Testing Hereditary Properties of Sequences

Cody R. Freitag[1], Eric Price[2], and William J. Swartworth[3]

1 **Department of Computer Science, UT Austin, Austin, TX, USA**
`cody@rdfriday.com`
2 **Department of Computer Science, UT Austin, Austin, TX, USA**
`ecprice@cs.utexas.edu`
3 **Department of Computer Science, UT Austin, Austin, TX, USA**
`wswartworth@gmail.com`

Abstract

A hereditary property of a sequence is one that is preserved when restricting to subsequences. We show that there exist hereditary properties of sequences that cannot be tested with sublinear queries, resolving an open question posed by Newman et al. [20]. This proof relies crucially on an infinite alphabet, however; for finite alphabets, we observe that any hereditary property can be tested with a constant number of queries.

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity

Keywords and phrases Property Testing

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

1 Introduction

Property testing is the problem of distinguishing objects x that satisfy a given property P from ones that are "far" from satisfying it in some distance measure [13], with constant (say, 2/3) success probability. The most basic questions in property testing are which properties can be tested with constant queries; which properties cannot be tested without reading almost the entire input x; and which properties lie in between.

This paper considers property testing of sequences under the edit distance. We say a length n sequence x is ϵ-far from another (not necessarily length-n) sequence y if the edit distance is at least ϵn. One of the key problems in property testing is testing if a sequence is monotone; a long line of work (see [10, 5, 7, 8] and references therein) showed that $\Theta(\frac{1}{\epsilon}\log n)$ queries are necessary and sufficient.

One can generalize monotonicity by considering properties defined by forbidden order patterns. For instance, avoiding the $(1, 3, 2)$ pattern would mean that x contains no length-3 subsequence with the first smaller than the third element and the third element smaller than the second. Monotonicity would correspond to avoiding the $(2, 1)$ sequence. Pattern free sequences have a long history of study in combinatorics, such as the (now proven) Stanley-Wilf conjecture [19, 12]. In property testing, Newman et al. recently showed (among other results) that every length-k pattern can be tested with $O(n^{1-1/k}/\epsilon^{1/k})$ nonadaptive queries [20], and that $\Omega(n^{1-2/(k+1)})$ queries are necessary for testers that make non-adaptive queries.

Properties defined by forbidden order patterns can be further generalized to hereditary properties of sequences. We say a sequence property P is *hereditary* if, for any sequence x satisfying P, any subsequence of x also satisfies P. Newman et al. [20] pose as an open problem the question we consider in this work: *can any hereditary property of sequences be tested with sublinear query complexity*?

Hereditary properties have long been studied for graphs. It was shown by [2] that hereditary properties of dense graphs are essentially precisely the ones that are testable with a constant number of queries. Similar results have been shown for hypergraphs [3] and certain sparse graphs [9].

Hereditary properties are also testable for permutations, under multiple notions of distance measure [16, 4, 17]. Since hereditary properties on graphs and permutations are testable, might they also be testable on sequences? For sequences the query complexity cannot be independent of n, since (for example) monotonicity testing requires $\Omega(\frac{1}{\epsilon}\log n)$ queries, but one could hope for something sublinear.

Our results. Our main result is to resolve the open question in the negative: there exist hereditary properties of sequences that cannot be tested with sublinear queries. We show how to reduce an arbitrary sequence property to a hereditary property over a larger alphabet. Since there exist sequence properties that require $\Omega(n)$ queries for constant ϵ, the same must hold true for hereditary properties:

▶ Theorem 1. Let $\epsilon \leq 1/40$. There exist hereditary properties of sequences for which no ϵ-tester with two-sided error exists that uses $o(n)$ queries.

Our reduction makes the sequence alphabet grow with n. While large alphabets often makes sense for sequence testing problems—for instance, forbidden order patterns typically expect all n sequence elements to be distinct—one may wonder if hereditary properties over finite alphabets behave differently. They do. We show that every hereditary property of sequences over a finite alphabet can be tested with a constant number of queries:

▶ Theorem 2. Every hereditary property over a finite alphabet is testable with query complexity independent of n.

Related work. A recent concurrent work [1] studies hereditary properties of edge-colored vertex-ordered graphs. They show that any hereditary property, for a fixed finite alphabet of edge colors, is testable with a constant number of queries. This is analogous to our upper bound for finite alphabets, but in the setting of ordered dense graphs rather than sequences.

Our Theorem 1 relies on finding a property that requires $\Omega(n)$ queries. The existence of such a property was shown in [6] for quantum property testers under Hamming distance, building on techniques in [14]. These techniques could be converted into our setting of classical property testers under edit distance. Instead, we choose to give an *explicit* property requiring $\Omega(n)$ queries for our setting, which may be of independent interest.

1.1 Overview of Techniques

This paper consists of three technical pieces: a reduction from arbitrary properties to hereditary properties over a larger alphabet; a lower bound for arbitrary properties; and an upper bound for hereditary properties over finite alphabets. We briefly outline each part in turn.

The reduction. In Section 3, we give a reduction showing that given a blackbox tester for hereditary properties using $q(n, \epsilon)$ queries, we can test arbitrary properties with $q(n, \epsilon/2)$ queries. The key to this transformation is making new, disjoint alphabets for each sequence length for the original property. Then, we can make that property hereditary by adding all subsequences. Because all alphabets are disjoint, the fact that the new property is hereditary doesn't make the property much easier to test.

Explicit hard properties. We construct an explicit property P of integer sequences which requires linear queries to test. Our construction consists of sequences over $\mathbb{F}_p$ where p grows linearly with the length of the sequence. We construct P such that a random sequence in P of length n is indistinguishable (in the information-theoretic sense) from a uniformly random sequence over $\mathbb{F}_p$ to any algorithm making fewer than $n/2$ queries. By making our property small enough, we ensure that almost all sequences over $\mathbb{F}_p$ of length n are ϵ-far from P. Thus we show that a correct tester would be able to distinguish a uniform sample from P from a uniform sample over the total space with good probability. Since this requires $n/2$ queries, we obtain a linear lower bound for testing P.

Finite-alphabet hereditary properties are easy. In Section 4, we show that testing for a hereditary property over a finite alphabet is equivalent to testing for the avoidance of a finite set of forbidden subsequences. If a sequence is ϵ-far from avoiding m subsequences under edit distance, then it must be at least ϵ/m-far from avoiding one such subsequence. This subsequence has some finite length k, which we show means that a uniform sample of $O(\frac{m}{\epsilon}k^2 \log k)$ indices finds this subsequence with constant probability.

2 Notation

A sequence of length n over an alphabet Σ is a function $S\colon [n] \to \Sigma$, often written as $(S_1, \ldots, S_n)$. A property P is a set of sequences, and we say a particular sequence S has property P if S is in P. We say that a sequence S of length n is ϵ-far from P if for all $x \in P$, $d(S, x) > \epsilon n$ for some distance measure d. In this paper we consider edit distance, i.e., $d(x, y)$ is the minimum number of symbol deletions, insertions, or substitutions needed to transform x into y.

A property P is *hereditary* if for all sequences S in P, every subsequence of S is also in P. For every property P, there is a smallest hereditary property containing P, which consists of all subsequences of elements in P. We call this property the *hereditary closure* of P and denote it by P^*.

An ϵ-tester for a property P is a randomized algorithm that on an input sequence S queries a set of indices of S (possibly adaptively) and accepts with probability at least $2/3$ if $S \in P$ and rejects with probability at least $2/3$ if S is ϵ-far from P. Such a tester is said to have *two-sided error*. If the tester is instead required to accept with probability 1 on all inputs in P, we say that the tester has *one-sided error*. We say that a property P is *testable* with $q(n, \epsilon)$ queries if for every $\epsilon > 0$ there is an ϵ-tester for P using at most $q(n, \epsilon)$ queries on sequences of length n with two-sided error.

3 Hereditary Properties over Arbitrary Alphabets

Our goal in this section is to prove Theorem 1:

▶ Theorem 1. Let $\epsilon \leq 1/40$. There exist hereditary properties of sequences for which no ϵ-tester with two-sided error exists that uses $o(n)$ queries.

We first give a reduction from arbitrary property testing on sequences to hereditary property testing. The result then follows from the existence of sequence properties that cannot be tested with sublinear queries.

3.1 Reduction from Testing Arbitrary Properties to Hereditary Properties

▶ **Lemma 3.** *Fix an arbitrary infinite alphabet Σ. If every hereditary property of sequences over Σ is testable with $q(n, \epsilon)$ queries, then every property of sequences over Σ is testable with $q(n, \epsilon/2)$ queries.*

Proof. Let P be an arbitrary property over the alphabet Σ. Since Σ is infinite, there is a countably infinite collection,$\{\Sigma_1, \Sigma_2, \ldots\}$, of disjoint subsets of Σ where each Σ_m has the same cardinality as Σ [1]. For each m, let $f_m : \Sigma \to \Sigma_m$ be a fixed bijection from Σ to Σ_m.

We construct a property Q by converting every sequence in P of length m to the corresponding alphabet Σ_m. More formally, let $Q_m = \{f_m(S) \mid S \in P, S \text{ is of length } m\}$ for each $m \in \mathbb{N}$, and let $Q = \bigcup_{m \in \mathbb{N}} Q_m$.

We claim that if S is in P, then $f_m(S)$ is in the hereditary closure Q^* of Q, and if S is ϵ-far from P, then $f_m(S)$ is $\epsilon/2$-far from Q^*. It will follow from this that an $\epsilon/2$ tester for the hereditary property Q^* suffices to test for P.

Suppose S is length n and has property P. Then $f_n(S) \in Q \subseteq Q^*$, so $f_n(S)$ is in Q^*. Now suppose that S is ϵ-far from P. Trivially $f_n(S)$ is ϵ-far from every subsequence of a sequence in Q_i^* with $i \neq n$ since Σ_i and Σ_n are disjoint. Also, $f_n(S)$ is ϵ-far from every sequence in Q_n since f_n is a bijection between Σ and Σ_n. If $f_n(S)$ were $\epsilon/2$-close to a subsequence x' of some $x \in Q_n$, then x' must have length at least $n - \epsilon n/2$. This means x' is $\epsilon/2$-close to x in edit distance. It then follows that $f_n(S)$ is ϵ-close to $x \in Q_n$, which is a contradiction. Therefore, $f_n(S)$ must be $\epsilon/2$-far from Q^*. ◀

3.2 An Explicit Property Requiring Linear Queries

Related work uses a nonconstructive argument to show that there exists properties of binary sequences which require linear queries to test with two-sided error [6]. Here we construct an explicit class of sequences over $\mathbb{Z}$ which require linear queries. Specifically we show that testing whether a vector in $\mathbb{F}_p^{2n}$ lies in the space of codewords of a Reed-Solomon code requires at least n queries.

For $p \geq k$, let Reed-Solomon$_p(l, k)$ denote the space of codewords for the Reed-Solomon code over $\mathbb{F}_p$ with message length l and codeword length k. Explicitly we define Reed-Solomon$_p(l, k)$ to be the column span of the following matrix taken over $\mathbb{F}_p$:

$$\begin{bmatrix} 1^0 & 1^1 & \ldots & 1^{l-1} \\ 2^0 & 2^1 & \ldots & 2^{l-1} \\ 3^0 & 3^1 & \ldots & 3^{l-1} \\ \vdots & \vdots & \ddots & \vdots \\ k^0 & k^1 & \ldots & k^{l-1} \end{bmatrix}.$$

Our main result is that when k is larger than l by a constant factor, testing for membership in Reed-Solomon$_p(l, k)$ requires linear queries.

▶ **Lemma 4.** *Let P be the space of codewords for Reed-Solomon$_p(n, 2n)$, and set $\epsilon = 1/40$. An adaptive two sided tester (with $2/3$ success probability), which ϵ-tests for P must make at least n queries.*

[1] For arbitrary Σ, this result requires the axiom of choice. However in the case $\Sigma = \mathbb{N}$ we may be explicit by setting $\Sigma_m = \{(m+i)^2 + i | i \in \mathbb{N}\}$.

We require the following well-known property of the Reed-Solomon matrix M.

▶ **Lemma 5.** *Let M be the $2n \times n$ matrix with $M_{i,j} = i^{j-1}$. Each $n \times n$ submatrix of M has full rank.*

Proof. Let $v = [v_0, \dots v_{n-1}]^T$, and let M_i denote the i^{th} row of M. Set

$$q_v(x) = v_0 + v_1 x^1 + \dots + v_{n-1} x^{n-1},$$

and observe that that $M_i v = q_v(i)$. If some n rows of M were dependent then for some nonzero v we would have $M_i v = q_v(i) = 0$ for n different values of i. But this cannot happen since q_v is a nonzero polynomial of degree at most $n-1$. ◀

Our main argument proceeds by showing that a tester for P would be able distinguish a sequence drawn from the uniform distribution on P from a sequence drawn from the uniform distribution on $\mathbb{F}_p^{2n}$ with good probability. We will first argue this fact, and then show that any algorithm which distinguishes these distribution with probability greater than $1/2$ must make at least n queries.

The first step amounts to bounding the size of an ϵ-ball in $\mathbb{F}_p^{2n}$.

▶ **Lemma 6.** *The size of an ϵ-ball in F_p^n under edit distance is at most $(ep/\epsilon)^{2\epsilon n}$.*

Proof. Recall that under our definitions, edit distance allows for insertions, deletions, and replacements. A replacement may be simulated with a deletion, followed by an insertion. Therefore, if $d(\cdot,\cdot)$ is the analogue of edit distance allowing only insertions and deletions as moves, it suffices to bound the size of a 2ϵ-ball under the metric d.

Fix $x \in F_p^n$. Any element in $B_d(2\epsilon, x)$ may be constructed from x by the following procedure. First we select a subset of ϵn indices of x to delete. Then we choose a multiset of indices in $\{0, 1, \dots n - \epsilon n\}$ of size ϵn corresponding to the locations in the resulting sequence where we will perform our insertions. Finally we choose a sequence of length ϵn to insert into those locations.

There are $\binom{n}{\epsilon n}$ ways to choose the ϵn elements to delete. Then there are $\binom{(n-n\epsilon)+n\epsilon}{n\epsilon} = \binom{n}{n\epsilon}$ ways to select the multiset of indices of size ϵn. Finally there are $p^{\epsilon n}$ ways to choose a sequence of length ϵn. It follows that

$$\begin{aligned} |B_d(2\epsilon, x)| &\leq \binom{n}{n\epsilon} \cdot \binom{n}{n\epsilon} \cdot p^{\epsilon n} \\ &\leq \left(\frac{e}{\epsilon}\right)^{2\epsilon n} \cdot p^{\epsilon n} \\ &\leq \left(\frac{ep}{\epsilon}\right)^{2\epsilon n}. \end{aligned}$$

◀

▶ **Lemma 7.** *Set $\epsilon = 1/40$, and let T be an ϵ-tester for P. For $x \sim$ Uniform$(\mathbb{F}_p^{2n})$, T will accept with probability strictly less than $1/2$ (for large enough n).*

Proof. The argument is that a uniformly random vector in $\mathbb{F}_p^{2n}$ is ϵ-far from P (in edit distance) with high probability. We first observe that an ϵ-neighborhood of P is small. In

particular we have

$$\begin{aligned}|\{x \in \mathbb{F}_p^{2n} : x \text{ is } \epsilon\text{-close to } P\}| &\leq |B_\epsilon| \cdot |P| \\ &\leq \left(\frac{ep}{\epsilon}\right)^{4\epsilon n} \cdot p^{2n/2} \\ &\leq (60p)^{n/10} \cdot p^n \\ &\leq p^{7n/10} \cdot p^n \\ &\leq p^{1.7n},\end{aligned}$$

where we used that $p \geq 2$.

The probability that a vector drawn uniformly from $\mathbb{F}_p^{2n}$ is ϵ-close to P is at most $p^{1.7n}/p^{2n}$ which in turn is at most $2^{-0.3n}$. Therefore for $x \sim \text{Uniform}(\mathbb{F}_p^{2n})$, and $n > 6$, we have

$$\Pr[T \text{ rejects on } x] \geq (2/3) \cdot (1 - 2^{-0.3n}) > 1/2,$$

since T must reject, with probability $2/3$, every point which is ϵ-far from P. ◀

The next step is to argue that any tester which makes fewer than n queries, cannot distinguish the distributions $\text{Uniform}(\mathbb{F}_p^{2n})$ and $\text{Uniform}(P)$. In fact we have the following:

▶ **Lemma 8.** *Let x and y be random vectors draw from $\text{Uniform}(\mathbb{F}_p^{2n})$ and $\text{Uniform}(P)$ respectively. For any collection $\mathcal{I} \subseteq [2n]$ of indices with $|\mathcal{I}| \leq n$, the distributions on $x|_\mathcal{I}$ and $y|_\mathcal{I}$ are both uniform over vectors of length $|\mathcal{I}|$*

Proof. It is immediately clear that $x|_\mathcal{I}$ is uniform. That $y|_\mathcal{I}$ is uniform follows from the construction of the matrix A. To be precise, first recall that the restriction of A to any collection n rows is an invertible matrix. It follows that for any $m \leq n$, the restriction of A to any m rows has rank m. The column span of a full-rank $m \times n$ matrix over $\mathbb{F}_p$ is exactly $\mathbb{F}_p^m$. Therefore $y|_\mathcal{I}$ is uniform over vectors of length $|\mathcal{I}|$. ◀

Putting these facts together completes the proof of Theorem 4.

Proof. Let x be a vector in $\mathbb{F}_p^{2n}$ sampled either from $\text{Uniform}(\mathbb{F}_p^{2n})$ or $\text{Uniform}(P)$. Suppose that our tester T makes at most n queries on x, possibly adaptively. By Lemma 8, the value at each index in x after fewer than n queries is uniformly random over $\mathbb{F}$, and independent of the values of all previous queries. Hence for either distribution we may simulate T's behavior by returning uniformly random values for each of its queries. Therefore T must have the same probability of acceptance on both of the two distributions for x. Lemma 7 shows that a correct T must accept on $\text{Uniform}(\mathbb{F}_p^{2n})$ with probability smaller than $1/2$. But by correctness, T must accept on $\text{Uniform}(P)$ with at least $2/3$ probability. It follows that a T which makes fewer than n queries cannot be correct. ◀

4 Hereditary Properties over Finite Alphabets

We now show that the reduction of Section 3.1 relied heavily on the fact the the resulting hereditary property was over an infinite alphabet. In fact, hereditary properties over a finite alphabet can be tested with sublinear query complexity.

▶ Theorem 2. Every hereditary property over a finite alphabet is testable with query complexity independent of n.

We begin with the following standard definition:

▶ **Definition 9.** A partial order $(P, \preceq)$ is said to be a *well partial order* if for every infinite sequence $p_1, p_2, \ldots$ of elements in P, there exists $i < j$ such that $p_i \preceq p_j$.

As mentioned in [18], the following result is well-known. We present a proof here mostly for completeness. A similar proof is presented in [15] but we provide a different exposition which exploits some general structural properties of well partial orders.

▶ **Lemma 10.** *Finite length sequences over a finite alphabet form a well partial order with respect to the subsequence relation.*

The proof of Lemma 10 relies on the following two lemmas.

▶ **Lemma 11.** *Let P be a well partially ordered set, and let $X = x_1, x_2, \ldots$ be a sequence of elements from P. Then there is a subsequence $Y = y_1, y_2, \ldots$ of X, such that $y_i \leq y_j$ for all $i \leq j$.*

Proof. First we argue that there exists an x_i which is (weakly) dominated by infinitely many elements of X. Suppose not. Then for each x_i, let i' be the largest integer satisfying $x_i \leq x_{i'}$. Let S denote the sequence of X corresponding to the set $\{x_{i'} : i \in \mathbb{N}\}$. Since S is necessarily infinite, there exists elements $s_i \leq s_j$ with $i < j$. But this contradicts the maximality of the $x_{i'}$'s.

To construct the sequence Y, we take y_1 to be x_{i_1}, where x_{i_1} is dominated by infinitely many elements in X. Set $S_1 = \{x_k : k > i_1, x_k \geq x_{i_1}\}$. Since S_1 is infinite, we may take y_2 to be x_{i_2} where x_{i_2} is dominated by infinitely many elements of S_1. By iterating this procedure we obtain our sequence Y. ◀

▶ **Lemma 12.** *Let $P_1, \ldots P_n$ be sets which are well partially ordered. Order the set $P_1 \times \ldots \times P_n$ by termwise domination. That is we say that $(p_1, \ldots, p_n) \leq (p'_1, \ldots p'_n)$ if and only if $p_i \leq p'_i$ for all $i \in [n]$. With this order, $P_1 \times \ldots \times P_n$ is a well partial order.*

Proof. By a straightforward induction, it suffices to prove the result when $n = 2$. Consider a sequence $S = \{(a_i, b_i)\}$ with $a_i \in P_1$ and $b_i \in P_2$. By Lemma 11 applied to P_1, there is an infinite subsequence of tuples S' such the first entries in each element of S' are (weakly) increasing. Now since P_2 is a well partial order, there exists elements $s'_i \leq s'_j$ in S' with $i < j$. Since S' is a subsequence of S it follows that S is a well partial order. ◀

Now we present a proof of Lemma 10.

Proof. Let $\mathcal{A}_k = \{a_1, \ldots, a_k\}$ be our finite alphabet of size k. Our proof is by induction on k. When $k = 1$ the result follows from $\mathbb{N}$ being a well partial order.

Now fix an alphabet of size $k+1$. Consider an infinite sequence $X = x_1, x_2, \ldots$ consisting of finite strings over the alphabet $\mathcal{A}_{k+1}$. Given a finite string $S = s_1, \ldots s_n$ over the alphabet $\mathcal{A}_{k+1}$ we represent it as a tuple $(u_1, \ldots, u_m)$ satisfying the following considerations:

- u_i is a finite sequence over the alphabet $\mathcal{A}_{k+1} - \{a_{i \mod (k+1)}\}$
- S is the concatenation of the strings $u_1, \ldots u_n$.
- each u_i is as long as possible, i.e. the first character of u_{i+1} is $a_{i \mod (k+1)}$.

Using the final property listed above, we observe that if this tuple has size at least $r(k+1)+1$, then S contains the subsequence $(a_1, a_2, \ldots, a_{k+1})^r$, where the exponent means that we repeat the string inside the parentheses r times.

Now represent each element of the sequence X as a tuple in this way. If x_1 is contained as a subsequence in some x_i with $i > 1$ then we are finished. Otherwise, let x_1 have length

l. Then x_1 is contained as a substring in $(a_1 a_2 \dots a_{k+1})^l$. The tuple associated to each x_i with $i > 1$ must have length at most $l(k+1)+1$. Otherwise, by our previous observation, x_i would contain $(a_1 a_2 \dots a_{k+1})^l$ as a substring, and hence also x_1. We may represent each x_i with a tuple of length exactly $l(k+1)+1$ by padding x_i's tuple with empty strings as necessary. By induction, the elements of these tuples are well partially ordered. But then Lemma 12 implies that the tuples of length $l(k+1)+1$ also form a well partial order. Since the ordering on strings respects the ordering on tuples, it follows that there exists $i < j$ with $x_i \leq x_j$. Therefore X is well partially ordered.

◀

We are now ready to prove the following key fact.

▶ **Lemma 13.** *Let P be a hereditary property of sequences over a finite alphabet Σ. Then there exists a finite set $\mathcal{S}$ of sequences over Σ such that P consists exactly of the sequences which do not contain any sequence in $\mathcal{S}$ as a subsequence.*

Proof. First observe that since P is hereditary, P consists of all sequences which do not contain any sequence in $\overline{P}$, the complement of P, as a subsequence. Since $\overline{P}$ is countable, we may enumerate it as $\overline{P} = \{q_1, q_2, \dots\}$. We construct $\mathcal{S}$ inductively, by setting $s_1 = q_1$, and setting $s_{i+1} = q_j$ where j is the minimum value such that q_j does not contain any of the sequences $s_1, \dots s_i$ as a subsequence. Lemma 10 implies that this process must halt at some point by the definition of a well partial order, so $\mathcal{S}$ will be finite. From the construction, it is clear that each sequence in $\overline{P}$ contains a sequence in $\mathcal{S}$ as a subsequence. Therefore, P is exactly the set of sequences that avoid sequences in $\mathcal{S}$ as a subsequence. ◀

With these results, we give a short proof of Theorem 2.

Proof. By Lemma 13 it suffices to construct a tester that tests whether an input x avoids a finite collection of forbidden subsequences. In fact it is enough to construct a tester for each such sequence individually. This is because if x is ϵ-far from avoiding a collection of m sequences, then x must be ϵ/m-far from avoiding one of these subsequences. This relies on the fact that we are using edit distance, so to avoid a particular subsequence, we can just delete a subset of indices that contain that subsequence.

Suppose x were ϵ/m-close to avoiding m subsequences, $y_1, \dots, y_m$, individually. Let S_i be the smallest set of indices such that deleting S_i from x causes x to avoid y_i. Note that by assumption of x being ϵ/m-close to avoiding y_i, $|S_i| \leq \epsilon n/m$. Then deleting $\cup_{i=1}^m S_i$ from x will cause x to avoid all m subsequences, but $|\cup_{i=1}^m S_i| \leq m \cdot (\epsilon n/m) = \epsilon n$. This contradicts that x is ϵ-far from avoiding all of $y_1, \dots, y_m$. Therefore constructing an ϵ/m-tester for avoiding a particular sequence suffices.

Let u be a forbidden subsequence of size k. If x is ϵ-far from avoiding u, x must have at least $\epsilon n/k$ disjoint copies of u as subsequences. It was noted in [20] that a uniform sample of $O(\epsilon^{1/k} n^{1-1/k})$ entries contains one of these subsequences with constant probability by a second moment bound. However, we show in Lemma 14 that over a finite alphabet, this can be improved to just a uniform sample of $O(\frac{1}{\epsilon} k^2 \log k)$ entries.

Then to test whether x has a hereditary property over a finite alphabet, we compute the m forbidden subsequences, each of length say at most k. Then after sampling $O(\frac{m}{\epsilon} k^2 \log k)$ random indices, if x is ϵ-far from avoiding all forbidden subsequences, we will find the subsequence that x is ϵ/m-far from avoiding with at least $2/3$ probability. ◀

▶ **Lemma 14.** *There exists an ϵ-tester with one-sided error for avoiding a fixed subsequence s of length k using $O(\frac{1}{\epsilon} k^2 \log k)$ queries.*

Proof. We first assume that k is a power of 2 and then reduce to the case of general k. We also use the fact that if a sequence x is ϵ-far from avoiding s as a subsequence, then there must be a set T consisting of $\epsilon n/k$ disjoint copies of s in x [20].

Let i be minimal such that the restriction of x to T contains at least $|T|/2 = \epsilon n/2k$ disjoint instances of the subsequence $s_1, \dots s_{k/2}$ strictly to the left of i. By minimality of i it follows that $x_i, x_{i+1}, \dots, x_n$ contains at least $\epsilon n/2k - 1$ disjoint copies of $s_{k/2+1}, \dots, s_k$. By iterating this procedure, we divide x into k blocks $X_1, \dots X_k$ such that each X_i contains at least $\epsilon n/k^2 - \log k$ copies of s_i, which is $\Omega(\epsilon n/k^2)$ as long as $k = o(n^{1/2})$.

Our algorithm is to sample a uniform subset of x of size u. The probability any individual sample will be an instance of s_i from the block X_i is at least $\Omega(\epsilon/k^2)$. Thus with constant probability, we will select a corresponding s_i from each of the blocks X_i after $O(\frac{1}{\epsilon}k^2 \log k)$ samples.

We now reduce the case where the length of the subsequence is a power of 2 to general k. Let s be of length k, and k' be the smallest power of 2 larger than k. Let c be any character not in the alphabet of the sequence. We will construct s' of length k' by adding $k' - k$ copies of c to the end of s. We also construct the sequence x' by adding $(k' - k) \cdot \epsilon n/k$ copies of c to the end of x.

Note that x' avoids s' if and only if x avoids s since c is disjoint from the original alphabet. Also $k' - k < k$, so the length of x' is at most $2n$. This means x is ϵ-far from avoiding s if and only if x' is at least $\epsilon/2$-far from avoiding s'. Also, we can simulate any property testing algorithm on x' since any query for an index greater than n must return c. Therefore we can test x for s-avoidance by testing x' for s'-avoidance using $O(\frac{1}{\epsilon/2}(k')^2 \log k') = O(\frac{1}{\epsilon}k^2 \log k)$ queries. ◀

5 Conclusions and Open Problems

We showed that there exist hereditary properties that require linear query complexity. However, we also show that when we restrict to hereditary properties over a finite alphabet, there are testers using queries independent of n. What can we say about other natural restrictions on hereditary properties? Sequences over an infinite alphabet don't form a well-partial order under the subsequence relation, as shown in [21], so we need different techniques to see if other interesting restrictions over infinite alphabets can be tested using sublinear queries.

One natural restriction is to order-based hereditary properties [11]. [20] considers testing the avoidance of permutation patterns, which is a subclass of order-based hereditary properties. A sequence S avoids a pattern π of length k if there is no set of indices $i_1 < i_2 < \ldots < i_k$ such that $S_{i_x} > S_{i_y}$ if and only if $\pi_x > \pi_y$. It is unknown whether testing the avoidance of constant length patterns requires more than polylog(n) queries with adaptive algorithms.

References

1 Noga Alon, Omri Ben-Eliezer, and Eldar Fischer. Testing hereditary properties of ordered graphs and matrices. *arXiv preprint arXiv:1704.02367*, 2017.

2 Noga Alon and Asaf Shapira. A characterization of the (natural) graph properties testable with one-sided error. *SIAM Journal on Computing*, 37(6):1703–1727, 2008.

3 Tim Austin and Terence Tao. Testability and repair of hereditary hypergraph properties. *Random Structures & Algorithms*, 36(4):373–463, 2010.

4 Antônio JO Bastos, Carlos Hoppen, Yoshiharu Kohayakawa, and Rudini M Sampaio. Every hereditary permutation property is testable. *Electronic Notes in Discrete Mathematics*, 38:123–128, 2011.

5 Arnab Bhattacharyya, Elena Grigorescu, Kyomin Jung, Sofya Raskhodnikova, and David P Woodruff. Transitive-closure spanners. *SIAM Journal on Computing*, 41(6):1380–1425, 2012.

6 Harry Buhrman, Lance Fortnow, Ilan Newman, and Hein Röhrig. Quantum property testing. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 480–488. Society for Industrial and Applied Mathematics, 2003.

7 Deeparnab Chakrabarty and C Seshadhri. Optimal bounds for monotonicity and lipschitz testing over hypercubes and hypergrids. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 419–428. ACM, 2013.

8 Deeparnab Chakrabarty and C Seshadhri. An optimal lower bound for monotonicity testing over hypergrids. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 425–435. Springer, 2013.

9 Artur Czumaj, Asaf Shapira, and Christian Sohler. Testing hereditary properties of nonexpanding bounded-degree graphs. *SIAM Journal on Computing*, 38(6):2499–2510, 2009.

10 Yevgeniy Dodis, Oded Goldreich, Eric Lehman, Sofya Raskhodnikova, Dana Ron, and Alex Samorodnitsky. Improved testing algorithms for monotonicity. In *Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques*, pages 97–108. Springer, 1999.

11 Eldar Fischer. On the strength of comparisons in property testing. *Information and Computation*, 189(1):107–116, 2004.

12 Jacob Fox. Stanley-wilf limits are typically exponential. *arXiv preprint arXiv:1310.8378*, 2013.

13 Oded Goldreich. Combinatorial property testing (a survey). *Randomization Methods in Algorithm Design*, 43:45–59, 1999.

14 Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

15 Leonard H Haines. On free monoids partially ordered by embedding. *Journal of Combinatorial Theory*, 6(1):94–98, 1969.

16 Carlos Hoppen, Yoshiharu Kohayakawa, Carlos Gustavo Moreira, and Rudini Menezes Sampaio. Testing permutation properties through subpermutations. *Theoretical Computer Science*, 412(29):3555–3567, 2011.

17 Tereza Klimošová and Daniel Král. Hereditary properties of permutations are strongly testable. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1164–1173. Society for Industrial and Applied Mathematics, 2014.

18 Joseph B Kruskal. The theory of well-quasi-ordering: A frequently discovered concept. *Journal of Combinatorial Theory, Series A*, 13(3):297–305, 1972.

19 Adam Marcus and Gábor Tardos. Excluded permutation matrices and the stanley–wilf conjecture. *Journal of Combinatorial Theory, Series A*, 107(1):153–160, 2004.

20 Ilan Newman, Yuri Rabinovich, Deepak Rajendraprasad, and Christian Sohler. Testing for forbidden order patterns in an array. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1582–1597. SIAM, 2017.

21 Daniel A Spielman and Miklós Bóna. An infinite antichain of permutations. *Electron. J. Combin*, 7:N2, 2000.