Complete Words and Superpatterns

by

Yonah Biers-Ariel

A thesis submitted in partial fulfillment of the requirements
for graduation with Honors in Mathematics.

Whitman College
2015

Certificate of Approval

This is to certify that the accompanying thesis by Yonah Biers-Ariel has been accepted in partial fulfillment of the requirements for graduation with Honors in Mathematics.

David Guichard, Ph.D.

Whitman College
May 11, 2015

Contents

Abstract

This thesis surveys the most important results regarding complete words, by which we mean words containing as subsequences every permutation of some set, and then considers two variants of the problem of finding the shortest complete words. The first of these asks for the shortest superpattern, which is a complete word with some additional requirements, and the second asks for the expected number of timesteps before a particular random process generates a complete word.

Introduction

Consider a string of elements from the set $[k] = \{1, 2, ..., k\}$. We will call this string a *word* and we will say that it is on the *alphabet* $[k]$. Suppose W is a word on $[k]$, and let w_i denote the i^{th} letter of W. We are interested in the *subsequences* of W. We say that a word T is a subsequence of W if there exists a strictly increasing sequence $x_1, x_2, ..., x_n$ such that $T = w_{x_1} w_{x_2} ... w_{x_n}$. In particular, we are interested in subsequences which are *permutations*. A permutation of the set $[k]$ is a word which contains each letter in $[k]$ exactly once. We illustrate these definitions in the example below.

Example 0.1. The string $W = 1442321$ is a word on the alphabet $[4]$. $T = 1432$ is a subsequence of W because $1, 2, 5, 6$ is a strictly increasing sequence and $T = w_1 w_2 w_5 w_6$. Further, T is a permutation of $[4]$ because it contains each of $1, 2, 3$, and 4 exactly once.

Section 1 considers words on the alphabet $[k]$ which contain each permutation of this set as a subsequence; we call such words $[k]$-*complete* following the terminology of [7]. Knuth proposed the problem of finding the shortest $[k]$-complete words in [2] (although he attributes it to R. M. Karp). The problem attracted significant attention in the 1970s; the length of the shortest $[k]$-complete word was found for $1 \leq k \leq 7$, and general upper and lower bounds were also proven. Since then, the problem has been more or less ignored, although two papers published in the last five years have disproved a long-standing conjecture that the upper bound discovered in the 1970s is sharp. We present no original results in Section 1; rather we simply survey the most important results regarding complete words and identify some interesting open questions.

Section 2 considers *superpatterns*, a variant of complete words introduced by Burstein et al. in [1]. The questions of interest in this section are essentially the same as in Section 1; we want to find upper and lower bounds on the lengths of the shortest $[k]$-superpatterns, and ideally we would like to know those lengths' exact values. This section does contain original work; in particular we prove the surprising result that $[k]$-complete words and $[k]$-superpatterns are actually equivalent. Because much more is known about $[k]$-complete words than $[k]$-superpatterns, this result provides a substantial amount of new information about $[k]$-superpatterns.

Finally, Section 3 considers a random process in which we begin with an empty word and then repeatedly choose a letter uniformly at random from $[k]$ and append it to the word. While no one has yet found an asymptotic value for the length of the shortest $[k]$-complete word, we are able to find an asymptotic value for the expected

number of letters our random process must add before a $[k]$-complete word appears. This process was first considered in [3]; here we provide shortened proofs of two results from that paper before proving the asymptotic result.

1 Complete Words

In this section, we describe the most important results regarding $[k]$-complete words and provide two conjectures for further research. The first significant result regarding $[k]$-complete words comes from [6], although the construction is due to Burstein et al. in [1]. The construction is provided there without proof, and so the proof that the construction is valid is original here.

Theorem 1.1 (Newey)**.** *There exist $[k]$-complete words of length $k^2 - 2k + 4$ for $k \geq 3$.*

Proof: Consider a word $\overline{W}$ formed by be concatenating $k-2$ copies of the string $123...(k-2)(k-1)$ and then placing a final 12 after this concatenation. Now, insert one copy of k at the beginning of $\overline{W}$, and, for $i = 1, 2, ..., k-1$, insert the $(i+1)^{\text{th}}$ copy of k immediately after the i^{th} copy of $k-i$. Call this word W, and note that it has length $(k-1)(k-2) + k + 2 = k^2 - 2k + 4$. As an example, when $k = 5$, W is given by 51234512354125341 52.

Now, we show that W is $[k]$-complete. We will first divide W into $k-2$ blocks, where block i contains the i^{th} copy of $123...(k-1)$ from $\overline{W}$ as well as the $(i+1)^{\text{th}}$ copy of k. Notice that all blocks contain k consecutive letters, and that the first copy of k as well as the final copies of $k, 1$, and 2 are not in any block. Now, fix some

permutation Σ. We will say that Σ has an *ascent* at position i if $\sigma_i > \sigma_{i-1}$ where σ_i is the i^{th} letter of σ.

Let k be the i^{th} letter of Σ; we will first handle the special case of when $i = k$. Now, k is the last letter of Σ, so consider the prefix of Σ preceding k. Suppose this prefix contains an ascent at position j. Then, for $m < j$, we can find σ_m in block m, for j we can find σ_j after σ_{j-1} in block $m-1$, and for $m > j$ we can find σ_m in block σ_{m-1}. Then, we can find the entire prefix by the end of block $k-2$, and since this block is followed by the subword $1k2$ which contains k, Σ is a subsequence of W. Otherwise, this prefix contains no ascents, and so Σ is the string $(k-1)(k-2)(k-3)...321k$; we can find the m^{th} letter in block m for $m \leq k-2$, and we can find the subword $1k$ after block $k-2$. Therefore, Σ is a subsequence of W.

Now, suppose that $i < k$. We will first show that the prefix of Σ ending in k can be completed by the end of the $(i-1)^{\text{th}}$ block if we refer to the k preceding the 1^{st} block as the 0^{th} block. If $i = 1$, then k is the first letter of Σ, and k appears in what we just defined to be the 0^{th} block. Otherwise, suppose that the prefix under consideration contains at least one ascent (other than the one ending with k), and suppose one such ascent occurs at position j. Then, for $m < j$, we can find σ_m in block m, for j we can find σ_j after σ_{j-1} in block $j-1$, and for $m > j$ we can find σ_m in block $m-1$. Therefore, we can find $k = \sigma_i$ in block $i-1$. Otherwise, this prefix contains no ascents (other than the one ending with k). Then, $\sigma_1 \leq k-1$, and for each m with $1 < m < i$ we get $\sigma_m < \sigma_{m-1}$. Thus, $\sigma_{i-1} \leq k-i+1$, and in our construction of W, we put the k in block $i-1$ after all letters which are less than or equal to $k-i+1$. Thus, for each $m < i$, we can find σ_m in block m, and we can find

$k = \sigma_i$ in the $i - 1^{\text{th}}$ block after σ_{i-1}.

Now that we know this, we need to find the postfix of Σ beginning after k as a subsequence of the portion of W which follows block $i-1$. Suppose that this postfix contains an ascent at position j. Then, for $i < m < j$, we can find σ_m in block $m-1$, for j we can find σ_j following σ_{j-1} in block $j-2$, and for $m > j$, we can find σ_m in block $m-2$, so we will have found Σ by the end of block $k-2$. Otherwise, this postfix contains no ascents. If $\sigma_{i+1} > k - i + 1$, then σ_{i+1} occurs after k in block $i-1$ by our construction of W. Therefore, we can find σ_{i+1} after k in block $i-1$, and, for $m > i+1$, we can find σ_m alone in block $m-2$. The final case we must consider is if the postfix contains no ascents and $\sigma_{i+1} \leq k - i + 1$. Then, since each subsequent letter in Σ is at least one less than the letter before, we have that $\sigma_k \leq k - i + 1 - (k - (i+1)) = 2$. Thus, for each $m > i$ we can find σ_m in block $m-1$, and we can find σ_k, which is either 1 or 2, after block $k-2$. Therefore, in every case we have found Σ as a subsequence of W. □

This particular construction assumes that 1,2, and k are distinct letters, and so we must have $k \geq 3$. The result holds for $k = 1, 2$ as well, however, as 1 is a [1]-complete word and 121 is a [2]-complete word, and both words, while shorter than the theorem guarantees, can be lengthened arbitrarily without changing their completeness.

The primary aim of most researchers in this topic is to find the length of the shortest k-complete word, which we will call $\rho(k)$, and so Theorem 1.1 serves as an upper bound on $\rho(k)$. Newey in [6] finds a lower bound equal to this upper bound for $1 \leq k \leq 7$. Here we prove this lower bound for $2 \leq k \leq 5$. Our proof is somewhat different from Newey's in that it demonstrates a lower bound on $\rho(k)$ for all k which

agrees with the upper bound of Theorem 1 when $k = 3, 4$.

Proposition 1.2. *The inequality* $\rho(k) \geq \frac{k^2}{2} + \frac{3k}{2} - 2$ *holds for all* $k \geq 2$.

Proof: We will proceed by induction. As a base case, consider $k = 2$. Then, 121 is a word on [2] containing 12 and 21, but in any two-letter word, either the first letter would greater than or equal to the second, and so 12 would be absent, or else the second letter would be greater than or equal to the first, and so 21 would be absent. Thus, $\rho(2) = 3 = \frac{2^2}{2} + \frac{3 \cdot 2}{2} - 2$.

Now, suppose Proposition 1.2 holds for some $k \geq 2$, and let W be an arbitrary word on the alphabet $[k+1]$ with length $\frac{(k+1)^2}{2} + \frac{3(k+1)}{2} - 3$; we will show that W does not contain all permutations of $[k+1]$. Recall that we denote the first letter of W by w_1, the second letter by w_2 and so on. Clearly, each letter in $[k+1]$ must appear somewhere in W. Let a be the last letter to make its first appearance in W, i.e. let a be the letter whose first appearance in W is preceded by an appearance of every other letter. Then, the first appearance of a occurs at the earliest as the $k+1^{\text{th}}$ letter of W. We will consider two cases: when a first appears as the $k+1^{\text{th}}$ letter of W and when a first appears after the $k+1^{\text{th}}$ letter.

In the first case, the subword $w_1 w_2 ... w_{k+1}$ contains all elements of $[k+1]$ exactly once, so the a appearing as the $k+1^{th}$ letter of W cannot be a part of any permutation beginning with $w_2 w_1 a$. Since $k + 1 \geq 3$, permutations of this form must exist, so a appears later on in W as well. Thus, there are at most $\frac{(k+1)^2}{2} + \frac{3(k+1)}{2} - 3 - (k+2) = \frac{k^2}{2} + \frac{3k}{2} - 3$ letters following the first a which are not a. However, W contains all permutations of $[k+1]$ beginning with a, so it must contain all permutations of

$[k+1]\backslash\{a\}$ following the first a. But, by the induction hypothesis, $\frac{k^2}{2} + \frac{3k}{2} - 3$ are insufficiently many letters to contain all the permutations of k letters.

In the second case, a first occurs at the earliest as the $k+2^{\text{th}}$ letter of W, so it has at most $\frac{(k+1)^2}{2} + \frac{3(k+1)}{2} - 3 - (k+2) = \frac{k^2}{2} + \frac{3k}{2} - 3$ letters following it. As before, W must contain all permutations of $[k+1]\backslash\{a\}$ following the first a, but, again, $\frac{k^2}{2} + \frac{3k}{2} - 3$ are insufficiently many letters to contain all the permutations of k letters. Thus, W does not contain all permutations of $[k+1]$, and so we have arrived at a contradiction, and, by induction, $\rho(k) \geq \frac{k^2}{2} + \frac{3k}{2} - 2$ for all $k \geq 2$. □

As noted, this bound is equivalent to $k^2 - 2k + 4$ for $k = 3$ and 4. Next, we will next consider $k = 5$.

Proposition 1.3. *The equality* $\rho(5) = 19$ *is valid.*

Proof: Theorem 1.1 gives $\rho(5) \leq 19$, and Proposition 1.2 gives $\rho(5) \geq 18$, so suppose $\rho(5) = 18$, and let W be an 18-letter word on the alphabet [5] containing all permutations of [5]. As in the previous proof, we look at the last letter to make its first appearance in W. Let this letter be a and note again that a appears at the earliest as the 5^{th} letter of W. We will examine three cases.

First, suppose there is only one a in W; then W has a subword containing all permutations of $[5]\backslash\{a\}$ on either side of a. Each such subword contains at least $\rho(4) = 12$ letters, and so W contains 25 letters in total, which contradicts our assumption that W contained just 18 letters.

Next, suppose there is more than one copy of a, and either a first appears after the 5^{th} letter of W, or there are more than two copies of a. Then, since W has 18

letters, there are fewer than 12 letters following the first a which are not a, and so the portion of W following the first a does not contain all permutations of $[5]\backslash\{a\}$, and W does not contain all permutations which begin with a.

Finally, suppose there are two copies of a, and one is the 5^{th} letter of W, and consider the position of the last copy. It must have 13 letters (including the first a) before it in order for W to contain every permutation ending in a, and so the final a has at most 4 letters following it. Then, we can find a two-letter word with distinct letters on $[5]\backslash\{a\}$ that does not occur after this a. Suppose this word is bc; let d, e be the two letters in $[5]\backslash\{a, b, c\}$. Both d and e occur exactly once before the first copy of a in W, so without loss of generality suppose d occurs before e. Then, the permutation $edabc$ does not occur in W; it cannot occur using the first a of W because that a is not preceded by ed, and it cannot occur using the second a of W because that a is not followed by bc. In every case, we have a contradiction, so $\rho(5) = 19$. $\square$

Since the upper bound of Theorem 1 is sharp for $1 \leq k \leq 7$, it was conjectured for more than thirty years that this bound was, in fact, the true value of $\rho(k)$ for all k. This conjecture has been shown to be false by anonymous counterexamples for $k = 10$ and 11, and the conjectured bound is shown to be false for all $k \geq 8$ by Zălinescu in [8]. The upper bound that this paper established, though, is as little an improvement as is possible; it shows that for all $k \geq 8$, there exist $[k]$-complete words of length $k^2 - 2k + 3$.

A more significant improvement comes from Radomirović in [7], and he makes this improvement using two lemmas which are interesting in their own right. Before we can give these results, however, we need some new notation. The notation follows

that of [7] except where noted. Suppose $[k]$ is an alphabet, W is a word on that alphabet, and $c, d \in [k]$. Then, $W_{c^i \ldots d^j}$ is the subword of W strictly between the i^{th} copy of c and the j^{th} copy of d. Similarly, $W_{\ldots c^i}$ is the prefix of W before the i^{th} copy of c (not including that copy of c), and $W_{d^j \ldots}$ is the suffix of W following the j^{th} copy of d (not including that copy of d). Also, c^f is understood to mean the final copy of c in a given word W.

Further, $\#_c W$ is the number of copies of c in W. Also, $\mathcal{P}_W$ is the set of all permutations of the set of elements of W (ignoring repeated elements); this is the one deviation from the notation of [7] as that paper uses $[W]$ to denote the set of permutations of elements of W. Finally, if U, V, W are all words, then (U, W, V) is the word formed by concatenating U, W, and V, while $\{(U, \mathcal{P}_W, V)\}$ is the set of all words which can be formed by concatenating U, some permutation of the set of elements of W, and V. To make this notation clearer, we provide the following examples.

Example 1.4. Suppose W is given by 1231231. Then, $W_{\ldots 2^2} = 1231$, $W_{1^2 \ldots} = 231$, and $W_{2^1 \ldots 3^2} = 312$.

Example 1.5. Suppose W is given by 121, U is given by 123, and V is given by 321. Then, $\#_1 W = 2, \#_2 W = 1$, and $\mathcal{P}_W = \{12, 21\}$. Also, $\{(U, \mathcal{P}_W, V)\} = \{12312321, 12321321\}$.

Now, we are ready to state the first lemma. Informally, it tells us that, given consecutive occurrences of a letter c in a $[k]$-complete word, we can permute all of the elements between these occurrences in any way we choose and still have a word which is $[k]\backslash\{c\}$-complete. Because a $[k]\backslash\{c\}$-complete word must not contain any

c's, we technically must delete all the c's from the resulting word to make it $[k]\backslash\{c\}$ complete, but these deletions clearly have no impact on the permutations of $[k]\backslash\{c\}$ that the word contains.

Lemma 1.6 (Radomirović)**.** *Let $c \in [k]$. If W is a $[k]$-complete word, then for every $1 \leq i < \#_c W$, every word in $\{(W_{...c^i}, \mathcal{P}_{W_{c^i...c^{i+1}}}, W_{c^{i+1}...})\}$ is, once all c's are deleted from it, a $[k]\backslash\{c\}$-complete word.*

Proof: Fix some $i \in \{1, ..., \#_c W - 1\}$. Let $W_1 = W_{...c^i}$, choose some $W_2 \in \mathcal{P}_{W_{c^i...c^{i+1}}}$, and let $W_3 = W_{c^{i+1}...}$. Let $V = (W_1, W_2, W_3)$, and consider some permutation $\Phi = d_1 d_2 ... d_{k-1}$ over $[k]\backslash\{c\}$. Let j be the greatest integer such that $d_1 d_2 ... d_j$ is a subsequence of W_1. We now consider two cases: when $d_{j+1} \in W_2$, and when $d_{j+1} \notin W_2$.

Case 1: Suppose $d_{j+1} \in W_2$. Then, let $U = W_{c^i...c^{i+1}}$ and note that $W = (W_1, c, U, c, W_3)$. By assumption, $d_{j+1} \in W_2$, and so $d_{j+1} \in U$. Then, since W is $[k]$-complete, it contains $d_1 d_2 ... d_{j+1} c d_{j+2} ... d_{k-1}$, and since W_1 does not contain $d_1 d_2 ... d_{j+1}$, the shortest prefix of W containing $d_1 d_2 ... d_{j+1} c$ is (W_1, c, U, c). Therefore, W_3 contains $d_{j+2} d_{j+3} ... d_{k-1}$, and so (W_1, d_{j+1}, W_3), which is a subsequence of V, contains Φ, and so V does as well.

Case 2: Suppose $d_{j+1} \notin W_2$. Then, since W is $[k]$-complete, it is certainly $[k]\backslash\{c\}$-complete, and so W contains Φ. Since $d_{j+1} \notin W_2$, it follows that $d_{j+1} d_{j+2} ... d_{k-1}$ is contained in W_3. Therefore, (W_1, W_3), which is a subsequence of V, contains Φ, and so V does as well. Therefore, V is $[k]\backslash\{c\}$ complete. □

Example 1.7. Consider the word 412341243142 which is [4]-complete by Theorem

1.1. We now choose the 1st and 2nd 1's, and apply Lemma 1.6. This gives us the that the following words are $[4]\backslash\{1\}$-complete: 423424342, 424324342, 432424342, 434224342, 442324342, 443224342. Note that several of these words now have consecutive letters which are identical; we can delete a copy of each such repeated letter to obtain, for instance, that 4324342 is $[4]\backslash\{1\}$-complete (we simply deleted one of the repeated 4's and one of the repeated 2's from 443224342).

The next lemma establishes two conditions which together are necessary and sufficient to show that a word is $[k]$-complete. This is the primary tool used in [7] to prove that the paper's construction is $[k]$-complete.

Lemma 1.8 (Radomirović)**.** *Let $c \in [k]$. A word W is $[k]$-complete if and only if both of the following hold:*

(1) $W_{c^1\ldots}$ and $W_{\ldots c^f}$ are $[k]\backslash\{c\}$-complete;

(2) For every $1 \leq i < \#_c W$ every word in $\{(W_{\ldots c^i}, \mathcal{P}_{W_{c^i\ldots c^{i+1}}}, W_{c^{i+1}\ldots})\}$ is $[k]\backslash\{c\}$-complete after all copies of c are removed.

Proof: Fix $c \in [k]$. Now, let W be a $[k]$-complete word. Then, W contains every permutation of $[k]$ of the form $cd_1d_2...d_{k-1}$ where the d_i are all the elements of $[k]\backslash\{c\}$, and it follows that $W_{c^1\ldots}$ must contain each possible arrangement of the d_i, i.e. every permutation of $[k]\backslash\{c\}$. Similarly, the fact that W contains every permutation of $[k]$ of the form $d_1d_2...d_{k-1}c$ implies that $W_{\ldots c^f}$ is $[k]\backslash\{c\}$ complete, and so condition (1) holds for any $[k]$-complete word. Meanwhile, Lemma 1.6 states that the condition (2) holds for any $[k]$-complete word.

Now, suppose that conditions (1) and (2) hold for some word W; we will show that

W is $[k]$-complete. Let $\Phi = d_1 d_2 ... d_k$ be a permutation of $[k]$; we will find Φ as a subsequence of W in each of 3 cases:

Case 1: Suppose c is the first letter of Φ. By condition (1), $W_{c^1...}$ contains $d_2 d_3 ... d_k$, and so W contains $c d_2 d_3 ... d_k = \Phi$.

Case 2: Suppose c is the last letter of Φ. By condition (1), $W_{...c^f}$ contains $d_1 d_2 ... d_{k-1}$, and so W contains $d_1 d_2 ... d_{k-1} c$.

Case 3: Suppose c is neither the first nor the last letter of Φ. If c appears only once in W, then we are done again by condition (1). Otherwise, c appears at least twice in W. Suppose $d_j = c$, noting that we must have $1 < j < k$, and write $W = (W_1, W_2)$ where W_1 is the shortest prefix of W which contains $d_1 d_2 ... d_{j-1}$ as a subsequence. Such a prefix must exist because $d_1 d_2 ... d_{j-1}$ is a word on $[k] \backslash \{c\}$ with no repeated letters, and so must be contained in any $[k] \backslash \{c\}$-complete word, which W contains by condition (1). If c does not appear in W_1, then we are done by condition (1). Otherwise, we can let $W_1 = (W_1', c, W_1'')$ where W_1'' does not contain c. Similarly, we can write $W_2 = (W_2', c, W_2'')$ where c does not appear in W_2'. Since $W = (W_1', c, W_1'', W_2', c, W_2'')$, each sequence in $S = \{(W_1', \mathcal{P}_{(W_1'', W_2')}, W_2'')\}$ is $[k] \backslash \{c\}$-complete by condition (2).

Now, the last letter of W_1 is d_{j-1}, and so W_1'' contains d_{j-1}, and we can find a $\pi \in \mathcal{P}_{(W_1'', W_2')}$ such that the final letter of π is d_{j-1}. Then, $(W_1', \pi, W_2'') \in S$ and so it contains $d_1 d_2 ... d_{j-1} d_{j+1} ... d_k$. Further, W_1' is strictly shorter than W_1 which was the shortest prefix of W containing $d_1 d_2 ... d_{j-1}$, and so it follows that W_1' does not contain $d_1 d_2 ... d_{j-1}$. Then, π is a permutation, and so it contains d_{j-1} just once - as its last element. Therefore, (W_1', π) does not contain $d_1 d_2 ... d_{j-1} d_{j+1}$, and so W_2'' must contain $d_{j+1} d_{j+2} ... d_k$. Recall that $W = (W_1, W_2', c, W_2'')$; since W_1 contains

$d_1 d_2 ... d_{j-1}$, c is d_j, and W_2'' contains $d_{j+1} d_{j+2} ... d_k$, it follows that W contains Φ, and so W is $[k]$-complete. $\square$

This lemma allows for the construction of a $[k]$-complete word of length $\lceil k^2 - \frac{7}{3}k + \frac{19}{3} \rceil$ for every $k > 4$, which is an improvement over the $k^2 - 2k + 4$ construction for all $k \geq 10$. The proof of the final theorem is beyond the scope of this thesis.

Theorem 1.9 (Radomirović)**.** *Suppose $k > 4$. Then, there exists a $[k]$-complete word of length $\lceil k^2 - \frac{7}{3}k + \frac{19}{3} \rceil$.*

As noted, this bound shows that $\rho(k) < k^2 - 2k + 4$ for all $k \geq 10$. Unfortunately, since $\rho(k) = k^2 - 2k + 4$ for at least three k values (indeed, the equality actually holds for all $3 \leq k \leq 7$), if $\rho(k) \neq k^2 - 2k + 4$, then $\rho(k)$ is not a quadratic. Moreover, because $\rho(k)$ is bounded above and below by the quadratics $k^2 - 2k + 4$ and $\frac{k^2}{2} + \frac{3k}{2} - 2$ respectively, $\rho(k)$ is not a polynomial at all. In [6], the first paper to conjecture that $\rho(k) = k^2 - 2k + 4$, Newey also provides an alternative conjecture as to the value of $\rho(k)$. This second conjecture was largely ignored by most of the subsequent literature because it is much messier, but, now that we know that $\rho(k)$ is not a polynomial, it is worth revisiting.

Conjecture 1.10 (Newey)**.** *The following is a valid formula for $\rho(k)$:*

$$\rho(k) = \begin{cases} k^2 & \text{for } k = 1 \\ k^2 - k + 1 & \text{for } 2 \leq k \leq 3 \\ k^2 - 2k + 4 & \text{for } 4 \leq k \leq 7 \\ ... & \\ k^2 - m \cdot k + \sum_{i=1}^{m} i 2^{m-i} & \text{for } 2^m \leq k \leq 2^{m+1} - 1. \end{cases}$$

Conjecture 1.10 implies that not only is $\rho(k)$ not a quadratic, but $\rho(k)$ eventually is strictly less than every quadratic with leading coefficient 1. Due to its complexity,

though, it seems appropriate to offer a simpler but weaker conjecture which would similarly show that $\rho(k)$ is eventually strictly less than every such quadratic.

Conjecture 1.11. *For every $m \in \mathbb{Z}$ there exist $K_m, c_m \in \mathbb{Z}$ such that $\rho(k) < k^2 - mk + c_m$ for all $k \geq K_m$.*

Mostly, the literature surrounding the shortest $[k]$-complete words involves constructing short $[k]$-complete words to find upper bounds on the lengths of the shortest such words. Fewer researchers have considered lower bounds on the lengths of $[k]$-complete words; however, there is one result from [4] which we will include here.

Theorem 1.12 (Kleitman and Kwiatkowski)**.** *For all k and for all $\epsilon > 0$, $\rho(k) > k^2 - c_\epsilon k^{7/4+\epsilon}$ where c_ϵ is a constant which depends on ϵ.*

2 Superpatterns

A natural generalization of a $[k]$-complete word is a word on $[k]$ which contains every possible length-k word as a subsequence (whereas a $[k]$-complete word only need contain every length-$[k]$ word where each letter is distinct). Unfortunately, if we try to construct the shortest such word for a given k, we immediately find that the concatenation of k blocks, each of which contains all k letters of $[k]$ exactly once, is such a word, and no shorter word with the desired property is possible. Therefore, the length of the shortest such word is k^2, and we will need to be a bit more creative to obtain an interesting variant of the problem of the shortest $[k]$-complete word.

To make this problem more interesting, we will consider an equivalence relation called

order isomophism. We say that two words S and T are order isomorphic (denoted by $S \sim T$) if the elements of S occur in the same relative order as the elements of T. Formally, $S \sim T$ if and only if $|S| = |T| = n$ and for any $i, j \leq n$, we have $s_i < s_j$ if and only if $t_i < t_j$. This is an equivalence relation, and we call the equivalence classes *preferential arrangements*. Then, an occurrence of a preferential arrangement is a subsequence of a word which is a member of the relevant equivalence class. Now, rather than look for words containing every possible k-letter word on $[k]$ as a subsequence, we will instead look for words containing occurrences of every possible length-k preferential arrangement. These words were first introduced by Burstein, Hästö, and Mansour in [1], and were called superpatterns; here we emphasize the alphabet from which they come by calling them $[k]$-superpatterns.

Example 2.1. The word 1122 is a length-4 preferential arrangement. It occurs in a word W on the alphabet [4] if W contains as a subsequence any of 1122, 1133, 1144, 2233, 2244, 3344.

Example 2.2. The word 1231231 is a superpattern for length-3 preferential arrangements because it contains an occurrence of each of 111, 112, 121, 122, 123, 132, 211, 212, 213, 221, 231, 312, 321, which are the 13 length-3 preferential arrangements.

As with $[k]$-complete words, the most frequently asked question regarding $[k]$-superpatterns is how short they can be. Just as we called the length of the shortest $[k]$-complete word $\rho(k)$, we will call the length of the shortest $[k]$-superpattern $\nu(k)$. In [1], the authors simply rephrase the classical conjecture for $[k]$-complete words by conjecturing that $\nu(k) = k^2 - 2k + 4$. Since that conjecture was later found to be

false for $\rho(k)$, it seems likely to be false for $\nu(k)$ as well. In this paper we will prove that this conjecture is indeed false, as a corollary to the main result of this section.

That result will be motivated by the following assertion in [1]. Because of the difficulty in proving it directly and the fact that it too is an immediate corollary of this section's main result, we will defer its proof for now.

Theorem 2.3 (Burstein, Hästö, and Mansour)**.** *For all k, the construction provided in Theorem 1.1 is not only a $[k]$-complete word, but is actually a $[k]$-superpattern.*

Theorem 2.3 gives us that $\nu(k) \leq k^2 - 2k + 4$ for all $k \geq 3$. Since each permutation of $[k]$ is a length-k word on $[k]$ which is order isomorphic only to itself, it follows that every $[k]$-superpattern contains every permutation of $[k]$ and is therefore a $[k]$-complete word. Therefore, $\nu(k) \geq \rho(k)$ for all k, and since $\rho(k) = k^2 - 2k + 4$ for $3 \leq k \leq 7$, we also have that $\nu(k) = k^2 - 2k + 4$ for $3 \leq k \leq 7$. The fact that $\nu(k) = \rho(k)$ for these values suggests the possibility that $\nu(k) = \rho(k)$ for all k. As it turns out, not only does this equality hold, but every $[k]$-complete word actually is a $[k]$-superpattern. Proving this, though, will require defining a new kind of superpattern with additional restrictions.

Before we actually define this new kind of superpattern, which we will call a *regular superpattern*, we will try to motivate it. Occurrences of permutations in a superpattern have certain nice properties that occurrences of general preferential arrangements lack; most importantly, each permutation is order isomorphic only to itself, and so we know exactly what any occurrence of a permutation of $[k]$ in a word on the alphabet $[k]$ looks like: it looks exactly like the permutation. So while an

occurrence of 122 in a word on the alphabet [3] might have the form 122, 133, or 233, an occurrence of 123 can only have the form 123. Then, if Π is a permutation of $[k]$ which occurs in a word on $[k]$, the smallest letter of Π is represented by the smallest letter of $[k]$, the second smallest letter of Π is represented by the second smallest letter of $[k]$, and so on. While we cannot hope to have exactly this correspondence between an occurrence of a preferential arrangement in a word on $[k]$ and the alphabet $[k]$, a regular superpattern will have something similar. If a preferential arrangement Σ has two equally small smallest letters, like, for example 112232, then these letters must be represented in a regular occurrence of Σ by one of the two smallest letters of $[k]$, i.e. 1 or 2. If Σ then has three equally small second-smallest letters, again like 112232, then these letters must be represented in a regular occurrence of Σ by one of the third, fourth, or fifth smallest letters of $[k]$, i.e. 3,4, or 5. This definition is formalized below.

Definition 2.4. Let $A_k = \{a_1, a_2, ..., a_k\}$ be a totally ordered alphabet with $a_1 < a_2 < ... < a_k$. A regular occurrence of a preferential arrangement in a word on the alphabet A_k is an occurrence of that arrangement such that for each letter, supposing there are i letters in the preferential arrangement that are less than that letter and j copies of that letter in the preferential arrangement, that letter is represented in the word by some element of $\{a_{i+1}, a_{i+2}, ..., a_{i+j}\}$. Then, an A_k-regular superpattern is a word on the alphabet A_k containing a regular occurrence of every length-$|A_k|$ preferential arrangement.

To illustrate this definition, consider the alphabet [6]. Then, a regular occurrence

of 112232 is one in which the $1s$ are represented by $1s$ or $2s$, the $2s$ are represented by $3s$, $4s$, or $5s$, and the 3 is represented by a 6. So, 113363 and 225565 are regular occurrences of 112232, but 113343 is not. Then, an A_k-regular superpattern is a word on the alphabet A_k containing a regular occurrence of every length-$|A_k|$ preferential arrangement.

Now, let $\mathcal{C}_{A_k}$ be the set of A_k-complete words, let $\mathcal{S}_{A_k}$ be the set of A_k-superpatterns, and let $\mathcal{R}_{A_k}$ be the set of A_k-regular superpatterns.

Theorem 2.5. *For all A_k with $k \geq 2$, $\mathcal{C}_{A_k} = \mathcal{S}_{A_k} = \mathcal{R}_{A_k}$.*

Proof: It is clear that $\mathcal{R}_{A_k} \subseteq \mathcal{S}_{A_k} \subseteq \mathcal{C}_{A_k}$, so it remains to show that $\mathcal{C}_{A_k} \subseteq \mathcal{R}_{A_k}$. We proceed by induction. As a base case, note that for A_2, any $\sigma_{\mathcal{C}} \in \mathcal{C}_{A_2}$ contains either the subsequence $a_1a_2a_1$ or $a_2a_1a_2$, so $\sigma_{\mathcal{C}} \in \mathcal{R}_{A_2}$. Now suppose that $\mathcal{C}_{A_{k-1}} \subseteq \mathcal{R}_{A_{k-1}}$. Choose any A_k (hereafter, we simply call this set A), choose some $\sigma_{\mathcal{C}} \in \mathcal{C}_A$, and let π be an arbitrary preferential arrangement of length k. Let π' be the portion of π following its first letter. We will now find a regular occurence of π in $\sigma_{\mathcal{C}}$ in both of two cases.

Case 1: Suppose that the first letter in π occurs just once in π. Call this first letter c, and let i be the number of letters in π less than c. Then, any regular occurrence of π represents c using a_{i+1}. Now, let $B = A\backslash\{a_{i+1}\} = \{b_1, b_2, ..., b_{k-1}\}$ where $b_1 < b_2 < ... < b_{k-1}$. Note that $b_j = a_j$ for $j < i+1$ and $b_j = a_{j+1}$ for $j \geq i+1$. Now, let $\sigma'_{\mathcal{C}}$ be the portion of $\sigma_{\mathcal{C}}$ following its first occurrence of a_{i+1} with all the a_{i+1}'s removed. Since $\sigma_{\mathcal{C}} \in \mathcal{C}_A$, it follows that $\sigma'_{\mathcal{C}} \in \mathcal{C}_B$. By the induction hypothesis, then, $\sigma'_{\mathcal{C}} \in \mathcal{R}_B$, so it contains a regular occurrence of π'. We claim that appending

a_{i+1} to the beginning of this occurrence gives a regular occurrence of π.

First, consider any letter $d < c$. Suppose there are j instances of d in π and d is greater than l other letters in π noting that $l + j \leq i$ must hold. Then, d must be represented in our regular occurrence of π by some element of $\{a_{l+1}, ..., a_{l+j}\}$. Since there are also j instances of d and l letters less than d in π', we know that d is represented in the regular occurrence of π' by some element of $\{b_{l+1}, ..., b_{l+j}\}$, and this set is equivalent to $\{a_{l+1}, ..., a_{l+j}\}$ because all the indices are less than $i+1$. Now consider c. For our occurrence to be regular, c must be represented using a_{i+1}, and it is. Finally consider $d > c$. Again, suppose there are j instances of d in π and d is greater than l other letters in π noting that, this time, $l \geq i + 1$. As before, d must be represented in our regular occurrence by some element of $\{a_{l+1}, ..., a_{l+j}\}$. Now, though, there are j instances of d and $l - 1$ letters less than d in π', so d is represented in the regular occurrence of π' by some element of $\{b_l, ..., b_{l+j-1}\}$, and this set is equivalent to $\{a_{l+1}, ..., a_{l+j}\}$ because all indices are at least $i + 1$. Thus, each letter in π is correctly represented, and we have a regular occurrence.

Case 2: Suppose that the first letter in π occurs p times with $p > 1$. Call this first letter c, and let i be the number of letters in π less than c. Then, any regular occurrence of π represents c using an element of $\{a_{i+1}, ..., a_{i+p}\}$. Let a_t be the last of those elements to make its first appearance in $\sigma_{\mathcal{C}}$, and let $\sigma'_{\mathcal{C}}$ be the portion of $\sigma_{\mathcal{C}}$ following the first occurrence of a_t with all subsequent a_t's removed. As in case 1, let $B = A \backslash \{a_t\} = \{b_1, b_2, ..., b_{k-1}\}$ where $b_1 < b_2 < ... < b_{k-1}$ and note that $b_j = a_j$ for $j < t$ and $b_j = a_{j+1}$ for $j \geq t$. Now, $\sigma'_{\mathcal{C}} \in \mathcal{C}_B$, and by the induction hypothesis, $\mathcal{C}_B \subseteq \mathcal{R}_B$, so $\sigma'_{\mathcal{C}}$ contains a regular occurrence of π'. Since there are $p-1$ occurrences

of c and i letters less than c in π', c must be represented in our regular occurrence by some element of $\{b_{i+1}, ..., b_{i+p-1}\}$ which is equivalent to $\{a_{i+1}, ..., a_{i+p}\}\backslash\{a_t\}$. Let a_s be the element in this set which represents c, and note that it must occur before the first appearance of a_t by our choice of a_t. We will show that appending a_s to the beginning of the regular occurrence of π' gives a regular occurrence of π.

As already noted, c is represented by $a_s \in \{a_{i+1}, ..., a_{i+p}\}$, and for any $d < c$, the proof that d is correctly represented is identical to the proof in case 1. For $d > c$, suppose there are j instances of d in π and d is greater than l other letters in π noting that $l \geq i + p$. Then, d must be represented in our regular occurrence by some element of $\{a_{l+1}, ..., a_{l+j}\}$. There are j instances of d and $l - 1$ letters less than d in π', so d is represented in the regular occurrence of π' by some element of $\{b_l, ..., b_{l+j-1}\}$, and this set is equivalent to $\{a_{l+1}, ..., a_{l+j}\}$ because all indices are at least $i+p \geq t$. Thus, we have found a valid regular occurrence of π in $\sigma_{\mathcal{C}}$, so $\sigma_{\mathcal{C}} \in \mathcal{R}_{A_k}$ and $\mathcal{C}_{A_k} = \mathcal{S}_{A_k} = \mathcal{R}_{A_k}$. □

Theorem 2.5 allows us to apply all of the facts about $[k]$-complete words from Section 1 to $[k]$-superpatterns. In particular, it is immediately apparent that, while the construction of Theorem 1.1 does indeed give a $[k]$-superpattern, it is not optimal when $k \geq 8$, and so the conjecture that $\nu(k) = k^2 - 2k + 4$ is false, just like the conjecture that $\rho(k) = k^2 - 2k + 4$.

A natural generalization of a $[k]$-superpattern is a word over a larger alphabet $[l]$ which contains all length-k preferential arrangements. Specifically, we will call a word on the alphabet $[l]$ an (l, k)-superpattern if it contains as a subsequence each length-k preferential arrangement and also contains each letter in the alphabet $[l]$. Then, we

will let $\mu(l,k)$ be the length of the shortest (l,k)-superpattern.

Now, when $l < k$, no (l,k)-superpatterns can exist because no words on $[l]$ contain any permutation of $[k]$. Also, when $l = k$, an (l,k)-superpattern is simply a $[k]$-superpattern, and so $\mu(k,k) = \nu(k)$. Therefore, we are only interested in $\mu(l,k)$ when $l > k$. Given the requirement that each (l,k)-superpattern contain each letter in the alphabet $[l]$, it seems that (l,k)-superpatterns will, if k is held constant and l grows large, be much longer than simple $[k]$-superpatterns. The following proposition formalizes that intuition.

Proposition 2.6. *For all k, there exists an L such that for all $l \geq L$, $\mu(l+1,k) = \mu(l,k)+1$.*

Proof: Define $d_k(l) = \mu(l,k) - l$. Given an (l,k)-superpattern, it is easy to construct an $(l+1,k)$-superpattern by simply appending the letter $l+1$ to the existing superpattern. Therefore $\mu(l+1,k) \leq \mu(l,k)+1$ for all k, so $d_k(l)$ is a non-increasing function. Since $d_k(l)$ is finite and $d_k(l) \geq 0$ for all k, it follows that there are just finitely many k such that $d_l(k+1) < d_l(k)$. Choose K to be one greater than the largest such k. For all $k \geq K$, we have $d_l(k+1) = d_l(k)$, so $\mu(l+1,k) = \mu(l,k)+1$. □

In the $k = 3$ case, at least, it is straightforward to see how low this value of L can be. Note that 1231231, 1231241, and 2353134 are all 7-letter $(l,3)$-superpatterns with $l = 3, 4$, and 5. Also note that any $(l,3)$-superpattern must contain at least three copies of some letter (in order to contain the preferential arrangement 111), and so $\mu(l,3) = l+2$ for $l \geq 5$. To show that $\mu(4,3) = 7$, note that the previous observation gives $\mu(4,3) \geq 6$, and suppose there exists a 6-letter $(4,3)$-superpattern W. W has

three identical letters a, and only three other letters; if the other three letters are not all distinct, then we could find a 6-letter word on [3] order-isomorphic to W and contradict Proposition 1.2. If the other three letters are all distinct, then there is either no more than one letter greater than a or else no more than one letter less than a; without loss of generality, assume there is no more than one letter greater than a. If there are no such letters, W fails to contain $112, 121$, and 211. If there is just one such letter, then it either occurs before two copies of a and W fails to contain 112, or it occurs after two copies of a and W fails to contain 211, contradicting the assumption that W is a $(4, 3)$-superpattern.

The preceding paragraph establishes the following formula for $\mu(l, 3)$:

$$\mu(l,3) = \begin{cases} 7 & \text{for } l = 3, 4, 5 \\ l+2 & \text{for } l > 5. \end{cases}$$

If we allow k to be any larger than 3, finding $\mu(l, k)$ becomes much more difficult. This is partly because, when $k > 3$, there is evidence that enlarging the available alphabet, i.e. increasing l, can sometimes actually decrease the value of $\mu(l, k)$. For instance, $\mu(4, 4) = 12$ by Theorems 1.2 and 2.5, but the word 43514342634 is an 11-letter $(6, 4)$-superpattern, so $\mu(6, 4) \leq 11$.

Note that a result of Miller in [5] guarantees words of length $\frac{k^2+k}{2}$ on the alphabet $[k+1]$ which contain all permutations of length k, but they unfortunately do not necessarily contain all preferential arrangements. While it seems too much to hope that adding additional letters to a superpattern's alphabet will allow us to cut the length of the superpattern in half, we do propose the following more modest conjecture:

Conjecture 2.7. *For all $k \geq 4$, there exists an $l > k$ such that $\mu(l, k) < \nu(k)$.*

3 Random Complete Words

Finally, we consider random words on the alphabet $[k]$. In particular, we are interested in the following random process: beginning with an empty word $\overline{W}$, at each timestep we choose a letter, uniformly at random, from the alphabet $[k]$. We then concatenate this value onto the end of $\overline{W}$ and check to see if $\overline{W}$ is a $[k]$-complete word (or, equivalently, a $[k]$-superpattern). We are interested in the value of $E[X_k]$ where X_k is the first timestep at which $\overline{W}$ is a $[k]$-complete word. This problem was first considered by Godbole and Liendo in [3]. There, the authors find values for $E[X_2]$ and $E[X_3]$; here we will provide different proofs for $E[X_2]$ and $E[X_3]$ which will illustrate the ideas we use to later find general upper and lower bounds on $E[X_k]$.

Theorem 3.1 (Godbole and Liendo)**.** *The equality $E[X_2] = 5$ holds.*

Proof: Note that a word is [2]-complete if and only if it contains two copies of a single letter separated by a copy of the other letter. Suppose we construct a word using the random process described above, and without loss of generality, suppose that the first letter is a 1. Then, in order to get a [2]-complete word, we must first wait to get a 2, and then wait to get another 1. Because each letter after the first one is equally likely to be a 2 or a 1, the wait time until the first 2 appears follows a geometric distribution with parameter $\frac{1}{2}$, and so does the wait time before a 1 appears after that first 2. Therefore, if T is a random variable following a geometric distribution

with parameter $\frac{1}{2}$, we have that

$$E[X_2] = 1 + 2 \cdot E[T] = 1 + 4 = 5.$$

□

Theorem 3.2 (Godbole and Liendo)**.** *The equality* $E[X_3] = 13.5625$ *holds.*

Proof: First, if a word $\overline{W}$ is [3]-complete, then it must contain as a subsequence a word W which meets the following conditions: every letter on the alphabet [3] appears in W, and the last letter to make its first appearance in W must be followed by every possible permutation of the remaining letters. For instance, if $\overline{W} = 112313231$, then $W = 123121$. In this proof, we will find the expected number of letters needed to form a word W with those conditions, then find the probability that after forming W we already formed a [3]-complete word, and, if not, we finally find the expected number of letters needed to turn W into a [3]-complete word.

In order to form W, we first need to see every letter once, i.e. we need to see three distinct letters. The first distinct letter will simply be the first letter. Then, of the three possible choices for each letter, two of them are distinct from the first letter. Therefore, the second distinct letter will appear after T_1 letters, where T_1 follows a geometric distribution with parameter $\frac{2}{3}$, and the third letter will appear after T_2 letters, where T_2 follows a geometric distribution with parameter $\frac{1}{3}$. Let a_1 be the first distinct letter to appear, a_2 be the second, and a_3 be the third. Then, we have a word with $a_1a_2a_3$ as a subsequence.

Now, we need to construct every possible permutation, i.e. a complete word, of

$\{a_1, a_2\}$. As noted in the previous proof, this means we need two copies of either a_1 or a_2, with a copy of the other between these two. We do not care whether a_1 or a_2 appears first, and so one of these will appear after T_3 letters, where T_3 follows a geometric distribution with parameter $\frac{2}{3}$. The other will appear afterwards after T_4 letters and the first one will appear again after T_5 letters, where T_4 and T_5 both follow a geometric distribution with parameter $\frac{1}{3}$. Therefore, if we let Y be the number of letters needed before the construction of W is finished, we have

$$E[Y] = 1 + E[T_1] + E[T_2] + E[T_3] + E[T_4] + E[T_5] = 1 + \frac{3}{2} + 3 + \frac{3}{2} + 3 + 3 = 13.$$

Now, W is either the subsequence $a_1a_2a_3a_1a_2a_1$, or else it is $a_1a_2a_3a_2a_1a_2$, and we will consider these two cases separately. Suppose W is $a_1a_2a_3a_1a_2a_1$. Then, it is straightforward to verify that it contains every permutation of $\{a_1, a_2, a_3\}$ except $a_2a_1a_3$. However, as we randomly added letters to $\overline{W}$ in order to find W as a subsequence of $\overline{W}$, we probably added several letters to $\overline{W}$ that we did not need for W. There are three cases in which these additional letters will make $\overline{W}$ $[k]$-complete as soon as we have finished constructing W. First, there may be a copy of a_1 between the first copy of a_2 and the copy of a_3; this will happen with probability $\frac{1}{2}$ because the first letter (besides possibly a_2) to follow this copy of a_2 is equally likely to be a_1 or a_3. Second, there may be a copy of a_3 between the second copies of a_1 and a_2; this also will happen with probability $\frac{1}{2}$. Finally, there may be a copy of a_3 between the second copy of a_2 and the third of a_1, which again will happen with probability $\frac{1}{2}$. Since all these events are independent, $\overline{W}$ contains $a_2a_1a_3$ and is therefore already

[3]-complete with probability $1 - \left(\frac{1}{2}\right)^3 = \frac{7}{8}$.

Next, suppose W is $a_1a_2a_3a_2a_1a_2$. Again, W contains each permutation of $\{a_1, a_2, a_3\}$ except possibly $a_2a_1a_3$, but now there are just two cases in which W actually turns out to contain $a_2a_1a_3$ as well. First, there may be a copy of a_1 between the first copy of a_2 and the copy of a_3, and again this will happen with probability $\frac{1}{2}$. Also, there may be a copy of a_3 between the second copy of a_1 and the third copy of a_2, which will also happen with probability $\frac{1}{2}$. Since these two events are independent $\overline{W}$ contains $a_2a_1a_3$ with probability $1 - \left(\frac{1}{2}\right)^2 = \frac{3}{4}$ in this case. Further, W is equally likely to have either of the forms $a_1a_2a_3a_1a_2a_1$ and $a_1a_2a_3a_2a_1a_2$, and, so in general, $\overline{W}$ is a [3]-complete word as soon as W is completed with probability $\frac{1}{2} \cdot \frac{7}{8} + \frac{1}{2} \cdot \frac{3}{4} = \frac{13}{16}$.

If $\overline{W}$ is not already a [3]-complete word, which happens with probability $\frac{3}{16}$, then we must continue adding letters until we add another copy of a_3. Because at each timestep we are equally likely to add any of three possible letters, we will add a_3 after T_6 letters where T_6 follows a geometric distribution with parameter $\frac{1}{3}$. Therefore,

$$E[X_3] = E[Y] + \frac{3}{16}E[T_6] = 13 + .5625 = 13.5625.$$

□

For $k > 3$, finding exact values for $E[X_k]$ becomes very difficult, and so we will provide upper and lower bounds rather than exact quantities. While an exact formula would be ideal, the following upper and lower bounds are good enough that together they demonstrate that $E[X_k]$ is asymptotically equivalent to $k^2 \ln(k)$.

Theorem 3.3. *For all k, $E[X_k] \leq k^2(\ln(k) + 1)$.*

Proof: Fix k. We will calculate $E[Y]$, where Y is the number of timesteps needed to construct a word W which contains k blocks, each of which contains all k letters in $[k]$. Since W must be a $[k]$-complete word, $E[Y] \geq E[X_k]$. In constructing each block of W, we repeatedly add letters until we have obtained k distinct letters. The first distinct letter of each block is simply the first letter. Then, of the k possible letters, $k-1$ are distinct from the first letter, and so the second distinct letter appears after T_2 timesteps where T_2 follows a geometric distribution with parameter $\frac{k-1}{k}$. In general, the i^{th} distinct letter appears after T_i timesteps where T_i follows a geometric distribution with parameter $\frac{k-i+1}{k}$. Since there are k blocks, we find that

$$E[Y] = k\sum_{i=1}^{k} \frac{k}{k-i+1} = k^2 \sum_{i=1}^{k} \frac{1}{k-i+1} = k^2 \sum_{i=1}^{k} \frac{1}{i} \leq k^2(1+\int_1^k \frac{1}{t}dt) = k^2(\ln(k)+1).$$

Since $E[X_k] \leq E[Y]$, we have $E[X_k] \leq k^2(\ln(k)+1)$. □

Theorem 3.4. *For all k, $E[X_k] \geq k^2(\ln(k)-1)$.*

Proof: Fix k, and fix some $[k]$-complete word $\overline{W}$. We will define a word W which $\overline{W}$ must contain, and then we will calculate $E[Y]$, where Y is the number of letters used before W appears. Let k_1 be the last element of $[k]$ to make its first appearance in $\overline{W}$ and let this appearance be the ${p_1}^{\text{th}}$ letter of $\overline{W}$. Then let k_2 be the last element of $[k]\backslash\{k_1\}$ to make its first appearance after $\overline{w}_{p_1}$, and let this appearance be the ${p_2}^{\text{th}}$ letter of $\overline{W}$. In general, let k_i be the last element of $[k]\backslash\{k_1, k_2, ..., k_{i-1}\}$ to make its first appearance after $\overline{w}_{p_{i-1}}$, and let this appearance occur at the ${p_i}^{\text{th}}$ letter of $\overline{W}$. Now, W consists of k blocks; the first block contains all letters in $[k]$, the second block contains all letters in $[k]\backslash\{k_1\}$, and so on. Because $\overline{W}$ is a $[k]$-complete word,

in particular because it contains $k_1k_2...k_k$ as a subsequence, it must contain W as a subsequence.

Now, let Y_i be the number of timesteps needed to form the i^{th} block of W. This block must contain $k-i+1$ distinct letters from the set $[k]\backslash\{k_1, k_2, ..., k_{i-1}\}$. At each timestep, we add one of k possible letters; there are $k-i+1$ possibilities for the first distinct letter, and so it appears after $T_{i,1}$ timesteps where $T_{i,1}$ follows a geometric distribution with parameter $\frac{k-i+1}{k}$. Then, there are $k-i$ possibilities for the second distinct letter and so on. Therefore, to find the j^{th} distinct letter requires waiting $T_{i,j}$ timesteps, where $T_{i,j}$ follows a geometric distribution with parameter $\frac{k-i-j+2}{k}$. Thus, we have that

$$E[Y_i] = \sum_{j=1}^{k-i+1} E[T_{i,j}] = \sum_{j=1}^{k-i+1} \frac{k}{k-i-j+2} = \sum_{j=i}^{k} \frac{k}{k-j+1}.$$

Using the fact that $E[Y] = \sum_{i=1}^{k} E[Y_i]$, we now get

$$E[Y] = \sum_{i=1}^{k}\sum_{j=i}^{k} \frac{k}{k-j+1} = \Big(\frac{k}{k} + \frac{k}{k-1} + ... + \frac{k}{1}\Big) + \Big(\frac{k}{k-1} + \frac{k}{k-2} + ... \frac{k}{1}\Big) + ... + \Big(\frac{k}{1}\Big)$$

$$= \Big(\frac{k}{k}\Big) + 2 \cdot \Big(\frac{k}{k-1}\Big) + ... + k \cdot \Big(\frac{k}{1}\Big) = \sum_{i=1}^{k} i\frac{k}{k-i+1} = k\sum_{i=1}^{k} \frac{i}{k-i+1}$$

$$= k\sum_{i=1}^{k} \frac{k-i+1}{i} = k^2\sum_{i=1}^{k} \frac{1}{i} - k\sum_{i=1}^{k} 1 + k\sum_{i=1}^{k} \frac{1}{i} \geq k^2 \int_1^k \frac{1}{t}dt - k^2 = k^2(\ln(k) - 1).$$

Therefore, $E[X_k] \geq E[Y] \geq k^2(\ln(k) - 1)$. □

Further Questions

There are a number of open questions regarding complete words and superpatterns. A few of them will be very difficult to prove; Conjecture 1.10, which proposes a formula for $\rho(k)$, has been outstanding for over 40 years. Still, there are simpler alternative conjectures, like Conjecture 1.11, which would give much of the same information. Regarding superpatterns, Conjecture 2.7 seems both tractable and interesting; it has not previously been studied, but it would give substantial insight into the lengths of superpatterns as their alphabet size increases.

Acknowledgements

The author would like to thank Patrick Keef and Aanand Sharma for their help editting and revising this thesis. He would also like to thank his thesis advisor David Guichard for his insight in suggesting topics of inquiry as well as his REU advisor Anant Godbole for suggesting this topic in the first place. Much of the research in this thesis was done in the summer of 2014 at an REU at East Tennessee State University with support from the National Science Foundation.

References

[1] A. Burstein, P. Hästö, and T. Mansour. Packing patterns into words. *Electronic Journal of Combinatorics*, 9, 2003.

[2] V. Chvatal, D. A. Klarner, and D. Knuth. Selected combinatorial research problems. 1972.

[3] Anant Godbole and Martha Liendo. Waiting time distribution for the emergence of superpatterns. *Methodology and Computing in Applied Probability*, 2015.

[4] D. J. Kleitman and D. J. Kwiatkowski. A lower bound on the length of a sequence containing all permutations as subsequences. *Journal of Combinatorial Theory, Series A*, 21:129136, 1976.

[5] Alison Miller. Asymptotic bounds for permutations containing many different patterns. *Journal of Combinatorial Theory*, 116:92108, 2009.

[6] Malcolm Newey. Notes on a problem involving permutations as subsequences. 1973.

[7] Saša Radomirović. A construction of short sequences containing all permutations of a set as subsequences. *Electronic Journal of Combinatorics*, 19, 2012.

[8] E. Zălinescu. Shorter strings containing all k-element permutations. *Information Processing Letters*, 111:605–608, 2011.