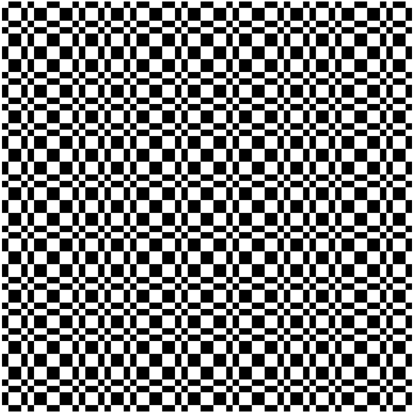
Pseudorandom sequences derived from automatic sequences

Arne Winterhof
(joint work with László Mérai)

Austrian Academy of Sciences
RICAM, Linz

IWSDA'22
August 3, 2022

Is this random?



... or is this more random?

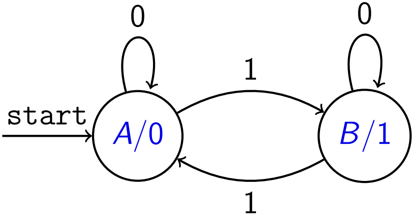
Pseudorandom sequences

- generated by deterministic algorithms which simulate randomness
- not random at all but guarantee certain desirable features (depending on application)
- mathematical/cryptographic point of view: as many desirable features as possible
- many different measures: linear complexity, correlation, normality, ...

Thue-Morse sequence

$$t_0 = 0, \quad t_n = \begin{cases} t_{n/2} & \text{if } n \text{ is even,} \\ 1 - t_{(n-1)/2} & \text{if } n \text{ is odd,} \end{cases} \qquad n = 1, 2, \ldots$$

- t_n is the sum of digits of n modulo 2, $n = 0, 1, \ldots$
- automatic sequence generated by the Thue-Morse automaton



- $t_0 \ldots t_{11} = 011010011001\ldots$
  $t_{11} = 1 - t_5 = t_2 = t_1 = 1 - t_0 = 1$
  or $11 = 8 + 2 + 1 = (1011)_2$: $t_{11} = 1 + 0 + 1 + 1 \equiv 1 \bmod 2$

The first 4096 sequence elements

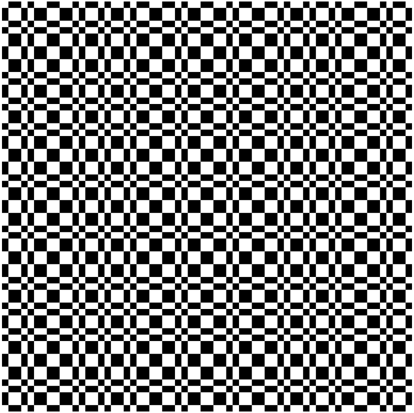


Figure: The first 4096 elements of the Thue-Morse sequence split into 64 rows of each 64 sequence elements. Zeros are represented by white, ones are represented by black.

Features

pseudorandom/desirable:

- large Nth linear complexity
- large Nth maximum-order complexity
- balance
- small well-distribution measure

not pseudorandom/undesirable:

- very large correlation measure of order 2
- very small expansion complexity
- there are short patterns such as 000 and 111 which do not appear in the sequence
- subword complexity is only linear

Subsequences

- may destroy the non-random structure of the original sequence
- may keep the desirable features of pseudorandomness

promising candidates:

- along squares, cubes, bi-squares, ... or along the values of any polynomial f of degree at least 2 with $f(\mathbb{N}_0) \subset \mathbb{N}_0$
- along primes
- along the Piatetski-Shapiro sequence $\lfloor n^c \rfloor$, $1 < c < 2$,
- along geometric sequences such as 3^n

Thue-Morse sequence along squares

inherits:

- large maximum-order complexity and thus a large linear complexity
- asymptotically balanced/simply normal

in contrast to the original sequence:

- unbounded expansion complexity
- normal, that is, asymptotically each pattern appears with the right frequency in the sequence

Roughly speaking: looks much more random than the original sequence.

Figure: The first 4096 elements of the Thue-Morse sequence along squares split into 64 rows of each 64 sequence elements. Zeros are represented by white, ones are represented by black.

Linear complexity

The Nth linear complexity $L((s_n), N)$ of a sequence (s_n) over $\mathbb{F}_2 = \{0, 1\}$ is the length L of a shortest linear recurrence relation satisfied by the first N elements of (s_n),

$$s_{n+L} \equiv c_{L-1} s_{n+L-1} + \cdots + c_1 s_{n+1} + c_0 s_n \bmod 2, \quad 0 \leq n \leq N - L - 1,$$

for some $c_0, \ldots, c_{L-1} \in \mathbb{F}_2$.
linear complexity: $L((s_n)) = \sup_{N \geq 1} L((s_n), N)$

- expected value $\frac{N}{2} + O(1)$ (Gustavson, 1976)
- deviations of magnitude $\log N$ must appear for infinitely many N (Niederreiter, 1988)
- $L((s_n)) < \infty \iff (s_n)$ ultimately periodic

Linear complexity of Thue-Morse sequence

$$L((t_n), N) = 2 \left\lfloor \frac{N+2}{4} \right\rfloor \quad \text{(Mérai/W., 2018)}$$

Proof of $L((t_n), N) \geq \frac{N-1}{2}$:

- (t_n) is not (ultimately) periodic
- generating function $G(x) = \sum_{n=0}^{\infty} t_n x^n$ is not rational
- $G(x)$ is algebraic over $\mathbb{F}_2(x)$:
  $h(x, G(x)) := (x+1)^3 G(x)^2 + (x+1)^2 G(x) + x = 0$

$$\sum_{\ell=0}^{L} c_\ell t_{n+\ell} = 0 \quad \text{for } 0 \le n \le N-L-1$$

$$f(x) = \sum_{\ell=0}^{L} c_\ell x^{L-\ell} \quad \text{and} \quad g(x) = \sum_{m=0}^{L-1} \left(\sum_{\ell=L-m}^{L} c_\ell t_{m+\ell-L} \right) x^m$$

$$f(x)G(x) \equiv g(x) \bmod x^N$$

$$\begin{aligned} f(x)^2 h(x, g(x)/f(x)) &= (x+1)^3 g(x)^2 + (x+1)^2 f(x) g(x) + x f(x)^2 \\ &= K(x) x^N, \quad K(x) \neq 0 \end{aligned}$$

$$2L+1 \ge N$$

Christol's theorem

Let

$$G(x) = \sum_{n=0}^{\infty} s_n x^n$$

be the generating function of the sequence (s_n) over $\mathbb{F}_q$. Then (s_n) is q-automatic if and only if $G(x)$ is algebraic over $\mathbb{F}_q(x)$, that is, there is a polynomial $h(x,y) \in \mathbb{F}_q[x,y] \setminus \{0\}$ such that $h(x, G(x)) = 0$.

Linear complexity of automatic sequences

Mérai/W., 2018:

Let q be a prime power and (s_n) be a q-automatic sequence over $\mathbb{F}_q$ which is not ultimately periodic. Let $h(x,y) = h_0(x) + h_1(x)y + \cdots + h_d(x)y^d \in \mathbb{F}_q[x,y]$ be a non-zero polynomial $h(x, G(x)) = 0$ with no rational function $r(x) \in \mathbb{F}_q(x)$ satisfying $h(x, r(x)) = 0$.

Put

$$M = \max_{0 \le i \le d} \{\deg h_i - i\}.$$

Then we have

$$\frac{N-M}{d} \le L((s_n), N) \le \frac{(d-1)N + M + 1}{d}.$$

- Upper bound comes from (Berlekamp-Massey algorithm) $L((s_n), N+1) \in \{L((s_n), N), N+1-L((s_n), N)\}$.

- (not ultimately periodic) automatic sequences have large Nth linear complexity
- Thue-Morse sequence: $L((t_n), N) = \frac{N}{2} + O(1)$
- never $\left|L((t_n), N) - \frac{N}{2}\right| \approx \log N$
- idea does not work for (t_{n^2})

Next we study a finer measure than linear complexity.
As a consequence we get

$$L((t_{n^2}), N) \geq cN^{1/2}$$

for some $c > 0$.

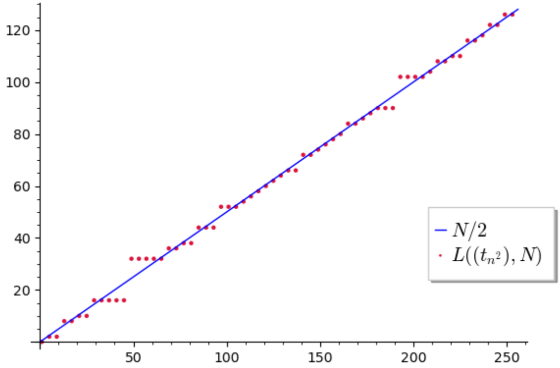


Figure: Nth linear complexity of the Thue-Morse sequence along squares.

Problem

Prove that

$$L((t_{n^2}), N) = \frac{N}{2} + o(N).$$

Maximum order complexity

The Nth maximum order complexity $M((s_n),N)$ is the smallest positive integer M with

$$s_{n+M} = f(s_{n+M-1},\ldots,s_n), \quad 0 \le n \le N-M-1,$$

for some mapping $f : \mathbb{F}_2^M \to \mathbb{F}_2$.

- $L((s_n),N) \ge M((s_n),N)$
- Jansen, 1990: expected value $\approx \log N$
- $(s_n, s_{n+1},\ldots,s_{n+M-2}) = (s_m, s_{m+1},\ldots,s_{m+M-2})$, $s_{n+M-1} \ne s_{m+M-1}$ for some $0 \le n < m \le N-M$ $\Longrightarrow M((s_n),N) \ge M$
- $C_2((s_n),N) \ge M((s_n),N) - 1$
- Chen/Gomez/Gomez/Tirkel, 2022: $C_2((s_n),N) \ge N - 2^{M((s_n),N)} + 1$
- desirable: $\log N \ll M((s_n),N) = o(N)$

Maximum order complexity of Thue-Morse sequence

Sun/W., 2019:

$$M((t_n), N) = 2^\ell + 1, \quad \text{where} \quad \ell = \left\lceil \frac{\log(N/5)}{\log 2} \right\rceil, \quad N \geq 4.$$

- $\frac{N}{5} + 1 \leq M((t_n), N) \leq 2\frac{N-1}{5} + 1$
- proof with Walnut (J. Shallit, personal communication)
- $M((t_n), N)$ too large: implies very large correlation measure of order 2 (aperiodic autocorrelation)

Maximum order complexity of Thue-Morse sequence along squares

Sun/W., 2019:

$$L((t_{n^2}),N) \ge M((t_{n^2}),N) \ge \sqrt{\frac{2N}{5}}, \quad N \ge 21.$$

- similar bounds for Rudin-Shapiro sequence, pattern sequences with the all 1 pattern
  (Rudin-Shapiro sequence (r_n):

$$r_n = \sum_{i=0}^{\infty} n_i n_{i+1}, \quad n = \sum_{i=0}^{\infty} n_i 2^i, \; n_i \in \{0,1\})$$

- Popoli, 2020: extension to polynomials of degree ≥ 2
  lower bound of order of magnitude $N^{1/d}$

Correlation measure of order k

For $k \geq 1$, the Nth correlation measure of order k of a binary sequence (s_n) is

$$C_k((s_n), N) = \max_{M,D} \left| \sum_{n=0}^{M-1} (-1)^{s_{n+d_1}} \cdots (-1)^{s_{n+d_k}} \right|,$$

where the maximum is taken over all $D = (d_1, d_2, \ldots, d_k)$ with integers satisfying $0 \leq d_1 < d_2 < \cdots < d_k$ and $1 \leq M \leq N - d_k$.

- introduced by Mauduit and Sárközy, 1997
- Alon et al., 2007: expected value $\Theta\left(\sqrt{N \log \binom{N}{k}}\right)$, $2 \leq k \leq N/4$
- $C_2((s_n), N) \geq M((s_n), N) - 1$
- Mauduit/Sárközy, 1998: $C_2((t_n), N) > \frac{N}{12}$, $N \geq 5$
- $C_2((t_n), N) \geq M((t_n), N) - 1 \geq \frac{N}{5}$

Correlation measure of order 2 of Thue-Morse sequence along squares

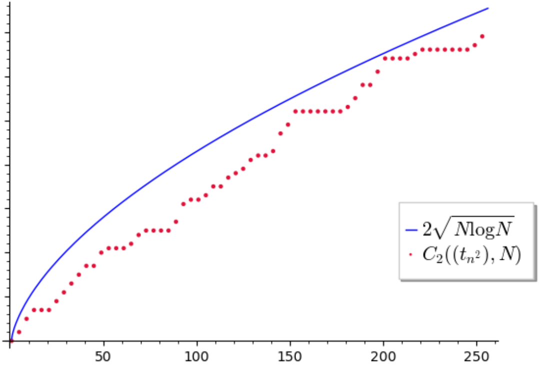


Figure: The Nth second order correlation measure of the Thue-Morse along squares.

Problem

For fixed $k = 2, 3, \ldots$ *show that*

$$C_k((t_{n^2}), N) = o(N).$$

Expansion complexity

Let (s_n) be a sequence over $\mathbb{F}_q$ with generating function

$$G(x) = \sum_{n=0}^{\infty} s_n x^n.$$

For a positive integer N, the Nth expansion complexity $E((s_n), N)$ of (s_n) is $E((s_n), N) = 0$ if $s_0 = \cdots = s_{N-1} = 0$ and otherwise the least total degree of a non-zero polynomial $h(x, y) \in \mathbb{F}_q[x, y]$ such that

$$h(x, G(x)) \equiv 0 \bmod x^N. \qquad (1)$$

$$E((s_n)) = \sup_{N \geq 1} E((s_n), N)$$

is the expansion complexity of (s_n).

Properties

- introduced by C. Diem, 2012
- $E((s_n)) < \infty \iff (s_n)$ is automatic (Christol)
- $E((t_n)) = 5$: $h(x, y) = (x+1)^3 y^2 + (x+1)^2 y + x$
- $E((t_{n^2})) = \infty$
- typical value $E((s_n), N) \approx N^{1/2}$ (Gomez/Mérai, 2020)
- $E((s_n), N) \leq \min\{L((s_n), N) + 1, N - L((s_n), N) + 2\}$ (Mérai/Niederreiter/W., 2017)

Expansion complexity of (t_{n^2})

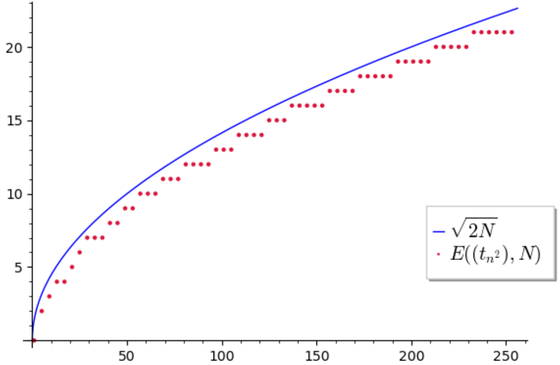


Figure: The Nth expansion complexity of the Thue-Morse sequence along squares.

Subword complexity

For a sequence (s_n) over the alphabet Δ the subword complexity $p((s_n), k)$ is the number of distinct subsequences of length k.

- $1 \le p((s_n), k) \le |\Delta|^k$
- (s_n) automatic (not ultimately periodic): $p((s_n), k)$ is of order of magnitude k

Normality

A sequence (s_n) is called normal if for any fixed length k and any pattern $\mathbf{e} \in \Delta^k$

$$\frac{\#\{0 \le n < N : (s_n, s_{n+1}, \ldots, s_{n+k-1}) = \mathbf{e}\}}{N} \to \frac{1}{|\Delta|^k},$$

as $N \to \infty$.

- (t_{n^2}) is normal (Drmota/Mauduit/Rivat, 2019)
- open: Is $(t_{f(n)})$ normal for $\deg(f) \ge 3$?
- $p((t_{n^2}), k) = 2^k$
- implies small correlation of fixed order with bounded lags $d_1 < \ldots < d_k \le B$

Analogs for finite fields

For a prime p and $q = p^r$ with $r \geq 2$ let $(\beta_1, \ldots, \beta_r)$ be an ordered basis of $\mathbb{F}_q$ over $\mathbb{F}_p$.

Thue-Morse function:

$$T\left(\sum_{i=1}^{r} x_i\beta_i\right) = \sum_{i=1}^{r} x_i, \quad x_1, \ldots, x_r \in \mathbb{F}_p$$

Rudin-Shapiro function

$$R\left(\sum_{i=1}^{r} x_i\beta_i\right) = \sum_{i=1}^{r-1} x_i x_{i+1}, \quad x_1, \ldots, x_r \in \mathbb{F}_p$$

Balance of Thue-Morse function

Dartyge/Sárközy, 2013 (using the Weil bound):
Let $f \in \mathbb{F}_q[x]$ be of degree d with $\gcd(d,q)=1$. Then for all $c \in \mathbb{F}_p$, we have

$$N_c := \left|\#\{\xi \in \mathbb{F}_q : T(f(\xi)) = c\} - p^{r-1}\right| \leq (d-1)p^{r/2}.$$

Sketch of Proof. $T(f(\xi)) = Tr(\underbrace{\sum_{i=1}^{r} \delta_i}_{=:\delta\neq 0} f(\xi))$, $\{\delta_1,\ldots,\delta_r\}$ dual basis

$$N_c = \frac{1}{p}\sum_{a\in\mathbb{F}_p}\sum_{\xi\in\mathbb{F}_q}\underbrace{\psi_p(Tr(a\delta f(\xi))-c)}_{\psi_q(a\delta f(\xi)-\eta)},$$

ψ_u additive canonical character of $\mathbb{F}_u$

Normality

Makhul/W., 2022:
Assume $1 \leq d < p$ and $s \leq d$. For any polynomial $f \in \mathbb{F}_q[x]$ of degree d and any pairwise distinct $\alpha_1, \ldots, \alpha_s \in \mathbb{F}_q$ and any $c_1, \ldots, c_s \in \mathbb{F}_p$ we have

$$\left|\#\{\xi \in \mathbb{F}_q : T(f(\xi + \alpha_i)) = c_i, 1 \leq i \leq s\} - p^{r-s}\right| \leq (d-1)p^{r/2}.$$

Balance of Rudin-Shapiro function

Dartyge/Mérai/W., 2021 (using Hooley-Katz bound):
Let $f \in \mathbb{F}_q[x]$ be of degree d with $\gcd(d, q) = 1$. Then for all $c \in \mathbb{F}_p$, we have

$$\left|\#\{\xi \in \mathbb{F}_q : R(f(\xi)) = c\} - p^{r-1}\right| \leq C_{d,r} p^{(3r+1)/4},$$

where the constant $C_{d,r}$ depends only on the degree d of f and r.

- Weil fails (degree (as univariate polynomial) too large)
- Deligne fails ((multivariate) polynomial has singular points)
- Hooley-Katz is generalization of Deligne for non-singular polynomials

The Hooley-Katz Theorem, 1991

We denote by $\overline{\mathbb{F}_p}$ the algebraic closure of $\mathbb{F}_p$.
The (affine) singular locus $\mathcal{L}(F)$ of a polynomial F over $\mathbb{F}_p$ in r variables is the set of common zeros in $\overline{\mathbb{F}_p}^r$ of the polynomials

$$F, \frac{\partial F}{\partial X_1}, \ldots, \frac{\partial F}{\partial X_r}.$$

Let Q be a polynomial over $\mathbb{F}_p$ in r variables of degree $D \geq 1$ such that the dimensions of the singular loci of Q and its homogeneous part Q_D of degree D satisfy

$$\max\{\dim(\mathcal{L}(Q)), \dim(\mathcal{L}(Q_D)) - 1\} \leq s.$$

Then the number N of zeros of Q in $\mathbb{F}_p^r$ satisfies

$$\left|N - p^{r-1}\right| \leq C_{D,r} p^{(r+s)/2}.$$

$s = -1$: Deligne

Problem

Study the normality of the Rudin-Shapiro function at $f(x)$. *Namely, show that*

$$\frac{\#\{\xi \in \mathbb{F}_q : R(f(\xi+\alpha_i)) = c_i, 1 \le i \le s\}}{p^{r-s}} \to 1 \quad \textit{as } p \to \infty$$

for some $s \geq 2$ *and any* $f \in \mathbb{F}_q[x]$ *of fixed degree.*

Further reading

L. Mérai, A. Winterhof, Pseudorandom sequences derived from automatic sequences. Cryptogr. Commun. 14 (2022), no. 4, 783–815.

Further reading

L. Mérai, A. Winterhof, Pseudorandom sequences derived from automatic sequences. Cryptogr. Commun. 14 (2022), no. 4, 783–815.

Thank you for your attention!