On Long Words Avoiding Zimin Patterns

Arnaud Carayol[1] and Stefan Göller[*2]

1 Université Paris-Est, LIGM (UMR 8049), CNRS, ENPC, ESIEE, UPEM, Marne-la-Vallée, France
carayol@u-pem.fr

2 LSV, CNRS & ENS Cachan, Université Paris-Saclay, Paris, France
goeller@lsv.fr

Abstract

A pattern is encountered in a word if some infix of the word is the image of the pattern under some non-erasing morphism. A pattern p is unavoidable if, over every finite alphabet, every sufficiently long word encounters p. A theorem by Zimin and independently by Bean, Ehrenfeucht and McNulty states that a pattern over n distinct variables is unavoidable if, and only if, p itself is encountered in the n-th Zimin pattern. Given an alphabet size k, we study the minimal length $f(n, k)$ such that every word of length $f(n, k)$ encounters the n-th Zimin pattern. It is known that f is upper-bounded by a tower of exponentials. Our main result states that $f(n, k)$ is lower-bounded by a tower of $n - 3$ exponentials, even for $k = 2$. To the best of our knowledge, this improves upon a previously best-known doubly-exponential lower bound. As a further result, we prove a doubly-exponential upper bound for encountering Zimin patterns in the abelian sense.

1998 ACM Subject Classification G.2.1 Combinatorics, F.4.3. Formal Languages

Keywords and phrases Unavoidable patterns, combinatorics on words, lower bounds

Digital Object Identifier 10.4230/LIPIcs.STACS.2017.19

1 Introduction

A pattern is a finite word over some set of pattern variables. A pattern matches a word if the word can be obtained by substituting each variable appearing in the pattern by a non-empty word. The pattern xx matches the word *nana* when x is replaced by the word *na*. A word encounters a pattern if the pattern matches some infix of the word. For example, the word *banana* encounters the pattern xx (as the word *nana* is one of its infixes). The pattern xyx is encountered in precisely those words that contain two non-consecutive occurrences of the same letter, as e.g., the word *abca*.

A pattern is unavoidable if over every finite alphabet every sufficiently long word encounters the pattern. Equivalently, by Kőnig's Lemma, a pattern is unavoidable if over every finite alphabet all infinite words encounter the pattern. If it is not the case, the pattern is said to be avoidable. The pattern xyx is easily seen to be unavoidable since every sufficiently long word over a non-empty finite alphabet must contain two non-consecutive occurrences of the same letter. On the other hand, the pattern xx is avoidable as Thue [19] gave an infinite word over a ternary alphabet that does not encounter the pattern xx.

A precise characterization of unavoidable patterns was found by Zimin [20] and independently by Bean, Ehrenfeucht and McNulty [6], see also [13] for a more recent proof. This

[*] Supported by Labex Digicosme, Univ. Paris-Saclay, project VERICONISS.

34th Symposium on Theoretical Aspects of Computer Science (STACS 2017).
Editors: Heribert Vollmer and Brigitte Vallée; Article No. 19; pp. 19:1–19:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

characterization is based on a family $(Z_n)_{n\geq 0}$ of unavoidable patterns, called the Zimin patterns. The Zimin patterns over the pattern variables $\{x_1, x_2, \ldots\}$ are defined by $Z_0 = \varepsilon$ and $Z_{n+1} = Z_n x_{n+1} Z_n$ for all $n \geq 0$. A pattern over n distinct pattern variables is unavoidable if, and only if, the pattern itself is encountered in the n-th Zimin pattern Z_n. Zimin patterns can therefore be viewed as the canonical patterns for unavoidability.

Due to the canonical status of Zimin patterns, it is natural to investigate the smallest word length $f(n,k)$ that guarantees that every word over a k-letter alphabet of this length encounters the n-th pattern Z_n. Computing the exact value of $f(n,k)$ for $n \geq 1$ and $k \geq 2$, or at least giving upper and lower bounds on its value, has been the topic of several articles in recent years [2, 18, 12, 3]. For small values of n and k, known results from [12, 11] are summarized in the following table, taken from [12] and enriched with results from [11].

	2	3	4	5	k
1	1	1	1	1	1
2	5	7	9	11	$2k+1$
3	29	≤ 319	≤ 3169	≤ 37991	$\leq \sqrt{e}2^k(k+1)! + 2k + 1$
4	$\in [10483, 236489]$				
n					

In general, Cooper and Rorabaugh [2, Theorem 1.1] showed that the value of $f(n,k)$ is upper-bounded by a tower of exponentials of height $n-1$. To make this more precise let us define the tower function $\mathrm{Tower} : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ inductively as follows: $\mathrm{Tower}(0,k) = 1$ and $\mathrm{Tower}(n+1,k) = k^{\mathrm{Tower}(n,k)}$ for all $n, k \in \mathbb{N}$.

▶ **Theorem 1** (Cooper/Rorabaugh [2]). *For all $n \geq 1$ and $k \geq 2$, $f(n,k) \leq \mathrm{Tower}(n-1, K)$, where $K = 2k+1$.*

In stark contrast with this upper bound, Cooper and Rorabaugh showed that $f(n,k)$ is lower-bounded doubly-exponentially in n for every fixed $k \geq 2$. To our knowledge, this is the best known lower bound for f.

▶ **Theorem 2** (Cooper/Rorabaugh [2]). $f(n,k) \geq k^{2^{n-1}(1+o(1))}$.

This lower bound is obtained by estimating the expected number of occurrences of Z_n in long words over a k-letter alphabet using the first moment method.

In the conclusion, we address the limitation of this method to show non-elementary lower bounds for f.

Our contributions. Our main contribution is to prove a lower bound for $f(n,k)$ that is non-elementary in n even for $k = 2$. We use Stockmeyer's yardstick construction [17] to construct for each $n \geq 1$, a family of words of length at least $\mathrm{Tower}(n-1, 2)$ (that we call higher-order counters here). We then show that a counter of order n does not encounter Z_n (for $n \geq 3$). As these words are over an alphabet of size $2n-1$, this immediately establishes that $f(n, 2n-1) \geq \mathrm{Tower}(n-1, 2)$ (cf. Corollary 12).

Stockmeyer's yardstick construction is a well-known technique to prove non-elementary lower bounds in computer science, for instance it is used to show that the first-order theory of binary words with order, is non-elementary, see for instance [10] for a proof.

By using a carefully chosen encoding we are able to prove a lower bound for f over a binary alphabet. Namely for all $n \geq 4$, it holds $f(n,2) \geq \mathrm{Tower}(n-3, 2)$ (cf. Corollary 14).

As a spin-off result, we also consider the abelian setting. Matching a pattern in the abelian sense is a weaker condition, where one only requires that all infixes that are matching a pattern variable must have the same number of occurrences of each letter (instead of being the same words). This gives rise to the notions of avoidability and unavoidability of patterns in the abelian sense. Every pattern that is unavoidable is in particular unavoidable in the abelian sense. However, the converse does not hold in general as witnessed by the pattern $xyzxyxuxyxzyx$, as shown in [5]. Even though Zimin patterns lose their canonical status in the abelian setting, the function $g(n, k)$, which is an abelian analog of the function $f(n, k)$, has been studied [18]. For this function, Tao [18] establishes a lower bound that turns out to be doubly-exponential from the estimations in [9]. The upper bound is inherited from the non-abelian setting and is hence non-elementary. We improve this upper bound to be doubly-exponential (Theorem 22). We also provide a simple proof using the first moment method that g admits a doubly-exponential lower bound which does not require the elaborate estimations of [9].

Connection to the equivalence problem of deterministic pushdown automata. The equivalence problem for deterministic pushdown automata (dpda) is a famous problem in theoretical computer science. Its decidability has been established by Sénizergues in 1997 and Stirling proved in 2001 the first complexity-theoretic upper bound, namely a tower of exponentials of elementary height [16] (in $\mathbf{F}_3$ in terms of Schmitz' classification [14]), see also [8] for a more recent presentation.

In [15, p. 24, C1] Sénizergues outlines a link between the complexity of Stirling's algorithm and the function $f(n, k)$ and remarks that $f(n, k)$ seems to be non-elementary in n for all fixed $k \geq 2$. The present article substantiates this remark.

Organization of the paper. We introduce necessary notations in Section 2. We show that $f(n, 2n-1) \geq \text{Tower}(n-1, 2)$ in Section 3. We lift this result to unavoidability over a binary alphabet in Section 4, where we show that $f(n, 2) \geq \text{Tower}(n-3, 2)$ for all $n \geq 4$. Our doubly-exponential bounds on abelian avoidability are presented in Section 5. We conclude in Section 6.

2 Preliminaries

For every two integers i, j, we write $[i, j]$ for the set $\{i, i+1, \ldots, j\}$ and $[j]$ for $\{1, \ldots, k\}$. By $\mathbb{N}$ we denote the non-negative integers and by $\mathbb{N}^+$ the positive integers.

If A is a finite set of symbols, we denote by A^* the set of all words over A and by A^+ the set of all non-empty words over A. We write ε for the empty word. For a word $u \in A^*$, we denote by $|u|$ its length. For two words u and v, we denote by $u \cdot v$ (or simply uv) their concatenation. A word v is a *prefix* of a word u, denoted by $v \sqsubseteq u$, if there exists a word z such that $u = vz$. If z is non-empty, we say that v is a *strict prefix*[1] of u. A word v is a suffix of a word u if there exist a word z such that $u = zv$. If z is non-empty, we say that v is a *strict suffix* of u. A word v is an infix of a word u if there exists z_1 and z_2 such that $u = z_1 v z_2$. If both z_1 and z_2 are non-empty, v is a *strict infix* of u. If v is an infix u and u can be written as $z_1 u z_2$, the integer $|z_1|$ is called an *occurrence* of v in u. For $a \in A$, we denote by $|u|_a$ the number of occurrences of the symbol a in u.

Given two non-empty sets A and B, a *morphism* is a function $\phi : A^* \to B^*$ that satisfies $\phi(a_1 a_2) = \phi(a_1)\phi(a_2)$ for all $a_1, a_2 \in A$. Thus, every morphism can simply be written as a

[1] Our definition of strict prefix is slightly non-standard as ε is a strict prefix of any non-empty word.

function from A to B^*. A morphism ϕ is said to be *non-erasing* if $\phi(a) \neq \varepsilon$ for all $a \in A$ and ϕ is *alphabetic* if $\psi(a) \in B$ for all $a \in A$.

Let us fix a countable set $\mathcal{X} = \{x_1, x_2, \ldots\}$ of *pattern variables.* A *pattern* is a finite word over $\mathcal{X}$. Let $\rho = \rho_1 \cdots \rho_n$ be a pattern of length n. A finite or infinite word w *matches* ρ if $w = \psi(\rho)$ for some non-erasing morphism ψ. A finite or infinite word w *encounters* ρ if some infix of w matches ρ. A pattern ρ is said to be *unavoidable* if for all $k \geq 1$ all but finitely many finite words (equivalently every infinite word, by Kőnig's Lemma) over the alphabet $[k]$ encounter ρ. Otherwise we say ρ is *avoidable.* Unavoidable patterns are characterized by the so-called Zimin patterns. For all $n \geq 0$, the n-th *Zimin pattern* Z_n is given by:

$$Z_0 = \varepsilon \qquad \text{and} \quad Z_{n+1} = Z_n x_{n+1} Z_n \quad \text{for all } n \geq 0.$$

For instance, we have $Z_1 = x_1$, $Z_2 = x_1 x_2 x_1$ and $Z_3 = x_1 x_2 x_1 x_3 x_1 x_2 x_1$.

The following theorem gives a decidable characterization of unavoidable patterns.

▶ **Theorem 3** (Bean/Ehrenfeucht/McNulty [6], Zimin [20]). *A pattern ρ containing n different variables is unavoidable if, and only if, Z_n encounters ρ.*

For instance, the pattern $x_1 x_2 x_1 x_2$ is avoidable because it matches $x_1 x_1$ which itself is not encountered in Z_n for all $n \in \mathbb{N}$. This characterization justifies the study of the following Ramsey-like function.

▶ **Definition 4.** Let $n, k \geq 1$. We define $f(n,k) = \min\{\ell \geq 1 \mid \forall w \in [k]^\ell : w \text{ encounters } Z_n\}$.

As we mainly work with Zimin patterns, we introduce the notions of Zimin type (i.e. the maximal Zimin pattern that matches a word) and Zimin index (i.e. the maximal Zimin pattern that a word encounters) and their basic properties.

The *Zimin type* $\mathrm{ZType}(w)$ of a word w is the largest n such that $w = \varphi(Z_n)$ for some non-erasing morphism φ. For instance, we have $\mathrm{ZType}(aaab) = 1$, $\mathrm{ZType}(aba) = 2$ and $\mathrm{ZType}(a^7 b a^7) = 4$. Note that the Zimin type of any non-empty word is greater or equal to 1 and the Zimin type of the empty word is 0.

Following the definition of the Zimin patterns, the Zimin type of a word can be inductively characterized as follows:

▶ **Fact 5.** *For any non-empty word w, $\mathrm{ZType}(w) = 1 + \max\{\mathrm{ZType}(\alpha) \mid w = \alpha\beta\alpha : \alpha, \beta \neq \varepsilon\}$, with the convention that the maximum of the empty set is 0.*

▶ **Definition 6.** The Zimin index $\mathrm{Zimin}(w)$ of a non-empty word w is the maximum Zimin type of an infix of w.

For instance, we have $\mathrm{Zimin}(aaab) = 2$ and $\mathrm{Zimin}(bbaba) = 2$. As a further example note that $\mathrm{Zimin}(baaabaaa) = 3$ although $\mathrm{ZType}(baaabaaa) = 1$.

▶ **Lemma 7.** *For any word w, we have the following properties:*

- $\mathrm{ZType}(w) \leq \mathrm{Zimin}(w)$,
- *for any infix w' of w,* $\mathrm{Zimin}(w') \leq \mathrm{Zimin}(w)$,
- $\mathrm{Zimin}(w) \leq \lfloor \log_2(|w| + 1) \rfloor$.

Proof. The first two points directly follow from the definition. For the last point, note that for a word w to encounter the n-th Zimin pattern Z_n, it must be of length a least $|Z_n|$. As Z_n has length $2^n - 1$, we have $2^{\mathrm{Zimin}(w)} - 1 \leq |w|$, which implies the announced bound. ◀

3 The Zimin index of higher-order counters

In this section we show that there is a family of words, that we refer to as “higher-order counters”, whose lengths are non-elementary in n and whose Zimin index is $n-1$, allowing us to show that $f(2n-1,n) \geq \mathrm{Tower}(n-1,2)$. In Section 3.1 we introduce higher-order counters and in Section 3.2 we show that their Zimin index is precisely $n-1$ including the mentioned lower bound on f.

3.1 Higher-order counters à la Stockmeyer

In this section we introduce counters that encode values ranging from 0 to a tower of exponentials. To the best of our knowledge this construction was introduced by Stockmeyer to show non-elementary complexity lower bounds and is often referred to as the “yardstick construction” [17]. We refer to such counters as “higher-order counters” in the following.

To make the notation less cluttered, we define $\boldsymbol{\tau} : \mathbb{N} \to \mathbb{N}$, the *tower of twos function* which satisfies $\boldsymbol{\tau}(n) = \mathrm{Tower}(n,2)$ for all $n \geq 0$. For all $n \geq 1$, we define an alphabet Σ_n by taking $\Sigma_1 = \{0_1, 1_1\}$ and for all $n > 1$, $\Sigma_n = \Sigma_{n-1} \cup \{0_n, 1_n\}$. We say the symbols 0_n and 1_n have *order* n. We define $\Sigma = \cup_{n \geq 1} \Sigma_n$ to be set of all these symbols.

For all $n \geq 1$ and for all $i \in [0, \boldsymbol{\tau}(n)-1]$, we define a word over Σ_n called *the* i*-th counter of order* n and denoted by $\llbracket i \rrbracket_n$. The definition proceeds by induction on n. For $n = 1$, there are only two counters $\llbracket 0 \rrbracket_1$ and $\llbracket 1 \rrbracket_1$ (recall that $\boldsymbol{\tau}(1) = 2$). We define $\llbracket 0 \rrbracket_1 = 0_1$ and $\llbracket 1 \rrbracket_1 = 1_1$ and for $n \geq 1$ and $i \in [0, \boldsymbol{\tau}(n+1)-1]$ we define

$$\llbracket i \rrbracket_{n+1} = \llbracket 0 \rrbracket_n b_0 \llbracket 1 \rrbracket_n b_1 \cdots \llbracket \boldsymbol{\tau}(n)-1 \rrbracket_n b_{\boldsymbol{\tau}(n)-1},$$

where $b_0 b_1 \cdots b_{\boldsymbol{\tau}(n)-2} b_{\boldsymbol{\tau}(n)-1}$ is the binary decomposition of i over the alphabet $\{0_{n+1}, 1_{n+1}\}$ with b_0 the least significant bit (i.e. $i = \sum_{j=0}^{\boldsymbol{\tau}(n)-1} \overline{b_j} \cdot 2^j$ where $\overline{b_j} = 0$ if $b_j = 0_{n+1}$ and $\overline{b_j} = 1$ if $b_j = 1_{n+1}$).

For $\llbracket 11 \rrbracket_3$, we have $11 = 1 \cdot 2^0 + 1 \cdot 2^1 + 0 \cdot 2^2 + 1 \cdot 2^3$ and hence

$$\llbracket 11 \rrbracket_3 = \underbrace{0_1 0_2 1_1 0_2}_{\llbracket 0 \rrbracket_2} \mathbf{1_3} \underbrace{0_1 1_2 1_1 0_2}_{\llbracket 1 \rrbracket_2} \mathbf{1_3} \underbrace{0_1 0_2 1_1 1_2}_{\llbracket 2 \rrbracket_2} \mathbf{0_3} \underbrace{0_1 1_2 1_1 1_2}_{\llbracket 3 \rrbracket_2} \mathbf{1_3}.$$

The following lemma can easily be proven by induction on n.

▶ **Lemma 8.** *Let* $n \geq 1$*.*

1. *For all* $i \neq j \in [0, \boldsymbol{\tau}(n)-1]$ *we have* $\llbracket i \rrbracket_n \neq \llbracket j \rrbracket_n$*.*
2. *If* $n > 1$*, then for all* $i \in [0, \boldsymbol{\tau}(n)-1]$ *and* $j \in [0, \boldsymbol{\tau}(n-1)-1]$ *the counter* $\llbracket j \rrbracket_{n-1}$ *has exactly one occurence in* $\llbracket i \rrbracket_n$*.*

The length L_n of an order-n counter satisfies the following equations: $L_1 = 1$ and $L_{n+1} = \boldsymbol{\tau}(n) \cdot (L_n + 1)$ for $n \geq 1$. Note that in particular we have $L_n \geq \boldsymbol{\tau}(n-1)$ for all $n \geq 1$.

3.2 Higher-order counters have small Zimin index

The aim of this section is to give an upper bound on the Zimin index of counters of order n. A first simple remark is that the Zimin index of any counter of order n is bounded by the Ziminindex of $\llbracket 0 \rrbracket_n$.

▶ **Lemma 9.** *For all* $n \geq 1$ *and for all* $i \in [0, \boldsymbol{\tau}(n)-1]$*,* $\mathrm{Zimin}(\llbracket i \rrbracket_n) \leq \mathrm{Zimin}(\llbracket 0 \rrbracket_n)$*.*

Proof Sketch. Let $n \geq 1$ and $i \in [0, \boldsymbol{\tau}(n)-1]$. Consider the morphism φ defined by $\varphi(0_n) = \varphi(1_n) = 0_n$ and $\varphi(x) = x$ for $x \in \Sigma_{n-1}$. We have $\varphi(\llbracket i \rrbracket_n) = \llbracket 0 \rrbracket_n$. Moreover as φ is alphabetic, if an infix α of $\llbracket i \rrbracket_n$ matches Z_ℓ for some $\ell \geq 0$ then $\varphi(\alpha)$ is an infix of $\varphi(\llbracket i \rrbracket_n) = \llbracket 0 \rrbracket_n$ that also matches Z_ℓ. The inequality claimed follows. ◀

This leads us to the main result of this section.

▶ **Theorem 10.** *For all $n \geq 3$,* $\mathrm{Zimin}(\llbracket 0 \rrbracket_n) \leq n-1$.

Proof. The proof proceeds by induction on $n \geq 3$. The base case can be checked using a computer program.

Assume that the property holds for some $n \geq 3$. Let us show that $\mathrm{Zimin}(\llbracket 0 \rrbracket_{n+1}) \leq n$. Let $\alpha\beta\alpha$ be an infix of $\llbracket 0 \rrbracket_{n+1}$ for some non-empty words α and β. It is enough to show that $\mathrm{ZType}(\alpha) \leq n-1$. We distinguish the following cases depending on the number occurrences of 0_{n+1} in α.

Case 1: $\boldsymbol{\alpha}$ contains no occurrences of $\mathbf{0_{n+1}}$. Then α is an infix of some $\llbracket i \rrbracket_n$ for some $i \in [0, \boldsymbol{\tau}(n)-1]$. Using Lemma 9 and the induction hypothesis, we have $\mathrm{ZType}(\alpha) \leq \mathrm{Zimin}(\alpha) \leq \mathrm{Zimin}(\llbracket i \rrbracket_n) \leq \mathrm{Zimin}(\llbracket 0 \rrbracket_n) \leq n-1$.

Case 2: $\boldsymbol{\alpha}$ contains at least two occurrences of $\mathbf{0_{n+1}}$. By definition of counters, α has an infix of the form $0_{n+1}\llbracket i \rrbracket_n 0_{n+1}$ for some $i \in [0, \boldsymbol{\tau}(n)-1]$. Hence $\llbracket 0 \rrbracket_{n+1}$ would contain two occurrences of $0_{n+1}\llbracket i \rrbracket_n 0_{n+1}$ which is not possible (cf. Lemma 7).

Case 3: $\boldsymbol{\alpha}$ contains exactly one occurrence of $\mathbf{0_{n+1}}$. By definition of $\llbracket 0 \rrbracket_{n+1}$, there exists $i \neq j \in [0, \boldsymbol{\tau}(n)-1]$ such that α is of the form $u0_{n+1}v$ with u a suffix of both $\llbracket i \rrbracket_n$ and $\llbracket j \rrbracket_n$ and v a prefix of both $\llbracket i+1 \rrbracket_n$ and $\llbracket j+1 \rrbracket_n$.

Consider the morphism ψ the erases all symbols in Σ_{n-1} and replaces 0_n and 1_n by 0 and 1, respectively. Let us assume that

$$\psi(u) = b_{\boldsymbol{\tau}(n-1)-\ell_0} \cdots b_{\boldsymbol{\tau}(n-1)-1} \quad \text{and} \quad \psi(v) = c_0 \cdots c_{\ell_1 - 1}$$

for some $\ell_0 \in [0, \boldsymbol{\tau}(n-1)]$ and $\ell_1 \in [0, \boldsymbol{\tau}(n-1)]$ and $b_k \in \{0,1\}$ for all $k \in [\boldsymbol{\tau}(n-1) - \ell_0, \boldsymbol{\tau}(n-1)-1]$ and $c_k \in \{0,1\}$ for all $k \in [0, \ell_1 - 1]$.

Let us start by showing that $\ell_0 + \ell_1 < \boldsymbol{\tau}(n-1)$.

By definition of counters, $b_{\boldsymbol{\tau}(n-1)-1}$ is the most significant bit of the binary representation (of length $\boldsymbol{\tau}(n-1)$) of i and j and c_0 is the least significant bit of the binary representation of both $i+1$ and $j+1$. More formally, there exist x_i and $x_j \in [0, 2^{\boldsymbol{\tau}(n-1)-\ell_0} - 1]$ and y_i and $y_j \in [0, 2^{\boldsymbol{\tau}(n-1)-\ell_1} - 1]$ such that

$$i = x_i + 2^{\boldsymbol{\tau}(n-1)-\ell_0} \cdot B \quad j = x_j + 2^{\boldsymbol{\tau}(n-1)-\ell_0} \cdot B \quad i+1 = C + 2^{\ell_1} y_i \quad j+1 = C + 2^{\ell_1} y_j$$

with $B = \sum\limits_{k=0}^{\ell_0 - 1} b_{\boldsymbol{\tau}(n-1)-\ell_0-k} \cdot 2^k$ and $C = \sum\limits_{k=0}^{\ell_1 - 1} c_k \cdot 2^k$.

Assume by way of contradiction that $\ell_0 + \ell_1 \geq \boldsymbol{\tau}(n-1)$. In particular, this implies $2^{\ell_1} \geq 2^{\boldsymbol{\tau}(n-1)-\ell_0}$. And hence,

$$\begin{array}{rcll} x_i & = & i \mod 2^{\boldsymbol{\tau}(n-1)-\ell_0} & \text{by definition of } i \\ & = & C - 1 + 2^{\ell_1} y_i \mod 2^{\boldsymbol{\tau}(n-1)-\ell_0} & \\ & = & C - 1 \mod 2^{\boldsymbol{\tau}(n-1)-\ell_0} & \text{as } 2^{\boldsymbol{\tau}(n-1)-\ell_0} \text{ divides } 2^{\ell_1}. \end{array}$$

A similar reasoning shows that $x_j = C - 1 \mod 2^{\boldsymbol{\tau}(n-1)-\ell_0}$. Hence $x_i = x_j$ and hence $i = j$ which brings the contradiction.

As $\ell_0 + \ell_1 < \boldsymbol{\tau}(n-1)$, there exists some $i_0 \in [0, \boldsymbol{\tau}(n) - 1]$ such that v is a prefix and u is a suffix of $[\![i_0]\!]_n$. That is, the binary representation of i_0 of length $\boldsymbol{\tau}(n-1)$ has $c_0 \cdots c_{\ell_1 - 1}$ as ℓ_1 least significant bits and $b_{\boldsymbol{\tau}(n-1)-\ell_0} \cdots b_{\boldsymbol{\tau}(n-1)-1}$ as ℓ_0 most significant bits. In particular, as $\ell_0 + \ell_1 < \boldsymbol{\tau}(n-1)$, we have that $[\![i_0]\!]_n = vru$ for some non-empty r.

We claim that $\mathrm{ZType}(\alpha) \leq \mathrm{Zimin}([\![i_0]\!]_n)$ by which we would be done since then $\mathrm{ZType}(\alpha) \leq \mathrm{Zimin}([\![i_0]\!]_n) \leq \mathrm{Zimin}([\![0]\!]_n) \leq n - 1$ by Lemma 9 and the induction hypothesis.

Assume that α can be written as $\gamma\delta\gamma$ for non-empty γ and δ. Using Fact 5, it is enough to show that $\mathrm{ZType}(\gamma) + 1 \leq \mathrm{Zimin}([\![i_0]\!]_n)$. Recall that $\alpha = u0_{n+1}v$ and α contains only one occurrence of 0_{n+1}. It follows that γ must be a prefix of u and a suffix of v. In particular, $[\![i_0]\!]_n = vru$ contains $\gamma r \gamma$ as an infix. And hence, we have $\mathrm{ZType}(\gamma) + 1 \leq \mathrm{ZType}(\gamma r \gamma) \leq \mathrm{Zimin}([\![i_0]\!]_n)$. ◀

The upper bound on the Zimin index of higher-order counters established in the previous theorem is tight.

▶ **Theorem 11.** *For all $n \geq 3$ and $i \in [0, \boldsymbol{\tau}(n) - 1]$, $\mathrm{Zimin}([\![i]\!]_n) = n - 1$.*

For $n \geq 3$, the counter $[\![0]\!]_n$ over the alphabet $\{0_1, 1_1, \ldots, 0_{n-1}, 1_{n-1}, 0_n\}$ of size $2n - 1$ has Zimin index at most $n - 1$. In particular, $[\![0]\!]_n$ does not encounter the pattern Z_n. Therefore its length $L_n \geq \boldsymbol{\tau}(n-1)$ gives a lower bound for the value of $f(n, 2n-1)$.

▶ **Corollary 12.** *For all $n \geq 3$, $f(n, 2n-1) \geq \boldsymbol{\tau}(n-1) = \mathrm{Tower}(n-1, 2)$.*

4 Reduction to the binary alphabet

In this section, we show how to encode a higher-order counter seen in Section 3 over the binary alphabet $\{0, 1\}$ while still preserving a relatively low Zimin index. For this we apply to higher-order counters, the morphism ψ defined for all $n \geq 1$, as follows

$$\psi(0_n) = 00\,(01)^{n-1}\,00 \qquad \psi(1_n) = 11\,(01)^{n-1}\,11.$$

For all $n \geq 1$ and $i \in [0, \boldsymbol{\tau}(n) - 1]$, we define $\{\!\{ i \}\!\}_n = \psi([\![i]\!]_n)$.

The set of images of the letters in Σ by this morphism forms what is known as an infix code, i.e. $\psi(a)$ is not an infix of $\psi(b)$ for any two letters $a, b \in \Sigma$ with $a \neq b$. In addition to being an infix code, the morphism was designed so that we are able to attribute a non-ambiguous partial decoding to most infixes of an encoded word (cf. Lemma 17) and the encoding of 0_n and 1_n differ by their first and last symbol.

Applying a non-erasing morphism to a word can only increase its Zimin index. We will see in the remainder of this section that the Zimin index of higher-order counters is increased by at most 2 when the morphism ψ is applied. It is possible that another choice of morphism would bring a better upper bound. However, note that the proof we present is tightly linked to the above-mentioned properties of ψ that are decisive for the proof to work.

This section is devoted to establishing the following theorem.

▶ **Theorem 13.** *For all $n \geq 2$ and for all $i \in [0, \boldsymbol{\tau}(n) - 1]$, $\mathrm{Zimin}(\{\!\{ i \}\!\}_n) \leq n + 1$.*

Recalling that an order-n counter has length at least $\boldsymbol{\tau}(n-1)$ and that applying ψ can only increase the length, we immediately obtain from Theorem 13 a non-elementary lower bound for $f(n, 2)$ whenever $n \geq 4$.

▶ **Corollary 14.** *For all $n \geq 4$, $f(n,2) \geq \boldsymbol{\tau}(n-3) = \mathrm{Tower}(n-3,2)$.*

The proof of Theorem 13 which is essentially a reduction to the proof of Theorem 10 is given in Section 4.2. To perform this reduction, we first establish basic properties of the morphism ψ and its decoding in Section 4.1.

4.1 Parsing the code ψ

A word $w \in \{0,1\}^*$ is *coded* by ψ (or simply a *coded word*) if it is the image by ψ of some word v over Σ^*. As the image of ψ is an infix code, the word $v \in \Sigma^*$ is unique. However in our proof, we need to take into consideration all infixes of a coded word. To be able to reuse the proof techniques of Theorem 10, it is necessary to associate to an infix of a coded word a partial decoding called a *parse*.

Let us therefore consider the following sets, $C = \psi(\Sigma) = \{\psi(0_k) \mid k \geq 1\} \cup \{\psi(1_k) \mid k \geq 1\}$ the image of the morphism, $L = \{v \in \{0,1\}^* \mid \exists u \in \{0,1\}^+, uv \in C\}$ the set of strict suffixes of C, $R = \{u \in \{0,1\}^* \mid \exists v \in \{0,1\}^+, uv \in C\}$ the set of strict prefixes of C and $F = \{v \in \{0,1\}^* \mid \exists u, w \in \{0,1\}^+, uvw \in C\}$ the set of strict infixes of C.

It is easy to see that every infix of a coded word belongs to $F \cup LC^*R$. This leads us to define a *parse* p as a triple (ℓ, u, r) in $L \times \Sigma^* \times R$. The word u will be called the *center* of the parse p. The *value* of the parse (ℓ, u, r) is the word $\ell\psi(u)r \in \{0,1\}^*$. We say that α *admits a parse* p if α is the value of p.

By the above fact, for all coded words all of its infixes not belonging to F have at least one parse. However, the parse is not necessarily unique. For instance, consider the infix $\alpha = 0000000000 = 0^{10}$ which appears in $\psi(0_1 0_1 0_1)$. It can be parsed as $(\varepsilon, 0_1 0_1, 00)$, $(0, 0_1 0_1, 0)$ and as $(00, 0_1 0_1, \varepsilon)$. However, we will provide sufficient conditions on an infix to admit a unique parse.

▶ **Definition 15.** A word $\alpha \in \{0,1\}^*$ is *simple* if either $|\alpha| < 11$, or α belongs to F or 0^{10} is an infix of α or 1^{10} is an infix of α.

This definition will be justified by the fact that for non-simple infixes there is exactly one possible parse. Moreover the term simple is justified in the context of this proof by the fact that simple infixes of $\{\!\!\{\, i \,\}\!\!\}_n$ can be shown to have Zimin index at most $n-1$ for all $n \geq 4$ and all $i \in [0, \boldsymbol{\tau}(n) - 1]$.

▶ **Lemma 16.** *For all $n \geq 4$, for all $i \in [0, \boldsymbol{\tau}(n) - 1]$, all simple infixes of $\{\!\!\{\, i \,\}\!\!\}_n$ have Zimin index at most $n-1$.*

▶ **Lemma 17.** *Any non-simple infix of a coded word admits a unique parse.*

Thus, we will refer to the unique parse of a non-simple infix α of a coded word as *the parse of α*.

The notion of occurrence naturally extends to parses. Let $w = w_0 \cdots w_{|w|-1} \in \Sigma^*$ and $p = (\ell, u = u_0 \cdots u_{|u|-1}, r)$ be a parse, an *occurrence of p in w* is an occurrence m of u in w such that whenever ℓ is non-empty we have $m \neq 0$ and ℓ is suffix of $\psi(w_{m-1})$ and similarly whenever r is non-empty we have $m + |u| < |w|$ and r is a prefix of $\psi(w_{m+|u|})$. For a word $w \in \Sigma^*$ and a non-simple infix α of $\psi(w)$, there is a one-to-one correspondence between the occurrences of α in $\psi(w)$ and the occurrences of its parse p_α in $\psi(w)$.

▶ **Definition 18.** For an occurence m of a parse $p = (\ell, u, r)$ in w, we define its context as the word in Σ^* equal to $w[m - \delta_0, m + |u| + \delta_1]$, where $\delta_0 = 0$ if $\ell = \varepsilon$ and $\delta_0 = 1$ otherwise and $\delta_1 = 0$ if $r = \varepsilon$ and $\delta_1 = 1$ otherwise.

By definition the context c of some occurrence of a parse $p = (\ell, u, r)$ in w is an infix of w containing u as an infix. Moreover the value α of p is an infix of $\psi(c)$.

4.2 Upper bound on the Zimin index

To establish Theorem 13, we prove the following stronger statement.

▶ **Theorem 19.** *For all $n \geq 2$ and for all $i \in [0, \boldsymbol{\tau}(n) - 1]$,*

$$\begin{array}{rcl@{\qquad}rcl} \mathrm{Zimin}(\{\!\{\, i \,\}\!\}_n \psi(0_{n+1})) & \leq & n+1, & \mathrm{Zimin}(\{\!\{\, i \,\}\!\}_n \psi(1_{n+1})) & \leq & n+1, \\ \mathrm{Zimin}(\psi(0_{n+1}) \{\!\{\, i \,\}\!\}_n) & \leq & n+1, & \mathrm{Zimin}(\psi(1_{n+1}) \{\!\{\, i \,\}\!\}_n) & \leq & n+1. \end{array}$$

Proof Sketch. We proceed by induction on n. For the case $n = 2$ and $n = 3$, the property is checked using a computer program. For the induction step assume that the property holds for some $n \geq 3$ and let us show that it holds for $n + 1$. Let $i \in [0, \boldsymbol{\tau}(n + 1) - 1]$, we will only show that $\mathrm{Zimin}(\{\!\{\, i \,\}\!\}_{n+1}) \leq n + 2$. The upper bound on $\mathrm{Zimin}(\{\!\{\, i \,\}\!\}_{n+1}, \psi(0_{n+2}))$, $\mathrm{Zimin}(\{\!\{\, i \,\}\!\}_{n+1} \psi(1_{n+2}))$, $\mathrm{Zimin}(\psi(0_{n+2}) \{\!\{\, i \,\}\!\}_{n+1})$ and $\mathrm{Zimin}(\psi(1_{n+2}) \{\!\{\, i \,\}\!\}_{n+1})$ can be deduced from this, but still require a tedious case analysis.

Let $\alpha\beta\alpha$ be an infix of $\{\!\{\, i \,\}\!\}_{n+1}$ for some non-empty α and β. It is enough to show that $\mathrm{ZType}(\alpha) \leq n + 1$. By Lemma 16, we only need to consider the case when α is non-simple. Let (ℓ, u, r) be the parse of α.

Let m be an occurrence of $\alpha\beta\alpha$ in $\{\!\{\, i \,\}\!\}_{n+1}$. In particular, m and $m + |\alpha\beta|$ are two occurrences of α in $\{\!\{\, i \,\}\!\}_{n+1}$. Hence there are two corresponding occurrences m_1 and m_2 of the parse p in $[\![\, i \,]\!]_{n+1}$. Consider the contexts c_1 and c_2 of p that correspond to the occurrences m_1 and m_2, respectively. Note that without further hypothesis c_1 and c_2 are not necessarily equal.

We distinguish cases depending on the number of occurrences of a symbol of order $n + 1$ in c_1. The cases where c_1 contains 0 or more than 2 symbols of order $n + 1$ are treated in a similar fashion as in the proof of Theorem 10. Assume that c_1 contains one and only one symbol of order $n + 1$.

As c_1 is an infix of $[\![\, i \,]\!]_{n+1}$ with one symbol of order $n + 1$, there exists $k_0 \in [0, \boldsymbol{\tau}(n) - 2]$ and some $b \in \{0_{n+1}, 1_{n+1}\}$ such that $c_1 = xby$ where $x \in \Sigma_n^*$ is a suffix of $[\![\, k_0 \,]\!]_n$ and $y \in \Sigma_n^*$ is a prefix of $[\![\, k_0 + 1 \,]\!]_n$.

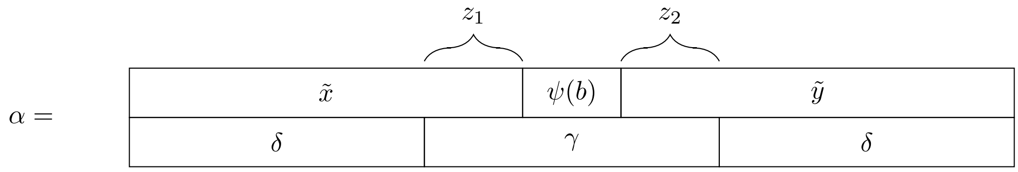
Note that if x or y are empty, we can conclude using the induction hypothesis. From now on, we assume that x and y are non-empty. In particular, the center u of the parse $p = (\ell, u, r)$ contains b and can therefore be uniquely written as $u = \underline{x}b\underline{y}$. Thus, $c_1 = xby$, $\alpha = \ell\psi(\underline{x})\psi(b)\psi(\underline{y})r$, $x = s\underline{x}$ and $y = \underline{y}t$ for some s and t such that $s = \varepsilon$ if $\ell = \varepsilon$ and $s \in \Sigma$ otherwise, where ℓ is a suffix of $\psi(s)$ and $t = \varepsilon$ if $r = \varepsilon$ and $t \in \Sigma$ otherwise, where r is a prefix of $\psi(t)$.

Claim 1. The context c_2 (of the second occurrence m_2 of α) is equal to c_1.

As $c_1 = c_2 = xby$ and as b belongs to the center u of the parse, the infix α can be written as $\alpha = \tilde{x}\psi(b)\tilde{y}$ where $\tilde{x}$ is a suffix of $\psi(x)$ and $\tilde{y}$ is a prefix of $\psi(y)$.

Claim 2. There exists $j_0 \in [0, \boldsymbol{\tau}(n) - 1]$ and a non-empty χ such that $\tilde{y}\chi\tilde{x} = \{\!\{\, j_0 \,\}\!\}_n$.

Let us now consider an arbitrary decomposition of α as $\delta\gamma\delta$ for non-empty δ and γ. Recall that it is enough to show that $\mathrm{ZType}(\alpha) \leq n + 1$ or that $\mathrm{ZType}(\delta) \leq n$. There are several cases to consider depending on how the decompositions of α as $\tilde{x}\psi(b)\tilde{y}$ and $\delta\gamma\delta$ overlap. We only present here one of the 6 cases where $|\delta| < |\tilde{x}| \leq |\delta\gamma|$ and $|\tilde{x}\psi(b)| \leq |\delta\gamma|$.



In this case $\gamma = z_1\psi(b)z_2$ with $z_1 \neq \varepsilon$ such that $\tilde{x} = \delta z_1$ and $\tilde{y} = z_2\delta$. Recall that there exists $j_0 \in [0, \boldsymbol{\tau}(n) - 1]$ and a non-empty χ such that $\tilde{y}\chi\tilde{x} = \{\!\{\, j_0 \,\}\!\}_n$. We have $\mathrm{Zimin}(\{\!\{\, j_0 \,\}\!\}_n) \leq \mathrm{Zimin}(\{\!\{\, j_0 \,\}\!\}_n\psi(0_{n+1})) \leq n+1$, where the last inequality follows from the induction hypothesis. Hence, $\{\!\{\, j_0 \,\}\!\}_n = \tilde{y}\chi\tilde{x} = z_2\delta\chi\delta z_1$ has Zimin index at most $n+1$. This implies that δ has Zimin type of at most n which concludes this case. ◀

5 Avoiding Zimin patterns in the abelian sense

Matching a pattern in the abelian sense is a weaker condition, where one only requires that all infixes that are matching a pattern variable must have the same number of occurrences of each letter (instead of being the same words). Hence, for two words $x, y \in A^*$ we write $x \equiv y$ if $|x|_a = |y|_a$ for all $a \in A$. Let $\rho = \rho_1 \cdots \rho_n$ be a pattern, where $\rho_i \in \mathcal{X}$ is a pattern variable for all $i \in [k]$. An *abelian factorization of a word $w \in A^*$ for the pattern ρ* is a factorization $w = w_1 \cdots w_n$ such that $w_i \neq \varepsilon$ for all $i \in [n]$ and $\rho_i = \rho_j$ implies $w_i \equiv w_j$ for all $i, j \in [n]$. A word $w \in A^*$ *matches the pattern ρ in the abelian sense* if there is an abelian factorization of w for ρ. The definitions when a word encounters a pattern in the abelian sense and when a pattern is unavoidable in the abelian sense are as expected.

We note that every pattern that is unavoidable is in particular unavoidable in the abelian sense. However, the converse does not hold in general as witnessed by the pattern $xyzxyxuxyxzyx$ as shown in [5].

To the best of the authors' knowledge abelian unavoidability still lacks a characterization in the style of general unavoidability in terms of Zimin patterns; we refer to [4] for some open problems and conjectures. Although being possibly less meaningful as for general unavoidability, the analogous Ramsey-like function for abelian unavoidability has been studied. For $n, k \geq 1$ we define $g(n,k) = \min\{\ell \geq 1 \mid \forall w \in [k]^\ell : w \text{ encounters } Z_n \text{ in the abelian sense}\}$. Clearly, $g(n,k) \leq f(n,k)$ and to the best of the authors' knowledge no elementary upper bound has been shown for g so far. By applying a combination of the probabilistic method [1] and of analytic combinatorics [7] Tao showed a lower bound for g given by the first inequality below. Unfortunately, it was not clear to us what the asymptotic behavior of this lower bound is. However Jugé [9] provided us with an estimate of its asymptotic behavior.

▶ **Theorem 20** (Tao [18], Corollary 3. Jugé [9]). *Let $k \geq 4$. Then*

$$g(n,k) \geq (1+o(1))\sqrt{2\prod_{j=1}^{n-1}\left[\sum_{\ell=1}^{\infty}\frac{1}{k^{2^j\ell}}\sum_{i_1+\cdots+i_k=\ell}\binom{\ell}{i_1,\ldots,i_k}\right]^{-1}} \geq \left(\frac{1}{\sqrt{21}}+o(1)\right)\frac{k^{2^{n-1}}}{k^{(n+1)/2}}.$$

In Section 5.1 we prove another doubly-exponential lower bound on g by applying the first moment method [1]. Our lower bound on g is not as good as the one obtained by Theorem 20 but its proof seems more direct. The proof follows a similar strategy as the (slightly better) doubly-exponential lower bound for f from [2], but again, seems to be more direct. Our novel contribution is to provide a matching doubly-exponential upper bound on g in Section 5.2. Note that Tao in [18] only provides a non-elementary upper bound for the non-abelian case.

5.1 A simple lower bound via the first-moment method

For all $n \geq 1$ let $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ denote the set of the first n pattern variables. Note that the variable x_i appears precisely 2^{n-i} times in Z_n and its first occurrence is at position 2^{i-1} for all $i \in [1, n]$. An *abelian occurrence of* Z_n in a word w is a pair $(j, \lambda) \in [0, |w|-1] \times \mathbb{N}^{\mathcal{X}_n}$ for which there is an factorization $w = uvz$ with $|u| = j$ and an abelian factorization $v_1 \cdots v_{2^n-1}$ of v for Z_n satisfying $\lambda(x_i) = |v_{2^{i-1}}|$.

By applying the probabilistic method [1] we show a lower bound for $g(n,k)$ that is doubly-exponential in n for every fixed $k \geq 2$. The proof is similar the lower bound proof from [2].

▶ **Theorem 21.** *For all $n \geq 1$ and all $k \geq 2$, $g(n,k) > k^{\lfloor \frac{2^n}{n+2} \rfloor - 1}$.*

Proof. For $n, \ell \geq 1$ let $\Delta_{n,k,\ell}$ denote the expected number of abelian occurrences of Z_n in a random word in the set $[k]^\ell$. Note that we always consider the uniform distribution over words. If $\Delta_{n,k,\ell} < 1$, then by the probabilistic method [1] there exists a word of length ℓ over the alphabet $[k]$ that does *not* encounter Z_n in the abelian sense; hence we can conclude that $g(n,k) > \ell$. Therefore we investigate those $\ell = \ell(n,k)$ for which we can guarantee that $\Delta_{n,k,\ell} < 1$. We need two intermediate claims.

Claim 1. Let $A_{k,h}$ denote the event that two independent random words u and v in $[k]^h$ satisfy $u \equiv v$. Then $\Pr(A_{k,h}) \leq 1/k$ for all $h \geq 1$.

It follows that the probability that m random words $w_1, w_2 \ldots, w_m \in [k]^h$ satisfy $w_1 \equiv w_2 \equiv \cdots \equiv w_m$ is at most $(1/k)^{m-1}$. Recall that $Z_n = y_1 \cdots y_{2^n-1}$, where $y_i \in \{x_1, \ldots, x_n\}$ for all $i \in [2^n - 1]$ and that the variable x_i appears precisely 2^{n-i} times in Z_n. We recall that we would like to bound the expected number of occurrences (in the abelian sense) of Z_n in a random word of length ℓ over the alphabet $[k]$. To account for this, we define for each mapping $\lambda : \mathcal{X}_n \to \mathbb{N}^+$ its *width* as $\mathrm{width}(\lambda) = \sum_{i=1}^n 2^{n-i} \cdot \lambda(x_i)$. For every word v of length $\mathrm{width}(\lambda)$ its (unique) *decomposition with respect to* λ is the unique factorization $v = v_1 \cdots v_{2^n-1}$ such that $y_j = x_i$ implies $|v_j| = \lambda(x_i)$ for all $j \in [2^n - 1]$ and all $i \in [n]$. Using the estimations in Claim 1 one can now show the following claim.

Claim 2. Let $\lambda : \mathcal{X}_n \to \mathbb{N}^+$ of width d and let B_λ denote the event that $(0, \lambda)$ is an occurrence in the abelian sense of Z_n in a random word from $[k]^d$. Then $\Pr(B_\lambda) \leq k^{n-2^n+1}$.

The probability that (j, λ) is an occurrence of Z_n in a random word from $[k]^\ell$ with $\ell \geq j + \mathrm{width}(\lambda)$ is equal to the probability that $(0, \lambda)$ is an occurrence of Z_n in a random word from $[k]^d$ (which is $\Pr(B_\lambda)$). Thus, this probability does not depend on j.

We are ready to prove an upper bound for $\Delta_{n,k,\ell}$, keeping in mind that any occurrence (j, λ) of Z_n in a random word of length ℓ must satisfy $\mathrm{width}(\lambda) \geq 2^n - 1$.

$$\begin{aligned}
\Delta_{n,k,\ell} &\leq \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} \sum_{\substack{\lambda:\mathcal{X}_n \to \mathbb{N}^+ \\ \mathrm{width}(\lambda)=d}} \Pr\left[(j,\lambda) \text{ is an ab. occ. of } Z_n \text{ in a random word in } [k]^\ell\right] \\
&\leq \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} \sum_{\substack{\lambda:\mathcal{X}_n \to \mathbb{N}^+ \\ \mathrm{width}(\lambda)=d}} \Pr(B_\lambda) \quad \overset{\text{Claim 2}}{\leq} \quad \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} \sum_{\substack{\lambda:\mathcal{X}_n \to \mathbb{N}^+ \\ \mathrm{width}(\lambda)=d}} k^{n-2^n+1} \\
&\leq \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} d^n \cdot k^{n-2^n+1} \quad \leq \quad \frac{\ell^2 \cdot \ell^n}{k^{2^n-n-1}} \quad = \quad \frac{\ell^{n+2}}{k^{2^n-n-1}} \qquad (1)
\end{aligned}$$

It remains to determine a sufficiently large ℓ that guarantees $\Delta_{n,k,\ell} < 1$. Using the previous inequalities it is not difficult to show that $\ell = k^{\lfloor \frac{2n}{n+2} \rfloor - 1}$ ensures that $\Delta_{n,k,\ell} < 1$. ◀

5.2 A doubly-exponential upper bound

Let us finally prove an upper bound for $g(n,k)$ that is doubly-exponential in n.

▶ **Theorem 22.** *For all $n, k \geq 1$, $g(n,k) \leq 2^{(4k)^n (n-1)!}$.*

Proof. For all $n \geq 1$, we will show the following inequality:

$$g(n+1,k) \quad \leq \quad (g(n,k)+1)(g(n,k)^{kn}+1) \tag{2}$$

A simple induction on n then shows the claimed upper bound.

To show (2), we consider words w over $[k]$ of length at least $(g(n,k)+1)(g(n,k)^{kn}+1)$. Such words can always be written as $w_1 a_1 w_2 a_2 \cdots w_m a_m z$, where $m = g(n,k)^{kn} + 1$, $|w_j| = g(n,k)$ and $a_j \in [k]$ for all $j \in [m]$ and $z \in [k]^*$.

By definition of $g(n,k)$, for all $j \in [m]$, the word w_j encounters Z_n in the abelian sense, witnessed in some infix v_j by some abelian factorization $v_j = v_j^{(1)} \cdots v_j^{(2^n-1)}$ for Z_n. For each such abelian factorization, it is natural to associate with every $i \in [n]$, the (unique) Parikh image of the words $v_j^{(h)}$ of the factorization that correspond to the different occurrences of the variable x_i in Z_n. Formally, each of the above abelian factorizations $v_j = v_j^{(1)} \cdots v_j^{(2^n-1)}$ induces a mapping $\psi_j : \mathcal{X}_n \to \mathbb{N}^{[k]}$ such that $\psi_i(x_i)(a) = |v_j^{(2^i-1)}|_a$ for all $j \in [m]$, all $i \in [n]$ and all $a \in [k]$. As expected, we write $\psi_j \equiv \psi_h$ if $\psi_j(x_i) = \psi_j(x_i)$ for all $i \in [n]$. Note that if there are distinct $j, h \in [1,m]$ with $\psi_j \equiv \psi_h$, then w encounters $Z_{n+1} = Z_n x_{n+1} Z_n$ in the abelian sense. It is easy to see that there are at most $g(n,k)^{kn}$ different equivalence classes for the ψ_j with respect to $\equiv$. Therefore as $m = g(n,k)^{kn} + 1$, there are always two distinct indices $i, j \in [1,m]$ that satisfy $\psi_i \equiv \psi_j$ and we have established (2). ◀

6 Conclusion

We have established a lower bound for $f(n,k)$ that is already non-elementary when $k = 2$. A natural question is whether the non-elementary lower bound for $f(n,k)$ obtained by an explicit construction in this article can be obtained using the probabilistic method. A first hint of an answer is that the first moment method used in [2] cannot be used to obtain a lower bound that is asymptotically above doubly-exponential. Indeed, as for a length $\ell \geq k^{2^n-n-2} + 2^n$, the expected number $\Delta_{n,k,\ell}$ of occurrences Z_n in a random word in $[k]^\ell$ is greater than 1.

To see this, recall that $|Z_n| = 2^n - 1$ and hence there is at most one possible occurrence of Z_n in any word of length $2^n - 1$. Let A_n denote the event that Z_n is encountered in a random word in $[k]^{2^n-1}$. We have $\Pr(A_n) = \prod_{i=1}^{n} (1/k)^{2^{n-i}-1} = k^{-2^n+n+2}$.

Assume that $\ell \geq k^{2^n-n-2} + 2^n$. For each $i \in [0, k^{2^n-n-2}]$, let X_i be the indicator random variable marking that the infix, of a random word in $[k]^\ell$, occurring at i and of length $2^n - 1$ matches Z_n. By linearity of the expectation, the following lower bound holds, $\Delta_{n,k,\ell} \geq \sum_{i=0}^{K} E(X_i) \geq (K+1)\Pr(A_n) = 1 + \dfrac{1}{K} \geq 1$, where $K = k^{2^n-n-2}$.

Acknowledgments. The authors would like to thank Vincent Jugé for his help and the anonymous reviewers for their valuable feedback.

References

1 N. Alon and J. Spencer. *The Probabilistic Method.* Wiley, 2015.

2 J. Cooper and D. Rorabaugh. Bounds on Zimin word avoidance. *CoRR*, abs/1409.3080, 2014. URL: http://arxiv.org/abs/1409.3080.

3 J. Cooper and D. Rorabaugh. Asymptotic density of Zimin words. *Discrete Mathematics & Theoretical Computer Science*, Vol. 18, no 3, 2016. URL: http://dmtcs.episciences.org/1414.

4 J. D. Currie. Pattern avoidance: themes and variations. *Theor. Comput. Sci.*, 339(1):7–18, 2005. doi:10.1016/j.tcs.2005.01.004.

5 J. D. Currie and V. Linek. Avoiding patterns in the abelian sense. *Canadian J. Math.*, 51(4):696–714, 2001. doi:10.4153/cjm-2001-028-4.

6 G. F. McNulty D. R. Bean, A. Ehrenfeucht. Avoidable Patterns in Strings of Symbols. *Pac. J. of Math.*, 85:261–294, 1979. doi:10.2140/pjm.1979.85.261.

7 P. Flajolet and R. Sedgewick. *Analytic combinatorics.* Cambridge University Press, Cambridge, 2009.

8 P. Jancar. Equivalences of pushdown systems are hard. In *Proceedings of FOSSACS 2014*, volume 8412 of *Lecture Notes in Computer Science*, pages 1–28. Springer, 2014. doi:10.1007/978-3-642-54830-7_1.

9 V. Jugé. Abelian Ramsey length and asymptotic lower bounds. *CoRR*, abs/1609.06057, 2016. URL: http://arxiv.org/abs/1609.06057.

10 K. Reinhardt. The complexity of translating logic to finite automata. In *Automata, Logics, and Infinite Games: A Guide to Current Research*, volume 2500 of *Lecture Notes in Computer Science*, pages 231–238. Springer, 2001. doi:10.1007/3-540-36387-4_13.

11 D. Rorabaugh. Toward the combinatorial limit theory of free words. *CoRR*, abs/1509.04372, 2015. URL: http://arxiv.org/abs/1509.04372.

12 W. Rytter and A. M. Shur. Searching for Zimin patterns. *Theor. Comput. Sci.*, 571:50–57, 2015. doi:10.1016/j.tcs.2015.01.004.

13 M. V. Sapir. Combinatorics on words with applications. Technical report, LITP, 1995.

14 S. Schmitz. Complexity hierarchies beyond elementary. *CoRR*, abs/1312.5686, 2014. URL: http://arxiv.org/abs/1312.5686.

15 G. Sénizergues. The equivalence problem for t-turn DPDA is co-NP. Technical Report 1297-03, LaBRI, 2003. available at http://dept-info.labri.u-bordeaux.fr/~ges.

16 C. Stirling. Deciding DPDA equivalence is primitive recursive. In *In Proceedings of ICALP 2002*, volume 2380 of *Lecture Notes in Computer Science*, pages 821–832. Springer, 2002. doi:10.1007/3-540-45465-9_70.

17 L. J. Stockmeyer. *The complexity of decision problems in automata and logic.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1974.

18 J. Tao. Pattern occurrence statistics and applications to the ramsey theory of unavoidable patterns. *CoRR*, abs/1406.0450, 2014. URL: http://arxiv.org/abs/1406.0450.

19 A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Skrifter I Mat.-Nat. Kl. Christiania*, 7:1—22, 1906.

20 A. I. Zimin. Blocking sets of terms. *Math. USSR Sbornik*, 47:50–57, 1984. doi:10.1070/sm1984v047n02abeh002647.