Well distributed occurrences property in infinite words

S. Puzynina (joint work with M. Bucci, A. de Luca, L. Dvořáková, J. Hladký, V. Schavelev)

Outline

Outline:

- WellDoc: abelian-type property of infinite words
- Motivation: PRNGs, lattice structure
- Welldoc for Sturmian and AR words
- Welldoc for morphic words

The talk is based on

- L. Balková, M. Bucci, A. De Luca, J. Hladký, S. Puzynina: Aperiodic pseudorandom number generators based on infinite words. Theor. Comput. Sci. 647: 85-100 (2016)
- S. Puzynina, V. Schavelev, Welldoc property of morphic words, 2023 [in preparation]

WellDoc Property

Alphabet: $\mathcal{A} = \{0, 1, \ldots, d-1\}$

w a finite word

Parikh vector of w: $\mathrm{PV}(w) = (|w|_0, |w|_1, \ldots, |w|_{d-1})$.

Example: $\mathrm{PV}(0102210) = (3, 2, 2)$.

u finite or infinite word

$\mathrm{Pref}_n\, u$ the prefix of length n of u: $\mathrm{Pref}_n\, u = u_0 u_1 \cdots u_{n-1}$.

WellDoc Property

Alphabet: $\mathcal{A} = \{0, 1, \ldots, d-1\}$
u an aperiodic infinite word

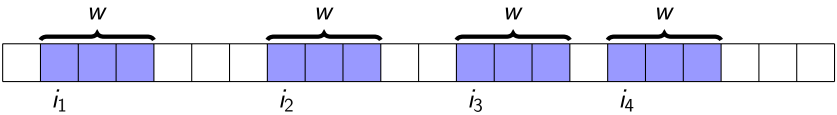
Definition (The WELLDOC property)

u has well distributed occurrences (or has the WELLDOC property) if for each $m \in \mathbb{N}$ and each factor w of u we have

$$\left\{ \left(|\operatorname{Pref}_{i_j} u|_0, \ldots, |\operatorname{Pref}_{i_j} u|_{d-1} \right) \bmod m \mid j \in \mathbb{N} \right\} = \mathbb{Z}_m^d ,$$

where $i_0, i_1, \ldots$ are the positions of the occurrences of w in u.

that is, the Parikh vectors of $\operatorname{Pref}_{i_j} u$ for $j \in \mathbb{N}$, when reduced modulo m, give the whole set $\mathbb{Z}_m^d$.



Example: Fibonacci word

Example

The Fibonacci word has the WellDoc property.

E.g., for 001 and $m = 2$:

$0100101001001010010100100101001001010\cdots$

i_1 i_2 i_3 i_4 i_5 i_6 i_7 i_8

i_j	$\mathrm{Pref}_{i_j}\, u$	$\mathrm{PV}\ (\mathrm{Pref}_{i_j}\, u)$	$\mathrm{PV} \mod 2$
$i_1 = 2$	01	$(1, 1)$	$(1, 1)$
$i_2 = 7$	0100101	$(4, 3)$	$(0, 1)$
$i_3 = 10$	0100101001	$(6, 4)$	$(0, 0)$
$i_4 = 15$	010010100100101	$(9, 6)$	$(1, 0)$

Example: Thue-Morse word

Example

The Thue-Morse word

$01101\underbrace{00}_{5}11\underbrace{00}_{9}101101\underbrace{00}_{17}1011\underbrace{00}_{23}1101\underbrace{00}_{29}1\cdots$

does not satisfy the WellDOc property.

Indeed, e.g. for $m = 2$ the factor $w = 00$ occurs only in odd positions i_j, so that $(|\operatorname{Pref}_{i_j} u|_0 + |\operatorname{Pref}_{i_j} u|_1) = i_j$ is odd. Thus

$$\{(|\operatorname{Pref}_{i_j} u|_0, |\operatorname{Pref}_{i_j} u|_1) \bmod 2 \mid j \in \mathbb{N}\} \neq \mathbb{Z}_2^2.$$

Universal word

An infinite word u on an alphabet $\mathcal{A}$ is universal if it contains all finite words over $\mathcal{A}$ as its factors.

Example

Any universal word u satisfies the WellDOc property:

- given m and w, arrange a word v containing w at positions with prefixes congruent to all vectors from $\mathbb{Z}_m^d$;
- w is universal $\Rightarrow$ v is its factor
- $\Rightarrow$ u also has all vectors from $\mathbb{Z}_m^d$, just shifted.

Recurrent word

Recurrent word = each factor occurs infinitely often.

Remark

If a recurrent infinite word u has the WellDOc property, then for each vector $\mathbf{v} \in \mathbb{Z}_m^d$ there are infinitely many values of j such that $\mathrm{PV}(\mathrm{Pref}_{i_j}\, u) \equiv \mathbf{v} \mod m$.

Pseudoranom Number Generators

Pseudorandom number generators:

- aim to produce random numbers using a deterministic process.
- not truly random, because it is completely determined by an initial value (seed)

For us:
Pseudorandom number generator (PRNG) with output $M \subset \mathbb{N}$, M finite, is an infinite word $Z = (Z_n)_{n \in \mathbb{N}}$ on the alphabet M.

Class of PRNGs:
A linear congruential generator (LCG) $(Z_n)_{n \in \mathbb{N}}$ with parameters $a, m, c \in \mathbb{N}$ is defined by the recurrence relation

$$Z_{n+1} = aZ_n + c \mod m.$$

Lattice Structure

$Z = (Z_n)_{n \in \mathbb{N}}$: a PRNG with output $M \subset \mathbb{N}$, M finite.

Z has the lattice structure if

- there exists $t \in \mathbb{N}$ such that the set

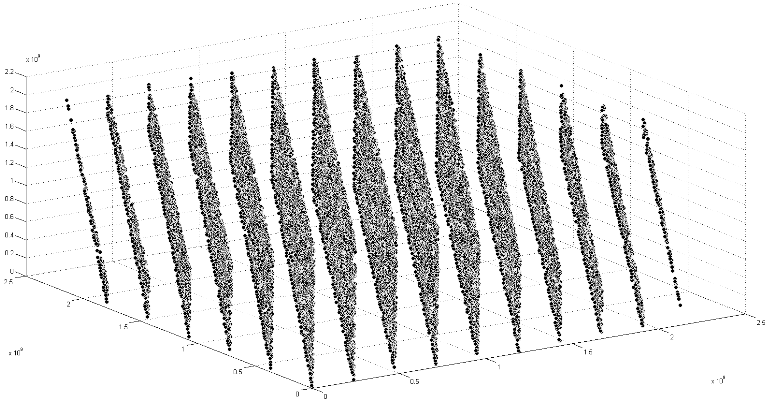
$$\{(Z_i, Z_{i+1}, \ldots, Z_{i+t-1}) \mid i \in \mathbb{N}\}$$

  is covered by a family of parallel equidistant hyperplanes and
- this family does not cover the whole lattice M^t.

Lattice Structure: example

Example: RANDU, the LCG with $a = (2^{16} + 3), m = 2^{31}, c = 0$.

For $t = 3$, the triples of RANDU, i.e., $\{(Z_i, Z_{i+1}, Z_{i+2}) \mid i \in \mathbb{N}\}$, are covered by 15 parallel equidistant planes:



Combining PRNGs using infinite words

- $X = (X_n)_{n\in\mathbb{N}}$ and $Y = (Y_n)_{n\in\mathbb{N}}$ PRNGs with the same output $M \subset \mathbb{N}$ and the same period $m \in \mathbb{N}$
- $u = u_0 u_1 u_2 \cdots$ a binary infinite word

The PRNG $Z = (Z_n)_{n\in\mathbb{N}}$ based on u is obtained as follows:

- replace the occurrences of 0's in u with the word X
- replace the occurrences of 1's in u with the word Y

Example

0100101001001010010100 1 $\cdots$

$X_0 Y_0 X_1 X_2 Y_1 X_3 Y_2 X_4 X_5 Y_3 X_6 X_7 Y_4 X_8 Y_5 X_9 X_{10} Y_6 X_{11} Y_7 X_{12} X_{13} Y_8 \cdots$

In the same way one can take a non-binary word and combine several PRNGs.

Welldoc and PRNGs

Theorem (Bucci, De Luca, Dvořáková, Hladký, P., 2016)

Let Z be the PRNG based on an infinite word u with the WellDOc property. Then Z has no lattice structure.

Welldoc and PRNGs

Theorem (Bucci, De Luca, Dvořáková, Hladký, P., 2016)

Let Z be the PRNG based on an infinite word u with the WellDOc property. Then Z has no lattice structure.

Remark

WellDOc is not necessary for absence of the lattice structure.

Example

Consider a modified Fibonacci word $\hat{u}$ where the letter 2 is inserted after each letter, i.e., $\hat{u} = 021202021202120 2 \ldots$.

- $\hat{u}$ does not have WellDOc.
- PRNG obtained by combining three generators according to $\hat{u}$ has no lattice structure.

Sturmian words

Theorem (Bucci, De Luca, Dvořáková, Hladký, P., 2016)

Let u be a Sturmian word. Then u has the WELLDOC property.

Definition

The rotation by angle α is the mapping $R_\alpha : [0,1) \mapsto [0,1)$ defined by $R_\alpha(x) = \{x + \alpha\}$, where $\{x\}$ is the fractional part of x.

$I_0 = [0, 1 - \alpha)$, $I_1 = [1 - \alpha, 1)$, $[0,1) = I_0 \cup I_1$.

Definition of Sturmian words via rotations

$$s_{\alpha,\rho}(n) = \begin{cases} 0 & \text{if } R_\alpha^n(\rho) = \{\rho + n\alpha\} \in I_0, \\ 1 & \text{if } R_\alpha^n(\rho) = \{\rho + n\alpha\} \in I_1. \end{cases}$$

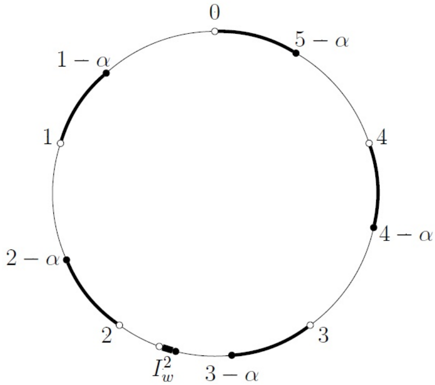
(One can also take $I_0' = (0, 1 - \alpha]$, $I_1' = (1 - \alpha, 1]$).

Sturmian words: Proof

- We will prove $\{(i_j, |\operatorname{Pref}_{i_j} u|_1)\} \mod m = (\mathbb{Z}_m)^2$ (this is equivalent to $\{PV(\operatorname{Pref}_{i_j} u)\} \mod m = (\mathbb{Z}_m)^2$).

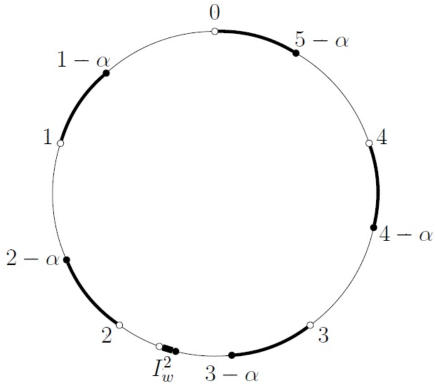
Sturmian words: Proof

- We will prove $\{(i_j, |\operatorname{Pref}_{i_j} u|_1)\} \mod m = (\mathbb{Z}_m)^2$ (this is equivalent to $\{PV(\operatorname{Pref}_{i_j} u)\} \mod m = (\mathbb{Z}_m)^2$).
- Take a circle of length m by copying m times $[0, 1)$:



Sturmian words: Proof

- We will prove $\{(i_j, |\text{Pref}_{i_j} u|_1)\} \mod m = (\mathbb{Z}_m)^2$ (this is equivalent to $\{PV(\text{Pref}_{i_j} u)\} \mod m = (\mathbb{Z}_m)^2$).
- Take a circle of length m by copying m times $[0, 1)$:



- Any factor w of u corresponds to an interval I_w in $[0, 1)$, so that rotating starting from I_w gives w.
- We define m intervals corresponding to w in $[0, m)$.

Sturmian words: Proof

- Take arbitrary $(j, i) \in \mathbb{Z}_m^2$.
  We need to find l such that
  - $u_l \dots u_{l+|w|-1} = w$,
  - $|\mathrm{Pref}_l u|_1 \mod m = i$,
  - $l \mod m = j$.

Sturmian words: Proof

- Take arbitrary $(j, i) \in \mathbb{Z}_m^2$.
  We need to find l such that
  - $u_l \ldots u_{l+|w|-1} = w$,
  - $|\mathrm{Pref}_l u|_1 \mod m = i$,
  - $l \mod m = j$.
- Consider rotation $R_{m\alpha,m}(j\alpha + \rho)$ by $m\alpha$ in m-circle.
- This rotation will put us to positions $mk + j$, $k \in \mathbb{N}$, in the Sturmian word
- The points in the orbit of this rotation on the m-circle are dense, and hence the rotation comes infinitely often to each interval.
- Pick k when $j\alpha + mk\alpha + \rho \in I_w^i \subset [i, i+1)$.
- We have $l = km + j$.

Arnoux-Rauzy words

Theorem (Bucci, De Luca, Dvořáková, Hladký, P., 2016)

Let u be an Arnoux-Rauzy word over the d-letter alphabet $\mathcal{A}$. Then u has the WellDOc property.

The proof is based on the definition via palindromic closures.

Binary morphic words

For a morphism φ, its matrix is defined by $A_\varphi = (|\varphi(j)|_i)_{i,j\in\mathcal{A}}$.

Theorem (P., Schavelev, 2023)

Let u be an infinite binary word generated by a primitive morphism φ. Then u satisfies WellDOc if and only if $\det A_\varphi = \pm 1$.

Example

Thue-Morse word: $\tau : 0 \mapsto 01, 1 :\mapsto 10$, $A_\tau = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $\det A_\tau = 0$, so the Thue-Morse word does not have WellDOc.

Example

Fibonacci word: $f : 0 \mapsto 01, 1 :\mapsto 0$, $A_f = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, $\det A_\tau = -1$, so the Fibonacci has WellDOc.

Nonbinary morphic words

Definition

u a recurrent infinite word, v its factor,

let $i_0 < i_1 < \dots$ be all integers i_j such that $v = u_{i_j} \cdots u_{i_j+|v|-1}$

$u_{i_j} \cdots u_{i_{j+1}-1}$ is a return word of v in u.

Theorem (P., Schavelev, 2023)

Let u be an infinite word generated by a primitive morphism φ. Then u satisfies WellDOc if and only if $\det A_\varphi = \pm 1$ and Parikh vectors of returns to u_0 generate $\mathbb{Z}^{|\mathcal{A}|}$ as additive group.

Morphic words: condition on returns

Example

Consider a morphism φ:

$$\varphi : \begin{array}{l} 0 \to 02, \\ 1 \to 101, \\ 2 \to 102, \end{array} \qquad A_\varphi = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

$\varphi^\infty(0) = 021021010210210102101021021010210 2\cdots$

- $\det A_\phi = 1$,
- the returns to 0 are only factors 01, 021, so the prefixes before 0 are in $\langle (1,1,0),(1,1,1) \rangle$.
- hence no WellDOc by definition

Morphic words

For necessity: recognizability for primitive morphisms:

Recognizability

Morphism φ generating an infinite word u is called recognizable if there exists $L > 0$ such that for each factor v of u of length at least $2L$ there exist integers $i, j \in \mathbb{N}, 0 \leq i < L, |v| - L \leq j < |v|$ and a factor w such that $u[i, j) = \varphi(w)$ and for each m such that $u[m, m + |v|) = v$ there exist i', j' such that $m + i = |\varphi(Pref_u(i'))|$ and $m + j = |\varphi(Pref_u(j'))|$ and $u[i', j') = w$.

Theorem (Mossé, 1992)

Every primitive aperiodic morphism is recognizable.

Conclusion

- Statistical tests show that mixing PRNGs according to a word with WellDOc gives better PRNGs.
  Apart from lattice structure, what other statistical properties are improved by WellDOc-mixing? Can we use some other words properties?
- Can we prove the characterization without using recognizability (also for non-primitive)?