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

Elena Grigorescu*, Madhu Sudan†, Minshen Zhu*
*Department of Computer Science, Purdue University. {elena-g, zhu628}@purdue.edu
†School of Engineering and Applied Sciences, Harvard University. madhu@cs.harvard.edu

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

**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.**

*A full version of this paper is accessible at:* https://arxiv.org/abs/2011.13737

## I. Introduction

In the trace reconstruction problem, a string $x \in \{0,1\}^n$ is sent over a deletion channel which deletes each entry independently with probability $p \in [0,1)$, resulting in a *trace* $\tilde{x} \in \{0,1\}^\ell$ of smaller length. The goal is to reconstruct $x$ exactly from a small set of independent traces. The trace reconstruction problem was introduced by Batu et al. [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 $x, y \in \{0,1\}^n$ 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 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 $x = (01)^k 101 (01)^k$ and $y = (01)^k 011 (01)^k$). 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

E.G and M.Z. were supported in part by NSF CCF-1910659 and NSF CCF-1910411. M.S. was supported in part by a Simons Investigator Award and NSF Award CCF 1715187.

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]. However, these are existential results, revealing little explicit structure that may help in further analysis.

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 $\tilde{x}_i$ over the sample set $S$ of independent traces, namely $E_S(\tilde{x}_i)$. While they appear restrictive, mean-based algorithms are in fact a very powerful class of algorithms – the upper bounds of [8], [7] are obtained via mean-based algorithms.

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

Our main results here prove that there exist *explicit* strings $x, y \in \{0,1\}^n$ at edit distance only 4 for which every mean-based algorithm requires a super-polynomial in $n$ number of samples. 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 of explicit hard-to-distinguish strings will eventually lead to a better understanding of the complexity of the trace reconstruction problem.

### 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 which 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. 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 the full paper [24] 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 $p = 1/2$. 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) [25], [26] problem from number theory. In the PTE problem, given an integer $k \geq 0$ one would like to find integer solutions $\{x_1, x_2, \dots, x_s\}$, and $\{y_1, y_2, \dots, y_s\}$ to the system $\sum_{i\in[s]} x_i^j = \sum_{i\in[s]} y_i^j$, for all $j \in [k]$, with $x_i \neq y_j$ for all $i, j \in [s]$. The goal is to find such solutions with $s$ as small as possible compared to $k$. 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.

As a consequence of our results we show that explicit strings that are exponentially hard to distinguish by mean-based algorithms imply explicit solutions of small size to a PTE system, which is the main open problem in the study of the PTE problem.

**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 $\exp\left(\Omega(n^\varepsilon)\right)$ 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 $n = (k \log k)^{2/\varepsilon}$.*

We also note that weak versions of our results follow from simple applications of the Descartes rule of signs. As an application of this rule to larger edit distances we also obtain the following theorem.

**Theorem 4.** *(Informal) Strings $x, y \in \{0,1\}^n$ with $d_E(x, y) = d \geq 1$ with some special block structures are distinguishable by mean-based algorithms running in time $n^{O(d^2)}$. In particular, the statement holds for every pair of strings at edit distance 2.*

We defer all missing proofs, discussions, and further related work to the full version of this paper [24].

### B. Our techniques

*a) The [8], [7] reduction to complex analysis:* Our techniques focus on analyzing the modulus of Littlewood polynomials with $\{-1,0,1\}$ coefficients over a shifted unit circle. The reduction to complex analysis was established in [8], [7]. In particular, since a mean-based algorithm only works with $E(\tilde{x}_i)$ for each $i$, 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

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

Here $p$ is the deletion probability and $q = 1-p$. Then [8], [7] show that the sample complexity upper bound for mean-based algorithms is roughly (up to squaring) the inverse of $\sup\{|Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)| : w \in \partial B(p;q)\}$, where $\partial B(p;q)$ is the complex unit circle shifted to the complex circle of radius $q$ centered at $p$. Hence, the problem reduces to understanding the maximum modulus of a polynomial on a shifted circle. 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 supremum of $|Q(w)|$ over the shifted complex circle to the multiplicity of the root $w=1$ of the polynomial $Q(w)$. Specifically, we show that as long as $w=1$ is a root with multiplicity no more than $k$, the supremum over the shifted complex circle is at least $n^{-O(k^2)}$.

**Theorem 5.** *Let $f(z)$ be a polynomial of degree $n$ with coefficients in $\{-1,0,1\}$. Suppose $z=1$ is a root of $f(z)$ with multiplicity at most $k$. Let $0<p<1$ and $q=1-p$. Then*

$$\sup\{|f(z)| : z \in \partial B(p;q)\} \geq \frac{1}{2}(4n^{k+2})^{-k} = n^{-O(k^2)}.$$

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** (Descartes' Theorem)**.** *[27] 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 $C(p) \geq Z(p)$.*

*Remark:* If $p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$ is a polynomial, we say a pair $(i,j)$ $(0 \leq i < j \leq n)$ is a *sign change* if $a_ia_j < 0$ and $a_{i+1} = a_{i+2} = \ldots = a_{j-1} = 0$. $C(p)$ exactly counts the number of such pairs $(i,j)$.

We note that prior work that we are aware of on understanding the structure of polynomials with many roots at 1 (e.g., [28], [29]) do not appear to imply our bounds on the complex unit circle.

We use this rule to prove the formal version of Theorem 4, and a weaker version of Theorem 1. See the full version.

*c) Complex analysis over shifted circles: the strong version of our results:* A main stepping stone in our derivation of Theorem 1 and Theorem 4 is that low multiplicity of the root $w=1$ implies large supremum over the shifted complex circle (i.e. Theorem 5). The idea is to find a point on the shifted circle that has a very small (but carefully chosen) distance to 1, and argue that the polynomial has large modulus at that point. The analysis is mostly elementary, including manipulations and estimations of trigonometric functions, and deriving bounds on the coefficients and derivatives of certain polynomials.

For the negative result (i.e. Theorem 2), the difficulty comes from the fact that the converse of Theorem 5 is not true. In fact, the strongest barrier is the fact that constructing small-size solutions to high-degree PTE systems is a well-known difficult problem in number theory. Previous non-constructive methods do not appear to be the correct tools either, because of the edit distance constraint, which cannot be accounted for in their arguments. Our construction is inspired by properties of product of cyclotomic polynomials and their relation to PTE solutions with special structures.

## 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 $p(w) = a_0 + a_1w + \ldots + a_nw^n$ be a polynomial where the coefficients are real. 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 $p$ 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 supremum from low multiplicity of root 1

Fix two strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ and denote $f(w) = Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)$. We are interested in $\sup|f(w)|$ where $w$ is on the circle $\partial B\,(p;q)$. It turns out that we can lower bound this supremum by upper bounding the multiplicity of zero of $f(w)$ at 1.

**Lemma 3** (Lemma 5.4 of [30]). *Suppose*

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

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

*Then* $\sum_{j=0}^{n-k} |b_j| \leq (n+1)\left(\frac{en}{k}\right)^k$.

**Lemma 4.** *Let $f(z) = \sum_{k=0}^{n} a_k z^k$ be a polynomial. Let $u(t) = \operatorname{Re} f(z(t))$ be a real-valued function, where $z(t) = p + qe^{it}$. Then*

$$|u'(t)| \leq q \cdot \sum_{k=0}^{n} k\,|a_k| \leq qn \cdot \sum_{k=0}^{n} |a_k| .$$

We can now prove Theorem 5, which will be useful to obtain several of our results in their weaker form. The proof also works for the more general class of polynomials with bounded integer coefficients.

*Proof of Theorem 5.* We can write $f(z) = (z-1)^k \cdot g(z)$ where $g(z)$ is a polynomial and $g(1) \neq 0$. It is not hard to see that $g(z)$ is a polynomial with integer coefficients, hence $g(1)$ is an integer and $|g(1)| \geq 1$.

Since the coefficients of $f$ are absolutely bounded by 1, by Lemma 3 the absolute values of the coefficients of $g(z)$ sum up to at most $n^{k+1}$.

Define a real-valued function $u(t) = \operatorname{Re} g\left(p + qe^{it}\right)$. We have $|u(0)| = |\operatorname{Re} g(1)| = |g(1)| \geq 1$. Lemma 3 and Lemma 4 together give the bound $|u'(t)| \leq q \cdot n^{k+2}$. Now we take $\theta = 1/2qn^{k+2}$ and $z = p + qe^{i\theta}$ so that

$$|z-1| = q\left|e^{i\theta} - 1\right| = 2q \sin\frac{\theta}{2} > \frac{q\theta}{2} = \frac{1}{4n^{k+2}}.$$

Here we uses the fact $\sin x > x/2$ for small $x$. The Mean Value Theorem implies that for some $\tilde{t} \in (0,\theta)$ we have

$$\begin{aligned} |g(z)| &\geq |\operatorname{Re} g(z)| = |u(\theta)| \geq |u(0)| - \theta\left|u'\left(\tilde{t}\right)\right| \\ &\geq 1 - \frac{1}{2qn^{k+2}} \cdot qn^{k+2} = \frac{1}{2}. \end{aligned}$$

Overall we have

$$|f(z)| = |z-1|^k \cdot |g(z)| \geq \left(\frac{1}{4n^{k+2}}\right)^k \cdot \frac{1}{2} = n^{-O(k^2)}.$$

Since $z \in \partial B\,(p;q)$, $\sup\{|f(z')| : z' \in \partial B\,(p;q)\} \geq |f(z)| \geq n^{-O(k^2)}$. □

### *A. Connection to the Prouhet-Tarry-Escott problem*

The following is a classical statement about the PTE problem.

**Theorem 6.** *(e.g. [31]) 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 $\exp\left(\Omega(n^\varepsilon)\right)$ 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* $n = (k \log k)^{2/\varepsilon}$.

## IV. 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.

Before proving Theorem 2, we need the following lemma about estimation of trigonometric functions.

**Lemma 5.** *The following bounds hold:*

1) *For integer* $d \geq 1$ *and* $\varphi$ *such that* $|d\varphi| \leq \frac{\pi}{3}$*, we have* $\cos(d\varphi) \leq (\cos\varphi)^d$.
2) *Let integer* $d \geq 1$ *and* $\theta$ *be such that* $\left|d \cdot \frac{\theta}{2}\right| < \frac{\pi}{3}$*. Let* $w = \left(1 + e^{i\theta}\right)/2$*. Then* $\left|w^d - 1\right| < 2\left|\sin\left(\frac{d\theta}{4}\right)\right|$.

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

**Theorem 7.** *Assume the deletion probability* $p = 1/2$*. 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* $Q_{\mathbf{e}}(w) := R(w) + E_n(w)$ *is a 0/1-coefficient polynomial which corresponds to a string* $\mathbf{e} \in \{0,1\}^n$*. 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* $n$.

*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 $R(w) + E_n(w)$ is a polynomial with 0/1 coefficients.

We can write

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

Consider a point $w = \left(1 + e^{i\theta}\right)/2$ on the circle $\partial B(1/2; 1/2)$, where $\theta \in [-\pi, \pi]$. We consider two cases.

1) $|\theta/2| \geq 3^{-k/4}\pi$. Using the bound $|\cos x| \leq 1 - (2x/\pi)^2$ for $|x| \leq \pi/2$, we have in this case $|w| = \left|\cos\frac{\theta}{2}\right| \leq 1 - 4 \cdot 3^{-k/2}$, and

$$\begin{aligned}
|P(w)| &\leq |w|^n \cdot 2(n+1) \leq \left(1 - 4 \cdot 3^{-k/2}\right)^n \cdot 2(n+1) \\
&\leq \exp\left(-\Omega\left(\sqrt{n}\right)\right).
\end{aligned}$$

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

2) $|\theta/2| < 3^{-k/4}\pi$. Note that when $j \leq k/4 - 1$, $|3^j \cdot \frac{\theta}{2}| < \frac{\pi}{3}$. Therefore by item (2) of Lemma 5, we have $|w^{3^j} - 1| \leq 2\left|\sin\left(3^j\theta/4\right)\right|$. Using the fact that $|\sin x| \leq |x|$, we have

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

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

□

## V. 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.

Many open questions remain, besides whether the worst-case reconstruction problem is solvable in polynomial time. We state below a few questions stemming from our work here.

**Problem 1** Can the connection between the multiplicity of zero at 1 and the supremum over the circle $\partial B(p;q)$ be tightened? Specifically, is the bound in Theorem 5 tight?

**Problem 2** Is the converse of Theorem 5 true? Namely, does high multiplicity of zero at 1 necessarily imply a small supremum over $\partial B(p;q)$? An affirmative answer to this question would establish an equivalence between PTE solutions and hard instances against mean-based trace reconstruction.

# References

[1] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *SODA*, 2004, pp. 910–918.

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

[3] K. Viswanathan and R. Swaminathan, "Improved string reconstruction over insertion-deletion channels," in *SODA*, 2008, p. 399–408.

[4] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *SODA*, 2008, pp. 389–398.

[5] A. McGregor, E. Price, and S. Vorotnikova, "Trace reconstruction revisited," in *ESA*, 2014, pp. 689–700.

[6] Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice," in *FOCS*, 2017, pp. 228–239.

[7] F. Nazarov and Y. Peres, "Trace reconstruction with $\exp(O(n^{1/3}))$ samples," in *STOC*, 2017, pp. 1042–1046.

[8] A. De, R. O'Donnell, and R. A. Servedio, "Optimal mean-based algorithms for trace reconstruction," in *STOC*, 2017, pp. 1047–1056.

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

[10] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability," in *COLT*, vol. 75, 2018, pp. 1799–1840.

[11] N. Holden and R. Lyons, "Lower bounds for trace reconstruction," *Ann. Appl. Probab.*, vol. 30, no. 2, pp. 503–525, 2020.

[12] L. Hartung, N. Holden, and Y. Peres, "Trace reconstruction with varying deletion probabilities," in *ANALCO*, 2018, pp. 54–61.

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

[14] M. Cheraghchi, J. L. Ribeiro, R. Gabrys, and O. Milenkovic, "Coded trace reconstruction," in *ITW*, 2019, pp. 1–5.

[15] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Trace reconstruction: Generalized and parameterized," in *ESA*, vol. 144, 2019, pp. 68:1–68:25.

[16] J. Brakensiek, R. Li, and B. Spang, "Coded trace reconstruction in a constant number of traces," p. (to appear), 2020.

[17] X. Chen, A. De, C. H. Lee, R. A. Servedio, and S. Sinha, "Polynomial-time trace reconstruction in the smoothed complexity model," 2020.

[18] Z. Chase, "New upper bounds for trace reconstruction," 2020.

[19] S. Narayanan and M. Ren, "Circular trace reconstruction," 2020.

[20] Z. Chase, "New lower bounds for trace reconstruction," *arXiv preprint arXiv:1905.03031*, 2019.

[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. Grigorescu, M. Sudan, and M. Zhu, "Limitations of mean-based algorithms for trace reconstruction at small distance," *CoRR*, vol. abs/2011.13737, 2020. [Online]. Available: https://arxiv.org/abs/2011.13737

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

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

[28] T. Erdélyi, "Coppersmith-rivlin type inequalities and the order of vanishing of polynomials at 1," *arXiv preprint arXiv:1406.2560*, 2014.

[29] ——, "On the multiplicity of the zeros of polynomials with constrained coefficients," 2019.

[30] 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.

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