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 [3], they look at sum of digits of prime numbers, and provide asymptotics for the size of the set $\{p \leq x, p \text{ prime } s_q(p) = \alpha (q-1) \log_q x\}$ where $\alpha$ lies in the range

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

for some constant $K$. In this note, we examine the tails of this distribution, and prove that

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

for $\frac{1}{2} \leq \alpha < 0.7375$. This proves that there are infinitely many primes with more than twice as many ones than zeros in their binary expansion.

1. Introduction

A prime number which can be written in the form $2^n - 1$ will have only ones in its binary expansion, and is called a Mersenne prime. The first few such primes are 3, 7, 31, and 127. Currently, the largest known prime is of this form, and it has over 12.9 million digits. These numbers have been looked at for centuries, and date back to Euclid who was interested in them for their connection with perfect numbers, something that we will not explore here. It is a long standing conjecture that there are infinitely many Mersenne primes, and currently this 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 the following motivational question:

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

If we let $s_q(n)$ denote the sum of the digits of $n$ written in base $q$, then we are asking if there are infinitely many primes $p$ which satisfy $s_2(p) \geq \frac{2}{3} \log_2 p$. Moving to a slightly more general setting, we will look at the sum of digits base $q$ rather than just the binary case. The average of $s_q(n)$ is roughly $\frac{q-1}{2}$ 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 much 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
\[ \{ p \leq x, \ p \text{ prime } s_q(p) = \alpha (q - 1) \log_q x \} \]
where \( \alpha \) lies in the range
\[ \alpha \in \left( \frac{1}{2} - K \frac{\log \log x}{\sqrt{\log x}}, \ \frac{1}{2} + K \frac{\log \log x}{\sqrt{\log x}} \right), \]
and is chosen so that \( \alpha (q - 1) \log_q x \) is an integer which avoids certain congruence conditions. However, these results don’t allow us to make any conclusions about problem 1. In [3] they also asked 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, lower bounds for the size of sets of primes of the form
\[ \{ p \leq x, \ p \text{ prime } : \ s_q(n) \geq \alpha (q - 1) \log_q x \} \]
where \( \alpha > \frac{1}{2} \). 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 2.** Given \( 0.2625 < \beta \leq \frac{1}{2} \) and \( \frac{1}{2} \leq \alpha < 0.7375 \), for sufficiently large \( x \) we have that
\[ \left| \{ p \leq x, \ p \text{ prime } : \ s_q(n) \geq \alpha (q - 1) \log_q x \} \right| \gg \epsilon x^{2(1-\alpha)} e^{-c(\log x)^{1/2+\epsilon}} \]
and
\[ \left| \{ p \leq x, \ p \text{ prime } : \ s_q(n) \leq \beta (q - 1) \log_q x \} \right| \gg \epsilon x^{2\beta} e^{-c(\log x)^{1/2+\epsilon}}. \]

We do not examine the sum \( \sum_{p \leq x} 2^{s_q(p)} \), rather we note that the multinomial distribution is sharply peaked, so results regarding primes in small intervals allow us to attain such a lower bound. From theorem 2, 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 \( \alpha \).

2. **The Tail Distribution**

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

**Lemma 3.** (Chernoff bound) Given \( \frac{1}{2} < a < 1 \), we have that
\[ \left| \{ n \leq q^k : \ a(q - 1) k \leq s_q(n) \} \right| \leq \exp \left( -\frac{k}{18} \left( a - \frac{1}{2} \right)^2 \right). \]

**Proof.** On the interval \([0, q^k]\) each digit can be thought of as an independent random variable which corresponds to the roll of a \( q \) sided dice with sides 0, 1, \ldots, \( q - 1 \). Normalizing, let \( \xi \) be a random variable where
\[ P \left( \xi_i = \frac{2}{q-1} j - 1 \right) = \frac{1}{q} \]
for \( 0 \leq j \leq q - 1 \), and for each \( i \) let \( \xi_i = \xi_i \). Our goal is then to examine
\[ P \left( \gamma \leq \frac{\xi_1 + \xi_2 + \cdots + \xi_k}{k} \right). \]
For any nonnegative $t$,
\[
\mathbb{P} \left( \gamma \leq \frac{\xi_1 + \xi_2 + \cdots + \xi_k}{k} \right) \leq \frac{\mathbb{E} \left( e^{t(\xi_1 + \cdots + \xi_k)} \right)}{e^{tk\gamma}} = (e^{-t\gamma} \mathbb{E} \left( e^{t\xi} \right))^k = e^{-kI(t, \gamma)}
\]
where
\[
I(t, \gamma) = t\gamma - \log \mathbb{E} \left( e^{t\xi} \right).
\]
Evaluating the expectation, we find that
\[
\mathbb{E} \left( e^{t\xi} \right) = \sum_{j=0}^{q-1} \frac{1}{q} e^{t \left( \frac{2j}{q-1} - 1 \right)} = e^{-t} \sum_{j=0}^{q-1} \left( e^{\frac{2j}{q-1}} \right)^j = \frac{1}{q} \frac{\sinh \left( t + \frac{t}{q-1} \right)}{\sinh \left( \frac{t}{q-1} \right)}.
\]
This gives rise to the series expansion
\[
\log \left( \frac{1}{q} \frac{\sinh \left( t + \frac{t}{q-1} \right)}{\sinh \left( \frac{t}{q-1} \right)} \right) = \frac{(q+1)}{6(q-1)} t^2 - \frac{q^3 + q^2 + q + 1}{180(q-1)^3} t^4 + O(t^6),
\]
allowing us to prove that
\[
\log \mathbb{E} \left( e^{t\xi} \right) \leq \frac{(q+1)}{6(q-1)} t^2.
\]
To maximize $I(t, \gamma)$, we choose $t = \frac{2q-1}{3q+1}$, and obtain the upper bound
\[
\mathbb{P} \left( \gamma \leq \frac{\xi_1 + \xi_2 + \cdots + \xi_k}{k} \right) \leq \exp \left( -k \frac{q-1}{6 \left( q+1 \right)} \gamma^2 \right),
\]
which proves the lemma since $q \geq 2$. □

Next, we will need the best existing results on prime gaps. In 2001, Baker, Harman and Pintz proved that
\[
\pi \left( x + x^\theta \right) - \pi(x) \gg \frac{x^\theta}{\log x}
\]
for any $\theta \geq 0.525$ [1]. Armed with equation 2.1 and lemma 3, we are now ready to prove theorem 2.

Proof. Let $\alpha' = \alpha + r(x)$ where $r(x)$ is chosen so that $\alpha' < 0.7375$. Let $k = \lfloor \log_q x \rfloor$, so that $q^k \leq x$, and let $l = \lfloor 2 \left( 1 - \alpha' \right) k \rfloor$. Consider the interval $[q^k - q^l, q^k - 1]$, which is an interval whose first $k - l$ digits base $q$ are equal to $q - 1$. By Baker, Harman and Pintz, there will be
\[
\gg q^l \frac{q^l}{\log (q^k)} \gg \frac{q^l}{\log x}
\]
primes in this interval, where the constant is explicit. By Lemma 3, there are at most $\exp \left( -\frac{\delta^2}{18} \right)$ integers between 0 and $q^l$ which have digit sum less than $(q-1)l \left( \frac{1}{2} - \delta \right)$. Letting $\delta = \frac{\log l}{\sqrt{l}}$, it follows that there are at most $q^l e^{-\left(\log l\right)^2}$ integers in the interval $[q^k - q^l, q^k - 1]$ whose digit sum is less than
\[
(q - 1)(k - l) + (q - 1)l \left( \frac{1}{2} - \frac{\log l}{\sqrt{l}} \right).
\]
As $q^l e^{-(\log l)^2}$ is significantly smaller than $\frac{q^l}{k \log q}$, almost all of the primes in this interval will have a digit sum greater than the above, and so we see that there

$$\gg \frac{q^l}{\log (x)}$$

primes with digit sum larger than

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

Expanding $\alpha' = \alpha + r(x)$, and taking $r(x) = c \frac{\log \log x}{\sqrt{\log x}}$ for the appropriate constant $c$ yields a digit sum greater than

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

which proves the result since

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

The proof for the lower bound of 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.

Remark. The reader may note that for any $\alpha < 0.7375$ there are more possible choices for the first $k - l$ digits other than all 1’s. It is conceivable that if we looked at multiple intervals where the first $k - l$ digits had many 1’s that we would be able to increase the density by a small factor, and possibly a significant factor for smaller $\alpha$. While such an approach seems promising, and while it seems logical to sum over multiple intervals, the end result and lower bound for the number of primes is roughly the same. The exponent of $x$ is no different, so we opted to present the simpler argument above.

Acknowledgements

I would like to thank Didier Piau for helping me understand the Chernoff bound.

References

[1] R. C. Baker, G. Harman, and J. Pintz. The difference between consecutive primes. II. Proc. London Math. Soc. (3), 83(3):532–562, 2001.
[2] Arthur H. Copeland and Paul Erdös. Note on normal numbers. Bull. Amer. Math. Soc., 52:857–860, 1946.
[3] Michael Drmota, Christian Mauduit, and Joël Rivat. Primes with an average sum of digits. Compos. Math., 145(2):271–292, 2009.
[4] Iekata Shiokawa. On the sum of digits of prime numbers. Proc. Japan Acad., 50:551–554, 1974.