Reconstructing sequences

A.D. Scott

Department of Pure Mathematics
and Mathematical Statistics,
University of Cambridge,
16 Mill Lane, Cambridge, CB2 1SB, England.

Abstract. We prove that every sequence of length n can be reconstructed from the multiset of all its subsequences of length k, provided $k \geq (1 + o(1))\sqrt{n \log n}$. This is a substantial improvement on previous bounds.

§1. Introduction

The large amount of research on the Graph Reconstruction Conjecture of S. Ulam and P. Kelley has led to interest in reconstruction problems for various other combinatorial structures (such as digraphs and posets). In this paper, we consider the reconstruction problem for sequences. A sequence S of length n contains $\binom{n}{k}$ subsequences of length k; the multiset of these sequences is called the *k-deck* of S (our notation follows [6]). A sequence that is uniquely defined by its k-deck is called *k-reconstructible*. Thus S is k-reconstructible iff no other sequence has the same k-deck as S. For instance, 1001 is not 2-reconstructible, since it has the same 2-deck as 0110. However, it is easily seen that all sequences of length 4 are 3-reconstructible, and a few moments' thought shows that all sequences of length n are $(n-1)$-reconstructible. In fact, it is easy to prove that all sequences of length n are $(\lfloor n/2 \rfloor)+1$-reconstructible, by considering subsequences of length $\lfloor n/2 \rfloor + 1$ which contains all occurrences of whichever symbol occurs fewest times.

The problem of determining for which k every sequence of length n can be reconstructed from its k-deck was raised by Kalashnik [3], who apparently proved that every sequence can be reconstructed from its $\lfloor n/2 \rfloor$-deck (see [6] for this, and for an incorrect assertion claimed by Aleksanjan [1]). Zenkin and Leont'ev [10] proved that we may be unable to reconstruct with $k = \log n / \log\log n$; they also gave some related results, including the fact that if $k = o(n)$ then almost every sequence cannot be reconstructed from the set of sequences (without multiplicity) found in its k-deck. Recently, Manvel, Meyerowitz, Schwenk, Smith and Stockmeyer [6] gave another proof that it is possible to reconstruct from the $\lfloor n/2 \rfloor$-deck, for $n \geq 7$, and proved that it is not necessarily possible to reconstruct from the $\log n$-deck; Schwenk [9] has improved the lower bound by a construction giving $\frac{5}{4}\log n$ for sufficiently large n. Leont'ev and Smetanin [5] remarked that determining whether a given vector can be uniquely reconstructed from a given set of subsequences is an NP-complete problem; Kubicka and Schwenk [4] have also investigated algorithmic aspects of the problem, and calculated precise bounds for small values of k.

It will be useful to define some notation. For a positive integer n, let $f(n)$ be the smallest k such that every sequence of length n is k-reconstructible. Let us note

that, for $k < l$, it is easy to deduce a sequence's k-deck from its l-deck, so $f(n)$ is nondecreasing. The bounds in [6] and [9] are

$$\frac{5}{4}\log_2(n) < f(n) \leq \left\lfloor \frac{n}{2} \right\rfloor, \tag{1}$$

for $n \geq 7$.

In the main result of this paper, we improve the upper bound substantially to

$$f(n) \leq (1 + o(1))\sqrt{n \log n}$$

(all logarithms will be natural, unless otherwise indicated). We remark that this is in sharp contrast to the case for graph reconstruction. Indeed, Nýdl [8] has shown that for every $c \in (0, 1)$, there exist graphs that cannot be reconstructed from their $\lfloor cn \rfloor$-decks.

Note that, as observed in [6], every sequence of length n is k-reconstructible iff every binary sequence of length n is k-reconstructible, since we can always choose to ignore the difference between certain symbols. We shall therefore assume that all our sequences are *binary* sequences.

We prove our result in two stages. We first show that in order to reconstruct the binary sequence $\mathbf{a} = (a_1, \ldots, a_n)$ it is enough to know the values of certain polynomials in $a_1, \ldots, a_n$, and then show that if k is large enough we can deduce these values from the k-deck of $\mathbf{a}$. We give the proof in §2, except for the proof of Lemma 1, which is given in §3. We make some further remarks in §4, and indicate how the same methods might be applied to the reconstruction of permutations and matrices.

§2. Main Results

We begin with a problem that is closely related to the problem of reconstructing a sequence from its subsequences of a given length. Given nonnegative integers $s_0, \ldots, s_k$, consider the equations

$$\begin{aligned} \sum_{j=1}^{n} a_j j^0 &= s_0 \\ \sum_{j=1}^{n} a_j j^1 &= s_1 \\ &\vdots \\ \sum_{j=1}^{n} a_j j^{k-1} &= s_{k-1}, \end{aligned} \tag{2}$$

where we demand $\mathbf{a} = (a_1, \ldots, a_n) \in \{0,1\}^n$. Under what conditions do the integers $s_0, \ldots, s_{k-1}$ uniquely determine $\mathbf{a}$?

Let us put this more formally. For a sequence of integers $\mathbf{a} = (a_1, \ldots, a_n)$, let

$$s_i(\mathbf{a}) = \sum_{j=1}^{n} a_j j^i,$$

and let $S_k(\mathbf{a})$ be the sequence $(s_0(\mathbf{a}), \ldots, s_{k-1}(\mathbf{a}))$; note that S_k is a linear function from $\mathbb{Z}^n$ to $\mathbb{Z}^k$. We define $f^*(n)$ to be the largest integer k such that we can solve $S_k(\mathbf{a}) = S_k(\mathbf{b})$ with distinct $\mathbf{a}, \mathbf{b} \in \{0,1\}^n$. Equivalently, k is the largest integer such that $S_k(\mathbf{a}) = \mathbf{0}$ has a non-zero solution with $\mathbf{a} \in \{-1,0,1\}^n$. Thus for $l > k$, we can reconstruct $\mathbf{a} \in \{0,1\}^n$ from $S_l(\mathbf{a})$.

In §3, we shall prove the following bounds for $f^*(n)$.

Lemma 1.

$$(1+o(1))\sqrt{\frac{2n}{\log_2 n}} \le f^*(n) \le (1+o(1))\sqrt{n \log n}. \tag{3}$$

□

Our main result now follows immediately, since we shall show that $S_k(\mathbf{a})$ can easily be calculated from the multiset of subsequences of $\mathbf{a}$ with length k.

Theorem 2. *For positive integers n, we have*

$$f(n) \leq (1 + o(1))\sqrt{n \log n}$$

.

Proof. Suppose we are given the k-deck of a sequence $\mathbf{a} = (a_1, \ldots, a_n)$. For $i \leq k$, let n_i be the number of subsequences of $\mathbf{a}$ of length i that terminate with a 1. Thus

$$n_i = \sum_{j=1}^{n} a_j \binom{j-1}{i-1} = \sum_{j=1}^{n} a_j p_i(j),$$

where $p_i(x)$ is a polynomial of degree $i-1$. It is easily checked that the polynomials $p_1(x), \ldots, p_{i+1}(x)$ form a basis for the space of polynomials of degree at most i; in particular, the polynomial x^i is in their span. Thus by taking a linear combination of the n_i we can determine the value of $s_i(\mathbf{a}) = \sum_{j=1}^{n} a_j j^i$, for $i < k$. Therefore we can reconstruct $\mathbf{a}$, provided that $k > f^*(n)$; the result follows from Lemma 1. □

§3. Proof of Lemma 1

In this section we give a proof of Lemma 1.

(i) We begin with the lower bound. Suppose $n > k > 1$ and $S_k(\mathbf{a}) = S_k(\mathbf{b})$ is not solvable with distinct $\mathbf{a}, \mathbf{b} \in \{0,1\}^n$. Note that

$$s_i(\mathbf{a}) \leq \sum_{j=1}^{n} j^i \leq \frac{(n+1)^{i+1}}{(i+1)}.$$

Therefore, since $k < n$, there are at most

$$\prod_{i=0}^{k-1} \frac{(n+1)^{i+1}}{(i+1)} < (n+1) \prod_{i=1}^{k-1} n^{i+1} < n^{(k+2)^2/2}$$

possible values for $S_k(\mathbf{a})$. However, there are 2^n sequences in $\{0,1\}^n$, so if $S_k(\mathbf{a}) = S_k(\mathbf{b})$ is not solvable with distinct $\mathbf{a}, \mathbf{b} \in \{0,1\}^n$ then we must have

$$n^{(k+2)^2/2} \geq 2^n,$$

and so

$$(k+2)^2 \geq \frac{2n}{\log_2 n},$$

which gives the lower bound in (3).

(ii) For the upper bound, let $\epsilon > 0$ and let $k = \lceil (1+\epsilon)\sqrt{n \log n} \rceil$. We show that $f^*(n) \leq k$ for sufficiently large n. We use standard elementary number theoretic results (see [2]).

Suppose $\mathbf{a} \in \{0,1\}^n$. For positive integers i, j define

$$n_{i,j}(\mathbf{a}) = \sum_{s \equiv i \mod j} a_s. \tag{4}$$

Thus $n_{0,1}(\mathbf{a}) = \sum_{i=1}^n a_i$ and $n_{i,n}(\mathbf{a}) = a_i$. We claim that, for any prime p and any integer i,

$$n_{i,p}(a) \equiv s_0(\mathbf{a}) - \sum_{j=0}^{p-1} \binom{p-1}{j} s_j(\mathbf{a})(-i)^{p-1-j} \qquad \mod p. \tag{5}$$

Indeed, since p is prime, $i^{p-1} \equiv 1 \bmod p$ for any $i \not\equiv 0$. Thus

$$\begin{aligned} s_0(\mathbf{a}) - \sum_{j=0}^{p-1} \binom{p-1}{j} s_j(\mathbf{a})(-i)^{p-1-j} &= \sum_{r=1}^n a_r - \sum_{j=0}^{p-1} (-i)^{p-1-j} \binom{p-1}{j} \sum_{r=1}^n a_r r^j \\ &= \sum_{r=1}^n a_r - \sum_{r=1}^n a_r \sum_{j=0}^{p-1} \binom{p-1}{j} r^j (-i)^{p-1-j} \\ &= \sum_{r=1}^n a_r - \sum_{r=1}^n a_r (r-i)^{p-1} \\ &\equiv \sum_{r \equiv i \mod p} a_r \qquad \mod p, \end{aligned}$$

as claimed.

It is clear from (4) that $0 \leq n_{i,p}(\mathbf{a}) \leq \lceil n/p \rceil$, so if $p > \sqrt{n} + 1$ then we have $0 \leq n_{i,p}(\mathbf{a}) < p$ and we can therefore determine $n_{i,p}(\mathbf{a})$ from (5). Thus we can calculate $n_{i,p}(\mathbf{a})$ from $S_k(\mathbf{a})$ for all primes p with $\sqrt{n}+1 < p < k$ and all integers i.

Now define the vector

$$\mathbf{v}_{i,j} = (v_{i,j}^{(1)}, \ldots, v_{i,j}^{(n)})$$

by

$$v_{i,j}^{(r)} = \begin{cases} 1 & \text{if } r \equiv i \bmod j \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

and let $\tilde{\mathbf{v}}_{i,j} = (\tilde{v}_{i,j}^{(r)})_{r=1}^{\infty}$ be the extension of this to all positive integers, defined in the same way. Note that $\mathbf{v}_{i,j} = \mathbf{v}_{i',j}$ or $\tilde{\mathbf{v}}_{i,j} = \tilde{\mathbf{v}}_{i',j}$ iff $i \equiv i' \bmod j$.

It is clear that, for any i, j, we have

$$n_{i,j}(\mathbf{a}) = \mathbf{a} \cdot \mathbf{v}_{i,j}.$$

Now suppose that $\mathbf{a}$, $\mathbf{b}$ are distinct vectors in $\{0,1\}^n$ with $S_k(\mathbf{a}) = S_k(\mathbf{b})$. As we have noted, if $\sqrt{n} + 1 < p < k$ and i is any integer, then $n_{i,p}(\mathbf{a})$ and $n_{i,p}(\mathbf{b})$ are uniquely defined by (5), and so we must have $n_{i,p}(\mathbf{a}) = n_{i,p}(\mathbf{b})$. Therefore

$$\mathbf{a} \cdot \mathbf{v}_{i,p} = \mathbf{b} \cdot \mathbf{v}_{i,p}. \tag{7}$$

We now work over $\mathbb{F}_2$. In order to prove the required bound it is enough to show that (7) implies $\mathbf{a} = \mathbf{b}$ for sufficiently large n (dependent on ϵ). We do this by showing that the set of vectors

$$S = \{\mathbf{v}_{i,p} : p \text{ prime}, \sqrt{n} + 1 < p < k, 0 \le i \le p - 1\} \tag{8}$$

spans $\mathbb{F}_2^n$. Then $\mathbf{a} \cdot \mathbf{v} = \mathbf{b} \cdot \mathbf{v}$ for any $\mathbf{v} \in \{0,1\}^n$, so $\mathbf{a} = \mathbf{b}$.

Clearly S is not in general an independent set. Let T be the subset of S defined by

$$T = \{\mathbf{v}_{i,p} : p \text{ prime}, \sqrt{n} + 1 < p < k, 1 \le i \le p - 1\} \tag{9}$$

Suppose first that the elements of T are not linearly independent, so that

$$\sum_{j=1}^{m} \mathbf{v}_{i_j,p_j} = \mathbf{0}$$

for some $m > 0$ and distinct $(i_1, p_1), \ldots, (i_m, p_m)$. Note that, by (9), $i_j \not\equiv 0 \bmod p_j$, for $j = 1, \ldots, m$. By the Chinese Remainder Theorem, we can find an integer $r > 0$ such that

$$r \equiv i_1 \qquad \bmod p_1$$

and, for each j such that $p_j \neq p_1$,

$$r \equiv 0 \qquad \mod p_j.$$

Therefore $\tilde{v}^{(r)}_{i_1,p_1} = 1$ and $\tilde{v}^{(r)}_{i_j,p_j} = 0$ for $j = 2, \dots, m$ (note that if $p_j = p_1$ then $i_j \not\equiv i_1 \bmod p_1$, so $\tilde{v}^{(r)}_{i_j,p_j} = 0$), and so, defining

$$\tilde{\mathbf{v}} = \sum_{j=1}^{m} \tilde{\mathbf{v}}_{i_j,p_j},$$

we have

$$\tilde{\mathbf{v}}^{(r)} = \sum_{j=1}^{m} \tilde{v}^{(r)}_{i_j,p_j} \neq 0.$$

Let s be the smallest positive integer such that $\tilde{\mathbf{v}}^{(s)} = 1$; clearly $s > n$. Then, for $1 \leq t \leq n$, consider the vector $\mathbf{v}_t$ defined by

$$\mathbf{v}_t = \sum_{j=1}^{m} \mathbf{v}_{i_j - s + t, p_j}.$$

In effect, we shift $\tilde{\mathbf{v}}$ to the left by $s - t$ places. We get

$$\mathbf{v}_t^{(i)} = 0$$

for $1 \leq i < t$ and

$$\mathbf{v}_t^{(t)} = 1.$$

Thus $\mathbf{v}_1, \dots, \mathbf{v}_n$ are in the span of S (though not necessarily of T) and span $\mathbb{F}_2^n$.

The other possibility is that the elements of T are independent, in which case we must have $|T| \leq n$. Let P be the set of primes p with $\sqrt{n} + 1 < p < k$. Then

$$|T| = \sum_{p \in P} (p - 1) = (1 + o(1)) \sum_{p \in P} p.$$

It follows from the Prime Number Theorem (see (22.19.1) in [2]) that, for $\eta > 0$,

$$\pi(x + \eta x) - \pi(x) = \frac{\eta x}{\log x} + o\left(\frac{x}{\log x}\right),$$

and so

$$\sum_{p \le x} p \sim \frac{x^2}{2 \log x},$$

where the sum is taken over primes $p \le x$. Therefore

$$|T| \sim \sum_{p \in P} p \sim \frac{k^2}{2 \log k} - \frac{n}{\log n},$$

so $|T| > n$ provided $k > (1 + o(1))\sqrt{n \log n}$. □

§4. Remarks

While Theorem 2 represents a substantial improvement on previous bounds, it is probably a long way from being best possible. Indeed, it seems likely that $f(n) = (1 + o(1))c \log n$, for some constant c. The method we have used above could perhaps be slightly improved by strengthening Lemma 1; however, the lower bound in (3) shows that we cannot hope for an improvement of more than a factor of $\log n$. Any significantly better bound on $f(n)$ would require a new idea.

The approach of §2 should also work for the problem of reconstructing matrices (and for the analogoue higher-dimensional problems). For an $n \times n$ matrix $A = (a_{ij})_{1,j=1}^{n}$, we define the k-deck of A to be the multiset of $\binom{n}{k}^2$ $k \times k$ submatrices of A (for the problem of reconstructing matrices from their principal minors, see Manvel and Stockmeyer [7]). By considering the $a \times b$ submatrices of A, for $1 \le a, b \le k$, we can determine the values of $\sum_{i,j=1}^{n} a_{ij} i^c j^d$, for $1 \le c, d \le k$. A similar argument to that used in §3 should give a bound of the form $O(n^{2/3} \log n)$. However, proving that the set of matrices equivalent to (8) forms a spanning set for $\mathbb{F}^{n^2}$ appears to be more difficult than in the one-dimensional case.

The same approach could be used for reconstructing permutations, which can be seen as a special case of the matrix reconstruction problem. Given a permutation σ of $[n] = \{1, \ldots, n\}$ and a subset $S \in [n]^{(k)}$ (i.e. $|S| = k$), σ rearranges the order of elements in S and thus induces a permutation $\sigma|S$. The k-deck of σ is the multiset $\{\sigma|S : S \in [n]^{(k)}\}$. It is fairly easy to reconstruct the k-deck of the

permutation matrix of σ from the k-deck of σ, and it should therefore be possible to obtain similar upper bounds to those for the matrix reconstruction problem.

Finally, we mention the problem of reconstructing a *cyclic* sequence: we are given the k-deck of a sequence of length n up to cyclic permutation and seek to reconstruct the original sequence up to cyclic permutation. Equivalently, we want to reconstruct a necklace of n coloured beads from the multiset of necklaces obtained by removing $n-k$ of the beads. If every cyclic sequence of length n is reconstructible from its k-deck then clearly every sequence of length n is reconstructible from its k-deck. However, we do not have good bounds for the cyclic reconstruction problem.

Acknowledgment. I would like to thank the anonymous referee for a careful reading of the paper.

References

[1] P.G. Aleksanjan, The reconstruction of vectors by their fragments, *Akad. Nauk. Armyan. SSR Dokl.* **68** (1979), 39-41.

[2] Hardy and Wright, *An Introduction to the Theory of Numbers, 5th ed*, Oxford University Press (1983), *xvi*+426 pp.

[3] L.I. Kalashnik, The reconstruction of a word from fragments, *Numerical Mathematics and computer technology, Akad. Nauk Ukrain. SSR Inst. Mat. Preprint IV* (1973), 56-57.

[4] E. Kubicka and A.J. Schwenk, The ultimate algorithm for counting subsequences, *lecture notes (Western Michigan University),* March 1990.

[5] V.K. Leont'ev and Yu.G. Smetanin, On the recovery of vectors from a collection of their fragments, *Soviet Math. Dokl.* **38** (1989), 438-441.

[6] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith and P. Stockmeyer, Reconstruction of sequences, *Discrete Math.* **94** (1991), 209-219.

[7] B. Manvel and P. Stockmeyer, On reconstruction of matrices, *Math. Mag.* **44** (1971), 218-221.

[8] V. Nýdl, Irreconstructibility of finite undirected graphs from large subgraphs, in Fourth Czechoslovakian Symposium on Combinatorics, Graphs and Complexity, ed. J. Ne**v**setril and M. Fiedler, Elsevier (1992), *pp.* 241-244.

[9] A.J. Schwenk, *private communication*.

[10] Zenkin and Leont'ev, On a non-classical recognition problem, *USSR Comput. Maths Math. Phys.* **24 (3)** (1984), 189-193.