Abstract

We constructively prove the exact distribution of deletion sizes for unavoidable strings, under the reductive decidability method of Zimin and Bean et al. Bounds such as these on the unique initial reductions of unavoidable strings were instrumental in proving the computational intractability of the reduction algorithm. We also provide the necessary supporting results, including some useful approximations on the deletion sizes of individual strings. This work improves upon previous results that, although sufficient to establish the desired exponential lower bound, were far from optimal.

1. Introduction

The majority of recent research in the area of unavoidable strings has focused on the complementary problem of avoidability [6] [4]. However, determining the computational complexity of deciding string avoidability remains an interesting open question. As was recently demonstrated by the author in [5], the known reduction decision procedure of Zimin [7] and Bean et al. [3] has an exponential lower bound. Furthermore, the four other known necessary conditions are provably insufficient to decide string avoidability [5].

Proving the intractability of any algorithmic implementation of the reductive deletion method required demonstrating an exponential number of valid deletion possibilities. Instrumental to the combinatorial approach was a definition for the size of a unique deletion, closely related to the number of choices a deterministic reduction algorithm could face. Although the provably obtained progression of deletion sizes was sufficient to establish the exponential lower bound, the results were far from comprehensive. As proved in this paper, the possible deletion sizes for unavoidable strings with a unique initial reduction can be specified precisely as a function of the number of distinct symbols. Moreover, the constructive proof techniques guarantee the existence of strings achieving those deletion sizes.

We begin, in section 2, by reviewing some fundamental results on unavoidable strings. In section 3, we present proofs and supporting results not previously available for the necessary statements originally given in [5], as well as the definition of deletion size. We provide in section 4 some useful approximations on the deletions sizes of individual unavoidable strings. Finally, a constructive proof of the exact distribution of deletion sizes for unavoidable strings is given in section 5.

2. Unavoidable strings

The collection of all possible non-empty strings over a finite (and non-empty) set of symbols $S$ will be denoted $S^+$. An overbar distinguishes elements of $S^+$ from individual string symbols. $\varepsilon$ denotes the empty string, and $S^+ \cup \{\varepsilon\} = S^*$. Note that $S^*$ is the set of reduced strings, since $s \varepsilon s = s$ for all $s \in S$ and by extension for all $t \in S^*$. Finally, $\leq$ will refer to the substring relation on $S^+ \times S^+$, that is $t \leq s$ if there exist strings $\tilde{u}, \tilde{v} \in S^*$ so that $\tilde{u} \tilde{v} = s$.

The conception of unavoidable strings as a type of pattern consisting purely of “variable” symbols was first observed in [6]. The following formal definition has the advantage of consistency with the terminology and accepted usage of the pattern matching community [1]. Within this framework, standard pattern matching asks whether there exists an identity mapping between a pattern $\tilde{p}$ and a subword of the text word $\varnothing$, when both strings are over the same alphabet $A$. The notion of matching correspondence is explicitly broadened by introducing the new set $V$ of “variable” pattern symbols, which may map to non-empty words.
in $A^+$, while the symbols in $A$ remain “constant.”

**Definition 1 (Generalized pattern occurrence).** [5] For $p \in (V \cup A)^+$ and $w \in A^+$, $p \mid w$ if there exists a map $\phi : (V \cup A) \rightarrow A^+$ such that $\phi(a) = a$ for all $a \in A$ and $\phi(p) \leq w$ under the induced homomorphism.

In this context, we have the following rephrasing of one definition of unavoidable strings.

**Definition 2 (Unavoidable).** [3] [7] A string $p$ is called unavoidable if every infinite word on $n$ letters has an occurrence of $p$ as a generalized pattern consisting entirely of variable symbols.

Unavoidability is a decidable string characteristic, as was proved independently in [7] and [3]. The common result states that the unavoidability of a string can be determined by reducing it to the empty string under an appropriate sequence of deletions.\(^1\)

**Definition 3 (Free set).** [3] [7] $F \subseteq V$ is free for $p \in V^+$ if and only if there exist sets $A, B \subseteq V$ such that $F \subseteq B \setminus A$ where, for all $xy \leq p$, $x \in A$ if and only if $y \in B$.

The requirement on all substrings $xy$ is called the “two-window” criteria. By assumption, $F$ will always be nonempty. Also, the definition could equivalently require $F \subseteq A \setminus B$.

**Definition 4 ($\sigma$-deletion).** [3] [7] The mapping $\sigma_F : V \rightarrow V \cup \{\epsilon\}$ is defined by

$$
\sigma_F(x) = \begin{cases} 
  x & \text{if } x \notin F \\
  \epsilon & \text{if } x \in F
\end{cases}
$$

Note that $\sigma_F(p)$ always refers to the reduced string in $V^*$ under the induced homomorphism.

**Theorem 1.** [3] [7] $p$ is an unavoidable string if and only if $p$ can be reduced to $\epsilon$ by a sequence of $\sigma$-deletions.

Furthermore, [7] provides an equivalent unavoidability characterization in terms understood as generalized pattern occurrence.

**Definition 5 (Zimin words).** [7] Let $Z_1 = a_1$ and recursively define $Z_n$ on $A_n = \{a_1, a_2, \ldots, a_n\}$ as $Z_n = Z_{n-1}a_nZ_{n-1}$. Equivalently, define a mapping $\theta : A_{n-1} \rightarrow A_n$ where $\theta(a_i) = a_{i+1}a_i$. Then $Z_n = a_1\theta(Z_{n-1})$.

Let $\alpha(p)$ be the number of distinct symbols occurring in a string $p$.

\(^1\)Since the two sets of decision results use different terminology and notation, we have chosen whichever seemed the most appropriate for the purposes of this paper.

**Theorem 2.** [7] $p$ is an unavoidable string if and only if $p$ occurs in $Z_{\alpha(p)}$ as a generalized pattern consisting entirely of variable symbols.

The equivalence of the two characterizations follows from the fact that a matching correspondence lifts through a $\sigma$-deletion. Intuitively, the $\sigma$-deletion criteria permits the variable symbols to be “squeezed” into the mapping as needed.

### 3. Deletion size

A $\sigma$-deletion of $p \in V^+$ requires a free set $F \subseteq B \setminus A$, where $A$ and $B$ satisfy the two-window criteria. The minimal such sets for a given $p$ can be simultaneously constructed by considering its adjacency graph.

**Definition 6 (Minimal $\sigma$-graph).** [2] [5] Let $G(p)$ be the bipartite graph with vertices $[a, x]$ and $[x, b]$ for every $x$ in $p$ and $a, b \notin V$. $([a, x], [y, b])$ is an edge in $G(p)$ if and only if $a \leq p$.

The projections of the left and right sides of each connected component in $G(p)$ back down onto subsets of $V$ yield sets $A$ and $B$ minimally satisfying the two-window criteria for $p$. As initially stated in [5], although without the following proofs and accompanying supporting results, these minimal sets are sufficient to generate the necessary reducing free set.

Call a $\sigma$-deletion and its associated free set $F$ worthwhile if $\sigma_F(p)$ is unavoidable, that is if a $\sigma$-deletion sequence beginning with $\sigma_F$ reduces $p$ to the empty string. Also, let $\equiv$ be the equivalence relation on the vertices of $G(p)$ where two vertices are equivalent if and only if they are in the same connected component.

**Lemma 1.** [5] If $p$ is unavoidable, then there exists a worthwhile $\sigma$-deletion $\sigma_F$ of $p$ such that, for all $y_1, y_2 \in F$, $[y_1, b] \equiv [y_2, b]$ in $G(p)$.

**Proof.** If $p$ is unavoidable, then there exists at least one worthwhile $\sigma$-deletion $\sigma_F$. Suppose there exist $y_1, y_2 \in F$ such that $[y_1, b] \neq [y_2, b]$ in $G(p)$. Let $Y_1 = \{y \in F \mid [y, b] \equiv [y_1, b]\}$ and let $Y_2 = F \setminus Y_1$. We can prove that $\sigma_{Y_1}$ is worthwhile by showing that first $Y_1$ may be $\sigma$-deleted from $p$ and then $Y_2$. Since $Y_1 \subseteq F \subseteq B \setminus A$, $Y_1$ may be $\sigma$-deleted from $p$. Let $q = \sigma_{Y_1}(p)$. $Y_2$ may be $\sigma$-deleted from $q$ if there exist sets $A$ and $B$ satisfying the free set definition. By considering $G(q)$ and the contradictions which would result if such sets $A$ and $B$ did not exist, it can be seen that $Y_2$ may be $\sigma$-deleted from $q$. Since $\sigma_{Y_2}(q) = \sigma_{Y_1}(\sigma_F(p)) = \sigma_F(p)$, we conclude that $\sigma_{Y_2}$ was a worthwhile $\sigma$-deletion of $q$ and so $\sigma_{Y_1}$ must have been a worthwhile $\sigma$-deletion of $p$. \(\square\)
Given a string \( \sigma \), we may always consider its reverse, denoted \((\sigma)^R\). Since Zimin words are unchanged under reversal, \((Z_n)^R = Z_n\), we have the following immediate result.

**Lemma 2.** \( \bar{p} \) is an unavoidable string if and only if \((\bar{p})^R\) is.

Since \(xy \leq p\) if and only if \(yx \leq (p)^R\), \(G((p)^R)\) is just the mirror image, or “reversal”, of \(G(p)\) modulo swapping \([a, t]\) and \([t, b]\). Hence, not only would both \(\bar{p}\) and \((\bar{p})^R\) be unavoidable, but they would have the same sequences of worthwhile \(\sigma\)-deletions.

**Lemma 3.** \(\sigma_F\) is a worthwhile \(\sigma\)-deletion of \(\bar{p}\) if and only if it is \(\text{for} (p)^R\).

The next corollary follows immediately from the previous three lemmas. Given the symmetry possible in the free set definition, where \(B\backslash A\) was arbitrarily chosen over \(A\backslash B\), such duality should be expected.

**Corollary 1.** [5] If \(\bar{p}\) is unavoidable, then there exists a worthwhile \(\sigma\)-deletion \(\sigma_F\) of \(p\) such that, for all \(y_1, y_2 \in F\), \([a, y_1] \equiv [a, y_2]\) in \(G(\bar{p})\).

Hence, \(G(\bar{p})\) is justifiably considered the minimal \(\sigma\)-graph of \(p\); unavoidability can be determined by considering only the free sets arising from the connected components of \(G(\bar{p})\). Although an unavoidable string must have such a minimal \(\sigma\)-deletion, it is by no means unique.

For example, according to [7], if \(p \in V^+\) has a trivial \(\sigma\)-deletion with either \(\bar{p} = x\bar{q}\) or \(\bar{p} = \bar{q}x\) and \(\bar{q} \in (V \setminus \{x\})^+\), then \(\bar{p}\) is unavoidable if and only if \(\bar{q}\) is. If \(\bar{p}\) is unavoidable, then it must have at least two worthwhile \(\sigma\)-deletions: the one for \(\bar{q}\) and \(F = \{x\}\). Likewise, if \(\bar{p}\) has a linear substring sequence with \(\bar{p} = q_1xyq_2xyq_3 \ldots x y q_m\), where \(q_i \in (V \setminus \{x, y\})^+\) (unless \(q_1 = \varepsilon\) or \(q_m = \varepsilon\) in which case the initial \(x\) or terminal \(y\) respectively may also not appear), then both \(F = \{x\}\) and \(F = \{y\}\) would be worthwhile \(\sigma\)-deletions of \(\bar{p}\).

In each of the previous two special cases, \(G(\bar{p})\) has at least three connected components. The following theorem is a complete generalization of this connection between multiple connected components and different minimal worthwhile \(\sigma\)-deletions.

**Theorem 3.** [5] If \(\bar{p}\) is unavoidable and \(G(\bar{p})\) has at least three connected components, then there exists more than one minimal worthwhile \(\sigma\)-deletion of \(\bar{p}\).

**Proof.** If \(\bar{p}\) is unavoidable, there must exist a worthwhile \(\sigma\)-deletion \(\sigma_F\) which, by lemma 1, we may assume is minimal. Excluding the two connected components of \(G(\bar{p})\) containing symbols from \(F\) on the right and left sides, there is at least one connected component remaining. Ruling out the two special cases of trivial \(\sigma\)-deletions and linear substring sequences, there must be at least three symbols wholly unconnected to those from \(F\). Employing techniques similar to the proof of lemma 1, if one of those unconnected symbols appears in the next (minimal) worthwhile \(\sigma\)-deletion, then the order of the two \(\sigma\)-deletions may be reversed. Hence \(\bar{p}\) has more than one minimal worthwhile \(\sigma\)-deletion.

If the symbols from the following \(\sigma\)-deletion are connected to \(F\) in \(G(p)\), then \(G(\sigma_F(\bar{p}))\) has at least three connected components and the symbols unrelated to \(F\) in \(G(\bar{p})\) remain disconnected in \(G(\sigma_F(\bar{p}))\). Since the successful \(\sigma\)-deletion sequence which reduces \(\bar{p}\) to \(\varepsilon\) is finite, this argument can be recursively applied until the desired conclusion is obtained.

The theorem’s usefulness, at least as far as this paper is concerned, lies primarily in its contrapositive. Restricting our focus to unavoidable \(\bar{p}\) having a unique worthwhile \(\sigma\)-deletion insures that \(G(\bar{p})\) has exactly two connected components.

Note that the converse of the theorem is not true; even if \(G(\bar{p})\) has exactly two connected components, it may be that \(p\) has a worthwhile \(\sigma\)-deletion from each one or more than one from within the same connected component.

**Lemma 4.** Suppose \(p \in V^+\) and \(p = x\bar{q}x\) with \(\bar{q} \in (V \setminus \{x\})^+\). Then \(\bar{p}\) is unavoidable if and only if \(\bar{q}\) is.

**Proof.** Any substring \(\bar{q}\) of an unavoidable \(\bar{p}\) must be likewise unavoidable. Since a worthwhile \(\sigma\)-deletion of \(\bar{q}\) must be one of \(\bar{p}\) as well, the other direction follows inductively.

Let \(p \in V^+\) such that \(G(p)\) has exactly two connected components. Then there are precisely two pairs of sets minimally satisfying the two-window criteria for \(p\), denoted \(A, B, A', B'\). Suppose that \(\bar{p} = s\bar{q}t\) for \(s, t \in V\). Then, for \(x \notin V\), \(x\bar{p}x\) has at least two worthwhile \(\sigma\)-deletions: the one for \(\bar{p}\) and \(F = \{x\}\). Furthermore, \(G(x\bar{p}x)\) still has only two connected components.

When \(p\) has a unique worthwhile \(\sigma\)-deletion, \(G(p)\) has exactly two connected components. By convention, we will assume that \(F \subseteq B \setminus A\) for the unique worthwhile free set \(F\) and the corresponding sets arising from the connected components of \(G(p)\). In this case, we may unambiguously define the \(\sigma\)-deletion size, measured with respect to the particular \(p\) and its minimal \(\sigma\)-graph \(G(p)\).

**Definition 7 (\(\sigma\)-deletion size).** [5] Suppose \(p\) has a unique worthwhile \(\sigma\)-deletion \(\sigma_F\) with \(F \subseteq B \setminus A\). Let \(|\sigma_F| = k_1/k_2\) where \(k_1 = |F|\) and \(k_2 = |B \setminus A|\).

In terms of the most general bounds, clearly \(1 \leq k_1 \leq k_2\). Theorem 5 provides a statement of the precise relationship among possible \(k_1, k_2, \alpha(p)\). First, however, we will give more approximate bounds for individual \(p \in V^+\).
4. Approximate bounds for specific strings

Lemma 5. Suppose that \( p \in V^+ \) is unavoidable and that \( \alpha(p) \geq 3 \). Suppose further that \( \bar{p} \) does not have a trivial \( \sigma \)-deletion, nor is \( p \) of the form \( p = x \bar{p} x \) where \( \bar{q} \in (V \setminus \{x\})^+ \). Let \( \sigma_F \) be a worthwhile \( \sigma \)-deletion of \( p \), with associated minimal two-window sets \( A \) and \( B \). Then \( |B \setminus A| < \alpha(p) - 1 \).

Proof. Suppose not. Then \( \sigma_F \) is a worthwhile \( \sigma \)-deletion of \( p \) where \( |B \setminus A| = \alpha(p) - 1 \geq 2 \). Let \( A = \{x\} \). Then the elements of \( B \) must alternate with \( x \) to form \( \bar{p} \). Let \( y \in B \). Recall that \( \bar{p} \) does not have a trivial \( \sigma \)-deletion, nor is \( \bar{p} \) of the form \( \bar{p} = x \bar{q} x \) where \( \bar{q} \in (V \setminus \{x\})^+ \). Hence, \( x \bar{y} x \bar{n} \leq \bar{p} \). If \( y \in F \), then \( x \bar{x} \leq \sigma_F(y) \), which is avoidable. Since \( F \neq \emptyset \) by assumption, this is a contradiction.

It follows from lemmas 4 and 5, that if \( |B \setminus A| = \alpha(p) - 1 \geq 2 \), then \( \bar{p} \) has at least two worthwhile \( \sigma \)-deletions. In particular, when considering the size of a unique worthwhile \( \sigma \)-deletion of \( p \), then it must be that \( k_2 \leq \alpha(p) - 2 \).

We can show, by construction, that there exist \( p \) with unique worthwhile \( \sigma \)-deletions of size \( \frac{1}{\alpha(p)} \) for \( \alpha(p) \geq 3 \). For example, consider \( p = xyzezwztx \) and its generalization. However, it is by no means clear that there exist \( p \) with \( \sigma \)-deletion size \( \frac{1}{\alpha(p)} - 2 \) for all \( 1 < k_1 \leq n - 2 \).

Lemma 6. Let \( p \in V^+ \) be unavoidable with \( \alpha(p) = n \) and worthwhile \( \sigma \)-deletion, \( \sigma_F \). Then \( \log |F| - |F| \leq n \).

Proof. Let \( |F| = k \) and suppose instead that \( \log k + k > n \), in which case \( k < 2^{n-k} \). If \( q = \sigma_F(p) \), then \( \alpha(q) = n - k \), which means that \( n - k \leq |q| \leq 2^{n-k} - 1 \). For every \( y, z \in F \), \( yz \not\in p \). There are strictly more than \( 2^{n-k} \) different \( y \in F \); however there are at most \( 2^{n-k} - 1 \) (not necessarily distinct) symbols in \( q \). Thus, even if every instance in \( p \) of \( x \in V \setminus F \) from \( \bar{q} \) were preceded and followed by a \( y \in F \), there would not be enough to separate every symbol of \( F \) from all the others. But this contradicts the fact that \( F \) may be \( \sigma \)-deleted from \( p \).

Equality can be achieved when \( k \) is a power of 2. Suppose that \( k = 2^m \). The Zimin word with \( m \) symbols has length \( |Z_m| = 2^m - 1 \). Consider the string \( p \) with \( 2^m + m \) symbols, where \( p = Z_m \) with one of the \( 2^m \) remaining symbols inserted between consecutively symbols of \( Z_m \).

\[ p = x_{m+1}x_1x_{m+2}x_2x_{m+3}x_3x_{m+4}x_4x_{m+5}x_5 \ldots x_{2^m} \]

Let \( B = \{x_{m+1}, \ldots, x_{2^m-m}\} \) and \( A = \{x_1, \ldots, x_m\} \). By construction, \( A \) and \( B \) satisfy the two-window property for \( p \). For \( F = B \setminus A \), \( |F| = 2^m = k \) and \( \sigma_F(\bar{p}) = Z_m \). Thus, \( F \) is a worthwhile free set of \( \bar{p} \) of size \( k \) and \( \alpha(p) = k + m = 2^m + m \). Note that \( p \) is unique, up to isomorphism, since each of the \( 2^m \) symbols in \( F \) must be separated from one another in \( p \) and \( Z_m \) is the unique unavoidable string, modulo isomorphism, of maximal length \( 2^m - 1 \). Aside from these particular strings, however, the inequality is strict.

Hence, we have the bound that \( k_1 + \log k_1 < \alpha(p) \) for any \( p \) with a unique worthwhile \( \sigma \)-deletion. The strictness of the inequality is due to the fact that \( p \) with \( \alpha(p) = 2^m + m \) and a \( \sigma \)-deletion \( \sigma_F \) with \( |F| = 2^m \) has at least two other worthwhile trivial \( \sigma \)-deletions. In fact, as the following lemma shows, such a \( p \) would have many other worthwhile \( \sigma \)-deletions.

Lemma 7. Suppose that \( p \in V^+ \) is unavoidable. If \( \sigma_F \) is a worthwhile \( \sigma \)-deletion of \( p \), and there exists \( y \in F \) such that \( p = q_1 \bar{y} q_2 \) with \( q_1, q_2 \in (V \setminus \{y\})^+ \), then \( \sigma_F \setminus \{y\} \) is also a worthwhile \( \sigma \)-deletion of \( p \).

Proof. Let \( \bar{F} = F \setminus \{y\} \), \( \alpha(\bar{F}) = n \), and \( |\bar{F}| = k \). Clearly, \( \bar{F} \subseteq F \subseteq B \setminus A \) for some sets \( A \) and \( B \) satisfying the two-window criteria for \( p \). Hence, \( F \) may be \( \sigma \)-deleted from \( p \). To see that \( \sigma \)-deleting \( \bar{F} \) would be worthwhile, note that \( \sigma_F(p) \setminus Z_m = k \). But then it is a simple matter to “lift” the appropriate mapping so that \( \sigma_F(p) \setminus Z_m = k \). Since \( \sigma_F(p) \) is unavoidable, \( \sigma \)-deleting \( F \) from \( p \) is worthwhile.

Consequently, \( \bar{p} \) can have a unique worthwhile \( \sigma \)-deletion which contains an isolated symbol, that is a symbol with exactly one instance in \( \bar{p} \). Since we are primarily concerned with the distribution of sizes of unique worthwhile \( \sigma \)-deletions \( \sigma_F \) of \( \bar{p} \), we now know that for every \( y \in F \), there are at least two instances of \( y \) in \( p \).

5. Exact deletion size distribution

Rather than continuing to work with specific \( p \), we will be thinking about equivalence classes of unavoidable strings, having the same number of symbols and the same size unique worthwhile \( \sigma \)-deletion.

Definition 8. [5] Let \( (k_1/k_2)_n \) be the set of all unavoidable strings such that, for any \( \bar{p} \in (k_1/k_2)_n \), \( \alpha(\bar{p}) = n \) and \( \bar{p} \) has a unique worthwhile \( \sigma \)-deletion \( \sigma_F \) where \( |\sigma_F| = k_1/k_2 \).

From the combinatorial results of [5], we know there exist entire progressions of nonempty \( (k_1/k_2)_n \), for a given number of symbols. However, we wish to specify exactly which \( (k_1/k_2)_n \) is not \( \emptyset \), as a function of \( n \). From the results of the previous section, we have the general bounds that \( 1 \leq k_2 \leq n - 2 \) and \( \log k_1 + k_1 < n \). Furthermore, lemma 7 points to the importance of the frequency with which symbols appear in an unavoidable \( p \).

Theorem 4. If \( (k_1/k_2)_n \) is nonempty, then there exists an integer \( j \), with \( k_2 - k_1 \leq j \leq 2k_2 - k_1 - 1 \), such that

\[ 2k_1 + 2j \leq 2^{n-1}(j + 1) \]
Proof. Suppose that $p \in (k_1/k_2)_n$. Let $\sigma_F$ be the unique worthwhile $\sigma$-deletion of $\bar{p}$. Let $q = \sigma_F(p)$ with $|q| = n - k_1$. Let $j_i$ be the number of instances of the $i$th least frequently occurring symbol in $q$. Any unavoidable $q$ has an isolated symbol, so $j_1 = 1$. For $i \geq 2$, $j_i \leq 1 + \sum_{m=1}^{i-1} j_m$. Now let $j$ be the number of instances of the $(k_2 - k_1)$ least frequently occurring symbols in $q$. Then $k_2 - k_1 \leq j \leq 2^{k_2-k_1}-1$. The $j$ occurrences of the $(k_2 - k_1)$ least frequently occurring symbols in $q$ may be interspersed by at most $(j + 1)$ unavoidable strings containing at most $(n-k_1)-(k_2 - k_1)$ symbols, and thus having length at most $2^{n-k_1} - 1$. Thus, $|q| \leq (2^{n-k_1} - 1)(j + 1) + j = 2^{n-k_1}(j + 1) - 1$.

Now, rather than thinking of $q$ as $p$ with the symbols of $F$ removed, consider $\bar{p}$ to be an “in” by lemma 7, for every $y \in F \subseteq B \setminus A$, $y$ must occur at least twice in $\bar{p}$. For any $y \in B \setminus A$, it can not be that $yz \leq \bar{p}$ where $z \in B \setminus A$. Since every instance of $y \in F$ in $\bar{p}$ must be separated from any other occurrence of $z \in F$, there are at most $|q| + 1$ spaces, before and after every symbol in $q$, which could accommodate the symbols of $F$. Let $m$ be the number of instances of $q$ of the symbols from $(B \setminus A) \setminus F$. For $z \in (B \setminus A) \setminus F$, we can not use the two spaces on either side of an instance of $z$ in $q$. Hence, $2k_1 \leq (|q| + 1) - 2m$. But we also know that $j \leq m$ so that, for $k_2 - k_1 \leq j \leq 2^{k_2-k_1} - 1$, we have

$$2k_1 + 2j \leq 2k_1 + 2m \leq |q| + 1 \leq 2^{n-k_1}(j + 1)$$

The right side of the resulting inequality is linear in $j$ and hence maximal when $j = 2^{k_2-k_1} - 1$. There exist strings $q$ achieving this maximal length. Specifically $q = Z_{n-k_1}$ which has length $2^{n-k_1} - 1 = 2^{n-k_1} - 2^{k_2-k_1} - 1 = 2^{n-k_1}(j + 1) - 1$. Thus, we would expect that for nonempty $(k_1/k_2)_n$ there would exist $p \in (k_1/k_2)_n$ such that $\sigma_F(p) = Z_{n-k_1}$. The following theorem validates this expectation.

The constructive proof techniques do not cover the exceptional cases, whose emptiness has been exhaustively verified. Since this is the fundamental result, we provide a detailed proof.

Theorem 5. [5] With the exception of $(2/2)_4$ and $(4/4)_7$, for $1 \leq k_1 \leq k_2 \leq n - 2$, $(k_1/k_2)_n \neq \emptyset$ if and only if $1 + \sum_{i=k_2-k_1}^{n-k_2-k_1} 2^i \geq k_1$.

Proof. Suppose that $1 + \sum_{i=k_2-k_1}^{n-k_2-k_1} 2^i < k_1$. But then $k_1 > 2^{n-k_2-k_1} - 2^{k_2-k_1} + 1$, which means that there is no possible $j$ such that $k_2 - k_1 \leq j \leq 2^{k_2-k_1} - 1$ and $k_1 \leq 2^{n-k_2-k_1}(j + 1) - 1$. Hence, by theorem 4, we know that $(k_1/k_2)_n = \emptyset$ in this case.

Now then, suppose that $k_1 \leq 1 + \sum_{i=k_2-k_1}^{n-k_2-k_1} 2^i$. Given $k_1$ and $k_2$ which satisfy the bound, we can construct $p$ such that $\alpha(p) = n$ and $p$ has a unique worthwhile $\sigma$-deletion $\sigma_F$, where $|F| = k_1$ and $|B \setminus A| = k_2$. Let $\bar{q} = Z_{n-k_1}$ and $F = \{y_1, y_2, \ldots, y_{k_1}\}$. We will construct $p$ from $\bar{q}$ by “filling in” elements of $F$ into some of the $2^{n-k_1}$ available spaces between the symbols of $\bar{q}$, $x_1, \ldots, x_{n-k_1}$, in such a way that $\sigma_F$ is the unique worthwhile $\sigma$-deletion of $p$.

The first condition imposed on creating $p$ is that for every $y_i \in F, y_i x_1 \leq p$ and $x_1 y_i \leq p$. The second constraint has two parts, depending on the relative sizes of $k_1$ and $k_2$. For every $x_j$ with $1 \leq j \leq n - k_2$, there must exist $y_i \in F$ such that $y_i x_j y_i \leq p$ and no $x_1 x_j$ or $x_j x_1 \leq p$. For every $x_j$ where $j = n - k_2 + 1, \ldots, n - k_1$, only $x_1 x_j x_1 \leq p$, so that every occurrence of these $x_j$ in $\bar{p}$ is bracketed only by $x_1$. In the event that $k_2 = k_1$, we modify the first part of the second condition slightly by requiring that $y_i x_n-k_1 y_i \leq \bar{p}$ where $y_i \neq y_j$. Any $p$ which meets both of these conditions has a $G(p)$ with exactly two connected components. Under such a constrained construction, $F = \{y_1, y_2, \ldots, y_{k_1}\}$ would clearly be a worthwhile $\sigma$-deletion of $p$. Furthermore, the second condition insures that there is no other possible worthwhile $\sigma$-deletion $\sigma_F$ with $x_i \in F$. (Even if $k_1 = k_2$, as the only isolated symbol, $x_{n-k_1}$ still couldn’t be $\sigma$-deleted.)

Hence, in order to insure the uniqueness of $F$ as a worthwhile $\sigma$-deletion, we need only consider subsets of $F$ to rule out all other possibilities. Suppose that $\bar{F} \subset F$, and let $\bar{\sigma} = \sigma_F(p)$. Recall that every other symbol of $\bar{q}$ is an $x_i$ so that $G(\bar{\sigma})$ can have at most two connected components. If $G(\bar{\sigma})$ were to have only one connected component, then $\bar{\sigma}$ could have no further $\sigma$-deletions, worthwhile or otherwise, and so $\sigma_F$ could not be a worthwhile $\sigma$-deletion of $p$. Considering $G(\bar{\sigma})$, let $A$ and $B$ be the projections of the connected component containing $[x_1, b]$ with $A^c$ and $B^c$ corresponding to that containing $[a, x_1]$. For $y_i \in F \setminus \bar{F}$, it must still be that $y_i x_1$ and $x_1 y_i \leq p$ as it was in $p$, so that $p$, $y_i \in A$ and $y_i \in B^c$. Now for $x_i$ where $i = 2, \ldots, n - k_1$, there are four possible cases.

1. There exist $y_j, y_j' \in \bar{F}$ such that $x_1 y_j x_1 \leq p$ and $x_1 y_j' x_1 \leq p$, which implies that $x_i \in B$ and $x_i \in A^c$.
2. There exists $y_j \in \bar{F}$ such that $x_1 y_j x_1 \leq p$ and $x_1 x_i \leq p$, which implies that $x_i \in B$ and $x_i \in A$.
3. There exist $y_j \in \bar{F}$ such that $x_1 x_i \leq p$ and $x_1 y_j x_1 \leq p$, which implies that $x_i \in B^c$ and $x_i \in A^c$.
4. $x_1 x_i \leq p$ and $x_1 x_i \leq p$, which implies that $x_i \in B^c$ and $x_i \in A$.

We note that if there exists $x_i$ belonging to two or more of the preceding four cases, then $G(\bar{\sigma})$ has only one connected component. Thus, as the third and final constraint on the creation of $p$, the symbols of $\bar{F}$ must be filled into $\bar{q}$ in such a way that $\sigma$-deleting from $\bar{q}$ any proper subset $\bar{F}$ of $F$ would
Table 1. This table lists the minimal $n_0$ such that $(k_1/k_2)_{n_0}$ is nonempty, according to theorem 5. Note that by the constructive argument of that proof, we know there exists $p \in (k_1/k_2)_{n_0}$ satisfying the conditions of the progression theorems in [5]. Hence, it is easy to see that for $m \geq n_0$, $(k_1/k_2)_m$ is nonempty as well.
result in at least one of $x_i$ for $1 < i < n - k_1$ falling into more than one of the four categories above.

We will consider the two different cases, $k_1 = k_2$ and $k_1 < k_2$. Although we will present one possible method of inserting the symbols of $F$ so that the conditions above are satisfied, this is not to suggest that it is the only means.

Suppose then that we wish to have $k_1 = k_2$. In this case, $k_1 \leq 1 + \sum_{i=k_0-k_1}^{n-2-k_1} 2^i$ means that $k_1 \leq 2^{n-1-k_1}$, so we are assured of having sufficient spaces interbetween the symbols of $q$ to accommodate at least two instances of every symbol $y_i \in F$. Consider $q = \tilde{q} x_{n-k_1} \tilde{q}_r$ where $\tilde{q}_r = \tilde{q}_r = Z_{n-k_1}$. We may assume that $k_1 \geq 2$. Divide $F$ into two essentially equal disjoint sets, $L$ and $R$, and suppose that $y_i \in L$ with $y_i \in R$. Let $|L| = |R| = 1$, fill in all the spaces of $\tilde{q}_r$ with $y_i$ and all those of $\tilde{q}_i$ with $y_i$. Clearly this satisfies the first and second constraints. Additionally, for $n > 4$, if we $\sigma$-delete either $y_i$ or $y_2$, then there exists $x_i$ such that the first and fourth cases apply. But this means that $\sigma(y_i)$ and $\sigma(y_2)$ are not worthwhile $\sigma$-deletions, and the third constraint is also satisfied.

Hence, suppose that $y_i \in L$ as well. We begin sequentially filling in the spaces of $\tilde{q}_r$, starting from the right, with two instances of each $y_i \in R$ except for $y_2$. Once all the other $y_i \in R$ have been exhausted, the rest of $\tilde{q}_r$ is filled in with $y_2$. Over in $\tilde{q}_i$, we fill in the first three spaces on the left and last space on the right with $y_i$. We may do this as long as $k_1 \neq 2^{n-1-k_1}$. The rest of $\tilde{q}_i$ is filled in same manner as $\tilde{q}_r$, placing two sequential instances of every $y_i \in L \setminus \{y_i, y_2\}$, separated by the $x_i$ symbols of $q$, and then using $y_3$ to fill in the rest as needed. We claim this construction satisfies the conditions outlined above.

Every other symbol of $q$ must be an $x_1$ so sequentially filling in two (or more) instances of $y_i$ clearly results in $y_i x_1 \in R \setminus \bar{F}$. Moreover, by construction, $y_i x_1$, $y_2 x_2 y_1 \leq \bar{p}$ and for $x_i$ where $2 < i < n - k_1$, there exists $y_i \in L \setminus \{y_i\}$ such that $y_i x_1 y_j \leq \bar{p}$. For $(2/2)$ and $(4/4)$, there is no possible way of filling in $Z_{2}$ or $Z_2$ satisfactorily under any method, and we know experimentally that these sets are empty. For $k_1 \leq k_1 = 2^{n-1-k_1}$ this construction yields that there exists $y_i \in L \setminus \{y_i\}$ such that $y_i x_1 y_j \leq \bar{p}$. Hence, we may replace the rightmost occurrences of $y_i$ by a different symbol and so accommodate exactly two instances of every symbol $y_i$ when $\tilde{q} \leq k_1 = 2^{n-1-k_1}$ while satisfying the first and second constraints.

All that remains to be shown is that $\sigma$-deleting any proper subset $\tilde{F}$ of $F$ results in an $x_i$ falling into at least two of the cases outlined above. Suppose not. Observe first of all that for $y_j \in R$, $x_2 y_j x_1 y_j x_i \leq \bar{p}$ or $x_2 y_j x_1 y_j x_2 \leq \bar{p}$ where $2 < i \leq n - k_1$ except for the rightmost $y_j$ for which $x_i = \epsilon$. For $y_j \in R$, $y_j \in \tilde{F}$ means that $x_2$ must fall into case two, three, or four. To avoid the contradictory first case as well, it must be that $y_m \in \tilde{F}$ for $y_m \in L$ and $y_m x_2 y_m \leq \bar{p}$. But then $x_2$ can only be in case four, which means that $y_j \in \tilde{F}$ for all $y_j \in R$. However, then $x_i$ is in case four for all $2 \leq i < n - k_1$, which means that $y_m \in \tilde{F}$ for all $y_m \in L$. (This would not be true, however, if we had chosen to put the symbol also occurring as $x_2 y_m x_m x_1$, for $m \neq 1$ into the first space of $\tilde{q}_i$.) But, in order to avoid a contradiction, we must have that $\tilde{F} = L \cup R = F$ which contradicts $\tilde{F}$ being a proper subset of $F$. If we had begun with $y_m \in F$ for $y_m \in L$, then we would have been forced to have $x_i$ in case four for $2 \leq i < n - k_1$, or that $x_2$ in case two or three. In any case, to avoid a contradiction, we must $\sigma$-delete some $y_j \in R$ as well, and the chain of reasoning continues as above. So, the unique worthwhile $\sigma$-deletion of $q$ must be $\sigma_F$ where $F = B = \{y_i, \ldots, y_k\}$, $A = \{x_1, \ldots, x_{n-k_1}\}$, and $|F| = k_1 = |B \setminus A| = k_2$.

Suppose instead that $k_1 < k_2$. According to the second condition, for the $k_2 - k_1$ symbols $x_i$ where $n - k_2 + 1 \leq i \leq n - k_1$, $x_i x_1 x_i \leq \bar{p}$ for any instance of $x_i$. Thus, we may think of $q$ as $q = \tilde{q} x_1 x_{n-k_2+1} x_1 y_m \ldots y_m x_1 x_{n-k_2+1} \tilde{q}$, where $\tilde{q}_i x_1 = x_i \tilde{q}_m x_i = x_i \tilde{q}_r = Z_{n-k_1}$, and we will be filling in the symbols of $F$ only between the symbols of $\tilde{q}_i$, $\tilde{q}_r$, and the $(2^{n-k_2} - 2) q_m$. We know that $k_1 \leq 1 + \sum_{i=k_0-k_1}^{n-2-k_1} 2^i$ so that $2k_1 \leq 2 + 2^{n-k_1} - 2k_1 = 2(2^{n-k_1} - 1) + (2^{n-k_1} - 2)$, which means that we have enough space between the symbols of $\tilde{q}_i$, $\tilde{q}_r,$ and the $q_m$ to accommodate at least two instances of every symbol from $F$. As when $k_2 = k_1$, we will specify how to fill in the symbols in such a way that $\sigma_F$ is the unique worthwhile $\sigma$-deletion of the resulting string $\bar{p}$.

We may assume that $|F| \geq 2$ and that $y_1, y_2 \in F$. Beginning with the right side of the rightmost $\tilde{q}_m$, if such exists, sequentially fill in two instances of each $y_i \in F \setminus \{y_1, y_2\}$. If there is any remaining space available in the $q_m$, use $y_2$ to fill in the rest of the $\tilde{q}_m$ and $y_i$ for all of $\tilde{q}_i$ and $\tilde{q}_r$. Otherwise, use $y_2$ to fill in the rightmost space of $\tilde{q}_r$ and the rightmost space of $\tilde{q}_r$. Then sequentially fill in the rest of $\tilde{q}_r$, from the right, followed by $\tilde{q}_r$ from the right, with two instances of $y_i \in F$ not already used. Any remaining spaces, including the final two leftmost of $\tilde{q}_i$, should be filled in with $y_i$. Again, every other symbol in $\tilde{q}$ must be an $x_1$ so sequentially filling in consecutive segments of $\tilde{q}$ with two (or more) occurrences of symbols $y_i$ from $F$ satisfies the first condition. For every $x_i$ where $2 \leq i \leq n - k_2$, $x_i \leq \tilde{q}_r$ so there exists $y_j \in F$ such that $y_j x_i y_j \leq \bar{p}$ and we know that $y_i x_1 y_j \leq \bar{p}$. By only filling in around the symbols of $\tilde{q}_i$, $\tilde{q}_r$, and the $q_m$, we also have that for all $x_i$ where $n - k_2 + 1 \leq i \leq n - k_1$, we have only $x_i x_i x_1 x_1 \leq \bar{p}$. Consequently, the second condition is satisfied by this method as well.

Again, all that remains to be shown is that $\sigma$-deleting any proper subset $\tilde{F}$ of $F$ results in an $x_i$ falling into at least two of the cases listed above. Suppose not. Suppose that $y_j \in \tilde{F}$ where $y_j$ is one of the symbols used when filling in $\tilde{q}_m$ or $\tilde{q}_r$, and that $j \neq 2$. Then there exists
$x_i$ where $2 \leq i \leq n - k_2$ such that $y_j x_i \neq y_j$, which means that $x_i$ must fall under case four. $x_i$ appears in $\xi_i$ as well. If $i \neq 2$, there must exist $y_h$ and $y_g$ such that $x_2 y_h x_1 y_g x_2 \neq \Sigma$. Since $x_2$ falls under case four, to avoid an immediate contradiction, it must be that $y_h, y_g \in \hat{F}$. $\sigma$-deleting $y_i$ implies that $x_2$ is in case two or four, while $\sigma$-deleting $y_g$ implies that $x_2$ is in case three or four. Thus, avoiding contradiction, $x_2$ also falls into case four. However, for every $y_g$ filled into $\xi_i$, for some $x_i$ where $3 \leq i \leq n - k_2$, $x_2 y_g x_1 y_g x_2 \leq \Sigma$ or $x_1 y_g x_2 \leq \Sigma$. Without contradiction, though, this means that $y_g \in \hat{F}$ for all $y_g$ filled into $\xi_i$. But then, $x_i$ must be in case four for $3 \leq i \leq n - k_2$, which forces $y_h \in \hat{F}$ for all $y_h$ filled into $\xi_i$. However, this means that $\hat{F} = F$ which contradicts $\hat{F}$ being a proper subset of $F$. If instead, for $y_j \in \hat{F}$, $y_j$ was one of the symbols used in $\xi_i$, including $y_1$ or $y_2$, then either $x_1 y_j x_2 \leq \Sigma$ or $x_1 y_j x_1 \leq \Sigma$ which means that $x_2$ must fall under case two, three, or four. There exists $y_j \in F$ such that $y_j$ was used when filling in $\xi_i$ or $\xi_j$ and $x_1 y_j x_2 \leq \Sigma$. To avoid the contradiction of $x_2$ also being in case one, it must be that $y_j \in F$. But then $x_2$ must be in case four, and so the rest of the chain of reasoning above applies, leading to a contradiction.

Thus, it must be that $\sigma_F$ is the unique worthwhile $\sigma$-deletion of the string $\Sigma$ which was created. By construction, $B = \{y_1, \ldots, y_{k_1}, x_{n-k_2+1}, \ldots, x_{n-k_2}\}$, $A = \{x_1, \ldots, x_{n-k_2}\}$, and $F = \{y_1, \ldots, y_{k_1}\}$, with $|F| = k_1$ and $|B \setminus A| = k_2$.

Hence, for $k_1, k_2$, and $n$ with $1 \leq k_1 \leq k_2 \leq n - 2$ and $k_1 \leq 1 + \sum_{i=k_2-k_1}^{n-2-k_1} 2^i$ (but not $n = 4$ and $k_1 = k_2 = 2$ or $n = 7$ and $k_1 = k_2 = 4$), there exists an unavoidable string $\Sigma$ such that $\alpha(\Sigma) = n$ and $\Sigma$ has a unique worthwhile $\sigma$-deletion $\sigma_F$ where $|\sigma_F| = k_1/k_2$. \hfill \Box

Note that if $k_1 \leq 1 + 2^{n-2-k_1}$, then $k_1 \leq 1 + \sum_{i=k_2-k_1}^{n-2-k_1} 2^i$. Hence, for a given $k_1$, if $(k_1/n - 2)_n$ is nonempty, then for all $k_2 = k_1, \ldots, n - 2$, $(k_1/k_2)_n$ is nonempty. Also, suppose that $k_1 = k_2$ satisfies the bound, so that $k_1 \leq 1 + \sum_{i=0}^{n-2-k_1} 2^i = 2^{n-1-k_1}$. Then we have that

\[ k_1 - 1 \leq 2^{n-1-k_1} - 1 < 1 + \sum_{i=(n-2)-(k_1-1)}^{(n-2)-(k_1-1)} 2^i \]

so that there is at most one $k_1$ for which $(k_1/k_1)_n$ is nonempty but $(k_1/n - 2)_n = \emptyset$.

As a function of $n$, the maximal possible $(k_1/k_2)_n$ with $k_1 = k_2$ under theorem 5 exceeds that given by the results of [5]. Hence, beyond their intrinsic value as a precise quantitative statement about the sizes of unique worthwhile $\sigma$-deletion, these results have even more implications for the algorithmic complexity of deciding string unavoidability by a $\sigma$-deletion reduction.

References