The separating words, k-deck, and trace reconstruction problems

Zachary Chase

University of Oxford

July 11th, 2022

3 Difficult Questions

Let $x, y \in \{0,1\}^n$ be a pair of distinct $0-1$ strings of length n.

Question 1

Is there a DFA with $O(\log n)$ states that accepts x but not y?

Question 2

Is there a string of length $O(n^{1/3})$ that appears a different number of times as a subsequence in x compared to y?

Question 3

If someone secretly chooses x or y and gives us $O(n^2)$ random subsequences of length $n/2$, can we say with high probability whether the secret string is x, or y?

Outline of the Talk

We'll go through each of these questions, explaining

1. what is known about them
2. what is not known about them, and
3. how they appear to relate to one another.

Partial spoiler: there's *much* more work to be done; all three problems currently have an exponential gap between the best known bounds!

The Separating Words Problem

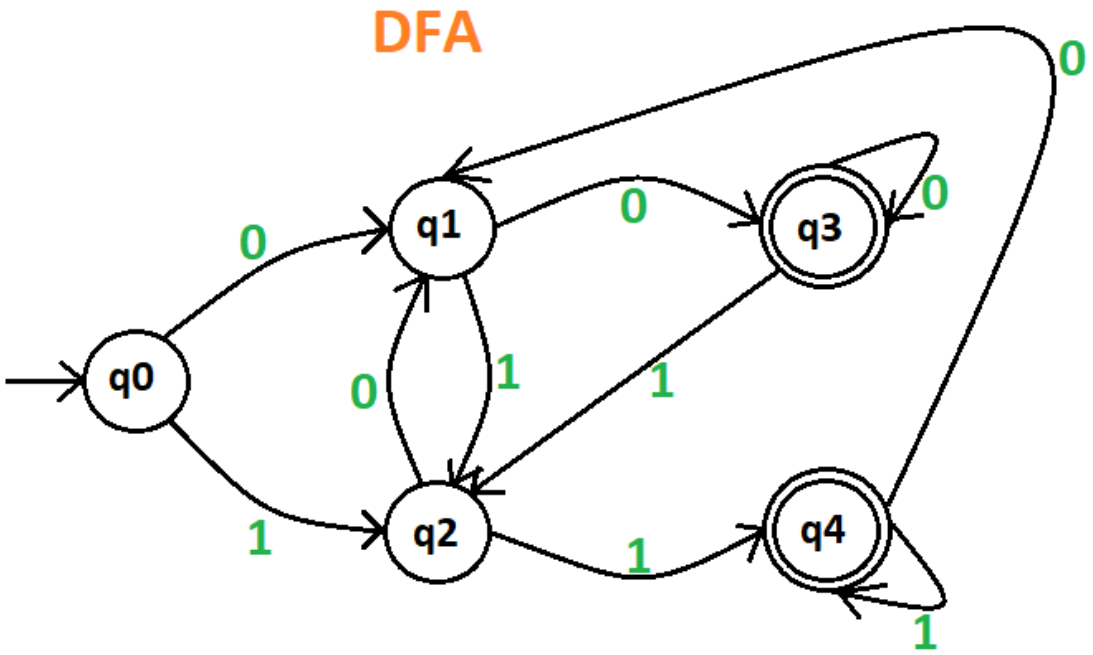
"Imagine a computing device with very limited powers. What is the simplest computational problem you could ask it to solve? It is not the addition of two numbers, nor sorting, nor string matching – it is telling two inputs apart." - Jeffrey Shallit

The Separating Words Problem

For distinct $x, y \in \{0,1\}^n$, what is the smallest size of a DFA that accepts x but not y?

The answer is just a function of n. In other words, we care about the "worst-case" pair $x, y \in \{0,1\}^n$.

The Separating Words Problem



This DFA accepts, for example, the strings 00, 111, and 1010011.

This DFA rejects, for example, the strings 0, 10, and 11001.

Say a DFA *separates* x, y if it accepts one but not the other.

The above DFA separates 00 and 10.

The Separating Words Problem

The Separating Words Problem

For distinct $x, y \in \{0,1\}^n$, what is the smallest size of a DFA that accepts x but not y?

The problem was introduced in 1986 by Goralcik and Koubek.

One could ask a similar question for strings with possibly different length, but this turns out to be too easy.

A trivial upper bound is $n+1$: simply put $n+1$ states "in a row" and have the transitions be such that when x is the input, we "move forward" every time.

The Separating Words Problem

Trivial upper bound is $n + 1$.

Theorem (Goralcik and Koubek, 1986)

Any distinct $x, y \in \{0,1\}^n$ can be separated by a DFA with $o(n)$ states.

They also showed a lower bound.

Theorem (Goralcik and Koubek, 1986)

There exist $x, y \in \{0,1\}^n$ for which no DFA with fewer than $\Omega(\log n)$ states can separate.

The example giving the $\log n$ lower bound is, for $k := \log n$,

$$0^k 1^{k+\mathrm{lcm}(1,2,\ldots,k)}$$

$$0^{k+\mathrm{lcm}(1,2,\ldots,k)} 1^k.$$

.

The Separating Words Problem

In 1989, Robson improved upon the upper bound.

Theorem (Robson, 1989)

Any distinct $x, y \in \{0,1\}^n$ can be separated by a DFA with $\widetilde{O}(n^{2/5})$ states.

We remark that he begins his paper with a shorter proof of an $\widetilde{O}(n^{1/2})$ upper bound.

We recently got a further improvement to the upper bound.

Theorem (C., 2020)

Any distinct $x, y \in \{0,1\}^n$ can be separated by a DFA with $\widetilde{O}(n^{1/3})$ states.

Remember, the lower bound is still $\Omega(\log n)$!

The k-deck Problem

The k-deck Problem

For distinct $x, y \in \{0,1\}^n$, what is the smallest length of a string w that appears a different number of times as a subsequence of x and y?

Once again, the answer is just a function of n – we care about the "worst-case" pair $x, y \in \{0,1\}^n$.

Example: For $w = 01$, $x = 0110$, and $y = 1001$, note that w appears twice as a subsequence in each of x and y:

$$x = 0110 = 0110$$

$$y = 1001 = 1001.$$

However, $w = 011$ appears 1 time in x and 0 times in y:

$$x = 0111 \qquad y = 1001.$$

The k-deck Problem

The trivial upper bound to the problem is, of course, n.

The example on the previous slide is actually an instance of the Thue-Morse word and its complement.

$$01101001100101101001011001101001$$

$$10010110011010010110100110010110$$

These pairs, for n a power of 2, give a lower bound of $\Omega(\log n)$ for the k-deck problem.

There were a couple of proofs of a $\lfloor \frac{n}{2} \rfloor$ upper bound and some multiplicative constant improvements of the $\Omega(\log n)$ lower bound.

The k-deck Problem

So the known bounds were basically $c \log n$ and Cn for certain constants $c, C > 0$.

Let's for ease introduce the notation $f(w; x)$ for the number of times that a string w appears as a subsequence in a string x.

Theorem (Scott, 1997) (Krasikov and Roddity, 1997)

For any distinct $x, y \in \{0,1\}^n$, there exists $|w| = \widetilde{O}(\sqrt{n})$ such that $f(w; x) \neq f(w; y)$.

There was also a major improvement to the lower bound.

Theorem (Dudik and Schulman, 2003)

For arbitrarily large n there exist distinct $x, y \in \{0,1\}^n$ such that for all $|w| \leq \exp\left(\Omega(\sqrt{\log n})\right)$, we have $f(w; x) = f(w; y)$.

The k-deck Problem

Theorem (Dudik and Schulman, 2003)

For arbitrarily large n there exist distinct $x, y \in \{0,1\}^n$ such that for all $|w| \leq \exp\left(\Omega(\sqrt{\log n})\right)$, we have $f(w; x) = f(w; y)$.

We briefly remark on their (random) construction.

First, note that the Thue-Morse sequence and its complement are generated by "repeated substitutions".

If $x^{(k)}, y^{(k)}$ are the Thue-Morse sequence and its complement of length 2^k, then

$$x^{(k+1)} = x^{(k)} \circ y^{(k)}$$

$$y^{(k+1)} = y^{(k)} \circ x^{(k)}.$$

The k-deck Problem

If $x^{(k)}, y^{(k)}$ are the Thue-Morse sequence and its complement of length 2^k, then

$$x^{(k+1)} = x^{(k)} \circ y^{(k)}$$

$$y^{(k+1)} = y^{(k)} \circ x^{(k)}.$$

In other words, to obtain $x^{(k+1)}$ and $y^{(k+1)}$ from $x^{(k)}$ and $y^{(k)}$, we substitute into the pattern

$$ab$$

$$ba$$

with $a = x^{(k)}, b = y^{(k)}$.

The Dudik-Schulman construction is by repeatedly substituting the previous pair into a cleverly chosen random pattern that can (and does) change as the construction goes on.

The k-deck Problem

The k-deck Problem

For distinct $x, y \in \{0,1\}^n$, what is the smallest length of a string w that appears a different number of times as a subsequence of x and y?

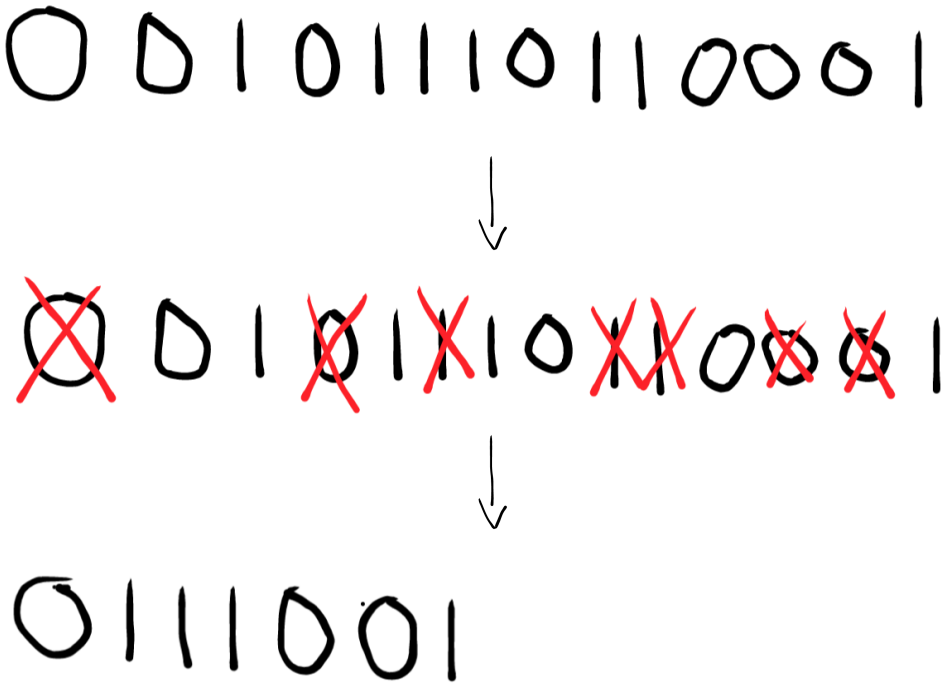
Those bounds are still the best to date: an upper bound of $\widetilde{O}(\sqrt{n})$ and a lower bound of $\exp\left(\Omega(\sqrt{\log n})\right)$.

The Trace Reconstruction Problem

Take a string $x \in \{0,1\}^n$.

A *trace* of x is obtained by deleting each bit with probability $1/2$ and concatenating the result.

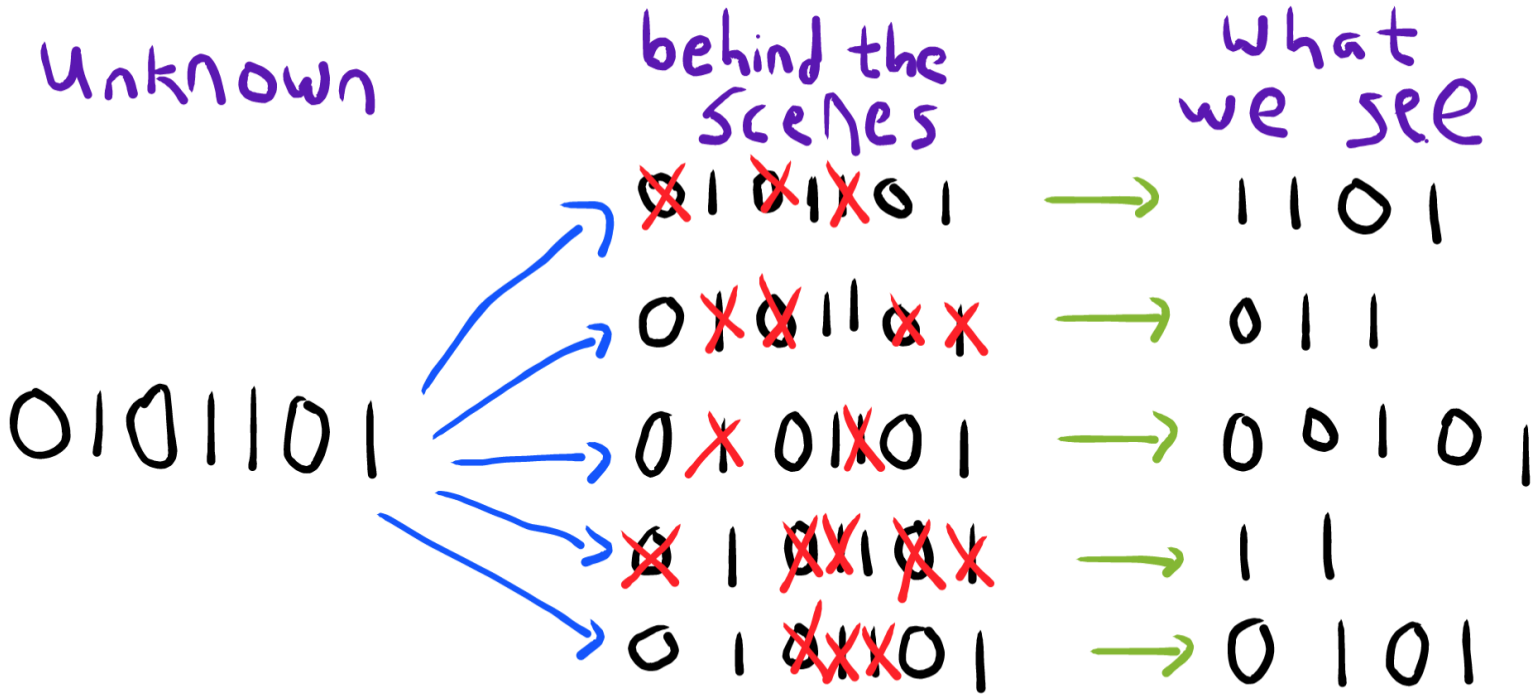
Example of x and a (random) trace:



The Trace Reconstruction Problem

There's an unknown string $x \in \{0,1\}^n$.

How many independently drawn traces of x do we need so that we can determine x with probability at least 0.9?



We want the answer for the worst-case x.

The Trace Reconstruction Problem

The optimal algorithm is quite obvious – choose the $x \in \{0,1\}^n$ that most likely generated the observed traces.

However, we're not (yet?) able to analyze this algorithm. So we instead come up with worse algorithms.

1st Batch of Bounds

Upper bound: $\exp(O(n))$
Lower bound: $\Omega(n)$

Proof: For upper bound, just wait until see the whole string. For lower bound, consider the pair

$$00\ldots001000\ldots00$$
$$00\ldots000100\ldots00$$

□

The main question is whether polynomially many traces suffice.

The Trace Reconstruction Problem

The trace reconstruction problem is essentially equivalent to its *pairwise version*.

If the unknown string is (correctly) promised to be either a certain x or y in $\{0,1\}^n$, then how many traces are needed to determine the truth (whp)?

As usual, we want the answer – a function of n – for the worst-case pair x, y.

The Trace Reconstruction Problem

The first bounds were $\Omega(n)$ and $\exp(O(n))$.

In 2008, Holenstein, Mitzenmacher, Panigrahy, and Wieder established an upper bound of $\exp\left(\widetilde{O}(n^{1/2})\right)$.

In 2016, De, O'donnell, Servedio, and Nazarov, Peres independently and concurrently established an upper bound of $\exp\left(O(n^{1/3})\right)$.

In 2018, Holden and Lyons improved the lower bound to $\widetilde{\Omega}(n^{5/4})$.

In 2019, C. improved the lower bound to $\widetilde{\Omega}(n^{3/2})$.

In 2020, C. improved the upper bound to $\exp\left(\widetilde{O}(n^{1/5})\right)$.

These are the best bounds to date. Much room to improve!

Connections

The 3 problems we just discussed seem to be more related than one might first expect.

The proofs of the

1. $\widetilde{O}(n^{1/2})$ bound for the separating words problem,
2. the $\widetilde{O}(n^{1/2})$ bound for the k-deck problem, and
3. the $\exp\left(\widetilde{O}(n^{1/3})\right)$ bound for the trace reconstruction problem

are all essentially the same!

And the proofs of the $\widetilde{O}(n^{1/3})$, $\exp\left(\widetilde{O}(n^{1/5})\right)$ bounds for the SW and TR problems, respectively, are also essentially the same.

More opaquely, I personally find it interesting that hard pairs for the SW and TR problems need to have common padding at the beginning, i.e., the first index at which they disagree must be "large".

A Number Theory Problem

For a subset $A \subseteq [n] := \{1, \ldots, n\}$, a prime p, and an integer i, define

$$A_{i,p} := \{a \in A : a \equiv i \pmod{p}\}.$$

Problem

Let $A, B \subseteq [n]$ be distinct. How small of a p can we find so that there is some i with

$$|A_{i,p}| \neq |B_{i,p}|?$$

Taking the negation, we're asking for the largest p so that there are two distinct subsets of $[n]$ with the same number of evens, the same number of odds, the same number of things 0 mod 3, the same number of things 1 mod 3, the same number of things 2 mod 3, ..., the same number of things $p-1$ mod p.

A Number Theory Problem

The answer is $\widetilde{O}(n^{1/2})$.

Theorem

For any distinct $A, B \subseteq [n]$, there is some $p \leq O(\sqrt{n \log n})$ and some i so that

$$|A_{i,p}| \neq |B_{i,p}|.$$

Proof: Suppose false. Then

$$\sum_{k=1}^{n} (1_A(k) - 1_B(k)) z^k$$

is a polynomial of degree at most n with roots at

$$z = e^{2\pi i \frac{m}{p}}$$

for all $m \in \{0, 1, \ldots, p-1\}$ for all $p \leq O(\sqrt{n \log n})$. So the polynomial must be identically 0, contradicting that A, B are distinct.

A Number Theory Problem

There's also a more obnoxious proof of the $\widetilde{O}(n^{1/2})$ bound that will actually be useful for later.

Theorem

For any distinct $A, B \subseteq [n]$, there is some non-negative integer $m = \widetilde{O}(n^{1/2})$ such that

$$\sum_{a \in A} a^m \neq \sum_{b \in B} b^m.$$

It's actually trivial to see that this implies the $\widetilde{O}(n^{1/2})$ bound for the number theory problem, since

$$\sum_{a \in A} a^m \equiv \sum_{i=0}^{p-1} |A_{i,p}| i^m \pmod{p}.$$

A 'Moments' Problem

Theorem

For any distinct $A, B \subseteq [n]$, there is some non-negative integer $m = \widetilde{O}(n^{1/2})$ such that

$$\sum_{a \in A} a^m \neq \sum_{b \in B} b^m.$$

This problem is susceptible to complex analytic techniques.

Indeed, it is relatively easy to see that it is equivalent to the following.

Theorem

Let $p(x)$ be a polynomial of degree n with coefficients in $\{-1, 0, 1\}$. Then, $(x-1)^{C\sqrt{n \log n}}$ does not divide $p(x)$.

If $(x-1)^{C\sqrt{n \log n}}$ did divide $p(x)$, then $p(x)$ will be extremely small uniformly near $x = 1$, which can be ruled out with complex analytic techniques.

Connections

The punchline: these three problems – the number theoretic one, the 'moments' one, and the complex analytic one – all arise from (and essentially are equivalent to) analyzing a type of "single-bit statistics" for the given pair x, y.

For the separating words problem, we want a small prime p and some integer i so that the number of 1s in x that are at indices congruent to i mod p differs from that of y.

For the k-deck problem, we look at the counts of the number of subsequences of x of length k that have a 1 in position j (for $j \leq k$).

For the trace reconstruction problem, we look at the probability that a trace of x will have a 1 in position j (for $j \leq n$).

Connections

For concreteness, running through the trace reconstruction argument with very large deletion parameter ($\approx 1 - n^{-1/2}$) recovers the $\widetilde{O}(n^{1/2})$ bound on the k-deck problem.

So given the above methods, it is natural to try looking at two indices instead of one, say.

For SW, we can look at the number of pairs of 1s in x, the first of which is at an index i_1 mod p and the second at an index i_2 mod p.

For k-deck, we can look at the number of subsequences of x of length k with 1s at positions j_1 and j_2.

For TR, we can look at the probability that a trace of x has 1s at positions j_1 and j_2.

Connections

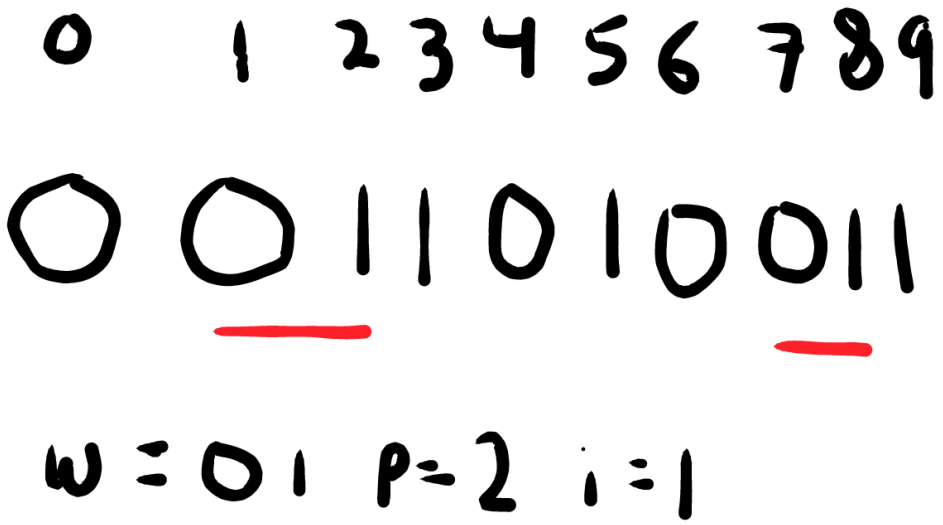
The just-mentioned idea of looking at two indices instead of one probably does give improvements in reality, but we currently don't know how to prove that it does.

However, that idea was looking at two indices that are potentially far away from one another. What if we look at contiguous indices?

It's easy to see that just looking at two contiguous indices won't be enough to get an improvement, so we have to look at many.

Contiguous Substring Appearances

So, for example, for SW, we'd be wanting to count the number of occurrences of a given string w, of length around $n^{1/3}$, as a contiguous substring in x and beginning at an index that is congruent to i mod p.



It's still relatively "cheap" to build a DFA with few states that counts such occurrences.

Contiguous Substring Appearances

If we take the string w to be *aperiodic*, then the occurrences of w in a given string x are well-separated from one another.

This translates to an extra assumption we are given, and can take advantage of, in the number-theoretic problem.

We say that $A \subseteq [n]$ is *d-separated* if $|a - a'| < d \implies a = a'$.

Theorem (C., 2020)

Let $A, B \subseteq [n]$ be distinct and each $n^{1/3}$-separated. Then there is some prime $p = \widetilde{O}(n^{1/3})$ and some i with

$$|A_{i,p}| \neq |B_{i,p}|.$$

Recall that without the $n^{1/3}$-separated assumption, the best we could do is $p = \widetilde{O}(n^{1/2})$.

A New Number-Theoretic Problem

Theorem (C., 2020)

Let $A, B \subseteq [n]$ be distinct and each $n^{1/3}$-separated. Then there is some prime $p = \widetilde{O}(n^{1/3})$ and some i with

$$|A_{i,p}| \neq |B_{i,p}|.$$

Unlike the original number theory problem, this one has no "proof from the book" (e.g. no strict cutoff).

So we take advantage of that more "obnoxious" proof of the previous number theory problem, by going to the "moments" problem, and then to the "polynomial near 1" problem.

The complex analytic techniques could be adapted to give an improvement under this "separated" assumption.

Connections

Unfortunately, it seems no (contiguous) substring methods can help with the k-deck problem.

Indeed, in that problem, padding is not needed and thus distinct strings with the same, say, $\log^2 n$-deck can have a different number of occurrences of 01, say, as a contiguous substring.

However, surprisingly (to me), substring methods can help with trace reconstruction.

Indeed, there are complicated formulae that recover good estimates for substring counts from few traces.

Summary

The Separating Words Problem

For any distinct $x, y \in \{0,1\}^n$, there is a DFA on $\cdot$ states that accepts x but not y.

$$\cdot = \widetilde{O}(n^{1/3}) \qquad \cdot = \Omega(\log n).$$

The k-deck Problem

For any distinct $x, y \in \{0,1\}^n$, there is a string $w \in \{0,1\}^{\cdot}$ that appears a different number of times as a subsequence in x and y.

$$\cdot = \widetilde{O}(n^{1/2}) \qquad \cdot = \exp\left(\Omega(\sqrt{\log n})\right).$$

The Trace Reconstruction Problem

For any distinct $x, y \in \{0,1\}^n$, $\cdot$ random subsequences of length $n/2$ suffice with high probability to distinguish between x and y.

$$\cdot = \exp\left(\widetilde{O}(n^{1/5})\right) \qquad \cdot = \Omega(n^{3/2}).$$

The End

Thanks for listening!