$\begin{aligned} h\left|\mathbb{E}\left[e^{2i\pi h\log X_N}\right]\right| &= h\left|\int_{a_N}^{b_N} x^{2i\pi h} f_N(x)\mathrm{d}x\right| \\ &= h\left|\frac{1}{2i\pi h}\cdot\left(\left[x^{2i\pi h+1}f_N(x)\right]_{a_N}^{b_N} - \int_{a_N}^{b_N} x^{2i\pi h+1} f_N'(x)\mathrm{d}x\right)\right| \\ &\leq \frac{1}{2\pi}\left(\sup_{x\in[a_N,b_N]}|xf_N(x)| + \int_{a_N}^{b_N}|xf_N'(x)|\mathrm{d}x\right). \end{aligned}$$

In particular, we have $\limsup_{N\to\infty}\sup_{h\in\mathbf{N}^*} h\left|\mathbb{E}\left[e^{2i\pi h\log X_N}\right]\right| < \infty$ provided that the three above properties hold. □

Notice that if g_n denotes the density of X_n^{-1}, we can easily show that g_n satisfies the above assumptions if and only if the ones are satisfied by the density of X_n. This suggests that our assumptions are not very restrictive. We give below three examples of distributions of random variables which satisfy the assumption (i) of Theorem 1 and the three conditions (a), (b) and (c) of Proposition 8. According to Theorem 1 and Proposition 8, the discrepancy for each example can be bounded as follows:

$$\widetilde{D}_N(X^{(d)}) \leq C_0'(\omega)\cdot(\log N)^2\cdot N^{-\frac{1}{2}} + c_0'\cdot\frac{1}{N}\sum_{n=1}^{N}(d_n)^{-1}.$$

To obtain the rate of the discrepancy, we have taken $\delta = 1$ and $\beta \to \infty$. In particular, if $(d_n) \to \infty$ with $d_n = O\left(n^\theta\right)$ for some $\theta > 0$, the sequence $X^{(d)} = (X_n^{d_n})$ is a.s. Benford.

Example 4. If X_n has an exponential distribution with parameter $\lambda_n > 0$, the properties (a), (b) and (c) hold simultaneously, with $k_n = 1$. Indeed, the first one is trivially satisfied and for the second and the third ones, we get:

$$\sup_{x\in\mathbf{R}_+}|xf_n(x)| = e^{-1} \quad \text{and} \quad \int_{\mathbf{R}_+}|xf_n'(x)|\mathrm{d}x = 1.$$

Besides, for each $\alpha > 0$, we have

$$\mathbb{P}\left(|\log X_n| > n^\alpha\right) = e^{-\lambda_n e^{n^\alpha}} + (1 - e^{-\lambda_n e^{-n^\alpha}}).$$

Hence the assumption (i) is satisfied if there exists α' such that $\lambda_n e^{n^{\alpha'}} \underset{n\to\infty}{\longrightarrow} \infty$ and $\lambda_n e^{-n^{\alpha'}} \underset{n\to\infty}{\longrightarrow} 0$.

Example 5. Assume that X_n has a standard Fréchet distribution with parameter $\alpha_n > 0$, i.e. $\mathbb{P}\left(X_n \leq x\right) = e^{-x^{-\alpha_n}}$ if $x \geq 0$ and $\mathbb{P}\left(X_n \leq x\right) = 0$ otherwise. The property (a) holds. Moreover, if $\inf_{n\geq 1}\alpha_n > 0$ and $\sup_{n\geq 1}\alpha_n < \infty$, we can easily prove that the properties (b) and (c) are satisfied. Besides, the assumption (i) is also satisfied since for each $\alpha > 0$, we have

$$\mathbb{P}\left(|\log X_n| > n^\alpha\right) \underset{n\to\infty}{\sim} e^{-\alpha_n\cdot n^\alpha} + e^{-e^{\alpha_n\cdot n^\alpha}},$$

where the right-hand side is the term of a convergent series.

Example 6. If X_n has a (continuous) uniform distribution on $[a_n, b_n]$, with $a_n < b_n$, the properties (a) and (c) hold. Moreover, the property (b) is satisfied when $\limsup \frac{a_n}{b_n} < 1$. Besides, a sufficient and few restrictive assumption on a_n, b_n to ensure that the assumption (i) holds is: $e^{-n^\alpha} = O(a_n)$ and $b_n = O(e^{n^\alpha})$ for some $\alpha > 0$. Unsurprisingly, the assumptions on b_n are very similar to those considered for a (discrete) uniform distribution.

3.3 A numerical illustration

In this section, we give a numerical illustration of a sequence of independent random variables (X_n) such that (X_n^d) is almost a Benford sequence. For each n, the distribution of X_n is assumed to be the (continuous) uniform distribution on $[1, n]$. This sequence satisfies the assumptions of Theorem 1 (see Example 6). In Table 1, we provide the frequencies of the first significant digit of $X_1^d, \ldots, X_N^d$, with $N = 1000$ and $d = 2$. It appears that the distribution of frequencies of (X_n^d) is close to the Benford's law.

First digit	(X_n^d)	Benford's law
1	0.293	0.306
2	0.183	0.184
3	0.130	0.116
4	0.099	0.106
5	0.081	0.082
6	0.065	0.055
7	0.058	0.050
8	0.047	0.053
9	0.043	0.048

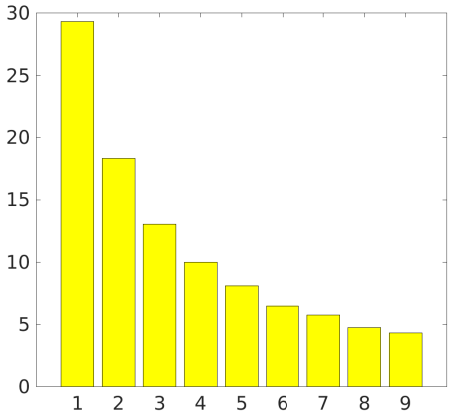

Table 1: a simulation of the frequencies of the first significant digits of $X_1^d, \ldots, X_N^d$, where X_n has a uniform distribution on $[1, n]$ for each $n \geq 1$, with $N = 1000$ and $d = 2$ (Scilab©).

References

[1] F. Benford. *The law of anomalous numbers. Proceedings of the American Philosophical Society*, (78): 551–572, 1938.

[2] A. Berger, T.P. Hill. An introduction to Benford's law. *Princeton University Press, Princeton, NJ*, 2015.

[3] N. Chenavier, B. Massé, and D. Schneider. Products of random variables and the first digit phenomenon, *available in https://arxiv.org/abs/1512.06049*, 2015

[4] G. Cohen, C. Cuny. On random almost periodic series and random ergodic theory. *Ergodic Theory Dynam. Systems*, (26): 683–709, 2006

[5] D. I. A. Cohen, T. M. Katz. Prime numbers and the first digit phenomenon. *J. Number Theory*, (18): 261–268, 1984

[6] P. Diaconis. The distribution of leading digits and uniform distribution mod 1. *Ann. Probability*, (5): 72–81, 1977

[7] D. Eliahou, B. Massé, D. Schneider. On the mantissa distribution of powers of natural and prime numbers. *Acta Math. Hungar.*, (139): 49–63, 2013

[8] R. W. Hamming. On the distribution of numbers. *Bell System Tech. J.*, (49): 1609–1625, 1970

[9] L. Kuipers, H. Niederreiter. Uniform distribution of sequences. *Wiley-Interscience [John Wiley & Sons], New York-London-Sydney. Reprint, Dover Publications, Mineola, NY 2006*, 1974

[10] B. Massé, D. Schneider. Random number sequences and the first digit phenomenon. *Electron. J. Probab.*, (17): no 86–17, 2012

[11] B. Massé, D. Schneider. Fast growing sequences of numbers and the first digit phenomenon. *Int. J. Number Theory*, (11): 705–719, 2015

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

[13] S. Newcomb. Note on the Frequency of Use of the Different Digits in Natural Numbers. *Amer. J. Math.*, 39–40, 1881

[14] M. J. Nigrini, L. J. Mittermaier. The use of Benford?s law as an aid in analytical procedures. *Auditing J. Pract. Th.* (2): 52–57, 1997

[15] J. Rivat, G. Tenenbaum. Constantes d'Erdös-Turán. *Ramanujan J.* (9): 111–121, 2005

[16] M. J. Sharpe. Limit laws and mantissa distributions. *Probab. Math. Statist.*, (26): 175–185, 2006

[17] B. Xu, and J. Wang, and G. Liu, and Y. Dai. Photorealistic computer graphics forensics based on leading digit law. *J. Electron.*, (28): 95–100, 2011