Limitations of Mean-Based Algorithms for Trace Reconstruction at Small Edit Distance

Elena Grigorescu, Madhu Sudan, Minshen Zhu

Abstract*—Trace reconstructionTrace reconstruction *considers the task of recovering an unknown string $\mathbf{x} \in \{0,1\}^n$ given a number of independent "traces", i.e., subsequences of $\mathbf{x}$ obtained by randomly and independently deleting every symbol of $\mathbf{x}$ with some probability p. The information-theoretic limit of the number of traces needed to recover a string of length n is still unknown. This limit is essentially the same as the number of traces needed to determine, given strings $\mathbf{x}$ and $\mathbf{y}$ and traces of one of them, which string is the source.

The most-studied class of algorithms for the worst-case version of the problem are "mean-based" algorithms. These are a restricted class of distinguishers that only use the mean value of each coordinate on the given samples. In this work we study limitations of mean-based algorithms on strings at small Hamming or edit distance.

We show that, on the one hand, distinguishing strings that are nearby in Hamming distance is "easy" for such distinguishers. On the other hand, we show that distinguishing strings that are nearby in edit distance is "hard" for mean-based algorithms. Along the way, we also describe a connection to the famous Prouhet-Tarry-Escott (PTE) problem, which shows a barrier to finding explicit hard-to-distinguish strings: namely such strings would imply explicit short solutions to the PTE problem, a well-known difficult problem in number theory. Furthermore, we show that the converse is also true, thus, finding explicit solutions to the PTE problem is equivalent to the problem of finding explicit strings that are hard-to-distinguish by mean-based algorithms.

Our techniques rely on complex analysis arguments that involve careful trigonometric estimates, and algebraic techniques that include applications of Descartes' rule of signs for polynomials over the reals.

***Index Terms*—Trace reconstruction, mean-based algorithms, complex analysis, multiplicity of zeros**

Elena Grigorescu and Minshen Zhu are with the Computer Science Department, Purdue University, West Lafayette, IN, 47907 USA e-mail: {elena-g, zhu628}@purdue.edu. Research supported in part by NSF CCF-1910659 and NSF CCF-1910411

Madhu Sudan is with the Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Boston, MA, 02134 e-mail: madhu@cs.harvard.edu. Research supported in part by a Simons Investigator Award and NSF Award CCF 1715187.

This paper was presented in part at *2021 IEEE International Symposium on Information Theory (ISIT).*

I. Introduction

In the trace-reconstruction problem, a string $\mathbf{x} \in \{0,1\}^n$ is sent over a deletion channel, which deletes each entry independently, with probability $p \in [0,1)$, resulting in a *trace* $\tilde{\mathbf{x}} \in \{0,1\}^\ell$ of smaller length. The goal is to reconstruct $\mathbf{x}$ exactly, from a small set of independent traces. The trace-reconstruction problem was introduced by Batu, Kannan, Khanna and McGregor [1], motivated by a natural problem in computational biology, in which a common ancestor DNA sequence is sought from a set of similar DNA sequences that might have resulted from the process of random deletions in the ancestor DNA. The information-theoretic limits and tight complexity of this problem have proven elusive so far, despite significant followup interest in a variety of relevant settings [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. The current upper bound in the worst-case formulation was recently improved by Chase [18], who showed that $\exp(\tilde{O}(n^{1/5}))$ traces are sufficient for reconstruction, thus beating the previous record of $\exp(O(n^{1/3}))$ traces due to [7], [8]. However, the most general lower bound is only $\tilde{\Omega}(n^{3/2})$ [11], [20], hence leaving the status of the problem widely open.

To gain more insight into the trace-reconstruction problem, we study the *trace-distinguishing* variant, in which, given two string $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$, the algorithm receives traces from one of the two trace distributions and is tasked to output the correct one. The trace-distinguishing problem is information theoretically equivalent to the classical trace-reconstruction problem [4]. From a computational standpoint, the same upper and lower bounds as for the general problem hold for the trace-distinguishing variant.

In this work we aim to get more insight into the worst-case trace distinguishing problem from understanding the role of *distance* in the complexity of the problem. We ask the following questions: Are all pairs of strings that are close in Hamming distance easily distinguishable?

Are all pairs of strings that are close in edit distance[1] easily distinguishable? Note that the strings used for showing the lower bounds in [11], [20] only differ in two locations, and are indeed efficiently distinguishable (these were the strings $\mathbf{x} = (01)^k 101(01)^k$ and $\mathbf{y} = (01)^k 011(01)^k$). On the other hand, it is also reasonable to believe that trace distributions of strings that are very different from each other are also easily distinguishable. In fact, there exist "codes", namely sets of strings that are very far from each other, whose elements (codewords) lead to trace distributions that are very easily distinguishable from each other [14], [16]. These codes can be constructed by efficient algorithms, leading to some partial notion of explicitness that may be later exploited in further algorithms for the trace-reconstruction problem.

Here we approach the above questions by analyzing a restricted class of algorithms, namely mean-based. Mean-based algorithms only use the empirical mean of individual bits, and hence they operate by disregarding the actual samples, and computing only with the information given by the averages of each bit $\tilde{\mathbf{x}}_i$ over the sample set S of independent traces, namely $\mathbb{E}_S[\tilde{\mathbf{x}}_i]$. While they appear restrictive, mean-based algorithms are in fact a very powerful class of algorithms – for example, the upper bounds of [8], [7] are obtained via mean-based algorithms.

However, there exist strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ [8], [7] that mean-based algorithms cannot distinguish with fewer than $\exp(\Omega(n^{1/3}))$ traces. This lower bound is based on a result in complex analysis [21], which only implies the existence of such strings $\mathbf{x}$ and $\mathbf{y}$, and not what such strings would look like structurally. In particular, we don't even have efficient algorithms for constructing such strings.

Our main results here prove that there exist *explicit* strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ at edit distance only 4, for which every mean-based algorithm requires a *super-polynomial* in n number of samples. By "explicit" strings we mean strings whose support set can be described mathematically, by algebraic equations (say, for example, $\mathbf{x} \in \{0,1\}^n$ is such that $\mathbf{x}_i = 1$ iff $i = 2^k$, for some integer k).

On the other hand, we identify some structural properties of strings at low edit distance that yield polynomial-time mean-based trace reconstruction. In [22], [15] the authors show that strings at small Hamming distance are efficiently distinguishable. We complement these results by observing that they are efficiently distinguishable even by mean-based algorithms.

We believe that understanding structural properties that are bottlenecks (such as explicit, hard-to-distinguish strings) for the algorithms we know of, as well as understanding structural properties that lead to fast algorithms, are necessary steps towards understanding the complexity of the trace-reconstruction problem.

We formalize our results next.

A. Our results

We start with an observation about strings at small Hamming distance.

Theorem 1. *Let $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ be two distinct strings within Hamming distance d from each other. There is a mean-based algorithm that distinguishes between $\mathbf{x}$ and $\mathbf{y}$ with high probability using $n^{O(d)}$ traces.*

The result is a slight strengthening of a recent result of [15], who proved exactly the same bounds for general algorithms. A weaker version was also shown in [22], [23], where it is proved that strings at Hamming distance $2k$ have distinct k-decks, i.e. multisets of all $\binom{n}{k}$ subsequences of length k. Our contribution here is essentially to notice that the techniques of [22], [23] imply that mean-based algorithms can in fact distinguish such trace distributions (see Appendix B for a more detailed discussion and the complete proof).

Our main results concern the negative results at small edit distance.

Theorem 2. *Assume the deletion probability $p \in (0,1)$. There exist (explicit) strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ within edit distance 4 of each other such that any mean-based algorithm requires $\exp\left(\Omega(\log^2 n)\right)$ traces to distinguish between $\mathbf{x}$ and $\mathbf{y}$ with high probability.*

Along the way, we also formalize a connection to the famous Prouhet-Tarry-Escott (PTE) [24], [25], [26] problem from number theory. The PTE problem is related to classical variants of the Waring problem and problems about minimizing the norm of cyclotomic polynomials, considered by Erdös and Szekeres [27], [28]. Perhaps not surprisingly, our explicit solution from Theorem 2 is based on products of cyclotomic polynomials.

In the PTE problem, given an integer $k \geq 0$, one would like to find sets A and B of integer solutions, with $A = \{\alpha_1, \alpha_2, \ldots, \alpha_s\}$ and $B = \{\beta_1, \beta_2, \ldots, \beta_s\}$, satisfying the system $\sum_{i \in [s]} \alpha_i^j = \sum_{i \in [s]} \beta_i^j$, for all $j \in [k]$, with $\alpha_i \neq \beta_j$ for all $i, j \in [s]$. The goal is to find such solutions with *size* s as small as possible compared to the *degree* k. It is easy to show that, most generally,

[1] The term "edit distance" in this paper refers to the number of insertions and deletions needed to transform one string into the other. This notion does not take into account substitutions, and is sometimes called "longest common subsequence (LCS) distance" in the literature.

it must be the case that $s \geq k+1$; and a pigeon-hole counting argument shows the existence of solutions with $s = O(k^2)$ [29]. With the additional constraint that the system is not satisfied for degree $k+1$, solutions of size $s = O(k^2 \log k)$ are known to exist [30]. However, all these are existential, non-constructive solutions, and the only general explicit solutions have size $s = \Theta(2^k)$ (e.g., [26], [28]).

We note that connections between the trace-reconstruction problem and the PTE problem have been previously made. In particular, Krasikov and Roditty [22] noticed that pairs of strings that have the same k decks yield solutions to PTE systems.

We first show that explicit strings that are exponentially hard to distinguish by mean-based algorithms imply solutions of small size to a PTE system, as follows. This can be viewed as a deeper reason for why the negative result for mean-based algorithms in [7], [8] is based on non-constructive arguments.

Theorem 3. *Fix any $\varepsilon \in (0, 1/3]$. Given distinct strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ such that any mean-based algorithm requires $\exp(\Omega(n^{\varepsilon}))$ traces to distinguish between $\mathbf{x}$ and $\mathbf{y}$, the following two sets constitute a solution to the degree-k PTE system*

$$D(\mathbf{x}) = \{i \colon x_i = 1\}, \quad D(\mathbf{y}) = \{i \colon y_i = 1\},$$

with size $n = (k \log^2 k)^{1/\varepsilon}$.

We also prove a converse of this result. However, the converse is in terms of an upper bound on the magnitude of the solutions to the PTE problem, rather than the size of the solution.

We note that the counting argument [29], [30] that shows existential results for the PTE solutions in terms of size and degree, in fact gives solutions in which the values of the integers are bounded from above by, say, an integer M. When the size of the solution is $s = \Omega(k^3)$, the proof [29], [30] shows that there exist solutions where M is polynomial in s. When the size s is exponential in the degree k, there are constructions with $M = O(s)$ [26]. Hence, the size of the solution and the magnitude of the solution lead to qualitatively similar bounds in interesting ranges of the parameters.

Theorem 4. *Suppose $A, B \subseteq \mathbb{N}$ form a solution to the degree-k PTE system, and let $n := \max A \cup B$. Define the following strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^{n+1}$:*

$$\forall i \in \{0, 1, \dots, n\},$$
$$x_i = \begin{cases} 0 & \text{if } i \notin A \\ 1 & \text{if } i \in A \end{cases}, \quad y_i = \begin{cases} 0 & \text{if } i \notin B \\ 1 & \text{if } i \in B \end{cases}.$$

Then for any $\varepsilon > 0$, $n^{\Omega(k)}$ traces are necessary for mean-based algorithms to distinguish between $0^{\ell}\mathbf{x}$ and $0^{\ell}\mathbf{y}$, where $\ell = n^{3+\varepsilon}$.

We remark that $n^{\Omega(k)} = N^{\Omega(k)}$ where $N = \ell + n = \Theta(n^{3+\varepsilon})$ is the length of strings $0^{\ell}\mathbf{x}$ and $0^{\ell}\mathbf{y}$. Since $k \leq n$ for any PTE solutions, the largest possible bound we could get via Theorem 4 is $N^{\Omega(N^{1/(3+\varepsilon)})} = \exp(N^{1/(3+\varepsilon)} \log N)$. This is consistent with the results of [8], [7], which showed that $\exp(N^{1/3})$ traces are sufficient for mean-based algorithms to distinguish between any two strings of length N. We also note that the hard strings obtained from general PTE solutions may have unbounded edit distance, thus they do not directly imply Theorem 2.

The strong connection with the PTE problem, which is believed to be a difficult problem in Number Theory, may be interpreted as evidence to the difficulty of finding explicit hard-to-distinguish strings for the trace-reconstruction problem. Such hard instances could be desirable when, for example, one wants to design instance-dependent algorithms to bypass the "mean-based barrier". We remark that a similar reduction from the problem of finding small-size explicit solutions for the PTE problem to the computational hardness of the Bounded Distance Decoding problem for Reed-Solomon codes from [31] revealed a similar barrier for the respective decoding problem.

PTE systems appear to be intimately connected to the trace-reconstruction problem. Indeed, the analysis of mean-based algorithms often reduces to the study Littlewood-type polynomials, namely polynomials with $\{-1, 0, 1\}$ coefficients, on the complex unit circle. This in turn often involves understanding the multiplicity of the root 1, which is again a question tightly related to the PTE problem (see discussion in Section III).

Finally, with the tools established in this paper, we apply the Descartes rule of signs [32] to complete the proofs of some of our results, e.g., the proof of Theorem 1. As another application of this rule to larger edit distances, we also obtain the following theorem, formalized in Section VI.

Theorem 5. *(Informal) Strings $x, y \in \{0,1\}^n$ with $d_E(x, y) = d \geq 1$ and certain special block structures are distinguishable by mean-based algorithms using $n^{O(d)}$ traces. In particular, the statement holds for every pair of strings at edit distance 2.*

This version This version of our paper includes several improvements over our previous version [33]. Some of these improvements are inspired by recent work of Sima and Bruck [34], which appeared after the previous version of our paper was published online. In particular, here we improve Theorem 7 due to a technical lemma in [34], whose proof we simplify further in Lemma 4. We explain the differences between the proofs in Section I-B after Theorem 7.

Other changes in this version include strengthening and simplification of theorems 1, 3, 5 and of Lemma 6. In addition, Theorem 4 is a new theorem that is essentially the converse of Theorem 3. This new theorem answers the open questions we raised in our previous version. We suggest new open problems in Section VII.

B. Our techniques

a) The [8], [7] reduction to complex analysis: We recall that a mean-based algorithm only works with "mean traces" of a string. Formally, the mean trace of string $\mathbf{x} \in \{0,1\}^n$ is a vector $\mathbf{E}(\mathbf{x}) = (E_0(\mathbf{x}), \cdots, E_{n-1}(\mathbf{x})) \in [0,1]^n$, where the j-th coordinate is defined as

$$E_j(\mathbf{x}) = \mathop{\mathbb{E}}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\mathbf{x}}} [\tilde{x}_j] .$$

It is not hard to see (e.g., [4]) that understanding the sample complexity for distinguishing between $\mathbf{x}$ and $\mathbf{y}$ by mean-based algorithms essentially amounts to understanding the ℓ_1-distance between the mean traces.

Proposition 1. *Given strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ with $\|\mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{y})\|_{\ell_1} = \varepsilon$, $\Omega(1/\varepsilon)$ traces are necessary, and $O(1/\varepsilon^2)$ traces are sufficient for mean-based algorithms to distinguish between $\mathbf{x}$ and $\mathbf{y}$.*

Our techniques focus on analyzing the modulus of Littlewood polynomials with $\{-1,0,1\}$ coefficients in certain regions of the complex plane. The reduction to complex analysis was established in [8], [7]. They define the associated polynomials $P_{\mathbf{x}}(z) = \sum_{j=0}^{n-1} E_j(\mathbf{x}) \cdot z^j$ and the related polynomial $Q_{\mathbf{x}}(p + qz) = q^{-1} P_{\mathbf{x}}(z) = \sum_{k=0}^{n-1} x_k \cdot (p + qz)^k$ (and hence $Q_{\mathbf{x}}(z) = \sum_{k=0}^{n-1} x_k \cdot z^k$), which is obtained from writing the E_j's explicitly as

$$\begin{aligned} E_j(\mathbf{x}) = \mathop{\mathbb{E}}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\mathbf{x}}} [\tilde{x}_j] &= \sum_{k=0}^{n-1} \Pr[\tilde{x}_j \text{ comes from } x_k] \cdot x_k \\ &= \sum_{k=0}^{n-1} \binom{k}{j} p^{k-j} q^{j+1} \cdot x_k, \end{aligned}$$

Here p is the deletion probability and $q = 1 - p$. The reduction is summarized in the following theorem.

Theorem 6 ([8], [7])**.** *The sample complexity of mean-based algorithms for the trace-distinguishing problem for strings $\mathbf{x}$ and $\mathbf{y}$ is lower bounded by (up to constants) the inverse of*

$$\sup\{|Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)| : w \in \partial B(p; q)\},$$

where $\partial B(p; q)$ denotes the circle of radius q centered at p in the complex plane.

For completeness, we include the details of the reduction in Appendix A. Note that all coefficients of $Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)$ belong to $\{-1, 0, 1\}$.

b) Applications of Descartes' Rule of Signs: Here we relate the ℓ_1-distance between the mean traces of $\mathbf{x}$ and $\mathbf{y}$ to the multiplicity of zero of the polynomial $Q_{\mathbf{x}} - Q_{\mathbf{y}}$ at 1. Specifically, we show that as long as 1 is a root with multiplicity no more than k, the ℓ_1-distance is at least $n^{-O(k)}$ (assuming the deletion probability p is a constant).

Theorem 7. *Let $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ be two distinct strings. Suppose the polynomial $f(z) = Q_{\mathbf{x}}(z) - Q_{\mathbf{y}}(z)$ has k roots at $z = 1$. Then for any deletion probability $p \in (0,1)$, we have*

$$\|\mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{y})\|_{\ell_1} \geq \frac{q}{e} \left(\frac{q}{n}\right)^k .$$

Here $q = 1 - p$ is the retention probability.

Our proof of Theorem 7 is inspired by the proof of Lemma 6 in [34]. The proof presented here is arguably simpler than the one that appeared in a preliminary version of this paper, and the bound is also improved from $n^{-O(k^2)}$ to $n^{-O(k)}$. The new idea of [34] was to find a point in the complex plane with nice properties on a circle *centered at* 1, whereas our initial proof revolved around finding such a point on a circle *touching* 1. Their analysis uses an averaging argument, which we simplify to an application of the Maximum Modulus Principle in our new proof (see Lemma 4 in Section III).

It is then desirable to upper bound the multiplicity of zero at 1 for various polynomials. Descartes' rule of sign changes provides a convenient tool to achieve this.

Lemma 1 ([32], Theorem 36, Chapter 1, Part Five of [35])**.** *Let $Z(p)$ be the number of real positive roots of the real polynomial $p(x)$ (counting with multiplicity) and $C(p)$ the number of changes of sign of the sequence of its coefficients. We then have $C(p) \geq Z(p)$.*

We note that prior work that we are aware of on understanding the structure of polynomials with many roots at 1 (e.g., [36], [37]) do not appear to imply our bounds on the complex unit circle.

c) Remark: If $p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$ is a polynomial, we say a pair (i, j) $(0 \leq i < j \leq n)$ is a *sign change* if $a_i a_j < 0$ and $a_{i+1} = a_{i+2} = \ldots =$

$a_{j-1} = 0$. $C(p)$ exactly counts the number of such pairs (i, j).

We use this rule to prove the formal version of Theorem 5, namely Theorem 10. We also use it to give a simple proof of Theorem 1.

d) Complex analysis over shifted circles: For the negative results (i.e., Theorem 2 and 4), our strategy is to apply Theorem 6 and analyze the supremum of $|Q_{\mathbf{x}} - Q_{\mathbf{y}}|$. For Theorem 4, we upper bound how much the modulus of an analytic function could change in a small neighbourhood of 1 by controlling its derivative. For Theorem 2, the strongest barrier is that it is unclear how to control the edit distance. This seems even more difficult for non-constructive arguments such as the ones in [21] and [29]. Our construction is inspired by properties of product of cyclotomic polynomials and their relation to PTE solutions with special structures.

C. Related work

The first formulations of the problem were proposed by [38], [39], and the precise formulation we study here were developed in [1], [4], motivated by the connection with DNA reconstruction. DNA sequencing recently motivated the model of "coded" trace reconstruction [14], [16], in which the goal is to reconstruct codewords of a known code, rather than an arbitrary string. Furthermore, the worst-case trace-reconstruction problem was also studied in the memoryless replication-insertion channel [40].

Besides the worst-case trace reconstruction discussed earlier in the introduction, a well-studied variant has been the average-case trace-reconstruction problem, studied in [4], [6], [5], [10]. Here, the best current lower bound is $\Omega(\log^{5/2} n)$ [11], [20], and the best algorithms run in time $\exp\left(O(\log^{1/3} n)\right)$.

A recent intriguing result [17] considers the smooth variant, which is an intermediate model between the worst-case and the average-case models. In the smooth model, the initial string is obtain from an arbitrary worst-case string perturbed so that each coordinate is replaced by a uniformly random bit with some constant probability $0 < \sigma < 1$. In [17], the authors show that in this case reconstruction can be done efficiently.

Other variants consider string reconstruction from the multiset of substrings [9], [13], population recovery variants [41], matrix reconstruction and parametrized algorithms [15], and circular trace reconstruction [19].

a) Recent Work.: After a preliminary version of this paper was published, Davies, Rácz, Rashtchian and Schiffer considered a relaxed problem named *approximate trace reconstruction* and provided efficient algorithms for several classes of strings [42]. Here the goal is to recover a string that is close to the true source string in edit distance. Soon after, [43], [44] and [45] showed that for random source strings an approximate solution can be found with high probability using very few traces. We remark that approximate solutions serve as distinguishers for pairs of strings (in the specified class) that are sufficiently far from each other in edit distance. On the other hand, for distinguishing strings that are close to each other, Sima and Bruck showed that $n^{O(d)}$ traces are also sufficient, where d is their edit distance [34]. Of course, the algorithm proposed in [34] is not mean-based (otherwise it would contradict Theorem 2). Nevertheless, the analysis, to a large extent, resembles that for mean-based algorithms. Indeed, one of their technical contributions is improving one of our technical lemma which relates mean-based trace reconstruction and the multiplicity of zeros of certain polynomials.

D. Organization of the paper

In Section II we develop the necessary notations and basic facts. In Section III we prove Theorem 7, which is a key factor in our analysis, and use it to prove Theorem 3. In Appendix B we prove Theorem 1. In Section IV we prove Theorem 4. In Section V we prove Theorem 9, which is a more concrete version of Theorem 2. In Section VI we prove Theorem 10, which is an equivalent and more concrete version of Theorem 5. In the Appendix, we also explain how to reduce the analysis of mean-based algorithms to understanding the supremum of certain polynomials over a circle in the complex plane.

II. Preliminaries

Given $z \in \mathbb{C}$ and $r \in \mathbb{R}_{\geq 0}$, we write

$$B(z; r) := \{w \in \mathbb{C} \colon |w - z| \leq r\}$$

for the disk centered at z with radius r, and write $\partial B(z; r)$ for its boundary.

Let $p(w) = a_0 + a_1 w + \ldots + a_n w^n$ be a polynomial where $a_j \in \mathbb{C}$. Let $A \subseteq \mathbb{C}$ be a set. We define the following norms.

$$\|p\|_1 = \sum_{j=0}^{n} |a_j|, \quad \|p\|_2 = \left(\sum_{j=0}^{n} |a_j|^2\right)^{1/2},$$
$$\|p\|_A = \sup_{w \in A} |p(w)|.$$

When $A = \partial B(0; 1)$ is the complex unit circle, we also write $\|p\|_A = \|p\|_\infty$. These norms are connected by the following inequalities.

Lemma 2. *Let p be a degree-n polynomial with real coefficients. Then*

$$\frac{1}{\sqrt{n+1}} \cdot \|p\|_1 \leq \|p\|_2 \leq \|p\|_\infty \leq \|p\|_1 .$$

Proof. The first and third inequalities are applications of Cauchy-Schwartz and the triangle inequality, respectively. The second inequality comes from the following identity

$$\|p\|_2^2 = \frac{1}{2\pi} \int_0^{2\pi} \left|p\left(e^{i\theta}\right)\right|^2 d\theta,$$

where the right-hand-side is clearly upper bounded by $\|p\|_\infty^2$. □

We will use the following bounds for a point $z \in \partial B(p;q)$.

Lemma 3. *Fix $p \in (0,1)$ and $q = 1-p$. Let $z = p+qe^{i\theta}$ where $\theta \in (-\pi,\pi]$. The following bounds hold.*

1) $|z| \leq 1 - 2pq(\theta/\pi)^2$.
2) $|z-1| \leq q|\theta|$.
3) *For any integer $d \geq 0$, $|z^d - 1| \leq dq|\theta|$.*

Proof. Item 1: By convexity of $\sin(\cdot)$ over $[0,\pi/2]$, we have

$$\begin{aligned}
1-\cos x &= 2\sin^2\frac{x}{2} \\
&\geq 2\left(\left(1-\frac{x}{\pi}\right)\cdot\sin 0 + \frac{x}{\pi}\cdot\sin\frac{\pi}{2}\right)^2 \\
&= 2\left(\frac{x}{\pi}\right)^2
\end{aligned}$$

for $x \in [0,\pi]$. By substituting $-x$ for x, we can obtain the same inequality for $x \in [-\pi, 0]$. Thus we have

$$\begin{aligned}
|w| &= \sqrt{\left(p+q\cos\theta\right)^2 + \left(q\sin\theta\right)^2} \\
&= \sqrt{p^2 + 2pq\cos\theta + q^2} \\
&= \sqrt{(p+q)^2 - 2pq(1-\cos\theta)} \\
&\leq \sqrt{1 - 4pq\left(\frac{\theta}{\pi}\right)^2} \\
&\leq 1 - 2pq\left(\frac{\theta}{\pi}\right)^2,
\end{aligned}$$

where the last line is due to $(1+x)^r \leq 1+rx$ for $r \in [0,1]$ and $x \geq -1$ (Bernoulli's inequality).

Item 2: By elementary identities for trigonometric functions, we have

$$\begin{aligned}
\left|e^{i\theta}-1\right| &= \left|\cos\theta - 1 + i\sin\theta\right| \\
&= \left|2\sin\frac{\theta}{2}\right| \cdot \left|-\sin\frac{\theta}{2} + i\cos\frac{\theta}{2}\right| \\
&= 2\sin\frac{|\theta|}{2}.
\end{aligned}$$

Therefore

$$\begin{aligned}
|z-1| &= \left|p+qe^{i\theta}-1\right| = q\left|e^{i\theta}-1\right| \\
&= q\cdot 2\sin\frac{|\theta|}{2} \leq q|\theta|.
\end{aligned}$$

The inequality is due to $\sin x \leq x$ for $x \geq 0$.

Item 3: Due to Item 1 and the triangle inequality, we have

$$\left|z^d - 1\right| = |z-1| \cdot \left|\sum_{j=0}^{d-1} z^j\right| \leq q|\theta| \cdot \sum_{j=0}^{d-1} |z|^j \leq dq|\theta|.$$

□

In this paper, "with high probability" means with probability at least $2/3$. We will use p for the deletion probability and $q = 1-p$. In this paper p and q will be constants. Given a string $\mathbf{a} \in \{0,1\}^n$, a trace $\tilde{\mathbf{a}} \in \{0,1\}^{\leq n}$ is a subsequence of $\mathbf{a}$ obtained by deleting each bit of $\mathbf{a}$ independently with probability p. The length of $\tilde{\mathbf{a}}$ is denoted by $|\tilde{\mathbf{a}}|$. For $0 \leq j \leq n-1$, the j-th bit of $\mathbf{a}$ and $\tilde{\mathbf{a}}$ are written as a_j and $\tilde{a}_j$, respectively. The distribution of $\tilde{\mathbf{a}}$ is denoted by $\mathcal{D}_{\mathbf{a}}$. We also associate to $\mathbf{a}$ the following polynomial

$$Q_{\mathbf{a}}(w) := a_0 + a_1 w + a_2 w^2 + \ldots + a_{n-1} w^{n-1}.$$

The degree of $Q_{\mathbf{a}}$ is at most $n-1$.

For strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$, we will write $d_{\mathsf{H}}(x,y)$ for the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$, where $d_{\mathsf{H}}(\mathbf{x},\mathbf{y}) = |\{i \in [n] \colon x_i \neq y_i\}|$; and write $d_{\mathsf{E}}(\mathbf{x},\mathbf{y})$ for the edit distance between $\mathbf{x}$ and $\mathbf{y}$, namely the minimum number of insertions and deletions that transform $\mathbf{x}$ into $\mathbf{y}$.

III. Large ℓ_1-distance between mean traces from low multiplicity of root 1

In this section we prove Theorem 7, which will be a key stepping stone to obtain our main results. We need the following lemma, which finds a point w in the neighbourhood of 1 with nice properties. This lemma is first proven in [34] with $p \leq 1/2$, and here we give a simpler proof which works for any $p \in (0,1)$.

Lemma 4. *Let $f(z)$ be a polynomial of degree n. Suppose we can write*

$$f(z) = (z-1)^k g(z)$$

for some polynomial g with $|g(1)| \geq 1$. Then for any $p \in (0,1)$ and $q = 1-p$, there exists $w \in \mathbb{C}$ such that $|(w-p)/q|^n \leq e$ and $|f(w)| \geq (q/n)^k$.

Proof. Let $\Gamma = B(1; q/n)$ denote the closed disk with radius q/n centered at 1 on the complex plane. By the

Maximum Modulus Principle (see, e.g., Theorem 1.3 in Chapter III, §1 of [46]), there exists a point $w \in \partial\Gamma$ such that

$$|g(w)| = \sup_{z\in\Gamma} |g(z)| \geq |g(1)| \geq 1.$$

We denote $w_0 \coloneqq w - 1$. Therefore $|w_0| = q/n$, and

$$\left|\frac{w-p}{q}\right|^n = \left|\frac{1+w_0-p}{q}\right|^n = \left|\frac{q+w_0}{q}\right|^n$$
$$\leq \left(1+\frac{|w_0|}{q}\right)^n = \left(1+\frac{1}{n}\right)^n \leq e.$$

Finally, we also have

$$|f(w)| = |w-1|^k \cdot |g(w)| \geq (q/n)^k.$$

□

Now we can prove Theorem 7. We recall the statement below.

Theorem 7. *Let $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ be two distinct strings. Suppose the polynomial $f(z) = Q_{\mathbf{x}}(z) - Q_{\mathbf{y}}(z)$ has k roots at $z = 1$. Then for any deletion probability $p \in (0,1)$, we have*

$$\|\mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{y})\|_{\ell_1} \geq \frac{q}{e}\left(\frac{q}{n}\right)^k.$$

Here $q = 1 - p$ is the retention probability.

Proof. We recall the definition of $P_{\mathbf{x}}$:

$$P_{\mathbf{x}}(z) \coloneqq \sum_{j=0}^{n-1} E_j(\mathbf{x}) \cdot z^j.$$

The following identity is proven in [8], [7] (see Appendix A for a proof):

$$P_{\mathbf{x}}\left(\frac{w-p}{q}\right) = q \cdot Q_{\mathbf{x}}(w). \tag{1}$$

Since $f(z) = Q_{\mathbf{x}}(z) - Q_{\mathbf{y}}(z)$ is a polynomial of degree n with k roots at $z = 1$, we can write $f(z) = (z-1)^k g(z)$ for some polynomial g such that $g(1) \neq 0$. We can also conclude that $|g(1)| \geq 1$ since g has integer coefficients (to see this, consider $g(z+1) = f(z+1)/z^k$). Therefore, we can apply Lemma 4 to f and obtain w such that $|(w-p)/q|^n \leq e$ and $|f(w)| \geq (q/n)^k$. By the triangle inequality, we have

$$\begin{aligned}
&\left|P_{\mathbf{x}}\left(\frac{w-p}{q}\right) - P_{\mathbf{y}}\left(\frac{w-p}{q}\right)\right| \\
=&\left|\sum_{j=0}^{n-1} (E_j(\mathbf{x}) - E_j(\mathbf{y})) \cdot \left(\frac{w-p}{q}\right)^j\right| \\
\leq& e \cdot \sum_{j=0}^{n-1} |E_j(\mathbf{x}) - E_j(\mathbf{y})|.
\end{aligned}$$

On the other hand, equation (1) gives

$$\begin{aligned}
&\left|P_{\mathbf{x}}\left(\frac{w-p}{q}\right) - P_{\mathbf{y}}\left(\frac{w-p}{q}\right)\right| \\
=&q\,|Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)| \\
\geq& q\left(\frac{q}{n}\right)^k.
\end{aligned}$$

Putting everything together, we have obtained

$$\sum_{j=0}^{n-1} |E_j(\mathbf{x}) - E_j(\mathbf{y})| \geq \frac{q}{e}\left(\frac{q}{n}\right)^k.$$

□

A. Connection to the Prouhet-Tarry-Escott problem

The following is a classical statement about the PTE problem.

Theorem 8 (e.g. [28], Proposition 1)**.** *Given $s, k \in \mathbb{N}$ and for $\alpha_i, \beta_i \in \mathbb{N}$, with $i \in [s]$, the following are equivalent:*

- *$\sum_{i=1}^s \alpha_i^j = \sum_{i=1}^s \beta_i^j$, for $1 \leq j \leq k$, and $\sum_{i=1}^s \alpha_i^{k+1} \neq \sum_{i=1}^s \beta_i^{k+1}$.*
- *$\sum_{i=1}^s x^{\alpha_i} - \sum_{i=1}^s x^{\beta_i} = (x-1)^{k+1} q(x)$ where $q \in \mathbb{Z}[x]$ and $q(1) \neq 0$.*

This connection allows us to prove Theorem 3.

Theorem 3. *Fix any $\varepsilon \in (0, 1/3]$. Given distinct strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ such that any mean-based algorithm requires $\exp(\Omega(n^\varepsilon))$ traces to distinguish between $\mathbf{x}$ and $\mathbf{y}$, the following two sets constitute a solution to the degree-k PTE system*

$$D(\mathbf{x}) = \{i \colon x_i = 1\}, \quad D(\mathbf{y}) = \{i \colon y_i = 1\},$$

with size $n = (k \log^2 k)^{1/\varepsilon}$.

Proof. Denote by m the multiplicity of root 1 of $Q_{\mathbf{x}} - Q_{\mathbf{y}}$. We consider two cases.

Case 1: $m \geq k + 1$. Let $\alpha_1, \alpha_2, \ldots, \alpha_s$ enumerate the set $D(\mathbf{x})$ where $s \leq n$ is the cardinality of $D(\mathbf{x})$. Similarly, we also let $\beta_1, \beta_2, \ldots, \beta_s$ enumerate $D(\mathbf{y})$. Note that $D(\mathbf{x})$ and $D(\mathbf{y})$ must have the same cardinality since otherwise $\mathbf{x}$ and $\mathbf{y}$ have different Hamming weights (and thus are distinguishable using constant traces). We have

$$\sum_{i=1}^s x^{\alpha_i} - \sum_{i=1}^s x^{\beta_i} = Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w) = (x-1)^m q(x)$$

for some $q \in \mathbb{Z}[x]$, $q(1) \neq 0$. Therefore, Theorem 8 implies that $D(\mathbf{x})$ and $D(\mathbf{y})$ form a solution to the degree-$(m-1)$ PTE system. In particular, they form a solution to the degree-k PTE system since $m - 1 \geq k$.

Case 2: $m \le k$. We will show by contradiction that this case never occurs. Otherwise, Theorem 7 gives us

$$\sum_{j=0}^{n-1} \left| \mathop{\mathbb{E}}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\mathbf{x}}} [\tilde{x}_j] - \mathop{\mathbb{E}}_{\tilde{\mathbf{y}} \sim \mathcal{D}_{\mathbf{y}}} [\tilde{y}_j] \right| \ge \frac{q}{e} \left(\frac{q}{n}\right)^m \ge \frac{q}{e} \left(\frac{q}{n}\right)^k$$
$$= \exp\left(-O(k \log n)\right).$$

On the other hand, the relation $n = (k \log^2 k)^{1/\varepsilon}$ also gives

$$n^\varepsilon = k \log^2 k, \quad \log n = \frac{1}{\varepsilon}\left(\log k + 2 \log\log k\right) = O(\log k),$$

which means $k \log n = O(k \log k) = o(n^\varepsilon)$ as $k, n \to \infty$. Therefore $\exp\left(o(n^\varepsilon)\right)$ traces are sufficient for a mean-based algorithm to distinguish between $\mathbf{x}$ and $\mathbf{y}$. However, this is a contradiction to the assumption that any mean-based algorithm requires $\exp(\Omega(n^\varepsilon))$ traces to distinguish between $\mathbf{x}$ and $\mathbf{y}$. □

IV. From PTE solutions to hard-to-distinguish strings

In this section, we prove Theorem 4, which says that PTE solutions imply “hard” strings for mean-based trace reconstruction.

The proof uses the following lemma.

Lemma 5 (Lemma 5.4 of [47])**.** *Suppose*

$$p(x) = \sum_{j=0}^{n} a_j x^j, \qquad |a_j| \le 1, a_j \in \mathbb{C}$$
$$p(x) = (x-1)^k q(x),$$
$$q(x) = \sum_{j=0}^{n-k} b_j x^j, \qquad b_j \in \mathbb{C}.$$

Then

$$\|q\|_1 = \sum_{j=0}^{n-k} |b_j| \le (n+1)\left(\frac{en}{k}\right)^k.$$

The following lemma is an analogue of the Mean Value Theorem for analytic functions.

Lemma 6. *Let $f(z)$ be an analytic function on an open set D, such that $|f'(z)| \le M$ for all $z \in D$. Then for z_0, z in the closure of D such that the line connecting z and z_0 is contained in D, we have*

$$|f(z)| \le |f(z_0)| + M \cdot |z - z_0|.$$

Proof. We write $f(x+yi) = u(x,y) + iv(x,y)$ for functions $u, v \colon \mathbb{R}^2 \to \mathbb{R}$.

Since $f(z)$ is an analytic function, it satisfies the Cauchy-Riemann equations (see, for instance, Chapter I, §6 of [46]):

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}, \text{ and } f'(x+yi) = \frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} i.$$

Let $r(x,y) \coloneqq |f(x+yi)|$. In other words, $r^2 = u^2 + v^2$. Taking partial derivatives of x and y on both sides gives

$$2r \cdot \frac{\partial r}{\partial x} = 2u \cdot \frac{\partial u}{\partial x} + 2v \cdot \frac{\partial v}{\partial x},$$
$$2r \cdot \frac{\partial r}{\partial y} = 2u \cdot \frac{\partial u}{\partial y} + 2v \cdot \frac{\partial v}{\partial y}.$$

Squaring both sides of both equations and combining give

$$\|\nabla r\|_2 \coloneqq \left\| \left(\frac{\partial r}{\partial x}, \frac{\partial r}{\partial y} \right) \right\|_2 = |f'(z)| \le M.$$

Now consider the auxiliary function

$$h(t) \coloneqq |f((1-t)z_0 + tz)| = r(x_t, y_t)$$

where $t \in [0,1]$, and $x_t, y_t \in \mathbb{R}$ are such that $(1-t)z_0 + tz = x_t + y_t i$. By the chain rule and Cauchy-Schwartz, for all $t \in (0,1)$ we have

$$h'(t) = \langle \nabla r(x_t, y_t), (x - x_0, y - y_0) \rangle$$
$$\le \|\nabla r\|_2 \cdot \|(x - x_0, y - y_0)\|_2$$
$$\le M \cdot |z - z_0|.$$

By the Mean Value Theorem, there exists $\bar{t} \in (0,1)$ such that

$$h'(\bar{t}) = \frac{h(1) - h(0)}{1 - 0} = |f(z)| - |f(z_0)|.$$

This implies

$$|f(z)| = |f(z_0)| + h'(\bar{t}) \le |f(z_0)| + M \cdot |z - z_0|.$$

□

Lemma 7. *Let $f(z)$ be a polynomial of degree n which can be factorized as $f(z) = (z-1)^{k+1} q(z)$ for some polynomial $q(z)$. Then for any $\alpha > 0$ we have*

$$\sup\left\{ \left| f\left(p + qe^{i\theta}\right) \right| : |\theta| < 1/\left(qn^{1+\alpha}\right) \right\} < 12n^{1-\alpha k}.$$

Proof. Let θ be such that $|\theta| < 1/(qn^{1+\alpha})$, and let $z = p + q^{i\theta}$. Item 2 of Lemma 3 implies $|z - 1| \le q|\theta| \le 1/n^{1+\alpha}$.

Denote $g(z) = (z-1)q(z)$. By Lemma 5, we have

$$\|g\|_1 \le (n+1)\left(\frac{en}{k}\right)^k < 6n^{k+1}.$$

Therefore $|g'(z)| \leq (n+1) \cdot \|g\|_1 \leq 12n^{k+2}$. Applying Lemma 6 with D being the open unit disk, $z_0 = 1$ and $z = p + qe^{i\theta}$, we have

$$\begin{aligned} |g(z)| &\leq |g(1)| + 12n^{k+2} \cdot |z-1| \\ &\leq 12n^{k+2} \cdot 1/n^{1+\alpha} < 12n^{k+1}. \end{aligned}$$

The lemma follows since

$$\begin{aligned} |f(z)| &= |z-1|^k \cdot |g(z)| \\ &\leq \left(1/n^{1+\alpha}\right)^k \cdot 12n^{k+1} \\ &= 12n^{1-\alpha k}. \end{aligned}$$

□

Now we are ready to prove Theorem 4. We recall the statement below.

Theorem 4. *Suppose $A, B \subseteq \mathbb{N}$ form a solution to the degree-k PTE system, and let $n := \max A \cup B$. Define the following strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^{n+1}$:*

$$\forall i \in \{0,1,\ldots,n\},$$
$$x_i = \begin{cases} 0 & \text{if } i \notin A \\ 1 & \text{if } i \in A \end{cases}, \quad y_i = \begin{cases} 0 & \text{if } i \notin B \\ 1 & \text{if } i \in B \end{cases}.$$

Then for any $\varepsilon > 0$, $n^{\Omega(k)}$ traces are necessary for mean-based algorithms to distinguish between $0^\ell \mathbf{x}$ and $0^\ell \mathbf{y}$, where $\ell = n^{3+\varepsilon}$.

Proof of Theorem 4. We write $f := Q_{\mathbf{x}} - Q_{\mathbf{y}}$. Due to Theorem 6, it suffices to show that

$$\sup\left\{\left|w^\ell f(w)\right| : w \in \partial B(p;q)\right\} \leq n^{-\Omega(k)}.$$

Writing $w = p + qe^{i\theta}$ where $\theta \in (-\pi, \pi]$, we prove the theorem in the following two cases.

Case 1: $|\theta| \geq 1/(qn^{1+\varepsilon/3})$.

By Item 1 of Lemma 3, we have

$$\begin{aligned} |w| &\leq 1 - 2pq\left(\frac{\theta}{\pi}\right)^2 \\ &\leq 1 - 2pq\left(\frac{1}{4qn^{1+\varepsilon/3}}\right)^2 \\ &\leq 1 - \frac{p}{8qn^{2+2\varepsilon/3}}. \end{aligned}$$

Therefore

$$\begin{aligned} \left|w^\ell f(w)\right| &= |w|^\ell \cdot |f(w)| \\ &\leq \left(1 - \frac{p}{8qn^{2+2\varepsilon/3}}\right)^{n^{3+\varepsilon}} \cdot (n+1) \\ &\leq \exp\left(-\Omega(pn^{1+\varepsilon/3}/q)\right) \\ &< n^{-\Omega(pk/q)}. \end{aligned}$$

The last inequality is due to $n^{1+\varepsilon/3} > n \ln n \geq k \ln n$ for large enough n.

Case 2: $|\theta| < 1/(qn^{1+\varepsilon/3})$.

We recall that A and B form a solution to the degree-k PTE system. According to the definition of $\mathbf{x}$, $\mathbf{y}$ and Theorem 8, the polynomial f can be factorized as $f(z) = (z-1)^{k+1} q(z)$ for some polynomial $q(z)$. Therefore, we can apply Lemma 7 with $\alpha = \varepsilon/3$ and obtain that

$$|w^\ell f(w)| \leq |f(w)| < 12n^{1-\varepsilon k/3}.$$

Combining the two cases, we have

$$\sup\left\{\left|w^\ell f(w)\right| : w \in \partial B(p;q)\right\} \leq n^{-\Omega(k)}.$$

□

V. Hard strings at edit distance 4

The goal of this section is to prove Theorem 2, and thus exhibit two strings at edit distance 4 such that every mean-based algorithm requires super-polynomially many traces.

We will prove the following theorem, which is a more concrete version of Theorem 2.

Theorem 9. *Let k be an odd integer and $n = \sum_{j=0}^{k} 3^j$ be an even integer, and $R(w) = \prod_{j=0}^{k}\left(1 - w^{3^j}\right)$ be a polynomial of degree n. Let $E_n(w) = \sum_{j=0}^{n/2} w^{2j}$. Then $Q_{\mathbf{e}}(w) := E_n(w) - R(w)$ is a 0/1-coefficient polynomial which corresponds to a string $\mathbf{e} \in \{0,1\}^{n+1}$. Moreover, any two strings $\mathbf{x}$, $\mathbf{y}$ of the form $\mathbf{x} = \mathbf{a}10\mathbf{e}$ and $\mathbf{y} = \mathbf{a}\mathbf{e}01$ satisfy*

$$\begin{aligned} &\sup\left\{|Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)| : w \in \partial B(p;q)\right\} \\ &\leq \exp\left(-\Omega(\log^2 n)\right), \end{aligned}$$

where $\mathbf{a}$ is an arbitrary string of length n. Here $p, q \in (0,1)$ are constants.

Proof. $R(w)$ has the following properties: (1) The coefficients of R belong to $\{-1, 0, 1\}$ since each monomial occurs only once in the expansion. (2) Odd-degree terms have negative signs, and even-degree terms have positive signs. It follows that $E_n(w) - R(w)$ is a polynomial with 0/1 coefficients.

We can write

$$\begin{aligned} P(w) &= Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w) \\ &= w^n\left((w^2-1)Q_{\mathbf{e}}(w) - \left(w^{n+2} - 1\right)\right) \\ &= w^n(w^2-1)\left(Q_{\mathbf{e}}(w) - E_n(w)\right) \\ &= w^n(1-w^2)R(w). \end{aligned}$$

Consider a point $w = p + qe^{i\theta}$ on the circle $\partial B(p;q)$, where $\theta \in (-\pi, \pi]$. We consider two cases.

Case 1: $|\theta| \geq 3^{-k/4}\pi$.

Due to Item 1 of Lemma 3, we have

$$|w| \leq 1 - 2pq\left(\frac{\theta}{\pi}\right)^2 \leq 1 - 2pq \cdot 3^{-k/2}.$$

Therefore

$$\begin{aligned}|P(w)| &\leq |w|^n \cdot 2(n+1)\\ &\leq \left(1 - 2pq \cdot 3^{-k/2}\right)^n \cdot 2(n+1)\\ &\leq \exp\left(-\Omega\left(pq\sqrt{n}\right)\right).\end{aligned}$$

The last inequality is because $1 - x < e^{-x}$ and $n = \sum_{j=0}^{k} 3^j > 3^k$.

Case 2: $|\theta| < 3^{-k/4}\pi$.

By Item 3 of Lemma 3, we have

$$\begin{aligned}|R(w)| &= \prod_{j=0}^{k/4-1}\left|w^{3^j} - 1\right| \cdot \prod_{j=k/4}^{k}\left|w^{3^j} - 1\right|\\ &\leq \prod_{j=0}^{k/4-1} 3^j q\,|\theta| \cdot 2^{3k/4}\\ &\leq \prod_{j=1}^{k/4}\left(3^{-j}\pi\right) \cdot 2^{3k/4}\\ &\leq 3^{-k^2/32} \cdot (8\pi)^{k/4}\\ &= \exp\left(-\Omega(k^2)\right)\\ &= \exp\left(-\Omega\left(\log^2 n\right)\right).\end{aligned}$$

Therefore $|P(w)| \leq 2|R(w)| \leq \exp\left(-\Omega\left(\log^2 n\right)\right)$. □

The edit distance between strings $\mathbf{x}$ and $\mathbf{y}$ constructed in the theorem above is clearly at most 4. Thus, Theorem 2 follows via Theorem 6 (see Appendix A for its proof).

Remark 1. *We make several remarks on the theorem. First, the bound is essentially tight for the constructed strings, since the polynomial $Q_{\mathbf{x}} - Q_{\mathbf{y}}$ has $k+2 = O(\log n)$ roots at 1, and Theorem 7 implies that $n^{O(k)} = \exp(O(\log^2 n))$ traces are also sufficient for distinguishing between $\mathbf{x}$ and $\mathbf{y}$ by mean-based algorithms. Second, the theorem exhibits two strings which attain the bound in Theorem 7 for $k = \Theta(\log n)$, meaning that Theorem 7 generally cannot be improved (at least in the regime $k = \Theta(\log n)$). Third, by Theorem 8 the constructed strings imply a solution to the degree-$(k+1)$ PTE system. However, the size of the solution is exponential, since the sparsity of $Q_{\mathbf{x}} - Q_{\mathbf{y}}$ is $\Theta(3^k)$.*

VI. Higher edit distance with special structures

In this section we show a more general result about strings at higher edit distance that have a special structure which leads to easy distinguishability. At the end we also discuss some implications about the edit distance 2 and 4 cases.

We consider pairs of strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ with the following block structure:

$$\mathbf{x} = \mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_d, \quad \mathbf{y} = \mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_d,$$

where for each $i = 1, 2, \ldots, d$, $\mathbf{x}_i$ and $\mathbf{y}_i$ are strings of length $\ell_i > 0$. Moreover, each block i falls into one of the following cases:

1) $\mathbf{x}_i = \mathbf{y}_i$;
2) $\mathbf{x}_i = a_i\mathbf{s}_i$ and $\mathbf{y}_i = \mathbf{s}_i b_i$ for bits $a_i, b_i \in \{0,1\}$ and string $\mathbf{s}_i$;
3) $\mathbf{x}_i = \mathbf{s}_i a_i$ and $\mathbf{y}_i = b_i\mathbf{s}_i$ for bits $a_i, b_i \in \{0,1\}$ and string $\mathbf{s}_i$;
4) $\mathbf{x}_i = a_i\mathbf{s}_i$ and $\mathbf{y}_i = b_i\mathbf{s}_i$ for distinct bits $a_i, b_i \in \{0,1\}$ and string $\mathbf{s}_i$;
5) $\mathbf{x}_i = \mathbf{s}_i a_i$ and $\mathbf{y}_i = \mathbf{s}_i b_i$ for distinct bits $a_i, b_i \in \{0,1\}$ and string $\mathbf{s}_i$.

We remark that $\mathbf{x}$ and $\mathbf{y}$ of the above form must be within edit distance $2d$ of each other, yet there are certainly strings at edit distance $2d$ which fail to follow this pattern (for example $\mathbf{x} = a_1 \ldots a_d\mathbf{s}$ and $\mathbf{y} = \mathbf{s}b_1 \ldots b_d$ generally do not have such a block decomposition).

We also note that if $\mathbf{x}$ and $\mathbf{y}$ have different Hamming weights (the Hamming weight of a string is the number of 1s in it), this makes them easily distinguishable by a mean-based algorithm. This is because their traces will exhibit a difference of q in expected Hamming weight. Using $O(1/q^2)$ traces, this difference will be noticeable by a mean-based algorithm. Therefore, we will focus on the more interesting case where $\mathbf{x}$ and $\mathbf{y}$ have the same Hamming weight.

Since $Q_{\mathbf{x}}(1)$ exactly equals to the Hamming weight of $\mathbf{x}$, we know that $w - 1$ is a factor of the polynomial $Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)$ if $\mathbf{x}$ and $\mathbf{y}$ have the same Hamming weight. It is thus natural to factor $Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w) = (w-1)R(w)$ for some polynomial R, and study the multiplicity of zeros of R. The special block structure described above allows us to explicitly write down the expression for R, and thus to study the number of sign changes in R. This is the main idea in proving the following theorem, which is the formal version of Theorem 5.

Theorem 10. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be strings with the special structure mentioned above. Then*

$$\|\mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{y})\|_{\ell_1} \geq \frac{q}{e}\left(\frac{q}{n}\right)^{3d}.$$

Proof. As a warm-up, let us first consider the case where all a_i's and b_i's are zero. Under this assumption cases 4 and 5 never arise in the above block decomposition. We can partition $[n]$ into three sets S_1, S_2, S_3, each of which collecting the indices of contiguous substrings of $\mathbf{x}$ (and therefore of $\mathbf{y}$) of lengths ℓ_i corresponding to the respective case of the first three special cases above. Let $t_i = \sum_{j=1}^{i-1} \ell_j$ be the starting index of block i (note that $t_1 = 0$ and $t_{d+1} = n$). As we are going to decompose the polynomials $Q_{\mathbf{x}}(w)$ and $Q_{\mathbf{y}}(w)$ using the block structure, these indices will come in handy later.

Recall that the polynomial $Q_{\mathbf{x}}(w)$ is defined as

$$Q_{\mathbf{x}}(w) = x_0 + x_1 w + x_2 w^2 + \ldots + x_{n-1} w^{n-1}.$$

Given the block structure of $\mathbf{x}$ and $\mathbf{y}$, we can express $Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)$ in terms of the polynomials $Q_{\mathbf{x}_i}(w) - Q_{\mathbf{y}_i}(w)$ as

$$Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w) = \sum_{i=1}^{d} w^{t_i} \left(Q_{\mathbf{x}_i}(w) - Q_{\mathbf{y}_i}(w)\right).$$

For $i \in S_2$, we have

$$\begin{aligned} Q_{\mathbf{x}_i}(w) - Q_{\mathbf{y}_i}(w) &= w Q_{\mathbf{s}_i}(w) - Q_{\mathbf{s}_i}(w) \\ &= (w-1) Q_{\mathbf{s}_i}(w). \end{aligned}$$

Similarly for $i \in S_3$, we have

$$\begin{aligned} Q_{\mathbf{x}_i}(w) - Q_{\mathbf{y}_i}(w) &= Q_{\mathbf{s}_i}(w) - w Q_{\mathbf{s}_i}(w) \\ &= (1-w) Q_{\mathbf{s}_i}(w). \end{aligned}$$

Putting everything together, we get

$$\begin{aligned} &Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w) \\ =&\sum_{i \in S_2} w^{t_i}(w-1) Q_{\mathbf{s}_i}(w) + \sum_{i \in S_3} w^{t_i}(1-w) Q_{\mathbf{s}_i}(w) \\ =&(w-1)\left(\sum_{i \in S_2} w^{t_i} Q_{\mathbf{s}_i}(w) - \sum_{i \in S_3} w^{t_i} Q_{\mathbf{s}_i}(w)\right). \end{aligned}$$

Towards applying Lemma 1, we are going to upper bound the number of sign changes in the second term of the above expression. We note that $Q_{\mathbf{s}_i}(w)$ is a polynomial with 0/1 coefficients. Each summand $w^{t_i} Q_{\mathbf{s}_i}(w)$ contains a set of monomials whose degrees are in an interval $[t_i, t_{i+1})$, and all these intervals are disjoint from each other. It follows that the number of sign changes is at most d. The lemma also follows by Theorem 7.

Now let us turn to the case where a_i's and b_i's are not necessarily zero. Due to "linearity" of the mapping $\mathbf{x} \mapsto Q_{\mathbf{x}}$ (i.e. $\mathbf{x} + \mathbf{y} \mapsto Q_{\mathbf{x}} + Q_{\mathbf{y}}$), it will be helpful to write $\mathbf{x} = \mathbf{x}_\emptyset + \mathbf{x}_\Delta$, where $\mathbf{x}_\emptyset$ is $\mathbf{x}$ but with all the a_i's replaced by zero, and $\mathbf{x}_\Delta$ contains only the a_i's. Similarly write $\mathbf{y} = \mathbf{y}_\emptyset + \mathbf{y}_\Delta$.

We recall that $\mathbf{x}$ and $\mathbf{y}$ have the same Hamming weight. That means the the following two sets

$$A = \{i \colon a_i = 1\} \text{ and } B = \{i \colon b_i = 1\}$$

have the same cardinality. Let $\pi \colon A \to B$ be a matching between the A and B. For $i \in A$ let $\sigma(i)$ be the index of a_i in $\mathbf{x}$ and let $\tau(i)$ be the index of $b_{\pi(i)}$ in $\mathbf{y}$. It follows that $\sigma(i) = t_i$ or $t_{i+1} - 1$, and $\tau(i) = t_{\pi(i)}$ or $t_{\pi(i)+1} - 1$, and that

$$Q_{\mathbf{x}_\Delta}(w) - Q_{\mathbf{y}_\Delta}(w) = \sum_{i \in A}\left(w^{\sigma(i)} - w^{\tau(i)}\right).$$

We then have

$$\begin{aligned} &Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w) \\ =&\left(Q_{\mathbf{x}_\emptyset}(w) - Q_{\mathbf{y}_\emptyset}(w)\right) + \left(Q_{\mathbf{x}_\Delta}(w) - Q_{\mathbf{y}_\Delta}(w)\right) \\ =&\sum_{i \in S_2} w^{t_i}(w-1) Q_{\mathbf{s}_i}(w) + \sum_{i \in S_3} w^{t_i}(1-w) Q_{\mathbf{s}_i}(w) \\ &\qquad + \sum_{i \in A}\left(w^{\sigma(i)} - w^{\tau(i)}\right) \\ =&(w-1)\left(\sum_{i \in S_2} w^{t_i} Q_{\mathbf{s}_i}(w) - \sum_{i \in S_3} w^{t_i} Q_{\mathbf{s}_i}(w) + \sum_{i \in A} J_i(w)\right), \end{aligned}$$

where each $J_i(w)$ is a polynomial of the form

$$J_i(w) = \begin{cases} w^{\tau(i)} + w^{\tau(i)+1} + \ldots + w^{\sigma(i)-1} & \text{if } \sigma(i) > \tau(i), \\ -w^{\sigma(i)} - w^{\sigma(i)+1} - \ldots - w^{\tau(i)-1} & \text{if } \sigma(i) < \tau(i). \end{cases}$$

Let us focus on the polynomial $R(w) = R_1(w) + R_2(w)$ where

$$\begin{aligned} R_1(w) &\coloneqq \sum_{i \in S_2} w^{t_i} Q_{\mathbf{s}_i}(w) - \sum_{i \in S_3} w^{t_i} Q_{\mathbf{s}_i}(w), \\ R_2(w) &\coloneqq \sum_{i \in A} J_i(w). \end{aligned}$$

Once more we are going to bound the number of sign changes in R. Fix an arbitrary i and consider two degrees $k_1 \neq k_2 \in [t_i + 1, t_{i+1} - 1)$. From previous discussions we know that w^{k_1} and w^{k_2} have the same sign in R_1. For R_2, we note that for each $j \in A$, w^{k_1} and w^{k_2} have the same coefficients in J_j. This is because the coefficients of J_j are identically 1 (or -1) in the degree interval $[\tau(i), \sigma(i))$ (or $[\sigma(i), \tau(i))$), which either contains or is disjoint with $\{k_1, k_2\}$. Therefore, the coefficients of w^{k_1} and w^{k_2} are the same in R_2. Finally, notice that the coefficients of R_1 belong to $\{0, 1, -1\}$, and that the coefficients of R_2 are integers. Therefore w^{k_1} and w^{k_2} have the same sign in $R = R_1 + R_2$.

Given a sign change (i,j) in R (cf. the remark below Lemma 1), we say an index k *cuts* (i,j) if $i \le k \le j$. The above argument shows that any sign change in R must be cut by some index in the set

$$\begin{aligned} C &= \bigcup_{i=1}^{d} \{t_i, t_{i+1}-1\} \\ &= \{t_1\} \cup \bigcup_{i=2}^{d} \{t_i - 1, t_i\} \cup \{t_{d+1}-1\} . \end{aligned}$$

As $t_1 = 0$ and $t_{d+1} - 1 = n-1$ each cuts at most 1 sign change, and for each i, $t_i - 1$ and t_i jointly cut at most 3 sign changes, it follows that R has at most $3(d-1)+2 = 3d-1$ sign changes.

Now we can apply Lemma 1 and get that the multiplicity of zero of $Q_{\mathbf{x}} - Q_{\mathbf{y}}$ at 1 is at most $3d$ (1 from the factor $(w-1)$, $3d-1$ from $R(w)$). By Theorem 7, we conclude that

$$\|\mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{y})\|_{\ell_1} \ge \frac{q}{e}\left(\frac{q}{n}\right)^{3d} .$$

□

a) What happens to edit distance 2 pairs?: Given that at edit distance 4 there are already hard strings for mean-based algorithms, this is indeed a natural question to ask. In fact, we will show that a mean-based algorithm can distinguish between $\mathbf{x}$ and $\mathbf{y}$ using only polynomially many traces, and this will be an application of Theorem 10.

Corollary 1. *Let $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ be two arbitrary (distinct) strings with $d_E(\mathbf{x}, \mathbf{y}) = 2$. Then $n^{O(1)}$ traces are sufficient for a mean-based algorithm to distinguish between $\mathbf{x}$ and $\mathbf{y}$.*

Proof. As before we assume $\mathbf{x}$ and $\mathbf{y}$ have the same Hamming weight. Due to the symmetry between 0 and 1, we may assume without loss of generality that $\mathbf{x} = \mathbf{a}0\mathbf{bc}$ and $\mathbf{y} = \mathbf{ab}0\mathbf{c}$ for strings $\mathbf{a}, \mathbf{b}, \mathbf{c}$ with lengths a, b, c, respectively, satisfying $a+b+c+1 = n$. We thus have the block decompositions $\mathbf{x} = \mathbf{x}_1\mathbf{x}_2\mathbf{x}_3$ and $\mathbf{y} = \mathbf{y}_1\mathbf{y}_2\mathbf{y}_3$ where

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{a}, \quad \mathbf{x}_2 = \mathbf{b}0, \quad \mathbf{x}_3 = \mathbf{c}, \\ \mathbf{y}_1 &= \mathbf{a}, \quad \mathbf{y}_2 = 0\mathbf{b}, \quad \mathbf{y}_3 = \mathbf{c}. \end{aligned}$$

Note that this falls into the special strucutre mentioned above for $d = 3$. Theorem 10 then implies

$$\sum_{j=0}^{n-1} \left| \mathop{\mathbb{E}}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\mathbf{x}}} [\tilde{x}_j] - \mathop{\mathbb{E}}_{\tilde{\mathbf{y}} \sim \mathcal{D}_{\mathbf{y}}} [\tilde{y}_j] \right| \ge \frac{q}{e}\left(\frac{q}{n}\right)^{9} ,$$

from which it follows that $n^{O(1)}$ traces are sufficient to distinguish between $\mathbf{x}$ and $\mathbf{y}$. □

In fact, with a more careful analysis one can nail down the constant and show that the sample complexity is $O(n^2)$, leading to a sharp transition in sample complexity from edit distance 2 to 4.

b) Other cases for edit distance 4: We also mention that the only hard pairs at edit distance 4 have the form

$$\begin{aligned} \mathbf{x} &= \mathbf{a}\ a_1\ \mathbf{b}\ a_2\ \mathbf{c}\ \mathbf{d}\ \mathbf{e}, \\ \mathbf{y} &= \mathbf{a}\ \mathbf{b}\ \mathbf{c}\ b_1\ \mathbf{d}\ b_2\ \mathbf{e}. \end{aligned}$$

The hard strings given in Theorem 2 are also in this form with $\mathbf{b} = \mathbf{d} = \varepsilon$ (the empty string). All other pairs not in this form will have the special structure mentioned earlier, and are thus easy.

VII. Conclusions and Open Problems

In this work we showed several results about the power and limitation of mean-based algorithms in distinguishing trace distributions of strings at small Hamming or edit distance.

Going beyond mean-based algorithms is obviously a major concern. A very natural next step is to incorporate "multi-bit statistics", namely the joint distribution of several bits of the traces. Indeed, the upper bound obtained in [20] is based on the joint distribution of roughly $n^{1/5}$ bits. Although this seems a much more general class of algorithms, the best bound they yield so far is still exponential. We leave as an open problem the power and limitation of algorithms based on multi-bit statistics.

Appendix A
Mean-based algorithms and connection to complex analysis

Fix a string $\mathbf{x} \in \{0,1\}^n$. The basic idea of [8] and [7] is to consider the average number of "1"s at index j in the traces of $\mathbf{x}$, i.e. the expectations $E_j(\mathbf{x}) := \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\mathbf{x}}}[\tilde{x}_j]$ for $j = 0, 2, \ldots, n-1$, where $\tilde{x}_j = 0$ for $j > |\tilde{\mathbf{x}}| - 1$. An algorithm is said to be *mean-based* if its output depends only on the statistical estimates of $E_j(\mathbf{x})$ where $j = 0, 1, \ldots, n-1$.

A. The reduction to complex analysis

[8] and [7] showed the following bound.

Theorem 11 ([8], [7]). *For all distinct $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ it is the case that*

$$\sum_{j=0}^{n-1} |E_j(\mathbf{x}) - E_j(\mathbf{y})| > \exp\left(-O\left(n^{1/3}\right)\right). \quad (2)$$

This result is sufficient to imply that $\exp\left(O(n^{1/3})\right)$ samples can tell the difference between $\mathcal{D}_{\mathbf{x}}$ and $\mathcal{D}_{\mathbf{y}}$ with high probability. To this end, they defined the following polynomial

$$P_{\mathbf{x}}(z) = \sum_{j=0}^{n-1} E_j(\mathbf{x}) \cdot z^j.$$

This makes the left-hand-side of (2) simply $\|P_{\mathbf{x}} - P_{\mathbf{y}}\|_1$. By writing explicitly

$$\begin{aligned} E_j(\mathbf{x}) &= \sum_{k=0}^{n-1} \Pr\left[\tilde{x}_j \text{ comes from } x_k\right] \cdot x_k \\ &= \sum_{k=0}^{n-1} \binom{k}{j} p^{k-j} q^{j+1} \cdot x_k, \end{aligned}$$

we have that

$$\begin{aligned} P_{\mathbf{x}}(z) &= \sum_{j=0}^{n-1} E_j(\mathbf{x}) \cdot z^j \\ &= \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} \binom{k}{j} p^{k-j} q^{j+1} \cdot x_k \cdot z^j \\ &= q \sum_{k=0}^{n-1} x_k \sum_{j=0}^{n-1} \binom{k}{j} p^{k-j} (qz)^j \\ &= q \sum_{k=0}^{n-1} x_k \cdot (p+qz)^k \\ &= q \cdot Q_{\mathbf{x}}(p+qz). \end{aligned}$$

In light of Lemma 2, one might as well bound $\|P_{\mathbf{x}} - P_{\mathbf{y}}\|_\infty$. Keeping in mind that the map $z \mapsto p+qz$ shifts the complex unit circle $\partial B(0;1)$ to $\partial B\left(p;q\right)$, so far we have reduced the problem to understanding the following supremum

$$\sup\left\{\left|Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)\right| : w \in \partial B\left(p;q\right)\right\}.$$

Using a result of [21], [8] and [7] proved that the above supremum is at least $\exp\left(-O(n^{1/3})\right)$, which is their main technical result.

To summarize, we have the following generic lemma.

Lemma 8. *Let* $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ *be two strings. Then*

$$\begin{aligned} &\frac{1}{\sqrt{n+1}} \|P_{\mathbf{x}} - P_{\mathbf{y}}\|_1 \\ \leq& q \cdot \sup\left\{\left|Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)\right| : w \in \partial B\left(p;q\right)\right\} \\ \leq& \|P_{\mathbf{x}} - P_{\mathbf{y}}\|_1. \end{aligned}$$

Proof. We have that $q^{-1} P_{\mathbf{x}}(z) = Q_{\mathbf{x}}(w)$ where $w = p + qz$. Applying Lemma 2 to the polynomial $q^{-1}\left(P_{\mathbf{x}} - P_{\mathbf{y}}\right)$ gives the lemma. □

A common bound in this paper is of the form

$$\|P_{\mathbf{x}} - P_{\mathbf{y}}\|_1 \geq n^{-O(d)}$$

for some parameter d. A standard Chernoff-Hoeffding bound argument shows that $n^{O(d)}$ traces are sufficient for a mean-based algorithm to distinguish between $\mathbf{x}$ and $\mathbf{y}$. On the other hand, if for some strings $\mathbf{x}$ and $\mathbf{y}$ one can show

$$\|P_{\mathbf{x}} - P_{\mathbf{y}}\|_1 \leq \varepsilon,$$

then it is the case that $\Omega(1/\varepsilon)$ traces are required for distinguishing between $\mathbf{x}$ and $\mathbf{y}$ by mean-based algorithms. For a formal discussion about the sample complexity versus various notions of distances related to the trace problem we refer the reader to [11].

Appendix B
Distinguishing between strings within small Hamming distance

We prove Theorem 1 in this section. We remark that the same result was proven in [15], which uses a previous result regarding reconstructing strings from their "k-decks" (i.e. the multi-set of subsequences of length k)[22]. One of the results in [22] states that strings within Hamming distance $2k$ have different k-decks. Therefore when the deletion probability $p \leq 1 - k/n$, the traces will have length at least k in expectation and we can reconstruct the k-deck with high probability in $n^{O(k)}$ traces. This is exactly the argument in [15], but we note here that this argument does not yield a mean-based algorithm.

Theorem 1 states that the same task can be accomplished also by mean-based algorithms. With the machinery established in this paper, this will be an immediate consequence of Descartes' rule of sign changes (Lemma 1).

Theorem 1. *Let* $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ *be two distinct strings within Hamming distance* d *from each other. There is a mean-based algorithm that distinguishes between* $\mathbf{x}$ *and* $\mathbf{y}$ *with high probability using* $n^{O(d)}$ *traces.*

Proof. Let $Q(w) = Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)$. We will show that the multiplicity of zero of Q at 1 is at most d.

We note that $Q(w)$ is a polynomial with at most d non-zero terms. Therefore the number of sign changes $C(Q)$ can never exceed d. By Lemma 1, the number of real positive roots of Q is at most d. In particular, the multiplicity of zero of Q at 1 is at most d. Thus by Theorem 7 we have

$$\|\mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{y})\|_{\ell_1} \geq \frac{q}{e}\left(\frac{q}{n}\right)^d.$$

The sample complexity bound $n^{O(d)}$ follows from Proposition 1. □

Acknowledgements

We are indebted to some anonymous reviewers for pointing us to several references that we previously missed and for many useful suggestions that have been incorporated in the current writeup.

References

[1] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004*. SIAM, 2004, pp. 910–918.

[2] S. Kannan and A. McGregor, "More on reconstructing strings from random traces: insertions and deletions," in *IEEE International Symposium on Information Theory, ISIT 2005*. IEEE, 2005, pp. 297–301.

[3] K. Viswanathan and R. Swaminathan, "Improved string reconstruction over insertion-deletion channels," in *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*. SIAM, 2008, pp. 399–408.

[4] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*. SIAM, 2008, pp. 389–398.

[5] A. McGregor, E. Price, and S. Vorotnikova, "Trace reconstruction revisited," in *22th Annual European Symposium on Algorithms, ESA 2014*, ser. Lecture Notes in Computer Science, vol. 8737. Springer, 2014, pp. 689–700.

[6] Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice," in *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*. IEEE Computer Society, 2017, pp. 228–239.

[7] F. Nazarov and Y. Peres, "Trace reconstruction with $\exp(O(n^{1/3}))$ samples," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017*. ACM, 2017, pp. 1042–1046.

[8] A. De, R. O'Donnell, and R. A. Servedio, "Optimal mean-based algorithms for trace reconstruction," *The Annals of Applied Probability*, vol. 29, no. 2, pp. 851–874, 2019.

[9] R. Gabrys and O. Milenkovic, "The hybrid k-deck problem: Reconstructing sequences from short and long traces," in *IEEE International Symposium on Information Theory, ISIT 2017*. IEEE, 2017, pp. 1306–1310.

[10] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability," in *Conference On Learning Theory, COLT 2018*, ser. Proceedings of Machine Learning Research, vol. 75. PMLR, 2018, pp. 1799–1840.

[11] N. Holden and R. Lyons, "Lower bounds for trace reconstruction," *The Annals of Applied Probability*, vol. 30, no. 2, pp. 503–525, 2020.

[12] L. Hartung, N. Holden, and Y. Peres, "Trace reconstruction with varying deletion probabilities," in *Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics, ANALCO 2018*. SIAM, 2018, pp. 54–61.

[13] R. Gabrys and O. Milenkovic, "Unique reconstruction of coded strings from multiset substring spectra," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 7682–7696, 2019.

[14] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6084–6103, 2020.

[15] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Trace reconstruction: Generalized and parameterized," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3233–3250, 2021.

[16] J. Brakensiek, R. Li, and B. Spang, "Coded trace reconstruction in a constant number of traces," in *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*. IEEE, 2020, pp. 482–493.

[17] X. Chen, A. De, C. H. Lee, R. A. Servedio, and S. Sinha, "Polynomial-time trace reconstruction in the smoothed complexity model," in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021*. SIAM, 2021, pp. 54–73.

[18] Z. Chase, "Separating words and trace reconstruction," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2021*. ACM, 2021, pp. 21–31.

[19] S. Narayanan and M. Ren, "Circular trace reconstruction," in *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[20] Z. Chase, "New lower bounds for trace reconstruction," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 57, no. 2. Institut Henri Poincaré, 2021, pp. 627–643.

[21] P. Borwein and T. Erdélyi, "Littlewood-type problems on subarcs of the unit circle," *Indiana University mathematics journal*, pp. 1323–1346, 1997.

[22] I. Krasikov and Y. Roditty, "On a reconstruction problem for sequences,," *J. Comb. Theory, Ser. A*, vol. 77, no. 2, pp. 344–348, 1997.

[23] A. D. Scott, "Reconstructing sequences," *Discrete Mathematics*, vol. 175, no. 1-3, pp. 231–238, 1997.

[24] E. Prouhet, "Mémoire sur quelques relations entre les puissances des nombres," *CR Acad. Sci. Paris*, vol. 33, no. 225, p. 1851, 1851.

[25] L. E. Dickson, *History of the Theory of Numbers, Volume II: Diophantine Analysis*. Courier Corporation, 2013, vol. 2.

[26] E. M. Wright, "Prouhet's 1851 solution of the Tarry-Escott problem of 1910," *The American Mathematical Monthly*, vol. 66, no. 3, pp. 199–201, 1959.

[27] P. Erdos and G. Szekeres, "On the product n k= 1 (1- zak), acad," *Serbe Sci. Publ. Inst. Math*, vol. 13, pp. 29–34, 1959.

[28] P. Borwein and C. Ingalls, "The Prouhet-Tarry-Escott problem revisited," *Enseign. Math*, vol. 40, pp. 3–27, 1994.

[29] E. M. Wright, "On Tarry's problem (I)," *The Quarterly Journal of Mathematics*, no. 1, pp. 261–267, 1935.

[30] L. K. Hua, *Introduction to number theory*. Springer, 1982.

[31] V. Gandikota, B. Ghazi, and E. Grigorescu, "Np-hardness of reed-solomon decoding, and the prouhet-tarry-escott problem," *SIAM J. Comput.*, vol. 47, no. 4, pp. 1547–1584, 2018.

[32] R. Descartes, *La géométrie*. Hermann, 1886.

[33] E. Grigorescu, M. Sudan, and M. Zhu, "Limitations of mean-based algorithms for trace reconstruction atsmall distance," *arXiv preprint arXiv:2011.13737v1*, 2020.

[34] J. Sima and J. Bruck, "Trace reconstruction with bounded edit distance," in *IEEE International Symposium on Information Theory, ISIT 2021*. IEEE, 2021, pp. 2519–2524.

[35] G. Pólya and G. Szegö, *Problems and Theorems in Analysis II: Theory of Functions. Zeros. Polynomials. Determinants. Number Theory. Geometry*. Springer Science & Business Media, 1997.

[36] T. Erdélyi, "Coppersmith–Rivlin type inequalities and the order of vanishing of polynomials at 1," *Acta Arithmetica*, vol. 172, pp. 271–284, 2014.

[37] ——, "On the multiplicity of the zeros of polynomials with constrained coefficients," *Approximation Theory and Analytic Inequalities*, 2020.

[38] V. I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences," *J. Comb. Theory, Ser. A*, vol. 93, no. 2, pp. 310–332, 2001.

[39] ——, "Efficient reconstruction of sequences," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 2–22, 2001.

[40] M. Cheraghchi, J. Downs, J. L. Ribeiro, and A. Veliche, “Mean-based trace reconstruction over practically any replication-insertion channel,” in *IEEE International Symposium on Information Theory, ISIT 2021*. IEEE, 2021, pp. 2459–2464.

[41] F. Ban, X. Chen, A. Freilich, R. A. Servedio, and S. Sinha, “Beyond trace reconstruction: Population recovery from the deletion channel,” in *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019*. IEEE, 2019, pp. 745–768.

[42] S. Davies, M. Z. Rácz, B. G. Schiffer, and C. Rashtchian, “Approximate trace reconstruction: Algorithms,” in *IEEE International Symposium on Information Theory, ISIT 2021*. IEEE, 2021, pp. 2525–2530.

[43] X. Chen, A. De, C. H. Lee, R. A. Servedio, and S. Sinha, “Near-optimal average-case approximate trace reconstruction from few traces,” *arXiv preprint arXiv:2107.11530*, 2021, (To appear in SODA 2022).

[44] Z. Chase and Y. Peres, “Approximate trace reconstruction of random strings from a constant number of traces,” *arXiv preprint arXiv:2107.06454*, 2021.

[45] D. Chakraborty, D. Das, and R. Krauthgamer, “Approximate trace reconstruction via median string (in average-case),” in *41st IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2021*, ser. LIPIcs, vol. 213, 2021, pp. 11:1–11:23.

[46] S. Lang, *Complex analysis*. Springer Science & Business Media, 2013, vol. 103.

[47] P. Borwein, T. Erdélyi, and G. Kós, “Littlewood-type problems on $[0, 1]$,” *Proceedings of the London Mathematical Society*, vol. 79, no. 1, pp. 22–46, 1999.

Elena Grigorescu Elena Grigorescu is an associate professor in the Computer Science department at Purdue University, which she joined in 2012. Before that, she graduated from MIT in 2010, and was a postdoctoral fellow at Georgia Tech 2010-2012.

Her research interests lie broadly in theoretical computer science, with a particular emphasis on sublinear algorithms, complexity theory, coding theory and learning theory. Her recent work has focused on designing algorithms that only use sublinear time or sublinear space, in computational models relevant to large data sets; on understanding the complexity of fundamental problems on error-correcting codes and point lattices, with applications to communications, cryptography and optimizations; and on understanding information-theoretical limits in diverse computational models.

Madhu Sudan Madhu Sudan is a Gordon McKay Professor in the John A. Paulson School of Engineering and Applied Sciences at Harvard University, where he has been since 2015. Madhu Sudan got his Bachelors degree from IIT Delhi in 1987 and his Ph.D. from U.C. Berkeley in 1992. Between 1992 and 2015, Madhu Sudan has been worked at IBM Research, MIT and Microsoft Research. Madhu Sudan is a recipient of the Nevanlinna Prize, the Infosys Foundation Prize in Mathematical Sciences, and the IEEE Hamming Medal. Madhu Sudan is a fellow of the ACM, the IEEE and the AMS. He is a member of the American Academy of Arts and Sciences and the National Academy of Sciences.

Madhu Sudan's research interests revolve around mathematical studies of communication and computation. Specifically his research focusses on concepts of reliability and mechanisms that are, or can be, used by computers to interact reliably with each other. His research draws on tools from computational complexity, which studies efficiency of computation, and many areas of mathematics including algebra and probability theory.

Minshen Zhu Minshen Zhu is a PhD student at Purdue University. His research interests lie in coding theory, particularly locally decodable codes, the trace reconstruction problem, and combinatorial problems on graphs.