Available online at www.sciencedirect.com
SCIENCE DIRECT®
ACADEMIC PRESS
Journal of Combinatorial Theory, Series A 103 (2003) 337–348
Journal of Combinatorial Theory
Series A
http://www.elsevier.com/locate/jcta

# Reconstruction from subsequences

Miroslav Dudík[a,1] and Leonard J. Schulman[b,2]

[a] *Princeton University, Princeton, NJ 08544, USA*
[b] *California Institute of Technology, Pasadena, CA 91125, USA*

Received 15 July 2002

**Abstract**

We consider the problem of reconstructing an $n$-sequence, given the multiplicities with which $k$-sequences occur as subsequences. We improve the lower bound on $k$ from $\Omega(\log n)$ to $\exp(\Omega(\log^{1/2} n))$.


## 1. Introduction

Fix an alphabet $\Sigma$. A sequence of length $k$ over $\Sigma$, or a $k$-sequence, is an element of $\Sigma^k$; the set of finite sequences over $\Sigma$ is $\Sigma^*$. The $k$-sequence $w = w_1 \dots w_k$ occurs as a subsequence of $x = x_1 \dots x_n \in \Sigma^*$ at $s$ if $s$ is a monotone increasing mapping $s : \{1, \dots, k\} \to \{1, \dots, n\}$ such that $w_j = x_{s(j)}$. Let $\mathcal{N}(w, x)$ be the set of mappings $s$ for which $w$ is a subsequence of $x$ at $s$; let $N(w, x) = |\mathcal{N}(w, x)|$ be the multiplicity with which $w$ occurs in $x$. (The empty sequence is assumed to match once into any sequence.)

**Definition 1.** For $x, y \in \Sigma^*$ write $x \overset{k}{\sim} y$ if $N(w, x) = N(w, y)$ for all $w \in \bigcup_{j=0}^{k} \Sigma^j$.

Observe that $|x| = |y|$ if $x \overset{k}{\sim} y$ for some $k > 0$. Observe also that if $|x| \geqslant k \geqslant |w|$, then $\binom{|x|-|w|}{k-|w|} N(w, x) = \sum_{w' \in \Sigma^k} N(w, w') N(w', x)$. Therefore if $|x| \geqslant k$, the weakened assumption that $N(w, x) = N(w, y)$ for all $w \in \Sigma^k$ suffices to imply $x \overset{k}{\sim} y$.

*E-mail addresses:* mdudik@princeton.edu (M. Dudík), schulman@caltech.edu (L.J. Schulman).
[1] Supported by a Gordon Wu fellowship. Part of this work was performed while the author was at Caltech, supported by a Milton and Jane Mohr scholarship.
[2] Supported in part by the National Science Foundation under Grant No. 0049092 (previously 9876172), and by the Charles Lee Powell Foundation.

0097-3165/03/$ - see front matter © 2003 Published by Elsevier Inc.
doi:10.1016/S0097-3165(03)00103-1

**Definition 2.** $S(k)$ is the least $n$ for which there exist distinct $x, y \in \Sigma^n$ such that $x \overset{k}{\sim} y$.

Without loss of generality, one can assume that $|\Sigma| = 2$ in Definition 2 (as was noted in [8]). Indeed, let $x, y$ satisfying Definition 2 contain more than two distinct characters of the alphabet. Pick two characters in which they differ first; restrict to the two subsequences consisting only of those characters. The subsequences satisfy Definition 2.

**Example 3.** For $k = 1, 2, 3$ we have $S(1) = 2$, $S(2) = 4$, $S(3) = 7$. Pairs of the shortest distinct $k$-equivalent sequences are $01 \overset{1}{\sim} 10$, $0110 \overset{2}{\sim} 1001$, $0111001 \overset{3}{\sim} 1001110$.

The problem of determining $S(k)$ is due to Kalashnik [5]. It is related to the information-theoretic study of noisy *deletion channels* in which characters of a transmitted sequence are randomly (but not necessarily independently) omitted. Kalashnik's problem addresses the variant in which, out of $n$ original characters, $n-k$ are chosen uniformly at random and deleted; the problem amounts to characterizing the greatest loss rate at which it is possible to determine the message using unlimited repeated sampling. This is similar to asking whether ancestral genomes can be inferred from modern specimens, although the genetic process is vastly more complicated than the one we consider, and cannot be exactly modeled mathematically.

The first bounds on $S(k)$ were $2k \leqslant S(k) \leqslant 2^k$, given in [8]. Up to now the best bounds for $S(k)$ were:

$k - 5 < \lfloor \frac{16}{7}\sqrt{S(k)} \rfloor$ due to Krasikov and Roditty [6];

$S(k) \leqslant \phi(k) \approx 1.7 \cdot 1.62^k$, where $\phi(1) = 2, \phi(2) = 5$ and $\phi(n) = \phi(n-1) + \phi(n-2)$ for $n \geqslant 3$, due to Choffrut and Karhumäki [4].

The lower bound on $S(k)$ depends on relatively few of the constraints imposed by subsequences (this is discussed further at the end of the paper). In spite of this, we show that the true asymptotics are closer to the lower bound than might have been suspected just from counting constraints. We improve the upper bound of $\approx 1.7 \cdot 1.62^k$ to:

**Theorem 4.** $S(k) \leqslant \exp(\frac{3+o(1)}{2 \log 3} \log^2 k)$.

*Since this is only an asymptotic bound we provide also the following explicit, albeit cumbersome, bound*:

$$S(k) \leqslant 1.2\Gamma(\log_3 k) 3^{3/2 \log_3^2 k - 1/2 \log_3 k} \quad \text{for } k \geqslant 5.$$

(*A list of best known upper bounds for small values of k is given in Section* 4.)

Thus, $\log_3 S(k) \leqslant (3/2 + o(1))\log_3^2 k$. Reconstruction of $n$-sequences from the multiplicities of $k$-subsequences therefore requires that $k \geqslant 3^{(\sqrt{2/3}-o(1))\log_3^{1/2} n}$.

## 2. Template sequences

Manvel et al. [8] showed that $S(k) \leqslant 2^k$ using the following lemma: if $x, y \in \Sigma^*$ are such that $x \overset{k}{\sim} y$, then $xy \overset{k+1}{\sim} yx$. Starting with the base case $0 \overset{0}{\sim} 1$, their construction yields the pairs

$$\begin{aligned} 01 &\overset{1}{\sim} 10 \\ 0110 &\overset{2}{\sim} 1001 \\ 01101001 &\overset{3}{\sim} 10010110 \\ &\vdots \end{aligned}$$

Note that these pairs are not minimal, as shown by Example 3. The two infinite sequences generated by this process are the prefixes of the Prouhet–Thue–Morse sequence and its complement. The Prouhet–Thue–Morse sequence is defined as $\{s(n)\}_{n=0}^{\infty}$, where for $n = b_N b_{N-1} \ldots b_0$ in binary, $s(n) = b_N + b_{N-1} + \cdots + b_0 \pmod 2$. This sequence has been extensively studied in combinatorics on words, differential geometry and other fields. The survey [1] lists several results.

One way to think of the above construction is that the sequences $x$ and $y$ are combined through a "template" corresponding to the case $01 \overset{1}{\sim} 10$. This perspective is the starting point for our work. The templates in our method will be more general and will take advantage of sequences with "wildcards". In the present section we introduce these sequences and show how to construct short templates. In the next section we will show how to use templates for an inductive construction demonstrating Theorem 4.

**Definition 5.** For an alphabet $\Gamma = \{X, Y\}$ of size 2, and for $k \geqslant r \geqslant 0$, let

$$U_r(k) = \left\{ w \in \bigcup_{j=r}^{k} (\Gamma \cup \{J\})^j : w \text{ has exactly } r \text{ non-} J \text{ characters} \right\}.$$

For $t \geqslant 1$ and $k_1 \geqslant \cdots \geqslant k_t \geqslant t$, let

$$U(k_1, \ldots, k_t) = \bigcup_{1 \leqslant r \leqslant t} U_r(k_r).$$

(We will actually only use the cases $t = 1$, namely $U(k) = U_1(k)$, and $t = 2$, namely $U(k_1, k_2) = U_1(k_1) \cup U_2(k_2)$.)

The character $J \notin \Gamma$ is used as a "wildcard" that matches to any single character of $\Gamma$. More precisely let the matching relation $M \subseteq (\Gamma \cup \{J\}) \times \Gamma$ be given by $M = \bigcup_{A \in \Gamma} \{(A, A) \cup (J, A)\}$. Say that the sequence $w = w_1 \ldots w_k \in (\Gamma \cup \{J\})^k$ occurs as a subsequence of $p = p_1 \ldots p_m \in \Gamma^*$ at $s$ if $s$ is a monotone increasing mapping $s : \{1, \ldots, k\} \to \{1, \ldots, m\}$ such that $(w_j, p_{s(j)}) \in M$ for all $1 \leqslant j \leqslant k$. As before, let $\mathcal{N}(w, p)$ denote the set of mappings $s$ for which $w$ is a subsequence of $p$ at $s$; let $N(w, p) = |\mathcal{N}(w, p)|$ be the multiplicity with which $w$ occurs in $p$.

**Definition 6.** Let $p, q \in \Gamma^*$. Write $p \overset{U_r(k)}{\sim} q$ if $N(w, p) = N(w, q)$ for all $w \in U_r(k)$. For $t \geqslant 1$ and $k_1 \geqslant \cdots \geqslant k_t \geqslant t$, write $p \overset{U(k_1, \ldots, k_t)}{\sim} q$ if $N(w, p) = N(w, q)$ for all $w \in U(k_1, \ldots, k_t)$.

Observe that $p \overset{k}{\sim} q$ implies $p \overset{U_r(k)}{\sim} q$, but that the latter appears to be substantially weaker if $r$ is small, expressing only the constraints due to the $2^r \binom{k+1}{r+1}$ strings of $U_r(k)$, rather than the $2^k$ strings of $\Gamma^k$. At least, however, $|p| = |q|$ if $p \overset{U_r(k)}{\sim} q$ for some $k \geqslant 1$ and any $r \leqslant k$. Of course, $\overset{k}{\sim}$ and $\overset{U_k(k)}{\sim}$ are the same.

**Definition 7.** Let $k_1 \geqslant 1$. Then $S_U(k_1)$ is the least $m \geqslant k_1$ for which there exist distinct $p, q \in \Gamma^m$ such that $p \overset{U(k_1)}{\sim} q$. Similarly, for $k_1 \geqslant \cdots \geqslant k_t \geqslant t$, $S_U(k_1, \ldots, k_t)$ is the least $m \geqslant k_1$ for which there exist distinct $p, q \in \Gamma^m$ such that $p \overset{U(k_1, \ldots, k_t)}{\sim} q$.

**Lemma 8.** *Let* $k_1 \geqslant k_2 \geqslant 2$ *and* $K = k_1^2 + k_2^2(k_2 - 1)/2$. *Then* $S_U(k_1, k_2) \leqslant K(\lg K + \lg\lg K + 1) = (1 + o(1))K \lg K$.

**Proof.** This is obtained by a counting argument. For $w, w' \in (\Gamma \cup \{J\})^*$ let $N_E(w, w')$ be the number of "exact matches" of $w$ into $w'$—in an exact match, $J$ is not treated as a wildcard but must match to another $J$. Then for $p \in \Gamma^*$ of length at least $k_1$, and for $w \in U_1(k_1)$,

$$\binom{|p| - |w|}{k_1 - |w|} N(w, p) = \sum_{w' \in U_1(k_1), |w'| = k_1} N_E(w, w') N(w', p).$$

Similarly for $p \in \Gamma^*$ of length at least $k_2$, and for $w \in U_2(k_2)$,

$$\binom{|p| - |w|}{k_2 - |w|} N(w, p) = \sum_{w' \in U_2(k_2), |w'| = k_2} N_E(w, w') N(w', p).$$

A sufficient condition for $p \overset{U(k_1,k_2)}{\sim} q$ is therefore that $N(w,p) = N(w,q)$ for all $w \in U_1(k_1)$ of length $k_1$ and for all $w \in U_2(k_2)$ of length $k_2$.

Observe that for each $w$ of the first type, the coefficient $N(w,p)$ is between 0 and $\binom{m}{k_1}$; there are $2k_1$ such $w$'s. However, these coefficients are not all independent: pairs which differ only in their non-$J$ character sum to $\binom{m}{k_1}$.

Similarly for each $w$ of the second type, the coefficient $N(w,p)$ is between 0 and $\binom{m}{k_2}$; there are $4\binom{k_2}{2}$ such $w$'s. Again these coefficients are not all independent: pairs which differ only in one non-$J$ character sum to $N(w',p)$ for some $w' \in U_1(k_2) \subseteq U_1(k_1)$.

Therefore the total number of equivalence classes of $\overset{U(k_1,k_2)}{\sim}$ among sequences of length $m$, is at most

$$\left[\binom{m}{k_1}+1\right]^{\binom{k_1}{1}}\left[\binom{m}{k_2}+1\right]^{\binom{k_2}{2}}.$$

For a bound on $S_U(k_1,k_2)$ it is sufficient that this number be less than $2^m$. Choosing $m = \lfloor K(\lg K + \lg\lg K + 1) \rfloor$, where $K = k_1\binom{k_1}{1} + k_2\binom{k_2}{2}$, we obtain the desired inequality:

$$\left[\binom{m}{k_1}+1\right]^{\binom{k_1}{1}}\left[\binom{m}{k_2}+1\right]^{\binom{k_2}{2}} \leqslant m^{k_1\binom{k_1}{1}} m^{k_2\binom{k_2}{2}} = m^K = 2^{K\lg m} < 2^m,$$

because

$$\begin{aligned} K\lg m &\leqslant K[\lg K + \lg(\lg K + \lg\lg K + 1)] \\ &= K\left[\lg K + \lg\lg K + \lg\left(1 + \frac{\lg\lg K + 1}{\lg K}\right)\right] \\ &\overset{(*)}{<} K[\lg K + \lg\lg K + 8/9] \overset{(**)}{<} \lfloor K[\lg K + \lg\lg K + 1] \rfloor = m. \end{aligned}$$

Inequalities $(*)$ and $(**)$ hold for $K \geqslant 9$.

For $K < 9$ we prove the lemma separately: there is only one such case, $k_1 = k_2 = 2$, and in that case $K = 6$ and $S_U(2,2) = S(2) = 4 \leqslant K(\lg K + \lg\lg K + 1)$. □

**Comment.** A similar counting argument directly gives an upper bound on $S(k)$. However, that bound is inferior to the $2^k$ obtained constructively in [8].

## 3. Proof of Theorem 4

The principal step in the proof of the theorem is a construction which combines sequences equivalent with respect to $\overset{k}{\sim}$, according to a template given by sequences equivalent with respect to $\overset{U(2k,k)}{\sim}$.

Let $h : \Gamma^* \to \Sigma^*$ be the map which replaces each $X$ by the sequence $x$, and each $Y$ by the sequence $y$.

**Lemma 9.** *Let* $x, y \in \Sigma^n$, $x \neq y$, *and*

$$x \overset{k}{\sim} y.$$

*Let* $p, q \in \Gamma^m$, $p \neq q$, $\varrho \in \{0,1,2\}$, *and*

$$p \overset{U(2k+\varrho,k+\varrho)}{\sim} q.$$

*Then* $h(p) \neq h(q)$ *and* $h(p) \overset{3k+\varrho}{\sim} h(q)$.

**Corollary 10.** $S(3k+\varrho) \leqslant S(k) S_U(2k+\varrho, k+\varrho)$.

**Proof.** By assumption $x \neq y$, so $h(p)$ and $h(q)$ differ in every interval corresponding to a character distinguishing $p$ from $q$. Hence $h(p) \neq h(q)$. We wish to show that $N(w, h(p)) = N(w, h(q))$ for all $w \in \Sigma^{3k+\varrho}$.

For positive integer $j$ let $[j] = \{1, \ldots, j\}$; for positive integers $j_1$ and $j_2$ denote by $\hat{L}(j_1, j_2)$ the set of strictly increasing functions from $\{0\} \cup [j_1]$ to $\{0\} \cup [j_2]$ which map 0 to 0 and $j_1$ to $j_2$. For a strictly increasing map $\ell$ let $w_{\ell,i}$ denote the sequence $(w_{\ell(i-1)+1}, \ldots, w_{\ell(i)})$. If $|w| = 3k+\varrho$ then each function $\ell \in \hat{L}(t, 3k+\varrho)$ specifies a partition of $w$ into $t$ non-empty segments $w_{\ell,1}, \ldots, w_{\ell,t}$.

Each map $s \in \mathcal{N}(w, h(p))$ (and similarly $s \in \mathcal{N}(w, h(q))$) defines a segmentation $w_1 w_2 \ldots w_m = w$, in which $w_i$ is the preimage of $h(p_i)$. (Some $w_i$ may be empty.) We can classify maps in $\mathcal{N}(w, h(p))$ according to this segmentation, and write $\mathcal{N}(w, h(p))$ as a disjoint union of direct products

$$\mathcal{N}(w, h(p)) = \bigcup_{w_1 w_2 \ldots w_m = w} \prod_{i=1}^{m} \mathcal{N}(w_i, h(p_i)).$$

Next reorganize this union according to the number of non-empty segments $w_i$. For each partition $\ell \in \hat{L}(t, 3k+\varrho)$ of $w$ into $t$ non-empty segments, there are $\binom{m}{t}$ ways (each labeled by an element $r \in \hat{L}(t+1, m+1)$) to map the segments $w_i$ into segments $h(p_i)$:

$$\mathcal{N}(w, h(p)) = \bigcup_{t \geqslant 1} \bigcup_{\ell \in \hat{L}(t, 3k+\varrho)} \bigcup_{r \in \hat{L}(t+1, m+1)} \prod_{i=1}^{t} \mathcal{N}(w_{\ell,i}, h(p_{r(i)})).$$

Convert to an enumeration:

$$N(w,h(p))=\sum_t \sum_{\ell\in\hat{L}(t,3k+\varrho)} \sum_{r\in\hat{L}(t+1,m+1)} \prod_{i=1}^{t} N(w_{\ell,i},h(p_{r(i)}))$$

The sequence $w$ can have at most 2 segments of length greater than $k$. Therefore, we can separate $\hat{L}(t,3k+\varrho)$ into the disjoint union of

$$\hat{L}_0(t,3k+\varrho)=\{\ell:\ell(i)-\ell(i-1)\leqslant k\ \forall i\},$$

$$\hat{L}_1(t,t',3k+\varrho)=\{\ell:\ell(t')-\ell(t'-1)>k \text{ and } \ell(i)-\ell(i-1)\leqslant k\ \forall i\neq t'\},$$

$$\text{where } 1\leqslant t'\leqslant t,$$

$$\hat{L}_2(t,t',t'',3k+\varrho)=\{\ell:\ell(t')-\ell(t'-1)>k,\ell(t'')-\ell(t''-1)>k,$$

$$\text{and } \ell(i)-\ell(i-1)\leqslant k\ \forall i\neq t',i\neq t''\},$$

$$\text{where } 1\leqslant t'<t''\leqslant t.$$

Observe that for $i$ such that $|w_{\ell,i}|=\ell(i)-\ell(i-1)\leqslant k$, it does not matter whether $h(p_{r(i)})=x$ or $h(p_{r(i)})=y$: in either case, $N(w_{\ell,i},h(p_{r(i)}))=N(w_{\ell,i},x)$.

$$\begin{aligned}N(w,h(p))=&\sum_t \sum_{\ell\in\hat{L}_0(t,3k+\varrho)} \sum_{r\in\hat{L}(t+1,m+1)} \prod_{i=1}^{t} N(w_{\ell,i},x)\\&+\sum_{t,t'} \sum_{\ell\in\hat{L}_1(t,t',3k+\varrho)} \sum_{r\in\hat{L}(t+1,m+1)} N(w_{\ell,t'},h(p_{r(t')})) \prod_{i\in[t]-\{t'\}} N(w_{\ell,i},x)\\&+\sum_{t,t',t''} \sum_{\ell\in\hat{L}_2(t,t',t'',3k+\varrho)} \sum_{r\in\hat{L}(t+1,m+1)} N(w_{\ell,t'},h(p_{r(t')}))N(w_{\ell,t''},h(p_{r(t'')}))\\&\times \prod_{i\in[t]-\{t',t''\}} N(w_{\ell,i},x).\end{aligned}$$

In the summation over $r$ in the first line, the argument does not depend on $r$, so the summation can be replaced by a multiplication by the factor $|\hat{L}(t+1,m+1)|=N(J^t,p)$. In the summation over $r$ in the second line, the argument depends on $r$ only through the value of $t'$ and the number of ways in which the segment $w_{\ell,t'}$ can map into the interval $h(p_{r(t')})$, which is to say, $N(w_{\ell,t'},h(p_{r(t')}))$. The latter in turn depends only on whether $p_{r(t')}$ is an $X$ or a $Y$; so we can rewrite the line by summing separately over the mappings consistent with each possibility. The third line can be

rewritten in a similar manner. So we have:

$$\begin{aligned}
N(w,h(p)) = &\sum_t \sum_{\ell\in\hat{L}_0(t,3k+\varrho)} N(J^t,p)\prod_{i=1}^{t} N(w_{\ell,i},x)\\
&+\sum_{t,t'}\sum_{\ell\in\hat{L}_1(t,t',3k+\varrho)}\sum_{A\in\{X,Y\}} N(J^{t'-1}AJ^{t-t'},p)N(w_{\ell,t'},h(A))\\
&\times\prod_{i\in[t]-\{t'\}} N(w_{\ell,i},x)\\
&+\sum_{t,t',t''}\sum_{\ell\in\hat{L}_2(t,t',t'',3k+\varrho)}\sum_{A,B\in\{X,Y\}}\\
&N(J^{t'-1}AJ^{t''-t'-1}BJ^{t-t''},p)N(w_{\ell,t'},h(A))N(w_{\ell,t''},h(B))\\
&\times\prod_{i\in[t]-\{t',t''\}} N(w_{\ell,i},x).
\end{aligned}$$

By assumption $p \overset{U(2k+\varrho,k+\varrho)}{\sim} q$, so that $N(J^t,p)=N(J^t,q)$ (this is simply $\binom{m}{t}$), $N(J^{t'-1}AJ^{t-t'},p)=N(J^{t'-1}AJ^{t-t'},q)$, and $N(J^{t'-1}AJ^{t''-t'-1}BJ^{t-t''},p)=N(J^{t'-1}AJ^{t''-t'-1}BJ^{t-t''},q)$. Hence $N(w,h(p))=N(w,h(q))$, and so $h(p) \overset{3k+\varrho}{\sim} h(q)$. □

We can now complete the proof of the theorem. Setting $K=(2k+\varrho)^2+(k+\varrho)^2(k+\varrho-1)/2=\frac{1+o(1)}{2}k^3$ in Lemma 8, we obtain

$$S_U(2k+\varrho,k+\varrho)\leqslant\frac{1+o(1)}{2}k^3\lg\left(\frac{1+o(1)}{2}k^3\right)\leqslant C(k)k^3\log_3 k, \tag{1}$$

where $C(k)=\frac{3\lg 3}{2}+o(1)$, or more precisely, $C(k)\leqslant 10$ for $k\geqslant 9$, and $C(k)\leqslant 3$ for $k\geqslant 3^5$.

Using the inequality $S(3k+\varrho)\leqslant S(k)S_U(2k+\varrho,k+\varrho)$ and bound (1) we establish the theorem by setting $k_0=k$ and, for $i>0$, $k_i=\lfloor k_{i-1}/3\rfloor$, stopping the series with the $i=i_0$ for which $k_{i_0}\leqslant 28$, i.e., $i_0=\lceil\log_3(k/28)\rceil$.

$$S(k)\leqslant S(k_{i_0})\prod_{i=1}^{i_0} S_U(2k_i+\varrho_i,k_i+\varrho_i) \tag{2}$$

$$\begin{aligned}
&\leqslant S(k_{i_0})\prod_{i=1}^{i_0} C(k_i)(k/3^i)^3\log_3(k/3^i),\\
&=3^{O(1)+\sum_{i=1}^{\lceil\log_3(k/28)\rceil}[O(1)+3(\log_3 k-i)+\log_3(\log_3 k-i)]}\\
&=3^{O(1)+O(\log k)+3/2\log_3^2 k+O(\log k)+O(\log k\log\log k)}\\
&=3^{(3/2+o(1))\log_3^2 k}.
\end{aligned} \tag{3}$$

In (2) we can bound $S(k_{i_0})$ by the best known upper bound (given in the next section), and $S_U(2k_i+\varrho_i, k_i+\varrho_i)$ by Lemma 8. The resulting upper bound improves on the previous best bound, $\phi(k)$, for $k \geqslant 29$. For a closed-form result, however, we offer the following:

$$S(k) \leqslant 1.2\Gamma(\log_3 k)3^{3/2\log_3^2 k - 1/2\log_3 k} \quad \text{for } k \geqslant 5. \tag{4}$$

The proof proceeds by induction on $k$. For $5 \leqslant k < 3^6$, we explicitly check that this bound is weaker than (2). For $k' \geqslant 3^6$, we let $k' = 3k+\varrho$, $\varrho \in \{0,1,2\}$ and prove the inductive step using Corollary 10 and bound (1):

$$\begin{aligned}
S(k') &= S(3k+\varrho) \leqslant S(k)S_U(2k+\varrho, k+\varrho) \\
&\leqslant S(k)C(k)k^3\log_3 k \quad \text{by (1)} \\
&\leqslant 1.2\Gamma(\log_3 k)3^{3/2\log_3^2 k - 1/2\log_3 k}3k^3\log_3 k \quad \text{by induction} \\
&= 1.2\Gamma(1+\log_3 k)3^{3/2\log_3^2 k - 1/2\log_3 k + 1 + 3\log_3 k} \\
&= 1.2\Gamma(1+\log_3 k)3^{3/2(1+\log_3 k)^2 - 1/2(1+\log_3 k)} \\
&\leqslant 1.2\Gamma(\log_3 k')3^{3/2\log_3^2 k' - 1/2\log_3 k'}.
\end{aligned}$$

Note that we assumed that $C(k) \leqslant 3$ since $k \geqslant 3^5$.

## 4. Small values of $k$

The following table contains the best known upper bounds on $S(k)$ for small values of $k$. Exact values of $S(k)$ for $k = 1, \ldots, 5$ are quoted from [7], the bound $S(6) \leqslant 30$ is from [8]. The best upper bound for $7 \leqslant k \leqslant 28$ is $S(k) \leq \phi(k)$, and for $k \geqslant 29$, it is bound (2).

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $S(k) \leq$ | 2 | 4 | 7 | 12 | 16 | 30 | 50 | 81 | 131 | 212 | 343 | 555 | 898 | 1 453 | 2 351 | 3 804 |
| | actual values of $S(k)$, [7] | | | | | [8] | $\phi(k)$ | | | | | | | | | |

| $k$ | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|
| $S(k) \leq$ | 6 155 | 9 959 | 16 114 | 26 073 | 42 187 | 68 260 | 110 447 | 178 707 | 289 154 | 467 861 |
| | $\phi(k)$ | | | | | | | | | |

| $k$ | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 |
|---|---|---|---|---|---|---|---|---|
| $S(k) \leq$ | 757 015 | 1 224 876 | 1 881 422 | 2 525 132 | 3 183 604 | 3 972 244 | 5 387 501 | 6 679 239 |
| | $\phi(k)$ | | bound (2) | | | | | |

The explicit bound (4) is much worse than the best known upper bounds for small values of $k$. The following table contains better bounds for $k \leqslant 85$. For $1 \leqslant k \leqslant 28$ we use the explicit form of $\phi(k) = O(1.62^k)$. For $29 \leqslant k \leqslant 84$ we use Corollary 10, with bounds $S(k/3) \leqslant \phi(k/3) = O(1.62^{k/3})$ and $S_U(2k3/3, k/3) = O(k^3 \log k)$. Leading coefficients are chosen to give bounds weaker than (2).

| $1 \leqslant k \leqslant 28$ | $29 \leqslant k \leqslant 84$ | $k \geqslant 85$ |
|---|---|---|
| $S(k) \leqslant 1.75 \cdot 1.62^k$ | $S(k) \leqslant 0.25 \cdot 1.17^k k^3 \log k$ | $S(k) \leqslant 1.2\Gamma(\log_3 k) 3^{3/2 \log_3^2 k - 1/2 \log_3 k}$ |

## 5. Discussion

### 5.1. The lower bound

In Section 2, we pointed out that $\overset{k}{\sim}$ implies $\overset{U_r(k)}{\sim}$ for all $0 \leqslant r \leqslant k$; in particular, it implies $\overset{U_1(k)}{\sim}$. Thus $S_U(k) \leqslant S(k)$. It seems likely that this is only a weak lower bound, if only because there are merely $k(k+1)$ sequences in $U_1(k)$ to impose constraints, as compared with $2^k$ sequences in $\Gamma^k$. Krasikov and Roditty [6] show that $S_U(k) \geq \Omega(k^2)$ while the present paper shows that $S(k) \leqslant k^{O(\log k)}$. Hence $S(k)$ grows much more slowly than would be predicted by a naïve dependence on the number of constraints. We briefly review the basis of the lower bound. In this section we assume that $\Gamma = \{0, 1\}$.

Let $p, q \in \{0,1\}^m$. From the proof of Lemma 8, it follows that $p \overset{U_1(k)}{\sim} q$ if and only if $N(w, q) = N(w, p)$ for all $w \in U_1(k)$, $|w| = k$; in fact it suffices to consider $w \in \{1, J\}^*$, because $N(J^h 1 J^{k-1-h}) = \binom{m}{k} - N(J^h 0 J^{k-1-h})$. The multiplicities $N(w, p)$ for $w = J^h 1 J^{k-1-h}$, $h = 0, 1, \ldots, k-1$, can be expressed as

$$N(w,p) = \sum_{i=1}^{m} \binom{i-1}{h} p_i \binom{m-i}{k-1-h}.$$

Polynomials $\binom{i-1}{h}\binom{m-i}{k-1-h}$ in $i$ are linearly independent of degree $k-1$, so they form a basis for the space of polynomials of degree at most $k-1$. Hence, $N(w,p) = N(w,q)$ for all $w \in U_1(k)$ if and only if

$$\sum_{1 \leqslant i \leqslant m} p_i i^h = \sum_{1 \leqslant i \leqslant m} q_i i^h \quad \text{for } 0 \leqslant h \leqslant k-1.$$

Finding distinct $m$-tuples $p_1, \ldots, p_m \in \{0,1\}$ and $q_1, \ldots, q_m \in \{0,1\}$ for some $m$ is then equivalent to solving the system

$$u_1^h + u_2^h + \cdots + u_s^h = v_1^h + v_2^h + \cdots + v_s^h, \quad h = 1, \ldots, k-1$$
$$1 \leqslant u_1 < u_2 < \ldots < u_s \leqslant m, \quad 1 \leqslant v_1 < v_2 < \ldots < v_s \leqslant m \tag{5}$$

for distinct $s$-tuples $u_1, \ldots, u_s$ and $v_1, \ldots, v_s$. This is known as the Prouhet–Tarry–Escott problem, which has been studied in Diophantine analysis [3]. The lower bound on $S_U(k)$ follows from the result in [2], which states that $k \leqslant \lfloor \frac{16}{7}\sqrt{m} \rfloor + 4$ for all solutions of (5).

The method of expressing constraints in $U_1(k)$ as polynomials can be generalized to $U_r(k)$. We obtain that $p \overset{U_r(k)}{\sim} q$ if and only if

$$\sum_{1 \leqslant i_1 < i_2 < \cdots < i_t \leqslant m} p_{i_1} i_1^{h_1} p_{i_2} i_2^{h_2} \cdots p_{i_t} i_t^{h_t} = \sum_{1 \leqslant i_1 < i_2 < \cdots < i_t \leqslant m} q_{i_1} i_1^{h_1} q_{i_2} i_2^{h_2} \cdots q_{i_t} i_t^{h_t}$$
$$\text{for } 1 \leqslant t \leqslant r, 0 \leqslant h_j, h_1 + h_2 + \cdots + h_t \leqslant k - t,$$

however, we do not know of a bound analogous to that of [2] for any $r \geqslant 2$.

### 5.2. General form of the upper bound method

Lemma 9 is a special case of the following, which is proven in the same manner:

**Lemma 11.** *Fix $t \geqslant 1$. Let $x, y \in \Sigma^*$, $x \neq y$, and $x \overset{k}{\sim} y$. Let $p, q \in \Gamma^*$, $p \neq q$, and suppose that $p \overset{U(k_1,\ldots,k_t)}{\sim} q$ for $k_r = (t-r)(k+1) + k + r$. Then $h(p) \neq h(q)$ and $h(p) \overset{t(k+1)+k}{\sim} h(q)$.*

This lemma gives an inductive bound on $S(k)$ analogous to Corollary 10. The case $t = 1$ (in other words the case in which we just use the best known upper bound on $S_U(k)$, namely $S_U(k) \in O(k^2 \log k)$) is already sufficient for a bound on $S(k)$ of the form $S(k) \leqslant \exp(c \log^2 k)$, albeit for a value of $c$ inferior to that derived above. For the known upper bounds on $S_U$, the case $t = 2$ gives a stronger upper bound on $S$ than any other value of $t$.

### 5.3. Open questions

Beyond the self-evident matter of closing the remaining gap between the upper and lower bounds on $S(k)$, several questions have caught our attention.

1. Does $\overset{U_r(h)}{\sim}$ imply $\overset{k}{\sim}$ for any $k > r+1$? Note that $\overset{U_{k-1}(k)}{\sim}$ implies $\overset{k}{\sim}$.
2. It is known that $\Omega(k^2) \leqslant S_U(k) \leqslant O(k^2 \log k)$, hence a tight bound on $S_U(k)$ will not improve the asymptotic bound on $S(k)$. However, for $t \geqslant 2$, the best bounds

are $\Omega(k^2) \leqslant S_U(tk+t,(t-1)k+t,\ldots,k+t) \leqslant O(k^{t+1}\log k)$. Is it possible to improve these bounds? Improving the upper bound on $S_U(tk+t,\ldots,k+t)$ could yield a better bound on $S(k)$ with the present method, and improving the lower bound on $S_U(k,\ldots,k)$ ($t$ arguments) for any $t \geqslant 1$ would yield a better lower bound on $S(k)$.

3. Is there an effective algorithm to find a string having the multiplicity list $(N(w,x))_{w \in \cup_{j=0}^{k} \Sigma^j}$? "Effective" may be taken to mean polynomial time in $2^k + n$. Manvel et al. [8] give an algorithm for $k > n/2$ which examines $O(n)$ entries in the multiplicity list. How many entries need to be examined for general $k$?
4. What can be said about higher-dimensional versions of this question? A natural version in $d$ dimensions is that for $k_1,\ldots,k_d > 0$, $N(w,x)$ is the multiplicity with which each "rectangular word" $w \in \Sigma^{k_1 \times \cdots \times k_d}$ occurs as a subrectangle (with any spacing) of the rectangular word $x \in \Sigma^{n_1 \times \cdots \times n_d}$.

## Acknowledgments

Thanks to Sridhar Rajagopalan for introducing us to this topic.

## References

[1] J.P. Allouche, J. Shallit, The ubiquitous Prouhet–Thue–Morse sequence, in: Sequences and their Applications (Singapore, 1998), Springer Series in Discrete Mathematics and Theoretical Computer Science, Springer, London, 1999, pp. 1–16.
[2] P. Borwein, T. Erdélyi, G. Kós, Littlewood-type problems on [0, 1], Proc. London Math. Soc. (3) 79 (1) (1999) 22–46.
[3] P. Borwein, C. Ingalls, The Prouhet–Tarry–Escott problem revisited, Enseign. Math. (2) 40 (1–2) (1994) 3–27.
[4] C. Choffrut, J. Karhumäki, Combinatorics of words, in: G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, Vol. I, Springer, Berlin, 1997, pp. 329–438.
[5] L.O. Kalashnik, The reconstruction of a word from fragments, Numerical Mathematics and Computer Technology, Akad. Nauk. Ukrain. SSR Inst. Mat., Preprint IV: 56–57, 1973.
[6] I. Krasikov, Y. Roditty, On a reconstruction problem for sequences, J. Combin. Theory Ser. A 77 (2) (1997) 344–348.
[7] J. Maňuch, Characterization of a word by its subwords, in: G. Rozenberg, W. Thomas (Eds.), Developments in Language Theory: Foundations, Applications, and Perspectives, Proceedings of the Fourth International Conference, World Scientific, Singapore, 2000, pp. 210–219.
[8] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, P. Stockmeyer, Reconstruction of sequences, Discrete Math. 94 (1991) 209–219.