AVERAGING ALONG UNIFORM RANDOM INTEGERS

ÉLISE JANVRESSE AND THIERRY DE LA RUE

Abstract. Motivated by giving a meaning to "The probability that a random integer has initial digit d", we define a *URI-set* as a random set $\boldsymbol{E}$ of natural integers such that each $n \geq 1$ belongs to $\boldsymbol{E}$ with probability $1/n$, independently of other integers. This enables us to introduce two notions of densities on natural numbers: The *URI-density*, obtained by averaging along the elements of $\boldsymbol{E}$, and the *local URI-density*, which we get by considering the k-th element of $\boldsymbol{E}$ and letting k go to ∞. We prove that the elements of $\boldsymbol{E}$ satisfy Benford's law, both in the sense of URI-density and in the sense of local URI-density. Moreover, if b_1 and b_2 are two multiplicatively independent integers, then the mantissae of a natural number in base b_1 and in base b_2 are independent. Connections of URI-density and local URI-density with other well-known notions of densities are established: Both are stronger than the natural density, and URI-density is equivalent to log-density. We also give a stochastic interpretation, in terms of URI-set, of the H_∞-density.

1. Introduction

1.1. **Benford's law and Flehinger's theorem.** Benford's law describes the empirical distribution of the leading digit of everyday-life numbers. It was first discovered by the astronomer Simon Newcomb in 1881 [11] and named after the physicist Franck Benford who independently rediscovered the phenomenon in 1938 [1]. According to this law, the proportion of numbers in large series of empirical data with leading digit $d \in \{1, 2, \ldots, 9\}$ is $\log_{10}(1 + 1/d)$. More generally, defining the *mantissa* $\mathscr{M}(x)$ of a positive real number x as the only real number in $[1, 10[$ such that $x = \mathscr{M}(x)10^k$ for some integer k, Benford's law states that for any $1 \leq \alpha < \beta < 10$, the proportion of numbers whose mantissa lies in $[\alpha, \beta]$ is $\log_{10}\beta - \log_{10}\alpha$.

Giving Benford's law a mathematical meaning requires to formalize the notion of "everyday-life numbers", which is far from obvious. However there have been many attempts to explain mathematically the ubiquity of this distribution in empirical datasets. One of them is Betty J. Flehinger's theorem, published in 1965 in an article entitled *On the probability that a random integer has initial digit A* [6]. It occurred to Flehinger that the most natural set of numbers on which we should verify Benford's distribution is the whole set of positive integers. Unfortunately, defining

$$P_n^1(d) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{[d,d+1[}\big(\mathscr{M}(j)\big)$$

(that is the proportion of integers between 1 and n with leading digit d), we see that the sequence $P_n^1(d)$ has no limit as $n \to \infty$: It oscillates over longer and longer

2000 *Mathematics Subject Classification.* 11A63, 11B05, 60G50, 60G55.
Key words and phrases. Benford's law, log-density, H_∞-density, uniform random integers.

periods. Flehinger's idea was then to seek the limit by iteration of the process of Cesaro averaging: She inductively set

$$P_n^{k+1}(d) := \frac{1}{n}\sum_{j=1}^{n} P_j^k(d),$$

and proved that

$$\lim_{k\to\infty} \limsup_{n\to\infty} P_n^k(d) = \lim_{k\to\infty} \liminf_{n\to\infty} P_n^k(d) = \log_{10}\left(1+\frac{1}{d}\right), \tag{1}$$

which is the proportion predicted by Benford's law. (Donald Knuth generalized Flehinger's theorem to the distribution of the whole mantissa in 1981 [9].)

In spite of its title, Flehinger's article has no probabilistic content. A good reason is that there is no way of picking an integer uniformly at random in the set of all natural numbers. The first motivation for the present work was nevertheless to translate Flehinger's theorem in the context of probability theory: How can we interpret the *probability* that a (random) integer has a given initial digit? Our purpose is thus to give a meaning to the sentence "An integer picked uniformly at random has such a property".

1.2. **Roadmap of the paper.** We construct in Section 2 a random infinite set $\boldsymbol{E}$ of integers, which we call a *URI-set*, such that averaging along the elements of $\boldsymbol{E}$ reflects the expected behaviour of a random integer. This random set enables us to introduce two notions of densities on natural numbers: The *URI-density* (see Section 3.1), obtained by averaging along the elements of $\boldsymbol{E}$, and the *local URI-density* (see Section 5.2), which we get by considering the k-th element of $\boldsymbol{E}$ and letting k go to ∞. We prove that the elements of $\boldsymbol{E}$ satisfy Benford's law, both in the sense of URI-density (Theorem 3.5), and in the sense of local URI-density (Theorem 5.1). Our point of view also enables to consider simultaneously the mantissae of a number in different bases, and in particular we prove a result which can be interpreted as follows: If b_1 and b_2 are two multiplicatively independent integers, then the mantissae of a natural number in base b_1 and in base b_2 are independent (Theorem 4.1). Connections of URI-density and local URI-density with other well-known notions of densities are established: We prove that both are stronger than the natural density (Theorems 3.3 and 5.2), and that in fact URI-density is equivalent to log-density (Theorem 3.6). We finish in Section 6 by giving a stochastic interpretation, in terms of URI-set, of the H_∞-density used in Flehinger's theorem, and by raising some open problems.

The construction of the random set of integers is inspired by a previous article by the same authors [8], where a probabilistic proof of Flehinger's theorem was provided. We summarize this proof in Section 2.1

Acknowledgements. The authors are grateful to Bodo Volkmann for stimulating questions.

2. Uniform random set of integers

2.1. **Flehinger's theorem through Markov chain.** We introduce a homogeneous Markov chain $(M_k)_{k\geq 0}$ taking values in $[1, 10[$, defined by its initial value M_0

(which can be deterministic or random) and the following transition probability: for any Borel set $S \subset [1, 10[$,

$$\mathbb{P}\big(M_{k+1} \in S \big| M_k = a\big) := \mathbb{P}\Big(\mathscr{M}(aU) \in S\Big), \tag{2}$$

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

Let us denote by μ^{B} the probability distribution on $[1, 10[$ given by Benford's law: For any $1 \le t < 10$,

$$\mu^{\mathrm{B}}([1, t]) := \log_{10} t.$$

It is proved in [8] by a standard coupling argument that

- μ^{B} is the only probability distribution on $[1, 10[$ which is invariant under the probability transition (2);
- Whatever choice we make for the initial condition M_0, we have for any Borel set $S \subset [1, 10[$ and for all $k \ge 1$

$$\big|\mathbb{P}(M_k \in S) - \mu^{\mathrm{B}}(S)\big| \le \left(\frac{9}{10}\right)^k. \tag{3}$$

A connection is made between the quantities $P_n^k(d)$, $n \ge 1$, and the k-th step of our Markov chain: It is established in [8] that for all $a \in [1, 10[$ and all $k \ge 1$,

$$\lim_{j\to\infty} P^k_{\lfloor a10^j \rfloor}(d) = \mathbb{P}\Big(M_k \in [d, d+1[\,\Big|\, M_0 = a\Big).$$

A proof of (1), with an estimation of the speed of convergence, follows: We get for all $k \ge 1$

$$\begin{aligned} &\Big|\liminf_{n\to\infty} P_n^k(d) - \mu^{\mathrm{B}}\Big([d, d+1[\Big)\Big| \le \left(\frac{9}{10}\right)^k \\ \text{and}\quad &\Big|\limsup_{n\to\infty} P_n^k(d) - \mu^{\mathrm{B}}\Big([d, d+1[\Big)\Big| \le \left(\frac{9}{10}\right)^k. \end{aligned} \tag{4}$$

2.2. **Construction of a random set of integers.** We can interpret the Markov chain (M_k) as the sequence of mantissae of positive random variables X_k, where the sequence (X_k) is itself a Markov chain such that, given $X_0, \dots, X_k$, X_{k+1} is uniformly distributed in $]0, X_k[$. Then $\boldsymbol{X} := \{X_k,\ k \ge 0\}$ is a discrete random set of real numbers which satisfies the following property: For any $t > 0$, conditionally to the fact that $\boldsymbol{X} \cap]t, \infty[\neq \emptyset$, $\max(\boldsymbol{X} \cap]0, t])$ is uniformly distributed in $]0, t]$, and independent of $\boldsymbol{X} \cap]t, \infty[$.

Our idea is thus to imitate the structure of this random set of reals, but inside the set of natural numbers. We are looking for a random infinite set of integers $\boldsymbol{E}$ satisfying

(U) for all $n \ge 1$, $\max(\boldsymbol{E} \cap \{1, \dots, n\})$ is uniformly distributed in $\{1, \dots, n\}$;
(I) for all $n \ge 1$, $\max(\boldsymbol{E} \cap \{1, \dots, n\})$ is independent of $\boldsymbol{E} \cap \{n+1, n+2, \dots\}$.

For such a random set $\boldsymbol{E}$, we must have by (U), for each $n \ge 1$,

$$\mathbb{P}(n \in \boldsymbol{E}) = \mathbb{P}\big(\max(\boldsymbol{E} \cap \{1, \dots, n\}) = n\big) = 1/n,$$

and (I) implies that all events $(n \in \boldsymbol{E})$ are independent.

Conversely, picking elements of $\boldsymbol{E}$ using independent Bernoulli random variables, with $\mathbb{P}(n \in \boldsymbol{E}) = 1/n$ for each $n \ge 1$ gives a random set satisfying the required

conditions. Indeed, for each $j \in \{1, \dots, n\}$, we get

$$\mathbb{P}\Big(\max(\boldsymbol{E} \cap \{1, \dots, n\}) = j\Big) = \mathbb{P}\Big(j \in \boldsymbol{E}, j+1 \notin \boldsymbol{E}, \dots, n \notin \boldsymbol{E}\Big) = \frac{1}{j}\frac{j}{j+1} \cdots \frac{n-1}{n} = \frac{1}{n}.$$

Observe also that, with probability 1, the cardinality of $\boldsymbol{E}$ is infinite.

Because of the uniformity property (U), such a random set $\boldsymbol{E}$ appears as a good way to modelize the uniform distribution in the set of natural numbers and will therefore be referred to as a set of *uniform random integers*, or *URI-set*.

3. URI-density and Benford's law

From now on, $\boldsymbol{E}$ denotes a URI-set, and we denote its ordered elements by

$$\boldsymbol{E} = \{N_1 = 1 < N_2 < \ldots < N_k < \ldots\}.$$

For each $n \geq 1$, we set $\boldsymbol{E}_n := \boldsymbol{E} \cap \{1, \dots, n\}$.

It will be useful to give the following estimation of $|\boldsymbol{E}_n|$.

Lemma 3.1.

$$\frac{|\boldsymbol{E}_n|}{\ln n} \xrightarrow[n\to\infty]{a.s.} 1.$$

We recall Theorem 12 page 272 in Petrov's book [12]:

Theorem 3.2. *Let (Z_n) be a sequence of independent centered real-valued random variables Z_n. If $t_n \uparrow \infty$ and*

$$\sum_{n\geq 1} \frac{\mathbb{E}\left[Z_n^p\right]}{t_n^p} < \infty,$$

for some p, $1 \leq p \leq 2$, then

$$\frac{\sum_{j=1}^n Z_j}{t_j} \xrightarrow[n\to\infty]{a.s.} 0.$$

Proof of Lemma 3.1. We consider the independent centered random variables $Z_n := \mathbb{1}_{\boldsymbol{E}}(n) - 1/n$. Since

$$\sum_{n\geq 1} \frac{\mathbb{E}\left[Z_n^2\right]}{(\ln n)^2} \leq \sum_{n\geq 1} \left(1 - \frac{1}{n}\right) \frac{1}{n(\ln n)^2} < \infty,$$

we get by Theorem 3.2

$$\frac{\sum_{j=1}^n Z_j}{\ln n} = \frac{1}{\ln n} \sum_{j=1}^n \mathbb{1}_{\boldsymbol{E}}(j) - \frac{1}{\ln n} \sum_{j=1}^n 1/j \xrightarrow[n\to\infty]{a.s.} 0\,. \tag{5}$$

This concludes the proof of the lemma. □

3.1. **URI-density.** We say that a subset A of the set of natural numbers has *URI-density* α if

$$\frac{1}{n} \sum_{k=1}^n \mathbb{1}_A(N_k) \xrightarrow[n\to\infty]{a.s.} \alpha.$$

Note that an equivalent formulation is

$$\frac{1}{|\boldsymbol{E}_n|} \sum_{j \in \boldsymbol{E}_n} \mathbb{1}_A(j) \xrightarrow[n\to\infty]{a.s.} \alpha.$$

As we expect, the URI-density generalizes the notion of natural density.

Theorem 3.3. *Let $A \subset \mathbb{Z}_+$. If $\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_A(j) \xrightarrow[n\to\infty]{} \alpha$, then A has URI-density α.*

Proof. When considering the elements of $\boldsymbol{E}_n$, it will be convenient to order them backwards:

$$\boldsymbol{E}_n = \boldsymbol{E} \cap \{1,\dots,n\} = \left\{ Y_1^{(n)} > Y_2^{(n)} > \ldots > Y_{|\boldsymbol{E}_n|}^{(n)} = 1 \right\}.$$

For each $n \geq 1$, let $a_n := \mathbb{1}_A(n)$. We are going to prove the result in the form

$$\frac{1}{|\boldsymbol{E}_n|} \sum_{j \in \boldsymbol{E}_n} a_j \xrightarrow[n\to\infty]{a.s.} \alpha. \tag{6}$$

We split the above average as

$$\frac{1}{|\boldsymbol{E}_n|} \sum_{j \in \boldsymbol{E}_n} a_j = \frac{1}{|\boldsymbol{E}_n|} \sum_{i=1}^{|\boldsymbol{E}_n|} \left(a_{Y_i^{(n)}} - K_i^n \right) + \frac{1}{|\boldsymbol{E}_n|} \sum_{i=1}^{|\boldsymbol{E}_n|} K_i^n, \tag{7}$$

where

$$K_i^n := \begin{cases} \mathbb{E}\left[a_{Y_i^{(n)}} \,\middle|\, Y_{i-1}^{(n)}, \dots, Y_1^{(n)} \right] & (2 \leq i \leq |\boldsymbol{E}_n|), \\ \mathbb{E}\left[a_{Y_1^{(n)}} \right] & (i = 1). \end{cases}$$

We first deal with the second term of (7). By Properties (U) and (I), $Y_1^{(n)}$ is uniformly distributed in $\{1,\dots,n\}$, and conditionally to $\left(Y_{i-1}^{(n)}, \dots, Y_1^{(n)}\right)$, $Y_i^{(n)}$ is uniformly distributed in $\left\{1, \dots, Y_{i-1}^{(n)} - 1\right\}$, as long as $Y_{i-1}^{(n)} > 1$. Hence,

$$K_1^n = \frac{1}{n}\sum_{j=1}^{n} a_j \quad \text{and} \quad K_i^n = \frac{1}{Y_{i-1}^{(n)} - 1} \sum_{j=1}^{Y_{i-1}^{(n)}-1} a_j, \quad (2 \leq i \leq |\boldsymbol{E}_n|).$$

By hypothesis, as soon as $Y_{i-1}^{(n)}$ is large, K_i^n is close to α. For any fixed $\varepsilon > 0$, the number of i's such that $|K_i^n - \alpha| > \varepsilon$ is bounded independently of n. Since $|\boldsymbol{E}_n| \to \infty$ a.s., it follows that

$$\frac{1}{|\boldsymbol{E}_n|} \sum_{i=1}^{|\boldsymbol{E}_n|} K_i^n \xrightarrow[n\to\infty]{a.s.} \alpha\,.$$

Let us now turn to the first term of (7). Define for any $n \geq 1$ and $i \geq 1$

$$A_i^n := \begin{cases} a_{Y_i^{(n)}} - K_i^n & (1 \leq i \leq |\boldsymbol{E}_n|), \\ 0 & (i > |\boldsymbol{E}_n|). \end{cases}$$

In order to prove (6), it remains to show that

$$\frac{1}{|\boldsymbol{E}_n|} \sum_{i=1}^{|\boldsymbol{E}_n|} A_i^n \xrightarrow[n\to\infty]{a.s.} 0. \tag{8}$$

Using a standard method, we first prove that (8) holds along a subsequence, then we control the oscillations to conclude. Consider $n_m := \lfloor \exp(m^2) \rfloor$. By lemma 3.1,

convergence along the subsequence (n_m) amounts to

$$A(n_m) := \frac{1}{\lfloor \ln n_m \rfloor} \sum_{i=1}^{\lfloor \ln n_m \rfloor} A_i^{n_m} \xrightarrow[m\to\infty]{a.s.} 0\,.$$

Observe now that the variance of $A_i^{n_m}$ is bounded, and that for $i \neq j$, $\mathbb{E}\left[A_i^{n_m} A_j^{n_m}\right] = 0$. Therefore, the variance of $A(n_m)$ is of order $(\ln n_m)^{-1} = 1/m^2$. By Tchebychev's inequality,

$$\sum_{m\geq 1} \mathbb{P}\left(A(n_m) > m^{-1/4}\right) \leq \sum_{m\geq 1} \frac{\mathrm{Var}(A(n_m))}{m^{-1/2}} = \sum_{m\geq 1} m^{1/2} O\left(\frac{1}{m^2}\right) < \infty.$$

Hence, by Borel-Cantelli, with probability 1 we have $A(n_m) \leq m^{-1/4}$ for m large enough. This proves that

$$\frac{1}{|\boldsymbol{E}_{n_m}|} \sum_{i=1}^{|\boldsymbol{E}_{n_m}|} A_i^{n_m} \xrightarrow[m\to\infty]{a.s.} 0.$$

Consider now an integer $n \in]n_m, n_{m+1}[$. We can write

$$\frac{1}{|\boldsymbol{E}_n|} \sum_{i=1}^{|\boldsymbol{E}_n|} A_i^n = \frac{1}{|\boldsymbol{E}_n|} \left(\sum_{i=1}^{|\boldsymbol{E}_n|} A_i^n - \sum_{i=1}^{|\boldsymbol{E}_{n_m}|} A_i^{n_m} \right) + \left\{ \frac{1}{|\boldsymbol{E}_n|} - \frac{1}{|\boldsymbol{E}_{n_m}|} \right\} \sum_{i=1}^{|\boldsymbol{E}_{n_m}|} A_i^{n_m} + \frac{1}{|\boldsymbol{E}_{n_m}|} \sum_{i=1}^{|\boldsymbol{E}_{n_m}|} A_i^{n_m} \tag{9}$$

Since $A_i^{n_m}$ is bounded and $|\boldsymbol{E}_{n_m}| - |\boldsymbol{E}_n| = o(|\boldsymbol{E}_n|)$, the second term on the RHS vanishes as $m \to \infty$. Moreover, $\boldsymbol{E}_{n_m} \subset \boldsymbol{E}_n$. Therefore each $A_i^{n_m}$ in the first term (except $A_1^{n_m}$ which has a slightly different definition) is annihilated by some A_j^n, and the first term of (9) reduces to

$$\frac{1}{|\boldsymbol{E}_n|} \sum_{i=1}^{|\boldsymbol{E}_n| - |\boldsymbol{E}_{n_m}| + 1} A_i^n - \frac{1}{|\boldsymbol{E}_n|} A_1^{n_m},$$

which goes to zero as $m \to \infty$. Since we already know that the third term goes to zero, this proves (8). □

3.2. **Benford's law.** We say that a sequence of positive real numbers (x_n) follows Benford's law if for all $1 \leq t < 10$

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\mathscr{M}(x_j) < t} \xrightarrow[n\to\infty]{} \log_{10} t.$$

Remark 3.4. *We recall that this is equivalent to the uniform distribution* $\bmod 1$ *of the sequence* $(\log_{10} x_k)$ *(see e.g.* [3]*). Therefore, this is also equivalent to the fact that the sequence* $(1/x_k)$ *follows Benford's law.*

The following theorem shows that the elements of the URI-set $\boldsymbol{E}$ almost surely follow Benford's law.

Theorem 3.5. *For any* $1 \leq t \leq 10$*, the URI-density of* $\{n \geq 1 : \ \mathscr{M}(n) < t\}$ *is* $\log_{10} t$*.*

Proof. By Duncan's work [5], this result can be viewed as a corollary of the equivalence between URI-density and log-density (see Theorem 3.6 below). However we provide a direct proof of it, in which useful ideas will be presented.

It is convenient to consider a coupling of the URI-set $\boldsymbol{E}$ and its continuous analog defined as follows: Let ξ be a Poisson process on $\mathbb{R}_+^*$ with intensity $1/x$. It can be viewed as a random set of points. For any interval I, let ξ_I denote the number of points in $I \cap \xi$: ξ_I is Poisson distributed with parameter $\int_I \frac{dx}{x}$. From ξ, define the random set $\boldsymbol{E}$ as the set of integers $n \geq 1$ such that $\xi_{]n-1,n]} \geq 1$. Since the random variables $(\xi_{]n-1,n]})_{n\geq 1}$ are independent and

$$\mathbb{P}(n \in \boldsymbol{E}) = 1 - \mathbb{P}\left(\xi_{]n-1,n]} = 0\right) = \frac{1}{n},$$

$\boldsymbol{E}$ is a URI-set.

The process ξ satisfies a property analogous to Property (U): For any $a > 0$, the largest point of $[0, a] \cap \xi$ is uniformly distributed in $[0, a]$. Indeed, for any $t < a$

$$\mathbb{P}\Big(\max([0, a] \cap \xi) \leq t\Big) = \mathbb{P}\left(\xi_{[t,a]} = 0\right) = \frac{t}{a}.$$

Let us order backwards the points of $\xi \cap [0, 1]$: $1 > Y_1 > Y_2 > \ldots$. Conditionally to Y_k, Y_{k+1} is uniformly distributed in $[0, Y_k]$. By Proposition 3.1 of [8], the mantissae $(\mathscr{M}(Y_k))$ constitute a Markov chain whose unique invariant distribution is μ^{B}. Moreover, Y_1 being uniformly distributed in $[0, 1]$, the distribution of $\mathscr{M}(Y_1)$ is the normalized Lebesgue measure on $[1, 10[$, hence is equivalent to μ^{B}. By the pointwise ergodic theorem, we have for any $1 \leq t < 10$

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\mathscr{M}(Y_k)<t} \xrightarrow[n\to\infty]{a.s.} \log_{10} t.$$

Consider now the points of $\xi \cap]1, +\infty[$: $X_1 < X_2 < \ldots$ They follow the same distribution as $(1/Y_1, 1/Y_2, \ldots)$. By Remark 3.4, we deduce that

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\mathscr{M}(X_k)<t} \xrightarrow[n\to\infty]{a.s.} \log_{10} t. \tag{10}$$

Now observe that

$$\sum_{n\geq 1} \mathbb{P}(\xi_{]n-1,n]} \geq 2) = \sum_{n\geq 1} \left(\frac{1}{n} + \frac{n-1}{n} \log \frac{n-1}{n}\right) < \infty,$$

hence by Borel-Cantelli, with probability one there is only a finite number of n's such that $\xi_{]n-1,n]} \geq 2$. Coming back to the URI-set $\boldsymbol{E} = \{N_1 = 1 < N_2 < \cdots\}$, this implies the almost-sure existence of R such that, for all large enough k, $0 \leq N_k - X_{k+R} < 1$. Since $N_k \to \infty$ a.s., we have $|\mathscr{M}(N_k) - \mathscr{M}(X_{k+R})| \to 0$ unless N_k be of the form 10^p (which, again by Borel-Cantelli, happens almost surely for only finitely many k's). It follows from (10) that

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\mathscr{M}(N_k)<t} \xrightarrow[n\to\infty]{a.s.} \log_{10} t.$$

□

3.3. **Equivalence with log-density.** To deal with the problem of non-existence of natural densities, several alternative densities have been introduced. Flehinger's theorem amounts to considering the so-called H^∞-*density*, obtained by iteration of Cesaro averages: A subset A of $\mathbb{Z}_+$ has H^∞-*density* α if

$$\lim_{k\to\infty} \limsup_{n\to\infty} P_n^k = \lim_{k\to\infty} \liminf_{n\to\infty} P_n^k = \alpha,$$

where the P_N^k's are inductively defined by $P_n^0 := \mathbb{1}_A(n)$ and $P_n^{k+1} := (1/n)\sum_{j=1}^n P_j^k$. Obviously, H^∞-density is *stronger* than natural density, in the sense used by Diaconis in [2]: If A has natural density α, then A has H^∞-density α. The example of the set A of integer whose initial digit (when written in base 10) is 1 shows that H^∞-density is *strictly* stronger than natural density.

Still stronger than H^∞-density is the notion of *log-density*. Recall that $A \subset \mathbb{Z}_+$ has *log-density* α if

$$\frac{1}{\ln n}\sum_{j=1}^n \frac{1}{j}\mathbb{1}_A(j) \xrightarrow[n\to\infty]{} \alpha.$$

A proof that log-density is stronger than H^∞-density can be found in [3], together with an example of a set A with a log-density, but for which the H^∞-density fails to exist.

Theorem 3.6. *Let $A \subset \mathbb{Z}_+$.*

$$\frac{1}{|\boldsymbol{E}_n|}\sum_{j=1}^n \mathbb{1}_A(j)\mathbb{1}_{\boldsymbol{E}}(j) - \frac{1}{\ln n}\sum_{j=1}^n \mathbb{1}_A(j)/j \xrightarrow[n\to\infty]{a.s.} 0\,.$$

In particular, A has URI-density α if and only if A has log-density α.

Proof. We apply Theorem 3.2: We consider the independent random variables $Z_n := \mathbb{1}_A(n)(\mathbb{1}_{\boldsymbol{E}}(n) - 1/n)$ which are centered. Since

$$\sum_{n\geq 1} \frac{\mathbb{E}\left[Z_n^2\right]}{(\ln n)^2} \leq \sum_{n\geq 1}\left(1-\frac{1}{n}\right)\frac{1}{n(\ln n)^2} < \infty,$$

we get

$$\frac{\sum_{j=1}^n Z_j}{\ln n} = \frac{1}{\ln n}\sum_{j=1}^n \mathbb{1}_A(j)\mathbb{1}_{\boldsymbol{E}}(j) - \frac{1}{\ln n}\sum_{j=1}^n \mathbb{1}_A(j)/j \xrightarrow[n\to\infty]{a.s.} 0\,. \tag{11}$$

Since, by Lemma 3.1

$$\frac{|\boldsymbol{E}_n|}{\ln n} \xrightarrow[n\to\infty]{a.s.} 1,$$

we can conclude the proof of Theorem 3.6. □

4. Independence of mantissae in different bases

All the previous results concerned numbers written in the base-10 numeral system, but extend straightforwardlly to any integer base b. We denote by $\mathscr{M}_b(x) \in [1, b[$ the *mantissa in base* b of a positive real number x and by μ_b^{B} the probability distribution over $[1, b[$ defined by $\mu_b^{\mathrm{B}}([1,t]) := \log_b t$ for all $1 \leq t < b$.

The purpose of this section is to prove the following theorem, which states that the mantissae in different bases of the elements of the URI-set $\boldsymbol{E}$ are independent, under some algebraic condition on the bases.

Theorem 4.1. *Let $(b_i)_{1\le i\le \ell}$ be positive integers, satisfying*

$$\forall a_1,\dots,a_\ell\in\mathbb{Z},\ \Big[\frac{a_1}{\ln b_1}+\cdots+\frac{a_\ell}{\ln b_\ell}=0\Big]\Longrightarrow a_1=\cdots=a_\ell=0. \tag{12}$$

Then for any $1\le t_i<b_i$ $(1\le i\le \ell)$, $\{n\in\mathbb{Z}_+:\mathscr{M}_{b_i}(n)\le t_i\ \forall 1\le i\le\ell\}$ has URI-density equal to $\prod_{i=1}^{\ell}\log_{b_i} t_i$.

Recall that the positive integers $(b_i)_{1\le i\le\ell}$ are said to be *multiplicatively independent* if $b_1^{s_1}\dots b_\ell^{s_\ell}=1$ where $(s_i)_{1\le i\le\ell}\subset\mathbb{Z}$ implies that $s_i=0$ for all i. Note that, in the case $\ell=2$, property (12) exactly means that b_1 and b_2 are multiplicatively independent. To our knowledge, it is unknown whether, in the general case, multiplicative independence of $b_1,\dots,b_\ell$ implies property (12). This question is related to the so-called *Schanuel's conjecture* in transcendental number theory (see [10], p. 30-31 or [13]).

Lemma 4.2. *Let $(Z_k)_{k\ge1}$ be i.i.d. random variables taking values in $(\mathbb{R}/\mathbb{Z})^\ell$ with common distribution ν. Assume that the only probability distribution μ such that $\mu*\nu=\mu$ is the Lebesgue measure on $(\mathbb{R}/\mathbb{Z})^\ell$. Then the random walk $(P_k)_{k\ge1}:=(Z_1+\cdots+Z_k)_{k\ge1}$ is uniformly distributed on $(\mathbb{R}/\mathbb{Z})^\ell$. In other words, it satisfies: For all cylinder $C=[u_1,v_1]\times\cdots\times[u_\ell,v_\ell]$ where $0\le u_i<v_i<1$,*

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_C(P_k)\xrightarrow[n\to\infty]{a.s.}\prod_{i=1}^{\ell}(v_i-u_i). \tag{13}$$

Proof. Denote by μ the Lebesgue measure on $(\mathbb{R}/\mathbb{Z})^\ell$. Let M_0 be a random variable with law μ, independent of $(Z_k)_{k\ge1}$. Setting

$$M_k:=M_0+Z_1+\cdots+Z_k=M_0+P_k,$$

we get a stationary random walk $(M_k)_{k\ge0}$. Since μ is the unique invariant measure under convolution by ν, the stationary process $(M_k)_{k\ge0}$ is ergodic, and by Birkhoff ergodic theorem, we get that for all cylinder $C=[u_1,v_1]\times\cdots\times[u_\ell,v_\ell]$ where $0\le u_i<v_i<1$,

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_C(M_k)\xrightarrow[n\to\infty]{a.s.}\prod_{i=1}^{\ell}(v_i-u_i).$$

(See *e.g.* [7], Corollary 2.5.2 page 38.) Therefore, we can find some $m_0\in(\mathbb{R}/\mathbb{Z})^\ell$ such that, with probability 1, for all cylinder $C=[u_1,v_1]\times\cdots\times[u_\ell,v_\ell]$ where $0\le u_i<v_i<1$ are rational numbers,

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_C(m_0+P_k)\xrightarrow[n\to\infty]{}\prod_{i=1}^{\ell}(v_i-u_i).$$

We thus obtain that, for all cylinder C with rational endpoints,

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{C-m_0}(P_k)\xrightarrow[n\to\infty]{a.s.}\prod_{i=1}^{\ell}(v_i-u_i)=\mu(C-m_0).$$

By density of the rationals, (13) is satisfied for any cylinder C. □

Proof of Theorem 4.1. As in the proof of Theorem 3.5, we consider the coupling of the URI-set with the Poisson process ξ, and we denote by $\cdots>X_2>X_1>1>Y_1>Y_2>\cdots$ the points of ξ. Define, for all $k\ge1$, $U_k:=Y_k/Y_{k-1}$ (where

$Y_0 := 1$): Then $(U_k)_{k\geq 1}$ is a sequence of i.i.d. uniform random variables in $[0,1]$, and $Y_k = U_1 U_2 \dots U_k$.

Set $Z_k := \Big(\log_{b_1}(U_k) \mod 1, \dots, \log_{b_\ell}(U_k) \mod 1\Big) \in (\mathbb{R}/\mathbb{Z})^\ell$, and let ν be the common law of the Z_k's. We claim that the only probability measure μ which is invariant under convolution by ν is the Lebesgue measure on $(\mathbb{R}/\mathbb{Z})^\ell$. Indeed, for such an invariant measure, the Fourier coefficients must satisfy

$$\forall (m_1, \dots, m_\ell) \in \mathbb{Z}^\ell, \quad \widehat{\mu}(m_1, \dots, m_\ell) = \widehat{\mu}(m_1, \dots, m_\ell)\widehat{\nu}(m_1, \dots, m_\ell).$$

We just have to check that $\widehat{\nu}(m_1, \dots, m_\ell) \neq 1$ when $(m_1, \dots, m_\ell) \neq (0, \dots, 0)$.

$$\begin{aligned}
\widehat{\nu}(m_1, \dots, m_\ell) &= \int_{(\mathbb{R}/\mathbb{Z})^\ell} e^{-i2\pi(m_1 t_1 + \dots + m_\ell t_\ell)}\, d\nu(t_1, \dots, t_\ell) \\
&= \int_{[0,1]} e^{-i2\pi\left(m_1 \log_{b_1} u + \dots + m_\ell \log_{b_\ell} u\right)}\, du \\
&= \int_{[0,1]} e^{-i2\pi\theta \ln u}\, du,
\end{aligned}$$

where $\theta := \frac{m_1}{\ln b_1} + \dots + \frac{m_\ell}{\ln b_\ell} \neq 0$ for $(m_1, \dots, m_\ell) \neq (0, \dots, 0)$ by (12). Hence,

$$\widehat{\nu}(m_1, \dots, m_\ell) = \frac{1}{1 - i2\pi\theta} \neq 1,$$

which proves the claim.

It follows by Lemma 4.2 that the sequence

$$\Big(\log_{b_1}(Y_k) \mod 1, \dots, \log_{b_\ell}(Y_k) \mod 1\Big)$$

is uniformly distributed in $(\mathbb{R}/\mathbb{Z})^\ell$, and the same is true if we replace Y_k by X_k. The end of the proof goes with similar arguments as for Theorem 3.5. □

Cassels-Schmidt-Benford sequences. Given two multiplicatively independent positive integers b_1 and b_2, Bodo Volkmann defined a *Cassels-Schmidt number of type* (b_1, b_2) as a number which is normal in base b_1 but not in base b_2. By analogy, he also proposed to define a *Cassels-Schmidt-Benford (CSB) sequence of type* (b_1, b_2) as a sequence of positive numbers (x_k) which follows Benford's law with respect to base b_1 but not with respect to base b_2.

It turns out that it is far easier to find explicit examples of CSB sequences. Indeed, take $x_k := (b_2)^k$, then (x_k) does certainly not follow Benford's law with respect to base b_2. But since $\ln b_2 / \ln b_1$ is not a rational number, the sequence $\big(\log_{b_1} x_k \mod 1\big) = (k \ln b_2 / \ln b_1 \mod 1)$ is uniformly distributed in $[0,1]$. Hence (x_k) follows Benford's law with respect to base b_1.

As an application of Theorem 4.1 in the case $\ell = 2$, we get that in almost every URI-set, we can find a CSB sequence (x_k) of type (b_1, b_2), such that the sequence $(\mathscr{M}_{b_2}(x_k))$ follow any probability distribution prescribed in advance on $[1, b_2[$.

5. Local URI-density

For the sake of simplicity, we now return to the classical base 10 numeration system (but obviously the following results also hold in any integer base).

5.1. **A single element of a URI-set satisfies Benford's law.** According to Theorem 3.6, Theorem 3.5 turns out to be weaker than Flehinger's theorem. However similar ideas to those developed in the proof can lead to somewhat stronger results than Theorem 3.5.

Theorem 5.1. *For all* $1 \leq \alpha < \beta < 10$

$$\lim_{k\to\infty} \mathbb{P}\Big(\mathscr{M}(N_k) \in [\alpha,\beta]\Big) = \mu^B([\alpha,\beta]).$$

Proof. As in the proof of Theorem 3.5, we consider the URI-set $\boldsymbol{E}$ constructed from the Poisson process ξ. For a fixed integer m, we number the points of $\xi\cap]m,+\infty[$:

$$\xi\cap]m,+\infty[= \{X_1^m < X_2^m < \ldots\}.$$

We also set $X_0^m := m$. Observe that the process $\left(1/X_k^m\right)_{n\geq 0}$ is again a Markov chain such that, given $(1/X_0^m,\ldots,1/X_k^m)$, $1/X_{k+1}^m$ is uniformly distributed in $]0,1/X_k^m[$. It follows by (3) and Remark 3.4 that, for any Borel set $S\subset[1,10[$,

$$\forall k\geq 1,\quad \left|\mathbb{P}\Big(\mathscr{M}\left(X_k^m\right)\in S\Big) - \mu^{\mathrm{B}}(S)\right| \leq \left(\frac{9}{10}\right)^k. \tag{14}$$

We now consider the event

$$A_m := \bigcap_{n>m}\left(\xi_{]n-1,n]}\leq 1\right)\cap \bigcap_{\ell:10^\ell>m}\left(\xi_{]10^\ell-1,10^\ell]}=0\right).$$

Defining the random variable J_m as the largest index such that $N_{J_m}\leq m$, the realization of A_m ensures that the shift of index between the process (X_k^m) and the integers N_k that are larger than $m+1$ remains constant, equal to J_m. Therefore, for any $k\geq 1$

$$0\leq N_{J_m+k} - X_k^m < 1.$$

Moreover, since A_m also forbids that any $N_k > m$ be of the form 10^ℓ, we get (conditionally to A_m)

$$\forall k\geq 1,\quad 0\leq \mathscr{M}\left(N_{J_m+k}\right) - \mathscr{M}\left(X_k^m\right) < \frac{10}{N_{J_m+k}} < \frac{10}{m}. \tag{15}$$

Note also that $\mathbb{P}(A_m)\to 1$ as $m\to\infty$, so that choosing m large enough will enable us to condition with respect to A_m without affecting too much the probability of any event. Indeed, we will make use of the following inequality, valid for any events A and B with $\mathbb{P}(A)>0$:

$$|\mathbb{P}(B\,|\,A) - \mathbb{P}(B)| \leq \frac{\mathbb{P}(A^c)}{\mathbb{P}(A)}. \tag{16}$$

Let us fix an arbitrary $\varepsilon>0$. We choose m large enough so that

$$\frac{\mathbb{P}(A_m^c)}{\mathbb{P}(A_m)} < \varepsilon \quad\text{and}\quad \frac{10}{m}<\varepsilon. \tag{17}$$

Conditioning with respect to J_m which takes values in $\{1,\ldots,m\}$, we get

$$\left|\mathbb{P}\Big(\mathscr{M}(N_k)\in[\alpha,\beta]\Big) - \mu^{\mathrm{B}}([\alpha,\beta])\right| \leq \sum_{j=1}^m \mathbb{P}(J_m=j)\left|\mathbb{P}\Big(\mathscr{M}(N_k)\in[\alpha,\beta]\,|\,J_m=j\Big) - \mu^{\mathrm{B}}([\alpha,\beta])\right|.$$

Then we write, for any $1 \leq j \leq m$,

$$\left|\mathbb{P}\Big(\mathscr{M}(N_k) \in [\alpha,\beta] \,|\, J_m = j\Big) - \mu^{\mathrm{B}}([\alpha,\beta])\right| \leq D_1 + D_2 + D_3 + D_4,$$

where

$$\begin{aligned}
D_1 &:= \left|\mathbb{P}\Big(\mathscr{M}(N_k) \in [\alpha,\beta] \,|\, J_m = j\Big) - \mathbb{P}\Big(\mathscr{M}(N_k) \in [\alpha,\beta] \,|\, A_m,\ J_m = j\Big)\right| \\
D_2 &:= \Big|\mathbb{P}\Big(\mathscr{M}(N_{J_m+k-j}) \in [\alpha,\beta] \,|\, A_m,\ J_m = j\Big) \\
&\qquad\qquad -\mathbb{P}\Big(\mathscr{M}(X^m_{k-j}) \in [\alpha,\beta] \,|\, A_m,\ J_m = j\Big)\Big|, \\
D_3 &:= \left|\mathbb{P}\Big(\mathscr{M}(X^m_{k-j}) \in [\alpha,\beta] \,|\, A_m,\ J_m = j\Big) - \mathbb{P}\Big(\mathscr{M}(X^m_{k-j}) \in [\alpha,\beta] \,|\, J_m = j\Big)\right| \\
D_4 &:= \left|\mathbb{P}\Big(\mathscr{M}(X^m_{k-j}) \in [\alpha,\beta] \,|\, J_m = j\Big) - \mu^{\mathrm{B}}([\alpha,\beta])\right|.
\end{aligned}$$

Observe that A_m, which is measurable with respect to the Poisson process on $]m,+\infty[$, is independent of $(J_m = j)$, which is measurable with respect to the Poisson process on $]1,m]$. Hence, using (16) and (17), we can bound $D_1 + D_3$ by 2ε.

Again, X^m_{k-j} is measurable with respect to the Poisson process on $]m,+\infty[$, hence is independent of $(J_m = j)$. By (14), the contribution of D_4 can be bounded by $(9/10)^{k-j}$, hence by $(9/10)^{k-m}$.

It remains to deal with D_2. Since everything is conditioned on A_m, we can use (15) to get

$$\begin{aligned}
D_2 \leq\ &\mathbb{P}\Big(\mathscr{M}(X^m_{k-j}) \in [\alpha - 10/m, \alpha] \,|\, A_m,\ J_m = j\Big) \\
&\qquad + \mathbb{P}\Big(\mathscr{M}(X^m_{k-j}) \in [\beta - 10/m, \beta] \,|\, A_m,\ J_m = j\Big).
\end{aligned}$$

Using again (16), (17) and (14) yields

$$D_2 \leq 2(9/10)^{k-m} + 2\varepsilon + \mu^{\mathrm{B}}\left([\alpha - 10/m, \alpha]\right) + \mu^{\mathrm{B}}\left([\beta - 10/m, \beta]\right) \leq 2(9/10)^{k-m} + 4\varepsilon.$$

□

Remark that the statement of Theorem 5.1 would not hold if we replace the interval $[\alpha,\beta]$ by any Borel set S. Indeed, since N_k is an integer, the probability that its mantissa belong to the set of irrational numbers is zero. In other words, the convergence of the distribution of $\mathscr{M}(N_k)$ to Benford's law is only a weak convergence. However, if we denote by $\tilde{X}_k$ the largest point of the Poisson process which is smaller than N_k, then $\tilde{X}_k \in]N_k - 1, N_k]$, and the distribution of $\mathscr{M}(\tilde{X}_k)$ converges to Benford's law in total variation norm.

5.2. **Local URI-density is stronger than natural density.** In view of Theorem 5.1, it is natural to introduce the *local URI-density* of $A \subset \mathbb{Z}_+$ as the limit, when $k \to \infty$, of $\mathbb{P}(N_k \in A)$ (whenever the limit exists). The purpose of this section is to prove that local URI-density is stronger than natural density:

Theorem 5.2. *If $A \subset \mathbb{Z}_+$ possesses a natural density, then the local URI-density of A exists and coincides with its natural density.*

In fact, local URI-density turns out to be *strictly* stronger than natural density, since Theorem 5.1 proves the existence of sets A without natural densities but for which the local URI-density exists.

The proof of Theorem 5.2 is based on the following lemma.

Lemma 5.3. *Let $(P_k)_{k\geq 1}$ be a sequence of probability distributions on $\mathbb{Z}_+$ satisfying: For any $k \geq 1$, there exists n_k such that*

- *The map $n \mapsto P_k(n)$ is non-decreasing on $\{1, \dots, n_k\}$ and non-increasing on $\{n_k, n_k + 1, \dots\}$;*
- *For any integer $m \geq 1$,*

$$\lim_{k\to\infty} \frac{P_k(mn_k)}{P_k(n_k)} = 1.$$

Then, if $A \subset \mathbb{Z}_+$ possesses a natural density α, $\lim_{k\to\infty} P_k(A) = \alpha$.

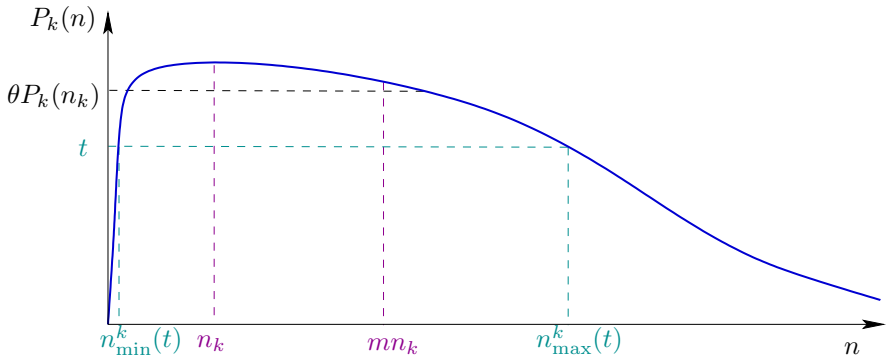

FIGURE 1. Profile of P_k

Proof. Let us fix $\varepsilon > 0$. Let $\theta \in]0, 1[$, close enough to 1 so that $(1 - \theta)/\theta < \varepsilon$. Let $m \in \mathbb{Z}_+$ be such that $m > 1/\varepsilon$ and such that, for any $n \geq m$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(i) \in]\alpha - \varepsilon, \alpha + \varepsilon[.$$

We choose k large enough such that

$$\frac{P_k(mn_k)}{P_k(n_k)} > \theta. \tag{18}$$

By a Fubini argument, we can write $P_k(A)$ as

$$P_k(A) = \int_0^{P_k(n_k)} |\{n \in A : P_k(n) > t\}|\, dt. \tag{19}$$

We split the integral into two terms

$$I_1 := \int_0^{\theta P_k(n_k)} |\{n \in A : P_k(n) > t\}|\, dt \text{ and } I_2 := \int_{\theta P_k(n_k)}^{P_k(n_k)} |\{n \in A : P_k(n) > t\}|\, dt.$$

Observe that

$$\theta P_k(n_k)\ |\{n \in \mathbb{Z}_+ : P_k(n) > \theta P_k(n_k)\}| \leq P_k\left(|\{n \in \mathbb{Z}_+ : P_k(n) > \theta P_k(n_k)\}|\right) \leq 1.$$

Therefore

$$I_2 \leq (1-\theta)P_k(n_k)\ \ |\{n \in \mathbb{Z}_+ : P_k(n) > \theta P_k(n_k)\}| \leq \frac{1-\theta}{\theta} < \varepsilon.$$

Let us turn to the estimation of I_1. By the hypothesis on the variations of $P_k(n)$, for any $0 < t < P_k(n_k)$, there exist $n^k_{\min}(t) \leq n_k \leq n^k_{\max}(t)$ such that

$$\{n : P_k(n) \geq t\} = \left\{n^k_{\min}(t), \dots, n^k_{\max}(t)\right\}.$$

(See Figure 5.2.) We can rewrite I_1 as

$$\int_0^{\theta P_k(n_k)} \Big(n^k_{\max}(t) - n^k_{\min}(t) + 1\Big)\varphi(t)\, dt,$$

where

$$\varphi(t) := \frac{1}{n^k_{\max}(t) - n^k_{\min}(t) + 1} \sum_{i=n^k_{\min}(t)}^{n^k_{\max}(t)} \mathbb{1}_A(i).$$

We prove that, for $0 < t < \theta P_k(n_k)$, $\varphi(t)$ is close to the natural density of A: By (18), for any $0 < t < \theta P_k(n_k)$, we have $n^k_{\max}(t) > mn_k$. Thus

$$1 \leq \frac{n^k_{\max}(t)}{n^k_{\max}(t) - n^k_{\min}(t) + 1} \leq \frac{m}{m-1} \leq \frac{1}{1-\varepsilon},$$

and

$$\frac{1}{n^k_{\max}(t)} \sum_{n=1}^{n^k_{\min}(t)-1} \mathbb{1}_A(n) \leq \frac{n^k_{\min}(t) - 1}{n^k_{\max}(t)} \leq \frac{1}{m} < \varepsilon.$$

Since $n^k_{\max}(t) > m$,

$$\frac{1}{n^k_{\max}(t)} \sum_{n=1}^{n^k_{\max}(t)} \mathbb{1}_A(n) \in]\alpha - \varepsilon, \alpha + \varepsilon[,$$

It follows that for $0 < t < \theta P_k(n_k)$, $\alpha - 2\varepsilon < \varphi(t) < (\alpha + \varepsilon)/(1-\varepsilon)$.

Hence we get the following estimation:

$$(\alpha - 2\varepsilon)I_3 < I_1 < \frac{\alpha + \varepsilon}{1-\varepsilon} I_3,$$

where

$$I_3 := \int_0^{\theta P_k(n_k)} \Big(n^k_{\max}(t) - n^k_{\min}(t) + 1\Big)\, dt.$$

Using (19) with $A = \mathbb{Z}_+$, we get

$$\int_0^{P_k(n_k)} \Big(n^k_{\max}(t) - n^k_{\min}(t) + 1\Big)\, dt = P_k(\mathbb{Z}_+) = 1,$$

and by the same argument as for the estimation of I_2, we have

$$\int_{\theta P_k(n_k)}^{P_k(n_k)} \Big(n^k_{\max}(t) - n^k_{\min}(t) + 1\Big)\, dt < \frac{1-\theta}{\theta} < \varepsilon.$$

hence $1 - \varepsilon < I_3 \leq 1$. □

Observe that the second condition in the lemma is crucial. Indeed, the sequence of binomial distributions of parameter $p \in]0,1[$ defined by

$$B_k(n) := \binom{k}{n} p^n (1-p)^{k-n}, 0 \le n \le k$$

satisfies the first assumption of the lemma. However there exists a set A possessing a natural density, but for which $B_k(A)$ fails to converge to this natural density as $k \to \infty$ (see [2], Theorem 3 page 25).

Proof of Theorem 5.2. Defining $P_k(n) := \mathbb{P}(N_k = n)$, we have to check the hypotheses of Lemma 5.3. We start by establishing an induction formula for $P_k(n)$. Recall that $\mathbb{P}(N_k = n) = 0$ if $n < k$, and that $\mathbb{P}(N_1 = n) = \mathbb{1}_{n=1}$. For $2 \le k \le n$, we decompose

$$\mathbb{P}(N_k = n) = \mathbb{P}(N_k = n, N_{k-1} = n-1) + \mathbb{P}(N_k = n, N_{k-1} \le n-2).$$

The first term in the RHS is $(1/n)\mathbb{P}(N_{k-1} = n-1)$. The second term can be written as

$$\begin{aligned}\mathbb{P}(N_k = n, N_{k-1} < n-1) &= \frac{1}{n}\left(1 - \frac{1}{n-1}\right)\mathbb{P}(|\boldsymbol{E}_{n-2}| = k-1) \\ &= \frac{n-2}{n}\frac{1}{n-1}\mathbb{P}(|\boldsymbol{E}_{n-2}| = k-1) \\ &= \frac{n-2}{n}\mathbb{P}(N_k = n-1).\end{aligned}$$

This yields, for any $2 \le k \le n$,

$$P_k(n) = \frac{n-2}{n}P_k(n-1) + \frac{1}{n}P_{k-1}(n-1). \tag{20}$$

For $2 \le k \le n-1$, dividing by $P_k(n-1)$ we obtain

$$\frac{P_k(n)}{P_k(n-1)} = 1 + \frac{1}{n}\Big(f_n(k) - 2\Big), \tag{21}$$

where

$$f_n(k) := \frac{P_{k-1}(n-1)}{P_k(n-1)}.$$

We prove by induction on $n \ge 4$ that $k \in \{2, \dots, n-1\} \mapsto f_n(k)$ is a non-decreasing function. Observe that $f_4(2) = 0$ and $f_4(3) = 1$, hence f_4 is non-decreasing. Assume that f_{n-1} is non-decreasing for some $n \ge 5$. Using (20), we get for $3 \le k \le n-1$

$$f_n(k) = \frac{\frac{n-3}{n-1}P_{k-1}(n-2) + \frac{1}{n-1}P_{k-2}(n-2)}{\frac{n-3}{n-1}P_k(n-2) + \frac{1}{n-1}P_{k-1}(n-2)}.$$

If $3 \le k \le n-2$, we get

$$f_n(k) = \frac{n-3+f_{n-1}(k-1)}{\frac{n-3}{f_{n-1}(k)} + 1}. \tag{22}$$

Hence, by induction, f_n is non-decreasing on $\{2, \dots, n-2\}$. Moreover, for $k = n-1$, since $P_{n-1}(n-2) = 0$,

$$f_n(n-1) = n-3+f_{n-1}(n-2) \ge n-3+f_{n-1}(n-3) \ge f_n(n-2).$$

Observe also that $f_n(2)=0$ and $f_n(n-1)=(n-3)(n-2)/2$ for all $n\geq 4$. Hence, for all $n\geq 5$, there exists an integer k_n such that $f_n(k)\leq 2$ for $2\leq k\leq k_n$ and $f_n(k)>2$ for $k>k_n$. Since $f_{n-1}(k-1)\leq f_{n-1}(k)$ for any $3\leq k\leq n-2$, we get by (22) that $f_n(k)\leq f_{n-1}(k)$, which proves that $n\mapsto k_n$ is non-decreasing.

For any fixed $k\geq 3$, let n_k be the smallest integer n such that $k_n\geq k$. By (21), $n\mapsto P_k(n)$ is non-decreasing upto n_k and non-increasing after n_k. Note that n_k exists, otherwise $n\mapsto P_k(n)$ would be non-decreasing, which is obviously impossible. Therefore, the first hypothesis of Lemma 5.3 is satisfied.

For all $k\geq 3$, observe that n_k is characterized by the following:

$$f_{n_k}(k)\leq 2, \quad \text{and } f_{n_k-1}(k)>2. \tag{23}$$

To check that (P_k) satisfies the second hypothesis, we need precise estimations of $f_n(k)$. We start by establishing a formula for $P_k(n)$. Observe that for all $1\leq j<n$, $\mathbb{P}(N_k=n|N_{k-1}=j)$ is equal to

$$\mathbb{P}(j+1\notin \boldsymbol{E},\dots n-1\notin \boldsymbol{E}, n\in \boldsymbol{E})=\frac{j}{j+1}\cdots\frac{n-2}{n-1}\frac{1}{n}=\frac{j}{n(n-1)}.$$

Hence, by conditioning, $P_k(n)=\mathbb{P}(N_k=n)$ can be rewritten as

$$\sum_{2\leq j_2<\dots<j_{k-1}\leq n-1}\mathbb{P}(N_k=n|N_{k-1}=j_{k-1})\mathbb{P}(N_{k-1}=j_{k-1}|N_{k-2}=j_{k-2})$$
$$\dots\quad \mathbb{P}(N_3=j_3|N_2=j_2)\mathbb{P}(N_2=j_2)$$

which yields

$$\begin{aligned}P_k(n)=&\sum_{2\leq j_2<\dots<j_{k-1}\leq n-1}\frac{j_{k-1}}{n(n-1)}\frac{j_{k-2}}{j_{k-1}(j_{k-1}-1)}\cdots\frac{j_2}{j_3(j_3-1)}\frac{1}{j_2(j_2-1)}\\ =&\frac{1}{n(n-1)}\sum_{2\leq j_2<\dots<j_{k-1}\leq n-1}\frac{1}{j_{k-1}-1}\cdots\frac{1}{j_3-1}\frac{1}{j_2-1}\\ =&\frac{1}{n(n-1)}\sum_{1\leq j_2<\dots<j_{k-1}\leq n-2}\frac{1}{j_2j_3\dots j_{k-1}}.\end{aligned}$$

We use this formula to estimate the denominator in the definition of $f_n(k)$:

$$\begin{aligned}P_k(n-1)=&\frac{1}{(n-1)(n-2)}\sum_{1\leq j_2<\dots<j_{k-1}\leq n-3}\frac{1}{j_2j_3\dots j_{k-1}}\\ =&\frac{1}{(n-1)(n-2)}\sum_{1\leq j_2<\dots<j_{k-2}\leq n-3}\frac{1}{j_2j_3\dots j_{k-2}}\quad C(j_2,\dots,j_{k-2}),\end{aligned}$$

where

$$C(j_2,\dots,j_{k-2}):=\frac{1}{k-2}\sum_{\substack{1\leq j\leq n-3\\ j\notin\{j_2,\dots,j_{k-2}\}}}\frac{1}{j}.$$

Observe that

$$\frac{1}{k-2}\sum_{k-2\leq j\leq n-3}\frac{1}{j}\leq C(j_2,\dots,j_{k-2})\leq\frac{1}{k-2}\sum_{1\leq j\leq n-3}\frac{1}{j},$$

which gives the following estimation

$$\frac{k-2}{\sum_{1\leq j\leq n-3}\frac{1}{j}}\leq f_n(k)=\frac{P_{k-1}(n-1)}{P_k(n-1)}\leq\frac{k-2}{\sum_{k-2\leq j\leq n-3}\frac{1}{j}}. \tag{24}$$

Let $m \geq 1$. Applying this estimation to f_{n_k-1} and f_{mn_k}, we get

$$0 \leq f_{n_k-1}(k) - f_{mn_k}(k) \leq \frac{k-2}{\sum_{k-2\leq j\leq n_k-4} \frac{1}{j}} - \frac{k-2}{\sum_{1\leq j\leq mn_k-3} \frac{1}{j}}.$$

The RHS of the above inequality can be written as a product AB, where

$$A := \frac{k-2}{\sum_{k-2\leq j\leq n_k-4} \frac{1}{j}} \quad \text{and } B := \frac{\sum_{1\leq j\leq mn_k-3} \frac{1}{j} - \sum_{k-2\leq j\leq n_k-4} \frac{1}{j}}{\sum_{1\leq j\leq mn_k-3} \frac{1}{j}}.$$

By (24) and (23), we get

$$\frac{k-2}{\sum_{1\leq j\leq n_k-3} \frac{1}{j}} \leq 2, \tag{25}$$

which ensures that A is bounded (say, by 4). Moreover, an easy computation shows that $B \sim \frac{\ln k}{\ln n_k}$, which goes to 0 as $k \to \infty$ by (25). Recalling that $f_{n_k-1}(k) > 2$ by (23), the above estimations prove the following property: For any $\varepsilon > 0$, for k large enough, $f_n(k) > 2 - \varepsilon$ for each $n_k \leq n \leq mn_k$. For such k, we get by (21)

$$\frac{P_k(mn_k)}{P_k(n_k)} = \prod_{n=n_k+1}^{mn_k} \frac{P_k(n)}{P_k(n-1)} = \prod_{n=n_k+1}^{mn_k} \left(1 + \frac{1}{n}(f_n(k) - 2)\right) \geq \left(1 - \frac{\varepsilon}{n_k}\right)^{(m-1)n_k}.$$

On the other hand, we know that $P_k(n_k) \geq P_k(mn_k)$, which proves that

$$\lim_{k\to\infty} \frac{P_k(mn_k)}{P_k(n_k)} = 1.$$

□

6. Open problems and discussion

6.1. **Connection with other densities.** We conjecture that the existence of local URI-density implies the existence of URI-density (and, in this case, that both coincide).

It is not clear either whether local URI-density is equivalent to the H^∞-density used by Flehinger. However we can provide the following interpretation of the H^∞-density in terms of our URI-set. Recall that in the proof of Theorem 3.3, we ordered the elements of $\boldsymbol{E}_n = \boldsymbol{E} \cap \{1, \ldots, n\}$ backwards:

$$\boldsymbol{E}_n = \left\{ Y_1^{(n)} > Y_2^{(n)} > \ldots > Y_{|\boldsymbol{E}_n|}^{(n)} = 1 \right\}.$$

Proposition 6.1 (Stochastic interpretation of H^∞-density)**.** *For $A \subset \mathbb{Z}_+$, A has H^∞-density α if and only if*

$$\lim_{k\to\infty} \liminf_{n\to\infty} \mathbb{P}\left(Y_k^{(n)} \in A\right) = \lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(Y_k^{(n)} \in A\right) = \alpha.$$

Proof. For each $n \geq 1$, we introduce a non-increasing sequence of random integers $(\widetilde{Y}_i^{(n)})_{i\geq 1}$, with the following distribution:

- $\widetilde{Y}_1^{(n)}$ is uniformly distributed in $\{1, \ldots, n\}$,
- conditionally to $\widetilde{Y}_1^{(n)}, \ldots, \widetilde{Y}_i^{(n)}$, the random variable $\widetilde{Y}_{i+1}^{(n)}$ is uniformly distributed in $\{1, \ldots, \widetilde{Y}_i^{(n)}\}$.

For $A \subset \mathbb{Z}_+$, we write $\mathbb{P}\left(\widetilde{Y}_k^{(n)} \in A\right)$ as

$$\sum_{1\le y_k\le y_{k-1}\le\cdots\le y_1\le n} \mathbb{1}_A(y_k)\mathbb{P}\left(\widetilde{Y}_k^{(n)} = y_k|\widetilde{Y}_{k-1}^{(n)} = y_{k-1}\right)\cdots \\ \mathbb{P}\left(\widetilde{Y}_2^{(n)} = y_2|\widetilde{Y}_1^{(n)} = y_1\right)\mathbb{P}\left(\widetilde{Y}_1^{(n)} = y_1\right),$$

which yields

$$\mathbb{P}\left(\widetilde{Y}_k^{(n)} \in A\right) = \frac{1}{n}\sum_{y_1=1}^{n}\frac{1}{y_1}\sum_{y_2=1}^{y_1}\cdots\frac{1}{y_{k-1}}\sum_{y_k=1}^{y_{k-1}}\mathbb{1}_A(y_k).$$

We recognize P_n^k used in the definition of the H^∞-density (see Section 3.3). Now, we observe that

$$\mathbb{P}\left(Y_k^{(n)} \in A\right) = \mathbb{P}\left(\widetilde{Y}_k^{(n)} \in A|D_k^n\right), \text{ where } D_k^n := \left\{\widetilde{Y}_1^{(n)} > \ldots > \widetilde{Y}_k^{(n)}\right\}.$$

It remains to prove that, for any fixed k, $\mathbb{P}(D_k^n) \to 1$ as $n \to \infty$. Fix $\epsilon > 0$ and choose $\delta > 0$ such that $(1-3\delta)^k > 1-\epsilon$. Observe that whenever $n > 1/\delta$, the proportion of integers $i \in \{1,\ldots,n\}$ such that $\delta < i/n < 1-\delta$ is larger than $1-3\delta$. Now, if $n > 1/\delta^k$, we have

$$\mathbb{P}\left(\frac{\widetilde{Y}_1^{(n)}}{n} \in]\delta, 1-\delta[, \frac{\widetilde{Y}_2^{(n)}}{\widetilde{Y}_1^{(n)}} \in]\delta, 1-\delta[, \ldots, \frac{\widetilde{Y}_k^{(n)}}{\widetilde{Y}_{k-1}^{(n)}} \in]\delta, 1-\delta[\right) > (1-3\delta)^k > 1-\epsilon.$$

Hence, $\mathbb{P}(D_k^n) > 1-\epsilon$. □

6.2. **Conditional URI-density.** Let P be a subset of $\mathbb{Z}_+$ with $\sum_{p\in P} 1/p = \infty$, so that the cardinality of $P \cap \boldsymbol{E}$ be almost surely infinite. We have two ways to define the *URI-density of A conditioned on P*. First, by averaging over $P \cap \boldsymbol{E}$, and consider (whenever it exists) the almost-sure limit of

$$\frac{\sum_{k=1}^n \mathbb{1}_{P\cap A}(N_k)}{\sum_{k=1}^n \mathbb{1}_P(N_k)}.$$

Second, by numbering the elements of $P = \{p_1 < p_2 < \cdots < p_n < \cdots\}$ and averaging over the random subset of P

$$\{p_{N_1} < p_{N_2} < \cdots < p_{n_k} < \cdots\},$$

that is by considering (whenever it exists) the almost-sure limit of

$$\frac{1}{n}\sum_{k=1}^n \mathbb{1}_A(p_{N_k}).$$

Question: are these two definitions equivalent?

6.3. **Asymptotic independence of successive elements in $\boldsymbol{E}$.** Another question concerning the URI-set $\boldsymbol{E}$ is the following: consider $A, B \subset \mathbb{Z}_+$, and assume that for both A and B we can define the density $d(A)$ and $d(B)$ (these could be the natural densities, the URI-densities, the H_∞-densities or maybe some other notions of densities). Under which condition do we have

$$\frac{1}{n}\sum_{k=1}^n \mathbb{1}_A(N_k)\mathbb{1}_B(N_{k+1}) \xrightarrow[n\to\infty]{\text{a.s.}} d(A)d(B) \ ?$$

We conjecture that it is true when both $d(A)$ and $d(B)$ are natural densities. However this is certainly not true for all A and B with URI-densities: As a counterexample, consider the set A of integers with leading digit 1 and the set B of integers with leading digit 9.

But what happens if for example $d(B)$ is the natural density whereas $d(A)$ is only the URI-density?

6.4. **Iterated URI-density.** Diaconis defined in [2] the *iterated log-density*: For a subset A of $\mathbb{Z}_+$, set

$$L(A,n,1) := \frac{1}{\ln n}\sum_{j=1}^{n}\frac{1}{j}\mathbb{1}_A(j),$$

and inductively for all $\ell \geq 2$:

$$L(A,n,\ell) := \frac{1}{\ln n}\sum_{j=1}^{n}\frac{1}{j}L(A,j,\ell-1).$$

The set A has *ℓ-th log-density* α if the limit of the above exists as $n \to \infty$ and is equal to α. In fact, this notion does not yield a new density, since Diaconis proved that A has an ℓ-th log-density if and only if A has a log-density (and then, of course, both coincide). Then he proposed to define the *L_∞-density*, which extends the log-density in much the same way as H_∞-density extends natural density: Consider $\lim_{\ell\to\infty} \liminf_{n\to\infty} L(A,n,\ell)$ and $\lim_{\ell\to\infty} \limsup_{n\to\infty} L(A,n,\ell)$. If the two limits are equal, call their common value the L_∞-density of A. As shown in [4], L_∞-density is strictly stronger than log-density.

It is natural in this context to study iterations of the URI-density, which can be defined as follows: Let $\left(\boldsymbol{E}^{(\ell)}\right)_{\ell\geq 1}$ be a sequence of independent URI-sets, and denote the random elements of $\boldsymbol{E}^{(\ell)}$ by $N_1^{(\ell)} = 1 < N_2^{(\ell)} < \cdots < N_k^{(\ell)} < \cdots$. We say that $A \subset \mathbb{Z}_+$ has *URI-density of order 2* equal to α if

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_A\left(N^{(1)}_{N^{(2)}_k}\right) \xrightarrow[n\to\infty]{a.s.} \alpha.$$

We define the *URI-density of order ℓ* in the same way, averaging along the subsequence $N^{(1)}_{N^{(2)}_{\iddots_{N^{(\ell)}_k}}}$.

We can also introduce the infinite iteration of the URI-method, considering almost-sure limsup and liminf in the above expressions, and see if they converge to the same limit as $\ell \to \infty$.

Although we have shown that URI-density and log-density coincide, it is not obvious if there are connections between iterated URI-density and iterated log-density. Can URI-densities of finite order ℓ be strictly stronger than URI-density? Can we compare the infinite iteration of both methods?

References

1. Franck Benford, *The law of anomalous numbers*, Proc. Amer. Phil. Soc. **78** (1938), 551–572.
2. Persi Diaconis, *Weak and strong averages in probability and the theory of numbers*, Ph.D. thesis, 1974.
3. ———, *The distribution of leading digits and uniform distribution* mod 1, Ann. Probability **5** (1977), no. 1, 72–81.

4. ———, *Examples for the theory of infinite iteration of summability methods*, Can. J. Math. **29** (1977), no. 3, 489–497.
5. Robert Luce Duncan, *Note on the initial digit problem*, Fibonacci Quat. **7** (1969), 474–475.
6. Betty J. Flehinger, *On the probability that a random integer has initial digit A*, Amer. Math. Monthly **73** (1966), 1056–1061.
7. Onésimo Hernández-Lerma and Jean Bernard Lasserre, *Markov chains and invariant probabilities*, Birkhäuser, 2003.
8. Élise Janvresse and Thierry de la Rue, *From uniform distributions to Benford's law*, J. Appl. Probab. **41** (2004), no. 4, 1203–1210.
9. Donald E. Knuth, *The art of computer programming. Vol. 2*, second ed., Addison-Wesley Publishing Co., Reading, Mass., 1981, Seminumerical algorithms, Addison-Wesley Series in Computer Science and Information Processing.
10. Serge Lang, *Introduction to transcendental numbers*, Addison-Wesley Publishing Company, 1966.
11. Simon Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881), 39–40.
12. Valentin V. Petrov, *Sums of independent random variables*, Springer-Verlag, 1975.
13. Michel Waldschmidt, *Séminaire sur les nombres transcendants 1972-73*, Publ. Math. Orsay **57** (1973).

Laboratoire de Mathématiques Raphaël Salem, UMR 6085 CNRS – Université de Rouen, Avenue de l'Université, B.P. 12, F76801 Saint-Étienne-du-Rouvray Cedex
E-mail address: Elise.Janvresse@univ-rouen.fr
E-mail address: Thierry.de-la-Rue@univ-rouen.fr