The level of distribution of the Thue–Morse sequence

Lukas Spiegelhofer

March 21, 2022, Rencontres de théorie analytique et élémentaire des nombres

The basic question

Let $q \geq 2$ be an integer. We know the base-q expansion of $n \in \mathbb{N}$:

$$n = \delta_0 q^0 + \delta_1 q^1 + \delta_2 q^2 + \cdots + \delta_{L-1} q^{L-1},$$

where $(\delta_j)_{0 \leq j < L} \in \{0, \ldots, q-1\}^L$ and ($L = 0$ or $\delta_{L-1} \neq 0$), and we write

$$[n]_q := (\delta_{L-1}, \delta_{\nu-1}, \ldots, \delta_0).$$

The level of distribution is concerned with arithmetic progressions, which in turn are given by repeated addition of a constant.

What happens to the base-q expansion of $n \in \mathbb{N}$ when a constant $d \in \mathbb{N}$ is added?

Carry propagation

Consider, for example, the following additions in base 2.

$$
\begin{array}{rr}
 & 11101001110110011 \\
+ & 10110001001101 \\
\hline
= & 100000000000000000. \\
 & \uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow
\end{array}
\qquad
\begin{array}{rr}
 & 11101001100110011 \\
+ & 10110001001101 \\
\hline
= & 11111111110000000. \\
 & \uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow
\end{array}
$$

$$
\begin{array}{rr}
 & 1110111101110110 \\
+ & 10110001001101 \\
\hline
= & 10001101111000011. \\
 & \uparrow\uparrow\uparrow\ \ \uparrow\uparrow\ \ \ \ \uparrow\uparrow\uparrow\uparrow\uparrow\ \
\end{array}
\qquad
\begin{array}{rrl}
 & 12 & \text{(digit sum)} \\
+ & 7 & \text{(digit sum)} \\
\hline
= & 9 & \text{(digit sum)} \\
+ & 10 & \text{(carries).}
\end{array}
$$

The appearance of *carries* in the addition $n + t$ causes many cases to be distinguished.

Addition of 1

The (possibly empty) block of 1s on the right of the binary expansion of n is replaced by 0s, and the 0 to the left of the block is replaced by 1.

$$* \, 011 \cdots 1 \mapsto *100 \cdots 0 \tag{1}$$

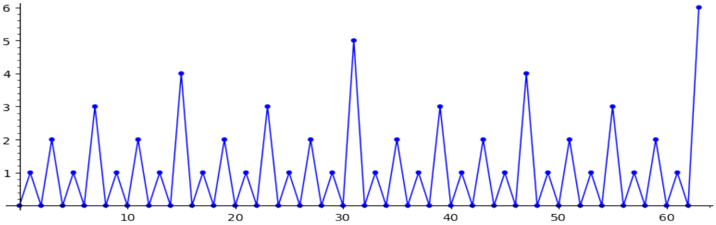
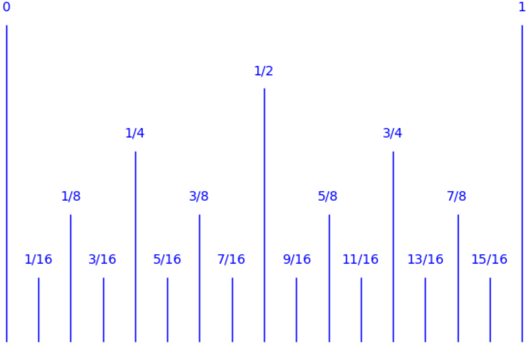


Figure: The number of carries in the addition $n + 1$

This is the *ruler sequence* $n \mapsto \nu_2(n+1)$, given by the 2-valuation $\nu_2(n+1)$.

The ruler sequence

The following picture is well known in countries using imperial units.



The case $t \geq 3$

For $d = 3$ we have the following cases:

$$*00 \mapsto *11; \qquad *01^k01 \mapsto *10^k00;$$
$$*01^k10 \mapsto *10^k01; \qquad *01^k11 \mapsto *10^k10.$$

This situation does not get better with growing d. Carries can propagate through many blocks of 1, and many cases occur.

The binary sum-of-digits function

As first step (in the quest of better understanding the base-q expansion) we consider the *base-q sum-of-digits function* s_q. The integer $\mathsf{s}_q(n)$ is the minimal number of powers of q needed to write n as their sum.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\mathsf{s}_2(n)$	0	1	1	2	1	2	2	3	1	2	2	3	2	3	3	4

The `POPCNT` instruction on modern microprocessors returns the binary sum of digits of an integer $n \in \{0, 2^{64} - 1\}$ within ~ 1ns.

The sum-of-digits function under addition

We have the important identity (Legendre)

$$\mathsf{s}_2(n) + \mathsf{s}_2(d) = \mathsf{s}_2(n+d) + \nu_2\left(\binom{n+d}{d}\right)\nu_2\left(\binom{n+d}{d}\right),$$

where the 2-valuation $\nu_2\bigl(\binom{n+d}{d}\bigr)$ equals the number of *carries* *carries* that appear in the addition $n+d$ in binary.

Let us define

$$\delta(j,d) \coloneqq \lim_{N\to\infty}\frac{1}{N}\#\bigl\{0 \le n < N : \mathsf{s}_2(n+d) - \mathsf{s}_2(n) = j\bigr\}.$$

T. W. Cusick conjectured that

$$c_d \coloneqq \delta(0,d) + \delta(1,d) + \cdots > 1/2.$$

SW2021

In other words,

When a constant d is added, does the binary sum of digits of n weakly increase, more often than not?

Theorem (S.–Wallner 2021, Ann. Scuola Norm. Sup. Pisa Cl. Sci.)

Assume that the positive integer d has at least M blocks of ones in its binary expansion (where M is an absolute constant). Then $c_d > 1/2$.

The remaining cases — few blocks of 1s — are the 'hard cases' according to Cusick, and the interesting ones for applications

$\longrightarrow$ more work to do!

The Thue–Morse sequence

The parity of the number of ones in the binary expansion yields the *Thue–Morse sequence*

$$\mathtt{T} = (\mathsf{s}_2(n) \bmod 2)_{n\geq 0} = \mathtt{01101001100101101001011001101001}\cdots$$

The sequence $\mathtt{T}$ is an *automatic sequence* and as such can be defined via a *uniform morphism on a finite alphabet*: Let us define

$$\varphi : \mathtt{0} \mapsto \mathtt{01}, \quad \mathtt{1} \mapsto \mathtt{10}.$$

Starting with $\mathtt{0}$, we obtain

$$\mathtt{0} \mapsto \mathtt{01} \mapsto \mathtt{0110} \mapsto \mathtt{01101001}\cdots$$

The behaviour of the Thue–Morse sequence under addition

Let us define

$$\tau(n) \coloneqq (-1)^{\mathsf{s}_2(n)} = 1 - 2\mathtt{T}(n) = (1, -1, -1, 1, -1, 1, 1, -1, \ldots)$$

and the *correlation*

$$\gamma_d \coloneqq \lim_{N\to\infty} \frac{1}{N} \sum_{0\le n<N} \tau(n)\overline{\tau(n+d)}.$$

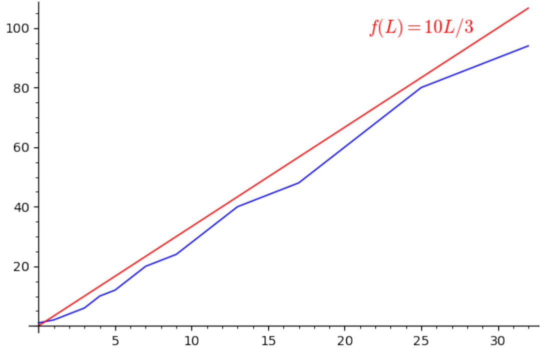
We have

$$\gamma_1 = -\frac{1}{3}, \quad \gamma_{2d} = \gamma_d, \quad \gamma_{2d+1} = \frac{-\gamma_d - \gamma_{d+1}}{2}.$$

We have $\mathsf{s}_2(n+1) - \mathsf{s}_2(n) = m$ for $m \le 1$ and $n \in 2^{1-m} - 1 + 2^{2-m}\mathbb{Z}$.
Therefore $\gamma_1 = -1/2 + 1/4 - 1/8 + \cdots = -1/3$.

The factor complexity of $\mathtt{T}$

There are only very few words over $\{0,1\}$ appearing as *factors* (contiguous finite subsequences) of $\mathtt{T}$: the number of factors of length L appearing in $\mathtt{T}$ is bounded by CL with an absolute constant C.



$p(L)$	2^L
1	1
2	2
4	4
6	8
10	16
12	32
16	64
20	128
22	256
24	512
28	1024

The sum-of-digits function along arithmetic progressions

Repeated addition of d leads to arithmetic progressions. How do the binary digits behave along $a + d\mathbb{N}$?
Avgustinovich, Fon-Der-Flaass, and Frid (2003) proved that *every* finite sequence $A \in \{0, 1\}^L$ appears as an arithmetic subsequence of $\mathtt{T}$.

$$\text{“The Thue–Morse sequence has} \left\{ \begin{array}{c} \text{low factor complexity} \\ \text{full arithmetical complexity} \end{array} \right\} \text{”}$$

This situation is very different from the Fibonacci word $F = 010010100100101001 \cdots$ defined by

$$0 \mapsto 01, \quad 1 \mapsto 0,$$

which only has cubic arithmetical complexity (Cassaigne–Frid 2007).

Arithmetic subsequences of $\mathtt{T}$

Theorem (Gel'fond)

Let q, m, d, b, a be integers and $q, m, d \geq 2$. Suppose that $\gcd(m, q-1) = 1$. Then

$$\left|\left\{1 \leq n \leq x : n \equiv a \bmod d, s_q(n) \equiv b \bmod m\right\}\right| = \frac{x}{dm} + \mathcal{O}\left(x^\lambda\right)$$

for some $\lambda < 1$ independent of x, d, a, and b.

We know that arbitrarily long sequences of $\mathtt{0}$s appear as arithmetic subsequences of $\mathtt{T}$, therefore the $\mathcal{O}$-constant cannot be uniform in d! This theorem does therefore not tell us much about *short* APs.

Very sparse arithmetic subsequences of $\mathtt{T}$

However, *for most d* the number of 0s and 1s will be balanced along short arithmetic sequences $(nd + a)_{0 \le n < N}$.

Theorem (S. 2020)

The Thue–Morse sequence has level of distribution 1. More precisely, for all $\varepsilon > 0$ we have

$$\sum_{1 \le d \le D} \max_{\substack{y,z \ge 0 \\ z-y \le x}} \max_{0 \le a < d} \left| \sum_{\substack{y \le n < z \\ n \equiv a \bmod d}} (-1)^{\mathsf{s}_2(n)} \right| \le C x^{1-\eta}$$

for some C and $\eta > 0$ depending on ε, where $D = x^{1-\varepsilon}$.

For all $\rho > 0$, most arithmetic subsequences A of $\mathtt{T}$ having N elements and common difference $\asymp N^\rho$ have about the same number of 0s and 1s.

Reduction of the theorem

Let $\mathrm{e}(x) = \exp(2\pi i x)$. The theorem follows from the following statement.

Proposition

For real numbers $N, D \ge 1$ and ξ set

$$S_0(N, D, \xi) = \sum_{D \le d < 2D} \max_{a \ge 0} \left| \sum_{0 \le n < N} \mathrm{e}\left(\frac{1}{2} \mathrm{s}_2(nd + a) + n\xi\right) \right|. \tag{2}$$

For all $\rho_2 \ge \rho_1 > 0$ there exist $\eta > 0$ and C such that the following holds. For all real numbers $N, D \ge 1$ such that $N^{\rho_1} \le D \le N^{\rho_2}$, and all ξ, we have

$$\frac{S_0(N, D, \xi)}{ND} \le C N^{-\eta}. \tag{3}$$

The inequality of van der Corput

Lemma

Let I be a finite interval in $\mathbb{Z}$ containing N integers and let z_n be a complex number for $n \in I$. For all integers $K \geq 1$ and $R \geq 1$ we have

$$\left|\sum_{n\in I} z_n\right|^2 \leq \frac{N+K(R-1)}{R} \sum_{0\leq |r|<R} \left(1-\frac{|r|}{R}\right) \sum_{\substack{n\in I\\ n+Kr\in I}} z_{n+Kr}\overline{z_n}.$$

The important thing is that we only need to estimate certain correlations $\sum z_{n+r}\overline{z_n}$ with "small r" instead of the original sum $\sum z_n$, and we can profit from *cancellation effects*.

Cancellation effects, part I: Mauduit–Rivat

"Adding a small integer mostly changes only digits at low positions." More precisely, assume that $r \in \{0, \ldots, 2^\mu - 1\}$, and that the positive integer n has at least one 0 in its binary expansion in the window $[\mu, \mu + \ell)$,

$$n = \bigl(\delta_\nu, \delta_{\nu-1}, \ldots, \delta_{\mu+\ell}, \underbrace{\delta_{\mu+\ell-1}, \ldots, \delta_\mu}_{\text{at least one 0}}, \delta_{\mu-1}, \ldots, \delta_0\bigr)_2.$$

In the addition $n + r$, there is no carry propagation into the digits with indices $\geq \mu + \ell$!

Writing

$$\mathsf{s}_2^A(n) := \sum_{j \in A} \delta_j(n)$$

for a set $A \subset \mathbb{N}$, we have

$$\mathsf{s}_2(n + r) - \mathsf{s}_2(n) = \mathsf{s}_2^A(n + r) - \mathsf{s}_2^A(n),$$

where $A = [0, \mu + \ell)$.

Cancellation effects, part II

Extending the Mauduit–Rivat idea, we may "cut out" an arbitrary interval $[a, b)$ of digits: Assume that

$$\left\| \frac{K}{2^b} \right\| < 2^{b-a+\ell}.$$

In other words, we have

$$\bigl(\delta_{a-\ell}(K), \ldots, \delta_{b-1}(K)\bigr) \in \{(0, \ldots, 0), (1, \ldots, 1)\}.$$

Assume that the binary expansion of n has at least one digit 0 and one digit 1 at indices $\in \{a - \ell, \ldots, a - 1\}$,

$$n = \bigl(\delta_\nu, \delta_{\nu-1}, \ldots, \delta_b, \delta_{b-1}, \ldots, \delta_a, \underbrace{\delta_{a-1}, \ldots, \delta_{a-\ell},}_{\text{both digits appear}} \delta_{a-\ell-1}, \ldots, \delta_0\bigr)_2.$$

Then

$$\mathsf{s}_2(n + K) - \mathsf{s}_2(n) = \mathsf{s}_2^A(n + K) - \mathsf{s}_2^A(n),$$

where $A = \mathbb{N} \setminus [a, b)$.

The core of the method: van der Corput, iterated

- The common difference d may be large compared to N. Addition of d changes up to $\rho_2 \log_2 N$ binary digits in each step. Applying van der Corput the first time, we cut away all digits above $M = \rho_2 \log_2 N + \ell$.
- The digits of $nd + a$ below M cannot attain all combinations, as n runs through $\{0, \ldots, N-1\}$ — too many digits are left!
- We apply van der Corput's inequality repeatedly on the sum

$$\sum \exp\bigl(\tfrac{1}{2}\mathsf{s}_2^M((n+r)d+a) - \mathsf{s}_2^M(nd+a)\bigr),$$

  cutting out a different interval of digits in each step.
- For this, we have to choose multiples K_j in such a way that the binary digits of $K_j d$ in a certain interval are all equal to 0 or all equal to 1 — a Diophantine approximation problem.

Gowers norms

- The remaining interval of digits is *short*, while the summation over n is *long*. For most d, we obtain uniform distribution of the binary digits of $nd + a$ in this interval. This enables us to replace the sum along the arithmetic progression $nd + a$ by a full sum!
- We now have to deal with *higher order correlations* — each application of van der Corput's inequality increases the order by 1. This leads us to a *Gowers norm* of the Thue–Morse sequence, for which an upper bound is available (Konieczny 2019).

We have to estimate

$$\left|\sum_{n<N} \mathrm{e}\left(\frac{1}{2}\mathsf{s}_2^M((n+r)d+a) - \frac{1}{2}\mathsf{s}_2^M(nd+a)\right)\right|^2$$

$$= \left|\sum_{n<N} \prod_{\varepsilon\in\{0,1\}} \mathrm{e}\left(\frac{1}{2}\mathsf{s}_2^M((n+\varepsilon r)d+a)\right)\right|^2$$

By iterating van der Corput, we are left with the expression

$$\left|\sum_{n<N} \prod_{\varepsilon,\varepsilon_1,\ldots,\varepsilon_m\in\{0,1\}} \mathrm{e}\left(\frac{1}{2}\mathsf{s}_2^M((n+\varepsilon r+\varepsilon_1 K_1 r_1+\cdots+\varepsilon_m K_m r_m)d+a)\right)\right|^2.$$

Each multiple K_j is responsible for eliminating an interval of μ digits, which is achieved by the condition

$$\left\|\frac{K_j d}{2^b}\right\| \le 2^{-\mu-\ell}.$$

Eliminating the very sparse arithmetic progression

This successive reduction of digits leaves us with only a short interval $[0,\sigma)$ of significant digits. Assuming for simplicity that d is odd, the expression $nd+a \bmod 2^\sigma$ traverses $[0,2^\sigma)$ in a uniform manner.

Now $nd+a$ may be replaced by n.

The very sparse arithmetic progression has been replaced by a full sum. In particular, the shift a has disappeared.
After some technicalities, we arrive at a *Gowers norm*, which we have to bound nontrivially.

A Gowers norm estimate

Proposition (Essentially Konieczny 2019)

Let $m \ge 2$ be an integer. There exist $\eta > 0$ and C such that

$$\frac{1}{2^{(m+1)\sigma}} \sum_{\substack{0 \le n < 2^\sigma \\ 0 \le r_1, \ldots, r_m < 2^\sigma}} \mathrm{e}\left(\frac{1}{2} \sum_{\varepsilon \in \{0,1\}^m} \mathsf{s}_2^{[0,\sigma)}(n + \varepsilon \cdot r) \right) \le C 2^{-\sigma\eta}$$

for all $\sigma \ge 0$, where $\varepsilon \cdot r = \sum_{1 \le i \le m} \varepsilon_i r_i$.

Restating the main theorem

Theorem (S. 2020)

The Thue–Morse sequence has level of distribution 1. More precisely, for all $\varepsilon > 0$ we have

$$\sum_{1\le d\le D} \max_{\substack{y,z\ge 0\\ z-y\le x}} \max_{0\le a<d} \left| \sum_{\substack{y\le n<z\\ n\equiv a \bmod d}} (-1)^{s_2(n)} \right| \le Cx^{1-\eta}$$

for some C and $\eta > 0$ depending on ε, where $D = x^{1-\varepsilon}$.

Fouvry and Mauduit (1996) obtained a level of distribution 0.5924 for the Thue–Morse sequence.

The Zeckendorf expansion

Every nonnegative integer n is the sum of different, non-consecutive Fibonacci numbers F_i and such a representation is unique $\rightsquigarrow$ Zeckendorf expansion.

0	0	0	8	10000	1	16	100100	2
1	1	1	9	10001	2	17	100101	3
2	10	1	10	10010	2	18	101000	2
3	100	1	11	10100	2	19	101001	3
4	101	2	12	10101	3	20	101010	3
5	1000	1	13	100000	1	21	1000000	1
6	1001	2	14	100001	2	22	1000001	2
7	1010	2	15	100010	2	23	1000010	2

- The number of 1s needed is the *Zeckendorf sum of digits* $z(n)$ of n.
- The Zeckendorf expansion is a generalization of the Fibonacci word, which is given by the lowest digit.

Theorem (Drmota, Müllner, S. 2021+)

1. *Let* $\vartheta \in \mathbb{R} \setminus \mathbb{Z}$. *The function* $n \mapsto \mathrm{e}(\vartheta \mathrm{z}(n))$ *has level of distribution* 1.
2. *Let* k *be a sufficiently large integer. There exists a prime number* p *with*
$$\mathrm{z}(p) = k.$$
In particular, p *can be represented as the sum of* k *pairwise different and non-consecutive Fibonacci numbers.*

Possible extensions and open problems

Consider other numeration systems, such as the Tribonacci expansion. Prove a level of distribution 1 and a prime number theorem for associated sum-of-digits functions.

Prove that for all $\varepsilon > 0$, most $D \le d < 2D$, all intervals I of length $\sim D^{\varepsilon}$, and all a,

$$m \mapsto \#\{n \in I : \mathsf{s}_q(n) = m, n \equiv a \bmod d\}$$

closely follows a Gaussian.

Prove that $\mathtt{T}(\lfloor n^c \rfloor)$ defines a normal sequence for all $c \in (1, 2)$ (*Simple normality* follows from the Compositio-paper).

Thank you!

[0] Supported by the Austrian Science Fund (FWF), Projects F55 and MuDeRa (jointly with ANR).