On long words avoiding Zimin patterns

Arnaud Carayol · Stefan Göller

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.

A. Carayol
Laboratoire d'informatique Gaspard Monge (UMR 8049)
Cité Descartes
5, boulevard Descartes
Champs-sur-Marne
77454 MARNE-LA-VALLEE Cedex 2
France
E-mail: carayol@u-pem.fr

S. Göller
Fachgebiet Theoretische Informatik / Komplexe Systeme
Fachbereich Elektrotechnik / Informatik
Universität Kassel
Wilhelmshöher Allee 73
34121 Kassel
Germany
E-mail: stefan.goeller@uni-kassel.de

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 finite alphabet must contain two non-consecutive occurrences of the same letter. On the other hand, the pattern xx is avoidable as Thue [23] 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 [24] and independently by Bean, Ehrenfeucht and McNulty [10], see also for a more recent proof [17]. This characterization is based on a family $(Z_n)_{n\geq 0}$ of unavoidable patterns, called the Zimin patterns, where

$$Z_1 = x_1 \quad \text{and} \quad Z_{n+1} = Z_n x_{n+1} Z_n \quad \text{for all } n \geq 1.$$

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

"what is 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 [6,22,16,7].

For small values of n and k, known results from [16,15] are summarized in the following table.

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

In general, Cooper and Rorabaugh [6, 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 $\text{Tower} : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ inductively as follows: $\text{Tower}(0, k) = 1$ and $\text{Tower}(n + 1, k) = k^{\text{Tower}(n,k)}$ for all $n, k \in \mathbb{N}$.

Theorem 1 (Cooper/Rorabaugh [6]) *For all $n \geq 1$ and $k \geq 2$, $f(n, k) \leq \text{Tower}(n - 1, K)$, 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 [6]) $f(n, k) \geq k^{2^{n-1}(1+o(1))}$.

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.

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 [21] to construct for each $n \geq 1$, a family of words of length at least $\text{Tower}(n - 1, 2)$ (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

$$f(n, 2n - 1) > \text{Tower}(n - 1, 2).$$

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 [14] 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

$$f(n, 2) > \text{Tower}(n - 3, 2).$$

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 when an infix matches a pattern variable it must only have the same number of occurrences of each letter (instead of being the same words). This gives rise to the notion of avoidable in the abelian sense and unavoidable pattern in the abelian sense. 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 [9]. 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 [22]. For this function, Tao [22] establishes a lower bound that turns out to be doubly-exponential from the estimations in [13]. The upper bound is inherited from the non-abelian setting and is hence non-elementary. We improve this

upper bound to doubly-exponential. 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 [13].

Comparison with [5]. While finalizing the present article, we became aware of the preprint [5] submitted in April 2017 in which Condon, Fox and Sudakov independently obtained non-elementary lower bounds for the function f. To keep the presentation clear, we present their contributions in this dedicated subsection.

Firstly the authors improve the upper-bound of f of Theorem 1 by showing [5, Theorem 2.1] that for all $n \geq 3$, $k \geq 35$,

$$f(n,k) \leq \text{Tower}(n-1,k).$$

They determine the value of $f(3,k)$ up to a multiplicative constant and show [5, Theorem 1.3] that:
$$f(3,k) = \Theta(2^k k!).$$

For the general case, they show [5, Theorem 1.1] for any fixed $n \geq 3$,

$$f(n,k) \geq k \underbrace{^{k^{k^{\cdots^{k-o(k)}}}}}_{n-1 \text{ times}}$$

They provide two proofs of this theorem. The first proof is based on the probabilistic method and positively answer a question we ask in conclusion of our conference paper [3]. The second proof uses a counting argument. For the case of the binary of alphabet, they show [5, Theorem 1.2] that:

$$f(n,2) \geq \text{Tower}(n-4,2).$$

In this last case, our bound is slightly better and has the extra advantage to provide a concrete family of words witnessing the bound.

Applications 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 [20] (in $\mathbf{F}_3$ in terms of Schmitz' classification [18]), see also [12] for a more recent presentation.

In [19] Sénizergues generalizes Stirling's approach by a the so-called "subwords lemma" allowing him both to prove a coNP upper bound for the equivalence problem of finite-turn dpda and to explicitly link the complexity of dpda equivalence with the growth of the function f: he shows that in case f is elementary, then the complexity of dpda equivalence is elementary.

Inspired by this insight, a closer look reveals that the above-mentioned function f has the same importance in all complexity upper bound proofs [20,19,12] for dpda equivalence. However, due to Theorem 7 one cannot hope to improve the computational complexity of dpda equivalence by proving an

elementary upper bound on f since $f(n, k)$ is shown to grow non-elementarily even for $k = 2$ (Theorem 7).

Organization of the paper. We introduce necessary notations in Section 2. We show that $f(n, 2n - 1) \geq \text{Tower}(2, n - 1)$ 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 define $[i, j] = \{i, i+1, \ldots, j\}$ and $[j] = \{1, \ldots, j\}$. 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*[2] of u. If v is an infix u and u can be written as $z_1 v z_2$, 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 $\psi : A^* \to B^*$ that satisfies $\psi(uv) = \psi(u)\psi(v)$ for all $u, v \in A^*$. Thus, every morphism can simply be given by a function from A to B^*. A morphism ψ is said to be *non-erasing* if $\psi(a) \neq \varepsilon$ for all $a \in A$ and ψ 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 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 $n \geq 1$, the n-th *Zimin pattern* Z_n is given by:

$$
\begin{aligned}
Z_0 &= \varepsilon \\
Z_1 &= x_1 \\
Z_{n+1} &= Z_n x_{n+1} Z_n \quad \text{for } n \geq 0.
\end{aligned}
$$

---

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

[2] Again, remark that our definition is slightly non-standard as strict prefixes or strict suffixes are in general not strict infixes.

For instance, we have $Z_1 = x_1$, $Z_2 = x_1x_2x_1$ and $Z_3 = x_1x_2x_1x_3x_1x_2x_1$.

The following statement gives a decidable characterization of unavoidable patterns.

Theorem 3 (Bean/Ehrenfeucht/McNulty [10], Zimin [24]) *A pattern ρ containing n different variables is unavoidable if, and only if, Z_n encounters ρ.*

For instance, the pattern $x_1x_2x_1x_2$ is avoidable because it is not encountered in Z_2 (not even in Z_n for any $n \in \mathbb{N}$).

Theorem 3 justifies the study of the following Ramsey-like function.

Definition 1 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.

Definition 2 The Zimin type $\text{ZType}(w)$ of a word w is the largest n such that $w = \varphi(Z_n)$ for some non-erasing morphism φ.

For instance, we have $\text{ZType}(aaab) = 1$, $\text{ZType}(aba) = 2$ and $\text{ZType}(a^7ba^7) = 4$. Remark 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 Zimin patterns, the Zimin type of a word can be inductively characterized as follows:

Fact 1 *For any non-empty word w,*

$$\text{ZType}(w) = 1 + \max\{\text{ZType}(\alpha) \mid w = \alpha\beta\alpha \text{ for non-empty } \alpha \text{ and } \beta\},$$

with the convention that the maximum of the empty set is 0.

Definition 3 The Zimin index $\text{Zimin}(w)$ of a non-empty word w the maximum Zimin type of an infix of w.

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

Lemma 1 *For any word w, we have the following properties:*

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

Proof The first two points directly follow from the definition. For the last point, remark 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

$$2^{\text{Zimin}(w)} - 1 \leq |w|,$$

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 length is non-elementary in n and whose Zimin index is $n-1$, allowing us to show that $f(2n-1, n) >$ 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 range 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" [21]. We refer to such counters as "higher-order counters" in the following.

We define the (unary) *tower function* $\boldsymbol{\tau} : \mathbb{N} \to \mathbb{N}$ as

$$\begin{aligned}\boldsymbol{\tau}(0) &= 1 && \text{and} \\ \boldsymbol{\tau}(n+1) &= 2^{\boldsymbol{\tau}(n)} && \text{for all } n \geq 0.\end{aligned}$$

Equivalently, $\boldsymbol{\tau}(n) = $ Tower$(n, 2)$ for all $n \in \mathbb{N}$. 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 the 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 $[\![\, i\,]\!]_n$. The definition proceeds by induction on n. For $n = 1$, there are only two counters $[\![\, 0\,]\!]_1$ and $[\![\, 1\,]\!]_1$ (recall that $\boldsymbol{\tau}(1) = 2$). We define

$$[\![\, 0\,]\!]_1 = 0_1 \text{ and } [\![\, 1\,]\!]_1 = 1_1.$$

For $n \geq 1$ and $i \in [0, \boldsymbol{\tau}(n+1) - 1]$ we define

$$[\![\, i\,]\!]_{n+1} = [\![\, 0\,]\!]_n b_0 [\![\, 1\,]\!]_n b_1 \cdots [\![\, \boldsymbol{\tau}(n) - 1\,]\!]_n b_{\boldsymbol{\tau}(n)-1},$$

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 instance, there are $\boldsymbol{\tau}(2) = 4$ counters of order 2, namely

$$\begin{aligned}[\![\, 0\,]\!]_2 &= 0_1 \mathbf{0_2} 1_1 \mathbf{0_2}, & [\![\, 1\,]\!]_2 &= 0_1 \mathbf{1_2} 1_1 \mathbf{0_2}, \\ [\![\, 2\,]\!]_2 &= 0_1 \mathbf{0_2} 1_1 \mathbf{1_2}, & [\![\, 3\,]\!]_2 &= 0_1 \mathbf{1_2} 1_1 \mathbf{1_2}.\end{aligned}$$

For $[\![\, 11\,]\!]_3$, we have $11 = 1 \cdot 2^0 + 1 \cdot 2^1 + 0 \cdot 2^2 + 1 \cdot 2^3$ and hence

$$[\![\, 11\,]\!]_3 = \underbrace{0_1 0_2 1_1 0_2}_{[\![\, 0\,]\!]_2} \mathbf{1_3} \underbrace{0_1 1_2 1_1 0_2}_{[\![\, 1\,]\!]_2} \mathbf{1_3} \underbrace{0_1 0_2 1_1 1_2}_{[\![\, 2\,]\!]_2} \mathbf{0_3} \underbrace{0_1 1_2 1_1 1_2}_{[\![\, 3\,]\!]_2} \mathbf{1_3}.$$

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

Lemma 2 *Let $n \geq 1$.*

1. *There are $\boldsymbol{\tau}(n)$ counters of order n.*
2. *For all $i \neq j \in [0, \boldsymbol{\tau}(n) - 1]$ we have $[\![\,i\,]\!]_n \neq [\![\,j\,]\!]_n$.*
3. *If $n > 1$, then for all $i \in [0, \boldsymbol{\tau}(n) - 1]$ and $j \in [0, \boldsymbol{\tau}(n - 1) - 1]$ the counter $[\![\,j\,]\!]_{n-1}$ has exactly one occurence in $[\![\,i\,]\!]_n$.*

The following lemma expresses that the order of a symbol in a counter of order n only depends on the distance of this symbol to an order n symbol. It is proven by induction on n by making use of the previous lemma.

Lemma 3 *Let $n \geq 2$, $i \in [0, \boldsymbol{\tau}(n) - 1]$ and p, p' and ℓ such that $p, p', p + \ell$ and $p' + \ell$ belong to $[0, |[\![\,i\,]\!]_n| - 1]$. If the symbols occurring at $p + \ell$ and $p' + \ell$ in $[\![\,i\,]\!]_n$ are of order n then the symbols occurring at p and p' in $[\![\,i\,]\!]_n$ have the same order.*

The length of an order-n counter, denoted by L_n is inductively defined as follows:

$$\begin{aligned} L_1 &= 1 \\ L_{n+1} &= \boldsymbol{\tau}(n) \cdot (L_n + 1) \qquad \text{for all } n \geq 1 \end{aligned}$$

Note that in particular for all $n \geq 1$ we have $L_n \geq \boldsymbol{\tau}(n - 1)$.

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 upper-bounded by the index of $[\![\,0\,]\!]_n$.

Lemma 4 *For all $n \geq 1$ and for all $i \in [0, \boldsymbol{\tau}(n) - 1]$,*

$$\mathrm{Zimin}([\![\,i\,]\!]_n) \leq \mathrm{Zimin}([\![\,0\,]\!]_n).$$

Proof Let $n \geq 1$ and let $i \in [0, \boldsymbol{\tau}(n) - 1]$. By definition of higher-order counters, we have

$$[\![\,0\,]\!]_n = \psi([\![\,i\,]\!]_n). \tag{1}$$

where ψ is the alphabetic morphism defined by $\psi(0_n) = \psi(1_n) = 0_n$ and $\psi(x) = x$ for all $x \in \Sigma_{n-1}$. Assume that $[\![\,i\,]\!]_n$ contains an infix of the form $\varphi(Z_\ell)$ for some non-erasing morphism φ and $\ell \geq 0$. By Eq. 1, $[\![\,0\,]\!]_n$ contains $\psi(\varphi(Z_\ell))$ as an infix. It follows that $\mathrm{Zimin}([\![\,0\,]\!]) \geq \mathrm{Zimin}([\![\,i\,]\!]_n)$.

This leads us to the main result of this section.

Theorem 4 *For all $n \geq 3$,*

$$\mathrm{Zimin}([\![\,0\,]\!]_n) \leq n - 1.$$

Proof The proof proceeds by induction on $n \geq 3$. For the base case, we have to show that

$$[\![0]\!]_3 = 0_1 0_2 1_1 0_2 \mathbf{0_3} 0_1 1_2 1_1 0_2 \mathbf{0_3} 0_1 0_2 1_1 1_2 \mathbf{0_3} 0_1 1_2 1_1 1_2 \mathbf{0_3}$$

has a Zimin index of at most 2. Assume towards a contradiction that $[\![0]\!]_3$ has Zimin index at least 3. Then it must contain an infix of the form $\alpha \beta \alpha$ for some non-empty α and β with α of Zimin type at least 2. In particular, α must be of length at least 3. A careful inspection shows that the only infixes of $[\![0]\!]_3$ of length at least 3 that appear twice are the following:

$$0_1 0_2 1_1 \quad 1_1 0_2 \mathbf{0_3} \quad 1_1 0_2 \mathbf{0_3} 0_1 \quad 0_2 \mathbf{0_3} 0_1$$
$$\mathbf{0_3} 0_1 1_2 \quad \mathbf{0_3} 0_1 1_2 1_1 \quad 0_1 1_2 1_1 \quad 1_1 1_2 \mathbf{0_3}$$

All these words have Zimin type 1 which concludes the base case.

Assume that the property holds for some $n \geq 3$. By Lemma 4 and induction hypothesis, we have that for all $i \in [0, \boldsymbol{\tau}(n) - 1]$,

$$\mathrm{Zimin}([\![i]\!]_n) \leq n - 1. \tag{2}$$

Let us show that $\mathrm{Zimin}([\![0]\!]_{n+1}) \leq n$. Let $\alpha \beta \alpha$ be an infix of $[\![0]\!]_{n+1}$ 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: α contains no occurrence of 0_{n+1}. Then α is an infix of some $[\![i]\!]_n$ for some $i \in [0, \boldsymbol{\tau}(n) - 1]$. By induction hypothesis (i.e. Eq. 2) and Lemma 1, $\mathrm{ZType}(\alpha) \leq \mathrm{Zimin}(\alpha) \leq \mathrm{Zimin}([\![i]\!]_n) \leq n - 1$.

Case 2: α contains at least two occurrences of 0_{n+1}. By definition of counters, α has an infix $0_{n+1} [\![i]\!]_n 0_{n+1}$ for some $i \in [0, \boldsymbol{\tau}(n) - 1]$. Hence $[\![0]\!]_{n+1}$ would contain two occurrences of $0_{n+1} [\![i]\!]_n 0_{n+1}$, which contradicts Lemma 2(3).

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

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

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

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

$$\ell_0 + \ell_1 < \boldsymbol{\tau}(n-1). \tag{3}$$

By definition of counters, $b_{\boldsymbol{\tau}(n-1)-1}$ is the most significant bit of the binary presentation (of length $\boldsymbol{\tau}(n-1)$) of i and j and c_0 is the least significant bit

of the binary presentation of both $i + 1$ and $j + 1$. More formally, there exist $x_i, x_j \in [0, 2^{\boldsymbol{\tau}(n-1)-\ell_0} - 1]$ and $y_i, y_j \in [0, 2^{\boldsymbol{\tau}(n-1)-\ell_1} - 1]$ such that:

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

with

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

Assume towards a 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,

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

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.

Having just shown $\ell_0 + \ell_1 < \boldsymbol{\tau}(n-1)$, 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:

$$
[\![i_0]\!]_n = vru \quad \text{for some non-empty } r. \tag{4}
$$

We claim that $\mathrm{ZType}(\alpha) \leq \mathrm{Zimin}([\![i_0]\!]_n)$ by which we would be done since then $\mathrm{ZType}(\alpha) \leq \mathrm{Zimin}([\![i_0]\!]_n) \leq n-1$ by Lemma 4 and induction hypothesis.

Assume that $\alpha = \gamma \delta \gamma$ for non-empty γ and δ. Using Fact 1, it is enough to show that $\mathrm{ZType}(\gamma) + 1 \leq \mathrm{Zimin}([\![i_0]\!]_n)$. Recall that $\alpha = u 0_{n+1} v$ 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 using Eq. 4, $[\![i_0]\!]_n$ contains $\gamma r \gamma$ as an infix.

$$
\begin{aligned}
\mathrm{ZType}(\gamma) + 1 &\leq \mathrm{ZType}(\gamma r \gamma) \\
&\leq \mathrm{Zimin}([\![i_0]\!]_n)
\end{aligned}
$$

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

Theorem 5

$$
\begin{aligned}
&\mathrm{Zimin}([\![0]\!]_1) = \mathrm{Zimin}([\![0]\!]_1) = 1 \\
&\mathrm{Zimin}([\![0]\!]_2) = \mathrm{Zimin}([\![3]\!]_2) = 2 \\
&\mathrm{Zimin}([\![1]\!]_2) = \mathrm{Zimin}([\![2]\!]_2) = 1 \\
&\mathrm{Zimin}([\![i]\!]_3) = 2 && \textit{for all } i \in [0, \tau(3) - 1] \\
&\mathrm{Zimin}([\![i]\!]_n) = n - 1 && \textit{for all } n \geq 4 \textit{ and all } i \in [0, \boldsymbol{\tau}(n) - 1]
\end{aligned}
$$

Proof The Zimin index of all higher-order counters of order at most 3 is checked by a computer program. For the last statement, Theorem 4 established that $\text{Zimin}(\llbracket i \rrbracket_n) \leq n-1$ for all $n \geq 3$ and all $i \in [0, \boldsymbol{\tau}(n)-1]$. We need to prove $\text{Zimin}(\llbracket i \rrbracket_n) \geq n-1$ for all $n \geq 4$. That is, we only need to show that for all $n \geq 4$, $\llbracket i \rrbracket_n$ contains an infix of Zimin type at least $n-1$. We prove the stronger property that for all $n \geq 3$, there exists a word $\alpha_n \in \Sigma^*$ of Zimin type at least $n-1$, which is an infix of both $\llbracket 2 \rrbracket_n$ and $\llbracket 3 \rrbracket_n$ (recall that in particular when $n \geq 3$, every higher-order counter $\llbracket i \rrbracket_{n+1}$ contains both $\llbracket 2 \rrbracket_n$ and $\llbracket 3 \rrbracket_n$ as infix).

We proceed by induction on $n \geq 3$.

For the base case $n = 3$, we take $\alpha_3 = 1_1 1_2 1_1$. Clearly $\text{ZType}(\alpha_3) = 2$ and as α_3 is an infix of $\llbracket 3 \rrbracket_2$, it is also an infix of $\llbracket 2 \rrbracket_3$ and $\llbracket 3 \rrbracket_3$ (and in fact of all order 3 counters).

Assume that the property holds for some $n \geq 3$ and let us show that it holds for $n + 1$. From the definitions, we have:

$$\llbracket 2 \rrbracket_{n+1} = \llbracket 0 \rrbracket_n \mathbf{0_{n+1}} \llbracket 1 \rrbracket_n \mathbf{1_{n+1}} \llbracket 2 \rrbracket_n \mathbf{0_{n+1}} \llbracket 3 \rrbracket_n \cdots$$
$$\llbracket 3 \rrbracket_{n+1} = \llbracket 0 \rrbracket_n \mathbf{1_{n+1}} \llbracket 1 \rrbracket_n \mathbf{1_{n+1}} \llbracket 2 \rrbracket_n \mathbf{0_{n+1}} \llbracket 3 \rrbracket_n \cdots$$

In particular, $\llbracket 2 \rrbracket_n 0_{n+1} \llbracket 3 \rrbracket_n$ is an infix of both $\llbracket 2 \rrbracket_{n+1}$ and $\llbracket 3 \rrbracket_{n+1}$. By induction hypothesis, α_n is an infix of $\llbracket 2 \rrbracket_n$ and of $\llbracket 3 \rrbracket_n$. Therefore there exist x_2, y_2, x_3 and y_3 such that:

$$\llbracket 2 \rrbracket_n = x_2 \alpha_n y_2$$
$$\llbracket 3 \rrbracket_n = x_3 \alpha_n y_3$$

In particular, the common infix can be written as:

$$\llbracket 2 \rrbracket_n 0_{n+1} \llbracket 3 \rrbracket_n = x_2 \alpha_n y_2 0_{n+1} x_3 \alpha_n y_3$$

We can take $\alpha_{n+1} = \alpha_n y_2 0_{n+1} x_3 \alpha_n$ which is therefore an infix of both $\llbracket 2 \rrbracket_{n+1}$ and $\llbracket 3 \rrbracket_{n+1}$. The Zimin type of $\alpha_{n+1} = \alpha_n \beta \alpha_n$ with $\beta = y_2 0_{n+1} x_3$ is at least 1 more than that of α_n. By induction hypothesis, the Zimin type of α_{n+1} is therefore at least n which concludes the proof.

Corollary 1 *For all $n \geq 3$,*

$$f(n, 2n - 1) > \boldsymbol{\tau}(n - 1).$$

Proof Let $n \geq 3$. The word $\llbracket 0 \rrbracket_n$ over the alphabet $\{0_1, 1_1, \ldots, 0_{n-1}, 1_{n-1}, 0_n\}$ of size $2n-1$ has length at least $L_n \geq \boldsymbol{\tau}(n-1)$ and Zimin index at most $n-1$. This word hence avoids the n-th Zimin pattern and witnesses the announced lower bound.

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 upper bound on the Zimin index. For this we apply to counters the morphism ψ, defined as follows

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

Definition 4 For all $n \geq 1$ and $i \in [0, \boldsymbol{\tau}(n) - 1]$, we define

$$\{\!\{\, i \,\}\!\}_n = \psi(\llbracket\, i \,\rrbracket_n).$$

The set of images of the letters 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 8),
- the encoding of 0_n and 1_n differ on their first and last symbol inter alia,
- the Zimin index of the encoding of an order n symbol is relatively low (we will show that it is at most $\lfloor \log_2(3 + 2n) \rfloor$).

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, remark that the proof we present is tightly linked to the above-mentioned properties of ψ that are decisive for the construction to work.

This section is devoted to establishing the following theorem.

Theorem 6 *For all $n \geq 2$ and for all $i \in [0, \boldsymbol{\tau}(n) - 1]$,*

$$\mathrm{Zimin}(\{\!\{\, i \,\}\!\}_n) \leq n + 1.$$

The proof, which is essentially an extension of the proof of Theorem 4, 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.

Recalling that an order n counter has length at least $\boldsymbol{\tau}(n-1)$, in particular so does its code, which is its image under the non-erasing morphism ψ.

Theorem 6 immediately implies the following non-elementary lower bound for $f(n, 2)$ whenever $n \geq 4$.

Theorem 7 *For all $n \geq 4$,*

$$f(n, 2) > \boldsymbol{\tau}(n - 3).$$

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 4, 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 of letters in Σ,
- $L = \{v \in \{0, 1\}^* \mid \exists u \in \{0, 1\}^+, uv \in C\}$ 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.

Remark that these four sets are all regular languages. Indeed,

- $C = 00(01)^*00 \cup 11(01)^*11$,
- $L = \varepsilon + 0 + 1 + (\varepsilon + 1)(01)^*11 + (\varepsilon + 0 + 1)(01)^*00$,
- $R = \varepsilon + 0 + 1 + 11(01)^*(\varepsilon + 0 + 1) + 00(01)^*(\varepsilon + 0)$, and
- $F = (\varepsilon + 0)(01)^*(\varepsilon + 0) + (\varepsilon + 1)(01)^*(\varepsilon + 0 + 1)$.

Thanks to this property, it is possible to reduce the proofs of our statements (namely Lemma 5 and Lemma 8) to computations on finite word automata and finite word transducers. For our tests, we used Awali[3] which is the latest version of the well-established Vaucanson platform.

We collect in the following lemma some observations on these sets which will be used through out the proofs in this section.

Lemma 5 *The sets C, R, L and F satisfy the following equations:*

1. $LC^*R \cap F = LR \cap F = \{0, 1\}^{\leq 2} \cup \{001, 110\}$,
2. $LR \cap C = \{0000, 1111\}$,

Lemma 6 *For every $n \geq 1$ and every infix of a word in $C_{\leq n}^*$ belongs to*

$$F_{\leq n} \cup L_{\leq n} C_{\leq n}^* R_{\leq n}$$

where $F_{\leq n}$, $C_{\leq n}$ and $R_{\leq n}$ respectively denote the restrictions of F, C and R to symbols of order at most n.

Proof Let us first remark that $\varepsilon \in L_{\leq n}$ and $\varepsilon \in R_{\leq n}$. We show, by induction on m, that every infix α of a word in $C_{\leq n}^m$ belongs to $F_{\leq n} \cup L_{\leq n} C_{\leq n}^* R_{\leq n}$. For $m = 0$, the property is immediate as $\varepsilon \in F_{\leq n}$ (and to $L_{\leq n} C_{\leq n}^* R_{\leq n}$). For $m = 1$, if α is an infix of $c \in C_{\leq n}$, then $c = x\alpha y$ for some words x and y. If $x \neq \varepsilon$ and $y \neq \varepsilon$ then α belongs to $F_{\leq n}$. If $x \neq \varepsilon$ and $y = \varepsilon$ then α belongs $L_{\leq n}$. If $x = \varepsilon$ and $y \neq \varepsilon$ then α belongs to $R_{\leq n}$. Finally, if $x = y = \varepsilon$, then

---

[3] http://vaucanson-project.org/AWALI/Awali.html

$\alpha = c \in C_{\leq n} \subseteq L_{\leq n}C_{\leq n}^*R_{\leq n}$. For the induction step, let α be an infix of some $c_1 \cdots c_{m+1} \in C_{\leq n}^{m+1}$, where $m \geq 1$. There are the following cases: either α is an infix of $c_1 \cdots c_m$ and we can conclude using the induction hypothesis, or α is an infix of c_{m+1} and we can conclude using the case $m = 1$. Finally it remains the case where $\alpha = xy$ with x a suffix of $c_1 \cdots c_m$ and y a prefix of c_{m+1}. Clearly x belongs to $L_{\leq n}C_{\leq n}^*$ and y belongs to $R_{\leq n} \cup C_{\leq n}$. This implies that α belongs to $L_{\leq n}C_{\leq n}^*R_{\leq n}$.

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* 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 $\psi(0_10_10_1)$. It can be parsed as

$$(\varepsilon, 0_10_1, 00) \, , \, (0, 0_10_1, 0) \text{ and as } (00, 0_10_1, \varepsilon).$$

However, we will provide sufficient conditions on an infix to admit a unique parse.

Definition 5 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.

On the one hand, the term "simple" is justified in the context of this proof as simple infixes of $\{\!\{ i \}\!\}_n$ can rather easily shown to have Zimin index at most $n - 1$ for all $n \geq 4$ and all $i \in [0, \boldsymbol{\tau}(n) - 1]$ (cf. Lemma 7).

Lemma 7 *For all $n \geq 4$ and all $i \in [0, \boldsymbol{\tau}(n) - 1]$ every simple infix of $\{\!\{ i \}\!\}_n$ has Zimin index at most $n - 1$.*

Proof Let $n \geq 4$, let $i \in [0, \boldsymbol{\tau}(n) - 1]$ and let α be a simple infix of $\{\!\{ i \}\!\}_n$. It is easy to check that $[\![i]\!]_n$ does not contain two consecutive occurrences of 0_1 or of 1_1. It follows that $\{\!\{ i \}\!\}_n$ cannot contain more than 9 consecutive zeros or 9 consecutive ones. Hence α is simple either because it belongs to F or because $|\alpha| < 11$.

First consider the case where $|\alpha| < 11$. We know, by Lemma 1, that $\mathrm{Zimin}(\alpha) \leq \lfloor \log_2(11) \rfloor = 3 \leq n - 1$.

Now consider the case where α belongs to F. By Lemma 6, we have that α either belongs to $F \cap F_{\leq n} = F_{\leq n}$ or to $F \cap L_{\leq n}C_{\leq m}^*R_{\leq n}$.

In the first case, $|\alpha| \leq 2n + 2$ and thus by Lemma 1, $\mathrm{Zimin}(\alpha) \leq \lfloor \log_2(2n + 2) \rfloor \leq n - 1$ since $n \geq 4$.

In the second case α belongs to $F \cap L_{\leq n}C_{\leq m}^*R_{\leq n} \subseteq LC^*R \cap F$. From Lemma 5, we have $LC^*R \cap F = LR \cap F = \{0,1\}^{\leq 2} \cup \{001, 100\}$. Hence in this case, we have $\mathrm{Zimin}(\alpha) \leq 1$.

On the other hand, the term "simple" will be justified by the fact that for non-simple infixes there is exactly one possible parse as shown in the following lemma.

Lemma 8 *Any non-simple infix of a coded word admits a unique parse.*

Proof The existence of the parse is immediate from Lemma 6. We will show that unicity of the parse can be be reduced to testing the functionality of a certain rational relation (i.e., a relation accepted by a word transducer), a property which can be decided in polynomial time [2].

It is natural to represent a parse $p = (\ell, u = u_1 \cdots u_k, r)$ by the word $w_p = \ell \sharp \psi(u_1) \sharp \cdots \psi(u_k) \sharp r \sharp$ over the alphabet $\{0, 1, \sharp\}$. The set of all words representing a parse is the regular set $L\sharp(C\sharp)^* R\sharp$. If α admits a parse p then the morphism π erasing the $\sharp$ symbol (i.e. defined by $\pi(0) = 0$, $\pi(1) = 1$ and $\pi(\sharp) = \varepsilon$) maps a w_p to α.

The relation R_{parse} containing all pairs (α, w_p) such that p is a parse of α is rational. Indeed R_1 is the restriction of the rational relation π^{-1} to regular image $L\sharp(C\sharp)^* R\sharp$. As rational relations are closed under restriction to a regular image (or domain), it follows that R_{parse} is a rational relation.

To check the unicity of the parse, it is enough to show the R_{parse} is functional when its domain is restricted to non-simple words. By Definition 5, the set of simple words is a regular set and hence so is the set of non-simple words. As rational relation are closed under restriction to a regular domain, we can construct a transducer for R_{parse} restricted to non-simple words and check its functionality with an automata framework.

A direct but rather tedious proof of this statement can be found in [4]. $\qquad\square$

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

The next lemma states that occurrences of codings of symbols of order strictly larger than one in a coded word can be related with occurrences of this symbol in the word that has been coded.

Lemma 9 *Let α be a non-simple infix of some coded word and let (ℓ, u, r) be its parse.*

If α contains $n > 1$ occurrences of $\psi(x)$ for some $x \in \Sigma \setminus \Sigma_1$ then u contains n occurrences of x.

Proof Let α a non-simple infix of some coded word and let $p = (\ell, u, r)$ be its parse. Let x be a letter in $\Sigma \setminus \Sigma_1$ such that α contains $n > 1$ occurrences of $\psi(x)$.

By definition of the parse p, we have $\alpha = \ell \psi(u) r$. If we write $u = u_1 \cdots u_m$ with $m \geq 0$ and $u_i \in \Sigma$ for all $i \in [1, m]$, we can write α as follows,

$$\alpha = \alpha_0 \alpha_1 \cdots \alpha_m \alpha_{m+1},$$

where

- $\alpha_0 = \ell$,
- $\alpha_i = \psi(u_i)$ for all $i \in [m]$,
- and $\alpha_{m+1} = r$.

Let $m_1 < \cdots < m_n$ be an enumeration of the n occurrences of $\psi(x)$ in α. For all $i \in [n]$, we denote by q_i the maximal integer satisfying $m_i \geq \sum_{j=0}^{q_i-1} |\alpha_j|$.

Claim. For all $i \in [n]$ we have $0 < q_i \leq m$ and $m_i = \sum_{j=0}^{q_i-1} |\alpha_j|$.

Proof of the claim. Let $i \in [1, n]$.

Let us first show that $q_i \neq 0$. Assume towards a contradiction that $q_i = 0$. By maximality of q_i, $\alpha_0 = \ell$ cannot be empty. This implies that $\psi(x)$ can be written as $\ell'\alpha_1 \cdots \alpha_k r'$ with ℓ' a non-empty suffix of ℓ, $k \geq 0$ and r' a prefix of α_{k+1}. As C is an infix code, k is necessarily equal to 0. Hence $\psi(x) = \ell'r'$. In particular $\psi(x) \in LR \cap C$. In Lemma 5, we remarked that $LR \cap C = \{0000, 1111\}$, which brings a contradiction with the fact that $x \in \Sigma \setminus \Sigma_1$.

Let us next show $q_i \neq m + 1$. Assume towards a contradiction that $q_i = m + 1$. In this case, $r \in R$ would contain $\psi(x)$ as an infix which contradicts the fact that C is an infix code.

Let us finally show $m_i \leq \sum_{j=0}^{q_i-1} |\alpha_j|$ (and thus $m_i = \sum_{j=0}^{q_i-1} |\alpha_j|$). Assume towards a contradiction that $m_i > \sum_{j=0}^{q_i-1} |\alpha_j|$. By definition of q_i, this implies that $\psi(x)$ can be written as $\ell'\alpha_{q_i+1} \cdots \alpha_{q_i+k} r'$ with ℓ' a non-empty suffix of α_{q_i}, $k \geq 0$ and r' a prefix of α_{q_i+k+1}. As C is an infix code, k is necessarily equal to 0. Hence $\psi(x) = \ell'r'$. In particular $\psi(x) \in LR \cap C$. As above, we have from Lemma 5, that $LR \cap C = \{0000, 1111\}$, which brings a contradiction with the fact that x is not of order 1.

End of the proof of the claim.

Using the claim, it follows that either $\psi(x)$ is a prefix of $\psi(u_{q_i})$ or conversely that $\psi(u_{q_i})$ is a prefix of $\psi(x)$. As C is an infix code, this is only possible if $\psi(x) = \psi(u_{q_i})$ and hence $u_{q_i} = x$.

Again using the claim, we have that $q_1 < \cdots < q_n$ (as $m_1 < \cdots < m_n$). Hence we have shown that u contains at least n occurrences of x. Clearly, u cannot contain more than n occurrences of x as each occurrence of x in u induces an occurrence of $\psi(x)$ in α.

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 *occurence 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 a suffix of $\psi(w_{m-1})$ and similarly whenever r is non-empty we have $m + |u| < |w|$ and r is a prefix of $\psi(w_{m+|u|})$.

Remark 1 In the previous lemma, the requirement that the order of the symbol is strictly greater than 1 is necessary. For instance consider the coded word $w = \psi(0_2 0_3) = 00010000010100$. If we take α to be w which is non-simple, α contains $\psi(0_1) = 0000$ as an infix but 0_1 does not occur in its unique parse of $(\varepsilon, 0_2 0_3, \varepsilon)$.

The next lemma shows that 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 $\psi(w)$.

Lemma 10 *For any word $w \in \Sigma^*$ and any non-simple infix α of $\psi(w)$, there is a unique order-preserving bijection between the occurrences of α in $\psi(w)$ and the occurrences of its parse p in w.*

Proof Let $w = w_0 w_1 \cdots w_{n-1}$, $n \geq 1$ a non-empty word over Σ and let α be a non-simple infix of the word $\psi(w)$. Consider the unique parse $p = (\ell, u, r)$ of the infix α. To each occurrence m of the parse p in w, we associate the occurrence $\rho(m) = (\sum_{i=0}^{m-1} |\psi(w_i)|) - |\ell|$ of the word α in $\psi(w)$. The mapping ρ, from the set of occurrences of p in w to the set of occurrences of α in $\psi(w)$, is order-preserving and injective. It remains to show that it is surjective.

Let h be an occurrence of α in $\psi(w)$. By definition, there exist two words $x, y \in \{0, 1\}^*$ such that $\psi(w) = x\alpha y$ and $|x| = h$. Consider the greatest integer $m_0 \in [0, n-1]$ such that

$$h \geq \sum_{i=0}^{m_0-1} |\psi(w_i)|.$$

We will show that m_0 is an occurrence of the parse p in w and hence that $\tau(m_0) = h$. Remarking that $\psi(w)$ is equal to both $x\alpha y$ and $\psi(w_0) \cdots \psi(w_{n-1})$, there must exist $k \geq 0$ such that $\alpha = \ell' \psi(w_{m_0}) \cdots \psi(w_{m_0+k-1}) r'$, where

- ℓ' is empty if $h = \sum_{i=0}^{m_0-1} |\psi(w_i)|$ and is the suffix of length $h - \sum_{i=0}^{m_0-1} |\psi(w_i)|$ of $\psi(w_{m_0-1})$ otherwise,
- and r' is empty if $m_0 + k - 1 = n - 1$ and a prefix of $\psi(w_{m_0+k})$ otherwise.

It follows that $(\ell', w_{m_0} w_{m_0+1} \cdots w_{m_0+k-1}, r')$ is a parse of α occurring at m_0 in w. The lemma now follows from the unicity of the parse.

Definition 6 For an occurrence m of a parse $p = (\ell, u, r)$ in w, we define its *context* $[p]_m$ 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, that itself contains u as an infix. Moreover, the value α of p is an infix of $\psi(c)$.

4.2 Upper bound on the Zimin index

We are now ready to upper-bound the Zimin index of the code of higher-order counters. Due to the nature of our coding ψ we need to prove a slightly stronger inductive statement that takes into the account the code of a symbol of order $n + 1$ directly before or directly after the code of a counter of order n.

Theorem 8 *For all $n \geq 2$ and for all $i \in [0, \boldsymbol{\tau}(n) - 1]$,*

$$\mathrm{Zimin}(\{\!\!\{\, i \,\}\!\!\}_n \psi(0_{n+1})) \leq n + 1, \qquad \mathrm{Zimin}(\{\!\!\{\, i \,\}\!\!\}_n \psi(1_{n+1})) \leq n + 1,$$
$$\mathrm{Zimin}(\psi(0_{n+1}) \{\!\!\{\, i \,\}\!\!\}_n) \leq n + 1, \qquad \mathrm{Zimin}(\psi(1_{n+1}) \{\!\!\{\, i \,\}\!\!\}_n) \leq n + 1.$$

Proof We proceed by induction on n. For the cases $n = 2$ and $n = 3$, the property is checked using a computer program. Remark that the reason we start the induction at 3 is to be able to apply the upper bound from Lemma 7.

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 have to show that

$$\text{Zimin}(\{\!\!\{\, i \,\}\!\!\}_{n+1}\psi(0_{n+2})) \leq n + 2, \quad \text{Zimin}(\{\!\!\{\, i \,\}\!\!\}_{n+1}\psi(1_{n+2})) \leq n + 2,$$
$$\text{Zimin}(\psi(0_{n+2})\{\!\!\{\, i \,\}\!\!\}_{n+1}) \leq n + 2, \quad \text{Zimin}(\psi(1_{n+2})\{\!\!\{\, i \,\}\!\!\}_{n+1}) \leq n + 2.$$

We start by showing that $\text{Zimin}(\{\!\!\{\, i \,\}\!\!\}_{n+1}) \leq n + 2$. Let $\alpha\beta\alpha$ be an infix of $\{\!\!\{\, i \,\}\!\!\}_{n+1}$ for some non-empty α and β. It is enough to show that $\text{ZType}(\alpha) \leq n+1$. By Lemma 7, we only need to consider the case when α is non-simple. Let $p = (\ell, u, r)$ be the parse of α, whose uniqueness is guaranteed by Lemma 8.

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 $\{\!\!\{\, i \,\}\!\!\}_{n+1}$. By Lemma 10, 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.

If c_1 does not contain any symbol of order $n + 1$. As c_1 is an infix of $[\![\, i \,]\!]_{n+1}$ and since by assumption c_1 does not contain any symbol of order $n + 1$, it must be an infix of some $[\![\, j \,]\!]_n$ with $j \in [0, \boldsymbol{\tau}(n) - 1]$. By definition of the context of a parse, α is an infix of $\psi(c_1)$ and hence of $\{\!\!\{\, j \,\}\!\!\}_n$. Thus,

$$\text{ZType}(\alpha) \leq \text{Zimin}(\{\!\!\{\, j \,\}\!\!\}_n) \leq \text{Zimin}(\{\!\!\{\, j \,\}\!\!\}_n\psi(0_{n+1})) \leq n + 1,$$

where the last inequality follows from induction hypothesis.

If c_1 contains at least two symbols of order $n + 1$. We will show that this situation cannot occur. By definition of $[\![\, i \,]\!]_{n+1}$, c_1 contains an infix of the form $b[\![\, j \,]\!]_n b'$ for some $j \in [0, \boldsymbol{\tau}(n) - 1]$ and $b, b' \in \{0_{n+1}, 1_{n+1}\}$. The center u of the parse $p = (\ell, u, r)$ must therefore contain $[\![\, j \,]\!]_n$ as an infix. As there are two occurrences of u in $[\![\, i \,]\!]_{n+1}$[4], this would imply that $[\![\, j \,]\!]_n$ has two occurrences in $[\![\, i \,]\!]_{n+1}$, which brings the contradiction (using Lemma 2).

If c_1 contains one and only one symbol of order $n + 1$.

As c_1 is an infix of $[\![\, i \,]\!]_{n+1}$ with one order $n + 1$ symbol, there exists $k_0 \in [0, \boldsymbol{\tau}(n) - 2]$ and some $b \in \{0_{n+1}, 1_{n+1}\}$ such that

$$c_1 = xby, \text{ where } x \in \Sigma_n^* \text{ is a suffix of } [\![\, k_0 \,]\!]_n \text{ and } y \in \Sigma_n^* \text{ is a prefix of } [\![\, k_0 + 1 \,]\!]_n.$$

Remark that if x or y are empty, we can conclude using induction hypothesis. Indeed in these cases, c_1 is an infix of either $[\![\, k_0 \,]\!]_n b$ or $b[\![\, k_0 + 1 \,]\!]_n$. Hence α, which is an infix of $\psi(c_1)$, is also an infix of either $\{\!\!\{\, k_0 \,\}\!\!\}_n\psi(b)$ or $\psi(b)\{\!\!\{\, k_0 + 1 \,\}\!\!\}_n$. As by induction hypothesis both have Zimin index at most $n + 1$, we can conclude using Lemma 1 that $\text{ZType}(\alpha) \leq n + 1$.

---

[4] Recall that there are two occurrences of p in $[\![\, i \,]\!]_{n+1}$.

From now on, we assume that both 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 $u = \underline{x} b \underline{y}$. In summary, we have

$$
\begin{aligned}
c_1 &= xby, \\
\alpha &= \ell \psi(\underline{x}) \psi(b) \psi(\underline{y}) r, \\
x &= s\underline{x}, \\
y &= \underline{y} t.
\end{aligned}
$$

for some s and t such that:

− $s = \varepsilon$ if $\ell = \varepsilon$ and otherwise $s \in \Sigma$ with ℓ is a suffix of $\psi(s)$.
− $t = \varepsilon$ if $r = \varepsilon$ and otherwise $t \in \Sigma$ with r is a prefix of $\psi(t)$.

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

Proof of the Claim 1: Similarly as for c_1, the context c_2 can be written as $s' \underline{x} b \underline{y} t'$ for some s' and t' such that:

− $s' = \varepsilon$ if $\ell = \varepsilon$ and otherwise $s' \in \Sigma$ with ℓ is a suffix of $\psi(s')$.
− $t' = \varepsilon$ if $r = \varepsilon$ and otherwise $t' \in \Sigma$ with t is a prefix of $\psi(t')$.

Towards a contradiction, assume c_1 and c_2 are different. It is either the case that $s \neq s'$ or the $t \neq t'$. As both cases can be shown analogously, we only consider the first one and assume that $s \neq s'$. In particular, without loss of generality we may assume that ℓ is non-empty.

The symbols s and s' occur in $[\![i]\!]_{n+1}$ at the same distance of an order $n + 1$ symbol and by Lemma 3 must have the same order. Furthermore the last symbol of their encoding by ψ is the same (it is the last symbol of ℓ). By the definition of ψ, s and s' are either both from $\{0_k \mid k \geq 1\}$ or both from $\{1_k \mid k \geq 1\}$. This proves that s and s' are equal which brings the contradiction.
End of the proof of Claim 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 $\tilde{y} \chi \tilde{x} = \{\!\{ j_0 \}\!\}_n$.

Proof of Claim 2. We proceed along the same lines as in the proof of Theorem 4. Consider the morphism φ that erases all symbols in Σ_{n-1} and replaces 0_n and 1_n by 0 and 1 respectively. That is, we can write $\varphi(x)$ and $\varphi(y)$ as follows,

$$
\begin{aligned}
\varphi(x) &= b_{\boldsymbol{\tau}(n-1)-\ell_0} \cdots b_{\boldsymbol{\tau}(n-1)-1} \\
\varphi(y) &= c_0 \cdots c_{\ell_1 - 1}
\end{aligned}
$$

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

With the same proof as in Theorem 4, we show that

$$\ell_0 + \ell_1 < \boldsymbol{\tau}(n-1). \tag{5}$$

By the same reasoning as in the proof of Theorem 4, there exists $j_0 \in [0, \boldsymbol{\tau}(n) - 1]$ and non-empty ξ such that $[\![\, j_0 \,]\!]_n = y\xi x$. By applying ψ and recalling that $\tilde{x}$ is a suffix of $\psi(x)$ and $\tilde{y}$ is a prefix of $\psi(y)$, we can conclude.

End of the proof of Claim 2.

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 $\text{ZType}(\alpha) \leq n + 1$ or that $\text{ZType}(\delta) \leq n$.

There are several cases to consider depending on how the two decompositions $\tilde{x}\psi(b)\tilde{y}$ and $\delta\gamma\delta$ overlap.

Case 1: $|\tilde{x}\psi(b)| \leq |\delta|$.

$$\alpha = \quad \begin{array}{|c|c|c|}\hline \tilde{x} & \psi(b) & \multicolumn{1}{c}{} \\ \hline \multicolumn{2}{|c|}{} & \\ \hline \end{array}$$

This situation cannot occur under our hypothesis. Indeed, α would contain two occurences of $\psi(b)$ which by Lemma 9 implies that the center of its parse contains two occurences of the order $n+1$ symbol b. This brings a contradiction with the fact that the context c_1 contains exactly one symbol of order $n + 1$.

Case 2: $|\tilde{x}| \leq |\delta|$ **and** $|\delta| < |\tilde{x}\psi(b)| \leq |\delta\gamma|$.

$$\alpha = \quad \begin{array}{|c|c|c|c|}\hline \tilde{x} & \psi(b) & \tilde{y} \\ \hline \end{array}$$

In this case, $\psi(b)$ can be written as $z_1 z_2$ such that $\delta = \tilde{x}z_1$ and γ as $\gamma_1\gamma_2$ such that $\tilde{y} = \gamma_2\delta$ and $z_1 \neq \varepsilon$.

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

By induction hypothesis, $\{\!\!\{\, j_0 \,\}\!\!\}_n\psi(b)$ has Zimin index at most $n + 1$. In particular, $\tilde{y}\chi\tilde{x}z_1$, which is a prefix of $\{\!\!\{\, j_0 \,\}\!\!\}_n\psi(b)$, also has Zimin index at most $n + 1$.

As $\tilde{y}\chi\tilde{x}z_1$ is equal to $\gamma_2\delta\chi\delta$ we have that $\delta\chi\delta$ is an infix of a word (i.e. $\{\!\!\{\, j_0 \,\}\!\!\}_nz_1$) of Zimin index at most $n+1$. By Fact 1, this implies that $\text{ZType}(\delta) \leq n$ which concludes the case.

Case 3: $|\tilde{x}| \leq |\delta|$ **and** $|\tilde{x}\psi(b)| > |\delta\gamma|$.

$$
\begin{array}{c}
\overbrace{\quad z_1 = r \quad}\ \ \overbrace{\qquad \gamma \qquad}\ \ \overbrace{\quad z_2 = \ell \quad} \\[2pt]
\alpha = \begin{array}{|c|c|c|}
\hline
\tilde{x} & \psi(b) & \tilde{y} \\
\hline
\end{array} \\[2pt]
\begin{array}{|c|c|c|}
\hline
\delta & \gamma & \delta \\
\hline
\end{array}
\end{array}
$$

In this case $\psi(b)$ can be written as $\psi(b) = z_1 \gamma z_2$ with z_2 non-empty such that:

- $\delta = \tilde{x} z_1$,
- $\delta = z_2 \tilde{y}$.

First recall that $\alpha = \ell \psi(\underline{x}) \psi(b) \psi(\underline{y}) r$. Next, recall that $\tilde{x} = \ell \psi(\underline{x})$, hence $\delta = \ell \psi(\underline{x}) z_1$. It follows that $(\ell, \underline{x}, z_1)$ is a parse of δ.

Finally, recall that $\tilde{y} = \psi(\underline{y}) r$, hence $\delta = z_2 \psi(\underline{y}) r$. It follows that $(z_2, \underline{y}, r)$ is a parse of δ.

By the unicity of the parse (Lemma 8), we have $(z_2, \underline{y}, r) = (\ell, \underline{x}, z_1)$ and hence $z_2 = \ell$, $z_1 = r$ and $\underline{x} = \underline{y}$.

We will now show that $\underline{x} = \underline{y}$ is empty.

Towards a contradiction, assume that $\underline{x}$ is not empty. We recall 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 $[\![\, k_0 + 1\,]\!]_n$.

Since $x = s\underline{x}$ it follows that $\underline{x}$ is a suffix of $[\![\, k_0\,]\!]_n$. By definition of $[\![\, k_0\,]\!]_n$ we have that $\underline{x}$ ends with an order n symbol. But $\underline{x} = \underline{y}$ is a also prefix of $[\![\, k_0 + 1\,]\!]_n$ (which contains an order n symbol) and hence starts with $[\![\, 0\,]\!]_{n-1}$. By Lemma 2, a suffix of $[\![\, k_0\,]\!]_n$ starting with $[\![\, 0\,]\!]_{n-1}$ is equal to $[\![\, k_0\,]\!]_n$. Hence $\tilde{x} = [\![\, k_0\,]\!]_n$ which is not a prefix of $[\![\, k_0 + 1\,]\!]_n$, which brings the contradiction.

Hence we have $\delta = \ell r = z_2 z_1$ and in particular $|\delta| < |\psi(b)| = 4 + 2n$. By Lemma 1, we can bound the Zimin index of δ by

$$
\mathrm{Zimin}(\delta) \le \lfloor \log_2(2n + 4) \rfloor.
$$

As for all $n \ge 3$, $\lfloor \log_2(2n + 4) \rfloor \le n$, we have shown that $\mathrm{ZType}(\delta) \le \mathrm{Zimin}(\delta) \le n$, which concludes this case.

Case 4: $|\delta| < |\tilde{x}| \le |\delta\gamma|$ **and** $|\tilde{x}\psi(b)| \le |\delta\gamma|$**.**

$$
\begin{array}{c}
\qquad\qquad\qquad \overbrace{\ z_1\ }\qquad\ \overbrace{\ z_2\ } \\[2pt]
\alpha = \begin{array}{|c|c|c|}
\hline
\tilde{x} & \psi(b) & \tilde{y} \\
\hline
\end{array} \\[2pt]
\begin{array}{|c|c|c|}
\hline
\delta & \gamma & \delta \\
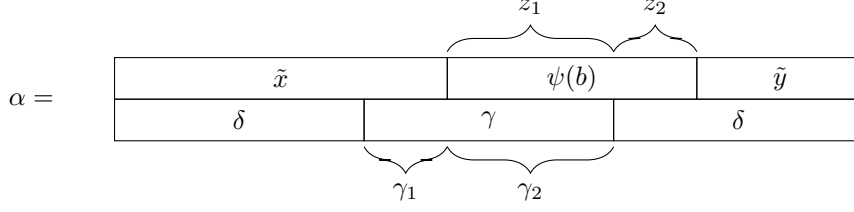\hline
\end{array}
\end{array}
$$

In this case $\gamma = z_1 \psi(b) z_2$ with $z_1 \ne \varepsilon$ such that $\tilde{x} = \delta z_1$ and $\tilde{y} = z_2 \delta$.

By Claim 2, there exists $j_0 \in [0, \boldsymbol{\tau}(n) - 1]$ and a non-empty χ such that $\tilde{y} \chi \tilde{x} = \{\!\!\{\, j_0\, \}\!\!\}_n$. We have $\mathrm{Zimin}(\{\!\!\{\, j_0\, \}\!\!\}_n) \le \mathrm{Zimin}(\{\!\!\{\, j_0\, \}\!\!\}_n \psi(0_{n+1})) \le n + 1$, where the last inequality follows from 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.

**Case 5:** $|\delta| < |\tilde{x}| \le |\delta\gamma|$ **and** $|\tilde{x}\psi(b)| > |\delta\gamma|$.

This case is the symmetric to Case 2. One can write $\psi(b)$ as $z_1 z_2$ such that $\delta = z_2\tilde{y}$ and γ as $\gamma_1\gamma_2$ such that $\tilde{x} = \delta\gamma_1$, where $\gamma_1 \neq \varepsilon$ and $z_2 \neq \varepsilon$.

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

By induction hypothesis, $\psi(b)\{\!\!\{\, j_0 \,\}\!\!\}_n$ has Zimin index at most $n + 1$. In particular, $z_2\tilde{y}\chi\tilde{x}$, which is a suffix, also has Zimin index at most $n + 1$.

As $z_2\tilde{y}\chi\tilde{x}$ is equal to $\delta\chi\delta\gamma_1$ we have that $\delta\chi\delta$ is an infix of the word $z_2\tilde{y}\chi\tilde{x} = z_2\{\!\!\{\, j_0 \,\}\!\!\}_n$ of Zimin index at most $n + 1$. By Fact 1, this implies that $\mathrm{ZType}(\delta) \le n$ which concludes the case.

Case 6: $|\tilde{x}| > |\delta\gamma|$
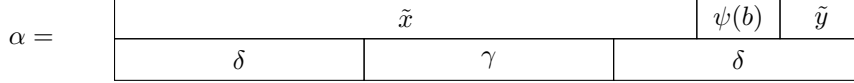
This situation cannot occur under our hypothesis. Indeed α would contain two occurrences of $\psi(b)$ which by Lemma 9 implies that the center of its parse contains two occurrences of the order $n + 1$ symbol b. This brings a contradiction to the fact that the context c_1 contains exactly one symbol of order $n + 1$.

We have shown that for all $i \in [0, \tau(n + 1) - 1]$ we have

$$\mathrm{Zimin}(\{\!\!\{\, i \,\}\!\!\}_{n+1}) \le n + 2.$$

Let us now show that for all $b \in \{0_{n+2}, 1_{n+2}\}$, we have

- $\mathrm{Zimin}(\psi(b)\{\!\!\{\, i \,\}\!\!\}_{n+1}) \le n + 2$ and
- $\mathrm{Zimin}(\{\!\!\{\, i \,\}\!\!\}_{n+1}\psi(b)) \le n + 2$.

We first consider the case of $\psi(b)\{\!\!\{\, i \,\}\!\!\}_{n+1}$. Let $\alpha\beta\alpha$ be an infix of $\psi(b)\{\!\!\{\, i \,\}\!\!\}_{n+1}$ for some non-empty α and β. By Fact 1, it is enough to show that $\mathrm{ZType}(\alpha) \le n + 1$. By Lemma 7, it is enough to consider the case when α is non-simple and hence by Lemma 8, α admits a unique parse $p = (\ell, u, r)$.

As $\alpha\beta\alpha$ is an infix of $\psi(b)\{\!\!\{\, i \,\}\!\!\}_{n+1}$, there exists z_1 and z_2 such that

$$\psi(b)\{\!\!\{\, i \,\}\!\!\}_{n+1} = z_1\alpha\beta\alpha z_2.$$

We distinguish different possible lengths of z_1.

Case 6A: $|z_1| \ge |\psi(b)|$. In this case, $\alpha\beta\alpha$ is an infix of $\{\!\!\{\, i \,\}\!\!\}_{n+1}$. We have already shown that $\mathrm{Zimin}(\{\!\!\{\, i \,\}\!\!\}_{n+1}) \le n + 2$ and thus $\mathrm{ZType}(\alpha) \le n + 1$.

Case 6B: $|z_1| = 0$. We will show that $|\alpha| < |\psi(b)|$ and hence by Lemma 1 we have $\text{Zimin}(\alpha) < |\psi(b)| = \lfloor \log_2(2n+6) \rfloor \leq n+1$ as $n \geq 3$.

Assume towards a contradiction that $|\alpha| \geq |\psi(b)|$. Hence $\psi(b)$ is a prefix of α. Therefore, $\psi(b)\{\!\{\, i \,\}\!\}_{n+1}$ would contain two occurrences of $\psi(b)$. By Lemma 9, $b[\![\, i \,]\!]_{n+1}$ would contain two occurrences of the order $n+2$ symbol b, which brings the contradiction.

Case 6C: $1 \leq |z_1| < |\psi(b)|$. We will show that α is an infix of $\psi(b')\{\!\{\, 0 \,\}\!\}_n$ for some $b' \in \{0_{n+1}, 1_{n+1}\}$. Note that this will be sufficient since then we can apply induction hypothesis to conclude that $\text{ZType}(\alpha) \leq n+1$.

As $1 \leq |z_1| < |\psi(b)|$, the parse $p = (\ell, u, r)$ is such that ℓ is a non-empty suffix of $\psi(b)$ and u is a prefix of $[\![\, i \,]\!]_{n+1}$. Let us first show that $[\![\, 0 \,]\!]_n$ is not a prefix of u. Assume towards a contradiction that $[\![\, 0 \,]\!]_n$, which is a prefix of $[\![\, i \,]\!]_{n+1}$, is also a prefix of u. By Lemma 10, this would imply that $[\![\, i \,]\!]_{n+1}$ contains two occurrences of $[\![\, 0 \,]\!]_n$, which brings the contradiction. Thus, u is not a prefix of $[\![\, 0 \,]\!]_n$ and hence $\psi(u)r$ is a prefix of $\{\!\{\, 0 \,\}\!\}_n$.

It remains to show that ℓ is a suffix of $\psi(b')$ for some $b' \in \{0_{n+1}, 1_{n+1}\}$. As there are two occurrences of the parse p in $\psi(b)\{\!\{\, i \,\}\!\}_{n+1}$, this implies that ℓ is the suffix of $\psi(b)$ and some $\psi(b'')$ for some symbol b'' of order $k \leq n+1$. From the definition of ψ, it follows that ℓ is a suffix of $(01)^{k-1}00$ or $(01)^{k-1}11$. Hence as announced, ℓ is a suffix of an order $n+1$ symbol.

We have shown that $\text{Zimin}(\psi(b)\{\!\{\, i \,\}\!\}_{n+1}) \leq n+2$.

It remains to consider the case of $\{\!\{\, i \,\}\!\}_{n+1}\psi(b)$. Remark that, as the definition of higher-order counters is not symmetrical with respect to left-right and right-left, this case is not identical to the previous one.

Let $\alpha\beta\alpha$ be an infix of $\{\!\{\, i \,\}\!\}_{n+1}\psi(b)$ for some non-empty α and β. By Fact 1, it is enough to show that $\text{ZType}(\alpha) \leq n+1$. By Lemma 7, it is enough to consider the case when α is non-simple and hence by Lemma 8, α has a unique parse $p = (\ell, u, r)$.

As $\alpha\beta\alpha$ is an infix of $\{\!\{\, i \,\}\!\}_{n+1}\psi(b)$, there exist z_1 and z_2 such that:

$$\{\!\{\, i \,\}\!\}_{n+1}\psi(b) = z_1\alpha\beta\alpha z_2$$

We distinguish cases on the length of z_2.

Case 6D: $|z_2| \geq |\psi(b)|$. In this case, $\alpha\beta\alpha$ is an infix of $\{\!\{\, i \,\}\!\}_{n+1}$. We have already shown that $\text{Zimin}(\{\!\{\, i \,\}\!\}_{n+1}) \leq n+2$.

Case 6E: $|z_2| = 0$. We will show that $|\alpha| < |\psi(b)|$ and hence by Fact 1, $\text{Zimin}(\alpha) \leq \lfloor \log_2(2n+6) \rfloor \leq n+1$ as $n \geq 3$.

Assume towards a contradiction that $|\alpha| \geq |\psi(b)|$. Hence $\psi(b)$ is a suffix of α. Therefore, $\{\!\{\, i \,\}\!\}_{n+1}\psi(b)$ would contain two occurrences of $\psi(b)$. By Lemma 9, $[\![\, i \,]\!]_{n+1}b$ would contain two occurrences of the order $n+2$ symbol b which brings the contradiction.

Case 6F: $1 \leq |z_2| < |\psi(b)|$.

Recall that $\{\!\{\, i \,\}\!\}_{n+1}$ ends with $\{\!\{\, \boldsymbol{\tau}(n)-1 \,\}\!\}_n\psi(b')$ for some $b' \in \{0_{n+1}, 1_{n+1}\}$. We now distinguish cases on the length of αz_2.

Subcase : $|\alpha z_2| \leq |\psi(b')\psi(b)$—$|$**.**

As b is an order $n+2$ symbol and b' an order $n+1$ symbol, we have that $|\alpha| < 4 + 2n + 4 + 2(n+1)$. By Lemma 1, it follows that $\mathrm{Zimin}(\alpha) \leq \lfloor \log_2(4n+10) \rfloor$. Furthermore as for all $n \geq 3$ we have $\lfloor \log_2(4n+10) \rfloor \leq n+1$, we can conclude this subcase.

Subcase : $|\psi(b')\psi(b)| < |\alpha z_2| \leq |\{\!\{\, \boldsymbol{\tau}(n) - 1 \,\}\!\}_n \psi(b')\psi(b)$—$|$**.**
In this case, the parse $p = (\ell, u, r)$ of α is such that:

- r is a non-empty prefix of $\psi(b)$,
- u ends with the order $n+1$ symbol b'.

By Lemma 10, the parse p has two occurences in $[\![\, i \,]\!]_{n+1} b$. Hence it has an occurrence in $[\![\, i \,]\!]_{n+1}$. As u ends with an order $n+1$ symbol and as any symbol of order $n+1$ can only be followed by a symbol of order 1 in $[\![\, i \,]\!]_{n+1}$, we have that r is a strict prefix of an order 1 symbol. In particular $|r| < 4$.

We have established that α is a suffix of $\{\!\{\, \boldsymbol{\tau}(n) - 1 \,\}\!\}_n \psi(b')r$ with $|r| < 4$. It remains to prove that $\mathrm{ZType}(\alpha) \leq n+1$.

Consider a decomposition of α as $\delta\gamma\delta$ for some non-empty δ and γ. Assume towards a contradiction that $|\delta| \geq |\psi(b')r|$. In this case, $\psi(b')$ is an infix of δ and hence α would have two occurrences of $\psi(b')$. By Lemma 9, the center of α's parse would contain two order $n+1$ symbols which contradicts the fact that α is a suffix of $\{\!\{\, \boldsymbol{\tau}(n) - 1 \,\}\!\}_n \psi(b')r$ which has precisely one occurrence of the code of one order $n+1$ symbol.

Hence we have $|\delta| < |\psi(b')r| \leq 2n+7$. By Lemma 1, $\mathrm{Zimin}(\delta) \leq \lfloor \log_2(2n+7) \rfloor$. As for all $n \geq 3$, it holds that $\lfloor \log_2(2n+7) \rfloor \leq n$. We have shown that $\mathrm{Zimin}(\delta) \leq n$ and hence $\mathrm{Zimin}(\alpha) \leq n+1$.

Subcase : $|\alpha z_2| > |\{\!\{\, \boldsymbol{\tau}(n) - 1 \,\}\!\}_n \psi(b')\psi(b)$—$|$**.**
This case cannot occur under our assumptions. Indeed, this would imply that the center u of the parse $p = (\ell, u, r)$ of α contains $[\![\, \boldsymbol{\tau}(n) - 1 \,]\!]_n$. As the parse p has at least two occurrences in $[\![\, i \,]\!]_{n+1} b$, it would imply that $[\![\, \boldsymbol{\tau}(n) - 1 \,]\!]_n$ has two occurrences in $[\![\, i \,]\!]_{n+1}$, which contradicts Lemma 2.

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 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 unvavoidable 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 [9].

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 [8] 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.

Definition 7 Let $n, k \geq 1$. 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 analytic combinatorics [11] Tao showed the following lower bound for g.

Theorem 9 (Tao [22], Corollary 3) *Let $k \geq 4$. Then*

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

Unfortunately, it was not clear to the authors what the asymptotic behavior of this lower bound is. However Jugé [13] provided us with an estimate of its asymptotic behavior.

Corollary 2 (Jugé [13]) *Let $k \geq 4$. The expression in Theorem 9, and hence $g(n,k)$, is lower-bounded by*

$$\left(\frac{1}{\sqrt{21}} + o(1) \right) \frac{k^{2^{n-1}}}{k^{(n+1)/2}} \quad .$$

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 combining Theorem 9 with Corollary 2 but its proof seems more direct (already more direct than the proof of Theorem 9 itself). The proof follows a similar strategy as the (slightly better) doubly-exponential lower bound for f from [6], but again, seems to be more direct. Our novel contribution is to provide a doubly-exponential upper bound on g in Section 5.2. Note that Tao in [22] 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. We 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 [6].

Theorem 10 *Let* $k \geq 2$. *Then*

$$g(n, k) > k^{\lfloor \frac{2^n}{n+2} \rfloor - 1} \quad .$$

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 $[k]^\ell$. Remark 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 $g(n, k) > \ell$. Therefore we investigate those $\ell = \ell(n, k)$ for which we can guarantee $\Delta_{n,k,\ell} < 1$. We need two intermediate claims.

Claim 1. The probability that m pairwise independent 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 $(1/k)^{m-1}$.

Proof of Claim 1. We only show the claim only for $m = 2$, the case when $m > 2$ can be shown analogously. 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 $h \geq 1$. For every word $w = w_1 \cdots w_h \in [k]^h$, let $\oplus_k w = \left(\sum_{i=1}^h w_i \right) \bmod k$. Remark that $u \equiv v$ implies that $\oplus_k u = \oplus_k v$. Let us fix any $j \in [k]$. Then we clearly have $\Pr[\oplus w = j] = 1/k$ for every random word $w = w_1 \cdots w_h \in [k]^h$. Thus,

$$\begin{aligned} \Pr(A_h) &\leq \sum_{j \in [1,k]}^k \Pr[\oplus_k u = \oplus_k v = j] \\ &= \sum_{j \in [1,k]}^k \Pr[\oplus_k u = j] \Pr[\oplus_k v = j] = \sum_{j \in [1,k]}^k 1/k^2 = 1/k \end{aligned}$$

End of the proof of Claim 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 $\text{width}(\lambda) = \sum_{i=1}^n 2^{n-i} \cdot \lambda(x_i)$. For every word v of length $\text{width}(\lambda)$ 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 $i \in [n]$.

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

Proof of Claim 2. Let $\lambda : \mathcal{X}_n \to \mathbb{N}^+$ with $d = \mathrm{width}(\lambda)$. For $i \in [n]$, let $j_1^{(i)} < \cdots < j_{2^{n-i}}^{(i)}$ be an enumeration of the 2^{n-i} indices corresponding to occurrences of x_i in Z_n. For all $i \in [n]$ consider the event $B_\lambda^{(i)}$ that a random word of $[k]^d$ has its decomposition with respect to λ of the form $v_1 \cdots v_{2^n-1}$ such that the words $v_{j_1^{(i)}}, \ldots, v_{j_{2^{n-i}}^{(i)}}$ (which are all of length $\lambda(x_i)$) are pairwise equivalent with respect to $\equiv$. The event B_λ is the intersection of the events $B_\lambda^{(1)}, \ldots, B_\lambda^{(n-1)}$ and $B_\lambda^{(n)}$. As $B_\lambda^{(1)}, \ldots, B_\lambda^{(n-1)}$ and $B_\lambda^{(n)}$ are mutually independent events, the probability $\Pr(B_\lambda)$ is equal to $\prod_{i=1}^{n} \Pr(B_\lambda^{(i)})$. We have

$$\Pr(B_\lambda) = \prod_{i=1}^{n} \Pr(B_\lambda^{(i)}) \overset{\text{Claim 1}}{\leq} \prod_{i=1}^{n} (1/k)^{2^{n-i}-1} = k^{-\left(\sum_{i=1}^{n} 2^{n-i}\right)+n} = k^{n-2^n+1} \quad (6)$$

End of the proof of Claim 2.

It is clear that for every (j, λ), where $d = \mathrm{width}(\lambda)$ and $j + d \leq \ell$, the probability that (j, λ) is an occurrence of a random word from $[k]^\ell$ equals to probability that $(0, \lambda)$ is such an occurrence and therefore equals $\Pr(B_\lambda)$. Thus, this probability does not depend on j.

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

$$
\begin{aligned}
\Delta_{n,k,\ell} \quad \leq \quad & \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} \sum_{\substack{\lambda:\mathcal{X}_n\to\mathbb{N}^+\\ \mathrm{width}(\lambda)=d}} \Pr\left[(j,\lambda) \text{ is an occ. in a random word in } [k]^\ell\right] \\
\leq \quad & \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} \sum_{\substack{\lambda:\mathcal{X}_n\to\mathbb{N}^+\\ \mathrm{width}(\lambda)=d}} \Pr(B_\lambda) \\
\overset{\text{Claim 2}}{\leq} \quad & \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} \sum_{\substack{\lambda:\mathcal{X}_n\to\mathbb{N}^+\\ \mathrm{width}(\lambda)=d}} k^{n-2^n+1} \\
\leq \quad & \sum_{d=2^n-1}^{\ell} \sum_{j=0}^{\ell-d} d^n \cdot k^{n-2^n+1} \\
\leq \quad & \sum_{d=2^n-1}^{\ell} \ell \cdot d^n \cdot k^{n-2^n+1} \\
\leq \quad & \frac{\ell^2 \cdot \ell^n}{k^{2^n-n-1}} \\
= \quad & \frac{\ell^{n+2}}{k^{2^n-n-1}} \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (7)
\end{aligned}
$$

We finally determine the largest value of ℓ that still guarantees that $\Delta_{n,k,\ell} < 1$.

$$
\begin{aligned}
\Delta_{n,k,\ell} < 1 \quad &\overset{(7)}{\Longleftarrow} \quad \frac{\ell^{n+2}}{k^{2^n-n-1}} < 1 \\
&\Longleftarrow \quad \ell^{n+2} < k^{2^n-n-1} \\
&\Longleftarrow \quad (n+2)\log_k \ell < 2^n - n - 1 \\
&\Longleftarrow \quad \log_k \ell < \frac{2^n - n - 1}{n+2} \\
&\Longleftarrow \quad \ell < k^{\frac{2^n-n-1}{n+2}} \\
&\Longleftarrow \quad \ell < k^{\frac{2^n}{n+2} - \frac{n+1}{n+2}} \\
&\Longleftarrow \quad \ell = k^{\lfloor \frac{2^n}{n+2} \rfloor - 1}
\end{aligned}
$$

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 11 $g(n,k) \leq 2^{(4k)^n (n-1)!}$.

Proof We prove the statement by induction on n. For $n = 1$ we have

$$
g(1,k) \quad = \quad 1 \quad \leq \quad 2^{(4k)^1 (1-1)!} \quad .
$$

For the induction step, let $n \geq 1$ and let us assume induction hypothesis for $g(n,k)$. To determine an upper bound $g(n+1,k)$ we consider any sufficiently long word $w \in [k]^+$ that we can factorize as $w = w_1 a_1 w_2 a_2 \cdots w_m a_m z$, where $|w_j| = g(n,k)$, $a_j \in [k]$ for all $j \in [m]$ and $z \in [k]^*$, where m is assumed sufficiently large for the following arguments to work. By induction hypothesis for all $j \in [m]$, w_j encounters Z_n in the abelian sense, witnessed in some infix v_j and some abelian factorization $v_j = v_j^{(1)} \cdots v_j^{(2^n-1)}$ for Z_n. To each such abelian factorization we can assign the Parikh image how the word v_j matches each variable x_i (with $i \in [n]$) that appears 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 $\psi_j(x_i)(t) = |v_j^{(2^i-1)}|_t$ for all $j \in [m]$, all $i \in [n]$ and all $t \in [k]$. 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 clearly w encounters $Z_{n+1} = Z_n x_{n+1} Z_n$ in the abelian sense. Let us therefore estimate a sufficiently large bound on m such that there are always two distinct indices $i, j \in [1, m]$ that satisfy $\psi_i \equiv \psi_j$.

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 by setting $m = g(n,k)^{kn} + 1$ we have shown

$$g(n+1,k) \quad \leq \quad (g(n,k)+1)(g(n,k)^{kn}+1) \quad . \tag{8}$$

Hence, we obtain

$$
\begin{aligned}
g(n+1,k) \quad &\overset{(8)}{\leq} \quad (g(n,k)+1)(g(n,k)^{kn}+1) \\
&\overset{g(n,k)\geq 1}{\leq} \quad 2 \cdot g(n,k) \cdot 2 \cdot g(n,k)^{kn} \\
&= \quad 4 \cdot g(n,k)^{kn+1} \\
&\overset{n \geq 1}{\leq} \quad 4 \cdot g(n,k)^{2kn} \\
&\overset{\text{IH}}{\leq} \quad 4 \cdot \left(2^{(4k)^n(n-1)!} \right)^{2kn} \\
&= \quad 4 \cdot 2^{2 \cdot 4^n k^{n+1} n!} \\
&= \quad 2^{2 \cdot 4^n k^{n+1} n! + 2} \\
&\overset{n \geq 1}{\leq} \quad 2^{2(2 \cdot 4^n k^{n+1} n!)} \\
&= \quad 2^{(4k)^{n+1} n!}
\end{aligned}
$$

6 Conclusion

We have established a lower bound for $f(n,k)$ that is already non-elementary when $k = 2$. A first element of an answer is that the first moment method used in [6] cannot be used to obtain a lower bound that is asymptotically above doubly-exponential. Indeed, as for a length $\ell \geq k^{2^n-n-1} + 2^n$, 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

$$\Pr(A_n) = \prod_{i=1}^{n} (1/k)^{2^{n-i}-1} = k^{-2^n + n + 1}.$$

Assume that $\ell \geq k^{2^n - n - 1} + 2^n$. For each $i \in [0, k^{2^n - n - 1}]$, 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, it follows that

$$\Delta_{n,k,\ell} \geq \sum_{i=0}^{k^{2^n - n - 1}} E(X_i) \geq (k^{2^n - n - 1} + 1) \Pr(A_n) = 1 + \frac{1}{k^{2^n - n - 1}} \geq 1.$$

Thus, more advanced probabilistic method techniques are necessary. Indeed, very recently [5] Condon, Fox and Sudakov have applied the local lemma to obtain non-elementary lower bounds on $f(n, k)$.

For the abelian case, an explicit family of words witnessing the doubly-exponential lower bound seems worth investigating.

References

1. N. Alon and J. Spencer. *The Probabilistic Method*. Wiley, 2015.
2. Marie-Pierre Béal, Olivier Carton, Christophe Prieur, and Jacques Sakarovitch. Squaring transducers: an efficient procedure for deciding functionality and sequentiality. *Theor. Comput. Sci.*, 292(1):45–63, 2003.
3. Arnaud Carayol and Stefan Göller. On long words avoiding zimin patterns. In Heribert Vollmer and Brigitte Vallée, editors, *34th Symposium on Theoretical Aspects of Computer Science, STACS 2017, March 8-11, 2017, Hannover, Germany*, volume 66 of *LIPIcs*, pages 19:1–19:13. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017.
4. Arnaud Carayol and Stefan Göller. On long words avoiding Zimin patterns. *CoRR*, abs/1902.05540, 2019.
5. David Conlon, Jacob Fox, and Benny Sudakov. Tower-type bounds for unavoidable patterns in words, 2017.
6. J. Cooper and D. Rorabaugh. Bounds on Zimin word avoidance. *CoRR*, abs/1409.3080, 2014.
7. J. Cooper and D. Rorabaugh. Asymptotic density of Zimin words. *Discrete Mathematics & Theoretical Computer Science*, Vol. 18, no 3, 2016.
8. J. D. Currie. Pattern avoidance: themes and variations. *Theor. Comput. Sci.*, 339(1):7–18, 2005.
9. J. D. Currie and V. Linek. Avoiding patterns in the abelian sense. *Canadian J. Math.*, 51(4):696–714, 2001.
10. G.F. McNulty D.R. Bean, A. Ehrenfeucht. Avoidable patterns in strings of symbols. *Pac. J. of Math.*, 85:261—-294, 1979.
11. P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.
12. P. Jancar. Equivalences of pushdown systems are hard. In *In Proceedings of FOSSACS 2014*, volume 8412 of *Lecture Notes in Computer Science*, pages 1–28. Springer, 2014.
13. V. Jugé. Abelian ramsey length and asymptotic lower bounds. *CoRR*, abs/1609.06057, 2016.

14. 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.
15. D. Rorabaugh. Toward the combinatorial limit theory of free words. *CoRR*, abs/1509.04372, 2015.
16. W. Rytter and A. M. Shur. Searching for Zimin patterns. *Theor. Comput. Sci.*, 571:50–57, 2015.
17. M. V. Sapir. Combinatorics on words with applications. Technical report, 1995.
18. S. Schmitz. Complexity hierarchies beyond elementary. *CoRR*, abs/1312.5686, 2014.
19. G. Sénizergues. The equivalence problem for t-turn DPDA is co-np. In *Proceedings of ICALP*, volume 2719 of *Lecture Notes in Computer Science*, pages 478–489. Springer, 2003.
20. 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.
21. L. J. Stockmeyer. *The complexity of decision problems in automata and logic*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1974.
22. J. Tao. Pattern occurrence statistics and applications to the ramsey theory of unavoidable patterns. *CoRR*, abs/1406.0450, 2014.
23. A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Skrifter I Mat.-Nat. Kl. Christiania*, 7:1—-22, 1906.
24. A. I. Zimin. Blocking sets of terms. *Math. USSR Sbornik*, 47:50–57, 1984.