Trace Reconstruction with Bounded Edit Distance

Jin Sima and **Jehoshua Bruck**

Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA

***Abstract*—The trace reconstruction problem studies the number of noisy samples needed to recover an unknown string $\mathbf{x} \in \{0,1\}^n$ with high probability, where the samples are independently obtained by passing $\mathbf{x}$ through a random deletion channel with deletion probability p. The problem is receiving significant attention recently due to its applications in DNA sequencing and DNA storage. Yet, there is still an exponential gap between upper and lower bounds for the trace reconstruction problem. In this paper we study the trace reconstruction problem when $\mathbf{x}$ is confined to an edit distance ball of radius k, which is essentially equivalent to distinguishing two strings with edit distance at most k. It is shown that $n^{O(k)}$ samples suffice to achieve this task with high probability.**

I. Introduction

The trace reconstruction problem seeks to recover an unknown string $\mathbf{x} \in \{0,1\}^n$, given multiple independent noisy samples or traces of $\mathbf{x}$. In this paper, a noisy sample is obtained by passing $\mathbf{x}$ through a deletion channel, which randomly and independently deletes each bit of $\mathbf{x}$ with probability p. We are interested in how many samples are needed to recover $\mathbf{x}$ with high probability.

The trace reconstruction problem was introduced in [2] and proposed earlier in [17] under an adversarial setting. It has been receiving increased attention recently due to its application in DNA sequencing and DNA storage [3], [20]. Also, there are many significant results on trace reconstruction and its variants and generalizations, such as coding for trace reconstruction [9] and population recovery [1]. For average case trace reconstruction, where the reconstruction error probability is averaged over all choices of $\mathbf{x} \in \{0,1\}^n$, the state of the art upper and lower bounds on the number of samples are $exp(O(\log^{\frac{1}{3}}(n)))$ [14] and $\Omega(\frac{\log^{\frac{5}{2}}(n)}{(\log\log n)^7})$ [7] respectively.

Despite the progress for average cases, the trace reconstruction problem proved to be highly nontrivial in worst cases, where the reconstruction error probability goes to zero for arbitrary choice of $\mathbf{x}$. For small deletion probabilities, the work in [10] showed that polynomial number of samples suffice when $p \leq n^{-(\frac{1}{2}+\epsilon)}$ for some $\epsilon > 0$, improving the result in [2] for $p \leq n^{-(\frac{1}{3}+\epsilon)}$ and some $\epsilon > 0$. When the deletion probability becomes constant, there is still an exponential gap between the upper and lower bounds on the number of samples needed. The first achievable sample size for constant deletion probability p is $exp(\tilde{O}(n^{\frac{1}{2}}))$ [15], which was improved to $exp(O(n^{\frac{1}{3}}))$ in independent and simultaneous works [12] and [19]. Both [12] and [19] studied mean-based algorithms, which use single bit statistics in traces, for reconstruction. They showed that $exp(O(n^{\frac{1}{3}}))$ is the best sample size achieved by mean-based algorithms. A novel approach in [12] and [19] is to relate single-bit statistics to complex polynomial analysis, and borrow results from [5] on complex analysis. This approach was further developed in [8], where multi-bit statistics were considered. The current best upper bound on the sample size is $exp(\tilde{O}(n^{\frac{1}{5}}))$ [8] for $p \leq \frac{1}{2}$, while the best lower bound $\Omega(\frac{n^{\frac{3}{2}}}{\log^7 n})$ is orders of magnitude away from the upper bound.

While the general trace reconstruction problem is hard to solve, in this paper, we focus on a variant of the trace reconstruction problem with an edit distance constraint. Specifically, the goal is to recover the string $\mathbf{x}$ by using its noisy samples and additional information of a given string $\mathbf{y}$, which is known to be within a bounded distance from $\mathbf{x}$. The edit distance between two strings is commonly defined as the minimum number of deletions, insertions, or substitutions that transform one string into another. In this paper, we consider only deletion/insertion for convenience, as a substitution is an insertion followed by a deletion. We say that a string $\mathbf{x}$ is within edit distance k to a string $\mathbf{y}$, denoted as $\mathbf{x} \in \mathcal{B}_k(\mathbf{y})$, if $\mathbf{x}$ can be obtained from $\mathbf{y}$ after at most k deletions and k insertions. Note that the general trace reconstruction problem consider cases when $k = n$.

The setting considered in this paper arises in many practical scenarios in genome sequencing, where one needs to recover an individual genome sequence of a species, given a reference genome sequence that represents the species [24]. Normally, the genome sequences of a species share some similarity and most of them can be considered to be within a bounded edit distance from the reference genome. One example is the Human Genome Project, where a human reference genome is provided to study the difference between individual genomes. Complementary to the problem we consider, the work in [11] studied approximate trace reconstruction, which aims to find an estimate within a given edit distance to the true string. Note that such an estimate, together with an algorithm to distinguish two strings within edit distance k, establishes a solution to the general trace reconstruction problem.

As indicated in [12], [13], [15], [16], [19], the problem of worst case trace reconstruction is essentially equivalent to a hypothesis testing problem of distinguishing any two strings using noisy samples. More specifically, the sample complexity needed for trace reconstruction is at most $poly(n)$ times the sample complexity needed to distinguish arbitrary two strings. The same equivalence holds in our setting as well, where a reference string $\mathbf{y}$ is known and close to $\mathbf{x}$ in edit distance. Hence, for convenience, we consider the problem in the form of distinguishing any two strings $\mathbf{x} \in \{0,1\}^n$ and $\mathbf{y} \in \{0,1\}^n$

when $\mathbf{x}$ is within edit distance k to $\mathbf{y}$. One special case of the problem is to distinguish two stings within Hamming distance k, which was addressed in [16] and $n^{O(k)}$ sample complexity was achieved. Recently, an independent work [13] studied the limitations of mean-based algorithms (see [12] and [19]) in distinguishing two strings with bounded edit distance. It was shown that mean-based algorithms need at least $n^{O(\log n)}$ traces to distinguish two strings with edit distance of even 4. The paper [13] also showed that $n^{O(k^2)}$ suffices to distinguish two strings $\mathbf{x} \in \{0,1\}^n$ and $\mathbf{y} \in \{0,1\}^n$ with special block structures, if $\mathbf{x} \in \mathcal{B}_k(\mathbf{y})$. Yet, as pointed out in [11], it is an open problem whether $n^{O(k)}$ samples suffice to recover a string that is within edit distance k to a known string.

The main contribution of this paper is an affirmative answer to this question. We show that distinguishing two sequences within edit distance k needs at most $n^{O(k)}$ samples. The result is stated in the following.

Theorem 1. *Let $\mathbf{x} \in \{0,1\}^n$ and $\mathbf{y} \in \{0,1\}^n$ be two strings satisfying $\mathbf{x} \in \mathcal{B}_k(\mathbf{y})$. Then strings $\mathbf{x}$ and $\mathbf{y}$ can be distinguished with high probability, given $n^{O(k)}$ noisy samples, each obtained by passing $\mathbf{x}$ through a deletion channel with deletion probability $p \leq \frac{1}{2}$.*

The approach we take follows a similar method to that in [8], [12], [13], [19], in the sense that we derive bounds on multi-bit statistics through complex analysis of a special class of polynomials. Yet, the complex analysis in this paper differs from those in [8], [12], [13], [19] in the following two ways. Firstly, we make use of the fact that the polynomial is related to a number theoretic problem called the Prouhet-Tarry-Escott problem [4], which is also noted in [13]. This allows us to link the problem to our previous result on deletion codes [23], where we showed that two constrained strings can be distinguished using weighted sums of powers, which is similar in form to the Prouhet-Tarry-Escott problem. Secondly, to find the maximum value of the polynomial, we let the complex variable take values on a small circle around the point 1, while the work in [8], [12], [13], [19] analyze the complex polynomial on a unit circle. By doing this, we are able to improve the $n^{O(k^2)}$ bound in [13] to $n^{O(k)}$.

The rest of the paper is organized as follows. In Section II we provide an introduction to the techniques and the lemmas needed to prove Theorem 1. In Section III, the proof of Theorem 1 is given. Section IV presents the proof of a critical lemma on complex analysis. Section V concludes the paper.

II. Preliminaries

In this section we present a brief introduction to the techniques and key lemmas needed in proving Theorem 1. In the following, it is assumed that the deletion probability $p = \frac{1}{2}$. For cases when $p \leq \frac{1}{2}$, it suffices to pass the samples through another deletion channel with deletion probability $\frac{1-2p}{2(1-p)}$. For strings $\mathbf{x} \in \{0,1\}^n$ and $\mathbf{y} \in \{0,1\}^n$, let $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ and $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ denote the sample obtained by passing $\mathbf{x}$ and $\mathbf{y}$ through the deletion channel respectively. We have $\tilde{X}_i = \emptyset$ or $\tilde{Y}_j = \emptyset$ if i or j is larger than the length of $\tilde{X}$ or $\tilde{Y}$, respectively.

The techniques we use were originated in [12], [19], which presented the following identity

$$\mathbb{E}_{\tilde{X}}[\sum_{i=1}^{n} \tilde{X}_i (2z-1)^i] = \frac{1}{2}\sum_{i=1}^{n} x_i z^i \triangleq f^s_{\mathbf{x}}(z), \qquad (1)$$

for a sequence $\mathbf{x}$ and a complex number z. The identity (1) links the analysis of single bit statistics $\{E_{\tilde{X}}[\tilde{X}_i]\}_{i=1}^n$ to that of complex polynomials. As a result, a lower bound on the maximal difference between single bit statistics $\max_{1\leq i\leq n}|E_{\tilde{X}}[\tilde{X}_i] - E_{\tilde{Y}}[\tilde{Y}_i]|$ can be obtained through analyzing the maximal value of the polynomial $f^s_{\mathbf{x}}(z) - f^s_{\mathbf{y}}(z)$ on a unit disk, a problem referred to as Littlewood type problems and studied in [5], [6]. Generalizing the approach in [12], [19], the paper [8] presented multi-bit statistics counterpart of (1), stated in the following lemma.

Lemma 1. *[8] For any complex number z and sequences $\mathbf{x} \in \{0,1\}^n$ and $\mathbf{w} \in \{0,1\}^\ell$, we have*

$$\begin{aligned}&\mathbb{E}_{\tilde{X}}[2^\ell \sum_{1\leq i_1<\ldots<i_\ell\leq n} \mathbb{1}_{\tilde{X}_{i_j}=w_j,\forall j\in[\ell]}(2z-1)^{i_1}(-1)^{i_\ell-i_1-\ell+1}]\\ &= \sum_{i=1}^{n-\ell+1} \mathbb{1}_{\mathbf{x}_{i:i+\ell-1}=\mathbf{w}} z^i\\ &\triangleq f^m_{\mathbf{x},\mathbf{w}}(z), \qquad (2)\end{aligned}$$

where $[\ell] = \{1,\ldots,\ell\}$ and $i : i+\ell-1 = \{i,\ldots,i+\ell-1\}$. For any statement E, the number $\mathbb{1}_E = 1$ iff E holds true.

Similar to the arguments in [8], we prove Theorem 1 by analyzing the polynomial $f^m_{\mathbf{x},\mathbf{m}}(z) - f^m_{\mathbf{y},\mathbf{m}}(z)$ associated with the multi-bit statistics in (2). However, the way in which the polynomial is analyzed in this paper deviates from that in [8]. While the paper [8] taylored the complex analysis arguments in [6] to obtain improved bounds, in this paper, we exploit number theoretic properties of two strings $\mathbf{x}$ and $\mathbf{y}$ within edit distance k.

In our previous paper [23], we showed implicitly that the weighted sums of powers $\sum_{i=1}^n i^j x_i$, $j \in \{0,\ldots,t\}$ can be used to distinguish two constrained strings $\mathbf{x}$ and $\mathbf{y}$ within bounded edit distance. The following lemma makes this statement explicit. For a sequence $\mathbf{x} \in \{0,1\}^n$, let $\mathcal{R}_{n,k}$ denote the set of length n strings such that any two 1 entries in each string are separated by a 0 run of length at least $k-1$.

Lemma 2. *For distinct strings $\mathbf{x}, \mathbf{y} \in \mathcal{R}_{n,5k}$, if $\mathbf{x} \in \mathcal{B}_{5k}(\mathbf{y})$, then there exists an integer $m \in [12k+1]$ such that $\sum_{i=1}^n i^m x_i \neq \sum_{i=1}^n i^m y_i$.*

Proof. Suppose on the contrary, we have that $\sum_{i=1}^n i^m x_i = \sum_{i=1}^n i^m y_i$ for all $m \in [12k+1]$. Then, we have that

$$\sum_{i=1}^{n}(\sum_{j=1}^{i} j^m) x_i = \sum_{i=1}^{n}(\sum_{j=1}^{i} j^m) y_i \qquad (3)$$

for all $m \in \{0,\dots,12k\}$. This is because $\sum_{j=1}^{i} j^m$ is a weighted sum of $i^1,\dots,i^{m+1}$ for any $m \in \{0,\dots,12k+1\}$ (Faulhaber's formula). Next, we borrow a result from [23].

Proposition 1. *[23] For sequences $\mathbf{x},\mathbf{y} \in \mathcal{R}_{n,3k}$, if $\mathbf{y} \in \mathcal{B}_{3k}(\mathbf{x})$ and $\sum_{i=1}^{n}(\sum_{j=1}^{i} j^m)x_i = \sum_{i=1}^{n}(\sum_{j=1}^{i} j^m)y_i$ for $m \in \{0,\dots,6k\}$, then $\mathbf{x}=\mathbf{y}$.*

Note that $\mathcal{R}_{n,5k} \subseteq \mathcal{R}_{n,6k}$ and that $\mathcal{B}_{5k}(\mathbf{x}) \subseteq \mathcal{B}_{6k}(\mathbf{x})$. Since (3) holds, we apply Proposition 1 with $k=2k$ and conclude that $\mathbf{x}=\mathbf{y}$, which contradicts the fact that $\mathbf{x}$ and $\mathbf{y}$ are distinct. □

Interestingly, the following result from [4] connects the sums of powers of two sets of integers that appear in Lemma 2 to the number of roots of a polynomial at 1. It allows us to combine the number theoretic result with further complex analysis, which will be given in Lemma 6. The lemma can be proved by checking the i-th, $i \in [m]$, derivative of the polynomial $\sum_{i=1}^{s} z^{\alpha_i} - \sum_{i=1}^{t} z^{\beta_i}$ at point $z=1$.

Lemma 3. *[4] Let $\{\alpha_1,\dots,\alpha_s\}$ and $\{\beta_1,\dots,\beta_t\}$ be two sets of integers. The following are equivalent:*

(a) $\sum_{i=1}^{s}\alpha_i^j = \sum_{i=1}^{s}\beta_i^j$ for $j \in [m-1]$.
(b) $(z-1)^m | \sum_{i=1}^{s} z^{\alpha_i} - \sum_{i=1}^{t} z^{\beta_i}$.

Remark 1. *The problem of finding two sets of integers $\{\alpha_1,\dots,\alpha_s\}$ and $\{\beta_1,\dots,\beta_s\}$ satisfying the statement (a) is called the Prouhet-Tarry-Escott problem [4]. This connection between the Prouhet-Tarry-Escott problem and the analysis of polynomials was also used in [13] and implicitly in [18].*

Lemma 2 requires that the strings $\mathbf{x}$ and $\mathbf{y}$ are within $\mathcal{R}(n,5k)$, which does not hold in general. Following the same trick as in [8] and [23], we define an indicator vector as follows. For any sequences $\mathbf{x} \in \{0,1\}^n$ and $\mathbf{w} \in \{0,1\}^\ell$, let

$$\mathbb{1}_{\mathbf{w}}(\mathbf{x})_i \triangleq \begin{cases} 1, & \text{if } \mathbf{x}_{i:i+\ell-1} = \mathbf{w}, \\ 0, & \text{else.} \end{cases}$$

It can be seen that the polynomial $f^m_{\mathbf{x},\mathbf{w}}(z)$ related to multi-bit statistics is exactly the polynomial $f^s_{\mathbb{1}_{\mathbf{w}}(\mathbf{x})}(z)$ related to single-bit statistics. To apply Lemma 2, we need to find a $\mathbf{w}$ such that $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{R}(n,5k)$. The same as what the paper [8] did, we find such a $\mathbf{w}$ by using the following lemma from [22]. A string $\mathbf{w} \in \{0,1\}^\ell$ is said to have period a, if and only if $w_i = w_{i+a}$ for $i \in [\ell-a]$. Moreover, a string $\mathbf{w} \in \{0,1\}^\ell$ is said to be non-periodic, iff $\mathbf{w}$ does not have period a for $a \in [\lceil\frac{\ell}{2}\rceil - 1]$.

Lemma 4. *For any sequences $\mathbf{w} \in \{0,1\}^{2p-1}$, either $(\mathbf{w},0)$ or $(\mathbf{w},1)$ is non-periodic, where $(\mathbf{w},0)$ and $(\mathbf{w},1)$ is the string obtained by appending 0 and 1 to $\mathbf{w}$, respectively.*

Lemma 4 can be proved by definition of period. The claim that $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{R}(n,p)$ follows from Lemma 4 and will be proved in Lemma 5. In addition, the edit distance between $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$ and $\mathbb{1}_{\mathbf{w}}(\mathbf{y})$ needs to be bounded to apply Lemma 2. This is proved in the following lemma.

Lemma 5. *Let $\mathbf{w} \in \{0,1\}^{2p}$ be a non-periodic string. For two strings $\mathbf{x}$ and $\mathbf{y} \in \mathcal{B}_k(\mathbf{x})$, we have that*

(a) $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{R}_{n,p}$.
(b) $\mathbb{1}_{\mathbf{w}}(\mathbf{y}) \in \mathcal{R}_{n,p}$.
(c) $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{B}_{5k}(\mathbb{1}_{\mathbf{w}}(\mathbf{y}))$.

Proof. The statements (a) and (b) follow from the definition of vectors $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$ and $\mathbb{1}_{\mathbf{w}}(\mathbf{y})$ and the fact that $\mathbf{w}$ is non-periodic. Suppose there are two 1 entries $\mathbb{1}_{\mathbf{w}}(\mathbf{x})_i$ and $\mathbb{1}_{\mathbf{w}}(\mathbf{x})_{i+a}$ in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$ that are separated by less than $p-1$ 0's, i.e., $a \le p-1$. Then by definition of $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$, we have that $\mathbf{x}_{i:i+2p-1} = \mathbf{w}$ and that $\mathbf{x}_{i+a:i+a+2p-1} = \mathbf{w}$. This implies that $w_j = x_{i+a+j-1} = w_{j+a}$ for $j \in [2p-a]$. Hence, the string $\mathbf{w}$ has period $a \le p-1$, contradicting to the fact that $\mathbf{w}$ is non-periodic. Hence, we have that $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{R}_{n,p}$, and similarly that $\mathbb{1}_{\mathbf{w}}(\mathbf{y}) \in \mathcal{R}_{n,p}$

We now prove statement (c). To this end, we first show that a deletion in $\mathbf{x}$ results in at most three deletions and two insertions in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$. Since $\mathbf{w}$ has length $2p$ and $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{R}_{n,p}$ as shown in (a), a deletion in $\mathbf{x}$ results in at most two deletions and two insertions of 1 entries in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$, respectively. Otherwise, suppose that a deletion in $\mathbf{x}$ deletes three 1 entries $\mathbb{1}_{\mathbf{w}}(\mathbf{x})_{i_1}$, $\mathbb{1}_{\mathbf{w}}(\mathbf{x})_{i_2}$, and $\mathbb{1}_{\mathbf{w}}(\mathbf{x})_{i_3}$ in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$, then we have that $i_3 - i_1 \ge 2p$ because (a) holds. This is impossible since $\mathbf{w} \in \{0,1\}^{2p}$ and the deletion in $\mathbf{x}$ can not affect the two occurrences $\mathbf{x}_{i_1:i_1+2p-1}$ and $\mathbf{x}_{i_3:i_3+2p-1}$ of $\mathbf{w}$ in $\mathbf{x}$ simultaneously. Hence a deletion causes at most two deletions of 1 entries in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$ and similarly, the same holds for insertions.

Moreover, at most one 0 entry is deleted in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$ because of the deletion in $\mathbf{x}$. Hence, a deletion in $\mathbf{x}$ causes at most three deletions and two insertions in total in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$, and k deletions in $\mathbf{x}$ results in at most $3k$ deletions and $2k$ insertions in $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$. The same holds for $\mathbf{y}$ and $\mathbb{1}_{\mathbf{w}}(\mathbf{y})$.

Since $\mathbf{x} \in \mathcal{B}_k(\mathbf{y})$, we conclude that $\mathbb{1}_{\mathbf{w}}(\mathbf{y})$ can be obtained from $\mathbb{1}_{\mathbf{w}}(\mathbf{x})$ by at most $5k$ deletions and $5k$ insertions, and hence, $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{B}_{5k}(\mathbb{1}_{\mathbf{w}}(\mathbf{y}))$. □

Finally, with Lemma 2 and Lemma 5 established, we present a lower bound on the maximal value of polynomial $f^m_{\mathbf{x},\mathbf{w}}(z) - f^m_{\mathbf{y},\mathbf{w}}(z)$ for z close to 1. This lower bound guarantees a gap between the multi-bit statistics of $\tilde{X}$ and $\tilde{Y}$, which makes $\mathbf{x}$ and $\mathbf{y}$ distinguishable by Hoeffding's inequality (See Section III). Note that it is important that z is located around the point 1 on the complex plain because of the scaling factor $(2z-1)^i$ in the multi-bit statistics in Eq. (2). To meet this requirement on z, existing works [8], [12], [13], [19] restrict z to lie on short subarcs of a unit circle around 1, a case also considered in [5] in the context of complex analysis. In this paper, we choose z from a small circle around 1. It turns out that this choice of z achieves a lower bound $\frac{1}{n^{O(k)}}$ on $f^m_{\mathbf{x},\mathbf{w}}(z) - f^m_{\mathbf{y},\mathbf{w}}(z)$, which improves the bound $\frac{1}{n^{O(k^2)}}$ established in [13]. The details will be given in the following lemma, which is a critical result in this paper. Its proof will be given in Section IV.

Lemma 6. *For distinct strings $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$, if $\sum_{i=1}^{n} i^m x_i \neq \sum_{i=1}^{n} i^m y_i$ for some non-negative integer m, then there exists a complex number z, such that $|2z-1|^n \leq 2$ and*

$$\sum_{k=1}^{n-\ell+1} \mathbb{1}_{\mathbf{x}_{k:k+\ell-1}=\mathbf{w}} z^k - \sum_{k=1}^{n-\ell+1} \mathbb{1}_{\mathbf{y}_{k:k+\ell-1}=\mathbf{w}} z^k \\ \geq \frac{1}{n^{2m}(2m+2)}. \tag{4}$$

III. Proof of Theorem 1

In this section we prove Theorem 1 based on the results from Lemma 1 to Lemma 6. Let ℓ be the smallest index such that $x_i \neq y_i$. If $\ell < 10k$, we have the following result from [21], which was also used in [8].

Proposition 2. *For sequences $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$, let ℓ be the smallest index such that $x_\ell \neq y_\ell$, i.e., $x_i = y_i$ for $i \in [\ell-1]$. Then, with high probability $\mathbf{x}$ and $\mathbf{y}$ can be distinguished using $exp(O(\ell^{\frac{1}{3}}))$ samples.*

According to Proposition 2, sequences $\mathbf{x}$ and $\mathbf{y}$ can be distinguished with high probability using $exp(O(\ell^{\frac{1}{3}})) < n^{O(k)}$ samples. Hence, it suffices to consider cases when $\ell \geq 10k$.

Let $\mathbf{w}' = \mathbf{x}_{\ell-10k+1:\ell-1}$. By Lemma 4, either $(\mathbf{w}', 0)$ or $(\mathbf{w}', 1)$ is non-periodic. Without loss of generality, assume that $\mathbf{w} = (\mathbf{w}', 0) \in \{0,1\}^{10k}$ is non-periodic. Then, similar to the arguments in [8], [12], [13], [16], [19], the core part of the proof is to show that the difference of multi-bit statistics $\mathbb{E}_{\tilde{X}}[\mathbb{1}_{\tilde{X}_{k_i}=w_i, \forall i \in [10k]}]$ and $\mathbb{E}_{\tilde{Y}}[\mathbb{1}_{\tilde{Y}_{k_i}=w_i, \forall i \in [10k]}]$, is at least $\frac{1}{n^{O(k)}}$ for some integers $1 \leq i_1 < \ldots < i_{10k} \leq n$, i.e.,

$$\max_{1 \leq i_1 < \ldots < i_{10k} \leq n} |\mathbb{E}_{\tilde{X}}[\mathbb{1}_{\tilde{X}_{i_j}=w_j, \forall j \in [10k]}] \\ - \mathbb{E}_{\tilde{Y}}[\mathbb{1}_{\tilde{Y}_{i_j}=w_j, \forall j \in [10k]}]| \geq \frac{1}{n^{O(m)}}. \tag{5}$$

Let

$$(i_1^*, \ldots, i_{10k}^*) = \text{argmax}_{1 \leq i_1 < \ldots < i_{10k} \leq n} |\mathbb{E}_{\tilde{X}}[\mathbb{1}_{\tilde{X}_{i_j}=w_j, \forall j \in [10k]}] \\ - \mathbb{E}_{\tilde{Y}}[\mathbb{1}_{\tilde{Y}_{i_j}=w_j, \forall j \in [10k]}]|,$$

which can be determined once $\mathbf{x}$ and $\mathbf{y}$ are given. Suppose that $\mathbf{x}$ is passed through the deletion channel N times, generating N independent samples $\{\tilde{T}^t\}_{t=1}^N$. Then, by using similar Hoeffding's inequality (or the Chernoff bound) arguments as in [19], we can show that with high probability, the empirical distribution $\frac{\sum_{t=1}^N \mathbb{1}_{\tilde{T}^t_{i_j^*}=w_j, \forall j \in [10k]}}{N}$ is closer to $E[\mathbb{1}_{\tilde{X}_{i_j^*}=w_j, \forall j \in [10k]}]$ than to $E[\mathbb{1}_{\tilde{Y}_{i_j^*}=w_j, \forall j \in [10k]}]$, if

$$N \geq O(\frac{1}{|\mathbb{E}_{\tilde{X}}[\mathbb{1}_{\tilde{X}_{i_j^*}=w_j, \forall j \in [10k]}] - \mathbb{E}_{\tilde{Y}}[\mathbb{1}_{\tilde{Y}_{i_j^*}=w_j, \forall j \in [10k]}]|^2)} \\ = n^{O(k)}.$$

Hence $\mathbf{x}$ and $\mathbf{y}$ can be distinguished using $n^{O(k)}$ samples. Therefore, it suffices to show (5) in the rest of the proof.

Since $\mathbf{w}$ is non-periodic and $\mathbf{x} \in \mathcal{B}_k(\mathbf{y})$, Lemma 5 implies that $\mathbb{1}_{\mathbf{w}}(\mathbf{x}), \mathbb{1}_{\mathbf{w}}(\mathbf{y}) \in \mathcal{R}(n, 5k)$ and that $\mathbb{1}_{\mathbf{w}}(\mathbf{x}) \in \mathcal{B}_{5k}(\mathbb{1}_{\mathbf{w}}(\mathbf{y}))$. Hence, we apply Lemma 2 and obtain an integer $m \in [12k+1]$ such that $\sum_{i=1}^n i^m x_i \neq \sum_{i=1}^n i^m y_i$. Then, according to Lemma 6, there exists a complex number z, such that $|2z-1|^n \leq exp(2)$ and (4) holds. Eq. (4) and Lemma 1 implies that

$$\sum_{1 \leq i_1 < \ldots < i_{10k} \leq n} |\mathbb{E}_{\tilde{X}}[\mathbb{1}_{\tilde{X}_{i_j}=w_j, \forall j \in [10k]}] - \mathbb{E}_{\tilde{Y}}[\mathbb{1}_{\tilde{Y}_{i_j}=w_j, \forall j \in [10k]}]| \\ \cdot 2^\ell (2z-1)^{k_1} (-1)^{k_\ell - k_1 - 10k + 1} \\ \geq \frac{1}{n^{O(m)}},$$

and thus that

$$\max_{1 \leq i_1 < \ldots < i_{10k} \leq n} |\mathbb{E}_{\tilde{X}}[\mathbb{1}_{\tilde{X}_{i_j}=w_j, \forall j \in [10k]}] - \mathbb{E}_{\tilde{Y}}[\mathbb{1}_{\tilde{Y}_{i_j}=w_j, \forall j \in [10k]}]| \\ \geq \frac{1}{n^{O(m)}} \cdot \frac{1}{2^\ell} \cdot \frac{1}{\binom{n}{10k}} \cdot \frac{1}{(2z-1)^n} \\ = \frac{1}{n^{O(k)}}.$$

Therefore, (5) holds and the proof is done.

IV. Proof of Lemma 6

Without loss of generality, assume that m is the smallest non-negative integer satisfying $\sum_{i=1}^n i^m x_i \neq \sum_{i=1}^n i^m y_i$. Let

$$f(z) = \sum_{k=1}^{n-\ell+1} \mathbb{1}_{\mathbf{x}_{k:k+\ell-1}=\mathbf{w}} z^k - \sum_{k=1}^{n-\ell+1} \mathbb{1}_{\mathbf{y}_{k:k+\ell-1}=\mathbf{w}} z^k \tag{6}$$

be a complex polynomial. The coefficients of $f(z)$ are within the set $\{-1, 0, 1\}$.

According to Lemma 3, we have that $f(z) = (z-1)^m q(x)$, where $q(z) = \sum_{i=0}^{n_1} c_i z^i$ is a complex polynomial with integer coefficients and $(z-1)$ does not divide $q(z)$, i.e., $q(1) \neq 0$. The following result was presented in [6]. It gives an upper bound on the norm of coefficients of $q(z)$.

Proposition 3. *[6] If a complex degree n polynomial $f(z)$ has all coefficients with norm not greater than 1, and can be factorized by*

$$f(z) = (z-1)^m q(x) = (z-1)^m (c_{n_1} z^{n_1} + \ldots + c_0),$$

then, we have that $\sum_{i=1}^{n_1} |c_i| \leq (n+1)(\frac{en}{m})^m$.

We are now ready to prove Lemma 6. Let $D \triangleq 2m+2$ $z_j = exp(\frac{2j\pi i}{D})$, $j \in [D]$ be a sequence of D complex numbers equally distributed on a unit circle. We first show that there exists a number $j \in [D]$ satisfying

$$q(1 + \frac{z_j}{n^2}) \geq \frac{1}{n^{O(m)}}.$$

Note that

$$\begin{aligned}
|\sum_{j=1}^{n} q(1+\frac{z_j}{n^2})| =& |\sum_{r=0}^{n_1} c_r[\sum_{j=1}^{D}(1+\frac{exp(\frac{2j\pi i}{D})}{n^2})^r]| \\
=& |\sum_{r=0}^{n_1} c_r \sum_{s=0}^{r}\binom{r}{s}\sum_{j=1}^{D}\frac{exp(\frac{2js\pi i}{D})}{n^{2s}}| \\
\overset{(a)}{=}& |\sum_{r=0}^{n_1} c_r \sum_{s=0}^{r}\binom{r}{s}\frac{D\mathbb{1}_{D|s}}{n^{2s}}| \\
=& |\sum_{r=0}^{n_1} c_r (D+\sum_{s=1}^{r}\binom{r}{s}\frac{D\mathbb{1}_{D|s}}{n^{2s}})| \\
=& |Dq(1)+\sum_{r=0}^{n_1}\sum_{s=1}^{r} c_r\binom{r}{s}\frac{D\mathbb{1}_{D|s}}{n^{2s}}| \\
\geq& D|q(1)| - \sum_{s=1}^{n_1}(\sum_{r=s}^{n_1}|c_r|)\binom{n}{s}\frac{D\mathbb{1}_{D|s}}{n^{2s}} \\
\overset{(b)}{\geq}& D-(n+1)(\frac{en}{m})^m\sum_{s=1}^{n_1}\frac{D\mathbb{1}_{D|s}}{n^{s}} \\
\geq& D-D(n+1)(\frac{en}{m})^m\frac{1}{n^D}\sum_{t=0}^{\infty}\frac{1}{n^{Dt}} \\
\overset{(c)}{\geq}& 1
\end{aligned}$$

as n goes to infinity, where (a) follows from the identity

$$\sum_{j=1}^{D} exp(\frac{2js\pi i}{D}) = \frac{exp(\frac{2sD\pi i}{D})-1}{exp(\frac{2s\pi i}{D})-1}=0,$$

(b) follows from Proposition 3 and the fact that $\binom{n}{s}\leq n^s$, and (c) follows from the fact that $D\triangleq 2m+2\geq 2$ and that $\sum_{t=0}^{\infty}\frac{1}{n^{Dt}}<2$. Therefore, there exists an integer j such that

$$|q(1+\frac{z_j}{n^2})|\geq\frac{1}{D},$$

and thus that

$$\begin{aligned}
|f(1+\frac{z_j}{n^2})| =& \frac{|q(1+\frac{z_j}{n^2})|}{n^{2m}} \\
\geq& \frac{1}{n^{2m}(2m+2)}.
\end{aligned}$$

Moreover, we have that

$$\begin{aligned}
|2(1+\frac{z_j}{n^2})-1|^n \leq& (1+\frac{4}{n^4}+4\frac{cos(\frac{2j\pi}{D})}{n^2})^n \\
\leq& 2
\end{aligned}$$

as n goes to infinity. Hence, $z=1+\frac{z_j}{n^2}$ satisfies the conditions in Lemma 6.

V. Conclusion

In this paper we study the trace reconstruction problem when the string to be recovered is within bounded edit distance to a known string. Our result implies that when the edit distance is constant, the number of traces needed is polynomial. The problem of whether polynomial number of samples suffices for the general trace reconstruction is open. However, it is interesting to see if the methods in this paper can be extended to obtain more general results.

References

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

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

[3] V. Bhardwaj, P. A. Pevzner, C. Rashtchian, and Y. Safonova, "Trace reconstruction problems in computational biology." *IEEE Transactions on Information Theory*, to appear.

[4] P. Borwein, "Computational excursions in analysis and number theory." *Springer Science & Business Media*, 2012.

[5] P. Borwein and T. Erdélyi, "Littlewood-type problems on subarcs of the unit circle." *Indiana University mathematics journal*, vol. 46, no. 4, pp. 1323—1346, 1997.

[6] Peter Borwein, Tamás Erdélyi, and Géza Kós. "Littlewood-type problems on [0, 1]." *Proceedings of the London Mathematical Society*, vol. 79, no. 1, pp. 22–46, 1999.

[7] Z. Chase, "New lower bounds for trace reconstruction." *arXiv:1905.03031*, 2020.

[8] Z. Chase, "New upper bounds for trace reconstruction." *arXiv:2009.03296*, 2020.

[9] 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.

[10] X. Chen, A. De, C. H. Lee, R. A. Servedio, and S. Sinha, "Polynomial-time trace reconstruction in the low deletion rate regime." *arXiv:2012.02844*, 2020.

[11] S. Davies, M. Z. Rácz, C. Rashtchian and B. G. Schiffer, "Approximate trace reconstruction." *arXiv:2012.06713*, 2020.

[12] 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.

[13] E. Grigorescu, M. Sudan, and M. Zhu, "Limitations of Mean-Based Algorithms for Trace Reconstruction at Small Distance." *arXiv:2011.13737*, 2020.

[14] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability." *Proceedings of the 31st Conference On Learning Theory (COLT)*, pp. 1799–1840, 2018.

[15] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results." *Proc. 19th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 389—398, 2008.

[16] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Trace reconstruction: Generalized and parameterized." *arXiv:1904.09618*, 2019.

[17] V. I. Levenshtein, "Efficient Reconstruction of Sequences." *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 2–22, 2001.

[18] I. Krasikov and Y. Roditty, "On a reconstruction problem for sequences." *Journal of Combina- torial Theory, Series A*, vol. 77, no. 2, pp. 344–348, 1997.

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

[20] L. Organick, S. D. Ang, Y. J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. Takahashi, S. Newman, H. Y. Parker, C. Rashtchian, G. G. K. Stewart, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, and K. Strauss, "Scaling up DNA data storage and random access retrieval," *bioRxiv*, 2017.

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

[22] J. M. Robson, "Separating strings with small automata." *Information Processing Letters*, vol. 30, no. 4, pp. 209–214, 1989.

[23] J. Sima and J. Bruck, "Optimal k-deletion correcting codes," *IEEE Transactions on Information Theory*, to appear.

[24] Wikipedia, "Reference genome", available at https://en.wikipedia.org/wiki/Reference_genome