Uniform Distribution Theory **10** (2015), no.1, 63–68

uniform distribution theory

THE TAIL DISTRIBUTION OF THE SUM OF DIGITS OF PRIME NUMBERS

Eric Naslund

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

ABSTRACT. Let $s_q(n)$ denote the base q sum of digits function, which for $n \leq x$, is centered around $\frac{q-1}{2}\log_q x$. In this paper we provide bounds on the tails of the distribution of $s_q(n)$, and prove that given α in the range $\frac{1}{2} \leq \alpha < 0.7375$, and any $\epsilon > 0$, there exists a constant c depending on ϵ such that

$$\left|\left\{p \leq x,\ p \text{ prime } :\ s_q(p) \geq \alpha(q-1)\log_q(x)\right\}\right| \geq \frac{2}{25}x^{2(1-\alpha)}e^{-c\sqrt{\log q}(\log x)^{1/2+\epsilon}}$$

for sufficiently large x. In particular, this shows that there are infinitely many primes with more than twice as many ones than zeros in their binary expansion.

Communicated by Michael Drmota

1. Introduction

A prime number of the form $2^n - 1$ is called a Mersenne prime and it will only have ones in its binary expansion. The first few such primes are 3, 7, 31, and 127, and currently the largest known prime is of this form with over 12.9 million digits. It is a long standing conjecture that there are infinitely many Mersenne primes, and this currently seems entirely out of reach of modern analytic methods. However, we may weaken the condition and ask about primes with a large number of 1's in their base 2 expansion. With this in mind, we ask:

Problem 1. *Are there infinitely many primes with more than twice as many ones than zeros in their binary expansion?*

2010 Mathematics Subject Classification: 11N05,11N36.
Keywords: Mersenne primes, Sum of Digits, Digits of primes, Prime numbers.
I would like to thank Didier Piau for his help, and Gil Kalai for his motivating Math Overflow question. I am also grateful to Greg Martin for his comments and suggestions.

The set of integers with more than twice as many ones than zero's is very small as most integers have approximately half of their digits equal to 1. If we let $s_q(n)$ denote the sum of the digits of n written in base q, then we are asking if there exists infinitely many primes p satisfying $s_2(p) \geq \frac{2}{3}\log_2 p$. Moving to a slightly more general setting, we will look at the sum of digits base q rather than just the binary case. On average $s_q(n)$ is $\frac{q-1}{2}$ multiplied by the number of digits, so we have the asymptotic

$$\sum_{n\leq x} s_q(n) \sim \frac{q-1}{2}\log_q x.$$

However, things become more complicated when we restrict ourselves to the prime numbers. In 1946 Copeland and Erdos [2] proved that

$$\frac{1}{\pi(x)}\sum_{p\leq x} s_q(p) \sim \frac{q-1}{2}\log_q(x)$$

where $\pi(x) = \sum_{p\leq x} 1$ is the prime counting function, and a more precise error term was subsequently given by Shiokawa [4]. In 2009, Drmota, Mauduit and Rivat [3] gave exact asymptotics for the set

$$\left\{p \leq x,\ p \text{ prime } s_q(p) = \alpha\,(q-1)\log_q x\right\}$$

where α lies in the range

$$\alpha \in \left(\frac{1}{2} - K\frac{(\log\log x)^{\frac{1}{2}-\epsilon}}{\sqrt{\log x}},\ \frac{1}{2} + K\frac{(\log\log x)^{\frac{1}{2}-\epsilon}}{\sqrt{\log x}}\right),$$

and is chosen so that $\alpha\,(q-1)\log_q x$ is an integer which avoids certain congruence conditions. However, these results only apply to the center of the distribution, and they don't allow us to make any conclusions about problem 1. In [3] they ask about finding non-trivial bounds for the sum $\sum_{p\leq x} 2^{s_q(p)}$, as this would yields results regarding the tail distribution of the sum of digits of primes. That is, they ask about lower bounds for the size of

$$\left\{p \leq x,\ p \text{ prime}:\ s_q(n) \geq \alpha(q-1)\log_q x\right\}$$

where $\alpha > \frac{1}{2}$ does not depend on x. These are exactly the type of bounds we are looking for in order to answer our question, as problem 1 is the case when $\alpha = \frac{2}{3}$ and $q = 2$. In this note, we provide such lower bounds, and prove the following theorem:

Theorem 1. *Given* $0.2625 < \beta \leq \frac{1}{2}$ *and* $\frac{1}{2} \leq \alpha < 0.7375$*, there exists a constant* c *depending on* ϵ *such that for sufficiently large* x *we have*

$$\left|\left\{p \leq x,\ p \text{ prime}:\ s_q(n) \geq \alpha(q-1)\log_q x\right\}\right| \geq \frac{2}{25}x^{2(1-\alpha)}e^{-c\sqrt{\log q}(\log x)^{1/2+\epsilon}}$$

and

$$\left|\left\{p \le x,\ p \text{ prime}:\ s_q(n) \le \beta(q-1)\log_q x\right\}\right| \ge \frac{2}{25}x^{2\beta}e^{-c\sqrt{\log q}(\log x)^{1/2+\epsilon}}.$$

To approach this problem we do not examine the sum $\sum_{p\le x} 2^{s_q(p)}$, and instead exploit the fact that the multinomial distribution is sharply peaked, using results regarding primes in small intervals to attain the desired lower bound. From theorem 1, problem 1 follows as a corollary. In fact, we have that for any $\alpha < 0.7375$ there are infinitely many primes where the proportion of 1's in their binary expansion greater than α.

2. The Tail Distribution

We start by providing bounds on the size of the tails of the multinomial distribution.

Lemma 1. *(Chernoff bound) Given $0 < a < \frac{1}{2} < b < 1$, we have that*

$$\left|\left\{n < q^k:\ b\,(q-1)\,k \le s_q(n)\right\}\right| \le q^k \exp\left(-\frac{2k}{9}\left(b-\frac{1}{2}\right)^2\right), \tag{1}$$

and

$$\left|\left\{n < q^k:\ s_q(n) \le a\,(q-1)\,k\right\}\right| \le q^k \exp\left(-\frac{2k}{9}\left(a-\frac{1}{2}\right)^2\right). \tag{2}$$

Proof. Each integer in the interval $\left[0, q^k-1\right]$ can be written so that it has exactly k digits base q, by adding zeros in front where neccesary. The distribution of each of the k digit's is an independent random variable which corresponds to the roll of a q sided dice with sides $0, 1, \ldots, q-1$. Normalizing, let ξ be a random variable where

$$\mathrm{P}\left(\xi_i = \frac{2}{q-1}j - 1\right) = \frac{1}{q}$$

for $0 \le j \le q-1$, and for each i let $\xi_i = \xi$. Our goal is then to examine

$$\mathrm{P}\left(\gamma \le \frac{\xi_1+\xi_2+\cdots+\xi_k}{k}\right).$$

For any nonnegative t,

$$\begin{aligned}\mathrm{P}\left(\gamma \le \frac{\xi_1+\xi_2+\cdots+\xi_k}{k}\right) &\le \frac{\mathbb{E}\left(e^{t(\xi_1+\cdots+\xi_k)}\right)}{e^{tk\gamma}} \\ &= \left(e^{-t\gamma}\mathbb{E}\left(e^{t\xi}\right)\right)^k\end{aligned}$$

$$= \quad e^{-kI(t,\gamma)}$$

where

$$I\left(t,\gamma\right) = t\gamma - \log \mathbb{E}\left(e^{t\xi}\right).$$

Evaluating the expectation, we find that

$$\mathbb{E}\left(e^{t\xi}\right) = \sum_{j=0}^{q-1} \frac{1}{q} e^{t\left(\frac{2j}{q-1}-1\right)} = \frac{e^{-t}}{q} \sum_{j=0}^{q-1} \left(e^{\frac{2t}{q-1}}\right)^{j} = \frac{1}{q} \frac{\sinh\left(t+\frac{t}{q-1}\right)}{\sinh\left(\frac{t}{q-1}\right)}.$$

This gives rise to the series expansion

$$\log\left(\frac{1}{q}\frac{\sinh\left(t+\frac{t}{q-1}\right)}{\sinh\left(\frac{t}{q-1}\right)}\right) = \frac{q^2-1}{6(q-1)^2}t^2 - \frac{q^4-1}{180(q-1)^4}t^4 + O\left(\frac{q^6}{(q-1)^6}t^6\right),$$

where the error term holds uniformely for $\frac{qt}{q-1} < 1$. This allows us to prove the inequality

$$\log \mathbb{E}\left(e^{t\xi}\right) \leq \frac{q^2t^2}{6(q-1)^2},$$

for q, t satisfying $\frac{qt}{q-1} < 1$. To maximize $I(t,\gamma)$, we choose $t = \frac{\gamma}{3}\frac{q-1}{q+1}$, and obtain the upper bound

$$\mathrm{P}\left(\gamma \leq \frac{\xi_1+\xi_2+\cdots+\xi_k}{k}\right) \leq \exp\left(-\frac{k}{6}\left(\frac{q-1}{q+1}\right)\gamma^2\right),$$

which proves equation (1) upon taking $\gamma = 2b-1$, and noting that $\frac{q-1}{q+1} \geq \frac{1}{3}$ since $q \geq 2$. The proof of equation 2 is identical, as the distribution is symmetric. $\square$

Next, we will need the best existing results on prime gaps. In 2001, Baker, Harman and Pintz proved that for $x \geq x_0$,

$$\pi\left(x+x^{\theta}\right) - \pi(x) \geq \frac{9}{100}\frac{x^{\theta}}{\log x} \tag{3}$$

for any $\theta \geq 0.525$ [1]. Armed with equation 3 and lemma 1, we are now ready to prove theorem 1.

P r o o f. Let $\alpha^{'} = \alpha + r(x)$ where $r(x)$ is chosen so that $\alpha^{'} < 0.7375$. Let $k = \left[\log_q x\right]$, $l = \lceil 2\left(1-\alpha^{'}\right)k\rceil$ so that $q^k \leq x$ and $q^l \geq x^{0.525}$. Consider the interval $\left[q^k - q^l,\ q^k - 1\right]$, which is an interval whose first $k-l$ digits base q are

equal to $q-1$. By Baker, Harman and Pintz, if x is sufficiently large, there will be

$$\geq \frac{9}{100}\frac{q^l}{\log\left(q^k\right)} \geq \frac{9}{100}\frac{q^l}{\log x}$$

primes in this interval, where the constant is explicit. By equation (2), there are at most $q^l \exp\left(-\frac{2l\delta^2}{9}\right)$ integers between 0 and q^l which have digit sum less than $(q-1)l\left(\frac{1}{2}-\delta\right)$. Letting $\delta = 9\sqrt{\frac{\log\log x}{l}}$, it follows that there are at most $q^l/\log^2 x$ integers in the interval $\left[q^k-q^l,\ q^k-1\right]$ whose digit sum is less than

$$(q-1)(k-l)+(q-1)l\left(\frac{1}{2}-\sqrt{\frac{\log l}{l}}\right).$$

For $x \geq e^{100}$, $\frac{1}{\log x} \leq \frac{1}{100}$, which implies that for sufficiently large x there are more than $\frac{2}{25}\frac{q^l}{\log x}$ primes in the interval $\left[q^k-q^l,\ q^k-1\right]$ with digit sum larger than

$$\alpha^{'}(q-1)k-(q-1)\sqrt{l\log\log x}.$$

Expanding $\alpha^{'} = \alpha + r(x)$, and taking $r(x) = \sqrt{\frac{\log\log x}{\log_q x}}$ yields a digit sum greater than

$$\alpha(q-1)\log_q(x),$$

which proves the result since

$$\frac{q^l}{\log x} \geq \frac{x^{2(1-\alpha)}x^{-2r(x)}}{\log x} \geq x^{2(1-\alpha)}\exp\left(-4\sqrt{\log q}\sqrt{\log x}\sqrt{\log\log x}\right).$$

The proof of the lower bound for the size of the corresponding set of primes with $s_q(p) \leq \beta(q-1)\log_q(x)$ for $0.2625 < \beta \leq \frac{1}{2}$ is identical. □

REFERENCES

[1] R. C. Baker, G. Harman, and J. Pintz. *The difference between consecutive primes. II.* Proc. London Math. Soc. (3), **83(3)** (2001),532–562.

[2] Arthur H. Copeland and Paul Erdös. *Note on normal numbers.* Bull. Amer. Math. Soc., **52** (1946), 857–860.

[3] Michael Drmota, Christian Mauduit, and Joël Rivat. *Primes with an average sum of digits.* Compos. Math., **145(2)** (2009), 271–292.

[4] Iekata Shiokawa. *On the sum of digits of prime numbers.* Proc. Japan Acad., **50** (1974), 551–554,.

Received May 8, 2013
Accepted August 20, 2013

Eric Naslund
Fine Hall, Washington Road
Princeton NJ 08544-1000 USA
E-mail: naslund@math.princeton.edu