Trace Reconstruction Problem

by

Aldo Pacchiano

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Engineering, MEng

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Signature redacted

Author .. ..
Department of Electrical Engineering and Computer Science
May 23, 2014

Signature redacted

Certified by.... .....
Constantinos Daskalakis
Associate Professor
Thesis Supervisor

Signature redacted

Accepted by..... .....
Albert R. Meyer
Chairman, Department Committee on Graduate Theses

Trace Reconstruction Problem

by

Aldo Pacchiano

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2014, in partial fulfillment of the
requirements for the degree of
Masters of Engineering, MEng

Abstract

In the setting of the trace reconstruction problem, a uniform random binary sequence $w \in \{0,1\}^n$ yields a collection of traces, such that each subsequence is obtained by independently deleting each bit with a public probability parameter p. In this thesis we explore a restricted version of this problem, in which each trace is a random subsequence of one of two original known sequences. Given a series of traces, we would like to device a method that allows to us to decide from which sequence, from the pair of known public sequences w, w', do all the traces come from. The question we will try to solve in this thesis is to know if such a method, operating with high probability and polynomially many samples, is possible in practice. Among other things, we show that if the two strings are drawn uniformly at random there is an algorithm that allows to efficiently distinguish with high probability the traces they produce, failing only on an exponentially small proportion of the random pairs. Additionally we explore variants of this problem and their connections with a number theoretic known as the Prouhet-Tarry-Escott problem.

Thesis Supervisor: Constantinos Daskalakis
Title: Associate Professor

Acknowledgments

I would like to extend a special note of gratitude to all those that accompanied me through this journey. A special thanks to my thesis adviser, Prof. Constantinos Daskalakis, whose valuable insight, time and guidance proved invaluable for the advancement of this project. I would not have been able to finish this nor any other stage of my education without the help I have received from my parents, whose great effort, dedication, and guidance has been the reason I have been able to finish this and any other project I have worked on so far. I dedicate this work to them.

Contents

1 Introduction **9**

1.1 Summary of results and thesis structure 14

2 Preeliminaries **17**

2.1 Subsequences and their properties 17

2.1.1 Sequence representation 17

2.1.2 Algorithms 18

2.1.3 Counting common subsequences 21

2.2 Variation distance 26

2.2.1 General properties 27

2.2.2 Hoeffding Inequality 28

2.2.3 Sampling from polynomially distinguishable families 29

2.2.4 Subsequences variation distance 30

3 Distinguishing tests and special string pairs **33**

3.1 Distinguishing tests 34

3.1.1 The number of ones test 34

3.1.2 The first one test 35

3.1.3 The marginals test 37

3.1.4 The failure of the marginals test 37

3.1.5 The subsequences of subsequences test 42

3.1.6 A subsequences of subsequences test 43

3.2 Special sequence pairs 45

3.2.1 Unbalanced sequences 46
3.2.2 Differing tails 46
3.2.3 Edit distance one are far 47
3.2.4 Alternating sequences 48
3.2.5 Cyclic shifted strings 50

4 Random strings are distinguishable **53**
4.0.6 Future directions 67
4.0.7 The marginals test 68

5 Extensions **71**
5.1 A new distance 71
5.2 Reconstruction of sequences 78
5.3 Searching for the minimising pair 87

6 Conclusion **91**

A Notation reference **93**

Chapter 1

Introduction

The large amount of research on the Graph Reconstruction Conjecture of S. Ulam and P. Kelley has led to an interest in reconstruction problems for a variety of other combinatorial structures (such as for example, digraphs and posets). In this thesis we focus on studying the reconstruction problem for sequences. In particular we present an overview of current advances on a simplified version of the Trace Reconstruction Problem.

The Trace reconstruction problem asks to reconstruct an unknown random sequence T over a fixed alphabet $\mathcal{X}$ by processing a series of subsequences of T, $S_1, \cdots, S_r$ drawn from a distribution over the subsequences of T.

We consider the unknown string to be drawn from the uniform distribution over sequences of $\mathcal{X}^n$. Let $p_n \in [0,1] \forall n$. A subsequence S of T (we denote $S \trianglelefteq T$ to say S is a subsequence of T) is drawn from a distribution $P_n(\cdot, p_n)$ where $P_n(S, p_n) = p_n^{n-|S|}(1-p_n)^{|S|}$.

In other words, for each symbol of T, p_n is the probability that symbol of T is deleted and the subsequence S is what remains after the deletions happen.

The authors of [2] study the trace reconstruction problem when the family $\{p_n\}$ is such that $p_n = \delta$ for some δ constant. Among other things they propose a polynomial time algorithm that is capable of reconstructing all but an exponentially small fraction of the strings in $\{0,1\}^n$ and after polynomial post processing time, whenever the deletion probability δ is less than some small universal constant γ. Unfortunately,

their reconstruction algorithm is quite complicated, and $\gamma << \frac{1}{2}$ therefore leaving some room for finding a simpler and more intuitive version, and one which could work for constant deletion probability δ for $\delta \geq \frac{1}{2}$. We will not survey the internal workings of this algorithm in this Thesis, although we will touch upon some of the results they present in their paper.

In the same paper, Mitzenmacher et al. sow the existence of an algorithm that uses $exp(\sqrt{n})$ many traces and that is able to reconstruct any sequence with high probability, where n is the size of the sequence to be reconstructed.

A related version of the same Trace Reconstruction Problem is that where T is an unknown sequence of length n over a fixed alphabet $\mathcal{X}$ and the goal is to reconstruct T after making questions of the form, "does the sequence S appear as a subsequence of T?

Dekel Tsur, in his paper "Tight bounds for string reconstruction using substring queries" [5] proves that every non adaptive algorithm must make $\Omega(\epsilon^{\frac{1}{2}}n^2)$ queries in order to reconstruct $1-\epsilon$ fraction of the strings of length n.

For the purposes of this Thesis, we consider the following version of the Trace reconstruction problem:

Problem 1. *The Trace Reconstruction Problem asks to devise an algorithm whereby after receiving a series of subsequences $t_1, \cdots, t_r$ from a hidden random sequence $w \in \{0,1\}$. The sequence w is assumed to be drawn from the uniform distribution. The sampling distribution over subsequences equals $\mathcal{P}_{w,p_n}(\cdot)$ for some probability parameter $p_n \in (0,1)$.*

The full version of the Trace Reconstruction Problem asks for the existence of an algorithm that after processing polynomially many samples and using polynomial post processing time, is capable of reconstructing, with exponentially small probability of error, the string w from which the samples were generated. The hidden string, w is assumed to be drawn uniformly at random from the space $\{0,1\}^n$.

A milder version of the Trace Reconstruction Problem asks to show that there is an algorithm that although only requires polynomially many samples, uses exponen-

tial post processing time with the samples to reconstruct, with exponentially small probability of error, the string w from which the samples were generated. As we mentioned before, Mitzenmacher et al. show in [2] show an algorithm capable of reconstructing with high probability a high proportion of the strings, but failing on the rest, and an algorithm capable of reconstructing all the strings but requiring both exponentially many samples, and therefore exponentially big post processing time, thus falling short from fully solving both the full and the milder version of the Trace Reconstruction Problem.

In this Thesis although we will talk about the Trace Reconstruction Problem, we will mainly focus on solving a very related problem, which we call the Pair Trace Identification Problem:

Problem 2. *Given two known sequences $w, w' \in \{0,1\}^n$, and a known parameter of probability $p \in [0,1]$. Design an algorithm that uses polynomially many samples and uses polynomial post processing time that allows to decide, with high probability and after receiving polynomially many traces from either w or w' (with a probability distribution over subsequences induced by the probability parameter p) which sequence generated the sequence of observed traces.*

In the context of the Pair Trace Identification Problem, we would like to show that the sequences are distinguishable, not only in theory, but in practice. This entails to show that the distributions induced by the deletion process are 'far' away from each other, enough to allow a sampling algorithm to distinguish the distribution originating the samples. To this purpose we will define an appropriate notion of distance between distance in Chapter 1.

The full version of the Pair Trace Identification Problem asks for the existence of an algorithm that after processing polynomially many samples and using polynomial post processing time, is capable of distinguishing, with exponentially small probability of error, if the samples came from w or w' for all possible pairs of strings $w, w' \in \{0,1\}^n$ with $w \neq w'$.

A milder version of the Pair Trace Identification Problems as to show that there

is an algorithm that requires only polynomially many samples, but fails only for a exponentially small proportion of the pairs w, w'.

In what follows, any algorithm capable to distinguish with high probability and only polynomially many samples if the observed traces are coming from either w or w', is referred to as a statistical test for the Pair Trace Identification Problem.

Our results will be mainly concerned with the case where $p = \frac{1}{2}$. The crux result that we achieve in this Thesis is a solution for the milder version of the Pair Trace Identification Problem. In other words, with high probability over the space of pairs of sequences, $w, w' \in \{0,1\}^n$, there exists an algorithm that distinguishes w from w' after only polynomially many traces and uses polynomial post processing time.

In our discussion, we will also talk about some other related problems to both the Trace Reconstruction Problem and the Pair Trace Identification Problem, in particular, the techniques and results we develop towards the solution of the Pair Trace Identification Problem will lead us to ask what is the shortest length of a sequence $k_0(n)$ such that for any two sequences $w, w' \in \{0,1\}^n$, there is a subsequence s with $s \in \{0,1\}^{k_0(n)}$ appears a different number of times in w than in w'. More formally:

A given sequence $w \in \{0,1\}^n$ contains $\binom{n}{k}$ subsequences of length k. Call the multiset of subsequences of size k of w, the $k-$deck of w. A sequence that is uniquely defined by its $k-$deck is called $k-$reconstructible. A sequence $w \in \{0,1\}^n$ is $k-$reconstructible if no other sequence has the same $k-$deck as w. For example, all sequences of length 4 are 3$-$reconstructible.

The following problem which in this Thesis we will refer to as the k-Trace reconstruction problem asks:

Problem 3. *What is the minimum k such that all sequences of $\{0,1\}^n$ are $k-$reconstructible. We call $k_0(n)$ to this minimum k.*

The current results on the $k-$Trace Reconstruction Problem can be summarized in the following:

$$c_1 \log^2(n) \leq k_0(n) \leq c_2\sqrt{n}$$

For some absolute constants $c_1, c_2 > 0$.

The lower bound is due to a constructive proof by Schulman and Dudik and it is shown in their paper 'Reconstruction from subsequences' [7]. The upper bound is a result by Krasikov and Roditty, and its proof can be found in their paper 'On a reconstruction problem for subsequences' [8].

Interestingly, the proof of the upper bound is closely related to the number theoretic result known as the Prouhet-Tarry-Escott problem. In Chapter 5, Extensions we will elucidate its connections with the Reconstruction Problems discussed in this Thesis.

Finally, as we mention before we will define a notion of distance between distributions, and will address the question of, finding a lower bound for the distance between any two distributions over subsequences induced by the deletion process here described and where $p = \frac{1}{2}$. The notion of distance between distributions that we will use in this Thesis is that of variation distance, which is defined in the following paragraphs.

For a given sequence $w \in \{0,1\}^n$ and a deletion parameter p, we define the distribution over subsequences of w induced by the deletion process here described as $\mathcal{P}_{w,p}(\cdot) : \{s \trianglelefteq w\} \to [0,1]$. $\mathcal{P}_{w,p}(\cdot)$ can be written out explicitly:

Definition 4. $\mathcal{P}_{w,p}(s) = C(w,s)p^{n-|s|}(1-p)^{|s|}$

Assume $w, w' \in \{0,1\}^n$. Let $d_{TV}(P_{w,p}(\cdot), P_{w',p}(\cdot)) = \frac{1}{2}\sum_{s \trianglelefteq w | s \trianglelefteq w'} |P_{w,p}(s) - P_{w',p}(s)|$ be the variation distance between two distributions induced by the deletion process with probability p.

Problem 5. *Find the minimum value of $d_{TV}(P_{w,p}(\cdot), P_{w',p}(\cdot))$ for all pairs of strings $w, w' \in \{0,1\}^n$. In particular, find the exact value of this lower bound when $p = \frac{1}{2}$ and the pair of strings w, w' that achieve it.*

We will discuss some advances and conjectures related to the solution of the Problem 5 in the last chapter of this Thesis, Extensions.

1.1 Summary of results and thesis structure

In this thesis we will focus on the Pair Trace Identification Problem for the case when the deletion probability equals $\frac{1}{2}$. The first chapter, titled "Preeliminaries", introduces a few efficient algorithms for computing a few combinatorial properties of subsequences.

In the second chapter, "Distinguishing tests and special string pairs", we explore a few statistical tests and their performance. Additionally we explore a few families of string pairs for which we can show their statistical distinguishability.

In the third chapter, "Random strings are distinguishable" we provide a proof that with high probability, pairs of uniformly random strings are statistically distinguishable. This is the main result of this section, and hinges upon some of the tests and results introduced in Chapter 2. In particular, the statistical test that is used in the proof of this theorem is a specialization of the test introduced in Chapter 2, which we call the "Subsequences of Subsequences Test".

The last chapter, "Extensions" discusses among other things, further directions towards the full solution of the Pair Trace Identification Problem, and the different and rich connections that the results exposed in Chapter 3 have with the k-trace reconstruction problem. We provide an alternative proof that matches the asymptotics of the existing upper bound for the following quantity:

$$k_0(n) = \max_{w \neq w' | w, w' \in \{0,1\}^n} \min_{k'} \text{s.t.} C(w, s_{k'}) \neq C(w', s_{k'})$$

Additionally we discuss a purely number theoretical implication of this proof, which to the knowledge of the author of this paper, shows an improvement on existing results concerning a particular class of Littlewood type problems on $[0, 1]$.

The reader might want to skip some of the sections. In particular, if the reader's intention is to read the Chapter titled 'Random Strings are Distinguishable', a sufficient read would go over the Preeliminaries Chapter, and the section on the subsequences of subsequences test in Chapter 3, and through all of the 'Random Strings are Distinguishable' Chapter.

Chapter 2

Preeliminaries

2.1 Subsequences and their properties

In this section we describe some of the properties of subsequences. The purpose of this discussion is to present some results about how to compute the number of subsequences of a given sequence. We will focus on sequences where the alphabet is $\mathcal{X} = \{0, 1\}$.

2.1.1 Sequence representation

Here we present an overview on the distinct ways in which we will represent sequences in the incoming sections and chapters.

Definition 6. *Every sequence $w \in \{0,1\}^n$ can be naturally identified with a vector in $\mathbb{R}^n$ whose entries are either 0 or 1. We call this the $0-1-$vector of w.*

Definition 7. *The $0-$vector of a sequence $w \in \{0,1\}^n$ with m ones is $(b_0, \cdots, b_m)$. Where b_i indicates the number of zeroes before the $i+1$-th and after the ith one. By convention, b_0 is the number of zeroes before the first one, b_m is the number of zeroes after the last one.*

Definition 8. *The $1-$vector of a sequence $w \in \{0,1\}^n$ with m zeroes is $(a_0, \cdots, a_m)$. Where a_i indicates the number of ones before the $i+1$-th and after the ith zero. By*

convention, a_0 is the number of ones before the first zero, a_m is the number of ones after the last zero.

Definition 9. *The position-of-ones of $w \in \{0,1\}^n$ equals $(x_1, \cdots, x_m)$ if w has m ones and the $i-th$ one of w is in position x_i. We define analogously the position-of-zeroes vector representation of w.*

2.1.2 Algorithms

We present a variety of algorithms over sequences and their subsequences that will prove handy in our discussion in the pages to come.

Subsequence appearances

Let $C(w, s)$ = number of times s is a subsequence of w.

We provide an algorithm that takes as inputs $w \in \{0,1\}$ and $s \in \{0,1\}^k$ where $k \leq n$ and outputs $C(w, s)$ in polynomial time.

The procedure hinges upon the following observation:

Observation 10. *If $s_k = w_n$ then:*

$$C(w, s) = C(w[: n-1], s) + C(w[: n-1], s[: k-1])$$

else:

$$C(w, s) = C(w[: n-1], s)$$

The observation 10 is enough to set the bases for a recursive algorithm capable of computing $C(w, s)$.

If for $j = 1, \cdots, i$ and $r = 1, \cdots, k$ we know $C(w[: k], s[: r])$ we can compute, by means of 10, the values of $C(w[: i+1], s[: r])$ for all $r = 1, \cdots, k$.

In fact, to compute the values $C(w[: i+1], s[: r])$ for all $r = 1, \cdots, k$ we only need to know the values of $C(w[: i], s[: r]) \forall r = 1, \cdots, k$.

The proposed algorithm computes the values of $C(w[:i], s[:r])$ for all values of $i \in [n]$ and $r \in [k]$ via memoization.

Algorithm 11. *1. Compute the array of values $C(w[:1], s[:r])$ for $r = 1, \cdots, k$.*

2. For all i use the observation in the previous paragraph to compute the array of values $C(w[:i+1], s[:r])\forall r \in [k]$ from the array of values $C(w[:i], s[:r])\forall r \in [k]$.

The basis of the algorithm is the computation of the array of values for $C(w[:1], s[:r])$. This equals:

$$C(w[:1], s[:1]) = 1 \text{ if } w[:1] = s[:1]$$
$$C(w[:1], s[:1]) = 0 \text{ o.w.}$$

Furthermore, by convention, we specify that $s[:0] = \emptyset$ and $C(w, \emptyset) = 1\forall w \in \{0,1\}^n$.

This algorithm takes $O(k)$ operations per step. Since the number of times these recursive calculations must be ran n times, the running time of this algorithm is $O(nk)$, which is polynomial in n, the size of w.

Subsequence count polynomials

Let $w \in \{0,1\}^n$, and $s \in \{0,1\}^k$ for some k. Let $C(w,s)$ be the number of times s appears as a subsequence of w.

Lemma 12. *Let $x_1, \cdots, x_m$, be the positions of the ones of w. For fixed m, n, we can write, $C(w,s)$ as $f_s(x_1, \cdots, x_m)$, a polynomial on $x_1, \cdots, x_m$.*

Proof. If s has no ones, then $f_s(x_1, \cdots, x_m) = \binom{n-m}{|s|}$ which is constant (by assumption m, n are fixed). Let $s = 0^k 1 s'$ where $k \geq 0$, need not be different from zero.

By summing over the position of the first one of s within w, either at x_1 or x_2, or

... x_m. We get that:

$$f_s(x_1, \cdots, x_m) = \sum_{j=1}^{m} f_{s'}(x_{j+1} - x_j, \cdots, x_m - x_j) * \binom{x_j - j}{k}$$

Where each $f_{s'}(x_{j+1} - x_j, \cdots, x_m - x_j)$ specifies the number of occurrences of s' in the region $w[x_j + 1 :]$ and $\binom{x_j - j}{k}$ is the number of ways to choose k zeros from the first $x_j - j$ zeros. Hence in overall each factor of the sum, $f_{s'}(x_{j+1} - x_j, \cdots, x_m - x_j) * \binom{x_j - j}{k}$ specifies the number of times s appears as a subsequence of w where the first one of s coincides with the j-th one of w.

The recursive nature of this definition finishes the proof.

□

From this result we can provide an alternative polynomial time algorithm to compute the value of $C(w, s)$.

By employing memoization, the recursion above yields a $O(n^2)$ algorithm for computing $f_s(w)$, for any given s. The advantage of this method is the fact that the polynomial expression $f_s(x_1, \cdots, x_m)$ can be reevaluated for different values of the array $(x_1, \cdots, x_m)$. For some sequences s, the expression $f_s(x_1, \cdots, x_m)$ can be evaluated in $O(n)$ time instead of $O(n^2)$.

The family of polynomials $\{f_s(x_1, \cdots, x_m)\}$ for all sequences s of size less or equal to n is an interesting object of study. Let $\mathcal{F}^{\Updownarrow}{}_k = \{f_s(x_1, \cdots, x_m)\}_{|s| \leq k}$ for a fixed m. The following question touches upon the independence structure of the $\mathcal{F}^{\Updownarrow}{}_k$.

Open Problem 13. *What is the span of $\mathcal{F}^{\Updownarrow}{}_k$ over the space of multivariate m variate polynomials?*

The motivation behind 13 stems from the observation that if $\mathcal{F}^{(\Updownarrow)}{}_k$ spans all the space of m variate polynomials $(x_1, \cdots, x_m)$, then there must exist a string of size k such that for every two strings $w, w' \in \{0,1\}^n$, with m ones in positions $x_1, \cdots, x_m$ and $x'_1, \cdots, x'_m$ respectively, $f_s(x_1, \cdots, x_m) \neq f_s(x'_1, \cdots, x'_m)$ or in other words, $C(w, s) \neq C(w', s)$

In the Chapter 'Extensions' we will provide some partial results regarding the Open Problem 13.

2.1.3 Counting common subsequences

In the case where the probability parameter of the Pair Trace Identification Problem is $p = \frac{1}{2}$, a given pair of sequences $w, w' \in \{0,1\}^n$ are identifiable, under the assumptions of the Pair Trace Identification Problem (Problem 2) if we consider the multiset of sequences of w that are not subsequences of w', and the cardinality of this set is of order $O(\frac{2^n}{n^c})$ for some universal constant c. If this was the case under the sampling procedure described by the Pair Trace Identification Problem statement, and with parameter $p = \frac{1}{2}$, the probability of observing a substring of w that is not a substring of w' given that the hidden sequence was w is at least $\frac{c_1}{n^c}$ for some constant c_1. In this case, the following algorithm distinguishes them after polynomially many samples (The condition and algorithm are symmetric if we exchange w by w'):

Algorithm 14. *Repeat n^c times: For any given sample s compute $C(w,s)$ and $C(w',s)$. If $C(w,s) \neq 0$ and $C(w',s)$ then output w as the belief for the hidden sequence, if after n^c times no such sequence s has appeared, output w' as the belief for the hidden sequence.*

In this section we present an algorithm that allows us to compute, for a given pair of sequences w and w', the cardinality of the multiset of subsequences of w that are not subsequences of w'. The spirit of investigating such a procedure lies in the observation above. If this set is "polynomially large" with respect to the space of 2^n possible traces, we hope to use the algorithm 14 to identify the source of the traces.

In the next section we present an algorithm that allows to compute the probability $p_{m(n)}$ of observing a subsequence of w that is not a subsequence of w' when the deletion parameter is an arbitrary value $p \in [0,1]$. If this probability $p_{m(n)}$ were of the form $\frac{c_1}{n^c}$ for some universal constants $c_1, c \in \mathbb{R}^+$, the same algorithmic procedure described in Algorithm 14 will solve the Pair Trace Identification problem between w and w'.

Unfortunately, the required separability property under which Algorithm 14 works

is not achieved by all sequence pairs w, w'. In particular, it is not achieved by the sequence pair, $w, w' = 0^k 10^{k-1}, 0^{k-1}10^k$. In this case, the sequences that are subsequences of w but not of w' are only those that are of the form $0^k 10^*$. There are 2^k of them, which is not a polynomial fraction of 2^{2k}.

Recall the following definition:

Definition 15. *Let $s \trianglelefteq w$ denote s is a subsequence of w.*

Let $w = w_1 \cdots w_m$ and $w' = w'_1 \cdots w'_n$.

Definition 16. *Let $\Phi(w, w') = \sum_{s \trianglelefteq w} 1_{s \trianglelefteq w'} = \sum_{s \trianglelefteq w} C(w', s)$. Where 1_A is the indicator function of A and $B = \{s | s \trianglelefteq w\}$ set of distinct sequences that are subsequences of w.*

Definition 17. *Let*

$$\begin{aligned}\Phi_0(w, w') &= \sum_{s \cup \{0\} \trianglelefteq w} 1_{s \cup \{0\} \trianglelefteq w'} \\ &= \sum_{s \cup \{0\} \trianglelefteq w} C(w', s \cup \{0\})\end{aligned}$$

We are interested in finding out a recursion for Φ. We will derive the desired recursion in several steps. The idea underlying the derivation is to construct the common subsequences of w and w' by observing their last character.

$$\Phi(w \cup \{0\}, w') = \Phi(w, w') + \text{ extra terms.}$$

Some subsequences of $w \cup \{0\}$ that are subsequences of w' as well, could lie only within w.

The remaining subsequences of $w \cup \{0\}$ that are subsequences of w' as well, all finish in the last 0 of $w \cup \{0\}$. Among these, there are two types of strings:

1. subsequences of $w \cup \{0\}$ having the added 0 as last character that are also subsequences of w and w'.

2. subsequences of $w \cup \{0\}$ having the added 0 as last character that are not subsequences of w but are subsequences of w'.

The first subset is exactly $\Phi_0(w, w')$. This is because one can take a subsequence of w finishing in zero, and substituting that last character with the added zero of $w \cup \{0\}$. This term comes into the sum by substituting Φ_0 last zero with the added zero in $w \cup \{0\}$.

$$\text{To count the second susbset of strings} = \sum_{\substack{s \trianglelefteq w \\ s \cup \{0\} \not\trianglelefteq w}} 1_{s \cup \{0\} \trianglelefteq w'}$$

In other words, the number of subsequences of w such that adding a zero they still belong to w' but not to w. To find a decomposition of the last desired term:

Say $s \trianglelefteq w$, $s \trianglelefteq w'$ is such that $s \cup \{0\} \trianglelefteq w$ and $s \cup \{0\} \trianglelefteq w'$ but $s \cup \{0\} \not\trianglelefteq w$. Call $w[:i] = w_1 \cdots w_i$. Then call r to the largest index of w such that $w_r = 0$. Define r' analogously for w'. Notice that $s \not\trianglelefteq w[:r-1]$, since otherwise $s \cup \{0\} \trianglelefteq w[:r] \trianglelefteq w$. Furthermore, since $s \cup \{0\} \trianglelefteq w'$, one must have that $s \trianglelefteq w'[:r'-1]$. And therefore we have the conditions:

1. $s \not\trianglelefteq w[:r-1]$
2. $s \trianglelefteq w$
3. $s \trianglelefteq w'[:r'-1]$

These conditions are necessary and sufficient for $s \cup \{0\}$ being in the set we are trying to enumerate. Necessary follows from above, sufficiency follows easily as well. The value we are looking for is $= \Phi(w, w'[r'-1]) - \Phi(w[:r-1], w'[r'-1])$ which completes the recursion. In synthesis:

$$\Phi(w \cup \{0\}, w') = \Phi(w, w') + \Phi_0(w, w') + \Phi(w, w'[:r'-1]) - \Phi(w[:r-1], w'[:r'-1])$$

We now provide a recursion for $\Phi_0(w, w')$. Observe first that $\Phi_0(w, w') = \Phi_0(w[:r], w'[:r'])$ meaning that if either of w or w' does not end in 0 then we have a recursion

into a shorter sequence. We are therefore interested in finding a recursive definition for the remaining case $\Phi(w \cup \{0\}, w' \cup \{0\})$. To obtain such a formula, we perform a similar process as before. The subsequences of $w \cup \{0\}$ ending in 0 common to $w'\{0\}$ can be divided into the following sets:

1. subsequences of w ending in zero common to $w' \cup \{0\}$.
2. subsequences of $w \cup \{0\}$ having the added 0 as last character that are subsequences of w and are subsequences of $w' \cup \{0\}$.
3. subsequences of $w \cup \{0\}$ having the added 0 as last character that are not subsequences of w and are subsequences of $w' \cup \{0\}$.

The size of the first subset is $\Phi_0(w, w' \cup \{0\})$. This follows merely by definition. The size of the second subset is $\Phi_0(w, w' \cup \{0\})$. This follows by considering all subsequences of w having a zero at the end and substituting that zero by the added zero in $w \cup \{0\}$. The third subset is $\{s \trianglelefteq w | s \cup \{0\} \not\trianglelefteq w, s \cup \{0\} \trianglelefteq w' \cup \{0\}\}$. To count this, notice that

$$\begin{aligned} \{s \trianglelefteq w | s \cup \{0\} \not\trianglelefteq w, s \cup \{0\} \trianglelefteq w' \cup \{0\}\} = \\ \{s \trianglelefteq w | s \cup \{0\} \not\trianglelefteq w, s \trianglelefteq w'\} = \\ \{s \trianglelefteq w | s \cup \{0\} \not\trianglelefteq w[:r-1], s \cup \{0\} \trianglelefteq w' \cup \{0\}\} \end{aligned}$$

The equivalence of the first and the second identity is trivial while the equivalence of the second and the third identity follows from the same argument for the first part since the condition that $s \not\trianglelefteq w$ is equivalent to $s \not\trianglelefteq w[:r-1]$ where r is the largest index of w such that $w_r = 0$. Therefore $\Phi_0(w \cup \{0\}, w' \cup \{0\}) = 2\Phi_0(w, w' \cup \{0\}) + \Phi(w, w') - \Phi(w[:r-1], w')$. In synthesis:

- $\Phi(w \cup \{0\}, w') = \Phi(w, w') + \Phi_0(w, w') + \Phi(w, w'[:r'-1]) - \Phi(w[:r-1], w'[:r'-1])$
- $\Phi_0(w \cup \{0\}, w' \cup \{0\}) = 2\Phi_0(w, w' \cup \{0\}) + \Phi(w, w') - \Phi(w[:r-1], w')$

Probabilistic count

In this section we devise a method to compute the probability, given the underlying sequences are w and w', of observing a trace (subsequence) that is common to both w and w', given the underlying sequence is (say) w.

Definition 18. *Define* $\Phi^{(p)}(w, w') = \sum_{s \trianglelefteq w} 1_{s \trianglelefteq w'} p(s|w)$. *Where* $p(s|w)$ = *probability of observing* s *given* w.

It is easy to see that $p(s|w) = p^{|w|-|s|}(1-p)^{|s|}$. $\Phi^{(p)}(w, w')$ is the probability that a common subsequence of w and w' is observed as a trace of w.

We can easily extend the previous' section recurrence relations to yield the following ones:

$$\Phi^{(p)}(w \cup \{0\}, w') = p\left(\Phi^{(p)}(w, w') + \Phi_0^{(p)}(w, w') + \Phi^{(p)}(w, w'[:r'-1])\right) - p^{|w|-r+1}\Phi^{(p)}(w[:r-1], w'[:r'-1])$$

Where

$$\Phi_0^{(p)}(w, w') = \sum_{s \cup \{0\} \trianglelefteq w} 1_{s \cup \{0\} \trianglelefteq w'} p(s \cup \{0\}|w)$$

Extending the recursion for $\Phi_0(w, w')$ into the probabilistic case, we get:

$$\Phi_0^{(p)}(w \cup \{0\}, w' \cup \{0\}) = p\left(2\Phi_0(w, w' \cup \{0\}) + \Phi(w, w')\right) - p^{|w|-r+1}\Phi(w[:r-1], w')$$

The recursions above allow us to compute these counts or probabilities in polynomial time by using memoization and noting the base cases:

1. $\Phi(0, w') = 1_{0 \in w'}$
2. $\Phi_0(0, w') = 1_{0 \in w'}$

3. $\Phi(1, w') = 1_{1 \in w'}$

4. $\Phi_0(1, w') = 0$

The general recursion needs a term:

$\Phi_1(w, w')$

This is defined analogously to $\Phi_0(w, w')$ both in the counting and probabilities case.

A potentially fruitful and interesting discussion could arise from generalizing the previous algorithm for the case of larger alphabets than $\{0, 1\}$. Because of lack of its relatedness to the main topic of this thesis we do not pursue that discussions here.

2.2 Variation distance

In order to establish the solvability of the Pair Trace Identification Problem, one would like first to show that the distributions over subsequences, $\mathcal{P}_{w,p}$ and $\mathcal{P}_{w',p}$ induced by the deletion process described for the Pair Trace Identification Problem, are far apart. The notion of distance between distributions that we will focus on in these notes is that which is known as Variation Distance.

A result we would like to establish is the theoretical possibility of distinguishing between every two given strings w and w', each in $\{0, 1\}^n$ when they are subject to the sampling conditions of the Pair Trace Identification Problem. In order to achieve this goal, we would like to show that the variation distance between the distributions over subsequences $\mathcal{P}_{w,p}, \mathcal{P}_{w',p}$ is "large". In other words, we would like to prove the distributions $\mathcal{P}_{w,p}$ and $\mathcal{P}_{w',p}$ are far apart.

In the context of this thesis we will say that two distributions are "far away" when there is a polynomial gap between the two of them. More formally:

Definition 19. *A sequence of families of distributions $\{\mathcal{F}_\setminus\}_{n=1}^{\infty}$ such that each $p_n \in \mathcal{F}_n$ is supported over a space of size $O(2^n)$ is called polynomially distinguishable if there is a universal constant c such that the variation distance between every two different distributions $p_n, q_n \in \mathcal{F}_n$ is at least $\Omega(\frac{1}{n^c})$.*

In the context of the trace identification problem, $\mathcal{F}_n = \{P_{w,p}\}_{w\in\{0,1\}^n}$. A question that will be explored in the incoming sections of this thesis is the following one:

Open Problem 20. *Is the family* $\{\mathcal{F}_n\}_{n=1}^{\infty}$ *where* $\mathcal{F}_n = \{P_{w,p}\}_{w\in\{0,1\}^n}$ *polynomially distinguishable?*

In the following section we will define the concept of Variation distance and explore some of its basic properties.

2.2.1 General properties

Given two probability distributions, the variation distance measures how far are the two of them from one another. In this thesis we will restrict ourselves to probability distributions supported over a discrete set.

Definition 21. *Given two distributions* μ, ν *supported over a discrete set* S *the variation distance between the two is defined as:*

$d_{TV}(\mu,\nu) = \| \mu - \nu \|_{TV} = \sup_{A\subset S} |\mu(A) - \nu(A)|$

One can understand the total variation distance as the maximal deviation between ν and μ when evaluated over any measurable set. We would like to prove that the variation distance between the induced distributions over subsequences for any pair of distinct sequences (w, w') is large enough so to allow the existence of a differentiating test.

It is easy to show that for a pair of discrete distributions μ, ν supported over a discrete set S, the variation distance $d_{TV}(\mu,\nu)$ can be written as:

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sum_{s\in S} |\mu(s) - \nu(s)|$$

Definition 22. *We say that the variation distance* $d_{TV}(\mu_n,\nu_n)$ *is polynomially large for a pair of distributions (indexed by n)* μ_n *and* ν_n *defined over an alphabet of size* 2^n *if* $d_{TV}(\mu_n,\nu_n) = \Omega(\frac{1}{n^c})$ *for some fixed constant* $c > 0$.

2.2.2 Hoeffding Inequality

The following inequality, known as Hoeffding inequality will be used multiple times in this Thesis.

Theorem 23. *Let $X_1, \ldots, X_n$ be independent random variables. Assume that the X_i are almost surely bounded; that is, assume for $1 \leq i \leq n$ that $\Pr(X_i \in [a_i, b_i]) = 1$. We define the empirical mean of these variables as $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. The following inequalities hold:*

$$\Pr(\overline{X} - \mathrm{E}[\overline{X}] \geq t) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$
$$\Pr(|\overline{X} - \mathrm{E}[\overline{X}]| \geq t) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

Hoeffding bound allows us to develop the following bounds for approximating the probability of a given event via sampling.

Suppose we are sampling from a family of random variables $\{\mathcal{X}_n\}_n$ and a family of events $\{A_n\}$. Let $\mathbb{P}(\mathcal{X}_n \in A_n) = p_{A_n}$

Suppose $p_{A_n} = \Omega(\frac{1}{n^k})$.

$$Y_n = \begin{cases} 1 & \text{if } \mathcal{X}_n \in A_n \\ 0 & \text{otherwise} \end{cases}$$

After generating R samples $Y_n^{(1)}, \cdots, Y_n^{(R)}$ from Y_n, since $Y_n \in [0, 1]$ the Hoeffding bound ensures that:

$$\Pr\left(\left|\frac{\sum_{i=1}^{R} Y_n^{(i)}}{R} - A_n\right| \geq \epsilon\right) \leq \exp\left(-\frac{2R^2\epsilon^2}{R(1)^2}\right) = \exp\left(-2R\epsilon^2\right)$$

In order to approximate $p_{A_n} = \Omega(\frac{1}{n^k})$ such that the probability of being far from its actual value by an additive factor of $\frac{1}{n^m}$ is of the form e^{-cn} for some constant $c > 0$

we need:

$$\begin{aligned}\exp(-2R\epsilon^2) &= \exp(-cn)\\ \rightarrow \epsilon &= \frac{1}{n^m}\\ \exp(-2R\frac{1}{n^{2m}}) &= \exp(-cn)\\ \rightarrow R &= O(n^{2m+1})\end{aligned}$$

This leads to the following lemma:

Lemma 24. *Let sequence of events* $\{A_n\}_{n=1}^{\infty}$ *defined over probability spaces* $\{\Omega_n\}_{n=1}^{\infty}$ *be such that* $\mathbb{P}(A_n) = \Omega(\frac{1}{n^k})$ *for some* $m \in \mathbb{N}$. U_n *be a uniform random variable over* Ω_n *for all* n. *It is possible to approximate* $\mathbb{P}(A_n)$ *up to an error of* $\frac{1}{n^{k+1}}$ *using* $O(n^{2(k+1)+1})$ *samples such that the probability of error is bounded by* $\exp(-2n)$.

2.2.3 Sampling from polynomially distinguishable families

In this section we prove a result concerning the distinguishability of two distributions whose variation distance is large.

Let p and q be two distributions over a set Ω. Consider the following four quantities:

$$\begin{aligned}A_{1,q} &= E_q[1_{q(i)>p(i)}]\\ A_{1,p} &= E_p[1_{q(i)>p(i)}]\\ A_{2,q} &= E_q[1_{p(i)>q(i)}]\\ A_{2,p} &= E_p[1_{p(i)>q(i)}]\end{aligned}$$

Notice that $d_{TV}(p,q) = \frac{1}{2}([A_{1,q} - A_{1,p}] + [A_{2,p} - A_{2,q}])$

This means that if $d_{TV}(p,q) \geq \frac{\delta}{2}$ then at least one between $[A_{1,q} - A_{1,p}]$ or $[A_{2,p} - A_{2,q}]$ is greater of equal than $\delta/2$. Wlog say $[A_{1,q} - A_{1,p}] \geq \delta/2$

If p and q are such that for every given $i \in \Omega$ it is possible to compute $p(i)$ and $q(i)$, we can use sampling and the previous section's Hoeffding bounds to estimate

to high accuracy $[A_{1,q} - A_{1,p}]$ by using Lemma 24. If $d_{TV}(p,q)$ is polynomially large, then by using Lemma 24 we will only need polynomially many samples.

If we have a protocol whereby we are receiving samples from either p or q, but we don't know which distribution are they coming from, we can estimate the empirical value of $[A_{1,q} - A_{1,p}]$ for each belief of p and q and compare it to the estimate we have already sampled.

In particular, when $d_{TV}(p,q)$ is polynomially large of order $\Omega(\frac{1}{n^c})$, then $O(n^{2c+1})$ samples are required to estimate up to accuracy $\Omega(\frac{1}{n^{c+1}})$ the value of $[A_{1,q} - A_{1,p}]$ or $[A_{2,p} - A_{2,q}]$. Call this estimate A. After sampling from the hidden distribution, we can estimate A assuming the two cases:

1. Hidden distribution is p
2. Hidden distribution is q

Call these estimates A_p and A_q respectively. After $O(n^{2c+1})$ samples, the estimate A_p or A_q yielding the closest estimate to the previously estimated value of A will provide us a value of the hypothesis. The algorithm fails with probability bounded by $\exp(-nc)$ for some constant $c > 0$.

2.2.4 Subsequences variation distance

As we have mentioned before, we will focus on the Pair Trace Identification Problem in the case where $p = \frac{1}{2}$. This is important for this section, since if the hidden sequence were $w \in \{0,1\}^n$ the probability that the sampled trace was made up of exactly the bits $\{w_{i_1}, \cdots, w_{i_t}\}$ for some set of values $\{i_1, \cdots, i_t\} \subset [n]$ is $\frac{1}{2^n}$ which is independent of the set $\{i_1, \cdots, i_t\}$.

Let $S_n = \{w \in \emptyset \cup \{0,1\} \cdots \{0,1\}^n\}$ be the set of binary sequences of length at most n. Notice that $|S_n| = 2^{n+1} - 1$. Let $\mathbb{S}_n$ be the $2^{n+1} - 1$ dimensional space whose basis vectors are all sequences of S_n.

Define the following map:

$$f_n : \{0,1\}^n \to \mathbb{S}_n$$

f_n maps any string $w \in \{0,1\}^n$ to the vector of frequencies of all its subsequences. For example:

$$f_3(001) = 1 \cdot e_\emptyset + 2 \cdot e_0 + 1 \cdot e_1 + 1 \cdot e_{00} + 2 \cdot e_{01} + 1 \cdot e_{001}$$

Where e_s denotes the basis vector in $\mathbb{S}_n$ corresponding to the coordinate associated with the sequence s.

Because we are dealing with the case $p = \frac{1}{2}$, instead of dealing with the distributional variation distance we will deal with the following more convenient distance:

Definition 25. *Given two binary sequences: w, w' let $d_s(w, w') = |f_n(w) - f_n(w')|_1$ be the l_1 distance between the vector images in $\mathbb{S}_n$ of each.*

This construction makes S_n into a Hilbert space. I.e. d_s induces a norm, and a dot product. Interestingly, $|f_n(w)| = 2^n$. All these vectors are on the l_1 sphere of radius 2^n.

Notice that $d_{TV}(\mathcal{P}_{w,\frac{1}{2}}, \mathcal{P}_{w',\frac{1}{2}}) = \frac{1}{2^{n+1}} d_s(w, w')$. In order to lower bound the variation distance of $d_{TV}(\mathcal{P}_{w,\frac{1}{2}}, \mathcal{P}_{w',\frac{1}{2}})$ is sufficient to find a lower bound for $d_s(w, w')$. Showing $d_{TV}(\mathcal{P}_{w,\frac{1}{2}}, \mathcal{P}_{w',\frac{1}{2}})$ is polynomially big is equivalent to show that there exists a constant $c \in \mathbb{R}^+$ such that for all pairs of distinct sequences $w, w' \in \{0,1\}^n$ we have that $d_s(w, w') = \Omega(n^c)$.

By our discussion on 'Sampling from polynomially distinguishable families' it follows that, a polynomially big variation distance for the distributions $\mathcal{P}_{w,\frac{1}{2}}, \mathcal{P}_{w',\frac{1}{2}}$ for any two sequences $w, w' \in \{0,1\}^n$ implies the existence of a statistical distinguishing test that allows to solve the Pair Trace Identification Problem with high probability for every pair of strings when $p = \frac{1}{2}$.

Chapter 3

Distinguishing tests and special string pairs

In this chapter we focus on developing additional machinery towards the solution of the Pair Trace Identification. In particular, we will propose a variety of statistical distinguishing tests that are here proposed to serve as partial solutions for the Pair Trace Identification Problem. Recall the Pair Trace Identification Problem statement:

Problem 26. *Given two known sequences $w, w' \in \{0,1\}^n$, and a known parameter of probability $p \in [0,1]$, design an algorithm that uses polynomially many samples and uses polynomial post processing time that allows to decide, with high probability and after receiving polynomially many traces, from which distribution are the samples coming from, either $\mathcal{P}_w^{(p)}(\cdot)$ or $\mathcal{P}_{w'}^{(p)}(\cdot)$.*

We would like to propose a statistical distinguishing test that achieves either one of the following:

(a) Polynomial number of traces, but unlimited processing time.

(b) Polynomial number of traces, polynomial processing time.

A solution to (a) would prove that the variation distance between the distributions over subsequences is polynomially large for every pair of sequences w, w'. The second, (b), would prove not only that the variation distance is polynomially large between

the two distributions induced by every pair of sequences w, w', but also, the existence of a polynomial time algorithm that processes these traces and decides if they are being sampled from $\mathcal{P}_w^{(p)}(\cdot)$ or $\mathcal{P}_{w'}^{(p)}(\cdot)$ in polynomial time.

3.1 Distinguishing tests

In this section we present a survey of different types of distinguishing tests and their properties.

Recall the following definition:

Definition 27. *Let* $C(w, s) = (\textit{number of times } s \textit{ is a subsequence of } w)$.

Definition 28. *We say that a pair of strings* w, w' *is distinguishable if there is a statistical test that allows to distinguish them in polynomial time after polynomially many samples.*

3.1.1 The number of ones test

We analyze the performance of the statistical distinguishing test based on counting the average observed number of ones (or zeroes) in the observed traces from the hidden sequence.

We show that all pairs of sequences (w, w') such that the number of ones in w is different from the number of ones in w', are distinguishable. That is, we provide a test that will distinguish w from w' after only polynomially many samples, and using polynomial post processing time.

Given a string $w \in \{0,1\}^*$, denote

1. $C_1(w)$ = number of ones of w.

2. $C_2(w)$ = number of zeros of w.

We say w and w' are $1-$unbalanced if $C_1(w) \neq C_1(w')$. For this section we will allow p to vary and attain any value in $(0,1)$ and not only $\frac{1}{2}$. We explore a sampling scheme

capable of discerning between sample traces coming from 2 unbalanced sequences. We will consider the following numbers:

$\alpha(i|w)$ =probability the observed sequence has i ones given w is the underlying seq.

It is easy to see that:

$$\alpha(i|w) = \begin{cases} \binom{C_1(w)}{i}(1-p)^i p^{C_1(w)-i} & \text{if } i \leq C_1(w) \\ 0 & \text{otherwise} \end{cases}$$

The distribution $\alpha(\cdot|w)$ is binomial with parameters $(1-p, C_1(w))$. The mean of $\alpha(\cdot|w)$ is $C_1(w) \cdot (1-p)$ If w and w' are a 1−unbalanced pair, the mean of the binomial distributions $\alpha(\cdot|w)$ and $\alpha(\cdot|w')$ differ by at least an additive factor of $1-p$. The argument for 0−unbalanced pairs yields a symmetric condition.

For a given w with m ones, we can compute the minimum number of samples s such that our estimate for the mean of $\alpha(\cdot|w_{hidden})$ will achieve a deviation from the mean of more than $\frac{1-p}{2}$ with probability less than $e^{-c|w|}$ for some constant c. A simple use of Hoeffding inequality yields:

$$s = O(|w|\frac{1}{(1-p)^2})$$

This value is polynomial in $|w|$ and in $\frac{1}{1-p}$. Since p is assumed to be fixed, this quantity yields a polynomial number of samples with respect to $|w|$.

We can immediately extend this result in the following way:

Lemma 29. *If p depends on the size of the sequence in such a way that $p = p_n = \Omega(\frac{1}{n^k})$ for some constant k, then this test needs only polynomially many samples.*

Corollary 30. *The sequence pair (w, w') with $|w| \neq |w'|$ is distinguishable in polynomial time and after polynomially many samples.*

3.1.2 The first one test

In this section we assume that $p = \frac{1}{2}$. The following is a purported test to differentiate between the two strings based on the idea of looking at the location of the first one.

In order to give the reader a feel for why the problem of devising a statistical distinguishing test is nontrivial we show that this test does not work as a distinguishing test for all pairs of sequences.

Let w and w' be two sequences of length n and let $I = \{i_1, \cdots, i_k\}$ be the positions for the ones in w and $I' = \{i'_1, \cdots, i'_k\}$ be the positions of the ones in w'. With $i_1 < i_2 < \cdots < i_k$ and $i'_1 < i'_2 < \cdots < i'_k$. Assume both w and w' have the same number of ones. Otherwise the "count the number of ones" and we could hope that the array of two tests ("number of ones" and "first one") would differentiate them. As mentioned before, and for ease of the discussion let $p = \frac{1}{2}$. Assume that $|I - I'| = |I' - I| = 1$.

The number of subsequences of w having a one in its first position is:

$$2^{n-i_1} + 2^{n-i_2} + \cdots + 2^{n-i_k}$$

More generally, the number of subsequences of w shaving a one in its i-th position is:

$$\binom{i_1 - 1}{i} 2^{n-i_1} + \binom{i_2 - 2}{i} 2^{n-i_2} + \cdots + \binom{i_k - k}{i} 2^{n-i_k}$$

The number of subsequences of w having no ones is 2^{n-k}. Because $|I - I'| = 1$, the two subsequences differ only in a single index. Assume the index they differ is the $j-$th position in which both w and w' have a one. Let them be (after relabeling of the indices) i_j and i'_j. Then:

$$d_s(w, w') = \sum_{i=1}^{n} \left| \binom{i_j - j}{i} 2^{n-i_j} - \binom{i'_j - j}{i} 2^{n-i'_j} \right|$$

The last sum is less or equal than:

$$\sum_{i=1}^{n} \left(\binom{i_j - j}{i} 2^{n-i_j} + \binom{i'_j - j}{i} 2^{n-i'_j} \right)$$

Now assume that j is roughly $\frac{n}{4}$ and i_j, i'_j are roughly $\frac{n}{2}$.

Then the sum above is roughly $n2^{\frac{3n}{4}}$. Which is exponentially far from 2^n.

3.1.3 The marginals test

We introduce the marginals test. We define the marginals test to be the following array of tests:

Let the positions of the ones in an n bit string w be $x_1, \cdots, x_m$. Define $p_i(w)$ as the probability that the i-th entry of an observed trace of w equals one.

$$p_i(w) = \frac{1}{2^n} \cdot \sum_{i=1}^{m} \binom{x_i - i}{i-1} * 2^{n-x_i}$$

In the following section we prove that there exists a pair of strings whose two arrays of marginals tests are exponentially close.

3.1.4 The failure of the marginals test

Let $P(i,j)$ = probability that the i-th symbol of w ends up in position j. By definition $P(i,j) = \frac{1}{2^i}\binom{i-1}{j-1}$. Consider the following matrix A where

$$A_{i,j} = \binom{j}{i-1} * 2^{n-i-1}$$

We define $\binom{n}{m} = 0$ for $m > n$. A is upper triangular. If we call $e_1, \cdots, e_n$ the indicator vectors on coordinates $1, \cdots, n$ then, if we identify a sequence $w \in \{0,1\}^n$ with a column vector v^w such that $v_i^w = 1$ if $w_i = 1$ and $v_i^w = 0$ if $w_i = 0$. Then,

$$p_i(w) = A \cdot v_i^w$$

In other words Pr(we observe a one in position j) $= \frac{1}{2^n} A v_j^w$. Let $v^{w'}$ be defined analogously.

The variation distance between the two summary statistics tests is:

$$|A(v^w - v^{w'})|_1$$

Let $v = v^w - v^{w'}$. By definition $v_i^w \in \{0, 1, -1\}$. The variation distance between the summary statistics of the tests applied to w and w' is:

$$|Av|_1$$

We prove the marginals test fails by exhibiting a pair of strings w_1 and $w_2 \in \{0,1\}^n$ for which the variation distance between the two arrays of marginals tests is exponentially small.

Theorem 31. *There are two strings w_1 and w_2 such that $|Aw_1 - Aw_2|_1 \leq n^{-\Omega_D elta(\log^2(n)))}$*

Because of **??** it is sufficient to find a vector v in $\{-1, 0, 1\}$ such that $|Av|_1$ is small enough.

For a given function f, let $\Delta`_d f$ to be the function so defined:

$$(\Delta_d f)(x) = f(x + d) - f(x)$$

In other words, the function Δ_d is the discrete derivative of f with step d. let's consider the following:

$$\Delta_{c_1, c_2, \cdots, c_k}(f) = \Delta_{c_1}(\Delta_{c_2, \cdots, c_k} f)$$

Consider the following recursively defined vector: $v^{(0)} = (1)$, and $v^{(i)} = (v^{(i-1)}, -v^{(i-1)})$. Let $w = (0^{n-1}, v^{(k)})$ where k be determined later. Then we have that if $y = A \cdot w$, the j component of y is $y_j = \sum_{i=1}^{\infty} \Delta_{2^{k-1}, 2^{k-2}, \cdots, 1} P(n + i, j)$.

In order to bound $|A \cdot w|$ it is sufficient to find a bound for $\Delta_{2^{k-1}, 2^{k-2}, \cdots, 1} P(n+i, j)$

Using the mean value theorem from calculus we can find a relation between $\Delta_{c_1, \cdots, c_k} f$ and $\frac{d^k}{dx^k} f$ through the following lemma:

Lemma 32. *Let $f : [0, \infty) \to \mathbb{R}$ a C^∞ function and $c_1, \cdots, c_k$ are given such that $c_i > 0$. Then we have that:*

$$(\Delta_{c_1, \cdots, c_k} f)(x) = (\prod_{i=1}^{k} c_i) \cdot \frac{d^k}{dx^k} f(x + X_x)$$

where $X_x \in [0, \sum_i c_i]$

The last lemma implies that in order to obtain the desired bounds, we would like to obtain some bounds for the absolute value of $\frac{d^k}{di^k}P(i,j)$.

Let's compute the derivatives of $P(i,j)$ respect to i.

$$\frac{d}{di}P(i,j) = \frac{d}{di}[\binom{i-1}{j-1}\frac{1}{2^i}]$$

By applying the product rule:

$$\frac{d}{di}[\binom{i-1}{j-1}\frac{1}{2^i}] =$$
$$[\frac{d}{di}\binom{i-1}{j-1}]\frac{1}{2^i} + \binom{i-1}{j-1}[\frac{d}{di}\frac{1}{2^i}] =$$
$$[\frac{d}{di}\binom{i-1}{j-1}]\frac{1}{2^i} + \binom{i-1}{j-1}\frac{1}{2^i}[\log(\frac{1}{2})]$$

Expanding $\frac{d}{di}\binom{i-1}{j-1} = \frac{d}{di}\frac{(i-j+1)\cdots(i-1)}{(j-1)!}$

Since $(j-1)!$ doesn't depend on i we can take the factor out.

$$\frac{d}{di}\binom{i-1}{j-1} = \frac{1}{(j-1)!}\frac{d}{di}[(i-j+1)\cdots(i-1)]$$

Let $y = (i-j+1)\cdots(i-1)$.

$$\frac{d}{di}\log(y) = \frac{1}{y}\frac{d}{di}y$$

On the other hand $\log(y) = \log(i-j+1) + \cdots + \log(i-1)$

Taking derivatives respect to i we obtain:

$$\frac{d}{di}\log(y) =$$
$$\frac{d}{di}[\log(i-j+1) + \cdots + \log(i-1)] =$$
$$\frac{1}{i-j+1} + \cdots + \frac{1}{i-1}$$

Let

$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ and let $\Psi^{(0)}(x) = \frac{d}{dx}ln(\Gamma(x))$ and $\Psi^{(k)}(x) = \frac{d^k}{dx^k}\Psi^{(0)}(x)$. The above observations yield the following formula for the derivative of $P(i,j)$:

$\frac{d}{di}P(i,j) = P(i,j)g(i,j)$ where

$$g(i,j) = \log(\frac{1}{2}) + \Psi^{(0)}(i) - \Psi^{(0)}(i-j+1)$$

Notice that this implies that:

$$\frac{d^k}{di^k}g(i,j) = \Psi^{(k)}(i) - \Psi^{(k)}(i-j+1)$$

The following lemmas provide some bounds for the value of $\frac{d^k}{di^k}g(i,j)$

Lemma 33. *Let* $|\epsilon|+ < \frac{1}{2} - \frac{2}{i}$. *For* $j = i(1 - \frac{1}{2} - \frac{\epsilon}{2})$ *and* $j \in \mathbb{N}$, *we have that* $|g(i,j)| \leq |2\epsilon| + \frac{4}{i}$.

Proof. We omit the proof. □

Lemma 34. *If* $k \geq 1, i > j > 1$ *we have*$|\frac{d^k}{di^k}g(i,j)| < \frac{4 \cdot k!}{(i-j)^k}$

Proof. We omit the proof. □

We can now use these lemmas to find a bound for $|\frac{d^k}{di^k}P(i,j)|$. The idea for a bound comes from the observation that we can write $|\frac{d^k}{di^k}P(i,j)|$ as a product of $P(i,j)$ and a polynomial of bounded size in derivatives of $g(i,j)$. More specifically:

Lemma 35. *Let* $f(x)$ *be a* C^∞ *function such that* $f(x) > 0$ *and* $f'(x) = f(x)g^{(1)}(x)$. *Then if we define* $g^{(k)}(x) = \frac{d}{dx}g^{(k-1)}(x)$ *we have that for all* $k \geq 1$*:*

$$\frac{d^k}{dx^k}f(x) = f(x)\sum_{\alpha=1}^{s(k)}\prod_{\beta} g^{(p_{\alpha,\beta})}(x)$$

Where we define $p_{\alpha,\beta} \in \mathbb{N}_{>0}, s(k) \leq (k+1)!$ *and* $\sum_\beta p_{\alpha,\beta} = k$ *for all* α. *(this is, for all fixed* α, *the* $p_{\alpha,\beta}$ *form a partition of* k.

Proof. The proof follows just by expansion and induction.

□

From these estimates on g, and its derivatives, we can find an upper bound for $|\frac{d^k}{di^k}P(i,j)|$.

Lemma 36. *Let* $i - j = (1+\epsilon)\frac{1}{2}i$, $i \geq 2$ *and* $j \in \mathbb{N}$. *Then*

$|\frac{d^k}{di^k}P(i,j)| \leq P(i,j)2^{2k\log(k)(1+o(1))}(\max(|\epsilon|, \frac{1}{\sqrt{\frac{1}{2}i}})^k$

Proof. We omit the proof. □

Now we can use this machinery to prove the desired theorem, the failure of the marginals.

Proof. As before, define $v^{(\alpha)}$ vectors of length 2^α, recursively as follows: $v^{(0)} = 1$, $v^{(\alpha)} = (v^{(\alpha-1)}, -v^{(\alpha-1)})$.

Set $w = (0^{n-1}, v^{(k)}, 0^\infty)$ where k will be determined later.

Let $j_- = n(\frac{1}{2}) - \sqrt{kn\log(n)/2}$ and $j_+ = n(\frac{1}{2}) + \sqrt{kn\log(n)/2} + 2^k$

Notice that:

$$\sum_{j=1}^{\infty} |\{Aw\}_j| = \sum_{j<j_-} |\{Aw\}_j| + \sum_{j=j_-}^{j_+} |\{Aw\}_j| + \sum_{j>j_+} |\{Aw\}_j|$$

Now, observe that:

$\sum_{j<j_-} |\{Aw\}_j| \leq 2^k Pr[\text{At least } \sqrt{kn\log(n)/2} \text{ deletions occur in the first } n \text{ bits}]$. By Hoeffdings bound, the last quantity is at most $2^k e^{-k\log(n)/2}$ Analogously one can prove that:

$$\sum_{j>j_+} |\{Aw\}_j| \leq 2^k n^{-k/2}$$

To bound for the remaining term, $\sum_{j=j_-}^{j_+} |\{Aw\}_j|$ works as follows:

First we notice that by the way in which the vector w is defined:

$$(Aw)_j = \Delta_{2^{k-1},\cdots,1}P(n,j)$$

Using the bounding lemmas above, and taking $\epsilon < \sqrt{\frac{k\log(n)}{2n^{\frac{1}{2^2}}}} + \frac{2\cdot 2^k}{n}$.

Then

$$\sum_{j=j_-}^{j_+} |\{Aw\}_j| \leq 2^{k^2}(\sqrt{\frac{k\log(n)}{2n^{\frac{1}{4}}}} + \frac{2\cdot 2^k}{n})^k$$

Now taking $k = \frac{\log(n)}{(\log(\log(n)))^2}$ and for n large enough,

$$2^{k^2}(\sqrt{\frac{k\log(n)}{2n^{\frac{1}{4}}}} + \frac{2\cdot 2^k}{n})^k \leq 2^{k^2}(\sqrt{\frac{k\log(n)}{n^{\frac{1}{4}}}})^k = 2^{-O(k\log(n))}$$

□

3.1.5 The subsequences of subsequences test

In this section we introduce a statistical test that relies on counts of subsequences of subsequences. The main idea of the subsequences of subsequences test is to count for a given 'model' substring s, the frequency with which s appears as a substring of the observed sequences. The quantity we would like to approximate is the expected number of appearances of s as substring of a substring of w.

For a given substring s, we call the associated test, T_s.

Recall that for a given sequence $w \in \{0,1\}^n$, and $s \in \{0,1\}^k$ for some k $C(w,s)$ denotes the number of times s appears as a subsequence of w. The subsequence of subsequences test aims to distinguish two sequences $w, w' \in \{0,1\}^n$ by sampling the number of times a model sequence $s \in \{0,1\}^k$ appears as subsequence of a subsequence of w or w' respectively. The following lemmas exemplify some of the characteristics this test enjoys.

Lemma 37. *Let $w \in \{0,1\}^n$. Let $s \in \{0,1\}^k$ for $k \leq n$. The number of times s appears as a subsequence of a subsequence of w equals $C(w,s) * 2^{n-k}$.*

Proof. For every instance of s as a subsequence of w, (say s'), there are 2^{n-k} sequences of w that contain s'. Since there are $C(w,s)$ instances of s as a subsequence of w,

the total number of times s appears as a subsequence of a subsequence of w equals $C(w,s) * 2^{n-k}$ as desired. □

Let s be a model sequence of $\{0,1\}^k$. Call

$$C'(w,s) = \text{ number of times } s \text{ is a subsequence of a subsequence of } w$$

Let w and w' be the two given sequences as input for the algorithm. The following lemma holds:

Lemma 38. *If* $|C(w,s) - C(w,s')| = O(2^k/n^r)$ *for some universal constant* $r = O(1)$, *then* $|C'(w,s) - C'(w',s)| = O(2^n/n^r)$. *In particular, sampling from the quantities* $C'(w,s)$ *and* $C(w',s)$ *we can get a distinguishing test that works with polynomially many samples and in polynomial time.*

Proof. For a given s, $e_1 = E_w[$ number of times s is a subsequence of a subsequence of $w] = \frac{C'(w,s)}{2^n}$. Analogously, $e_2 = E_{w'}[$ number of times s is a subsequence of a subsequence of $w'] = \frac{C'(w',s)}{2^n}$. If the assumption of the lemma was true, there is a polynomial gap of order $\frac{1}{n^r}$ between e_1 and e_2. Using the Hoeffding bounds we derived in the previous Chapter, we must be able to approximate the actual mean of the hidden sequence's distribution to within an additive error of $\frac{1}{n^r}$ using only $O(n^{2r+1})$ samples. Since the values $C'(w,s)$ and $C'(w',s)$ are easily computable knowing w, w' and s, this yields a distinguishing test that works with polynomially many samples and using polynomial time processing. □

3.1.6 A subsequences of subsequences test

We exhibit a particular array of subsequences of subsequences tests that satisfies various interesting properties and that will be utilized in the Chapter, 'Random Strings are Distinguishable' to prove the distinguishability of random pairs of strings. Consider the following family of sequences:

$$A_1 = \{s_l = 1^l 0\}, B_1 = \{s'_l = 1^{l-1}01\}$$

If $w \in \{0,1\}^n$ and its position-of-ones representation be $(x_1, \cdots, x_m)$ i.e. the positions of the ones in w are $x = x_1, \cdots, x_m$ for some $m \leq n$ then we can count $C(w, s_l)$ for all l as follows:

Lemma 39. $C(w, s_l) = \sum_{i=1}^{m} \binom{i-1}{l-1}(n - x_i)$

Proof. A simple counting argument gives the answer. Each term of the sum counts the number of appearances of s_l such that the one is in position x_i, that is in such a position there are $\binom{i-1}{l-1}$ options to take the initial $l-1$ ones. The $n - x_i - (m - i)$ factor counts the possibilities for the zero.

□

We can count $C(w, s'_l)$ for all l'.

Lemma 40.

$$C(w, s'_l) = \sum_{i=1}^{m}(\binom{i-1}{l-2}(\sum_{j=i}^{m}(x_j - x_i - (j-i))))$$

Proof. A simple counting argument gives us the answer. Each term of the sum counts the number of appearances of s'_l such that the $l-1$ one is in position x_i. That is, in such a position there are $\binom{i-1}{l-2}$ options to take the initial $l-2$ ones. The $\sum_{j=i}^{m}(x_j - x_i - (j-i))$ factor comes from counting the number of 01 substrings on the interval $w[x_i + 1, :]$ where x_j is the position of the last one of the subsequence.

□

The formulas above are linear on $x_1, \cdots, x_m$.

We now show the following lemma:

Lemma 41. *(a) There is an index l for which $|C(w, s_l) - C(w', s_l)| \neq 0$*

(b) There is an index for which $|C(w, s'_l) - C(w', s'_l)| \neq 0$.

Proof. Given $w \in \{0,1\}^n$, let $x \in \mathbb{R}^m$ be a vector such that x_i = position of i one in w. Define x' analogously for w'. The statement holds trivially if the number of ones in w is different from the number of ones in w'. The argument is symmetric for the zeroes. Each of the $C(w, s_i)$ and the $C(w, s_i')$ is a linear function (plus some fixed constant) on the x_i or the x_i'. More formally:

$$\exists \text{ affine funtion } F : \mathbb{R}^m \to \mathbb{R}$$
$$C(w, s_i) = F(x)$$

Their corresponding matrices for the linear part, M and M' are triangular with nonzero diagonal entries and therefore invertible. The last means that $Mx \neq Mx'$ if $x \neq x'$. □

As a corollary of this we have the following

Corollary 42. *If any of the sequences has $O(\log(n))$ ones (or zeros), then they are distinguishable.*

The test that separates them apart is the array of tests described above for the subsequences of subsequences test. If WLOG the number of zeroes is $c\log(n)$, then, the number of samples required will be of the order $O(n^{2c+1})$, which is polynomial in n, if c is constant.

3.2 Special sequence pairs

Here we explore the different properties of a variety of natural sequence pairs. We derive the polynomiality of the variation distance between a series of them. We also explore a variety of tests that for particular pairs of sequences act as distinguishing tests.

3.2.1 Unbalanced sequences

If w and w' have different number of ones (or zeroes) the Number of Ones test differentiates them in polynomial samples and time.

3.2.2 Differing tails

In this section we assume that $p = \frac{1}{2}$. We prove the following result:

Lemma 43. *If $p = \frac{1}{2}$. For pairs of sequences w, w' such that*

1. $w = w_1 01$
2. $w' = w_1 10$

$d_s(w, w')$ is polynomially large.

Proof. Let s be some string. The difference $C(w, s) - C(w', s)$ equals to the instances of the string s appearing as a substring of w using its last $[0, 1]$ symbols plus the instances of the string s appearing as a substring of w' using its last $[1, 0]$ symbols.

This is because all instances of the string s appearing as a substring of w completely contained in the w_1 part of w completely cancel with all the instances of the string s appearing as a substring of w' completely contained in the w_1 part of w'.

Similarly all instances of the string s appearing as a substring of w containing only the last 1 and not the $[0, 1]$ of w completely cancel with all the instances of the string s appearing as a substring of w' containing only the last 1 an not the $[1, 0]$ of w'. The same is true with all the instances of the string s appearing as a substring of w containing only the last 0 an not the $[0, 1]$ of w.

If $n = |w|$, the value we are looking for equals $2^{n-2} * 2 = 2^{n-1}$, which is polynomially big respect to the size of 2^n. □

Evidently, the last implies that for

1. $w = 01 w_1$
2. $w' = 10 w_1$

$d_s(w, w')$ is also polynomially big.

In the following section we prove a stronger result.

3.2.3 Edit distance one are far

We assume that $p = \frac{1}{2}$. In this section we prove the following result:

Lemma 44. *If $p = \frac{1}{2}$. For pairs of sequences w, w' such that*

1. $w = w_1 01 w_2$
2. $w' = w_1 10 w_2$

$d_s(w, w')$ is polynomially big.

Proof. Let P_m be the distribution of the position of the $m - th$ one in the observed sequences.

We will show that for the given sequences there is an m for which the variation distance of the distributions P_m and P'_m is polynomially large. P_m denotes the distribution of the position of the $m - th$ one in the observed sequences coming from w and P'_m the distribution of the position of the $m - th$ one in the observed sequences coming from w'.

Call $P_{m,r}$ the probability of observing a trace from w with its $m-$th one in the $r-$th position. Call $P'_{m,r}$ the probability of observing a trace from w' with its $m-$th one in the $r-$th position.

Say w has k ones in positions $i_1, \cdots, i_k$ respectively. With $i_1 < i_2 < \cdots < i_k$

Let l be the index of the middle one (that which lies after the w_1 and before w_2)

By construction w' has k ones in positions $i'_1, \cdots, i'_k$

with $i'_j = i_j$ for $j \neq l$ and $i'_l = i_l - 1$.

For the case where $p = \frac{1}{2}$

$$P_{m,r} \propto \sum_{j=m}^{k} \binom{j-1}{m-1} \binom{i_j - j}{r-m} 2^{n-i_j} P'_{m,r} \propto \sum_{j-m}^{k} \binom{j-1}{m-1} \binom{i'_j - j}{r-m} 2^{n-i'_j}$$

Where the proportionality denominator is $\frac{1}{2^n}$.

$$|P_{m,r} - P'_{m,r}| = |\binom{l-1}{m-1}2^{n-i_l}\left[\binom{i_l-l}{r-m} - 2\binom{i_l-l-1}{r-m}\right]|$$

By using pascal's identity and cancelling out some terms, we get:

$$|P_{m,r} - P'_{m,r}| = |\binom{l-1}{m-1}2^{n-i_l}\left[\binom{i_l-l-1}{r-m} - \binom{i_l-l-1}{r-m-1}\right]|$$

Notice that by unimodality of the binomial coefficients, if we consider

$$\sum_{r=m}^{i_l+m-l} |P_{m,r} - P'_{m,r}| = \binom{l-1}{m-1}2^{n-i_l}\left(\binom{i_l-l-1}{0} - 0 + \binom{i_l-l-1}{1} - \binom{i_l-l-1}{0} + \cdots\right)$$

The telescopic sum cancels all the terms except the middle ones, where the direction of the subtractions flips. Therefore:

$$\sum_{r=m}^{i_l+m-l} |P_{m,r} - P'_{m,r}| = \binom{l-1}{m-1}2^{n-i_l} * 2\binom{i_l-l-1}{\lfloor\frac{i_l-l-1}{2}\rfloor}$$

Now, notice that if $m-1 = \lfloor\frac{l-1}{2}\rfloor$ then

since $\binom{n}{\lfloor\frac{n}{2}\rfloor} = \theta(\frac{2^n}{\sqrt{n}})$

We get that under these assumptions:

$$\sum_{r=m}^{i_l+m-l} |P_{m,r} - P'_{m,r}| = \theta(\frac{2^n}{\sqrt{l*(i_l-l-1)}})$$

Which implies a polynomial lower bound for the variation distance for the types of sequences considered above. □

3.2.4 Alternating sequences

For the setting where $p = \frac{1}{2}$, we prove that the alternating sequences are far apart. More precisely:

Lemma 45. *If $w = 01\cdots01$ and $w' = 10\cdots10$, then $d_s(w, w')$ is polynomially big.*

For $n = 10$ the sequences are:

1. $w = 0101010101$

2. $w' = 1010101010$

Proof. If $n = |w|$ is odd, both sequences differ in the number of ones they contain, and therefore they are polynomially separated.

Assume n is even.

$w = 01 \cdots 01$ and $w' = 10 \cdots 10$

Notice that we can write w and w' in the following form:

$w = w_1 1$ and $w' = 1 w_1$

All instances of strings that are subsequences of w that are completely contained within w_1 cancel out with all those instances of strings that are subsequences of w' and are completely contained within w_1.

Among those instances of strings that are subsequences of w that have the last one of w and start with a zero cannot cancel out with any of the sequences of w' that have the first one of w' and end with a zero.

The remaining strings, are those that are subsequences of w that have the last one an start in a one. And those strings that are subsequences of w' that have the first one and end in a one.

These can be paired to each other. Every instance of such a sequence in w can be translated (shifted to the left) to correspond with an instance in w'. Since this operation is invertible, it gives us a bijection between them, and therefore, they do not add to the count of the variation distance.

The variation distance is therefore, the sum of the number A_1 of instances of strings that are subsequences of w that have the last one of w and start with a zero plus A_2 the number of subsequences of w' that have the first one of w' and end with a zero.

$w = 01 \cdots 01$ and $w' = 10 \cdots 10$

Let $n = 2k$.

A simple counting argument yields:

All zeroes of w are in positions $2j-1$ for $j = 1, \cdots, k$.

The number of symbols between a zero in position $2j-1$ and the last one of w is $2k-1-(2j-1) = 2k-2j$. Therefore the value of A_1 is:

$A_1 = \sum_{j=1}^{k} 2^{2k-2j} = \sum_{j=1}^{k} 4^{k-j} = \frac{4^{k+1}-1}{3} = \frac{2^{n+2}-1}{3}$

Analogously

$A_2 = \frac{2^{n+2}-1}{3}$

And therefore $A_1 + A_2 = \frac{2^{n+3}-2}{3}$ which is polynomially big in the size of 2^n. □

3.2.5 Cyclic shifted strings

We see that there is a test that distinguishes cyclic shifted strings. Let the two strings be the pair (w, w') where w' is a cyclic shift of w. WLOG we assume that the cyclic shift is to the right.

Let $s = 01$. The following holds:

Lemma 46. $C(w, s) \neq C(w', s)$

Proof. Let $x_1, \cdots, x_m$ be the positions of the ones in w. We assume the shift from w to w' is to the right.

There are two cases to analyze:

1. If $x_m = n$
2. $x_m \neq n$

In the first case, the first digit of w' is a one. $C(w, s) = \sum_{i=1}^{m} x_i - i$. Since there are $x_i - i$ instances of s ending in the $i-$th one. Because we are in the first case, the positions of the ones in w' are $1, x_1 - 1, \cdots x_{m-1} - 1$. The last implies that $C(w', s) = 1-1+\sum_{i=1}^{m-1} x_i - 1 - (i+1)$. Which reduces to $C(w', s) = \sum_{i=1}^{m-1} x_i - (i+2)$ which is clearly less than $C(w, s)$ if $m \neq 0$.

In the second case, there is no wrap around in w' and the positions of the ones in w' are, $x_1 + 1, \cdots, x_m + 1$. The value of $C(w', s) = \sum_{i=1}^{m} x_i + 1 - i$ which is clearly bigger than $C(w, s)$ since each term dominates the corresponding one of the other summation. □

This result implies that the test T_s succeeds, since $|s|$ is of constant size. The cyclic shifted strings are distinguishable.

Chapter 4

Random strings are distinguishable

In this Chapter we prove that there exists some universal constant $c \in [0,1)$ and a statistical test $\mathcal{T}$ such that the fraction of pairs the test $\mathcal{T}$ fails to distinguish over the space of pairs $\{0,1\}^n \times \{0,1\}^n$ is asymptotically of order less than c^n. In other words, given a uniform random pair of sequences, the probability that $\mathcal{T}$ fails for the array of sequences $\{s_{k'}\}$ is exponentially small.

Consider the 0−vector representation of a sequence w with m ones. Recall that the 0−vector of a sequence $w \in \{0,1\}^n$ is $(b_0, \cdots, b_m)$. Where b_i indicates the number of zeroes before the $i+1$-th and after the ith one. By convention, b_0 is the number of zeroes before the first one, b_m is the number of zeroes after the last one. Let $s_k = 1^k 0$ be the sequence that starts with k zeroes and finishes with a 1.

Lemma 47. *If w can be represented as $(b_0, \cdots, b_m)$ then for all k we have that $C(w, s_k) = \sum_{i=0}^{m} \binom{i}{k} \cdot b_i$.*

Proof. Every instance of s_k appearing as a subsequence of w is specified by the position of its ones within the sequence w. The last zero of s_k may lie in any of the regions, $b_0, \cdots, b_m$. The number of instances of s_k appearing as subsequences of w and having its last zero within b_i is:

$$\binom{i}{k} \cdot b_i$$

Summing over all the possible values of $i \in \{0, \cdots, m\}$ yields the desired result. □

When comparing two sequences w and w' and their subsequences counts for the family of sequences $\{s_k\}_{k=0}^m$, the previous lemma yields the following necessary condition for w and w' to have different counts for s_k.

Corollary 48. *If w can be represented as $(b_0, \cdots, b_m)$ and w' can be represented as $(b'_0, \cdots, b'_m)$ then $C(w, s_k) \neq C(w', s_k)$ if and only if $\sum_{i=0}^m \binom{i}{k} \cdot b_i \neq \sum_{i=0}^m \binom{i}{k} \cdot b'_i$*

Let k be the $\min_{k' \in \{0,\cdots,m\}}$ such that $C(w, s_{k'}) \neq C(w', s_{k'})$.

Corollary 49. *k as defined above if and only if $\sum_{i=0}^m \binom{i}{k'} \cdot b_i \neq \sum_{i=0}^m \binom{i}{k'} \cdot b'_i$ for all $k' \leq k-1$.*

Corollary 50. *If $|C(w, s_k) - C(w', s_k)| = O(2^k/n^r)$ for some absolute constant r, then $|C'(w, s_k) - C'(w', s_k)| = O(2^n/n^r)$. In particular, sampling from the quantities $C'(w, s_k)$ and $C(w', s_k)$ we can get a distinguishing test that works with polynomially many samples and in polynomial time. Additionally, this implies the variation distance between the distributions of subsequences, the one associated with w and the one associated with w' are polynomially far.*

In the following notes of this section we explore explore under what series of conditions under which the corollary above holds. The main conjecture of this section is the following:

Conjecture 51. *Let $k = \min_{k'} s.t. C(w, s_{k'}) \neq C(w', s_{k'})$ for $s_{k'} = 1^{k'}0$. $|C(w, s_k) - C(w', s_k)| = O(2^k/n^c)$ for some universal constant c.*

If the conjecture above were true, it would imply the following:

(a) The variation distance between the distribution induced by the deletion process with parameter $p = \frac{1}{2}$ [We believe this can be generalized to any constant] between the sequence w and the sequence w' is polynomially large.

(b) There exists a test that allows to differentiate samples from the distribution induced by w and the distribution induced by w' using polynomially many samples and polynomial post processing time.

As we defined before, let $k = \min_{k'} \text{s.t.} C(w, s_{k'}) \neq C(w', s_{k'})$ if and only if: $\sum_{i=0}^{m} \binom{i}{k'} \cdot b_i = \sum_{i=0}^{m} \binom{i}{k'} \cdot b'_i$ for all $k' \leq k-1$ and $\sum_{i=0}^{m} \binom{i}{k} \cdot b_i \neq \sum_{i=0}^{m} \binom{i}{k} \cdot b'_i$.

Consider the polynomial $p_w(x) = \sum_{i=0}^{m} b_i x^i$ and $p_{w'}(x) = \sum_{i=0}^{m} b'_i x^i$ Through the following lemmas we will relate the polynomials $p_w(x), p_{w'}(x)$ with the first index k for which the counts of $C(w, s_k)$ and $C(w', s_k)$ differ.

Lemma 52. *$(1-x)^{k-1} | p_w(x) - p_{w'}(x)$ but $(1-x)^k$ doesn't divide $f(x) = p_w(x) - p_{w'}(x)$ If and only if $\frac{\partial^{k'} f}{\partial x^{k'}}$ is divisible by $(1-x)$ for all $k' \leq k-2$.*

Proof. This follows from simple properties of a polynomial and its derivatives. In particular, the fact that $(x-a)^r$ divides $p(x)$ if and only if $\frac{\partial^{r-1} f(x)}{\partial x^{r-1}}$ is divisible by $x-a$ but $\frac{\partial^r f(x)}{\partial x^r}$ is not. □

Lemma 53. *Keeping the notation above, assuming that $k = \min_{k'} s.t. C(w, s_{k'}) \neq C(w', s_{k'})$, let $f(x) = (1-x)^{k-1} g(x)$, where $g(1) \neq 0$. Then $g(1) = \sum_{i=0}^{m} \binom{i}{k} \cdot b_i - \sum_{i=0}^{m} \binom{i}{k} \cdot b'_i$.*

Proof. The result follows from noting that by L'Hopital's rule: $\lim_{x \to 1} f(x)/(1-x)^{k-1} = \frac{\partial f}{\partial x} / \frac{\partial (1-x)^{k-1}}{\partial x} = \cdots = \frac{\partial^{k-1} f}{\partial x^{k-1}} / \frac{\partial^{k-1} (1-x)^{k-1}}{\partial x^{k-1}}$. It follows that: $\frac{\partial^{k-1} f}{\partial x^{k-1}} / \frac{\partial^{k-1} (1-x)^{k-1}}{\partial x^{k-1}} = \sum_{i=0}^{m} \binom{i}{k} \cdot b_i - \sum_{i=0}^{m} \binom{i}{k} \cdot b'_i$

Since $\lim_{x \to 1} f(x)/(1-x)^{k-1} = \lim_{x \to 1} g(x) = g(1)$ the result follows. □

The last few lemmas yield the following corollary:

Corollary 54. *$\sum_{i=0}^{m} \binom{i}{k'} \cdot b_i = \sum_{i=0}^{m} \binom{i}{k'} \cdot b'_i$ for all $k' \leq k$ and $\sum_{i=0}^{m} \binom{i}{k} \cdot b_i \neq \sum_{i=0}^{m} \binom{i}{k} \cdot b'_i$. Is equivalent to: $(1-x)^k | p_w(x) - p_{w'}(x)$ but $(1-x)^{k+1}$ doesn't divide $p_w(x) - p_{w'}(x)$.*

Conditions under which the subsequences of subsequences test yields a good result.

Let $|w| = |w'| = n$ and let m = number of ones in w. In general, if $n \geq \beta \cdot 2^{\alpha k}$ for some $\alpha, \beta > 0$, then $k \leq \frac{\log(n) + \log(\beta)}{\alpha}$ which implies that $2^k \leq \beta^{\frac{1}{\alpha}} \cdot n^{\frac{1}{\alpha}}$. Since $\beta^{\frac{1}{\alpha}}$, the last implies that $2^{n-k} = O(\frac{2^n}{n^{\frac{1}{\alpha}}})$. In fact, it implies that $2^{n-k} = \Theta(\frac{2^n}{n^{\frac{1}{\alpha}}})$.

From lemma above it follows that for this particular choice of α if there is a sequence s such that $|s| = k$ and $C(w,s) \neq C(w,s')$ then, the subsequences of subsequences test using model sequence s serves to differentiate the distributions of w and w' using only polynomially many samples.

In particular if $n = \Theta(2^{\alpha k})$ where k is the first $k = \min_{k'} \text{s.t.} C(w, s_{k'}) \neq C(w', s_{k'})$. For some universal constant α then, the subsequences of subsequences test with model sequence s_k will be able to distinguish between the to distributions of subsequences drawn from w or w' using only polynomially many samples and polynomial processing.

We now proceed to characterize some classes of sequences for which there is a distinguishing test based on a subsequences of subsequences test for the family $\{s_i\}$.

Let $k = \min_{k'} \text{s.t.} C(w, s'_k) \neq C(w', s'_k)$

Lemma 55. *If $g(-1) \neq 0$ then, $|w| = n \geq 2^{k-1}$. In particular this implies that $n = O(2^k)$.*

Proof. Let $f(x)$ be defined as before, and $f(x) = (1-x)^{k-1} g(x)$ where $g(1) \neq 0$. We can assume this is the case by invoking the use of lemma 53. Notice that for every $\omega \in \mathbb{C}$ such that $|\omega| = 1$, we have that, by the definition of $f(x)$ and a simple use of the triangle inequality:

$$|f(\omega)| \leq 2(n-m)$$

Where n is the size of the sequences and m is the number of ones in both w and w'. In particular, since we WLOG can assume that $m \geq \lfloor \frac{n}{2} \rfloor$. $|f(\omega)| \leq n$. We prove that if $g(-1) \neq 0$ then $|f(-1)| \geq 2^{k-1}$.

$|f(-1)| = |(1+1)^{k-1}||g(-1)|$. If $g(-1) \neq 0$, since g is an integer polynomial, $|g(-1)| \geq 1$. This means that $|f(-1)| \geq 2^{k-1}$. Hence, $n \geq |f(-1)| \geq 2^{k1}$ and therefore $|w| \geq 2^{k-1}$. □

Corollary 56. *If -1 is not a root of $p_w(x)p_{w'}(x)$, then, there is a distinguishing test that works with polynomially many samples and uses polynomial time. Such a test is the subsequences of subsequences test for the first subsequence in $\{s_k\}$, $s_{k'}$ for which*

$C(w, s_{k'}) \neq C(w', s_{k'})$. *The* α *exponent of this test is* 1. *For those pairs of sequences* (w, w') *such that* $f(-1) \neq 0$, *the gap is of order* $\Omega(\frac{1}{n})$. *Following our discussion on Hoeffding bounds, the algorithm will require* $O(n^{2*2+1}) = O(n^5)$ *samples to achieve a* $1/n^2$ *factor accuracy.*

Translating the above corollary to the $(b_0, \cdots, b_m)$ and $(b'_1, \cdots, b'_m)$ representation of w and w', we can establish the following result:

Corollary 57. *If* w *and* w' *are two sequences from* $\{0,1\}^n$ *with representations* $(b_0, \cdots, b_m)$ *and* $(b'_1, \cdots, b'_m)$, *then there is a test distinguishing them if the sum of the* b_i *in the even positions doesn't match the sum of the* b'_i *in the even positions.*

Proof. By assumption $\sum_{i=0}^{m} b_i = \sum_{i=0}^{m} b'_i$. By the corollary, a distinguishing test exists if $f(-1) \neq 0$. The condition $f(-1) = 0$ implies:

$$p_w(-1) = p_{w'}(-1)$$

Evaluating the polynomial $p_w(x) = \sum_{i=0}^{m} b_i x^i$ yields, $p_w(-1) = \sum_{i=0}^{m} b_i(-1)^i$. The equality $p_w(-1) = p_{w'}(-1)$ implies that:

$$\sum_{i=0}^{m} b_i(-1)^i = \sum_{i=0}^{m} b'_i(-1)^i$$

This equation along with the identity $\sum_{i=0}^{m} b_i = \sum_{i=0}^{m} b'_i$ yields the desired result.

□

It is possible to replicate a similar result by considering other roots of unity, distinct from -1.

Lemma 58. *If* $g(i) \neq 0$ *then,* $|w| = n \geq 2^{\frac{k-1}{2}}$. *For* $i = \sqrt{-1}$. *In particular,* $n = O(2^{\frac{k}{2}})$

Proof. Let $f(x)$ be defined as before, and $f(x) = (1-x)^{k-1}g(x)$ where $g(1) \neq 0$. If i is not a root of $g(x)$ then $|f(i)| = |(1-i)^{k-1}g(i)| = |(1-i)|^{k-1}|g(i)|$. Notice that

because $g(x)$ has integer coefficients, $g(i)$ is a member of the integer lattice generated over $\mathbb{C}$ by i and 1. The smallest nonzero vector in this lattice has norm 1. Since $|1-i| = \sqrt{2}$, this implies that $|f(i)| \geq 2^{\frac{k-1}{2}}$. Hence $n \geq |f(i)| \geq 2^{\frac{k-1}{2}}$. □

Corollary 59. *If i is not a root of $p_w(x) - p_{w'}(x)$, then, there is a distinguishing test that works with polynomially many samples and uses polynomial time. Such a test is the subsequences of subsequences test for the first subsequence in $\{s_k\}$ for which its counts first differ in w and w'. The α exponent of this test is $\frac{1}{2}$. For those pairs of sequences (w, w') such that $f(i) \neq 0$, the gap is of order $\Omega(\frac{1}{n^2})$. Because $f(x)$ has integer coefficients, $f(i) = 0$ implies that $f(-i) = 0$ as well. The test will need $O(n^{2*3+1}) = O(n^7)$ samples to achieve a $1/n^3$ factor accuracy.*

We can translate the above corollary to the $(b_0, \cdots, b_m)$ and $(b'_1, \cdots, b'_m)$ representation of w and w', we can establish the following result:

Corollary 60. *If w and w' are two sequences from $\{0,1\}^n$ with representations $(b_0, \cdots, b_m)$ and $(b'_1, \cdots, b'_m)$, then there is a test distinguishing them if the sum of the b_i in positions congruent to a modulo 4 doesn't match the sum of the b'_i in positions congruent to a modulo 4 for all a in the system of residues modulo 4.*

Proof. By the corollary 57 we can assume

$$\sum_{i \in [0,m] i \equiv 0 \bmod 2} b_i = \sum_{i \in [0,m] i \equiv 0 \bmod 2} b'_i$$
$$\sum_{i \in [0,m] i \equiv 1 \bmod 2} b_i = \sum_{i \in [0,m] i \equiv 1 \bmod 2} b'_i$$

Evaluating the polynomial $p_w(x) = \sum_{i=0}^{m} b_i x^i$ yields, $p_w(i) = \sum_{i=0}^{m} b_i (i)^i$. The equality $p_w(i) = p_{w'}(i)$ implies that:

$$\sum_{i=0}^{m} b_i (i)^i = \sum_{i=0}^{m} b'_i (i)^i$$

Equating the coefficients of i in the left hand side and the coefficients of i in the right hand side and combining the resulting identities with the result from lemma [x] yields the desired result.

$$\sum_{i\in[0,m] i\equiv 0 \bmod 4} b_i = \sum_{i\in[0,m] i\equiv 0 \bmod 4} b'_i$$
$$\sum_{i\in[0,m] i\equiv 1 \bmod 4} b_i = \sum_{i\in[0,m] i\equiv 1 \bmod 4} b'_i$$
$$\sum_{i\in[0,m] i\equiv 2 \bmod 4} b_i = \sum_{i\in[0,m] i\equiv 2 \bmod 4} b'_i$$
$$\sum_{i\in[0,m] i\equiv 3 \bmod 4} b_i = \sum_{i\in[0,m] i\equiv 3 \bmod 4} b'_i$$

□

It is possible to replicate a similar result by considering other roots of unity, distinct from -1 and i.

Lemma 61. *If $g(\omega) \neq 0$ then, $|w| \geq 2^{\frac{\log(3)}{2}(k-1)}$. For ω = third root of unity.*

Proof. Let $f(x)$ be defined as before, and $f(x) = (1-x)^{k-1}g(x)$ where $g(1) \neq 0$. If ω is not a root of $g(x)$ then $|f(\omega)| = |(1-\omega)^{k-1}g(\omega)| = |(1-\omega)|^{k-1}|g(\omega)|$. Notice that because $g(x)$ has integer coefficients, $g(\omega)$ is a member of the integer lattice generated over $\mathbb{C}$ by ω and 1. The smallest nonzero vector in this lattice has norm 1. Since $|1-\omega| = \sqrt{3}$, this implies that $|f(\omega)| \geq 2^{\frac{\log(3)}{2}(k-1)}$. Hence $n \geq |f(\omega)| \geq 2^{\frac{\log(3)}{2}(k-1)}$. □

Corollary 62. *If ω is not a root of $p_w(x)p_{w'}(x)$, then, there is a distinguishing test that works with polynomially many samples and uses polynomial time. Such a test is the subsequences of subsequences test for the first subsequence in $\{s_k\}$ for which its counts first differ in w and w'. The α exponent of this test is $\frac{\log(3)}{2}$. For those pairs of sequences (w, w') such that $f(\omega) \neq 0$, the gap is of order $\Omega(\frac{1}{n^{\frac{2}{\log(3)}}})$. Hence in order to approximate it up to an accuracy of $1/n^{\frac{2}{\log(3)}+1}$ we need $O(n^{(2*\frac{2}{\log(3)}+1)+1}) = O(n^{(\frac{4}{\log(3)}+2)})$ samples. Because $f(x)$ has integer coefficients, $f(\omega) = 0$ implies that $f(\bar{\omega}) = 0$ as well.*

We can translate the above corollary to the $(b_0, \cdots, b_m)$ and $(b'_1, \cdots, b'_m)$ representation of w and w', and establish the following result:

Corollary 63. *If w and w' are two sequences from $\{0,1\}^n$ with representations $(b_0, \cdots, b_m)$ and $(b'_1, \cdots, b'_m)$, then there is a test distinguishing them if the sum of the b_i in positions congruent to a modulo 3 doesn't match the sum of the b'_i in positions congruent to a modulo 3 for all a in the system of residues modulo 3.*

Proof. Let

$$A = \sum_{i \in [0,m] i \equiv 0 \bmod 3} b_i$$

$$A' = \sum_{i \in [0,m] i \equiv 0 \bmod 3} b'_i$$

$$B = \sum_{i \in [0,m] i \equiv 1 \bmod 3} b_i$$

$$B' = \sum_{i \in [0,m] i \equiv 1 \bmod 3} b'_i$$

$$C = \sum_{i \in [0,m] i \equiv -1 \bmod 3} b_i$$

$$C' = \sum_{i \in [0,m] i \equiv -1 \bmod 3} b'_i$$

Evaluating the polynomial $p_w(x) = \sum_{i=0}^{m} b_i x^i$ yields, $p_w(\omega) = \sum_{i=0}^{m} b_i (\omega)^i$. The equality $p_w(\omega) = p_{w'}(\omega)$ implies that:

$$\sum_{i=0}^{m} b_i (\omega)^i = \sum_{i=0}^{m} b'_i (\omega)^i$$

Along with the condition $\sum_{i=0}^{m} b_i = \sum_{i=0}^{m} b'_i$, equating the coefficients of i in the left hand side and the coefficients of i in the right hand side and equating the real coefficients in the right hand side and the real coefficients in the left hand side yields the following system of equations:

$$A + B + C = A' + B' + C'$$
$$A - B = A' - B'$$
$$A - C = A' - C'$$
$$B - C = B' - C'$$

After the appropriate elimination of variables the following equations hold:

$$A = A'$$
$$B = B'$$
$$C = C'$$

Therefore

$$\sum_{i\in[0,m]i\equiv 0 \bmod 3} b_i = \sum_{i\in[0,m]i\equiv 0 \bmod 3} b'_i$$
$$\sum_{i\in[0,m]i\equiv 1 \bmod 3} b_i = \sum_{i\in[0,m]i\equiv 1 \bmod 3} b'_i$$
$$\sum_{i\in[0,m]i\equiv -1 \bmod 3} b_i = \sum_{i\in[m]i\equiv -1 \bmod 3} b'_i$$

□

Remark 64. *We can establish a final result on this vein by considering the 6th root of unity.*

Unfortunately, these results do not extend to all the roots of unity. This is because the integral lattice generated by the powers of a root of unity other than the $2nd$, $3rd$, $4th$ or $6th$ does not have a minimal element. We have no hope to be able to bound the modulus of $g(\omega_n)$ below, where ω_n is any other root of unity.

As a consequence of these observations, the subsequences of subsequences test

for $k = \min_k \text{s.t.} C(w, s_k) \neq C(w', s_k)$ works with high probability after polynomially many samples and polynomial time if any of the following conditions holds:

$$\sum_{i\in[0,m] i\equiv 0 \bmod 2} b_i \neq \sum_{i\in[0,m] i\equiv 0 \bmod 2} b'_i$$

$$\sum_{i\in[0,m] i\equiv 1 \bmod 2} b_i \neq \sum_{i\in[0,m] i\equiv 1 \bmod 2} b'_i$$

$$\sum_{i\in[0,m] i\equiv 0 \bmod 3} b_i \neq \sum_{i\in[0,m] i\equiv 0 \bmod 3} b'_i$$

$$\sum_{i\in[0,m] i\equiv 1 \bmod 3} b_i \neq \sum_{i\in[0,m] i\equiv 1 \bmod 3} b'_i$$

$$\sum_{i\in[0,m] i\equiv -1 \bmod 3} b_i \neq \sum_{i\in[0,m] i\equiv -1 \bmod 3} b'_i$$

$$\sum_{i\in[0,m] i\equiv 0 \bmod 4} b_i \neq \sum_{i\in[0,m] i\equiv 0 \bmod 4} b'_i$$

$$\sum_{i\in[0,m] i\equiv 1 \bmod 4} b_i \neq \sum_{i\in[0,m] i\equiv 1 \bmod 4} b'_i$$

$$\sum_{i\in[0,m] i\equiv 2 \bmod 4} b_i \neq \sum_{i\in[0,m] i\equiv 2 \bmod 4} b'_i$$

$$\sum_{i\in[0,m] i\equiv 3 \bmod 4} b_i \neq \sum_{i\in[0,m] i\equiv 3 \bmod 4} b'_i$$

Remark 65. *These restrictions hold in both for the* $0-$ *vectors of* w, w' *and the* $1-$*vector of* w, w'*. By considering these extra restrictions it would be possible to (at least) refine the constants in the exponent of current upper bound on the probability of failure.*

The restriction of the sum over the even positions and etc needs to hold both ways, for the ones and for the zeroes. This is, for the $0-$vector and the $1-$vector.

In the following we analyze a model whereby a random pair of strings $(w, w') \in \{0,1\}^n \times \{0,1\}^n$ is drawn uniformly at random and, we are asked if we can solve the Pair Trace Identification Problem on the pair. We show that with exponentially small probability the array of subsequence of subsequences tests exhibited in this section

is unable to distinguish between w and w'. First we show that within the space of sequences (w, w') where w and w' have both the same number of ones, the fraction of pairs that do not pass the subsequences of subsequences test described above is vanishingly small.

We focus on the following conditions. There is a distinguishing test for w and w' if

$$\sum_{i\in[0,m] i\equiv 0 \bmod 2} b_i \neq \sum_{i\in[0,m] i\equiv 0 \bmod 2} b'_i$$
$$\sum_{i\in[0,m] i\equiv 1 \bmod 2} b_i \neq \sum_{i\in[0,m] i\equiv 1 \bmod 2} b'_i$$

Let $\mathcal{M} = (w, w') \in \{0,1\}^n \times \{0,1\}^n$ such that w and w' have the same number of ones.

Lemma 66.

$$|\mathcal{M}| = \sum_{i=0}^{n} \binom{n}{i}^2$$

Proof. The number of sequences $w \in \{0,1\}^n$ such that w has exactly m ones is $\binom{n}{m}$. The number of pairs (w, w') with $w, w' \in \{0,1\}^n$ is and exactly m ones is therefore $\binom{n}{m}^2$. □

We count the number of pairs (w, w') within $\mathcal{M}$ such that if w and w' have both m ones exactly, and their representations are $(b_0, \cdots, b_m)$ and $(b'_0, \cdots, b'_m)$, then $\sum_{i\in[0,m] i\equiv 0 \bmod 2} b_i \neq \sum_{i\in[0,m] i\equiv 0 \bmod 2} b'_i$.

Lemma 67. *The number of pairs $w, w' \in \{0,1\}^n$ such that w and w' have both exactly m ones equals:*

$$\sum_{\textit{target sum } t=0}^{n-m} \binom{t + \lfloor \frac{m}{2} \rfloor}{\lfloor \frac{m}{2} \rfloor}^2$$

Where the target sum is that over the even positions of $(b_0, \cdots, b_m)$ *and* $(b'_0, \cdots, b'_m)$.

Proof. The lemma follows immediately from the basic counting fact that the number of ordered tuples $(a_1, \cdots, a_r)$ of nonnegative numbers summing to a target value t equals:

$$\binom{t+r-1}{r-1}$$

Applying this fact to each possible value of the target sum yields the desired result. □

The fraction of pairs that are not distinguishable by the subsequences of subsequences test over pairs of strings (w, w') such that w and w' have both exactly m ones is at most:

$$D_m = \frac{\sum_{\text{target sum } t=0}^{n-m} \binom{t+\lfloor \frac{m}{2} \rfloor}{\lfloor \frac{m}{2} \rfloor}^2}{\binom{n}{m}^2}$$

We will try to upper bound D_m for $m \in [n/4, 3n/4]$. Let $r(t) = \binom{t+\lfloor \frac{m}{2} \rfloor}{\lfloor \frac{m}{2} \rfloor}^2$. Since $r(t)$ is an increasing function of t,

$$\sum_{\text{target sum } t=0}^{n-m} \binom{t+\lfloor \frac{m}{2} \rfloor}{\lfloor \frac{m}{2} \rfloor}^2 \leq (n-m)\binom{n-m+\lfloor \frac{m}{2} \rfloor}{\lfloor \frac{m}{2} \rfloor}^2$$

This estimate provides us with the following upper bound for D_m.

$$D_m \leq \frac{(n-m)\binom{n-m+\lfloor \frac{m}{2} \rfloor}{\lfloor \frac{m}{2} \rfloor}^2}{\binom{n}{m}^2}$$

After expanding out these expressions:

$$D_m \leq (n-m)\left(\frac{m(m-1)\cdots(\lfloor \frac{m}{2} \rfloor + 1)}{n(n-1)\cdots(n-\lfloor \frac{m}{2} \rfloor + 1)}\right)^2$$

Lemma 68. *If $m \in [n/4, \cdots, 3n, 4]$. Then,*

$$D_m \leq n\left(\frac{3}{4}\right)^{\frac{n}{4}}$$

Proof.

$$\begin{aligned}
D_m &\leq (n-m)\left(\frac{m(m-1)\cdots(\lfloor\frac{m}{2}\rfloor+1)}{n(n-1)\cdots(n-\lfloor\frac{m}{2}\rfloor+1)}\right)^2 \\
&\leq n\left(\frac{m(m-1)\cdots(\lfloor\frac{m}{2}\rfloor+1)}{n(n-1)\cdots(n-\lfloor\frac{m}{2}\rfloor+1)}\right)^2 \\
&= n\left(\frac{m}{n}\cdot\frac{m-1}{n-1}\cdots\frac{\lfloor\frac{m}{2}\rfloor+1}{n-\lfloor\frac{m}{2}\rfloor+1}\right)^2 \\
&= n\left(\frac{3}{4}\cdot\frac{3}{4}\cdots\frac{3}{4}\right)^2 \\
&= n\left(\frac{3}{4}\right)^m \\
&\leq n\left(\frac{3}{4}\right)^{\frac{n}{4}}
\end{aligned}$$

Where the last inequalities follow from the condition $n \in [n/4, 3n/4]$. □

To aid in our proof, we prove a concentration inequality for the case squares of the binomial coefficients.

Lemma 69. *Let $0 < \epsilon < \frac{1}{4}$ then: Let A, B be the following sets of indices:*

$$\begin{aligned}
A &= [\lfloor(\frac{1}{2}-\epsilon)n\rfloor, \lceil(\frac{1}{2}+\epsilon)n\rceil] \\
B &= [0, \lfloor(\frac{1}{2}-\epsilon)n\rfloor] \cup [\lceil(\frac{1}{2}+\epsilon)n\rceil, n]
\end{aligned}$$

The following inequality holds:

$$\sum_{i\in B}\binom{n}{i}^2 \leq \lambda(\epsilon)\sum_{i\in A}\binom{n}{i}^2$$

Where $\lambda = \frac{\frac{1}{\epsilon}e^{-\frac{\epsilon^2 n}{2}}}{1-\frac{1}{\epsilon}e^{-\frac{\epsilon^2 n}{2}}}$

Proof. Notice that $\max_{i\in B}\binom{n}{i} \leq \min_{i\in A}\binom{n}{i}$.

$$\begin{aligned}
\sum_{i\in B}\binom{n}{i} &\leq \left(\sum_{i\in B}\binom{n}{i}\right)\max_{i\in B}\binom{n}{i} \\
&\leq \lambda\left(\sum_{i\in A}\binom{n}{i}^2\right)\max_{i\in B}\binom{n}{i} \\
&\leq \lambda\left(\sum_{i\in A}\binom{n}{i}^2\right)\min_{i\in A}\binom{n}{i} \\
&= \lambda\sum_{i\in A}\binom{n}{i}^2
\end{aligned}$$

As desired. □

$$D = \frac{\sum_{m=0}^{n}\sum_{\text{target sum } t=0}^{n-m}\binom{t+\lfloor\frac{m}{2}\rfloor}{\lfloor\frac{m}{2}\rfloor}^2}{\sum_{m=0}^{n}\binom{n}{m}^2}$$

Using the same argument to upper bound D_m we get the following bound for D:

$$\begin{aligned}
D &\leq \frac{\sum_{m=0}^{n}(n-m)\binom{n-m+\lfloor\frac{m}{2}\rfloor}{\lfloor\frac{m}{2}\rfloor}^2}{\sum_{m=0}^{n}\binom{n}{m}^2} \\
&\leq n\left(\frac{\sum_{m=0}^{n}\binom{n-m+\lfloor\frac{m}{2}\rfloor}{\lfloor\frac{m}{2}\rfloor}^2}{\sum_{m=0}^{n}\binom{n}{m}^2}\right) \\
&= n\left(\frac{\sum_{m\in A}\binom{n-m+\lfloor\frac{m}{2}\rfloor}{\lfloor\frac{m}{2}\rfloor}^2 + \sum_{m\in B}\binom{n-m+\lfloor\frac{m}{2}\rfloor}{\lfloor\frac{m}{2}\rfloor}^2}{\sum_{m=0}^{n}\binom{n}{m}^2}\right) \\
&\leq \sum_{m\in A} D_m + n\left(\frac{\sum_{m\in B}\binom{n-m+\lfloor\frac{m}{2}\rfloor}{\lfloor\frac{m}{2}\rfloor}^2}{\sum_{m=0}^{n}\binom{n}{m}^2}\right)
\end{aligned}$$

The first summand can be bounded above by a simple use of the Lemma 68, and the second lemma can be bounded above by a simple use of Lemma 69 with $\epsilon = \frac{1}{4}$. Adding the two bounds yields:

$$D \leq \frac{n}{2}\left(n\frac{3}{4}\right)^{\frac{n}{4}} + n\lambda$$

Where $\lambda = \frac{4e^{-\frac{n}{32}}}{1-4e^{-\frac{n}{32}}}$

This implies that for some universal constant $c \in [0,1)$ the fraction of pairs for which the subsequences of subsequences test over the space of pairs $\mathcal{M}$ fails is asymptotically of order less than c^n. In other words, given a uniform random pair of sequences drawn from $\mathcal{M}$, the probability the subsequences of subsequences test for the array of sequences $\{s_{k'}\}$ fails is exponentially small.

If w and w' are both from $\{0,1\}^n$ but $(w,w') \notin \mathcal{M}$, then, the test, counting the number of ones in both sequences is a distinguishing test that works with polynomially many samples and polynomial processing.

A more exact statement can be enunciated by first noticing that a simple combinatorial identity yields the following asymptotic estimate $|\mathcal{M}| = \binom{2n}{n}$, and therefore $\frac{|\mathcal{M}|}{2^{2n}} \equiv o(\frac{1}{\sqrt{n}})$. Therefore,

Theorem 70. *There exists some universal constant $c \in [0,1)$ such that the fraction of pairs for which the subsequences of subsequences test over the space of pairs $\{0,1\}^n \times \{0,1\}^n$ fails is asymptotically of order less than c^n. In other words, given a uniform random pair of sequences, the probability the subsequences of subsequences test fails for the array of sequences $\{s_{k'}\}$ is exponentially small.*

4.0.6 Future directions

In this section we discuss how the techniques introduced in this section could lead to the solution of the Pair Trace Identification Problem. The following conjecture, if proven right would imply the polynomial distinguishability of all pairs of distinct sequences w, w' and therefore a full solution of the Pair Trace Identification Problem.

Open Problem 71. *For every $w, w' \in \{0,1\}^n$, there exists a $k \leq n$ and a subsequence $s \in \{0,1\}^k$ such that $|C(w,s) - C(w',s)| = O(\frac{2^k}{n^c})$ for some universal constant c.*

A more detailed condition, whereby the desired string specified in 71 is conjectured to come from a polynomially large family of strings is the following:

Open Problem 72. *Let* $\{s_i\}_{i=0}^m = \{0^i1\}_{i=0}^m$ *and* $\{s'_i\}_{i=0}^m = \{1^i0\}_{i=0}^m$. *If*

$$k = \max_{w \neq w' | w,w' \in \{0,1\}^n} \min_i s.t. C(w, s_i) \neq C(w', s_i)$$

Define k' *analogously for the family* $\{s'_k\}$ *Then,* $|C(w, s_k) - C(w', s_k)| = O(\frac{2^k}{n^c})$ *or* $|C(w, s'_{k'}) - C(w', s'_{k'})| = O(\frac{2^{k'}}{n^c})$

The proof exhibited above does not work for the full case because of the crystallographic restriction. For any given primitive root ω_q, and polynomial $g \in \mathbb{Z}[x]$, there is no simple lower bound for $|g(\omega_q)|$, even if $g(\omega_q) \neq 0$, if ω_q is not a second, third, fourth, or sixth root of unity.

4.0.7 The marginals test

The marginals test is related to the array of subsequences of subsequences tests introduced in this section.

Recall that for a pair of sequences $w, w' \in \{0,1\}^n$, each with exactly m ones and with 0−vectors $(b_1, \cdots, b_m)$ and $(b'_1, \cdots, b'_m)$ we define the polynomials:

$$p_w(x) = \sum_{i=0}^{m} b_i x^i$$
$$p_{w'}(x) = \sum_{i=0}^{m} b'_i x^i$$

And the polynomial $f(x) = p_w(x) - p_{w'}(x) = \sum_{i=0}^m (b_i - b'_i)x^i$.

Recall that in the proof of the failure of the marginals test that the construction of the vectors that achieve subpolynomiallity in the l_1 distance of their summary statistics goes as follows:

$$v^{(0)} = 1, v^{(i)} = (v^{(i-1)}, -v^{(i-1)})$$

Let $f_{v^{(i)}}$ be the polynomial corresponding to the difference vector $v^{(i)}$. It is easy to see that $f_{v^{(1)}}(x) = -(x-1)$. Furthermore, a simple computation yields the following result:

Observation 73.

$$f_{v^{(k)}}(x) = (x^{2^{k-1}} - 1) f_{v^{(k-2)}}(x)$$

Furthermore if $v = (0^m, v^{(k)})$ *is a difference vector, then* $f_v(x) = f_{v^{(k)}}(x)$

We can solve the recurrence on $f_{v^{(k)}}(x)$ explicitly:

Observation 74.

$$f_{v^{(k)}}(x) = -(x-1) \prod_{i=0}^{k-2} (x^{2^i} - 1)$$

Notice that $(x-1)^k | f_{v^{(k)}}(x)$ but $(x-1)^{k+1}$ does not divide $f_{v^{(k)}}(x)$. And that $(x+1)^{k-2} | f_{v^{(k)}}(x)$.

Notice that even though the difference vector $v^{(k)}$ makes the marginals test fail, it does not make the subsequences of subsequences test fail. This is simply because, if $v = (0^m, v^{(k)})$ where $v = w - w'$, where w and w' as seen as vectors, then the observation in the previous paragraph this means that $2^{k-2} || C(w, s_k) - C(w', s_k)|$, which implies that it does not violate the conjecture stated in 71.

In the next chapter we will explore another way in which the marginals test is related to the results and techniques exhibited in this section.

Chapter 5

Extensions

In this Chapter we explore several problems related to the Pair Trace Reconstruction. In particular, we talk about its relation with the $k-$Trace Reconstruction Problem and with the full Trace Reconstruction. We show how the techniques and methods developed in this thesis can help gain some new insights in both problems.

5.1 A new distance

In this section we will define a new distance between sequences, and explain how could this be used to generate a statistical distinguishing test between any two strings $w, w' \in \{0,1\}^n$.

Definition 75. *Let w, w' be two sequences in $\{0,1\}$. Let $d^1(w,w')$ be the minimum length of a sequence s for which $C(w,s) \neq C(w',s)$.*

If w and w' have distinct quantities of ones, (or zeros) then $d^1(w,w') = 1$. Notice also that $d^1(011,110) = 2$.

Definition 76. *The $k-$deck of a sequence $w \in \{0,1\}^n$ is the multiset of all subsequences of size k of w.*

We prove first if $w \neq w'$, $d^1(w,w') - 1$ is the last size for which the $k-$decks of w and w' differ. In other words:

Lemma 77. *If two sequences w, w' have the same $k-$deck, then they also have the same $k'-$deck for all $k' \in \{1, \cdots, k\}$.*

Proof. A given subsequence s of w of size $k-i$ will appear exactly:

$$\binom{n-(k-i)}{i} C(w,s)$$

times in the k subsequences of w. If w and w' had the same $k-$deck, we should have that $C(w,s) = C(w',s)$, and therefore they will also have the same $k-i$ deck. □

Let $k_0(n) = \max_{w,w' \in \{0,1\}^n} d^1(w,w')$. We would like to understand the asymptotic behavior of $k_0(n)$.

There exists a strong relationship between the asymptotic behavior of k_0, the search for a bound on the variation distance, subsequence count polynomials and the subsequences of subsequences test.

By definition $k_0(n)$ is the smallest value such that $\mathcal{F}^{(m)}_{k_0(n)}$ for all values of m forms a basis of all the m variables multivariate polynomials of degree up to $k_0(n)$.

In particular, let the subsequence of subsequences test be ran on all sequences s such that $|s| \leq k_0(n)$.

In particular, if the following conjecture was true, then the procedure of running the subsequences of subsequences test for all subsequences of size up to $k_0(n)$ would yield a polynomial samples, polynomial post processing time, procedure to solve the Pair Trace Identification Problem, proving that the variation distance between $\mathcal{P}_{w',\frac{1}{2}}$ and $\mathcal{P}_{w,\frac{1}{2}}$ for every pair $w, w' \in \{0,1\}^n$ is polynomially large.

Conjecture 78. *The max of $d^1(w,w')$ over all pairs $w, w' \in \{0,1\}^n$ such that $w \neq w'$ is of order $O(\log(n))$.*

Unfortunately, this conjecture is false, as we will see in the next section. Nevertheless, it is easy to prove that:

Lemma 79. ?? $d^1(w,w') \leq \lfloor \frac{n}{2} \rfloor + 1$

Proof. Let a sequence $w \in \{0,1\}^n$ have m ones and a 0−vector representation $(b_0, \cdots, b_m)$. Let $s_{i,m-i} = 1^i 01^{m-i}$. Then $\forall i \in \{0, \cdots, m\}, C(w, s_{i,m-i}) = b_i$. We define $s_{0,m} = 01^m$ and $s_{m,0} = 1^m 0$. Notice that $|s_{i,m-i}| = m + 1$. Knowing $C(w, s_{i,m-1})$ for all i, we can know the 0−vector representation of w, and therefore the array of values $C(w, s_{i,m-i})$ is unique to every w. A symmetric argument can be made for the 1−vector representation of w. This means that $d^1(w, w') \leq (\max_m \min_{m,n-m}) + 1 = \lfloor \frac{n}{2} \rfloor + 1$ as desired.

□

Lemma 80. *If w and w' end in the same symbol. WLOG say $w = w_1 0$ and $w' = w'_1 0$ Then $d^1(w, w') = d^1(w_1, w'_1)$.*

Proof. Let s be such that the number of occurrences of s in w and w' are the same. And such that for all s' with $|s'| < |s|$, $C(w, s') = C(w', s')$.

If the last symbol of s is different from the last symbol of w:

$$C(w, s) = C(w_1, s)$$

and

$$C(w', s) = C(w'_1, s)$$

This is because s when considered as a substring of w or w' cannot possibly use the very last one in either string. In this case $C(w_1, s) = C(w'_1, s)$ are the same.

If the last symbol of s is the same as the last symbol of w. Let wlog $s = s_1 0$

Then

$$C(w, s) = C(w_1, s) + C(w_1, s_1)$$

This formula reads, the number of appearances of s in w equals the number of appearances of s where s doesn't contain the last 0 of w plus the number of appearances of s where it contains the last 0 of w.

Analogously,

$$C(w', s) = C(w_1', s) + C(w_1', s_1)$$

If s_1 doesn't finish in a zero, if $s_1 = s_2 1$, then $C(w_1, s_1) = C(w, s_1)$. Analogously $C(w_1', s_1) = C(w', s_1)$. If s_1 finishes in a zero, $s_1 = s_2 0$, then

Notice that

$$C(w, s_1) = C(w_1, s_1) + C(w_1, s_2)$$

and

$$C(w', s_1) = C(w_1', s_1) + C(w_1', s_2)$$

These imply that:

$$C(w_1, s_1) = C(w, s_1) - C(w_1, s_2)$$

and

$$C(w_1', s_1) = C(w', s_1) - C(w_1', s_2)$$

By assumption, since for all s'' with $|s''| < |s|$, $C(w, s'') = C(w', s'')$ and $|s_1| < s$ we have that:

$C(w_1, s_1) = C(w_1', s_1)$ as long as $C(w_1, s_2) = C(w_1', s_2)$

Applying the same argument over and over again, until we reach the first one in s. Gluing these observations together we get that, if $s = s_k 10^k$

$$C(w_1, s) = C(w_1', s)$$

Provided that

$C(w_1, s_k1) = C(w'_1, s_k1)$ which is true because $|s_k1| < |s|$ which implies that

$$C(w, s_k1) = C(w', s_k1)$$

And because w, w' both end in 0, $C(w_1, s_k1) = C(w, s_k1)$ and $C(w'_1, s_k1) = C(w', s_k1)$.

Which implies that $C(w_1, s_k1) = C(w'_1, s_k1)$, and therefore that $C(w_1, s) = C(w'_1, s)$, as desired.

The last implies that if s is such that $C(w, s) = C(w', s)$ and for all s' with $|s'| < |s|$, $C(w, s') = C(w', s)$ then, $C(w_1, s) = C(w'_1, s)$. Which means that $d^1(w, w') \leq d^1(w_1, w'_1)$.

The inequality $d^1(w_1, w'_1) \leq d^1(w, w')$ is much simpler. If s is such that $C(w_1, s) = C(w'_1, s)$ and such that $\forall s'$ with $|s'| < |s|$, $C(w_1, s') = C(w'_1, s')$ then, because $C(w, s) = C(w_1, s) + C(w_1, s_1)$, and $C(w', s) = C(w'_1, s) + C(w'_1, s_1)$.

If $C(w_1, s) = C(w'_1, s)$ then, it implies that $C(w_1, s_1) = C(w'_1, s_1)$, which in turn implies that $C(w, s) = C(w', s)$. This chain of arguments imply that $d^1(w_1, w'_1) \leq d^1(w, w')$. Combining the two inequalities yields:

$$d^1(w_1, w'_1) = d^1(w, w')$$

□

Lemma 81. *If w and w' are two sequences such that there is no intermediate index i such that $w[:i]$ and $w'[:i]$ have the exact same number of symbols (Except for the extremes.) Then the number of appearances of* 01 *in w and the number of appearances of* 01 *in w' differ.*

Proof. Let $x_1, \cdots, x_m$ the positions of the ones in w and let $y_1, \cdots, y_m$ be the positions of the ones in w'.

The condition imposed on w and w' is equivalent to:

$$x_i > y_i \forall i$$

The number of instances of 01 in w is:

$$A == \sum_{i=1}^{m} x_i - i$$

The number of instances of 01 in w' is:

$$B = \sum_{i=1}^{m} y_i - i$$

Since $x_i > y_i \ \forall i$ then

$A - B > 0$. By observing there is a difference of one per term subtraction, it can be proven that:

$$A - B \geq m$$

□

Lemma 82. *Let w, w' be two strings of length n such that the number of ones in $w[:i]$ equals the number of ones in $w'[:i]$. If $d^1(w[:i], w'[:i]) \neq d^1(w[i+1:], w'[i+1:])$ then $d^1(w, w') = \min(d^1(w[:i], w'[:i]), d^1(w[i+1:], w'[i+1:]))$.*

Proof. WLOG let the left hand pair $w[:i]$, $w'[:i]$. be such that $d^1(w[:i], w'[:i]) < d^1(w[i+1:], w'[i+1:])$, and let s be the first string such that $C(w[:i], s) \neq C(w'[:i], s)$.

By assumption $|s| = d^1(w[:i], w'[:i])$, and $C(w[i+1:], s') = C(w'[i+1:], s')$, for $s' = s$ and for all s' with $|s'| < |s|$. Furthermore, $C(w[:i], s') = C(w'[:i], s')$ for all s' with $|s'| < |s|$.

Notice that

$$C(w, s) = \sum_{j=1}^{|s|-1} C(w[:i], s[:j]) * C(w[i+1:], s[j+1:]) + C(w[:i], s) + C(w[i+1:], s)$$

and

$$C(w', s) = \sum_{j=1}^{|s|-1} C(w'[:i], s[:j]) * C(w'[i+1:], s[j+1:]) + C(w'[:i], s) + C(w'[i+1:], s)$$

Since all of the corresponding summands are the same, except for $C(w[i+1:], s)$ and $C(w'[i+1:], s)$, the two quantities differ. This implies directly that $d^1(w, w') \leq \min(d^1(w[:i], w'[:i]), d^1(w[i+1:], w'[i+1:]))$.

A similar argument tells us that for all s_0 such that $|s_0| < |s|$, since

$$C(w, s_0) = \sum_{j=1}^{|s_0|-1} C(w[:i], s_0[:j]) * C(w[i+1:], s_0[j+1:]) + C(w[:i], s_0) + C(w[i+1:], s_0)$$

and

$$C(w', s_0) = \sum_{j=1}^{|s_0|-1} C(w'[:i], s[:j]) * C(w'[i+1:], s_0[j+1:]) + C(w'[:i], s_0) + C(w'[i+1:], s_0)$$

and all of the corresponding RHS terms are the same, then $C(w, s_0) = C(w', s_0)$ which implies that $d^1(w, w') \geq \min(d^1(w[:i], w'[:i]), d^1(w[i+1:], w'[i+1:]))$. Combining both results, yields the desired result.

□

Let $k_0 = \min_{w \neq w' \in \{0,1\}^n} d^1(w, w')$

Lemma 83. *Let $w, w' \in \{0,1\}$ be such that $d^1(w, w') \geq 1$ then, $d^1(ww', w'w) \geq d^1(w, w') + 1$*

Proof. Let $w^{(1)} = ww'$ and $w^{(2)} = w'w$. Suppose $s \trianglelefteq w$ and $|s| = d^1(w, w')$. Notice that $C(w^{(1)}, s) = C(w^{(2)}, s)$. This is because all instances of s as subsequence of $w^{(1)}$ is either fully contained in w, w' or split between w and w'. That is, $C(w^{(1)}, s) = C(w, s) + C(w, s') + \sum_{i=1}^{|s|-1} C(w, s[:i]) \cdot C(w', s[i+1:|s|])$. Since $|s[:i]|, |s[i+1:]| < |s| = d^1(w, w')$ this means that $C(w, s[:i]) = C(w', s[:i])$ and $C(w, s[i+1:]) = C(w', s[i+1:])$. Therefore $C(w^{(1)}, s) = C(w, s) + C(w, s') + \sum_{i=1}^{|s|-1} C(w, s[:$

$i]) \cdot C(w', s[i+1:|s|]) = C(w,s) + C(w,s') + \sum_{i=1}^{|s|-1} C(w', s[:i]) \cdot C(w, s[i+1:|s|]) =$ $C(w'w, s)$, implying that $d^1(w^{(1)}, w^{(2)}) \geq d^1(w, w') + 1$. □

5.2 Reconstruction of sequences

The previous discussion on the subsequences of subsequences test is closely related to the following trace reconstruction problem:

Suppose we are given a multiset of sequences of length k. Does that multiset come from some sequence of length n? If so, is the source sequence unique?

This problem is very much reminiscent of other reconstruction problems such as the the problem of reconstructing graphs from vertex deleted subgraphs. [INSERT REFERENCE]

In this section we consider the problem of reconstructing an sequence of length n from the multiset of its subsequences of size k. As it was mentioned in the introduction we refer to this problem as the $k-$trace reconstruction problem:

Problem 84. *What is the minimum k such that all sequences of $\{0,1\}^n$ are $k-$reconstructible.*

In other words, we are interested in understanding the asymptotics for the following quantity.

$$k = \max_{w \neq w' | w, w' \in \{0,1\}^n} \min_{k'} \text{s.t.} C(w, s_{k'}) \neq C(w', s_{k'})$$

The best bounds known to date for the value of k are:

$$c \log(n)^2 \leq k \leq c' \sqrt(n)$$

For some universal constants $c, c' \in \mathbb{R}^+$. The upper bound is proven in [8], the best constant c' known is $\lfloor \frac{16}{7} \sqrt{n} \rfloor + 4$. The proof for the lower bound $c \log(n)$ is constructive and can be found in [7]. In this section we show a simplified version of

the proof for the lower bound, and an alternative version which relies on the methods introduced in the previous chapter.

Loose bounds

For the lower bound it is easy to show $k = O(\log(n))$. The proof follows by construction see [7]:

Let $w_0 = 0, w_0' = 1$. And define w_n and w_n' recursively by:

$$w_{n+1} = w_n w_n'$$
$$w_{n+1}' = w_n' w_n$$

The validity of the construction follows from the lemma 83.

A.D. Scott in his paper 'Reconstructing sequences' proves a lower bound of the order $(1 + o(1))\sqrt{n \log(n)}$ [11]

Subsequence count polynomials

Recall that if $w \in \{0,1\}^n$ and $x_1, \cdots, x_m$ are the positions of the ones of w. For a fixed m, n, $C(w,s) = f_s(x_1, \cdots, x_m)$. Let $F_k^{(m)} = \{f_s(x_1, \cdots, x_m)\}_{|s| \leq k}$.

A consequence of the result by Shulman see [7] it follows immediately that the system of polynomials made up of all subsequence count polynomials for subsequences of length up to $c\log(n)$ for any $c \in \mathbb{R}^+$ cannot be a basis for the space of all polynomials. More formally:

Lemma 85. *For every n, there is an m such that the span of $\mathcal{F}^{(\updownarrow)}{}_r$ for $r = O(log(n))$ is never the whole space of multivariate polynomials of m variables.*

Moreover, there exists a fundamental connection between finding the asymptotic behavior of k and problem 13:

Open Problem 86. *What is the span of $\mathcal{F}_r^{(m)}$ over the space of multivariate m variate polynomials?*

The span of $\mathcal{F}_r^{(m)}$ equals the space of multivariate polynomials of m variables if $r \geq k$.

The upper bound

First we show a simple proof of the upper bound for $k_0(n)$. We show that the following bound holds:

$$k_0(n) \leq \lfloor \frac{16}{7}\sqrt{n} \rfloor + 5$$

The proof we present here is a simplified section of the one presented by Krasikov, Roditty in [8] and referenced by Schulman in [7].

Identify a sequence $w \in \{0,1\}^n$ with a vector in $\mathbb{R}^n$ whose entries are only 0 or 1.

For a sequence $w \in \{0,1\}^n$ and any given k and for all values of $i \in \{1, \cdots, k\}$, let $f_i^{(k)}(w)$ be the number of times the $i-$th entry of a subsequence of size k of w equals 1.

Remark 87.

$$f_i^{(k)}(w) = \sum_{j=1}^{n} \binom{j}{i-1} \binom{n-j}{k-i} w_j$$

In particular $f_k^{(k)}(w) = \sum_{j=1}^{n} \binom{j}{k-1} w_j$ Consider the array of values $\{f_k^{(k)}(w)\}_{k=1}^{n}$ associated with w and $\{f_k^{(k)}(w')\}$ with w'. It follows that if w and w' have the same $k-$deck for a given k' then $f_k^{(k)}(w) = f_k^{(k)}(w')$ for all $k \in \{1, \cdots, k'\}$. This follows from the definition of $f_k^{(k)}$ and from **??**.

Let $k_0 = \min_k$ for which $f_k^{(k)}(w) \neq f_k^{(k)}(w')$, then

$$\sum_{j=1}^{n} \binom{j}{k-1} w_j = \sum_{j=1}^{n} \binom{j}{k-1} w_j' \forall k \leq k_0 - 1$$

The following theorem from [9] will be auxiliary to our results.

Theorem 88. *Every polynomial p of the form*

$$p(x) = \sum_{j=0}^{n} a_j x^j, |a_0| = 1, |a_j| \leq 1, a_j \in \mathbb{C}$$

has at most $\lfloor \frac{16}{7}\sqrt{n} \rfloor + 4$ *zeroes at* 1.

Let $p_w(x)^{(1)} = \sum_{i=1}^{n} x^{i-1} w_i$ and $p_{w'}^{(1)}(x) = \sum_{i=1}^{n} x^{i-1} w_i$

Following similar ideas to the ones exhibited by P. Borwein and C. Ingalls in their survey paper "The Prouhet-Tarry-Escott problem revisited", [10], we can establish the following lemma:

Lemma 89. *The following conditions are equivalent:*

(i) $\sum_{j=1}^{n} \binom{j}{k-1} w_j = \sum_{j=1}^{n} \binom{j}{k-1} w'_j \forall k \leq k_0 - 1$ *and* $\sum_{j=1}^{n} j^{k_0-1} w_j \neq \sum_{j=1}^{n} j^{k_0-1} w'_j$

(ii) $(x-1)^k | p_w^{(1)}(x) - p_{w'}^{(1)}(x)$ *and* $(x-1)^{k_0}$ *does not divide* $p_w^{(1)}(x) - p_{w'}^{(1)}(x)$

Proof. The proof is the same process by which the Corollary 54 is produced. □

Let $f^{(1)}(x) = p_w^{(1)}(x) - p_{w'}^{(1)}(x)$. Let x^α be the first power of x appearing with a nonzero coefficient in $f^{(1)}(x)$. Define $f_1^{(1)}(x) = \frac{f^{(1)}}{x^\alpha}$. Notice that all the nonzero coefficients of $f_1^{(1)}$ have modulus 1, and that the degree of $f_1^{(1)}$ is at most $n-1$. Applying the result of Theorem 88 to bound the multiplicity of $x-1$ as a root of $f_1^{(1)}$, we obtain the following corollary:

Corollary 90.

$$k_0 - 1 \leq \lfloor \frac{16}{7}\sqrt{n-1} \rfloor + 4$$

or more succinctly:

$$k_0 \leq \lfloor \frac{16}{7}\sqrt{n-1} \rfloor + 5$$

Since this bound is independent of the underlying pair of strings w, w' and only depends on their length, it must be the case that:

$$k_0(n) \leq \lfloor \frac{16}{7}\sqrt{n-1} \rfloor + 5$$

The condition $\sum_{j=1}^{n} \binom{j}{k-1} w_j = \sum_{j=1}^{n} \binom{j}{k-1} w'_j \forall k \leq k_0 - 1$ implies, by taking linear combinations of the coefficients, that:

$$\sum_{j=1}^{n} j^{k-1} w_j = \sum_{j=1}^{n} j^{k-1} w'_j \forall k \leq k_0 - 1$$

The later condition is closely related with the number theoretic problem known as the Prouhet-Tarry-Escott problem.

The Prouhet-Tarry-Escott problem asks for finding tuples of integers $u_1, \cdots, u_s$ and $v_1, \cdots, v_s$ such that:

$$u_1^h + u_2^h + \cdots + u_s^h = v_1^h + v_2^h + \cdots v_s^h h = 1, \cdots, k_0 - 1$$
$$1 \leq u_1 < u_2 \cdots < u_s \leq n, 1 \leq v_1 < v_2 < \cdots < u_{\leq} n$$

It is known that $k_0 \leq \lfloor \frac{16}{7}\sqrt{n} \rfloor + 4$. A simple application of this fact to the problem at hand yields the desired bound for $k_0(n)$.

An alternative proof for the upper bound

In what follows we derive an alternative proof for the upper bound for $k_0(n)$ that hinges upon the $0-$vector representation that lead to the proof of distinguishability of random sequences in Chapter 4.

Let $w \in \{0,1\}^n$ a sequence with m ones. Let w be represented, as in Chapter 4, by the vector $(b_0, \cdots, b_m)$. Where b_i indicates the number of zeroes before the $i+1$-th and after the ith one. As before, b_0 is the number of zeroes before the first one, b_m is the number of zeroes after the last one. Let $s_k = 1^k 0$ be the sequence that starts

with k zeroes and finishes with a 1. Recall that:

$$C(w, s_k) = \sum_{i=0}^{m} \binom{i}{k} \cdot b_i$$

Recall that if w can be represented as $(b_0, \cdots, b_m)$ and w' can be represented as $(b'_0, \cdots, b'_m)$, then

$$C(w, s_k) \neq C(w', s_k) \text{ iff}$$
$$\sum_{i=0}^{m} \binom{i}{k} \cdot b_i \neq \sum_{i=0}^{m} \binom{i}{k} \cdot b'_i$$

Let $k_0 = \min_{k' \in \{0,\cdots,m\}}$ such that $C(w, s_{k'}) \neq C(w', s_{k'})$. The following condition holds:

Lemma 91. *Let k_0 defined as above, then:*

$$\sum_{i=0}^{m} i^r \cdot b_i = \sum_{i=0}^{m} i^r \cdot b'_i$$

for all $r \in \{0, \cdots, k_0 - 1\}$.

This condition follows immediately from noticing that the equations $\sum_{i=0}^{m} i^r \cdot b_i = \sum_{i=0}^{m} i^r \cdot b'_i$ can be obtained as a linear combination of the conditions: $\sum_{i=0}^{m} \binom{i}{k} \cdot b_i \neq \sum_{i=0}^{m} \binom{i}{k} \cdot b'_i$ for all $i \leq k_0 - 1$.

Following the same notation as in Chapter 4, let $w, w' \in \{0,1\}^n$, define $p_w(x) = \sum_{i=0}^{m} b_i x^i$ and $p_{w'}(x) = \sum_{i=0}^{m} b'_i x^i$ and $f(x) = p_w(x) - p_{w'}(x)$.

We cam establishe the following corollary, a rewriting of the conditions in Corollary 54:

Corollary 92. *(i) $C(w, s_{k'}) = C(w', s_{k'})$ for $k' \in \{0, \cdots, k_0 - 1\}$ but $C(w, s_{k_0}) \neq C(w', s_{k_0})$*

(ii) $\sum_{i=0}^{m} \binom{i}{k'} \cdot b_i = \sum_{i=0}^{m} \binom{i}{k'} \cdot b'_i$ for all $k' \leq k_0 - 1$ and $\sum_{i=0}^{m} \binom{i}{k_0} \cdot b_i = \sum_{i=0}^{m} \binom{i}{k_0} \cdot b'_i$.

(iii) $(1-x)^{k_0} | p_w(x) - p_{w'}(x)$ *but* $(1-x)^{k_0+1}$ *doesn't divide* $p_w(x) - p_{w'}(x)$.

The three conditions are equivalent.

The following theorem from [9] will be auxiliary to our results.

Theorem 93. *There is an absolute constant* $c > 0$ *such that every polynomial* p *of the form:*

$$p(x) = \sum_{j=0}^{n} a_j x^j, |a_j| \leq 1, a_j \in \mathbb{C}$$

has at most $c(n(1 - \log(|a_0|)))^{\frac{1}{2}}$ *zeroes at* 1.

Let x^α be the first power of x in $f(x)$ with a nonzero coefficient, and define $f_1(x) = \frac{1}{nx^\alpha} f(x)$. All the coefficients of $f_1(x)$ have modulus less than 1, hence we can apply to $f_1(x)$ the Theorem 93. Let a_0 be the first nonzero coefficient of $f_1(x)$. Since $|a_0| \geq \frac{1}{n}$ this implies that $-\log(|a_0|) \leq \log(n)$. Therefore,

$$k_0 \leq c(n(1 + \log(n)))^{\frac{1}{2}} = O(\sqrt{n})$$

We cannot use the same Theorem 88 in this case because $|a_0|$ need not equal 1.

Because by considering either the 0−vector or the 1−vector representation of w, w' we can assume the degree of $f_1(x)$ be at most $\lfloor \frac{n}{2} \rfloor$. This gives us an improved bound of

$$k_0 \leq c(n/2(1 + \log(n/2)))^{\frac{1}{2}} = O(\sqrt{n})$$

Which implies:

$$k_0(n) \leq c(n/2(1 + \log(n/2)))^{\frac{1}{2}} = O(\sqrt{n})$$

This bound is better than the previous one, provided $c < \frac{16}{7}$. It might be possible to obtain a better bound for $k_0(n)$ than the existing value, $\lfloor \frac{16}{7}\sqrt{n} \rfloor + 5$ via sharpening

the constant c. We leave that as an open problem for the reader.

Number theoretic connections

The results of the previous section imply that given two sequences $w, w' \in \{0,1\}^n$, the multiplicity of $(x-1)$ as a root of $f(x) = p_w(x) - p_{w'}(x)$ is less than or equal to $\lfloor \frac{16}{7}\sqrt{n} \rfloor + 4$.

This prompts the following result:

Theorem 94. *Let $f(x) \in \mathbb{Z}[x]$, an n degree polynomial with integer coefficients. If $A_1 = \sum_{i=0}^{n} a_i 1_{a_i \geq 0}$ and $A_2 = \sum_{i=0}^{n} -a_i 1_{a_i \leq 0}$. Then the multiplicity of $(x-1)$ as a root of $f(x)$, $k_0 - 1$ follows:*

$$k_0 \leq \lfloor \frac{16}{7}\sqrt{n + \max A_1, A_2} \rfloor + 5$$

Proof. Let $f(x) = \sum_{i=0}^{n} a_i x^i$. Let $b = (b_0, \cdots, b_n), b' = (b'_0, \cdots, b'_n)$ be two vectors defined as:

$$b_i = 1_{a_i \geq 0} a_i$$
$$b'_i = -1_{a_i \leq 0} a_i$$

As above, let $A_1 = \sum_{i=0}^{n} a_i 1_{a_i \geq 0}$ and $A_2 = \sum_{i=0}^{n} -a_i 1_{a_i \leq 0}$. Let $w \in \{0,1\}^{n+A_1}$ and $w' \in \{0,1\}^{n+A_2}$ be the sequences defined by the $0-$vectors b and b' respectively. Notice that $f(x) = p_w(x) - p_{w'}(x)$ which, together with the previous remarks, completes the proof. □

Notice that the coefficients of $f(x)$ are integers and follow it is realizable by two sequences w, w' of size n such as

Reconstruction algorithm

Consider the array of subsequences of subsequences tests which uses the model sequences $\{s_i\}_{i=0}^{m}$.

The Trace Reconstruction Problem asks to devise an algorithm whereby after receiving a series of subsequences $t_1, \cdots, t_r$ from a hidden random sequence $w \in \{0,1\}$. The sequence w is assumed to be drawn from the uniform distribution. The sampling distribution over subsequences equals $\mathcal{P}_{w,p_n}(\cdot)$ for some probability parameter $p_n \in (0,1)$.

Current developments towards the solution of the Trace Reconstruction Problem include a reconstruction algorithm developed by Thomas Holenstein, Michael Mitzenmacher, Rina Panigrahy, Udi Wieder, in their paper Trace reconstruction with constant deletion probability and related results [2] which proposes an efficient algorithm that, whenever p_n is a constant smaller than $\frac{1}{3}$, only fails to uniquely reconstruct an exponentially small fraction of the sequences of $\{0,1\}^n$.

In this thesis we have focused on the case where the deletion probability is a constant p. We proposed an efficient algorithm that, whenever $p = \frac{1}{2}$, fails to distinguish between w, w' drawn uniformly at random from $\{0,1\}^n$ with only an exponentially small probability.

The subsequences of subsequences test described in the previous chapter provides a natural way to approach the reconstruction problem. Let $\{s_i\}_{i=0}^{n-1}$ be the family of sequences $s_i = 1^i 0$

The proposed algorithm takes the following form:

Algorithm 95.

Draw polynomially many sample traces. In order to obtain a estimate of the expected number of appearances that s_i for all i has as a subsequence of a sequence of the hidden string. Let the estimates be $\{\bar{E}_i\}_{i=0}^{n-1}$

Loop through all sequences $w \in \{0,1\}^n$, compute the real expected values $\{\bar{E}_i(w)\}_{i=0}^{n-1}$ for each s_i and output the sequence $w \in \{0,1\}$ such that $\max_{i \in \{0,\cdots,n-1\}} |\bar{E}_i - \bar{E}_i(w)|$ is minimal.

Unfortunately this algorithm although utilizes polynomially many samples, takes an exponential amount of time.

Conjecture 96. *The algorithm predicts the structure of the hidden sequence with high probability.*

An alternative version of the algorithm 97 is the following:

Algorithm 97.

Draw polynomially many sample traces. In order to obtain a estimate of the expected number of appearances that s_i for all i has as a subsequence of a sequence of the hidden string. Let the estimates be $\{\bar{E}_i\}_{i=0}^{n-1}$. Let $e = [\bar{E}_0, \cdots, \bar{E}_{n-1}]$ be a vector in $\mathbb{R}^n$.

Notice that if $w \in \{0,1\}^n$ and its $0-$vector is of the form $b = (b_0, \cdots, b_m)$ then there exists an invertible operator A such that $Ab = e$. Try $A^{-1}e$ and recover the closest possible values of b in the resulting vector.

This algorithm requires only a polynomial number of samples and it runs in polynomial time. Nevertheless, it requires that the array of subsequences of subsequences test $\{s_i\}_{i=0}^m$ be always able to distinguish a pair of sequences $w, w' \in \{0,1\}^n$.

5.3 Searching for the minimising pair

In this section we explore a lower bound for the pair of strings $w, w' \in \{0,1\}^n$ minimising the variation distance between $\mathcal{P}_{w,p}$ and $\mathcal{P}_{w',p}$ where $p = \frac{1}{2}$. We conjecture the asymptotics of the minimal distance and the structure of the pairs of strings that achieve it. The content of this section is speculative. The results and conjectures here exhibited have been produced by a series of computer simulations, the code for which is available through the author's github account, Pacchiano.

Cyclic shift

As a result from the simulations we have the following conjecture.

Conjecture 98. *For a given string w, the string w' that minimizes the variation distance $d_s(w, w')$ is such that w' is a cyclic shift of w.*

The minimizing pair

As a result from the simulations we also have the following conjecture. Consider the following binary strings:

1. $w = 0^n 10^{n-1}$
2. $w' = 0^{n-1}10^n$

Lemma 99. *The distance* $d_s(w, w') = \binom{2n}{n}$

Proof. Notice that all subsequences s of w that do not contain the middle 1 'cancel' with corresponding subsequences of w' that do not contain the middle 1. Among the remaining sequences we have the following relations:

$$C(w, 0^i10^j) = \binom{n}{i} \cdot \binom{n-1}{j}$$
$$C(w', 0^i10^j) = \binom{n-1}{i} \cdot \binom{n}{j}$$

Notice that:

$$d_s(w, w') = \sum_{i,j} |C(w, 0^i10^j) - C(w', 0^i10^j)|$$

It can be shown that

$$\begin{aligned} |C(w, 0^i10^j) - C(w', 0^i10^j)| &= |\binom{n}{i} \cdot \binom{n-1}{j} - \binom{n-1}{i} \cdot \binom{n}{j}| \\ &= \frac{1}{n}\binom{n}{i}\binom{n}{j}|i-j| \end{aligned}$$

This means that :

$$d_s(w, w') = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\binom{n}{i}\binom{n}{j}|i-j|$$

It can be shown that this sum equals $\binom{2n}{n}$. □

The following is the crux conjecture of this section:

Conjecture 100. $\min_{w,w' \in \{0,1\}^{2n}} d_s(w,w')$ *is exactly* $\binom{2n}{n}$ *Furthermore: It is achieved only for the pairs of strings:* $w = 0^n 10^{n-1}$ $w' = 0^{n-1}10^n$ *and* $w = 1^n 01^{n-1}$ $w' = 1^{n-1}01^n$

In the following we exhibit the possible structure of a proof of the result above, with a series of steps completed and some others pending for completion.

Proof. We divide the proof into two sections.

1. First we see that for a given string w, by a result hinted above, $d_s(w,w')$ is minimal for w' a cyclic shift of w.

2. Second we conjecture that among all pairs of sequences $d_s(w,w')$ such that w' is a cyclic shift of w, the pair that minimizes $d_s(w,w')$ is precisely $(w,w') = (0^n10^{n-1}, 0^{n-1}10^n)$ or $(w,w') = (0^{n-1}10^n, 0^n10^{n-1})$

□

The following lemma shows that the variation distance of a pair of sequences w, w' such that w' is a cyclic shift of w is polynomially large.

Lemma 101. *The variation distance* $d_s(w,w')$ *between* w *and a cyclic shift* w' *is "polynomially big".*

Proof. In Chapter 3 we provided a test that distinguishes pairs of cyclic shifted strings. That shows the variation distance is big. □

Unfortunately, the steps towards a proof of the conjectures exhibited in this section is far from being possible with the machinery developed in this Thesis so far. We leave the interested reader with the task of pursuing those further directions.

Chapter 6

Conclusion

In this thesis we explored a few different aspects of the Trace Reconstruction Problem. In particular, we focused on solving the Pair Trace Identification Problem, and managed to obtain a partial result in the form of the distinguishability of random strings when the deletion parameter equals $p = \frac{1}{2}$. We showed that .

There is plenty of room for improvement and future work. In particular regarding the extension of the existing results that led towards the distinguishability of random traces that could lead to a full proof of the Pair Trace Identification Problem at least for the case where $p = \frac{1}{2}$. Additionally, proving true the conjecture for the lower bound on the variation distance is not only an important problem towards the full solution of the Trace Identification Problem, but also would mean the solution of a problem with an undeniable aesthetic appeal.

In synthesis, the Trace Identification Problem is a rich source of interesting problems regarding the combinatorics of strings and its asymptotic properties, and one which has strong connections with areas ranging from error correcting codes to Computational Biology. It was a delight to have had the chance to work on it.

Appendix A

Notation reference

Definition 102. *Let* $C(w, s) = (number\ of\ times\ s\ is\ a\ subsequence\ of\ w)$.

Definition 103. *Let* $s \trianglelefteq w$ *denote* s *is a subsequence of* w.

Definition 104. *Let* $w = w_1 \cdots w_n$ *be a sequence in* $\{0,1\}^n$. $w[:i]$ *denotes* $w_1 \cdots w_i$.

Definition 105. *Let* $w = w_1 \cdots w_n$ *be a sequence in* $\{0,1\}^n$. $w[i:]$ *denotes* $w_i \cdots w_n$.

Definition 106. *Let* $w \in \{0,1\}^n$, *and* $s \in \{0,1\}^k$ *for some let* $f_s(w)$ *be the number of times* s *appears as a subsequence of* w. *Where* w *is represented as a vector* $(x_1, \cdots, x_m)$ *(the positions of the ones in* w*).*

Definition 107. $d^1(w, w')$ *be the minimum length of a sequence* s *for which* $C(w, s) \neq C(w', s)$

Definition 108. *The* $k-$*deck of a sequence* $w \in \{0,1\}^n$ *is the multiset of all subsequences of size* k *of* w.

Definition 109. $k_0(n) = \max_{w,w' \in \{0,1\}^n} d^1(w, w')$.

Definition 110. $\mathcal{P}_{w,p}(s) = C(w, s)p^{n-|s|}(1-p)^{|s|}$ *be the distribution over subsequences of* w *induced by the deletion process here studied.*

Definition 111. *The* $0-$*vector of a sequence* $w \in \{0,1\}^n$ *with* m *ones is* $(b_0, \cdots, b_m)$. *Where* b_i *indicates the number of zeroes before the* $i+1$*-th and after the* i*th one. By convention,* b_0 *is the number of zeroes before the first one,* b_m *is the number of zeroes after the last one.*

Definition 112. *The* $1-$*vector of a sequence* $w \in \{0,1\}^n$ *with* m *zeroes is* $(a_0, \cdots, a_m)$. *Where* a_i *indicates the number of ones before the* $i+1$*-th and after the* i*th zero. By convention,* a_0 *is the number of ones before the first zero,* a_m *is the number of ones after the last zero.*

Bibliography

[1] I.B. Leader. Extremal Combinatorics. Class notes Combinatoric, Cambridge, Michelmas 2012

[2] Thomas Holenstein, Michael Mitzenmacher, Rina Panigrahy, Udi Wieder, Trace reconstruction with constant deletion probability and related results.

[3] T. Batu, S. Kannan, S. Khanna and A. McGregor. Reconstructing strings from random traces. 2004

[4] V.I. Levenshtein. Efficient reconstruction of sequences. IEEE Transactions on Information Theory, vol 47. 2001

[5] Dekel Tsur. Tight bounds for string reconstruction using substring queries.

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

[7] Miroslav Dudik, Leonard J. Schulman, Reconstruction from subsequences, Journal of Combinatorial Theory, Series A 103 (2003) 337âĂŞ348

[8] I. Krasikov, Y. Roditty, On a reconstruction problem for subsequences, J. Combin Theory, Ser. A 77 (2) (1997) 344-348

[9] P. Borwein, T. Erdelyi, G. Kos, Littlewood-type problems on [0,1], Proc. London Math. Soc.(3) 79 (1) (1999) 22-46

[10] P. Bowein, C. Ingalls, The Prouhet-Tarry-Escott problem revisited, Enseign. Math. (2) 40 (1-2) (1994) 22-46

[11] A.D. Scott, Reconstructing Sequences, Discrete Mathematics Volume 175, Issues 1 âĂŞ 3, 15 October 1997, Pages 231âĂŞ238