On the distribution of the digits of quotients of integers and primes

Alessandro Gambini, Remis Tonon, and Alessandro Zaccagnini

Abstract. We investigate the distribution of the digits of quotients of randomly chosen positive integers taken from the interval $[1, T]$, improving the previously known error term for the counting function as $T \to +\infty$. We also resolve some natural variants of the problem concerning points with prime coordinates and points that are visible from the origin.

1 General introduction

The first attempts to study number-theoretic problems by means of probabilistic methods date back to the 19th century and involve mainly two mathematicians: first Gauss, who was interested in the number of products of exactly k distinct primes below a certain threshold (and the solution of this problem for $k = 1$ is the famous prime number theorem, proved by Hadamard and de la Vallée Poussin, building on the ideas of Riemann, only in 1896; see [7, Chapter 8]) and then Cesàro, who showed in 1881 that the probability that two randomly chosen integers are coprime is $\frac{6}{\pi^2}$. In 1885, he then published a book [3] collecting some interesting problems from some articles he published in *Annali di matematica pura ed applicata*: the article we are interested in is *Eventualités de la division arithmétique*, which originally appeared as [2]. There he states that, dividing two random integers, the probability that the ith digit after the decimal point is r is given by

$$(1.1) \qquad \frac{1}{20} + \frac{10^i}{2} \int_0^1 \frac{1 - \varphi}{1 - \varphi^{10}} \varphi^{10^i - 1 + r} \, \mathrm{d}\varphi.$$

As a consequence, we have the somewhat surprising discovery that if one takes two "random" positive integers n and m and considers the distribution of the first decimal digit of their ratio $\frac{n}{m}$, it is slightly more likely that this turns out to be 0 rather than, say, 1.

Taking as a starting point, this result and the Benford law [1], which has a similar behavior with regard to the frequency distribution of the main digit in many real-life numerical datasets, Gambini, Mingari Scarpello, and Ritelli [8] studied the problem

in more detail, and recognized that the integral can be expressed in terms of the digamma function

$$\psi(x) := \frac{\Gamma'(x)}{\Gamma(x)}.$$

The representation

$$\psi(z) = -\gamma - \frac{1}{z} + \sum_{k \geq 1} \frac{z}{k(z+k)} = -\gamma + \sum_{k \geq 0}\left(\frac{1}{k+1} - \frac{1}{k+z}\right),$$

which is (5.7.6) of [10], led the authors to a different form for the integral in (1.1), which made them suspect that an elementary proof of the result was possible. Indeed, they were able to find it, and this is the starting point for our study.

In order to be more precise, we start giving some definitions. We consider a number base $b \geq 2$ and the corresponding set of digits $S_b = \{0, \ldots, b-1\}$. Given a positive real number x and a positive integer i, we are concerned with the ith digit to the right of the point in the representation in base b of x: we will call it $\phi(x; b; i)$. We remark that $\phi(x; b; i)$ can be computed by means of

$$\phi(x; b; i) = \lfloor b\{b^{i-1}x\}\rfloor = \lfloor b^i x - b\lfloor b^{i-1}x\rfloor\rfloor.$$

In fact, this formula is correct for any $i \in \mathbb{Z}$, with the obvious interpretation if $i < 0$. We recall that $\lfloor x\rfloor \in \mathbb{Z}$ and $\{x\} \in [0,1)$ denote the integer and the fractional part of the real number x, respectively, so that $x = \lfloor x\rfloor + \{x\}$.

With b and i as above and a digit $r \in S_b$, we also define $\Phi(T; b, r; i) := |\mathcal{A}(T; b, r; i)|$, where

$$\mathcal{A}(T; b, r; i) := \{(n, m) \in \mathbb{N}^2 \cap [1, T]^2 : \phi(n/m; b; i) = r\}.$$

Throughout the paper, for brevity, we often drop the dependency on b, r, and i of our functions, whenever there is no possibility of misunderstanding. We recall that Gambini et al. [8] implicitly obtained the asymptotic formula

$$\Phi(T; b, r; i) = c(b, r; i)T^2 + O\left(T^{3/2}\right),$$

as $T \to +\infty$, where b, r, and i are fixed. The authors considered couples of real numbers, both taken from $[1, T]$, whereas we are only interested in points with integral coordinates. The constant $c(b, r; i)$ is defined as an infinite series and can be expressed by means of the digamma function, as follows:

$$c(b, r; i) = \frac{1}{2b} + \frac{1}{2}b^i \int_0^1 \frac{1-\varphi}{1-\varphi^b}\varphi^{b^i+r-1}\,\mathrm{d}\varphi$$

(1.2)
$$= \frac{1}{2b} + \frac{1}{2}b^{i-1}\left(\psi\left(\frac{b^i+r+1}{b}\right) - \psi\left(\frac{b^i+r}{b}\right)\right).$$

This agrees with (1.1) by (5.9.16) of [10]. We remark here that this problem is appropriately situated among the classical problems of counting lattice points that belong to some region of the plane. One of the most famous of these results is Minkowski's

theorem, which states that every closed convex set in $\mathbb{R}^n$ that is symmetric with respect to the origin and with volume greater than 2^n contains at least one point with integer coordinates distinct from the origin. This theorem, proved in 1889, gave birth to a new branch of number theory: the Geometry of Numbers. Another famous result in this direction is Pick's theorem; proved in 1899 (see [11]), it relates the area of a simple polygon with integer coordinates with the number of lattice points in its interior and the number of lattice points on its boundary. Finally, it is mandatory to refer to two of the most famous problems in analytic number theory, which are related to ours: Gauss' circle problem and Dirichlet's divisor problem (see [9, Chapter 18]). Both of them deal with counting points with integer coordinates belonging to some region delimited by conics: a circle and a hyperbola. The two mathematicians were able to provide a formula with the area of the region as a main term plus some error due to the integer points near the boundary. For example, Gauss proved that in the circle of radius r, there are

$$\pi r^2 + E(r)$$

points with integer coordinates, where $|E(r)| \leq 2\sqrt{2}\pi r$. By this simple formulation, one could think that guessing the right order of magnitude of the error term should not be a hard problem; actually, although many (slow) improvements have been done, both problems remain open.

The interest in the random integer quotients also embraces other fields in analytic number theory. Recently, there have been several papers dealing with the cardinalities of A/A for subsets A of the set of the first n positive integers getting general lower and upper bounds (see Cilleruelo and Guijarro-Ordóñez [4] and Cilleruelo, Ramana, and Ramaré [5, 6]). Another line of research has to do with the search for prime numbers with a positive proportion of preassigned digits in base b and the estimation of the number of these prime numbers (see Swaenepoel [12]).

2 Results

After this excursus, which gives some motivation to our research for a better error term, we turn to our results. In this paper, we introduce some number-theoretic devices which allow us to improve upon the result by Gambini et al., and specifically to obtain a better error term. In all statements, we consider b, r, and i fixed, so that, here and throughout the paper, implicit constants may depend on them. We also recall that $c(b, r; i)$ is the constant defined in (1.2).

Theorem 2.1   *As $T \to +\infty$, we have*

$$\Phi(T; b, r; i) = c(b, r; i) T^2 + O\left(T \log(T)\right).$$

Our improvement stems largely from the fact that we evaluate more carefully the error terms arising from computing ratios of integers with the desired digit and that we introduce a variable threshold, to be chosen at the end of the proof, which allows us to ignore some points in $\mathcal{A}(T; b, r; i)$.

In the second part of the paper, we deal with a variation of the same problem: we consider the case of primes. We let $\mathfrak{P}$ denote the set of positive prime integers. For the sake of clarity, for $X \geq 2$, we let

$$(2.1) \qquad \mathcal{E}(X) := \sup_{2 \leq x \leq X} \left| \pi(x) - \mathrm{li}(x) \right|,$$

so that $\mathcal{E}$ is nonnegative and increasing, and can be bounded by means of the Prime Number Theorem, which is Lemma 5.1 below.

Theorem 2.2   *In the case of primes, as $T \to +\infty$, we have*

$$\sum_{(p,q) \in \mathcal{A}(T;b,r;i) \cap \mathfrak{P}^2} \log(p) \log(q) = c(b,r;i) T^2 + O\left(T\mathcal{E}(T) \log(T) \right).$$

It is also possible to study the corresponding problem where only the numerator or the denominator is restricted to being a prime, and the result is similar. See Section 5 for some comments and details on the error term in Theorem 2.2.

We also tackle another variant of this problem where we consider only points that are visible from the origin, casting out multiplicities. We obtain our last result as a corollary of Theorem 2.1, via Möbius inversion.

Theorem 2.3   *As $T \to +\infty$, we have*

$$\sum_{\substack{(n,m) \in \mathcal{A}(T;b,r;i) \\ (n,m)=1}} 1 = \frac{c(b,r;i)}{\zeta(2)} T^2 + O\left(T \log^2(T) \right),$$

where ζ denotes the Riemann ζ-function.

3   Basic strategy of the proofs

We follow [8] quite closely. The proofs share some common features, and it is probably clearer if we deal with them at the outset. We remark that we may assume that T is an integer, because the total error involved in changing T by a bounded amount is small: see the end of this section. We first decompose the set $\mathcal{A}(T;b,r;i)$ as an appropriate union of sets. For $r \in S_b$, we define

$$y_k = y_k(b,r;i) := \frac{b^i}{bk+r} = \frac{b^{i-1}}{k + \frac{r}{b}},$$

$$x_k = y_k(b,r+1;i) := \frac{b^i}{bk+r+1} = \frac{b^{i-1}}{k + \frac{r+1}{b}}.$$

With these definitions and Figure 1 in mind, we write

$$\mathcal{A}_k(T) = \mathcal{A}_k(T;b,r;i) = \left\{ (n,m) \in \mathbb{N}^2 \cap [1,T]^2 : \frac{n}{m} \in \left[\frac{kb+r}{b^i}, \frac{kb+r+1}{b^i} \right) \right\}$$

$$(3.1) \qquad\qquad = \left\{ (n,m) \in \mathbb{N}^2 \cap [1,T]^2 : m \in (nx_k, ny_k] \right\}$$

$$(3.2) \qquad\qquad = \left\{ (n,m) \in \mathbb{N}^2 \cap [1,T]^2 : n \in [m/y_k, m/x_k) \right\},$$
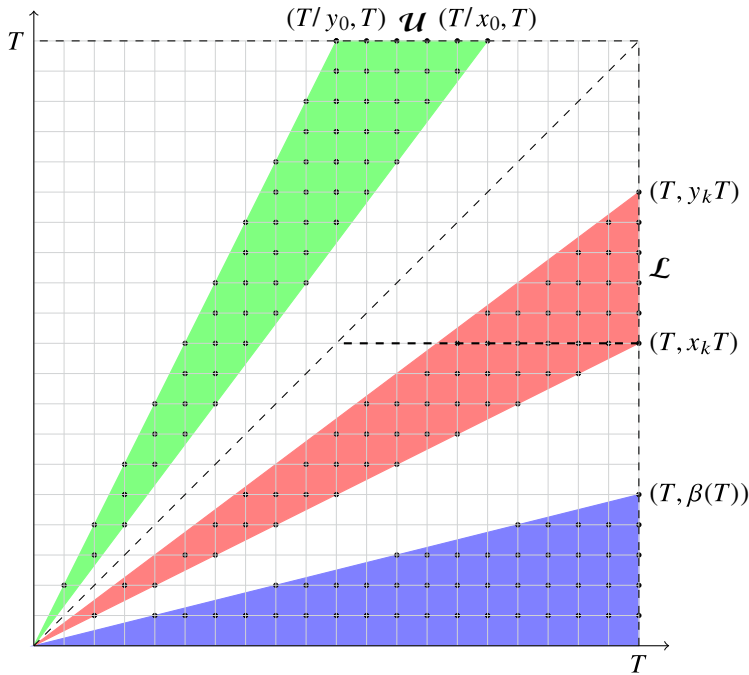
Figure 1: How to split the sets: here, we illustrate the case $i = 1$. The set $\mathcal{U}$ is a triangle. The set $\mathcal{L}$ is an infinite union of triangles; we estimate trivially the contribution from triangles in the shaded region at the bottom. In the paper, we consistently use n for the abscissa and m for the ordinate of the points.

so that $\mathcal{A}(T; b, r; i) = \bigcup_{k \geq 0} \mathcal{A}_k(T; b, r; i)$. This is easily checked using the definition of ϕ. The sets $\mathcal{A}_k(T)$ are pairwise disjoint and correspond to the lattice points contained in triangles with a vertex at the origin and the other vertices either on the segment $[1, T] \times \{T\}$, when $0 \leq k < b^{i-1}$ (i.e., $n < m$), or on $\{T\} \times [1, T]$, when $k \geq b^{i-1}$ (i.e., $n \geq m$), provided that k satisfies (3.3) below. Therefore, we split the infinite union above accordingly as

$$\mathcal{U}(T; b, r; i) := \bigcup_{k=0}^{b^{i-1}-1} \mathcal{A}_k(T; b, r; i) \quad \text{and} \quad \mathcal{L}(T; b, r; i) := \bigcup_{k \geq b^{i-1}} \mathcal{A}_k(T; b, r; i).$$

At this point, it is worth looking back at the definition of $\mathcal{A}_k(T)$. When k is such that

$$\frac{1}{y_k} = \frac{kb + r}{b^i} > T,$$

the set $\mathcal{A}_k(T)$ is empty. This means that we can bound the range for k, taking

(3.3) $$k \leq b^{i-1}T - \frac{r}{b} \leq b^{i-1}T.$$

Hence, we have that

$$\mathcal{L}(T; b, r; i) = \bigcup_{b^{i-1} \le k \le b^{i-1} T} \mathcal{A}_k(T; b, r; i).$$

For k, above a certain threshold depending on T, it is difficult to evaluate the cardinality of $\mathcal{A}_k(T)$ exactly: this will be the source of our first error term.

We notice here, even though we will need this later, that for $k \ge 1$ and $r \in S_b \cup \{b\}$, we have

$$\frac{1}{3k} \le \frac{1}{k} \cdot \frac{1}{2 + 1/b} = \frac{1}{k} \cdot \frac{1}{1 + \frac{b+1}{b}} < \frac{1}{k} \cdot \frac{1}{1 + \frac{r+1}{bk}} = \frac{1}{k + \frac{r+1}{b}} < \frac{1}{k + \frac{r}{b}} \le \frac{1}{k},$$

so that

$$\begin{cases} x_k \asymp b^{i-1}/k, \\ y_k \asymp b^{i-1}/k. \end{cases}$$

3.1 Weights

In order to treat our problems in a unified fashion, we introduce weights associated to points with integral coordinates in $[1, T]^2$. We will eventually choose the following weights: $\omega(n, m) := 1$ for all n and m in the problem with all integers;

$$\omega(n, m) := \begin{cases} \log(n) \log(m) & \text{if } n \text{ and } m \text{ are both prime numbers,} \\ 0 & \text{otherwise,} \end{cases}$$

in the problem with primes; and

$$\omega(n, m) := \begin{cases} 1 & \text{if } (n, m) = 1, \\ 0 & \text{otherwise,} \end{cases}$$

in the problems with "reduced" couples. We have to evaluate

$$\sum_{(n,m) \in \mathcal{U}(T;b,r;i)} \omega(n, m) \qquad \text{and} \qquad \sum_{(n,m) \in \mathcal{L}(T;b,r;i)} \omega(n, m).$$

With this choice of weights, we see that the error involved in changing T to the nearest integer is $O(T)$ in the case of integers and $O(T \log(T))$ in the case of primes.

3.2 The contribution from the set $\mathcal{U}(T; b, r; i)$

We refer to Figure 1, and remark that

$$\sum_{(n,m) \in \mathcal{U}(T;b,r,i)} \omega(n, m) = \sum_{k=0}^{b^{i-1}-1} \sum_{(n,m) \in \mathcal{A}_k(T;b,r,i)} \omega(n, m)$$

(3.4)
$$= \sum_{k=0}^{b^{i-1}-1} \sum_{m \in [1,T]} \sum_{n \in [m/y_k, m/x_k)} \omega(n, m).$$

3.3 The contribution from the set $\mathcal{L}(T; b, r; i)$

To a first approximation, $|\mathcal{A}_k(T)|$ is the area of the corresponding triangle, with an error proportional to the perimeter, that is $O(T)$. However, as $T \to +\infty$, the number of such triangles tends to infinity as well, and their area may be small, and we have to be much more careful than this. We choose an appropriate function $\beta = \beta(T)$ and estimate trivially the contribution from k satisfying (3.3) with $k > T/\beta(T)$. It is

$$(3.5) \qquad \ll \sum_{n \leq T} \sum_{m \leq n\beta(T)/T} \omega(n, m) \leq T\beta(T) \max_{n \leq T} \max_{m \leq n\beta(T)/T} \omega(n, m).$$

Recalling (3.1), we see that the contribution from $\mathcal{A}_k(T; b, r; i)$ is

$$(3.6) \qquad \begin{aligned} &\sum_{m \leq x_k T} \sum_{n \in [m/y_k, m/x_k)} \omega(n, m) + \sum_{x_k T < m \leq y_k T} \sum_{n \in [m/y_k, T]} \omega(n, m) \\ &= \sum_{m \leq x_k T} \sum_{n \in [m/y_k, m/x_k)} \omega(n, m) + \sum_{x_k T/y_k \leq n \leq T} \sum_{m \in [x_k T, ny_k]} \omega(n, m). \end{aligned}$$

We find it convenient to write these quantities as above in order to keep error terms under control.

3.4 The value of the constants $c(b, r; i)$

The following lemma will take care of the constant appearing in the main terms.

Lemma 3.1   *For $X \to +\infty$, we have*

$$\sum_{b^{i-1} \leq k \leq X} (y_k - x_k) = \sum_{k \geq b^{i-1}} \frac{b^i}{(bk + r)(bk + r + 1)} + O\left(X^{-1}\right).$$

Furthermore,

$$(3.7) \qquad \sum_{k \geq b^{i-1}} \frac{b}{(bk + r)(bk + r + 1)} = \psi\left(\frac{b^i + r + 1}{b}\right) - \psi\left(\frac{b^i + r}{b}\right).$$

Proof   Because $y_k - x_k \leq k^{-2}$, the error term arising from extending the sum over k to all positive integers is $\ll X^{-1}$. The results follow from the same property of the digamma function quoted above: see [10, equation (5.7.6)]. ∎

4 The proof of Theorem 2.1 for integers

Our improvement over the previous results depends on the fact that we manage to choose a specific threshold for k: above it, but below (3.3), we will estimate $|\bigcup_k \mathcal{A}_k(T)|$ trivially, while for k less than it, we will evaluate $|\mathcal{A}_k(T)|$ more precisely.

4.1 Estimate of $\mathcal{U}(T)$

Now, let us consider $\mathcal{U}(T)$. By (3.4), we have

$$|\mathcal{U}(T;b,r;i)| = \sum_{k=0}^{b^{i-1}-1} |\mathcal{A}_k(T;b,r;i)| = \sum_{k=0}^{b^{i-1}-1} \sum_{m\in[1,T]} \left(\frac{m}{x_k} - \frac{m}{y_k} + O\,(1)\right)$$

$$= b^{i-1} \sum_{m\in[1,T]} \frac{m}{b^i} + O\,(T).$$

If $T \in \mathbb{N}$, we have

$$(4.1) \qquad |\mathcal{U}(T;b,r;i)| = \frac{T(T+1)}{2b} + O\,(T) = \frac{T^2}{2b} + O\,(T).$$

4.2 Estimate of $\mathcal{L}(T)$

Using (3.6), we see that the number of lattice points in $\mathcal{A}_k(T;b,r;i)$ is

$$(4.2) \qquad = \sum_{n\leq x_k T} \left(\left[\frac{n}{x_k}\right] - \left[\frac{n}{y_k}\right]\right) + \sum_{x_k T/y_k \leq n \leq T} \left([ny_k] - [x_k T]\right).$$

The first term in (4.2) is

$$= \sum_{n\leq x_k T} \left(\frac{n}{x_k} - \frac{n}{y_k}\right) + O\,(x_k T) = \left(\frac{1}{x_k} - \frac{1}{y_k}\right) \sum_{n\leq x_k T} n + O\left(\frac{T}{k}\right)$$

$$(4.3) \qquad\qquad\qquad = \frac{y_k - x_k}{x_k y_k} \sum_{n\leq x_k T} n + O\left(\frac{T}{k}\right)$$

$$= \frac{y_k - x_k}{2x_k y_k} \left((x_k T)^2 + O\,(x_k T)\right) + O\left(\frac{T}{k}\right)$$

$$(4.4) \qquad\qquad\qquad = \frac{y_k - x_k}{2y_k} x_k T^2 + O\left(\frac{T}{k}\right).$$

We further rewrite the last summand in (4.2) as

$$\sum_{x_k T/y_k \leq n \leq T} \left([ny_k] - [x_k T]\right) = \sum_{x_k T/y_k \leq n \leq T} \left(ny_k - \{ny_k\}\right) - [x_k T]\left(T - \left[\frac{x_k T}{y_k}\right]\right)$$

$$= I_1 - I_2,$$

say. We have

$$I_1 = \sum_{x_k T/y_k \leq n \leq T} ny_k + O\left(1 + \left(1 - \frac{x_k}{y_k}\right)T\right) = y_k \sum_{x_k T/y_k \leq n \leq T} n + O\left(1 + \frac{T}{k}\right)$$

$$(4.5) \quad = \frac{1}{2}y_k\left(T^2 - \frac{x_k^2}{y_k^2}T^2 + O\,(T)\right) + O\left(1 + \frac{T}{k}\right) = \frac{1}{2}y_k\left(1 - \frac{x_k^2}{y_k^2}\right)T^2 + O\left(1 + \frac{T}{k}\right).$$

We also have

$$I_2 = [x_k T]\left(T - \left[\frac{x_k T}{y_k}\right]\right) = (x_k T + O(1))\left(T - \frac{x_k T}{y_k} + \left\{\frac{x_k T}{y_k}\right\}\right)$$

$$= \frac{(y_k - x_k)x_k}{y_k}T^2 + O\left(\frac{y_k - x_k}{y_k}T\right) + O(x_k T)$$

(4.6) $$= \frac{(y_k - x_k)x_k}{y_k}T^2 + O\left(\frac{T}{k}\right).$$

Summing up (4.2) and (4.4)-(4.6), we have

$$|\mathcal{A}_k(T; b, r; i)| = \frac{y_k - x_k}{2y_k}x_k T^2 + \frac{1}{2}y_k\left(1 - \frac{x_k^2}{y_k^2}\right)T^2 - \frac{(y_k - x_k)x_k}{y_k}T^2 + O\left(\frac{T}{k}\right)$$

$$= \frac{y_k - x_k}{2y_k}\left(x_k + (y_k + x_k) - 2x_k\right)T^2 + O\left(\frac{T}{k}\right)$$

(4.7) $$= \frac{1}{2}(y_k - x_k)T^2 + O\left(\frac{T}{k}\right).$$

We finally sum (4.7) over $b^{i-1} \le k \le b^{i-1}T$, obtaining

$$\sum_{b^{i-1} \le k \le b^{i-1}T} |\mathcal{A}_k(T; b, r; i)| = \frac{1}{2}T^2 \sum_{b^{i-1} \le k \le b^{i-1}T} (y_k - x_k) + O\left(\sum_{b^{i-1} \le k \le b^{i-1}T} \frac{T}{k}\right).$$

Using Lemma 3.1 with $X = b^{i-1}T$, we see that the error term arising from the completion of the series is $\ll T^{-1}$. The other error term contributes $\ll T \log T$. Hence,

$$\sum_{b^{i-1} \le k \le b^{i-1}T} |\mathcal{A}_k(T; b, r; i)| = \frac{1}{2}T^2 \sum_{k \ge b^{i-1}} (y_k - x_k) + O(T \log(T)).$$

The proof is complete by (3.7) and (4.1).

5 Primes with weights: approach via θ

In this section, we prove Theorem 2.2. With notation as in Section 3 and splitting the sets in the same way, we set

$$N_k(T) := \sum_{(p,q) \in \mathcal{A}_k(T)} (\log p)(\log q).$$

We write (p, q) for prime numbers in place of (n, m) and use logarithmic weights in order to exploit the linearity of the main term of the Chebyshev θ-function. This is critical in order to avoid the introduction of special functions whose behavior is hard to estimate carefully over the range of values of k that we need. We give more details at the end of this section.

Lemma 5.1 (Prime Number Theorem)   *There exists a positive constant c such that*

$$\pi(x) = \mathrm{li}(x) + O\left(xe^{-c\sqrt{\log x}}\right), \quad as\ x \to +\infty.$$

If the Riemann Hypothesis is true, then the error term on the right-hand side may be replaced by $O\left(x^{1/2}\log(x)\right)$.

We will need repeatedly the following simple lemma, whose proof by partial summation is straightforward.

Lemma 5.2   *Let $f\colon\mathbb{R}^+ \to \mathbb{R}^+$ be a smooth increasing function, and let $\mathcal{E}$ be defined by* (2.1). *Then,*

$$\sum_{p\leq X} f(p) = \int_2^X \frac{f(t)}{\log(t)}\,\mathrm{d}t + O\left(f(X)\mathcal{E}(X)\right), \qquad \text{as } X \to +\infty.$$

5.1   The contribution from the set $\mathcal{U}(T)$

Using Lemma 5.2 with $f(t) = t\log(t)$ and recalling (3.2) and (3.4), for $0 \leq k < b^{i-1}$, we have to evaluate

$$\sum_{q\leq T}\left(\theta\left(\frac{q}{x_k}\right) - \theta\left(\frac{q}{y_k}\right)\right)\log(q) = \frac{1}{b^i}\sum_{q\leq T} q\log(q) + \sum_{p\leq T} O\left(\mathcal{E}(p)\log(p)\right)$$

$$= \frac{1}{b^i}\int_2^T t\,\mathrm{d}t + O\left(T\mathcal{E}(T)\log(T)\right) + O\left(\theta(T)\mathcal{E}(T)\right)$$

$$= \frac{1}{2b^i}T^2 + O\left(T\mathcal{E}(T)\log(T)\right).$$

Summing over the values of k mentioned above, we find that the total contribution from the set $\mathcal{U}(T)$ is

(5.1)
$$\frac{1}{2b}T^2 + O\left(T\mathcal{E}(T)\log(T)\right).$$

5.2   The contribution from the set $\mathcal{L}(T)$

Using (3.6), we see that we have to evaluate

$$\sum_{q\leq x_k T}\left(\theta\left(\frac{q}{x_k}\right)\log(q) - \theta\left(\frac{q}{y_k}\right)\log(q)\right) + \sum_{x_k T/y_k\leq p\leq T}\left(\theta(py_k) - \theta(x_k T)\right)\log(p)$$

for $b^{i-1} \leq k \leq T/\beta(T)$. Using Lemma 5.2 with $f(t) = t\log(t)$, we deduce that the first summand is

$$\sum_{q\leq x_k T}\left(\theta\left(\frac{q}{x_k}\right)\log(q) - \theta\left(\frac{q}{y_k}\right)\log(q)\right)$$

$$= \left(\frac{1}{x_k} - \frac{1}{y_k}\right)\sum_{q\leq x_k T} q\log(q) + O\left(\sum_{q\leq x_k T}\mathcal{E}\left(\frac{q}{x_k}\right)\log(q)\right)$$

$$= \Big(\frac{1}{x_k} - \frac{1}{y_k}\Big) \int_2^{x_k T} t \, dt + O\left(\theta(x_k T)\mathcal{E}(T)\right)$$

$$= \frac{1}{2} x_k^2 T^2 \Big(\frac{1}{x_k} - \frac{1}{y_k}\Big) + O\left(\frac{T}{k}\mathcal{E}(T)\right).$$

The second summand is

$$\sum_{x_k T/y_k \leq p \leq T} \left(\theta(p y_k) - \theta(x_k T)\right)\log(p) = \sum_{x_k T/y_k \leq p \leq T} \left(y_k p - x_k T\right)\log(p)$$

$$+ O\left(\sum_{x_k T/y_k \leq p \leq T} \mathcal{E}(p y_k)\log(p)\right).$$

Using again Lemma 5.2, we see that the main term is

$$y_k \int_{x_k T/y_k}^{T} t \, dt - x_k T \int_{x_k T/y_k}^{T} dt + O\left(y_k T \mathcal{E}(T)\right) = \frac{1}{2} y_k \Big(1 - \frac{x_k}{y_k}\Big)^2 T^2 + O\left(\frac{T}{k}\mathcal{E}(T)\right).$$

The error term is

$$\sum_{x_k T/y_k \leq p \leq T} \mathcal{E}(p y_k)\log(p) \ll \left(\pi(T) - \pi\Big(\frac{x_k}{y_k}T\Big)\right)\mathcal{E}(T y_k)\log(T)$$

$$\ll \frac{(1 - x_k/y_k)T}{\log(T/k)}\mathcal{E}(T y_k)\log(T) \ll \frac{T}{k}\mathcal{E}(T y_k),$$

by the Brun–Titchmarsh inequality.

Summing up, the total contribution of the main terms is

$$\frac{1}{2}T^2 \sum_{b^{i-1} \leq k \leq T/\beta(T)} \left(x_k^2\Big(\frac{1}{x_k} - \frac{1}{y_k}\Big) + y_k\Big(1 - \frac{x_k}{y_k}\Big)^2\right)$$

$$= \frac{1}{2}T^2 \sum_{b^{i-1} \leq k \leq T/\beta(T)} (y_k - x_k)$$

$$(5.2) \qquad = \frac{b^{i-1}}{2}T^2\left(\psi\Big(\frac{b^i + r + 1}{b}\Big) - \psi\Big(\frac{b^i + r}{b}\Big)\right) + O\left(T\beta(T)\right),$$

by Lemma 3.1. The total error term is

$$(5.3) \qquad \ll T\mathcal{E}(T) \sum_{b^{i-1} \leq k \leq T/\beta(T)} \frac{1}{k} \ll T\mathcal{E}(T)\log(T).$$

In order to complete the proof, we just collect (5.1)-(5.3) and choose $\beta(T) = T^{1/2}$.

5.3 Comments on the choice of weights

We remark that using the characteristic function of the primes as the choice of weight in this problem leads to a number of technical complications. If we insist on counting couples of primes *without* weights, we would find a much weaker result, with the error term smaller than the main term just by a factor $\log(T)$.

6   On points which are visible from the origin

We now deal with a similar problem, where we only count points that are visible from the origin, that is, couples (n, m) with $(n, m) = 1$. We could use the familiar device of writing the characteristic function of such couples by means of the Möbius function μ. As we said in Section 3, we could write,

$$\omega(n, m) := \sum_{d \mid (n,m)} \mu(d) = \begin{cases} 1 & \text{if } (n, m) = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and then proceed along the same lines as in Section 3. This approach does not lead to any particular improvement, because the extra sum on d generates an additional logarithm in the error term. Nevertheless, numerical data suggest a slight regularization in this case (see Section 7).

6.1   Deduction of Theorem 2.3 from Theorem 2.1

We notice that

$$
\begin{aligned}
&\left|\{(n, m) \in [1, T]^2 : \phi(n/m) = r\}\right| \\
&\qquad = \sum_{d \leq T} \left|\{(n, m) \in [1, T]^2 : (n, m) = d \wedge \phi(n/m) = r\}\right| \\
&\qquad = \sum_{d \leq T} \left|\{(n, m) \in [1, T/d]^2 : (n, m) = 1 \wedge \phi(n/m) = r\}\right|.
\end{aligned}
$$

Hence, by the Möbius inversion formula, we have

$$
\begin{aligned}
&\left|\{(n, m) \in [1, T]^2 : (n, m) = 1 \wedge \phi(n/m) = r\}\right| \\
&\qquad = \sum_{d \leq T} \mu(d) \left|\{(n, m) \in [1, T/d]^2 : \phi(n/m) = r\}\right| \\
&\qquad = \sum_{d \leq T} \mu(d) c(b, r; i) \frac{T^2}{d^2} + \sum_{d \leq T} O\left(\frac{T}{d}(1 + \log(T/d))\right) \\
&\qquad = c(b, r; i) T^2 \sum_{d \leq T} \frac{\mu(d)}{d^2} + \sum_{d \leq T} O\left(\frac{T}{d}(1 + \log(T/d))\right).
\end{aligned}
$$

The leading term is

$$\frac{c(b, r; i)}{\zeta(2)} T^2 + O(T).$$

The first error term contributes $O(T \log(T))$. The second error term is $O\left(T \log^2(T)\right)$.

7 Remarks and numerical data

7.1 The limit of the method

Given an integer $b \geq 2$, pick a digit $r \in S_b$. We choose $i = 1$ for simplicity. We want to count the number of lattice points lying on the boundary of the region $\mathcal{A}_k(T; b, r; 1)$. We start with estimating

$$(7.1) \qquad \left| \left\{ (n, m) \in [1, T]^2 : (n, m) = 1 \wedge b \left\{ \frac{n}{m} \right\} = r \right\} \right|.$$

We obviously have

$$b \left\{ \frac{n}{m} \right\} = r \qquad \Longleftrightarrow \qquad \left\{ \frac{n}{m} \right\} = \frac{r}{b} = \frac{r/(b, r)}{b/(b, r)}.$$

The fractions at far left and far right are reduced, and this forces

$$m = \frac{b}{(b, r)} \qquad \text{and} \qquad n \equiv \frac{r}{(b, r)} \bmod m.$$

Let $m_1 = b/(b, r)$. Hence, the cardinality of the set in (7.1) is $\sim T/m_1$. A similar argument shows that

$$\left| \left\{ (n, m) \in [1, T]^2 : (n, m) = d \wedge b \left\{ \frac{n}{m} \right\} = r \right\} \right| \sim \frac{T}{d m_1},$$

for $d \leq T/m_1$. Hence, the total number of lattice points on the boundary of the set $\mathcal{A}_k$ is

$$(7.2) \qquad \sim \frac{T}{m_1} \sum_{d \leq T/m_1} \frac{1}{d} \sim \frac{(b, r)}{b} T \log(T).$$

This seems to be the limit of our method for the proof of Theorem 2.1.

In the next paragraph, we will show that for "small" T, actually there seems to be a bit of difference in the convergence rate and this is evident with peaks when $\frac{(b,r)}{b}$ is large.

7.2 Numerical data

We collect here some histograms, created using Wolfram Mathematica, obtained with the numerical computations that we performed for the problem of digits. Some variants were considered trying to understand if a different counting function could help to regularize the problem and reduce the error term. In every histogram, on the x-axis, we have the digit r, and the height of the bar represents $\Phi(T; b, r; 1)$ (in the cases with $i > 1$, the difference is less noticeable, so we did not report examples) normalized by dividing by $\sum_{r=0}^{b-1} \Phi(T; b, r; 1)$. We recall that $\Phi(T; b, r; 1) := |\mathcal{A}(T; b, r; 1)|$. The continuous line, instead, is the graph of $c(b, r; i)$ as a function of r; we indicated with a dot the values corresponding to integer values of r, which are the ones appearing in the theorems.
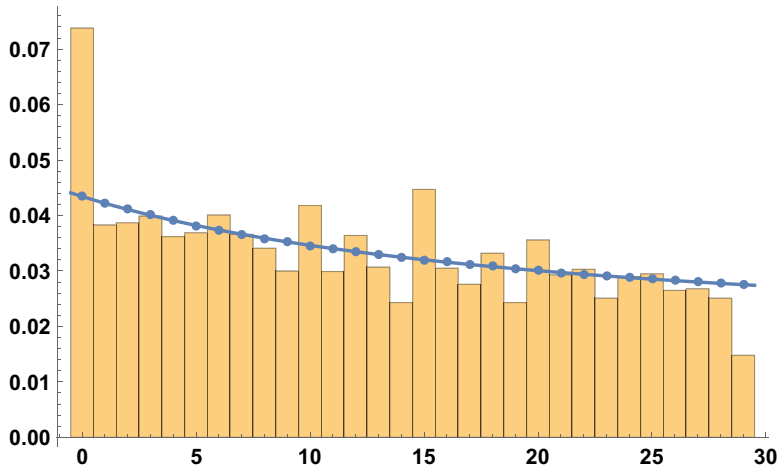
Figure 2: The histogram for $|\mathcal{A}(100; 30, r; 1)|$.

We do not report cases with a very large sample ($T = 10,000$), because in these circumstances, the histogram is really close to the expected value. But for smaller T, we can see some phenomenon taking place: for $T = 100$, the irregularities seem all but random. If you look at the case $b = 30$ (see Figure 2), which has many divisors, some numerical values noticeably exceed the expected ones. We do not report the plots where b is prime, because they actually show the behavior we expect and that we have calculated in (7.2): an anomalous peak in 0 and a monotonically decreasing trend.

Actually, the main problem for small T is a simple fact of multiplicities: fractions with small numerator and denominator (like 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{2}{3}$, etc.) will be counted many times: in both of them, we can recognize high counting values for the fractions that we have just mentioned.

To avoid this phenomenon, we can just count the fractions with multiplicity one, which means taking just the reduced ones: in Figure 3, we represent $\Phi(T; b, r; 1)$, with $(n, m) = 1$.

There is a regularization, but some other phenomenon emerged: it seems that the divisors and, in general, the numbers with prime factors in common with 30 tend to be have higher values. The explanation for this lies in the discontinuity of the system of digits: if we perturbed just a bit a number that has no digits to the right of the first one, we could reduce its first digit by one. To be more formal, if a rational number admits a finite representation (when, after reducing the fraction, the denominator divides the base), it admits also a periodic infinite one. Just to make an example, if we think about the base 10, we have $1 = 0.\bar{9}$.

So, in our case, any number that admits any ambiguity in its first digit in base b should morally be divided into the two digits with an equal weight: assigning half weight to two digits if $b\{n/m\} \in \mathbb{Z}$, the rate of convergence is very good already for small values of T. See Figure 4.
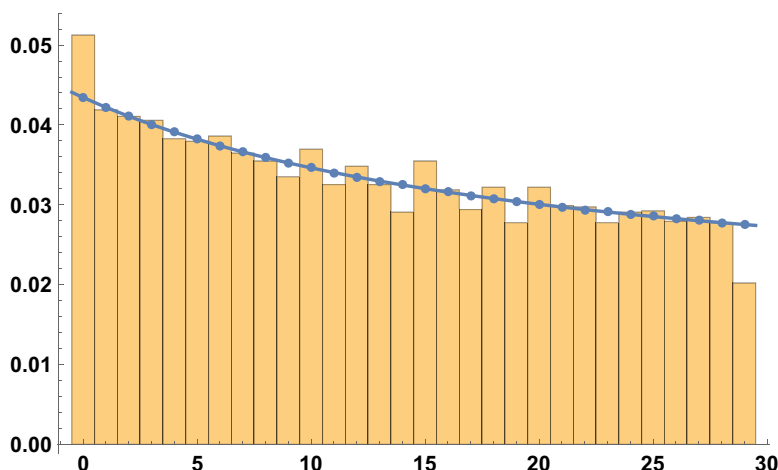
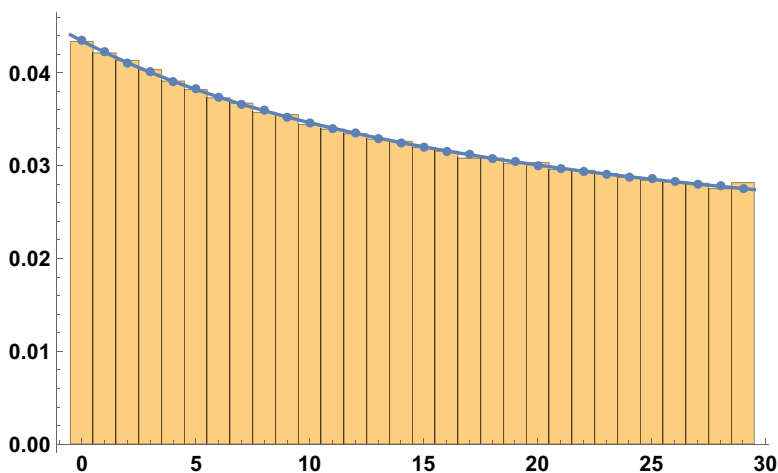*Figure 3*: The histogram for $|\mathcal{A}(100; 30, r; 1)|$ (with $(n, m) = 1$).

Figure 4: The histogram for $\mathcal{A}(100; 30, r; 1)$ with $(n, m) = 1$ and assigning half weight to two digits if $b\{n/m\} \in \mathbb{Z}$.

7.3 Primes

We made similar computations also in the case of prime numbers. Two positive primes are not coprime if and only if they are equal: to avoid such a possibility, we just ignore the diagonal $p = q$ in every histogram.

Figure 5 represents the counting function that we obtained assigning half weight to two digits if $b\{p/q\} \in \mathbb{Z}$.
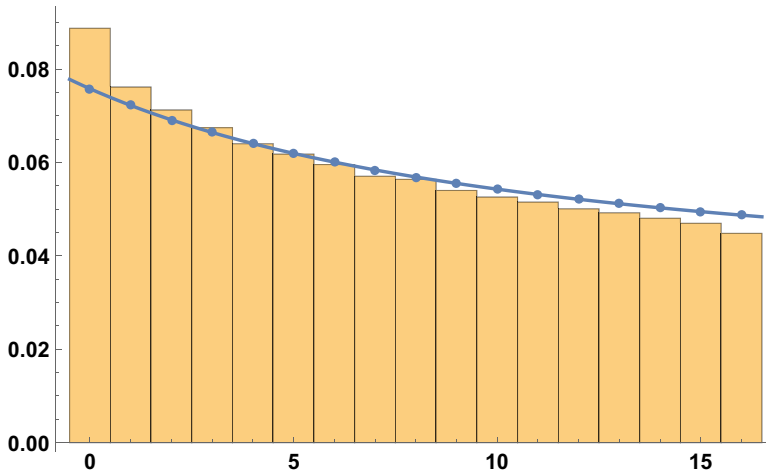
Figure 5: The histogram for $\mathcal{A}(1000; 17, r; 1)$ for primes $p \neq q$, assigning half weight to two digits if $b\{p/q\} \in \mathbb{Z}$.

The trend is, however, monotonic and decreasing also for primes; the assignment of the half weight influences only the cases $r = 0$ and $r = b - 1$ with the effect of a further regularization. This regularity is intrinsic in the fact that we automatically exclude points with multiplicity, so we would expect something better than $T \log T$ even if we cannot prove it. The leading constant is decreasing in r, and it is about 0 for r around $\frac{b}{2}$. This would explain why, in the above figures, for T sufficiently large, the first blocks are always above the trend line, while the latest are below.

Acknowledgment   We thank Sandro Bettin for many conversations on the subject, and for his help with the plots at the end of the present paper.

References

[1]   F. Benford, *The law of anomalous numbers*. Proc. Amer. Philos. Soc. **78**(1938), 551–572.
[2]   E. Cesàro, *Eventualités de la division arithmétique*. Ann. Mat. Pura Appl. (1867–1897) **13**(1885), no. 1, 269–290.
[3]   E. Cesàro, *Excursions arithmétiques à l'infini*. A. Hermann, Paris, 1885.
[4]   J. Cilleruelo and J. Guijarro-Ordóñez, *Ratio sets of random sets*. Ramanujan J. **43**(2017), no. 2, 327–345.
[5]   J. Cilleruelo, D. S. Ramana, and O. Ramaré, *The number of rational numbers determined by large sets of integers*. Bull. Lond. Math. Soc. **42**(2010), no. 3, 517–526.
[6]   J. Cilleruelo, D. S. Ramana, and O. Ramaré, *Quotient and product sets of thin subsets of the positive integers*. Proc. Steklov Inst. Math. **296**(2017), no. 1, 52–64.
[7]   H. Davenport, *Multiplicative number theory. Vol. 74*. Springer Science & Business Media, Berlin and Germany, 2013.
[8]   A. Gambini, G. M. Scarpello, and D. Ritelli, *Probability of digits by dividing random numbers: a ψ and ζ functions approach*. Expo. Math. **30**(2012), no. 3, 223–238.
[9]   G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*. Oxford University Press, Oxford, UK, 1979.

[10] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge University Press, Cambridge, MA, 2010.

[11] G. Pick, *Geometrisches zur zahlenlehre*. Sitzenber. Lotos (Prague) **19**(1899), 311–319.

[12] C. Swaenepoel, *Prime numbers with a positive proportion of pre-assigned digits*. Proc. Lond. Math. Soc. **121**(2020), no. 1, 83–151.

Dipartimento di Matematica, Sapienza Università di Roma, Piazzale Aldo Moro 5, Roma 00185, Italy
e-mail: alessandro.gambini@uniroma1.it

Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Università di Parma, Parco Area delle Scienze 53/a, Parma 43124, Italy
e-mail: remis.tonon@unimore.it   alessandro.zaccagnini@unipr.it