PATTERN OCCURRENCE STATISTICS AND APPLICATIONS TO THE RAMSEY THEORY OF UNAVOIDABLE PATTERNS

JIM TAO

To the memory of Philippe Flajolet and Paul Erdős

Abstract. As suggested by Currie, we apply the probabilistic method to problems regarding pattern avoidance. Using techniques from analytic combinatorics, we calculate asymptotic mean pattern occurrence and use them in conjunction with the probabilistic method to establish new results about the Ramsey theory of unavoidable patterns in the abelian full word case and in the nonabelian partial word case.

Keywords: Combinatorics on words; Partial words; Unavoidable patterns; Abelian patterns; Probabilistic method; Analytic combinatorics; Ramsey theory.

1. Introduction

In [13], Currie reviews results and formulates a large number of open problems concerning pattern avoidance as well as an abelian variation of it. Given a pattern p over an alphabet $\mathcal{V}$ and a word w over an alphabet $\mathcal{A}$, we say that *w encounters p* if there exists a nonerasing morphism $h : \mathcal{V}^* \to \mathcal{A}^*$ such that $h(p)$ is a factor of w; otherwise *w avoids p*. In other words, w encounters $p = p_1 \cdots p_n$, where $p_1, \ldots, p_n \in \mathcal{V}$, if w contains $u_1 \cdots u_n$ as a factor, where $u_1, \ldots, u_n$ are nonempty words in $\mathcal{A}^*$ satisfying $u_i = u_j$ whenever $p_i = p_j$. On the other hand, *w encounters $p = p_1 \cdots p_n$ in the abelian sense* if w contains $u_1 \cdots u_n$ as a factor, where u_i can be obtained from u_j by rearranging letters whenever $p_i = p_j$; otherwise *w avoids p in the abelian sense*.

Words avoiding patterns such as squares have been used to build several counterexamples in context-free languages [26], groups [1], lattice of varieties [19], partially ordered sets [33], semigroups [14, 20], symbolic dynamics [27], to name a few. Words avoiding squares in the abelian sense have also been used in the study of free partially commutative monoids [12,15], and have helped characterize the repetitive commutative semigroups [20]. In addition, words avoiding more general patterns find applications in algorithmic problems on algebraic structures [22].

In this paper, we meet the goal of Problem 4 as expressed by Currie in [13], which is to "explore the scope of application of the probabilistic method to problems in pattern avoidance." The probabilistic method [2], pioneered by Erdős, has recently become

Date: February 14, 2019.

1991 *Mathematics Subject Classification.* 05A15, 05A16.

This material is based upon work supported by the National Science Foundation under Grant No. DMS–1060775.

one of the most powerful techniques in combinatorics. It is used to demonstrate, via statistical means, the existence of certain combinatorial objects without constructing them explicitly. Analytic combinatorics [17], pioneered by Flajolet and Sedgewick and expanded to the multivariate case [28] by Pemantle and Wilson, allows precise calculation of the statistics of large combinatorial structures by studying their associated generating functions through the lens of complex analysis. Since analytic combinatorics calculates the statistics of large combinatorial structures, and the probabilistic method uses such statistics to infer the existence of specific combinatorial objects, we use both techniques in tandem to prove some Ramsey theoretic results about pattern avoidance.

We also extend some of our results to partial words, which allow for undefined positions represented by hole characters. In this context, given a pattern p over $\mathcal{V}$ and a partial word w over $\mathcal{A}$, we say that w *encounters* p if there exists a nonerasing morphism $h : \mathcal{V}^* \to \mathcal{A}^*$ such that $h(p)$ is *compatible with* a factor of w. Several results concerning (abelian) pattern avoidance have recently been proved in this more general context of partial words (see, for example, [3–9]).

The contents of our paper are as follows. In Section 2, we discuss some basic concepts, fix some notations, and mention some previous results from the literature. In Section 3, we discuss some tools, such as ordinary generating functions, and techniques from analytic combinatorics. In Section 4, we use those tools and techniques in conjunction with the probabilistic method to calculate asymptotic pattern occurrence statistics and to establish new results about the Ramsey theory of unavoidable patterns in the full word case (both nonabelian sense and abelian sense) and the partial word case. Finally in Section 5, we suggest additional possible uses of these data in applications such as cryptography and musicology. We also discuss a number of open problems.

2. Basic concepts, notations, and known results

A (full) *word* over an alphabet $\mathcal{A}$ is a sequence of characters from $\mathcal{A}$. We call the characters in $\mathcal{A}$ *letters*. The number of characters in a word is its *length*. We denote by $\mathcal{A}^*$ the set of all words over $\mathcal{A}$; when equipped with the concatenation or product of words, where the empty word ε serves as identity, it is called the free monoid generated by $\mathcal{A}$. A word w over $\mathcal{A}$ *encounters* the word p over an alphabet $\mathcal{V}$ if w contains $h(p)$ as a factor for some nonerasing morphism $h : \mathcal{V}^* \to \mathcal{A}^*$. Otherwise w *avoids* p and is *p-free*. In this case we interpret p to be a *pattern*. For example, the word *tennessee* encounters the pattern *abaca*, as witnessed by the morphism $h : \{a, b, c\}^* \to \{e, n, s, t\}^*$ with $h(a) = e$, $h(b) = nn$, and $h(c) = ss$. Thus *tennessee* contains $h(abaca) = ennesse$, a factor of tennessee.

We count multiple instances of a pattern in a word as follows: we say that w encounters p a total of $N > 0$ times if, for some maximal $m > 0$, there exist m distinct nonerasing morphisms $h_i : \mathcal{V}^* \to \mathcal{A}^*$ such that for some $t_1, \ldots, t_m > 0$, $h_i(p)$ is a factor of w exactly $t_i > 0$ times, and $\sum_{i=1}^{m} t_i = N$. For example, the word 11111111

encounters the pattern *aba* 34 times because for $3 \leq k \leq 8$, each of the $9-k$ factors of length k lies in the image of $\lfloor (k-1)/2 \rfloor$ nonerasing morphisms $\{a,b\}^* \to \{1\}^*$, and $6 \cdot 1 + 5 \cdot 1 + 4 \cdot 2 + 3 \cdot 2 + 2 \cdot 3 + 1 \cdot 3 = 34$. One may object to this definition on the basis that the factor 11111111 is counted as three occurrences of the pattern *aba*, but since pattern occurrences are defined in terms of nonerasing morphisms, it makes sense to count the same factor multiple times if it lies in the image of multiple distinct nonerasing morphisms. Patterns are an abstract idea that goes beyond the concrete words that they map to under these nonerasing morphisms; they are a kind of symmetry that exists in the words in which they appear. For that reason the *aba* pattern group of the factor 11111111 should be larger than that of a factor 12345671, just as the group $O(2)$ of symmetries of a circle is larger than the group D_4 of symmetries of a square.

A *partial word* over $\mathcal{A}$ is a sequence of characters from the extended alphabet $\mathcal{A} + \{\diamond\}$, where we refer to $\diamond$ as the *hole* character. Define the *hole density* of a partial word to be the ratio of its number of holes to its length, i.e. $d := h/n$ where d is the hole density, h is the number of holes, and n is the length of the partial word. A *completion* of a partial word w is a full word constructed by filling in the holes of w with letters from $\mathcal{A}$.

If $u = u_1 \cdots u_n$ and $v = v_1 \cdots v_n$ are partial words of equal length n, where $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ denote characters from $\mathcal{A} + \{\diamond\}$, we say that u is *compatible* with v, denoted $u \uparrow v$, if $u_i = v_i$ whenever $u_i, v_i \in \mathcal{A}$. A partial word w over $\mathcal{A}$ encounters the full word p over $\mathcal{V}$ if some factor f of w satisfies $f \uparrow h(p)$ for some nonerasing morphism $h : \mathcal{V}^* \to \mathcal{A}^*$. Otherwise w *avoids* p and is *p-free*. Again we interpret p to be a *pattern*. For example, the partial word *velve⋄ta* encounters *abab*, as witnessed by the morphism $h : \{a,b\}^* \to \{a,e,l,v,t\}^*$ with $h(a) = ve$ and $h(b) = l$. Thus $h(abab) = velvel$, which is compatible with *velve⋄*, a factor of *velve⋄ta*. We count multiple instances of a pattern in a partial word as follows: we say that w encounters p a total of $N > 0$ times if, for some maximal $m > 0$, there exist m distinct nonerasing morphisms $h_i : \mathcal{V}^* \to \mathcal{A}^*$ such that, for some $t_1, \ldots, t_m > 0$, there are t_i factors f_i of w that satisfy $f_i \uparrow h_i(p)$, and $\sum_{i=1}^{m} t_i = N$.

Suppose $p = p_1 \cdots p_n$ where $p_1, \ldots, p_n \in \mathcal{V}$. A full word w *encounters* p *in the abelian sense* if w contains $u_1 \cdots u_n$ as a factor, where word u_i can be obtained from word u_j by rearranging letters whenever $p_i = p_j$. Otherwise w *avoids* p *in the abelian sense* and is *abelian p-free*. For example, the full word *v al h al la* encounters *abaa* in the abelian sense. We count multiple instances of an abelian pattern in a word as follows: we say that w encounters p in the abelian sense $N > 0$ times if, for some maximal $m > 0$, there exist m distinct sequences of words S_i of the form $(u_1, \ldots, u_n)$ such that w contains $u_1 \cdots u_n$ as a factor $t_i > 0$ times, word u_j can be obtained from word u_k by rearranging letters whenever $p_j = p_k$, and $\sum_{i=1}^{m} t_i = N$.

A pattern p is *m-avoidable* if there are arbitrarily long words over an m-letter alphabet that avoid p. A pattern p is *m-avoidable over partial words* if for every $h \in \mathbb{N}$ there is a partial word with h holes over an m-letter alphabet that avoids p. A pattern p is *m-avoidable in the abelian sense* if there are arbitrarily long words over an m-letter alphabet that avoid p in the abelian sense. Otherwise, p is, respectively, *m-unavoidable*,

m-unavoidable over partial words, and *m-unavoidable in the abelian sense*. For example, the *Zimin patterns* Z_i where

$$Z_1 = a_1 \text{ and } Z_i = Z_{i-1}a_iZ_{i-1} \tag{2.1}$$

are m-unavoidable for all $m \geq 1$ [25]. They are also m-unavoidable over partial words for all $m \geq 1$ as well as m-unavoidable in the abelian sense for all $m \geq 1$. Indeed, since Z_i occurs in a partial word whenever it occurs in some completion of the partial word, Z_i is unavoidable over partial words, and since all occurrences of Z_i in the nonabelian sense are occurrences of Z_i in the abelian sense, Z_i is unavoidable in the abelian sense.

Define the *Ramsey length* $L(m, p)$ of an m-unavoidable pattern p to be the minimal length of a word over an m-letter alphabet that ensures the occurrence of p. Similarly, define the *partial Ramsey length* $L_d(m, p)$ of a pattern p that is m-unavoidable over partial words with hole density $\geq d$ to be the minimal length of a partial word with hole density d over an m-letter alphabet that ensures the occurrence of p, and define the *abelian Ramsey length* $L_{\text{ab}}(m, p)$ of a pattern p that is m-unavoidable in the abelian sense to be the minimal length of a word over an m-letter alphabet that ensures the occurrence of p in the abelian sense.

For small values of m and i, we have the following table of results for $L(m, Z_i)$, as compiled in [10] from papers [31] and [32]:

	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					

Currently, the best known lower bound for $L(m, Z_i)$ is a tower of $i-3$ exponentials, even for $m = 2$.

Theorem 2.1. *[10] For all $i \geq 1$ and $m \geq 2$, $L(m, Z_i) \geq 2 \uparrow\uparrow (i-3)$.*

In the paper we do not improve on this lower bound, but the techniques we develop and use also work for the abelian and partial word cases, helping us prove new lower bounds for abelian and partial Ramsey lengths.

The best known upper bound for $L(m, Z_i)$ is a tower of exponentials of height $i-1$.

Theorem 2.2. *[11] For all $i \geq 1$ and $m \geq 2$, $L(m, Z_i) \leq (2m+1) \uparrow\uparrow (i-1)$.*

We prove another exponential tower upper bound which is not as good but can gotten through brute force without a combinatorial argument.

Knuth's *up-arrow notation* [23] as used above is defined as follows: For all integers x, y, n such that $y \geq 0$ and $n \geq 1$:

$$x \uparrow^n y = \begin{cases} x^y & \text{if } n = 1, \\ 1 & \text{if } y = 0, \\ x \uparrow^{n-1} (x \uparrow^n (y-1)) & \text{otherwise.} \end{cases}$$

More specifically, we use the double up-arrow, i.e. the above operator where $n = 2$. For example, $3 \uparrow\uparrow 3 = 3^{3^3}$. We make use of the following identity regarding double up-arrows

$$x \uparrow\uparrow (n+1) = x^{x\uparrow\uparrow n}, \tag{2.2}$$

which follows by induction on $n \geq 0$. Using the same example as before, we observe that $3 \uparrow\uparrow 3 = 3^{3\uparrow\uparrow 2} = 3^{3^3}$.

3. Tools and techniques

For standard terms and theorems related to the symbolic method, we refer the reader to the book of Flajolet and Sedgewick [17].

We can often specify a combinatorial class by performing a series of operations on basic "atomic" objects of size 1: cartesian product $\mathcal{B} \times \mathcal{C}$, combinatorial sum (disjoint union) $\mathcal{B} + \mathcal{C}$, sequence construction $\mathrm{SEQ}(\mathcal{B})$, and substitution $\mathcal{B} \circ \mathcal{C}$, where $\mathcal{B}, \mathcal{C}$ are combinatorial classes [17, pp. 25–26, 87]. As it turns out, specifications of combinatorial classes translate directly into generating functions. According to the admissibility theorem for ordinary generating functions [17, pp. 27, 87], the OGFs of such classes admit convenient closed-form expressions.

We recall some basic constructions [17, p. 50]: The class $\mathcal{E} = \{\varepsilon\}$ consisting of the neutral object only, and the class $\mathcal{Z}$ consisting of a single "atomic" object (node, letter) of size 1 have OGFs $E(z) = 1$ and $Z(z) = z$, respectively. Let $\mathcal{A} = m\mathcal{Z}$ denote an alphabet of m letters and $\mathcal{W} = \mathrm{SEQ}(\mathcal{A})$ denote the set of all possible words over $\mathcal{A}$. Then $\mathcal{A}$ and $\mathcal{W}$ have associated OGFs $A(z) = mz$ and $W(z) = 1/(1 - mz)$, respectively.

Tuples or repetitions of letters and words make an appearance frequently in our arguments. We construct them as follows: Let $\mathcal{J}_k = \mathcal{A} \circ \mathcal{Z}^k$ be the set of all k-tuples of the same letter in $\mathcal{A}$ and let $\mathcal{K}_k = \mathcal{W} \circ \mathcal{Z}^k$ be the set of all k-tuples of the same word over $\mathcal{A}$. Then $\mathcal{J}_k$ and $\mathcal{K}_k$ have associated OGFs $J_k(z) = mz^k$ and $\mathcal{K}_k = 1/(1 - mz^k)$.

The following theorem greatly simplifies the process of finding the asymptotics of a sequence, given knowledge of its generating function.

Theorem 3.1. *[17, p. 258] Let $f(z)$ be a function meromorphic at all points of the closed disc $|z| \leq R$, with poles at points $\alpha_1, \alpha_2, \ldots, \alpha_r$. Assume that $f(z)$ is analytic at all points of $|z| = R$ and at $z = 0$. Then there exist r polynomials $\{P_j\}_{j=1}^r$ such that*

$$f_n := [z^n]f(z) = \sum_{j=1}^{r} P_j(n)\alpha_j^{-n} + O(R^{-n}).$$

Furthermore the degree of P_j is equal to the order of the pole of f at α_j minus one.

The following theorem formalizes our use of the probabilistic method.

Theorem 3.2. *[2, p. 18] Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ be a real-valued random variable, i.e. such that for all $x \in \mathbb{R}$, $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$. Let*

$$E[X] := \int_\Omega X(\omega)\, \mathrm{d}P(\omega)$$

denote the mathematical expectation of X. If $E[X] < \infty$, then for some $\omega \in \Omega$, $X(\omega) \leq E[X]$.

Straightforward applications of the probabilistic method often give crude results, as demonstrated below; nevertheless, they still provide important qualitative information.

Define the kth *Ramsey number* $R(k)$ to be the minimal value of n in the statement of Ramsey's theorem, Theorem 3.3, for a given value of k.

Theorem 3.3. *[29] For every positive integer k there is a positive integer n, such that if the edges of the complete graph on n vertices are all colored either red or blue, then there must be k vertices such that all edges joining them have the same color.*

In 1947, Erdős proved the following result.

Theorem 3.4. *$R(k) \geq 2^{k/2}$ for all $k \geq 2$.*

Erdős' lower bound, exponential with base $\sqrt{2}$, is rough (and so are all lower bounds on $R(k)$ proven since then) because the best known upper bounds are exponential with base 4. In fact, some of the major open problems in combinatorics, according to Gowers [18], are the following: Does there exist a constant $a > \sqrt{2}$ such that $R(k) \geq a^k$ for all sufficiently large k? Does there exist a constant $b < 4$ such that $R(k) \leq b^k$ for all sufficiently large k? Although Erdős' lower bound for $R(k)$ is crude, it tells us valuable information about $R(k)$, namely that it grows at least exponentially.

4. Mean pattern occurrence: the full word case

We calculate mean pattern occurrence and prove results about the Ramsey theory of unavoidable patterns in the following cases: nonabelian full words, abelian full words, and nonabelian partial words. We calculate the mean number of occurrences of a pattern in a word of a given length and use that statistic to establish a lower bound on Ramsey lengths.

Theorem 4.1 together with Corollary 4.3 answer the basic question as to when a full word can avoid a given pattern. Theorem 4.2 together with Corollary 4.4 answer the basic question as to when a full word can avoid a given pattern in the abelian sense.

Theorem 4.1. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $k = |p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. Then the mean number of occurrences of p in a full word of length n over an alphabet of m letters is*

$$\widehat{\Omega}_n \sim \frac{1}{\prod_{j=s+1}^{r}(m^{k_j-1}-1)} \frac{n^{s+1}}{(s+1)!}.$$

To illustrate Theorem 4.1, consider the pattern $p = abacaba$, where $r = 3$, $s = 1$, and where $k_1 = 1$, $k_2 = 2$, and $k_3 = 4$ denote, respectively, the number of occurrences of c, b, and a in p. Substituting these variables, $m = 12$, and $n = 100$, we find that

$$\widehat{\Omega}_{100} \approx 0.26319\cdots.$$

Theorem 4.2. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $k = |p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. Then the mean number of occurrences of p in the abelian sense in a word of length n over an alphabet of $m \geq 4$ letters is*

$$\widehat{\Omega}_n \sim \frac{n^{s+1}}{(s+1)!}\prod_{j=s+1}^{r}\left[\sum_{\ell=1}^{\infty}\frac{1}{m^{k_j\ell}}\sum_{i_1+\cdots+i_m=\ell}\binom{\ell}{i_1,\ldots,i_m}^{k_j}\right].$$

To illustrate Theorem 4.2, consider the pattern $p = aba$, where $r = 2$, $s = 1$, $k_1 = 1$, and $k_2 = 2$. Substituting these variables, $m = 12$, and $n = 100$, and applying [30, Theorem 4] we find that

$$\widehat{\Omega}_{100} \approx \frac{100^2 \cdot 12^{12/2}(4\pi)^{-11/2}}{2}\zeta\left(\frac{11}{2}\right) \approx 13778.87\cdots.$$

Corollary 4.3. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $|p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. If*

$$n < (1+o(1))\left[(s+1)!\prod_{j=s+1}^{r}(m^{k_j-1}-1)\right]^{\frac{1}{s+1}},$$

there is a word of length n over an alphabet of m letters that avoids p. If $p = Z_i$ then

$$L(m, Z_i) \geq (1+o(1))\sqrt{2\prod_{j=1}^{i-1}(m^{2^j-1}-1)}. \tag{4.1}$$

Substituting for example $m = 12$ and $i = 3$, we find that

$$L(12, Z_3) \geq 194.92\cdots.$$

Corollary 4.4. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $|p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. For*

$$n < (1+o(1))\left\{(s+1)!\prod_{j=s+1}^{r}\left[\sum_{\ell=1}^{\infty}\frac{1}{m^{k_j\ell}}\sum_{i_1+\cdots+i_m=\ell}\binom{\ell}{i_1,\ldots,i_m}^{k_j}\right]^{-1}\right\}^{\frac{1}{s+1}},$$

there is a word of length n over an alphabet of $m \geq 4$ letters that avoids p in the abelian sense. If $p = Z_i$ then

$$L_{\mathrm{ab}}(m, Z_i) \geq (1 + o(1)) \sqrt{2 \prod_{j=1}^{i-1} \left[\sum_{\ell=1}^{\infty} \frac{1}{m^{2^j \ell}} \sum_{i_1 + \cdots + i_m = \ell} \binom{\ell}{i_1, \ldots, i_m}^{2^j} \right]^{-1}}.$$

These corollaries are rather crude. The first, Corollary 4.3 says that the maximum length of words avoiding *aa* is at least $m - 2$, but m is the cardinality of the alphabet. Nevertheless, it says that, as a variable is repeated, the maximum length of words avoiding the associated pattern grows at least exponentially. For example, the maximum length of words avoiding the pattern a^k is at least $m^{k-1} - 2$. The maximum length of words avoiding the Zimin pattern Z_i, as defined in Equation (2.1), is at least $-1 + \sqrt{2 \prod_{j=1}^{i-1} (m^{2^j - 1} - 1)}$.

First we prove the nonabelian word case, Theorem 4.1.

Proof. For the *mean* number of occurrences of a pattern p, calculations similar to those employed for the number of occurrences of a word [17, p. 61] can be based on regular specifications. Each occurrence of p consists of a concatenation of nonempty words (represented by $\mathcal{W} \setminus \{\varepsilon\} = \mathrm{SEQ}(\mathcal{A}) \setminus \{\varepsilon\}$) repeated k_j times for the jth variable, surrounded by arbitrary sequences of letters. Thus all the occurrences of p as a factor are described by

$$\widehat{\mathcal{O}} = \mathrm{SEQ}(\mathcal{A}) \times \prod_{j=1}^{r} [(\mathcal{W} \setminus \{\varepsilon\}) \circ \mathcal{Z}^{k_j}] \times \mathrm{SEQ}(\mathcal{A}),$$

so we get

$$\begin{aligned} \widehat{O}(z) &= \frac{1}{(1 - mz)^2} \prod_{j=1}^{r} \left(\frac{1}{1 - mz^{k_j}} - 1 \right) \\ &= \frac{m^r z^k}{(1 - mz)^{2+s}} \prod_{j=s+1}^{r} \frac{1}{1 - mz^{k_j}}. \end{aligned} \tag{4.2}$$

We have a pole of order $2 + s$ at $z = 1/m$, and poles at the k_j different k_jth roots of $1/m$ for $k_j \geq 2$ (which have modulus greater than $1/m$). By Theorem 3.1, we know that for any $R > 1$, there exist polynomials $P_1, P_{s+1}, \ldots, P_r$ such that

$$[z^n] \widehat{O}(z) = P_1(n) m^n + \sum_{j=s+1}^{r} P_j(n) m^{n/k_j} + O(R^{-n}),$$

where the degree of P_1 is $s + 1$. For an asymptotic equivalent of $[z^n] \widehat{O}(z)$, only the pole at $z = 1/m$ needs to be considered because it is closest to the origin and corresponds to the fastest exponential growth; it is the dominant singularity. We plug in $z = 1/m$ in

Equation (4.2) for the nonsingular portion to obtain the first-order asymptotics of the OGF near $z = 1/m$:

$$\begin{aligned}\widehat{O}(z) &\sim \frac{m^{r-k}}{\prod_{j=s+1}^{r}(1-m^{1-k_j})}\frac{1}{(1-mz)^{2+s}}\\ &= \frac{1}{\prod_{j=s+1}^{r}(m^{k_j-1}-1)}\frac{1}{(1-mz)^{2+s}},\end{aligned}$$

which correspond to the first-order asymptotics of the associated sequence,

$$\begin{aligned}[z^n]\widehat{O}(z) &\sim \frac{1}{\prod_{j=s+1}^{r}(m^{k_j-1}-1)}\binom{n+s+1}{s+1}m^n\\ &\sim \frac{1}{\prod_{j=s+1}^{r}(m^{k_j-1}-1)}\frac{n^{s+1}m^n}{(s+1)!}.\end{aligned}$$

Therefore, the mean number of occurrences of a pattern p in a word of length n over an alphabet of m letters is

$$\widehat{\Omega}_n \sim \frac{1}{\prod_{j=s+1}^{r}(m^{k_j-1}-1)}\frac{n^{s+1}}{(s+1)!}.$$

□

Next we prove the abelian word case, Theorem 4.2.

Proof. Each occurrence of p consists of a concatenation of nonempty words repeated k_j times for the jth variable surrounded by arbitrary sequences of letters, with the additional $k_j - 1$ instances of each substituted word being allowed to permute their letters. Thus all the occurrences of p as a factor in the abelian sense are described by

$$\widehat{\mathcal{O}} = \mathrm{SEQ}(\mathcal{A}) \times \prod_{j=1}^{r}\left(\sum_{w\in\mathcal{W}\setminus\{\varepsilon\}}|\mathrm{Per}(w)|^{k_j-1}\mathcal{Z}^{k_j|w|}\right) \times \mathrm{SEQ}(\mathcal{A}),$$

where $\mathrm{Per}(w)$ denotes the set of distinct permutations of the word w. So we get

$$\begin{aligned}\widehat{O}(z) &= \frac{1}{(1-mz)^2}\prod_{j=1}^{r}\left(\sum_{\ell=1}^{\infty}z^{k_j\ell}\sum_{i_1+\cdots+i_m=\ell}\binom{\ell}{i_1,\ldots,i_m}^{k_j}\right)\\ &= \frac{m^s z^s}{(1-mz)^{2+s}}\prod_{j=s+1}^{r}\left(\sum_{\ell=1}^{\infty}z^{k_j\ell}\sum_{i_1+\cdots+i_m=\ell}\binom{\ell}{i_1,\ldots,i_m}^{k_j}\right).\end{aligned}$$

Since it is not obvious whether the generating function may be analytically continued beyond its radius of convergence, we treat it as though it is *lacunary*, i.e. not analytically continuable, and we use techniques from [16] to calculate asymptotics.

Note that $k_j \geq 2$ for $s+1 \leq j \leq r$ and

$$\binom{\ell}{i_1, \ldots, i_m} < \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m} = m^\ell$$

for $\ell \geq 2$. Applying [30, Theorem 4] and using our assumption that $m \geq 4$, we find that

$$\begin{aligned}
\left| \sum_{\ell=1}^{\infty} z^{k_j \ell} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \right| &\leq \sum_{\ell=1}^{\infty} |z|^{k_j \ell} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \\
&\leq \sum_{\ell=1}^{\infty} \frac{1}{m^{k_j \ell}} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \\
&= \sum_{\ell=1}^{\infty} \sum_{i_1+\cdots+i_m=\ell} \left[\frac{\binom{\ell}{i_1, \ldots, i_m}}{m^\ell} \right]^{k_j} \\
&\leq \sum_{\ell=1}^{\infty} \sum_{i_1+\cdots+i_m=\ell} \left[\frac{\binom{\ell}{i_1, \ldots, i_m}}{m^\ell} \right]^{2} \\
&= \sum_{\ell=1}^{\infty} \frac{1}{m^{2\ell}} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{2} \\
&\leq \sum_{\ell=1}^{\infty} (1+o(1)) m^{m/2} (4\pi\ell)^{(1-m)/2} \\
&\sim m^{m/2} (4\pi)^{(1-m)/2} \zeta\left(\frac{m-1}{2}\right) < \infty
\end{aligned}$$

so

$$\sum_{\ell=1}^{\infty} z^{k_j \ell} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j}$$

converges at all z in the closed disc $|z| \leq 1/m$, and the radius of convergence is at least $1/m$. In fact, $\sum_{\ell=1}^{\infty} z^{2\ell} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^2$ has radius of convergence $R = 1/m$ since, by the Cauchy–Hadamard theorem, the radius of convergence satisfies

$$\begin{aligned}
\frac{1}{R} &= \limsup_{\ell \to \infty} \left[\sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^2 \right]^{\frac{1}{2\ell}} \\
&= \limsup_{\ell \to \infty} \left[m^{2\ell + \frac{m}{2}} (4\pi\ell)^{(1-m)/2} \right]^{\frac{1}{2\ell}} \\
&= \lim_{\ell \to \infty} \left[m^{1 + \frac{m}{4\ell}} (4\pi\ell)^{\frac{1-m}{4\ell}} \right] \\
&= m.
\end{aligned}$$

We may factor $\widehat{O}(z)$ as $\widehat{O}(z) = P(mz) \cdot Q(mz)$, where

$$P(z) = \frac{1}{(1-z)^{2+s}}$$

and

$$Q(z) = z^s \prod_{j=s+1}^{r} \left[\sum_{\ell=1}^{\infty} \left(\frac{z}{m}\right)^{k_j \ell} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \right].$$

Note that $Q(z)$ is analytic in $|z| < 1$ and converges at all points on the unit disc. Also note that $Q(z)$ is $\mathcal{C}^\infty$-smooth on the unit circle; differentiating the power series any number of times does not make it diverge. In particular, $Q(z)$ is $\mathcal{C}^{2+s}$-smooth on the unit circle.

Note that $P(z)$ is of global order $-2-s$ and is its own log-power expansion of type $\mathcal{O}^t$ relative to $W = \{1\}$, where $t = \infty$. Since $t = \infty > u_0 = \lfloor ((2+s)+(-2-s))/2 \rfloor \geq 0$, the conditions of [16, Theorem 1] hold. Letting $c_0 = \lfloor ((2+s)-(-2-s))/2 \rfloor = 2+s$, we find that

$$[z^n](P(z) \cdot Q(z)) = [z^n](P(z) \cdot H(z)) + o(1),$$

where $H(z)$ is the Hermite interpolation polynomial such that all its derivatives of order $0, \ldots, 1+s$ coincide with those of $Q(z)$ at $w = 1$. Note that this implies that

$$H(1) = \prod_{j=s+1}^{r} \left[\sum_{\ell=1}^{\infty} \frac{1}{m^{k_j \ell}} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \right].$$

Scaling by a factor of m, since the singularity occurs at a radius $1/m$, we get

$$[z^n]\widehat{O}(z) = [z^n](P(mz) \cdot H(mz)) + o(1).$$

Since $H(z)$ is a polynomial, the only singularity of $P(mz) \cdot H(mz)$ is $z = 1/m$, so it dominates, and by Theorem 3.1,

$$\begin{aligned}
[z^n]\widehat{O}(z) &\sim \binom{n+s+1}{s+1} m^n \prod_{j=s+1}^{r} \left[\sum_{\ell=1}^{\infty} \frac{1}{m^{k_j \ell}} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \right] \\
&\sim \frac{n^{s+1} m^n}{(s+1)!} \prod_{j=s+1}^{r} \left[\sum_{\ell=1}^{\infty} \frac{1}{m^{k_j \ell}} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \right].
\end{aligned}$$

Therefore, the mean number of occurrences of a pattern p in the abelian sense in a word of length n over an alphabet of $m \geq 4$ letters is

$$\widehat{\Omega}_n \sim \frac{n^{s+1}}{(s+1)!} \prod_{j=s+1}^{r} \left[\sum_{\ell=1}^{\infty} \frac{1}{m^{k_j \ell}} \sum_{i_1+\cdots+i_m=\ell} \binom{\ell}{i_1, \ldots, i_m}^{k_j} \right].$$

□

When $\widehat{\Omega}_n < 1$, we may apply Theorem 3.2, so the corollaries follow.

According to [25, p. 101],

$$L(m, Z_i) \leq m^{L(m,Z_{i-1})}[L(m, Z_{i-1}) + 1] + L(m, Z_{i-1})$$

and $L(m, Z_2) = 2m + 1$. Through brute force, we can establish a crude upper bound for $L(m, Z_i)$, which we write using Knuth's up-arrow notation.

Theorem 4.5. *For $m \geq 2$ and $i \geq 2$,*

$$L(m, Z_i) < m \uparrow\uparrow (2i - 1). \tag{4.3}$$

Proof. Since $m \geq 2$, $L(m, Z_2) = 2m + 1 < m^{m^m} = m \uparrow\uparrow 3$ establishes our base case. For the inductive step, assume that for some $i \geq 2$,

$$L(m, Z_i) < m \uparrow\uparrow (2i - 1).$$

As stated earlier,

$$L(m, Z_{i+1}) \leq m^{L(m,Z_i)}[L(m, Z_i) + 1] + L(m, Z_i),$$

so

$$\begin{aligned}
L(m, Z_{i+1}) &< [m^{L(m,Z_i)} + 1][L(m, Z_i) + 1] \\
&< [m^{m\uparrow\uparrow(2i-1)} + 1][m \uparrow\uparrow (2i - 1) + 1] \\
&= [m \uparrow\uparrow (2i) + 1][m \uparrow\uparrow (2i - 1) + 1] \\
&< [m \uparrow\uparrow (2i) + 1][m \uparrow\uparrow (2i)] \\
&< [m \uparrow\uparrow (2i)]^3 \\
&= m^{3m\uparrow\uparrow(2i-1)} \\
&< m^{m\uparrow\uparrow(2i)} \\
&= m \uparrow\uparrow (2i + 1),
\end{aligned}$$

and our induction is complete. □

Our derived upper bound for $L(m, Z_i)$ in Equation (4.3), which uses tetration, is vastly greater than our derived lower bound for $L(m, Z_i)$ in Equation (4.1), which uses repeated squaring. Nevertheless, we have established concrete upper and lower bounds for $L(m, Z_i)$. The upper bound for $L(m, Z_i)$ in Theorem 4.5 also applies to $L_{\text{ab}}(m, Z_i)$.

5. Mean pattern occurrence: the partial word case

Next, we investigate the case of patterns in partial words. As in the case of full words, we calculate the mean number of pattern occurrences. First, we take the average over all partial words of a given length. Then we average over all strictly partial words of a given length, and finally, we take the average over all partial words of a given length with a given hole density. The last of these statistics, gotten through the calculation of bivariate asymptotics, allows us to prove a lower bound on partial Ramsey lengths.

5.1. **Mean over all partial words of a given length.** The following lemma, which can be proved by induction on k, will help us compare the distances of poles of the generating function from the origin and establish one of them as the dominant singularity.

Lemma 1. *If $m \geq 2$ and $k \geq 2$, then $m2^k - m + 1 < (m+1)^k$.*

Theorem 5.1 together with Corollary 5.2 answer the basic question as to when a partial word can avoid a given pattern.

Theorem 5.1. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $k = |p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. Then the mean number of occurrences of p in a partial word of length n over an alphabet of m letters is*

$$\widehat{\Omega}_n \sim \frac{n^{s+1}}{(s+1)!} \prod_{j=s+1}^{r} \frac{m2^{k_j} - m + 1}{(m+1)^{k_j} - (m2^{k_j} - m + 1)}.$$

Proof. For the *mean* number of occurrences of a pattern p in a partial word of length n, calculations similar to those employed for the number of occurrences of a pattern p in a full word of length n can be based on regular specifications. Each occurrence of p consists of a concatenation of nonempty full words repeated k_j times for the jth variable surrounded by arbitrary sequences of letters and hole characters, with the option of having some letters in the substituted words be replaced by $\diamond$'s. When a letter in a word is replaced in every instance by $\diamond$, that letter practically no longer exists, and we treat it like $\diamond$. Thus all the occurrences of p as a factor are described by

$$\widehat{\mathcal{O}} = \mathrm{SEQ}(\mathcal{A} + \{\diamond\}) \times \prod_{j=1}^{r} \mathrm{SEQ}(\{\diamond^{k_j}\} + \mathcal{A} \circ [(\mathcal{Z} + \{\diamond\})^{k_j} \backslash \{\diamond^{k_j}\}]) \backslash \{\varepsilon\} \times \mathrm{SEQ}(\mathcal{A} + \{\diamond\}),$$

so we get

$$\begin{aligned} \widehat{O}(z) &= \frac{1}{[1-(m+1)z]^2} \prod_{j=1}^{r} \left(\frac{1}{1 - z^{k_j} - m(2^{k_j} - 1)z^{k_j}} - 1 \right) \\ &= \frac{(m+1)^s z^k}{[1-(m+1)z]^{2+s}} \prod_{j=s+1}^{r} \frac{m2^{k_j} - m + 1}{1 - (m2^{k_j} - m + 1)z^{k_j}}. \end{aligned}$$

We have a pole of order $2 + s$ at $z = 1/(m+1)$, and poles at the k_j different k_jth roots of $1/(m2^{k_j} - m + 1)$ for $k_j \geq 2$. Those poles have modulus greater than $1/(m+1)$ by Lemma 1. The singularity at $z = 1/(m+1)$ dominates because it is closest to the origin,

so by Theorem 3.1,

$$\widehat{O}(z) \sim \frac{(m+1)^{s-k}}{[1-(m+1)z]^{2+s}} \prod_{j=s+1}^{r} \frac{m2^{k_j}-m+1}{1-(m2^{k_j}-m+1)/(m+1)^{k_j}}$$
$$= \frac{1}{[1-(m+1)z]^{2+s}} \prod_{j=s+1}^{r} \frac{m2^{k_j}-m+1}{(m+1)^{k_j}-(m2^{k_j}-m+1)}.$$

Taking the coefficient of z^n in the Taylor expansion, we get

$$[z^n]\widehat{O}(z) \sim \binom{n+s+1}{s+1}(m+1)^n \prod_{j=s+1}^{r} \frac{m2^{k_j}-m+1}{(m+1)^{k_j}-(m2^{k_j}-m+1)}$$
$$\sim \frac{n^{s+1}(m+1)^n}{(s+1)!} \prod_{j=s+1}^{r} \frac{m2^{k_j}-m+1}{(m+1)^{k_j}-(m2^{k_j}-m+1)}.$$

Therefore, the mean number of occurrences of a pattern p in a partial word of length n over an alphabet of m letters is

$$\widehat{\Omega}_n \sim \frac{n^{s+1}}{(s+1)!} \prod_{j=s+1}^{r} \frac{m2^{k_j}-m+1}{(m+1)^{k_j}-(m2^{k_j}-m+1)}.$$

□

To illustrate Theorem 5.1, consider the pattern $p = abacaba$, where $r = 3$, $s = 1$, $k_1 = 1$, $k_2 = 2$, and $k_3 = 4$. Substituting these variables, $m = 12$, and $n = 100$, we find that

$$\widehat{\Omega}_{100} \approx 8.9384\cdots.$$

When $\widehat{\Omega}_n < 1$, we may apply Theorem 3.2, so we get

Corollary 5.2. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $k = |p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. If*

$$n < (1+o(1)) \left[(s+1)! \prod_{j=s+1}^{r} \left(\frac{(m+1)^{k_j}}{m2^{k_j}-m+1} - 1 \right) \right]^{\frac{1}{s+1}},$$

there is a partial word of length n over an alphabet of m letters that avoids p.

Remark 1. *When a partial word avoids a pattern, all of its completions also do, so the above results also apply to the case of strictly partial words of a given length. We get the same asymptotics because the subtracted term has strictly lower order asymoptotically than the dominant first-order term.*

5.2. **Mean over all partial words of a given length with a given hole density.** For all terms and notations not defined here, we refer the reader to the book of Pemantle and Wilson [28, pp. 120, 127, 135, 143–145, 154, 174, 177, 192, 198, 336, 341].

Lemma 2 will help us compare the distances of poles of the bivariate version of the generating function from the origin and establish one of them as the dominant singularity.

Lemma 2. *Let $H_1 = 1 - (m+u)z$, and more generally let $H_j = 1 - [m(1+u)^j - mu^j + u^j]z^j$ for integers $j \geq 2$. Let B_1 be the component of $\mathbb{R}^2 \backslash \mathtt{amoeba}(H_1)$ containing a ray $(-\infty, b] \cdot (1,1)$, and more generally let B_j be the component of $\mathbb{R}^2 \backslash \mathtt{amoeba}(H_j)$ containing a ray $(-\infty, b] \cdot (1,1)$ for integers $j \geq 2$. Then*

$$\partial B_1 = \{(-\log(m+u), \log u) : u \in (0,\infty)\},$$

$$\partial B_j = \left\{ \left(-\frac{1}{j} \log[m(1+u)^j - (m-1)u^j], \log u \right) : u \in (0,\infty) \right\}$$

for $j \geq 2$, and $B_1 \subset B_j$ for all $j \geq 2$.

Proof. First, note that

$$\mathtt{amoeba}(H_1) = \{(-\log|m+u|, \log|u|) : u \in \mathbb{C}\}$$

and

$$\mathtt{amoeba}(H_j) = \left\{ \left(-\frac{1}{j} \log|m(1+u)^j - (m-1)u^j|, \log|u| \right) : u \in \mathbb{C} \right\}$$

for integers $j \geq 2$. Since the polynomial $m+u$ has all positive coefficients,

$$\log|m+u| \leq \log(m+|u|)$$

and

$$\partial B_1 = \{(-\log(m+u), \log u) : u \in (0,\infty)\}.$$

More generally, the polynomial $m(1+u)^j - (m-1)u^j$ has all positive coefficients, so $\log|m(1+u)^j - (m-1)u^j| \leq \log[m(1+|u|)^j - (m-1)|u|^j]$ and

$$\partial B_j = \left\{ \left(-\frac{1}{j} \log[m(1+u)^j - (m-1)u^j], \log u \right) : u \in (0,\infty) \right\}$$

for $j \geq 2$. For any parameter $u \in (0,\infty)$, the corresponding points on ∂B_1 and ∂B_j lie on the same horizontal line. However, the power means inequality gives us

$$\sqrt[j]{\tfrac{(m+u)^j + (m-1)u^j}{m}} > \tfrac{(m+u)+(m-1)u}{m} = 1+u$$

which implies

$$-\log(m+u) < -\frac{1}{j} \log[m(1+u)^j - (m-1)u^j],$$

so ∂B_1 lies strictly to the left of ∂B_j and $B_1 \subset B_j$ for all $j \geq 2$. □

We now calculate the mean number of occurrences of a pattern in a partial word with a given length and hole density. This requires the construction of a bivariate generating function, where a second variable, u, marks the number of $\diamond$'s in a partial word.

Theorem 5.3. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $k = |p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. Then the mean number of occurrences of p in a partial word of length n with hole density $d \in \mathbb{Q} \cap (0,1)$ over an alphabet of m letters is*

$$\widehat{\Omega}_{n,d} \sim \frac{n^{s+1}}{(s+1)!} \prod_{j=s+1}^{r} \frac{[1+d(m-1)]^{k_j} - \left(1-\frac{1}{m}\right)(md)^{k_j}}{m^{k_j-1} - [1+d(m-1)]^{k_j} + \left(1-\frac{1}{m}\right)(md)^{k_j}}.$$

Proof. Marking each $\diamond$ with the variable u,

$$\widehat{\mathcal{O}} = \mathrm{SEQ}(\mathcal{A} + \{\diamond\}) \times \prod_{j=1}^{r} \mathrm{SEQ}(\{\diamond^{k_j}\} + \mathcal{A} \circ [(\mathcal{Z} + \{\diamond\})^{k_j} \backslash \{\diamond^{k_j}\}]) \backslash \{\varepsilon\} \times \mathrm{SEQ}(\mathcal{A} + \{\diamond\}),$$

becomes

$$\begin{aligned} \widehat{O}(z,u) &= \frac{1}{(1-mz-uz)^2} \prod_{j=1}^{r} \left(\frac{1}{1 - u^{k_j} z^{k_j} - m[(1+u)^{k_j} - u^{k_j}] z^{k_j}} - 1 \right) \\ &= \frac{(m+u)^s z^k}{[1-(m+u)z]^{2+s}} \prod_{j=s+1}^{r} \frac{m(1+u)^{k_j} - mu^{k_j} + u^{k_j}}{1 - [m(1+u)^{k_j} - mu^{k_j} + u^{k_j}] z^{k_j}}. \end{aligned}$$

For convenience, write $G = (m+u)^s z^k \prod_{j=s+1}^{r} [m(1+u)^{k_j} - mu^{k_j} + u^{k_j}]$ and $H = H_1^{2+s} \prod_{j=s+1}^{r} H_{k_j}$, where $H_1 = 1 - (m+u)z$ and $H_j = 1 - [m(1+u)^j - mu^j + u^j] z^j$, so that $\widehat{O}(z,u) = \frac{G}{H}$.

Note that $\widehat{O} = \frac{G}{H}$ is singular where H is zero, i.e. on the singular variety

$$\mathcal{V} := \mathcal{V}_H = \left\{ (z,u) \in \mathbb{C}^2 : H_1^{2+s} \prod_{j=s+1}^{r} H_{k_j} = 0 \right\} = \mathcal{V}_1 \cup \left(\bigcup_{j=s+1}^{r} \mathcal{V}_{k_j} \right)$$

where we define $\mathcal{V}_1 := \{(z,u) \in \mathbb{C}^2 : 1 - (m+u)z = 0\}$ and $\mathcal{V}_j := \{(z,u) \in \mathbb{C}^2 : 1 - [m(1+u)^j - mu^j + u^j] z^j = 0\}$. Taking the log-modulus gives us the associated amoebas, $\mathtt{amoeba}(H) := \mathtt{amoeba}(H_1) \cup \left(\bigcup_{j=s+1}^{r} \mathtt{amoeba}(H_{k_j}) \right)$, where

$$\mathtt{amoeba}(H_1) = \{(-\log|m+u|, \log|u|) : u \in \mathbb{C}\}$$

and

$$\mathtt{amoeba}(H_j) = \left\{ \left(-\frac{1}{j} \log|m(1+u)^j - (m-1)u^j|, \log|u| \right) : u \in \mathbb{C} \right\}.$$

Let B be the component of $\mathbb{R}^2 \backslash \mathtt{amoeba}(H)$ containing a ray $(-\infty, b] \cdot (1,1)$. By Lemma 2 we know that $B = B_1$, where B_1 is the component of $\mathbb{R}^2 \backslash \mathtt{amoeba}(H_1)$ containing a ray

$(-\infty, b] \cdot (1,1)$. The strata are

$$S_1 = \mathcal{V}_1 \setminus \bigcup_{j=s+1}^{r} \mathcal{V}_{k_j},$$

$$S_{k_j} = \mathcal{V}_{k_j} \setminus \bigcup_{k_i \neq k_j} \mathcal{V}_{k_i},$$

and intersections of $\mathcal{V}_{k_i}$ and $\mathcal{V}_{k_j}$. By Lemma 2, only the critical points of S_1 may have log-moduli on

$$\partial B = \partial B_1 = \{(-\log(m+u), \log u) : u \in (0,\infty)\}.$$

The critical point on S_1 is described by the critical point equations:

$$\begin{aligned} 1 - (m+u)z &= 0 \\ -hz(m+u) &= -nuz. \end{aligned}$$

The solution to the above system of equations is $(z_*, u_*) = \left(\frac{n-h}{mn}, \frac{hm}{n-h}\right)$, and its log-modulus lies on ∂B. Since $\mathbf{x}_{\min} = \operatorname{Re}\log\left(\frac{n-h}{mn}, \frac{hm}{n-h}\right)$ is the unique minimizer in ∂B for $h = h_{\hat{\mathbf{r}}}$, i.e. $\mathbf{x}_{\min}$ minimizes $-\hat{\mathbf{r}} \cdot \mathbf{x}$, and the singleton set containing the critical point $E = \left\{\left(\frac{n-h}{mn}, \frac{hm}{n-h}\right)\right\} \subseteq \mathbf{T}(\mathbf{x}_{\min})$ is a finite nonempty set of quadratically nondegenerate smooth points, the intersection cycle

$$\sigma = \left[\sum_{\mathbf{z} \in W} \mathcal{C}(\mathbf{z})\right]$$

is the sum of quasi-local cycles $\mathcal{C}(\mathbf{z})$ for $\mathbf{z} \in E$, where $\mathcal{C}(\mathbf{z})$ is a homology generator of

$$(\mathcal{V}^{h(\mathbf{x}_{\min})+\varepsilon}, \mathcal{V}^{h(\mathbf{x}_{\min})-\varepsilon}),$$

for example the descending submanifold.

We get asymptotics, so $[z^n u^h]\widehat{O}(z,u)$

$$\sim (2\pi)^{\frac{1-2}{2}} \binom{-h}{s+1} (\det \mathcal{H}_1)^{-1/2} \cdot \frac{\left[G \Big/ \prod_{j=s+1}^{r} H_{k_j}\right]_{(z,u)=\left(\frac{n-h}{mn}, \frac{hm}{n-h}\right)}}{\left(\frac{hm}{n-h}\right)^{s+2} \left(-\frac{n-h}{mn}\right)^{s+2}} h^{\frac{1-2}{2}} \left(\frac{n-h}{mn}\right)^{-n} \left(\frac{hm}{n-h}\right)^{-h}$$

$$\sim \frac{1}{\sqrt{2\pi}} \frac{(-h)^{s+1}}{(s+1)!} \frac{-h}{\sqrt{n(n-h)}} \cdot \frac{\left[G \Big/ \prod_{j=s+1}^{r} H_{k_j}\right]_{(z,u)=\left(\frac{n-h}{mn}, \frac{hm}{n-h}\right)}}{\left(-\frac{h}{n}\right)^{s+2}} \frac{m^{n-h}}{\sqrt{h}} \left(1 - \frac{h}{n}\right)^{h-n} \left(\frac{n}{h}\right)^{h},$$

where

$$\begin{aligned}\det \mathcal{H}_1 &= \frac{Q}{(-uH_{1u})^3} \\ &= \frac{-u^2z^2z(-m-u) - u(-z)z^2(-m-u)^2 - z^2u^2(-2)(-m-u)(-z)(-1)}{[(-u)(-z)]^3} \\ &= \frac{n(n-h)}{h^2} \text{ at } u = \frac{hm}{n-h},\end{aligned}$$

and where $\left[G \Big/ \prod_{j=s+1}^{r} H_{k_j}\right]_{(z,u)=\left(\frac{n-h}{mn}, \frac{hm}{n-h}\right)}$

$$\begin{aligned}&= \left[\left(\frac{mn}{n-h}\right)^s \left(\frac{n-h}{mn}\right)^k\right] \cdot \prod_{j=s+1}^{r} \frac{m\left(\frac{n-h+hm}{n-h}\right)^{k_j} - (m-1)\left(\frac{hm}{n-h}\right)^{k_j}}{1 - \left[m\left(\frac{n-h+hm}{n-h}\right)^{k_j} - (m-1)\left(\frac{hm}{n-h}\right)^{k_j}\right]\left(\frac{n-h}{mn}\right)^{k_j}} \\ &= \prod_{j=s+1}^{r} \frac{m(n-h+hm)^{k_j} - (m-1)(hm)^{k_j}}{(mn)^{k_j} - m(n-h+hm)^{k_j} + (m-1)(hm)^{k_j}}.\end{aligned}$$

Substituting the latter quantity, we get

$$\begin{aligned}&[z^n u^h]\widehat{O}(z,u) \sim \frac{1}{\sqrt{2\pi}} \frac{(-h)^{s+1}}{(s+1)!} \frac{-h}{\sqrt{n(n-h)}} \left(-\frac{n}{h}\right)^{s+2} \frac{m^{n-h}}{\sqrt{h}} \left(1 - \frac{h}{n}\right)^{h-n} \frac{n^h}{h^h} \cdot \\ &\prod_{j=s+1}^{r} \frac{m(n-h+hm)^{k_j} - (m-1)(hm)^{k_j}}{(mn)^{k_j} - m(n-h+hm)^{k_j} + (m-1)(hm)^{k_j}} \\ &= \frac{m^{n-h} n^{s+1} (1-d)^{n(d-1)-\frac{1}{2}}}{(s+1)!\sqrt{2\pi n d}\, d^{nd}} \prod_{j=s+1}^{r} \frac{[1+d(m-1)]^{k_j} - (1-1/m)(md)^{k_j}}{m^{k_j-1} - [1+d(m-1)]^{k_j} + (1-1/m)(md)^{k_j}},\end{aligned}$$

where we let $d = h/n$ denote the density of holes.

Since the total number of partial words of length n with h holes over an alphabet of m letters is, by Stirling's approximation,

$$\begin{aligned}\binom{n}{h} m^{n-h} &\sim m^{n-h} \left(\frac{n}{h}\right)^h \left(\frac{n}{n-h}\right)^{n-h} \sqrt{\frac{2\pi n}{(2\pi h)2\pi(n-h)}} \\ &= \frac{m^{n-h}}{\sqrt{2\pi n d(1-d)}\,[d^d(1-d)^{1-d}]^n},\end{aligned}$$

we find that the mean number of occurrences of p in a partial word of length n with hole density d over an alphabet of m letters is

$$\widehat{\Omega}_{n,d} \sim \frac{n^{s+1}}{(s+1)!} \prod_{j=s+1}^{r} \frac{[1+d(m-1)]^{k_j} - \left(1 - \frac{1}{m}\right)(md)^{k_j}}{m^{k_j-1} - [1+d(m-1)]^{k_j} + \left(1 - \frac{1}{m}\right)(md)^{k_j}}.$$

□

To illustrate Theorem 5.3, consider the pattern $p = abacaba$, where $r = 3$, $s = 1$, $k_1 = 1$, $k_2 = 2$, and $k_3 = 4$. Substituting these variables, $m = 12$, $n = 100$, and $d = 1/10$ we find that

$$\widehat{\Omega}_{100,1/10} \approx 17.788\cdots.$$

When $\widehat{\Omega}_{n,d} < 1$, we may apply Theorem 3.2, so in that case we obtain the following corollary.

Corollary 5.4. *Suppose that a pattern p uses r distinct variables, where the jth variable occurs $k_j \geq 1$ times. Without loss of generality, let $k = |p| = k_1 + \cdots + k_r$ and $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. If*

$$n < (1+o(1)) \left[(s+1)! \prod_{j=s+1}^{r} \left(\frac{m^{k_j-1}}{[1+d(m-1)]^{k_j} - \left(1-\frac{1}{m}\right)(md)^{k_j}} - 1 \right) \right]^{\frac{1}{s+1}}$$

there is a partial word of length n with hole density d over an alphabet of m letters that avoids p.

The upper bound for $L(m, Z_i)$ in Theorem 4.5 also applies to $L_d(m, Z_i)$. For a lower bound, we get the following.

Corollary 5.5.

$$L_d(m, Z_i) \geq (1+o(1)) \sqrt{2 \prod_{j=1}^{i-1} \left(\frac{m^{2^j-1}}{[1+d(m-1)]^{2^j} - \left(1-\frac{1}{m}\right)(md)^{2^j}} - 1 \right)}.$$

Substituting for example $m = 12$, $d = 1/10$, and $i = 3$, we find that

$$L_{1/10}(12, Z_3) \geq 23.709\cdots.$$

6. Conclusion and open problems

Using techniques from analytic combinatorics, we have calculated asymptotic mean pattern occurrence and used these statistics in conjunction with the probabilistic method to establish new results about Ramsey theoretic pattern avoidance in the abelian full word case and the nonabelian partial word case. We have established, in particular, lower bounds for Ramsey lengths.

However, there may be more possible uses of these data in applications such as cryptography and musicology. Cryptanalysts may compare the pattern occurrence statistics of possible ciphertexts to those of random noise to detect the existence of hidden messages; see the definitions of semantic security and pseudorandom generator in [21, pp. 67, 70]. Musicologists may compare the pattern occurrence statistics of different musical compositions to further their understanding of musical forms; for previous work connecting music theory and theoretical computer science see [24].

We propose the following open problems.

Problem 1. *Can you adapt the techniques appearing in this paper to the following two cases, which we have not considered, and get similar results?*

- *When can a partial word avoid a given pattern in the abelian sense?*
- *When can a full necklace avoid a given pattern?*

Problem 2. *Can you find better lower and upper bounds for the Ramsey lengths $L(m, p)$, $L_d(m, p)$, and $L_{\text{ab}}(m, p)$ than the ones appearing in this paper?*

Acknowledgements

I thank Andrew Lohr from Rutgers University and Brent Woodhouse from The University of California at Los Angeles for their very valuable comments and suggestions. I thank Francine Blanchet-Sadri for giving me support to work under her supervision.

References

[1] S. Adjan and P. S. Novikov. Infinite periodic groups I–III. *Izv. Akad. Nauk SSSR Ser. Mat.*, 32, 1968.

[2] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., New York, 1992. With an appendix by Paul Erdős, A Wiley-Interscience Publication.

[3] F. Blanchet-Sadri, B. De Winkle, and S. Simmons. Abelian pattern avoidance in partial words. Submitted to RAIRO-Theoretical Informatics and Applications.

[4] F. Blanchet-Sadri, J. I. Kim, R. Mercaş, W. Severa, S. Simmons, and D. Xu. Avoiding abelian squares in partial words. *Journal of Combinatorial Theory, Series A*, 119:257–270, 2012. (`www.uncg.edu/cmp/research/abelianrepetitions`).

[5] F. Blanchet-Sadri, A. Lohr, and S. Scott. Computing the partial word avoidability indices of binary patterns. *Journal of Discrete Algorithms*, 23:113–118, 2013.

[6] F. Blanchet-Sadri, A. Lohr, and S. Scott. Computing the partial word avoidability indices of ternary patterns. *Journal of Discrete Algorithms*, 23:119–142, 2013.

[7] F. Blanchet-Sadri, R. Mercaş, S. Simmons, and E. Weissenstein. Avoidable binary patterns in partial words. *Acta Informatica*, 48(1):25–41, 2011.

[8] F. Blanchet-Sadri, S. Simmons, and D. Xu. Abelian repetitions in partial words. *Advances in Applied Mathematics*, 48:194–214, 2012.

[9] F. Blanchet-Sadri and B. Woodhouse. Strict bounds for pattern avoidance, 2013.

[10] Arnaud Carayol and Stefan Göller. On Long Words Avoiding Zimin Patterns. In Heribert Vollmer and Brigitte Vallelé, editors, *34th Symposium on Theoretical Aspects of Computer Science (STACS 2017)*, volume 66 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 19:1–19:13, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[11] Joshua Cooper and Danny Rorabaugh. Bounds on Zimin word avoidance. *Congr. Numer.*, 222:87–95, 2014.

[12] R. Cori and M. Formisano. Partially abelian square-free words. *RAIRO-Theoretical Informatics and Applications*, 24:509–520, 1990.

[13] James D. Currie. Pattern avoidance: themes and variations. *Theoret. Comput. Sci.*, 339(1):7–18, 2005.

[14] A. de Luca and S. Varricchio. Some combinatorial properties of the Thue-Morse sequence and a problem of semigroups. *Theoretical Computer Science*, 63:333–348, 1989.

[15] V. Dickert. Research topics in the theory of free partially commutative monoids. *Bulletin of the European Association for Theoretical Computer Science*, 40:479–491, 1990.

[16] Philippe Flajolet, Eric Fusy, Xavier Gourdon, Daniel Panario, and Nicolas Pouyanne. A hybrid of Darboux's method and singularity analysis in combinatorial asymptotics. *Electron. J. Combin.*, 13(1):Research Paper 103, 35, 2006.

[17] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, 2009.

[18] W. T. Gowers. The two cultures of mathematics. In *Mathematics: Frontiers and Perspectives*, pages 65–78. Amer. Math. Soc., Providence, RI, 2000.

[19] J. Jezek. Intervals in the lattice of varieties. *Algebra Universalis*, 6:147–158, 1976.

[20] J. Justin. Characterization of the repetitive commutative semigroups. *Journal of Algebra*, 21:87–90, 1972.

[21] Jonathan Katz and Yehuda Lindell. *Introduction to Modern Cryptography*. Chapman & Hall/CRC Cryptography and Network Security. Chapman & Hall/CRC, Boca Raton, FL, 2008.

[22] O. Kharlampovich and M. Sapir. Algorithmic problems in varieties, a survey. *International Journal of Algebra and Computation*, 12:379–602, 1995.

[23] Donald E. Knuth. Mathematics and computer science: coping with finiteness. *Science*, 194(4271):1235–1242, 1976.

[24] Donald E. Knuth. The complexity of songs. *Comm. ACM*, 27(4):344–346, 1984.

[25] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, Cambridge, 2002.

[26] M. G. Main, W. Bucher, and D. Haussler. Applications of an infinite squarefree co-CFL. *Theoretical Computer Science*, 49(2–3):113–119, 1987.

[27] M. Morse. Recurrent geodesics on a surface of negative curvature. *Transactions of the American Mathematical Society*, 22:84–100, 1921.

[28] Robin Pemantle and Mark C. Wilson. *Analytic Combinatorics in Several Variables*. Cambridge University Press, Cambridge, 2013.

[29] F. P. Ramsey. On a Problem of Formal Logic. *Proc. London Math. Soc.*, S2-30(1):264, 1930.

[30] L. B. Richmond and Jeffrey Shallit. Counting abelian squares. *Electron. J. Combin.*, 16(1):Research Paper 72, 9, 2009.

[31] D. Rorabaugh. Toward the Combinatorial Limit Theory of Free Words. *ArXiv e-prints*, September 2015.

[32] Wojciech Rytter and Arseny M. Shur. Searching for zimin patterns. *Theor. Comput. Sci.*, 571:50–57, 2015.

[33] W. T. Trotter and P. Winkler. Arithmetic progressions in partially ordered sets. *Order*, 4:37–42, 1987.

Department of Mathematics, Caltech MC 253-37, 1200 E California Blvd, Pasadena, California 91125, USA
E-mail address: jtao@caltech.edu