Bounds for the discrete correlation of infinite sequences on k symbols and generalized Rudin-Shapiro sequences

E. Grant, J. Shallit, T. Stoll

February 11, 2009

1 Introduction

Pseudorandom sequences, i.e., deterministic sequences on finite alphabets with properties reminiscent of random sequences, are an intensively studied subject. We refer to the series of papers by Mauduit, Sárközy and coauthors [1, 4, 5, 12, 13] among many others. A great part of the mentioned work deals with correlation measures for binary sequences and the problem to find large classes of finite pseudorandom binary sequences with small autocorrelation. Let $x = x_0x_1\cdots x_N \in \{-1,1\}^N$ be a finite word over the alphabet $\{-1,1\}$. Then the correlation measure of order m of x is defined as

$$U_m(x) = \max_{M,\boldsymbol{r}} \left| \sum_{n=0}^{M} x_{n+r_1}x_{n+r_2}\cdots x_{n+r_m} \right|, \tag{1.1}$$

where the maximum is taken over all $\boldsymbol{r} = (r_1, r_2, \ldots, r_m)$ with $0 \le r_1 < r_2 < \cdots < r_m$ and M such that $M + r_m \le N$. In case of infinite words $x = x_0x_1\cdots$ the correlation of order m is defined as

$$V_m(x, M) = \sum_{n=0}^{M} x_{n+r_1}x_{n+r_2}\cdots x_{n+r_m}, \tag{1.2}$$

with fixed $\boldsymbol{r}$. In contrast to $U_m(x)$, this definition does not take "large-range correlations" into account. In fact, r_m could be $\Omega(N)$ for the finite word correlation [12]. Recently, Mauduit and Sárközy [14] generalized several measures for pseudorandomness to finite sequences over k-letter alphabets. These distribution measures have been studied by Bérczi [3] from a probabilistic point of view.

The aim of the present paper is to study the *discrete correlation* among members of arbitrary infinite sequences over k symbols, where we just take into account whether two symbols are identical. In the sequel, we denote by $\mathbb{N}$ the set of non-negative integers, and we assume that sums start with index 0 (empty sums are supposed to be zero), unless otherwise stated. We further denote by $n \bmod k$ the unique integer n' with $0 \le n' \le k-1$ and $n \equiv n' \pmod{k}$. We use "word" and "sequence" interchangeably.

Let $x = x_0x_1\cdots$ be an infinite word over an alphabet of size k. Without loss of generality we may assume that $x_i \in \{0, 1, \ldots, k-1\}$ for $i \in \mathbb{N}$. For vectors $(i_1, i_2, \ldots, i_m)$ with integers i_j $(1 \le j \le m)$ satisfying $0 \le i_1 < i_2 < \cdots < i_m$, define the *discrete correlation coefficient* $\delta(i_1, i_2, \ldots, i_m)$ *of order* m by

$$\delta(i_1, i_2, \ldots, i_m) = \begin{cases} 0, & \text{if } x_{i_1} = x_{i_2} = \cdots = x_{i_m}; \\ 1, & \text{otherwise.} \end{cases}$$

Moreover, define $C_{\boldsymbol{r}}$ for all fixed $\boldsymbol{r} = (r_1, r_2, \ldots, r_m)$ with $0 \le r_1 < r_2 < \cdots < r_m$ by

$$C_{\boldsymbol{r}} = \liminf_{N\to\infty} \frac{1}{N} \sum_{n<N} \delta(n + r_1, n + r_2, \ldots, n + r_m). \tag{1.3}$$

It is important to remark that for a random sequence (where every symbol is independently chosen with probability 1/k) the quantity $C_{\boldsymbol{r}}$ equals $1 - 1/k^{m-1}$ with probability one. In this paper we investigate sequences with respect to this leading term. We first show by combinatorial means that for any infinite sequence on k symbols the quantity $C_{\boldsymbol{r}}$ cannot be too large for all $\boldsymbol{r}$ (Theorem 2.3). Our result, however, does not rule out the existence of deterministic sequences that actually attain our bound. We provide such a construction in the case of $m = 2$ by introducing *generalized Rudin-Shapiro sequences on* k *symbols*, which extends a construction by Queffélec [15] and Høholdt, Jensen and Justesen [7, 8]. The motivation stems from the fact that the autocorrelation $C_{(r_1,r_2)}$ of the infinite Rudin-Shapiro sequence on two symbols is small [13, Theorem 4]. Our construction, however, gives a large class of sequences with small autocorrelation for any alphabet with cardinality k, whenever k is prime or squarefree.

The paper is structured as follows. In Section 2 we state the general bounds for the discrete correlation in Theorems 2.3 and 2.4. In Section 3 we give the definition of generalized Rudin-Shapiro sequences. Sections 4 and 5 are devoted to the combinatorial proofs of Theorem 2.3 and 2.4, respectively. In Section 6 we give the proof of Theorem 2.6 by using the Lovász local lemma. Finally, in Sections 7 and 8 we give the proofs for Theorems 3.1 and 3.3 by means of exponential sums.

2 General bounds for the discrete correlation

We wish to establish upper bounds for $C_{\boldsymbol{r}}$ as $\boldsymbol{r}$ gets "large". To begin with, we normalize the vector $\boldsymbol{r}$. For an integer sequence $T = (t_0, t_1, \ldots)$ with $t_i + r_1 \ge 0$ for $i \in \mathbb{N}$, we define shifted versions of $C_{\boldsymbol{r}}$, namely,

$$C_{\boldsymbol{r},T} = \liminf_{N\to\infty} \frac{1}{N} \sum_{n<N} \delta(n + t_N + r_1, n + t_N + r_2, \ldots, n + t_N + r_m).$$

Proposition 2.1. *Let* $\boldsymbol{r} = (r_1, r_2, \ldots, r_m)$ *with* $0 \le r_1 < r_2 < \cdots < r_m$, *and let* $T = (t_0, t_1, \ldots)$ *be a sequence of integers with* $t_i + r_1 \ge 0$ *for all* i. *If* $t_N = o(N)$, *then* $C_{\boldsymbol{r},T} = C_{\boldsymbol{r}}$.

Proof. We note that

$$C_{\boldsymbol{r},T} = \liminf_{N\to\infty} \frac{1}{N} \sum_{n=t_N}^{N+t_N-1} \delta(n+r_1, n+r_2, \ldots, n+r_m).$$

Since $\delta(n+r_1, n+r_2, \ldots, n+r_m) \in \{0,1\}$ for all n, the above sum differs from the corresponding sum in (1.3) by at most $2t_N$. Thus if $t_N = o(N)$, then

$$C_{\boldsymbol{r},T} = \liminf_{N\to\infty} \frac{1}{N} \left(\sum_{n<N} \delta(n+r_1, n+r_2, \ldots, n+r_m) + o(N) \right) = C_{\boldsymbol{r}}. \qquad \square$$

By taking $T = (t, t, \ldots)$, Proposition 2.1 implies that $C_{\boldsymbol{r}+t\boldsymbol{1}} = C_{\boldsymbol{r}}$ for all constants $t \geq -r_1$. We shall say $\boldsymbol{r}$ is *normalized* whenever $r_1 = 0$ and $r_1 < r_2 < \cdots < r_m$, and henceforth only consider normalized $\boldsymbol{r}$. In the $m = 2$ case, we then have $\boldsymbol{r} = (0, r_2)$ and we can establish an upper bound by taking the limit as r_2 approaches infinity. We shall obtain the following result.

Theorem 2.2. *Let x be an infinite word over an alphabet of size k. Then*

$$\liminf_{r_2\to\infty} C_{(0,r_2)} \leq 1 - \frac{1}{k}. \tag{2.1}$$

In the next section we provide the construction of deterministic sequences with equality in (2.1). More precisely, we show that for generalized Rudin-Shapiro sequences (k prime or squarefree) we have

$$\inf_{r_2>0} \{C_{(0,r_2)}\} = 1 - \frac{1}{k}.$$

To generalize Theorem 2.2 to larger values of m, we must precisely define the notion of "$\boldsymbol{r}$ getting large". Let $||\cdot||$ be a norm on the finite dimensional vector space $\mathbb{R}^m$. We will prove the following upper bound on $C_{\boldsymbol{r}}$ as $||\boldsymbol{r}||$ tends to infinity:

Theorem 2.3. *Let x be an infinite word over an alphabet of size k. Then for any $m \geq 2$ and any norm $||\cdot||$, we have*

$$\lim_{\lambda\to\infty} \left(\inf \{C_{\boldsymbol{r}} : \boldsymbol{r} \in \mathbb{N}^m,\ \boldsymbol{r} \text{ normalized},\ ||\boldsymbol{r}|| \geq \lambda\} \right) \leq 1 - \frac{1}{k^{m-1}}. \tag{2.2}$$

We note that Theorem 2.2 is immediately implied by Theorem 2.3 by taking $m = 2$. Theorem 2.3 is proven via a combinatorial argument in Section 4.

In order to also consider the local autocorrelation properties of sequences, we define a related quantity. Again, let x be an infinite word over an alphabet of size k. For a given vector $\boldsymbol{r}$ and positive integers d, we define

$$D^d_{\boldsymbol{r}} = \min_{n\geq 0} \left(\frac{1}{d} \sum_{i=n}^{n+d-1} \delta(i+r_1, i+r_2, \ldots, i+r_m) \right). \tag{2.3}$$

Note that for a random sequence on k symbols, we necessarily have $D^d_{\boldsymbol{r}} = 0$ for all $\boldsymbol{r}$ and d. We will prove that for a given vector $\boldsymbol{r}$, the value of $C_{\boldsymbol{r}}$ of an infinite sequence is an upper bound for all of the values of $D^d_{\boldsymbol{r}}$:

Theorem 2.4. *Let x be an infinite word over an alphabet of size k, $\boldsymbol{r}$ be normalized and $d > 0$. Then $D_{\boldsymbol{r}}^d \le C_{\boldsymbol{r}}$.*

As an immediate consequence of Theorem 2.3 and Theorem 2.4, we obtain an upper bound on $D_{\boldsymbol{r}}^d$ as $||\boldsymbol{r}||$ tends to infinity.

Corollary 2.5. *Let x be an infinite word over an alphabet of size k. Then for any $m \ge 2$, $d > 0$, and norm $||\cdot||$, we have*

$$\lim_{\lambda\to\infty} \left(\inf\left\{D_{\boldsymbol{r}}^d : \boldsymbol{r} \in \mathbb{N}^m,\ \boldsymbol{r} \text{ normalized},\ ||\boldsymbol{r}|| \ge \lambda\right\}\right) \le 1 - \frac{1}{k^{m-1}}. \tag{2.4}$$

An interesting example occurs when we choose a fixed $d > 0$ and take

$$\boldsymbol{r} = (0, d, 2d, \ldots, (m-1)d).$$

Then for each subword $w_1 w_2 \cdots w_m$ of x with $|w_i| = d$ for all i, the number of indices j where $|\{w_i[j] : 1 \le i \le m\}| > 1$ is at least $dD_{\boldsymbol{r}}^d$. In this case, for sufficiently large d, we can get arbitrarily close to the bound in (2.4).

Theorem 2.6. *For all $\varepsilon > 0$ there exist an infinite word x over an alphabet of size k and $d_0 = d_0(\varepsilon)$ such that for all $d > d_0$ and $\boldsymbol{r} = (0, d, 2d, \ldots, (m-1)d)$ we have*

$$D_{\boldsymbol{r}}^d \ge 1 - \frac{1}{k^{m-1}} - \varepsilon.$$

3 Generalized Rudin-Shapiro sequences

The quantity $C_{\boldsymbol{r}}$ has been studied for various special sequences. A classical result of Mahler [10] states that for the Thue-Morse sequences over k symbols, the summatory correlation has no uniform leading term. On the contrary, Queffélec [15] noted (referring to an unpublished result by Kamae) that the Rudin-Shapiro sequence indeed has the desired leading term, whenever r is fixed. As for the hub of the present article, Mauduit and Sárközy [13, Corollary after Theorem 4] showed that for the correlation of order 2 one may let $r_2 = o(N)$ without losing this property. The following definition gives an extension to alphabets of size $k \ge 2$.

Definition 3.1. Let $g : \{0, 1, \ldots, k-1\} \times \mathbb{Z} \to \mathbb{Z}$, $(j, n) \mapsto g(j, n)$ be a function which is periodic in n with period k. Furthermore, let g be such that for all integers u, i with $0 \le u < u + i \le k - 1$ we have

$$\{\ (g(u+i, n) - g(u, n)) \bmod k :\ 0 \le n \le k-1\ \} = \{\ 0, 1, \ldots, k-1\ \}.$$

Then we call a sequence $(\hat{a}(n))_{n\ge 0}$ over the alphabet $\{0, 1, \ldots, k-1\}$ a *generalized Rudin-Shapiro sequence* if there exists a sequence of integers $(a(n))_{n\ge 0}$ such that $\hat{a}(n) \equiv a(n) \pmod k$ and

$$a(nk + j) = a(n) + g(j, n), \qquad 0 \le j \le k-1, \quad n \ge 1. \tag{3.1}$$

The function g is called an *admissible function*.

Example 1: A "canonical" admissible function g in the sense of Definition 3.1 is

$$g(j,n) = j \cdot (n \bmod k), \tag{3.2}$$

which is Queffélec's generalization for the ordinary Rudin-Shapiro sequence [15, Section 4]. In this case $g(u+i,n) - g(u,n) \equiv in \pmod k$, and $\{in : \ 0 \le n \le k-1\}$ runs for i with $0 \le i \le k-1$ through all residue classes mod k, provided k is prime. In particular, for $k = 2$ and

$$g(j,n) = \begin{cases} 1, & \text{if } j = 1,\ n \equiv 1 \pmod 2; \\ 0, & \text{otherwise} \end{cases}$$

we get the *Rudin-Shapiro sequence over the alphabet* $\{0,1\}$, namely,

$$(\hat{a}(n))_{n\ge 0} = 0,0,0,1,0,0,1,0,\ldots,$$

where the corresponding sequence $a(n)$ counts the number of subblocks $(1,1)$ in the binary expansion of n.

Example 2: For $k = 2$ and appropriate initial conditions, we get sequences which count *any* fixed block of size two. For instance, by setting

$$g(1,0) = 1, \qquad g(0,0) = g(1,1) = g(0,1) = 0,$$

the resulting sequence $(\hat{a}(n))_{n\ge 0}$ counts (mod 2) the number of subblocks (01) in the binary expansion of n.

Example 3: For $k = 3$ an admissible function other than (3.2) is given by

$$g(j,n) = \begin{cases} 1, & \text{if } j \equiv n \pmod 3; \\ 0, & \text{otherwise.} \end{cases}$$

Here, the resulting sequence $(\hat{a}(n))_{n\ge 0}$ (with initial conditions $\hat{a}(0) = \hat{a}(1) = \hat{a}(2) = 0$) gives the cumulative number of appearances (mod 3) of subblocks (00), (11) and (22) in the ternary expansion of integers.

The following theorem shows that generalized Rudin-Shapiro sequences resemble the discrete autocorrelation behavior of random sequences if $m = 2$.

Theorem 3.1. *Let*

$$\hat{a}(0), \hat{a}(1), \hat{a}(2), \ldots$$

be a generalized Rudin-Shapiro sequence over $\{0,1,\ldots,k-1\}$ *with* k *prime. Moreover, let* $0 \le r_1 < r_2$. *Then, as* $N \to \infty$, *we have*

$$\sum_{n<N} \delta(n+r_1, n+r_2) = \left(1 - \frac{1}{k}\right) N + O_k\left((r_2 - r_1)\log\frac{N}{r_2 - r_1} + r_2\right), \tag{3.3}$$

where the implied constant only depends on k.

In the proof, we give an explicit value for the implied constant. As an immediate consequence we note

Corollary 3.2. *In the setting of Theorem 3.1, if $r_2 = o(N)$ then*

$$\sum_{n<N} \delta(n+r_1, n+r_2) \sim \left(1 - \frac{1}{k}\right) N.$$

It seems natural to consider the cross product of two generalized Rudin-Shapiro sequences to prime bases to construct an extremal sequence for squarefree k. Let $k = p_1 p_2 \cdots p_d$ be a product of pairwise distinct primes, and put $c_1 = 1$, $c_i = p_1 p_2 \cdots p_{i-1}$ for $2 \le i \le d$. We define the sequence $(\hat{a}(n))_{n\ge 0}$ by

$$\hat{a}(n) = a(n) \bmod k, \tag{3.4}$$

where $(a(n))_{n\ge 0}$ is defined by

$$a(n) = c_1 a_1(n) + c_2 a_2(n) + \cdots + c_d a_d(n). \tag{3.5}$$

Herein, $(a_i(n))_{n\ge 0}$ satisfies the recursive relation

$$a_i(p_i n + j) = a_i(n) + g_i(j, n), \qquad 1 \le i \le d, \tag{3.6}$$

for $n \ge 1$ and $0 \le j \le p_i - 1$. Again, the functions g_i are admissible functions in the sense of Definition 3.1 for $1 \le i \le d$. Our next result gives an estimate for the correlation of order two.

Theorem 3.3. *Let $k = p_1 p_2 \cdots p_d$ with $d \ge 2$ be squarefree and denote by*

$$\hat{a}(0), \hat{a}(1), \hat{a}(2), \ldots$$

a generalized Rudin-Shapiro sequence over $\{0, 1, \ldots, k-1\}$ defined by (3.4), (3.5) and (3.6). Moreover, let $0 \le r_1 < r_2$ and $0 < \gamma < 1$. Then, as $N \to \infty$, we have

$$\sum_{n<N} \delta(n+r_1, n+r_2) = \left(1 - \frac{1}{k}\right) N + O_k\Big((r_2 - r_1) N^{1-\gamma/d} + (r_2 - r_1) N^{1-\gamma} \log \frac{N^{\gamma/d}}{r_2 - r_1} + N^{\gamma} + r_1\Big), \tag{3.7}$$

where the implied constant only depends on k.

Corollary 3.4. *In the setting of Theorem 3.3, if $r_2 = o(N^{\gamma/d})$ then*

$$\sum_{n<N} \delta(n+r_1, n+r_2) \sim \left(1 - \frac{1}{k}\right) N.$$

4 Proof of Theorem 2.3

We need the following lemma for our proof of Theorem 2.3.

Lemma 4.1. *Suppose we have a multiset of n distinct objects of k types, and let $d \leq n$ be a fixed constant. Then among the $\binom{n}{d}$ subsets of d objects, the number containing at least one pair of objects of different types is at most*

$$\frac{n^d}{d!}\left(1 - \frac{1}{k^{d-1}}\right).$$

Proof. Suppose we have b_i objects of type i for all $1 \leq i \leq k$. Then we have $\binom{b_i}{d}$ subsets consisting entirely of objects of type i. Thus the total number of subsets P that contain at least one pair of objects of different types is

$$\begin{aligned} P &= \binom{n}{d} - \sum_{i=1}^{k} \binom{b_i}{d} \\ &= \frac{1}{d!}\left(n(n-1)\cdots(n-d+1) - \sum_{i=1}^{k} b_i(b_i-1)\cdots(b_i-d+1)\right). \end{aligned}$$

Consider the polynomial $\phi(x) = x(x-1)\cdots(x-d+1) = e_1 x + \cdots + e_d x^d$. We rewrite our expression for P in terms of ϕ,

$$\begin{aligned} P &= \frac{1}{d!}\left(\phi(n) - \sum_{i=1}^{k} \phi(b_i)\right) \\ &= \frac{1}{d!}\left(\phi(n) - \left(e_1 \sum_{i=1}^{k} b_i + e_2 \sum_{i=1}^{k} b_i^2 + \cdots + e_d \sum_{i=1}^{k} b_i^d\right)\right). \end{aligned}$$

By the power means inequality,

$$\frac{n}{k} = \frac{1}{k}\sum_{i=1}^{k} b_i \leq \left(\frac{1}{k}\sum_{i=1}^{k} b_i^\nu\right)^{1/\nu} \qquad \text{for all } \nu \geq 1,$$

and thus

$$\left(\frac{n^\nu}{k^{\nu-1}}\right) \leq \sum_{i=1}^{k} b_i^\nu.$$

We apply this bound to our expression for P to yield the desired result,

$$\begin{aligned} P &\leq \frac{1}{d!}\left(\phi(n) - \left(e_1 n + e_2\frac{n^2}{k} + \cdots + e_d\frac{n^d}{k^{d-1}}\right)\right) \\ &= \frac{1}{d!}\left(n(n-1)\cdots(n-d+1) - k\cdot\frac{n}{k}\left(\frac{n}{k}-1\right)\cdots\left(\frac{n}{k}-d+1\right)\right) \\ &\leq \frac{1}{d!}\left(n^d - k\left(\frac{n}{k}\right)^d\right) \\ &= \frac{n^d}{d!}\left(1 - \frac{1}{k^{d-1}}\right). \end{aligned}$$

□

With our lemma in hand, we now prove Theorem 2.3. We proceed via contradiction. Suppose that for some $m \geq 2$ and some norm $||\cdot||$ on $\mathbb{R}^m$, there exists an $\varepsilon > 0$ such that

$$\lim_{\lambda\to\infty}\left(\inf\left\{C_{\boldsymbol{r}} : \boldsymbol{r}\in\mathbb{N}^m,\ \boldsymbol{r} \text{ normalized},\ ||\boldsymbol{r}||\geq\lambda\right\}\right) = 1 - \frac{1}{k^{m-1}} + \varepsilon.$$

We assume without loss of generality that $\varepsilon < \dfrac{1}{k^{m-1}}$. Our limit implies that there is some $\lambda_0 \in \mathbb{R}$ such that for all normalized $\boldsymbol{r} \in \mathbb{N}^m$ with $||\boldsymbol{r}|| \geq \lambda_0$ we have

$$\liminf_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}\delta(i+r_1, i+r_2, \ldots, i+r_m) \geq 1 - \frac{1}{k^{m-1}} + \frac{\varepsilon}{2}. \tag{4.1}$$

We define $\rho(\boldsymbol{r}) = \max\{r_j\} - \min\{r_j\}$ to be the *range* of $\boldsymbol{r}$ and note that $\rho(\boldsymbol{r}) = r_m$ whenever $\boldsymbol{r}$ is normalized. Let $\boldsymbol{r}^* = (0, \ldots, 0, 1) \in \mathbb{R}^m$ and let p be an integer such that $p||\boldsymbol{r}^*|| \geq \lambda_0$. Then whenever $\boldsymbol{r}$ is normalized with $\rho(\boldsymbol{r}) \geq p$, we have $||\boldsymbol{r}|| \geq ||p\boldsymbol{r}^*|| = p||\boldsymbol{r}^*|| \geq \lambda_0$. Hence, for all normalized $\boldsymbol{r}$ with $\rho(\boldsymbol{r}) \geq p$, we can pick $n_{\boldsymbol{r}} \in \mathbb{N}$ by (4.1) such that for all $N \geq n_{\boldsymbol{r}}$, we have

$$\frac{1}{N}\sum_{i=0}^{N-1}\delta(i+r_1, i+r_2, \ldots, i+r_m) \geq 1 - \frac{1}{k^{m-1}} + \frac{\varepsilon}{3}. \tag{4.2}$$

To construct our counterexample, we ensure that we have selected p such that

$$p \geq m, \tag{4.3}$$

and then pick $q \in \mathbb{N}$ such that the following both hold:

$$(a) \qquad q > \frac{18m^2(m-1)}{\varepsilon}; \tag{4.4}$$

$$(b) \qquad q^{m-1} > \frac{9m(m-1)p^{m-1}}{\varepsilon}. \tag{4.5}$$

Since there are finitely many normalized $\boldsymbol{r} \in \mathbb{N}^m$ with $p \leq \rho(\boldsymbol{r}) \leq q$, we can then pick an $n \in \mathbb{N}$ such that the following both hold:

$$(a) \qquad n \geq n_{\boldsymbol{r}} \text{ for all normalized } \boldsymbol{r} \text{ with } p \leq \rho(\boldsymbol{r}) \leq q. \tag{4.6}$$

$$(b) \qquad n > \frac{18qm!}{\varepsilon}. \tag{4.7}$$

Now, for any set $U \subset \mathbb{N}$ with $|U| = m$, there is a unique normalized vector $\boldsymbol{r}^U$ and integer offset $\mu(U)$ such that the vector $\boldsymbol{r}^U + \mu(U)\mathbf{1}$ is an ordering of the elements of U. We write $\delta(U)$ to denote the correlation coefficient associated to this vector, namely $\delta(U) = \delta(r_1^U + \mu(U), r_2^U + \mu(U), \ldots, r_m^U + \mu(U))$. We also write $\rho(U) = \max(U) - \min(U)$ for the range of U. It follows that $\rho(U) = \rho(\boldsymbol{r}^U) = r_m^U$, and $\mu(U) = \min(U)$. With these definitions in hand, we consider the following sum, which will be counted in two different ways to achieve our contradiction:

$$S = \sum_{a=0}^{n-1} \left(\sum_{\substack{U \subseteq \{a,\ldots,a+q-1\} \\ |U|=m}} \delta(U) \right).$$

We first use Lemma 4.1 to bound S from above. The sum

$$\sum_{\substack{U \subseteq \{a,\ldots,a+q-1\} \\ |U|=m}} \delta(U)$$

counts the number of subsets of m elements from the multiset $[x_a, x_{a+1}, \ldots, x_{a+q-1}]$ that contain at least one pair of distinct symbols of the k possible symbols. Thus Lemma 4.1 applies, yielding

$$S \leq \sum_{a=0}^{n-1} \frac{q^m}{m!} \left(1 - \frac{1}{k^{m-1}}\right) = \frac{nq^m}{m!} \left(1 - \frac{1}{k^{m-1}}\right). \tag{4.8}$$

Next, we will attempt to bound S from below by expressing it in terms of partial sums of the form seen in (4.2). Our first goal will be to rearrange this sum according to the multiplicity of $\delta(U)$ for each U. Sets U will be subsets of $\{a, \ldots, a+q-1\}$ for more values of a if they have lower range, so we sort the terms according to the value of $\rho(U)$, yielding

$$S = \sum_{b=m-1}^{q-1} \sum_{a=0}^{n-1} \left(\sum_{\substack{U \subseteq \{a,\ldots,a+q-1\} \\ |U|=m \\ \rho(U)=b}} \delta(U) \right).$$

For a given $U \subset \{0, \ldots, n+q-2\}$ with $|U| = m$, we have $U \subseteq \{a, \ldots, a+q-1\}$ if and only if $\min(U) \geq a$ and $\max(U) \leq a+q-1$. Thus $U \subseteq \{a, \ldots, a+q-1\}$ for precisely those a with $\mu(U) + \rho(U) - (q-1) \leq a \leq \mu(U)$. However, when we rearrange our sum, we must count only those a which also lie in the range $\{0, \ldots, n-1\}$. We rewrite our sum as

$$S = \sum_{b=m-1}^{q-1} \left(\sum_{\substack{U \subseteq \{0,\ldots,n+q-2\} \\ |U|=m \\ \rho(U)=b}} \left(\sum_{a=\max\{\mu(U)+\rho(U)-(q-1),0\}}^{\min\{\mu(U),n-1\}} \delta(U) \right) \right).$$

We drop all terms containing elements less than q or greater than $n-1$. All the sets U which remain will have $\mu(U)+\rho(U)-(q-1)\geq 0$ and $\mu(U)\leq n-1$, such that

$$\begin{aligned} S &\geq \sum_{b=m-1}^{q-1}\left(\sum_{\substack{U\subseteq\{q,\ldots,n-1\}\\ |U|=m\\ \rho(U)=b}}\left(\sum_{a=\mu(U)+\rho(U)-(q-1)}^{\mu(U)}\delta(U)\right)\right)\\ &= \sum_{b=m-1}^{q-1}\left(\sum_{\substack{U\subseteq\{q,\ldots,n-1\}\\ |U|=m\\ \rho(U)=b}}((q-\rho(U))\delta(U))\right)\\ &= \sum_{b=m-1}^{q-1}\left((q-b)\sum_{\substack{U\subseteq\{q,\ldots,n-1\}\\ |U|=m\\ \rho(U)=b}}\delta(U)\right). \end{aligned}$$

We now need to add back some of the terms we dropped and subtract away appropriate compensation. We can choose $U\subseteq\{0,\ldots,n-1\}$ with $|U|=m$, $\rho(U)=b$ and $U\nsubseteq\{q,\ldots,n-1\}$ by picking $\min(U)\in\{0,\ldots,q-1\}$, taking $\max(U)=\min(U)+b$, and then choosing the remaining $m-2$ elements from $\{\min(U)+1,\ldots,\min(U)+b-1\}$. There are $q\binom{b-1}{m-2}$ ways of doing this. It is convenient to instead use qb^{m-2} as an upper bound for this quantity; we then use the fact that $\delta(U)\in\{0,1\}$ to write

$$\begin{aligned} S &\geq \sum_{b=m-1}^{q-1}\left((q-b)\left(\left(\sum_{\substack{U\subseteq\{0,\ldots,n-1\}\\ |U|=m\\ \rho(U)=b}}\delta(U)\right)-qb^{m-2}\right)\right)\\ &> \sum_{b=m-1}^{q-1}\left((q-b)\sum_{\substack{U\subseteq\{0,\ldots,n-1\}\\ |U|=m\\ \rho(U)=b}}\delta(U)\right)-q^{m+1}. \end{aligned}$$

In a similar manner, we add back more terms so that we may consider all $U\subseteq\{0,\ldots,n+$

$q-1\}$ with $|U| = m$ and $\rho(U) = b$, and subtract off another multiple of q^{m+1} to compensate,

$$S > \sum_{b=m-1}^{q-1} \left((q-b) \sum_{\substack{U \subseteq \{0,\ldots,n+q-1\} \\ |U|=m \\ \rho(U)=b}} \delta(U) \right) - 2q^{m+1}.$$

We now associate each set U to its sorted vector $\boldsymbol{r}^U + \mu(U)\mathbf{1}$ and group them according to their $\boldsymbol{r}^U$ values. Since we count each subset of $\{0, \ldots, n+q-1\}$ having range $\leq q-1$, we are certain to include $\boldsymbol{r} + i\mathbf{1}$ for every normalized $\boldsymbol{r}$ of range $\leq q-1$ and every offset i from 0 to n. We drop any other terms and ignore those $\boldsymbol{r}$ with $\rho(\boldsymbol{r}) < p$ (recalling (4.3), where we ensured that $p \geq m$), leaving us with

$$S > \sum_{b=m-1}^{q-1} \left((q-b) \sum_{\substack{\boldsymbol{r} \in \mathbb{N}^m \\ \boldsymbol{r} \text{ normalized} \\ \rho(\boldsymbol{r})=b}} \sum_{i=0}^{n} \delta(\boldsymbol{r} + i\mathbf{1}) \right) - 2q^{m+1}.$$

Finally, we may use (4.2) to bound the inner sums from below, since for all $\boldsymbol{r}$ with $\rho(\boldsymbol{r}) \geq p$ we have $n \geq n_{\boldsymbol{r}}$ by (4.6). We then simply count the number of normalized $\boldsymbol{r}$ vectors of each range, obtaining

$$\begin{aligned} S &> \sum_{b=p}^{q-1} \left((q-b) \sum_{\substack{\boldsymbol{r} \in \mathbb{N}^m \\ \boldsymbol{r} \text{ normalized} \\ \rho(\boldsymbol{r})=b}} n\left(1 - \frac{1}{k^{m-1}} + \frac{\varepsilon}{3}\right) \right) - 2q^{m+1} \\ &= n\left(1 - \frac{1}{k^{m-1}} + \frac{\varepsilon}{3}\right) \sum_{b=p}^{q-1} \left((q-b)\binom{b-1}{m-2} \right) - 2q^{m+1} \\ &\geq \frac{n}{(m-2)!}\left(1 - \frac{1}{k^{m-1}} + \frac{\varepsilon}{3}\right) \sum_{b=p}^{q-1} \left((q-b)(b-m)^{m-2} \right) - 2q^{m+1}. \end{aligned} \tag{4.9}$$

We simplify and evaluate the remaining sum to get

$$\begin{aligned}
\sum_{b=p}^{q-1}\big((q-b)(b-m)^{m-2}\big) &\geq \sum_{b=p}^{q-1}\big((q+m-b)(b-m)^{m-2}\big) - mq^{m-1} \\
&\geq \sum_{b=p}^{q+m}\big((q+m-b)(b-m)^{m-2}\big) - 2mq^{m-1} \\
&= \sum_{b=p-m}^{q}\big((q-b)b^{m-2}\big) - 2mq^{m-1} \\
&\geq \sum_{b=0}^{q}\big((q-b)b^{m-2}\big) - 2mq^{m-1} - qp^{m-1} \\
&= q\sum_{b=0}^{q} b^{m-2} - \sum_{b=0}^{q-1} b^{m-1} - (2m+1)q^{m-1} - qp^{m-1} \\
&\geq q\int_0^q b^{m-2}\,\mathrm{d}b - \int_0^q b^{m-1}\,\mathrm{d}b - 2mq^{m-1} - qp^{m-1} \\
&= \frac{q^m}{m(m-1)} - 2mq^{m-1} - qp^{m-1}.
\end{aligned}$$

We substitute this back into (4.9) to obtain

$$S > \frac{nq^m}{m!}\left(1 - \frac{1}{k^{m-1}} + \frac{\varepsilon}{3}\right) - 2q^{m+1} - \frac{2mnq^{m-1}}{(m-2)!} - \frac{nqp^{m-1}}{(m-2)!}. \tag{4.10}$$

What remains is to eliminate the three leftover terms on the right hand side with the bounds we used when selecting q and n. First, by

$$\left(\frac{nq^m}{m!}\right)\left(\frac{\varepsilon}{9}\right) > \frac{2mnq^{m-1}}{(m-2)!}. \tag{4.11}$$

Second, by (4.5), we also picked q such that

$$\left(\frac{nq^m}{m!}\right)\left(\frac{\varepsilon}{9}\right) > \frac{nqp^{m-1}}{(m-2)!}. \tag{4.12}$$

Third, by (4.7), we picked n such that

$$\left(\frac{nq^m}{m!}\right)\left(\frac{\varepsilon}{9}\right) > 2q^{m+1}. \tag{4.13}$$

Adding (4.11), (4.12), and (4.13) together, we get

$$\left(\frac{nq^m}{m!}\right)\left(\frac{\varepsilon}{3}\right) > 2q^{m+1} + \frac{2mnq^{m-1}}{(m-2)!} + \frac{nqp^{m-1}}{(m-2)!}$$

and we substitute this into (4.10) to obtain

$$S > \frac{nq^m}{m!}\left(1 - \frac{1}{k^{m-1}}\right)$$

which contradicts (4.8), proving the desired result. □

5 Proof of Theorem 2.4

Suppose, for our sequence, that there exists some $m \geq 2$, $\boldsymbol{r} \in \mathbb{N}^m$, and $d > 0$ such that $D^d_{\boldsymbol{r}} > C_{\boldsymbol{r}}$. Let $\varepsilon = D^d_{\boldsymbol{r}} - C_{\boldsymbol{r}}$ and pick $p \in \mathbb{N}$ such that

$$p > \frac{2dD^d_{\boldsymbol{r}}}{\varepsilon}.$$

Then by our definition of $C_{\boldsymbol{r}}$, there is some $n \geq p$ such that

$$\frac{1}{n}\sum_{i=0}^{n-1} \delta(i + r_1, \ldots, i + r_m) < C_{\boldsymbol{r}} + \frac{\varepsilon}{2}.$$

Dividing n by d, we let $n = ad + b$, where a and b are non-negative integers and $b < d$. Then rearranging our expression and applying the definition of $D^d_{\boldsymbol{r}}$ yields:

$$\begin{aligned}
C_{\boldsymbol{r}} &> \frac{1}{n}\sum_{i=0}^{n-1} \delta(i + r_1, \ldots, i + r_m) - \frac{\varepsilon}{2} \\
&= \frac{1}{n}\left(\sum_{i=0}^{a-1}\sum_{j=id}^{id+d-1} \delta(j + r_1, \ldots, j + r_m) + \sum_{i=ad}^{ad+b-1} \delta(j + r_1, \ldots, j + r_m)\right) - \frac{\varepsilon}{2} \\
&\geq \frac{1}{n}\left(\sum_{i=0}^{a-1}\left(dD^d_{\boldsymbol{r}}\right) + \sum_{i=ar}^{ar+b-1} \delta(j + r_1, \ldots, j + r_m)\right) - \frac{\varepsilon}{2} \\
&\geq \frac{adD^d_{\boldsymbol{r}}}{n} - \frac{\varepsilon}{2} \\
&\geq D^d_{\boldsymbol{r}} - \frac{dD^d_{\boldsymbol{r}}}{n} - \frac{\varepsilon}{2}.
\end{aligned}$$

However, since

$$n \geq p > \frac{2dD^d_{\boldsymbol{r}}}{\varepsilon},$$

we then have

$$\frac{dD^d_{\boldsymbol{r}}}{n} < \frac{\varepsilon}{2},$$

and substituting this into the above yields

$$C_{\boldsymbol{r}} > D^d_{\boldsymbol{r}} - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} = D^d_{\boldsymbol{r}} - \varepsilon = C_{\boldsymbol{r}}.$$

Thus we have a contradiction, and so we have $D^d_{\boldsymbol{r}} \leq C_{\boldsymbol{r}}$ for all $\boldsymbol{r}$ and d. □

6 Proof of Theorem 2.6

It is sufficient to show that for all integers $k, m \geq 2$ and all real numbers $\varepsilon > 0$, there exist an integer d_0 and an infinite word $x = x_0 x_1 x_2 \cdots$ over a k-letter alphabet such that for every integer $d > d_0$ and $i \geq 0$ there are at least $(1 - \frac{1}{k^{m-1}} - \varepsilon)$ positions where the m words

$$x_i \cdots x_{i+d-1},\ x_{i+d} \cdots x_{i+2d-1},\ \ldots,\ x_{i+(m-1)d} \cdots x_{i+md-1}$$

do not all agree. We use the Lovász local lemma to show the existence of finite words of every sufficiently long length satisfying the condition. The existence of an infinite word then follows from the usual compactness argument.

Here is the statement of the Lovász local lemma, as taken from [2, Chap. 5].

Lemma 6.1. *Let $A_1, A_2, \ldots, A_T$ be events in a probability space, with a dependency digraph $D = (S, E)$. Suppose there exist real numbers $u_1, u_2, \ldots, u_T$ with $0 \leq u_i < 1$ for $1 \leq i \leq T$ such that*

$$\Pr(A_i) \leq u_i \prod_{(i,j) \in E} (1 - u_j) \tag{6.1}$$

for $1 \leq i \leq T$. Then the probability that none of the events $A_1, A_2, \ldots, A_T$ occur is $\geq \prod_{1 \leq i \leq T}(1 - u_i)$.

Let $A_{i,d}$ denote the event that there are $< t$ positions where the m words

$$x_i \cdots x_{i+d-1},\ x_{i+d} \cdots x_{i+2d-1},\ \ldots,\ x_{i+(m-1)d} \cdots x_{i+md-1}$$

do not all agree. Moreover, let S be the space of all such events $A_{i,d}$ and (S, E) the dependency digraph specifying when one event is dependent on another, which corresponds to overlapping ranges of the word being constructed.

To evaluate $\Pr[A_{i,d}]$ it suffices to count the number of such strings. First, we choose the values for the symbols of the first string, $x_i, \ldots, x_{i+d-1}$, which can be done in k^d ways. Next, we choose the precise number of positions j in which the m strings will fail to agree, and the positions themselves. This can be done in $\sum_{0 \leq j < t} \binom{d}{j}$ ways. For each such position, there are $k^{m-1} - 1$ ways to choose the symbols of the remaining $m - 1$ strings in such a way that they do not universally agree with the first string. The remaining symbols in the last $m - 1$ strings are now completely determined, as they must agree with the symbols in the corresponding position in the first string. The total number of such strings is therefore

$$P = k^d \sum_{0 \leq j < t} \binom{d}{j} (k^{m-1} - 1)^j.$$

We therefore find

$$\Pr[A_{i,d}] = \frac{P}{k^{md}} = \sum_{0 \leq j < t} \binom{d}{j} \left(\frac{k^{m-1} - 1}{k^{m-1}}\right)^j \left(\frac{1}{k^{m-1}}\right)^{d-j}.$$

To estimate this sum we use the following classical estimate on the tail of the binomial distribution, which is a version of Hoeffding's inequality [6]:

Lemma 6.2. *Suppose $0 < p < 1$, and let t, d be positive integers with $t \leq dp$. Then*

$$\sum_{0 \leq j \leq t} \binom{d}{j} p^j (1-p)^{d-j} \leq e^{-2(dp-t)^2/d}.$$

If we now take $t = (1 - \frac{1}{k^{m-1}} - \varepsilon)d$, $p = \frac{k^{m-1}-1}{k^{m-1}}$, we obtain

$$\Pr[A_{i,d}] \leq e^{-2d\varepsilon^2}.$$

Now fix n, the length of the string. We want none of the events $A_{j,s}$ for $d_0 \leq s \leq n/m$, $0 \leq j \leq n - ms$, to take place. Choose $u_{j,s} = e^{-\frac{1}{2}s\varepsilon^2}$. Then

$$\begin{aligned} \prod_{((i,d),(j,s)) \in E} (1 - u_{j,s}) &= \prod_{\substack{i-ms+1 \leq j \leq i+md-1 \\ 0 \leq j \leq n-ms \\ d_0 \leq s \leq n/m}} (1 - u_{j,s}) \\ &\geq \prod_{s \geq d_0} (1 - u_{j,s})^{md+ms-1}. \end{aligned}$$

Taking logarithms, we get

$$\sum_{((i,d),(j,s)) \in E} \log(1 - u_{j,s}) \geq \sum_{s \geq d_0} (md + ms - 1) \log(1 - u_{j,s}).$$

Provided $u_{j,s}$ is sufficiently small, we can bound $\log(1-u_{j,s})$ with $-cu_{j,s}$ for some constant c. Hence we get

$$\begin{aligned} &\sum_{s \geq d_0} (md + ms - 1) \log(1 - u_{j,s}) \\ &\qquad \geq \sum_{s \geq d_0} -(md + ms - 1) c e^{-\frac{1}{2}\varepsilon^2 s} \\ &\qquad = -(md-1)c \sum_{s \geq d_0} e^{-\frac{1}{2}\varepsilon^2 s} - mc \sum_{s \geq d_0} s e^{-\frac{1}{2}\varepsilon^2 s} \\ &\qquad = -(md-1)c \frac{e^{-\frac{1}{2}\varepsilon^2(d_0-1)}}{e^{\frac{1}{2}\varepsilon^2} - 1} - mc \frac{e^{-\frac{1}{2}\varepsilon^2(d_0-1)}(1-d_0) + d_0 e^{-\frac{1}{2}\varepsilon^2(d_0-2)}}{(e^{\frac{1}{2}\varepsilon^2} - 1)^2}. \end{aligned}$$

Now choose d_0 large enough so that

$$\frac{e^{-\frac{1}{2}\varepsilon^2(d_0-1)}}{e^{\frac{1}{2}\varepsilon^2} - 1} \leq \frac{\varepsilon^2}{2mc},$$

and also large enough so that

$$\frac{e^{-\frac{1}{2}\varepsilon^2(d_0-1)}(1-d_0) + d_0 e^{-\frac{1}{2}\varepsilon^2(d_0-2)}}{(e^{\frac{1}{2}\varepsilon^2} - 1)^2} \leq \frac{\varepsilon^2 d_0}{2mc}.$$

It follows that

$$\begin{aligned}\log\left(u_{i,d}\prod_{((i,d),(j,s))\in E}(1-u_{j,s})\right) &\geq -\frac{1}{2}\varepsilon^2 d-(md-1)c\frac{\varepsilon^2}{2mc}-mc\frac{\varepsilon^2 d_0}{2mc}\\ &\geq -\frac{1}{2}\varepsilon^2 d-\frac{1}{2}\varepsilon^2 d-\frac{1}{2}\varepsilon^2 d_0\\ &\geq -\frac{3}{2}\varepsilon^2 d\\ &\geq -2\varepsilon^2 d\\ &\geq \log\Pr[A_{i,d}],\end{aligned}$$

as desired. Hence, by the Lovász local lemma, it follows that the probability that none of the events $A_{j,s}$ occur is $\geq \prod_{((i,d),(j,s))\in E}(1-u_{j,s})>0$, and hence such a string of length n exists. □

7 Proof of Theorem 3.1

Before turning to the proof of Theorem 3.1, we need one auxiliary tool. We rewrite the left-hand-side expression of (3.3) in terms of exponential sums. As usual, set $\mathrm{e}(z)=e^{2\pi \mathrm{i}z}$ for $z\in\mathbb{R}$.

Proposition 7.1. *For any infinite word $x_0x_1x_2\cdots$ over $\{0,1,\dots,k-1\}$ we have*

$$\sum_{n<N}\delta(n+r_1,n+r_2)=N\left(1-\frac{1}{k}\right)-\frac{1}{k}\sum_{1\leq h<k}\sum_{n<N}e\left(\frac{h}{k}(x_{n+r_2}-x_{n+r_1})\right).$$

Proof. The proof is based on the relation

$$\sum_{0\leq h<k}\mathrm{e}\left(\frac{hu}{k}\right)=\begin{cases}0, & \text{if } k\nmid u;\\ k, & \text{if } k\mid u.\end{cases}\tag{7.1}$$

First, since $x_n\in\{0,1,2,\dots,k-1\}$ we notice that $k\mid(x_{n+r_2}-x_{n+r_1})$ if and only if $x_{n+r_2}=x_{n+r_1}$. Therefore,

$$\begin{aligned}\sum_{n<N}\delta(n+r_1,n+r_2)&=N-\sum_{n<N}\frac{1}{k}\sum_{0\leq h<k}\mathrm{e}\left(\frac{h}{k}\left(x_{n+r_2}-x_{n+r_1}\right)\right)\\ &=N\left(1-\frac{1}{k}\right)-\sum_{n<N}\frac{1}{k}\sum_{1\leq h<k}\mathrm{e}\left(\frac{h}{k}\left(x_{n+r_2}-x_{n+r_1}\right)\right).\end{aligned}$$

□

In view of Theorem 3.1 and Proposition 2.1 it suffices to show that for all $1 \le h \le k-1$ we have

$$\sum_{n<N} \mathrm{e}\left(\frac{h}{k}(\hat{a}(n+r)-\hat{a}(n))\right) = O_k\left(r\log\left(\frac{N}{r}\right)+r\right), \tag{7.2}$$

where the implied constant only depends on k. Since $\mathrm{e}(z+1)=\mathrm{e}(z)$, the left-hand-side sum in (7.2) can be rewritten in the form

$$\gamma_N(r) = \sum_{n<N} \mathrm{e}\left(\frac{h}{k}(a(n+r)-a(n))\right). \tag{7.3}$$

In the sequel we will need the generalized quantities

$$\gamma_N(r,f) = \sum_{n<N} \mathrm{e}\left(\frac{h}{k}(a(n+r)-a(n))\right)\mathrm{e}\left(\frac{hf(n)}{k}\right), \tag{7.4}$$

where $f:\ \mathbb{N}\to\mathbb{Z}$ is an arbitrary periodic function with period k. We first show that for all such f we have $\gamma_N(1,f)=O_k(\log N)$ for $N>k$. We will then use induction on r to prove (7.2), which in turn proves Theorem 3.1.

We follow the reasoning of Mauduit [11]. Regarding (7.4) we split the summation over $n<N$ up according to the residue class of n modulo k. We obtain

$$\begin{aligned}\gamma_{kN+j}(1,f) &= \sum_{n<kN+j} \mathrm{e}\left(\frac{h}{k}(a(n+1)-a(n))\right)\mathrm{e}\left(\frac{hf(n)}{k}\right)\\ &= \sum_{i=0}^{k-1}\sum_{kn+i<kN+j} \mathrm{e}\left(\frac{h}{k}(a(kn+i+1)-a(kn+i))\right)\mathrm{e}\left(\frac{hf(i)}{k}\right).\end{aligned}$$

Thus,

$$\gamma_{kN+j}(1,f) = \sum_{n=0}^{k-1} \mathrm{e}\left(\frac{h}{k}(a(n+1)-a(n))\right)\mathrm{e}\left(\frac{hf(n)}{k}\right) \tag{7.5}$$

$$+\sum_{u=0}^{j-1} \mathrm{e}\left(\frac{h}{k}(a(kN+u+1)-a(kN+u))\right)\mathrm{e}\left(\frac{hf(u)}{k}\right) \tag{7.6}$$

$$+\sum_{u=0}^{k-2} \mathrm{e}\left(\frac{hf(u)}{k}\right)\sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}(a(kn+u+1)-a(kn+u))\right) \tag{7.7}$$

$$+\mathrm{e}\left(\frac{hf(k-1)}{k}\right)\sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}(a(kn+k)-a(kn+k-1))\right). \tag{7.8}$$

The sums (7.5) and (7.6) are trivially bounded by $k+j\le 2k-1$. Concerning (7.7) we note

that for $0 \le u \le k-2$ we have

$$\begin{aligned}
&\sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}\left(a(kn+u+1)-a(kn+u)\right)\right)\\
&\qquad = \sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n)+g(u+1,n)-a(n)-g(u,n)\right)\right)\\
&\qquad = \sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}(g(u+1,n)-g(u,n))\right).
\end{aligned}$$

By our assumption $g(u+1,n)-g(u,n)$ runs through a complete residue system mod k for $1 \le n \le k$, so this sum is bounded in modulus by $k/2$. Therefore, (7.7) is bounded by $k(k-1)/2$. Finally, we rewrite the sum in (7.8) in the form

$$\begin{aligned}
&\sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}\left(a(kn+k)-a(kn+k-1)\right)\right)\\
&\qquad = \sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n+1)+g(0,n+1)-a(n)-g(k-1,n)\right)\right)\\
&\qquad = \sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n+1)-a(n)\right)\right)\mathrm{e}\left(\frac{h\hat{f}(n)}{k}\right),
\end{aligned}$$

where $\hat{f}(n) = g(0,n+1)-g(k-1,n)$ is again periodic with period k in n. Summing up, we get

$$|\gamma_{kN+j}(1,f)| \le |\gamma_N(1,\hat{f})| + \frac{k}{2}(k+3). \tag{7.9}$$

From (7.9) and $|\gamma_n(1,f)| \le k-1$ for $1 \le n \le k-1$ and all f we get by induction that for all k-periodic functions f and all $N > k$,

$$|\gamma_N(1,f)| \le \frac{k(k+3)}{2\log k}\ \log N + k - 1. \tag{7.10}$$

For our induction on r to work, we need one more initial value, namely

$$\gamma_N(0,f) = \sum_{n<N} \mathrm{e}\left(\frac{hf(n)}{k}\right)$$

which satisfies

$$|\gamma_N(0,f)| \le \frac{k}{2}, \quad \text{if } f(\{0,1,\ldots,k-1\}) = \{0,1,\ldots,k-1\}. \tag{7.11}$$

Now, let us consider the general case with $r = kM+i > 0$ where $M \geq 0$ and $0 \leq i \leq k-1$ but $(M,i) \neq (0,0)$. Similarly to (7.5)–(7.8) we have

$$\begin{aligned}
&\gamma_{kN+j}(kM+i,f) = \\
&\quad \sum_{u=0}^{k-2} \mathrm{e}\left(\frac{hf(u)}{k}\right) \sum_{1\leq n<N} \mathrm{e}\left(\frac{h}{k}(a(kn+u+kM+i)-a(kn+u))\right) \qquad (7.12)\\
&\quad + \mathrm{e}\left(\frac{hf(k-1)}{k}\right) \sum_{1\leq n<N} \mathrm{e}\left(\frac{h}{k}(a(kn+k-1+kM+i)-a(kn+k-1))\right) \qquad (7.13)\\
&\quad + O(1),
\end{aligned}$$

where the implied constant is bounded in modulus by $2k-1$. We again need a close inspection of the two infinite sums (7.12) and (7.13). First, suppose $i \neq 0$. We rewrite the sum (7.12) in the form

$$\begin{aligned}
&\sum_{u=0}^{k-1-i} \mathrm{e}\left(\frac{hf(u)}{k}\right) \sum_{1\leq n<N} \mathrm{e}\Big(\frac{h}{k}(a(n+M)+g(u+i,n+M) \\
&\qquad\qquad - a(n)-g(u,n))\Big) \\
&+ \sum_{u=k-i}^{k-2} \mathrm{e}\left(\frac{hf(u)}{k}\right) \sum_{1\leq n<N} \mathrm{e}\Big(\frac{h}{k}(a(n+M+1)+g(u+i-k,n+M+1) \\
&\qquad\qquad - a(n)-g(u,n))\Big) \\
= &\sum_{u=0}^{k-1-i} \mathrm{e}\left(\frac{hf(u)}{k}\right) \sum_{1\leq n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n+M)-a(n)\right)\right) \mathrm{e}\left(\frac{hf_1(n)}{k}\right) \\
&+ \sum_{u=k-i}^{k-2} \mathrm{e}\left(\frac{hf(u)}{k}\right) \sum_{1\leq n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n+M+1)-a(n)\right)\right) \mathrm{e}\left(\frac{hf_2(n)}{k}\right),
\end{aligned}$$

where

$$\begin{aligned}
f_1(n) &= g(u+i,n+M)-g(u,n), && \text{for } 0 \leq u \leq k-1-i,\\
f_2(n) &= g(u+i-k,n+M+1)-g(u,n), && \text{for } k-i \leq u \leq k-2.
\end{aligned}$$

Using (7.4) this yields

$$\begin{aligned}
&\sum_{u=0}^{k-2} \mathrm{e}\left(\frac{hf(u)}{k}\right) \sum_{1\leq n<N} \mathrm{e}\left(\frac{h}{k}(a(kn+u+kM+i)-a(kn+u))\right) \qquad (7.14)\\
&= \sum_{u=0}^{k-1-i} \mathrm{e}\left(\frac{hf(u)}{k}\right) \gamma_N(M,f_1) + \sum_{u=k-i}^{k-2} \mathrm{e}\left(\frac{hf(u)}{k}\right) \gamma_N(M+1,f_2) + O(1),
\end{aligned}$$

where the $O(1)$-term comes from including $n = 0$ into (7.14) and therefore is trivially bounded in modulus by $(k-i)+(i-1) = k-1$. Consider the second sum (7.13) and let $i \neq 0$. Then

$$\begin{aligned} &a(k(n+M+1)+i-1) - a(kn+k-1) \\ &\qquad = a(n+M+1) - a(n) + g(i-1, n+M+1) - g(k-1, n). \end{aligned}$$

Therefore,

$$\begin{aligned} \Big|\, \mathrm{e}\left(\frac{hf(k-1)}{k}\right) \sum_{1\le n<N} \mathrm{e}\left(\frac{h}{k}(a(kn+k-1+kM+i) - a(kn+k-1))\right) \,\Big| \\ \le \quad |\gamma_N(M+1, f_3)| + 1, \end{aligned} \tag{7.15}$$

where $f_3(n) = g(i-1, n+M+1) - g(k-1, n)$. Now, from (7.12), (7.13), (7.14) and (7.15) we see that

$$\begin{aligned} |\gamma_{kN+j}(kM+i, f)| \le |\gamma_N(M, f_1)| \cdot (k-i) + |\gamma_N(M+1, f_2)| \cdot (i-1) \\ + |\gamma_N(M+1, f_3)| + 1 + (2k-1) + (k-1). \end{aligned} \tag{7.16}$$

Plugging in $M = 0$, using (7.10) and (7.11) and observing that $f_1(n) = g(u+i, n) - g(u, n)$ permutes $\{0, 1, \dots, k-1\}$ by assumption, we get

$$|\gamma_{kN+j}(i, f)| \le \frac{k(k-1)(k+3)}{2\log k} \log N + \frac{k}{2}(2k+3), \qquad 1 \le i \le k-1.$$

This implies that for $1 \le i \le k-1$ and all functions f with period k we have

$$|\gamma_N(i, f)| \le \frac{k(k-1)(k+3)}{2\log k} \log\left(\frac{N}{k}\right) + \frac{k}{2}(2k+3), \qquad N > k. \tag{7.17}$$

On the other hand, if $0 \le u \le k-1$ then

$$a(k(n+M)+u) - a(kn+u) = a(n+M) - a(n) + g(u, n+M) - g(u, n),$$

so by joining (7.12) and (7.13) in case that $i = 0$ we directly get

$$|\gamma_{kN+j}(kM, f)| \le \sum_{u=0}^{k-1} (|\gamma_N(M, f_4)| + 1) + (2k-1), \tag{7.18}$$

where $f_4(n) = g(u, n+M) - g(u, n)$. Therefore, by (7.10) and (7.18) applied for $M = 1$ we get

$$|\gamma_N(k, f)| \le \frac{k^2(k+3)}{2\log k} \log\left(\frac{N}{k}\right) + k^2 + 2k - 1, \tag{7.19}$$

provided $N > k$. Therefore, for all $N > k$,

$$|\gamma_N(i, f)| \le \frac{k^2(k+3)}{2\log k} \log\left(\frac{N}{k}\right) + k^2 + 2k - 1, \tag{7.20}$$

for the whole range $1 \le i \le k$. We now start our induction on the parameter $r = kM+i$. We iterate (7.16) and (7.18) with (7.20) as an initial value to obtain for $r = k^s+1, k^s+2, \dots, k^{s+1}$ with $s \ge 0$ and for all $N > k^{s+1}$,

$$\begin{aligned}|\gamma_N(r,f)| &\le \frac{k^2(k+3)}{2\log k}\, k^s\, \log\left(\frac{N}{k^{s+1}}\right) + k^s(k^2+2k-1) + \sum_{j=0}^{s-1}(3k-1)k^j \\ &\le \frac{k^2(k+3)}{2\log k}\, k^s\, \log\left(\frac{N}{k^{s+1}}\right) + \frac{k^s(k^3+k^2)}{k-1}.\end{aligned}$$

This finishes the proof of Theorem 3.1. □

8 Proof of Theorem 3.3

For the proof of Theorem 3.3 it suffices to show that for all $1 \le h \le k-1$ and $0 < \gamma < 1$ we have

$$\sum_{n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n+r)-a(n)\right)\right) \ll N^{\gamma} + rN^{1-\gamma/d} + rN^{1-\gamma}\log\left(\frac{N^{\gamma/d}}{r}\right), \tag{8.1}$$

where the implied constant only depends on k. We follow Kim [9, Section 4], however suitably modifying the argument to deal with the function a not being k-additive in the usual sense. We need some more notation. Let $b = (b_1, b_2, \dots, b_d)$ and set

$$P_b = \{n \in \mathbb{N} : \quad n \equiv b_i \bmod p_i^{s_i}, \quad 1 \le i \le d\},$$

where s_i is the unique integer with $p_i^{s_i} \le N^{\gamma/d} < p_i^{s_i+1}$. Since the p_i's denote different primes by assumption, we have

$$\#\{n \in \mathbb{N} : \quad n \in P_b\} = \frac{N}{\prod_{i=1}^d p_i^{s_i}} + O(1).$$

Further set

$$\begin{aligned}\mathcal{B} &= \{(b_1, b_2, \dots, b_d) : \qquad 0 \le b_i < p_i^{s_i} \quad \text{for } 1 \le i \le d\}, \\ \mathcal{B}_0 &= \{(b_1, b_2, \dots, b_d) : \qquad 0 \le b_i < p_i^{s_i} - r \quad \text{for } 1 \le i \le d\}.\end{aligned}$$

Now, consider $n = n_i p_i^{s_i} + b_i$ where $0 \le b_i < p_i^{s_i} - r$. We may assume that $n_i \ge 1$, which is true for most n, i.e. $N^{\gamma/d} \le n < N$ (the error term of $N^{\gamma/d}$ is negligible in the final estimate). Write

$$\begin{aligned}b_i + r &= \beta'_{s_i-1}p_i^{s_i-1} + \beta'_{s_i-2}p_i^{s_i-2} + \cdots + \beta'_0, \\ b_i &= \beta_{s_i-1}p_i^{s_i-1} + \beta_{s_i-2}p_i^{s_i-2} + \cdots + \beta_0\end{aligned}$$

where $\beta_\nu, \beta'_\nu \in \{0, 1, \dots, p_i - 1\}$ for $0 \le \nu < s_i$. Furthermore, set

$$\begin{aligned}v_i &= \max(j : \quad \beta'_j \ne 0, \quad 0 \le j \le s_i - 1), \\ w_i &= \max(j : \quad \beta_j \ne 0, \quad 0 \le j \le s_i - 1),\end{aligned}$$

which indicate the uppermost nonzero coefficients in the expansions. Then by (3.6) we can rewrite $a_i(n+r)-a_i(n)$ in the form

$$\begin{aligned}
&a_i\left(n_ip_i^{s_i}+\beta'_{s_i-1}p_i^{s_i-1}+\cdots+\beta'_0\right)-a_i\left(n_ip_i^{s_i}+\beta_{s_i-1}p_i^{s_i-1}+\cdots+\beta_0\right)\\
&\quad=a_i(n_i)+g_i(\beta'_{s_i-1},n_i)+\sum_{\nu=0}^{s_i-2}g_i(\beta'_\nu,\beta'_{\nu+1})\\
&\qquad-\left(a_i(n_i)+g_i(\beta_{s_i-1},n_i)+\sum_{\nu=0}^{s_i-2}g_i(\beta_\nu,\beta_{\nu+1})\right)\\
&\quad=g_i(\beta'_{s_i-1},n_i)-g_i(\beta_{s_i-1},n_i)+\sum_{\nu=0}^{s_i-2}\left(g_i(\beta'_\nu,\beta'_{\nu+1})-g_i(\beta_\nu,\beta_{\nu+1})\right)\\
&\quad=a_i(b_i+r)-a_i(b_i)+\mu_i(b_i,r,n_i),
\end{aligned}$$

where

$$\mu_i(b_i,r,n_i)=g_i(\beta'_{s_i-1},n_i)-g_i(\beta_{s_i-1},n_i)+\sum_{\nu=v_i}^{s_i-2}g_i(\beta'_\nu,\beta'_{\nu+1})-\sum_{\nu=w_i}^{s_i-2}g_i(\beta_\nu,\beta_{\nu+1}).$$

Consequently,

$$\begin{aligned}
\sum_{n<N}\mathrm{e}\left(\frac{h}{k}\left(a(n+r)-a(n)\right)\right)=&\quad\sum_{n<N}\prod_{i=1}^{d}\mathrm{e}\left(\frac{h}{k}\,c_i\left(a_i(n+r)-a_i(n)\right)\right)\\
=&\quad\sum_{b\in\mathcal{B}_0}\sum_{\substack{n<N\\ n\in P_b}}\prod_{i=1}^{d}\mathrm{e}\left(\frac{h}{k}\,c_i\left(a_i(b_i+r)-a_i(b_i)+\mu_i(b_i,r,n_i)\right)\right)\\
&+\sum_{b\in\mathcal{B}\setminus\mathcal{B}_0}\sum_{\substack{n<N\\ n\in P_b}}\mathrm{e}\left(\frac{h}{k}\,\left(a(n+r)-a(n)\right)\right),
\end{aligned}$$

which equals

$$\sum_{b\in\mathcal{B}}\prod_{i=1}^{d}\mathrm{e}\left(\frac{h}{k}\,c_i\left(a_i(b_i+r)-a_i(b_i)\right)\right)\sum_{\substack{n<N\\ n\in P_b}}\prod_{i=1}^{d}\mathrm{e}\left(\frac{h}{k}\,c_i\mu_i(b_i,r,n_i)\right) \tag{8.2}$$

$$+\sum_{b\in\mathcal{B}\setminus\mathcal{B}_0}\sum_{\substack{n<N\\ n\in P_b}}\left(\mathrm{e}\left(\frac{h}{k}\left(a(n+r)-a(n)\right)\right)-\prod_{i=1}^{d}\mathrm{e}\left(\frac{h}{k}\,c_i\left(a_i(b_i+r)-a_i(b_i)+\mu_i(b_i,r,n_i)\right)\right)\right). \tag{8.3}$$

The second sum (8.3) is trivially bounded by (we follow [9])

$$\begin{aligned} 2\,|\mathcal{B}\setminus\mathcal{B}_0|\cdot\#\{n<N:\ n\in P_b\} &\ll \left(\sum_{i=1}^{d}\frac{r}{p_i^{s_i}}\prod_{j=1}^{d}p_j^{s_j}\right)\left(\frac{N}{\prod_{i=1}^{d}p_i^{s_i}}+O(1)\right)\\ &\ll rN^{1-\gamma/d}, \end{aligned} \tag{8.4}$$

which is one of the error terms in the estimate. Now, consider the first sum (8.2). Let

$$\mathcal{B}^r=\{b\in\mathcal{B}:\quad v_i=w_i \text{ and } \beta_{v_i}=\beta'_{w_i} \text{ for all } 1\le i\le d\}.$$

Obviously, for every $b\in\mathcal{B}^r$ we have $\mu_i(b_i,r,n_i)=0$ for all $n<N$, $n\in P_b$. We use a similar splitting as above, such that (8.2) satisfies

$$\begin{aligned} \ll\quad & \sum_{b\in\mathcal{B}}\prod_{i=1}^{d}\mathrm{e}\left(\frac{h}{k}\,c_i\,(a_i(b_i+r)-a_i(b_i))\right)\sum_{\substack{n<N\\ n\in P_b}}1\\ &+2\,|\mathcal{B}\setminus\mathcal{B}^r|\left(\frac{N}{\prod_{i=1}^{d}p_i^{s_i}}+O(1)\right). \end{aligned}$$

Our next task is to establish a bound for $|\mathcal{B}\setminus\mathcal{B}^r|$. Let $p_i^{t_i}\le r<p_i^{t_i+1}$. We have to count the number of b_i's with $0\le b_i<p_i^{s_i}$ such that performing the addition b_i+r gives rise to a carry propagation which is transported to the digits β_{v_i} of b_i, thus giving a contribution to $\mu_i(b_i,r,n_i)$. A necessary condition for this effect is that

$$\beta_{t_i+1}=\beta_{t_i+2}=\cdots=\beta_{s_i-2}=p_i-1.$$

Hence

$$\begin{aligned} |\mathcal{B}\setminus\mathcal{B}^r| &\le \sum_{i=1}^{d}\left(p_i^{t_i+1}+(s_i-1-t_i)p_i^{t_i+2}\right)\\ &\ll \sum_{i=1}^{d}\left(r+p_i r\left(\frac{\log N^{\gamma/d}}{\log p_i}-\log r\right)\right)\\ &\ll r\log N^{\gamma/d}. \end{aligned}$$

Summing up, we obtain

$$\begin{aligned}
&\sum_{n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n+r)-a(n)\right)\right) = \\
&\qquad \sum_{b\in\mathcal{B}}\prod_{i=1}^{d} \mathrm{e}\left(\frac{h}{k}\,c_i\left(a_i(b_i+r)-a_i(b_i)\right)\right) \sum_{\substack{n<N\\ n\in P_b}} 1 + O\left(rN^{1-\gamma/d} + rN^{1-\gamma}\log N^{\gamma/d}\right) \\
&= \quad \prod_{i=1}^{d}\sum_{b_i=0}^{p_i^{s_i}-1} \mathrm{e}\left(\frac{h}{k}\,c_i\left(a_i(b_i+r)-a_i(b_i)\right)\right)\left(\frac{N}{\prod_{i=1}^{d} p_i^{s_i}} + O(1)\right) + O\left(rN^{1-\gamma/d}\right) \\
&= \quad N\prod_{i=1}^{d}\frac{1}{p_i^{s_i}}\sum_{b_i=0}^{p_i^{s_i}-1} \mathrm{e}\left(\frac{h}{k}\,c_i\left(a_i(b_i+r)-a_i(b_i)\right)\right) + O\left(N^{\gamma} + rN^{1-\gamma/d}\right).
\end{aligned}$$

Finally, we show how to obtain the saving in the exponent, which again finishes the proof of Theorem 3.3. Since $c_i = p_1p_2\cdots p_{i-1}$, we see that for every h there exists an index l with $1 \le l \le d$ and

$$\frac{h}{k}\,c_l = \frac{hp_1p_2\cdots p_{i-1}}{p_1p_2\cdots p_d} = \frac{h'}{p_l},$$

with $\gcd(h', p_l) = 1$. Applying Theorem 3.1 with $k = p_l$ and estimating the other factors trivially, we get

$$\sum_{n<N} \mathrm{e}\left(\frac{h}{k}\left(a(n+r)-a(n)\right)\right) \ll N^{1-\gamma} r\log\frac{N^{\gamma/d}}{r} + N^{1-\gamma}r + N^{\gamma} + rN^{1-\gamma/d},$$

which gives the statement of the theorem. □

References

[1] R. Ahlswede, J. Cassaigne, A. Sárközy, *On the correlation of binary sequences*, Discrete Appl. Math. 156 (2008), no. 9, 1478–1487.

[2] N. Alon, J. Spencer, The Probabilistic Method, 2nd edition, John Wiley & Sons, 2000.

[3] G. Bérczi, *On finite pseudorandom sequences of k symbols*, Period. Math. Hungar. 47 (2003), 29–44.

[4] J. Cassaigne, C. Mauduit, A. Sárközy, *On finite pseudorandom binary sequences. VII. The measures of pseudorandomness*, Acta Arith. 103 (2002), 97–118.

[5] L. Goubin, C. Mauduit, A. Sárközy, *Construction of large families of pseudorandom binary sequences*, J. Number Theory 106 (2004), no. 1, 56–69.

[6] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Stat. Assoc. 58 (1963), l13–30.

[7] T. Høholdt, H.E. Elbrønd, J. Justesen, *Autocorrelation properties of a class of infinite binary sequences*, IEEE Trans. Inform. Theory 32 (1986), 430–431.

[8] T. Høholdt, H.E. Elbrønd, J. Justesen, *Aperiodic correlations and the merit factor of a class of binary sequences*, IEEE Trans. Inform. Theory 31 (1985), 549–552.

[9] D. H. Kim, *On the joint distribution of q-additive functions in residue classes*, J. Number Theory 74 (1999), 307–336.

[10] K. Mahler, *The spectrum of an array and its application to the study of the translation properties of a simple class of arithmetical functions, Part II*, On the translation properties of a simple class of arithmetical functions, J. of Mathematics & Physics 6 (1927), 158–163.

[11] C. Mauduit, *Multiplicative properties of the Thue-Morse sequence*, Period. Math. Hungar. 43 (2001), 137–153.

[12] C. Mauduit, A. Sárközy, *On finite pseudorandom binary sequences. I. Measure of pseudorandomness, the Legendre symbol*, Acta Arith. 82 (1997), 365–377.

[13] C. Mauduit, A. Sárközy, *On finite pseudorandom binary sequences. II. The Champernowne, Rudin-Shapiro, and Thue-Morse sequences, a further construction*, J. Number Theory 73 (1998), 256–276.

[14] C. Mauduit, A. Sárközy, *On finite pseudorandom sequences of k symbols*, Indag. Math. (N.S.) 13 (2002), 89–101.

[15] M. Queffélec, *Une nouvelle propriété des suites de Rudin–Shapiro*, Ann. Inst. Fourier (Grenoble) 37 (1987), no. 2, 115–138.

E. Grant (egrant@uwaterloo.ca), J. Shallit (shallit@cs.uwaterloo.ca), T. Stoll (tstoll@cs.uwaterloo.ca): Faculty of Mathematics, School of Computer Science, University of Waterloo, Waterloo, ON, Canada.