Circular Trace Reconstruction

Shyam Narayanan
Massachusetts Institute of Technology, Cambridge, MA, USA
shyamsn@mit.edu

Michael Ren
Massachusetts Institute of Technology, Cambridge, MA, USA
mren36@mit.edu

Abstract

Trace reconstruction is the problem of learning an unknown string x from independent traces of x, where traces are generated by independently deleting each bit of x with some deletion probability q. In this paper, we initiate the study of *Circular trace reconstruction*, where the unknown string x is circular and traces are now rotated by a random cyclic shift. Trace reconstruction is related to many computational biology problems studying DNA, which is a primary motivation for this problem as well, as many types of DNA are known to be circular.

Our main results are as follows. First, we prove that we can reconstruct arbitrary circular strings of length n using $\exp\left(\tilde{O}(n^{1/3})\right)$ traces for any constant deletion probability q, as long as n is prime or the product of two primes. For n of this form, this nearly matches what was the best known bound of $\exp\left(O(n^{1/3})\right)$ for standard trace reconstruction when this paper was initially released. We note, however, that Chase very recently improved the standard trace reconstruction bound to $\exp\left(\tilde{O}(n^{1/5})\right)$. Next, we prove that we can reconstruct random circular strings with high probability using $n^{O(1)}$ traces for any constant deletion probability q. Finally, we prove a lower bound of $\tilde{\Omega}(n^3)$ traces for arbitrary circular strings, which is greater than the best known lower bound of $\tilde{\Omega}(n^{3/2})$ in standard trace reconstruction.

2012 ACM Subject Classification Mathematics of computing → Probabilistic algorithms

Keywords and phrases Trace Reconstruction, Deletion Channel, Cyclotomic Integers

Digital Object Identifier 10.4230/LIPIcs.ITCS.2021.18

Related Version A full version of the paper is available at `https://arxiv.org/abs/2009.01346`.

Funding *Shyam Narayanan*: Supported by the MIT Akamai Fellowship, the NSF Graduate Fellowship, and a Simons Investigator Award.
Michael Ren: Supported by NSF-DMS grant 1949884 and NSA grant H98230-20-1-0009.

Acknowledgements The first author thanks Professor Piotr Indyk for many helpful discussions and feedback, Mehtaab Sawhney for pointers to some references, and Professor Bjorn Poonen for a helpful discussion on sums of roots of unity. The second author thanks Professor Joe Gallian for running the Duluth REU at which part of this research was conducted, as well as program advisors Amanda Burcroff, Colin Defant, and Yelena Mandelshtam for providing a supportive environment. The authors also thank Amanda Burcroff for helpful edits on the paper's writeup.

1 Introduction

The trace reconstruction problem asks one to recover an unknown string x of length n from independent noisy samples of the string. In the original setting, x is a binary string in $\{0,1\}^n$, and a random subsequence $\tilde{x}$ of x, called a *trace*, is generated by sending x through a deletion channel with deletion probability q, which removes each bit of x independently with some fixed probability q. The main question is to determine how many independent traces are needed to recover the original string with high probability. This question has become very well studied over the past two decades [26, 27, 7, 24, 23, 38, 30, 19, 32, 35, 20, 21, 22, 13, 15, 14],

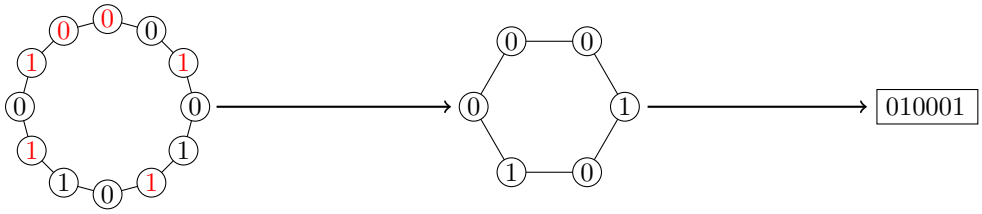

Figure 1 An example of a circular trace. We start with an unknown circular string (top left). Each bit of the string is randomly deleted (red bits are deleted, black bits are retained) and the order of the retained bits is preserved, so we are left with the smaller circular string. However, since there is no beginning or end of the circular string, we assume the string is seen in clockwise order starting from a randomly chosen bit.

with many results over various settings. For instance, people have studied the case where we wish to reconstruct x for any arbitrarily chosen $x \in \{0,1\}^n$ (worst-case) or the case where we just wish to reconstruct a randomly chosen string x (average-case). People have also studied the trace reconstruction problem for various values of the deletion probability q, such as if q is a fixed constant between 0 and 1 or decays as some function of n. People have also studied variants where the traces allow for insertions of random bits, rather than just deletions, and variants where the string is no longer binary but from a larger alphabet.

Finally, various generalizations or variants of the trace reconstruction problem have also been developed. These include error-correcting codes over the deletion channel (i.e., "coded" trace reconstruction) [16, 11], reconstructing matrices [25] and trees [18] from traces, and reconstructing mixtures of strings from traces [3, 4, 31].

In this paper, we develop and study a new variant of trace reconstruction that we call *Circular trace reconstruction*. In this variant, there is again an unknown string $x \in \{0,1\}^n$ that we can sample traces from, but this time, the string x is a cyclic string, meaning that there is no beginning or end to the string. Equivalently, one can imagine a linear string that undergoes a random cyclic shift before a trace is returned. See Figure 1 for an example. Our goal, like in the normal trace reconstruction, is to reconstruct the original circular string using as few random traces as possible.

1.1 Main Results and Comparison to Linear Trace Reconstruction

Perhaps the first natural question about circular trace reconstruction is the following: how does the sample complexity of circular trace reconstruction compare to the sample complexity of standard (linear) trace reconstruction? Intuitively, one should expect circular trace reconstruction to be at least as difficult as standard trace reconstruction, since given any trace of a linear string, we can randomly rotate it to get a trace of the corresponding circular string. This reasoning, however, is slightly flawed. For instance, perhaps the hardest instance of linear trace reconstruction comes from distinguishing between two strings x and y which are different as linear strings but equivalent up to a cyclic shift. In this case, the circular trace reconstruction problem does not even need to distinguish between x and y, because they are equivalent! However, by padding the trace with extra bits before randomly rotating, one can show that circular trace reconstruction is at least as hard as linear trace reconstruction in both the worst-case and average-case. Indeed, we have the following proposition. As its proof is quite simple, we defer it to Appendix A in the full version of this paper on arXiv.

▶ **Proposition 1.** *Suppose that we can solve worst-case (resp., average case) circular trace reconstruction over length m strings with deletion probability q using $T(m,q)$ traces. Then, we can solve worst-case (resp., average case) linear trace reconstruction over length n strings with deletion probability q using $\min_{m \geq 2n} T(m,q)$ traces.*

Given Proposition 1, any upper bounds for circular trace reconstruction imply nearly equivalent upper bounds for the linear trace reconstruction, and any lower bounds for linear trace reconstruction imply nearly equivalent lower bounds for circular trace reconstruction. This raises two natural questions. First, can we match or nearly match the best linear trace reconstruction upper bounds for circular trace reconstruction? Second, can we beat the best linear trace reconstruction lower bounds for circular trace reconstruction?

The first main result we prove is for worst-case circular strings. At the time of the initial release of this paper, the best known upper bound for worst-case linear trace reconstruction with deletion probability q, where q is a fixed constant between 0 and 1, is $\exp\left(O(n^{1/3})\right)$, where the unknown string has length n [19, 32]. Shortly afterwards, the upper bound was improved to $\exp\left(\tilde{O}(n^{1/5})\right)$ [14]. Our first main result, which we prove in Section 3, provides an upper bound for the circular trace reconstruction problem that nearly matches the results of [19, 32], but only if the length n has at most 2 prime factors.

▶ **Theorem 2.** *Let x be an unknown, arbitrary circular string of length n, let q be the deletion probability of each element in the string, and let $p = 1 - q$ be the retention probability. Then, if n is either a prime or a product of two (possibly equal) primes, using $\exp\left(O\left(n^{1/3}(\log n)^{2/3}p^{-2/3}\right)\right)$ random traces, we can determine x with failure probability at most 2^{-n}.*

The primary reason why our theorem fails for n having 3 or more prime factors is that we prove the following number theoretic result which is crucial in our algorithm.

▶ **Theorem 3.** *For any fixed integer $n \geq 2$, the following statement is true **if and only if** n has at most 2 prime factors, counting multiplicity.*

Define $\omega := e^{2\pi i/n}$, and suppose that $a_0, \ldots, a_{n-1}, b_0, \ldots, b_{n-1}$ are all integers in $\{0, 1\}$. Also, suppose that for all $0 \leq k \leq n-1$, there is some integer c_k such that $\sum_{i=1}^{n} a_i\omega^{i\cdot k} = \omega^{c_k} \cdot \sum_{i=1}^{n} b_i\omega^{i\cdot k}$. Then, the sequences $\{a_i\}$ and $\{b_i\}$ are cyclic shifts of each other.

In this conference version, we only prove the above theorem in the special case that n is prime. For the full proof of this theorem, please see the arXiv version of this paper at `https://arxiv.org/abs/2009.01346`.

The next main result we prove is for average-case circular strings: we show that a random circular string can be recovered using a polynomial number of traces. Formally, we prove the following theorem in Section 4.

▶ **Theorem 4.** *Let x be an unknown but randomly chosen circular string of length n and let $0 < q < 1$ be the deletion probability of each element. Then, there exists a constant C_q depending only on q such that we can determine x with failure probability at most n^{-10} using $O(n^{C_q})$ traces.*

The main lemma we need to prove Theorem 4 is actually a result that is true for worst-case strings. Specifically, we show how to recover the multiset of all consecutive substrings of length $O(\log n)$ using a polynomial number of traces. While this does not guarantee that we can recover an arbitrary circular string, it does allow us to recover what we will call *regular strings*, which we show comprise the majority of circular strings. The following lemma may be of independent interest for studying worst-case strings as well, as it allows one to gain information about all "consecutive chunks" of the unknown string using only a polynomial number of queries.

▶ **Lemma 5.** *Let $x = x_1 \cdots x_n$ be an arbitrary circular string of length n and let $0 < q < 1$ be the deletion probability of each element. Then, for $k = 100 \log n$, we can recover the multiset of all substrings $\{x_i x_{i+1} \cdots x_{i+k-1}\}_{i=1}^{n}$, where indices are modulo n, using $O(n^{C_q})$ traces with failure probability n^{-10}, where C_q is a constant that only depends on q.*

The best upper bound for average-case linear trace reconstruction is $\exp\left(O((\log n)^{1/3})\right)$ [22]. Unfortunately, we were not able to adapt their argument to circular strings. One major reason why we are unable to do so is that in the argument of [22] (as well as [35], which provides an $\exp\left(O((\log n)^{1/2})\right)$ sample algorithm), the authors recover the $(k+1)^{\text{st}}$ bit of the string assuming the first k bits are known using a small number of traces, and by reusing traces, they inductively recover the full string. However, since we are dealing with circular strings, even recovering the "first" bit does not make much sense. However, we note that even a polynomial-sample algorithm is quite nontrivial. In the linear case, a polynomial-sample algorithm for average-case strings was first proven by [23], and their algorithm only worked as long as the deletion probability q was at most some small constant, which when optimized is only about 0.07 [35].

Our final main result regards lower bounds for worst-case strings. For linear worst-case strings, the best known lower bound for trace reconstruction is $\tilde{\Omega}(n^{3/2})$ [13]. For circular trace reconstruction, we show an improved lower bound of $\tilde{\Omega}(n^3)$. Moreover, the proof of our lower bound is actually much simpler and cleaner than those of the known lower bounds for standard trace reconstruction [13, 21]. Specifically, we prove the following theorem, done in Section 5:

▶ **Theorem 6.** *Let x be the string $10^n 10^{n+1} 10^{n+k} = 1\underbrace{0\ldots0}_{n \text{ times}}1\ \underbrace{0\ldots0}_{n+1 \text{ times}}\ 1\ \underbrace{0\ldots0}_{n+k \text{ times}}$, where $n \geq 1$ and $2 \leq k \leq 4$. Likewise, let y be the string $y = 10^n 10^{n+k} 10^{n+1}$. Then, the strings x, y are not equivalent up to cyclic rotations, but for any constant deletion probability q, one requires $\Omega(n^3/\log^3 n)$ random traces to distinguish between the original string being x or y. Thus, for all integers n, worst-case circular trace reconstruction requires at least $\tilde{\Omega}(n^3)$ random traces.*

1.1.1 Concurrent Work

We note that a very similar statement to Lemma 5, but for linear strings, was proven in independent concurrent work by Chen et. al. [15, Theorem 2], which provides a polynomial-sample algorithm for a "smoothed" variant of worst-case linear trace reconstruction. Many ideas in our proof of Lemma 5 and their proof appear to overlap, though our proof is substantially shorter. We discuss the relation between our work and [15] further at the end of Section 4.

1.2 Motivation and Relation to Other Work

From a theoretical perspective, circular trace reconstruction can bring many novel insights to the theory of reconstruction algorithms, some of which may be useful even in the standard trace reconstruction problem. For instance, the proof of Theorem 2 combines analytic, statistical, and combinatorial approaches as in previous trace reconstruction papers, but now also uses ideas from number theory and results about cyclotomic integers. To the best of our knowledge, this paper is the first paper on trace reconstruction to utilize number theoretic ideas, though there is work on other problems about cyclic strings that uses ideas from number theory. Also, Lemma 5 shows a way to recover all contiguous sequences in the

original string of length $O(\log n)$ for arbitrary circular strings, which is a new result even in the linear case (concurrent with [15]) and has applications to problems in linear trace reconstruction as well (as done in [15]).

From an applications perspective, trace reconstruction is closely related to the multiple sequence alignment problem in computational biology. In the multiple sequence alignment problem, one is given DNA sequences from several related organisms, and the goal is to align the sequences to determine what mutations each descendant underwent from their common ancestor: the trace reconstruction problem is analogous to actually recovering the common ancestor. One can learn more about trace reconstruction's relation to the multiple sequence alignment problem (as well as to various other problems in computational biology) via the recent survey [8].

The multiple sequence alignment problem is also a key motivation for studying circular trace reconstruction. Many important types of DNA, such as mitochondrial DNA in humans and other eukaryotes, chloroplast DNA, bacterial DNA, and DNA in plasmids, are predominantly circular (see, e.g., [37, pp. 313, 397, 516-517], or [1]). Therefore, understanding circular trace reconstruction could prove useful in reconstructing ancestral sequences for mitochondrial or bacterial DNA. Another problem in computational biology that trace reconstruction may be applicable to is the DNA Data Storage problem, where data is stored in DNA and can be recovered through sequencing, though the stored DNA may mutate over time [17, 34]. Recently, long-term DNA data storage in plasmids has been successfully researched [33], which further motivates the study of circular trace reconstruction.

Besides the linear trace reconstruction problem, circular trace reconstruction is also closely related to the problem of population recovery from the deletion channel [3, 4, 31], where the goal is to recover an unknown mixture of ℓ strings from random traces. Indeed, receiving traces from a circular string is equivalent to receiving traces from a uniform mixture of a linear string along with all of its cyclic shifts, so circular trace reconstruction can be thought of as an instance of population recovery from the deletion channel with mixture size $\ell = n$.

Unfortunately, the best known algorithm for population recovery over worst-case strings requires $\exp\left(\tilde{O}(n^{1/3}) \cdot \ell^2\right)$ traces [31], which is not useful if $\ell = n$. However, to prove our worst-case upper bound, we will use ideas based on [19, 32, 31] to estimate certain polynomials that depend on the unknown circular string x. For the average case problem, i.e., if given a mixture over ℓ random strings, population recovery can be done with $\text{poly}\left(\ell, \exp\left((\log n)^{1/3}\right)\right)$ random traces. While this seemingly implies a $\text{poly}(n)$-sample algorithm for average-case circular trace reconstruction, the n cyclic shifts of the circular string are quite similar to each other and thus do not behave like a collection of n independent random strings. Indeed, our techniques for average-case circular trace reconstruction are very different from those developed in [4].

While circular strings have not been studied before in the context of trace reconstruction, people have studied circular strings and cyclic shifts in the context of edit distance [28, 2], multi-reference alignment [5, 6, 36], and other pattern matching problems [12]. We note that [2] also applies results from number theory and about cyclotomic polynomials, though their techniques overall are not very similar to ours.

1.3 Proof Overview

In this subsection, we highlight some of the ideas used in Theorems 2, 4, and 6.

The proof of Theorem 2 is partially based on ideas from [19, 32, 31]. The authors of [19, 32] consider two strings $x, y \in \{0,1\}^n$ and show how to distinguish between random traces of x and random traces of y. To do so, they construct an unbiased estimator for the

polynomial $P(z;x) := \sum_{i=1}^{n} x_i z^i$ (or $P(z;y) = \sum_{i=1}^{n} y_i z^i$) solely based on the random trace of either x or y, for some $z \in \mathbb{C}$. By showing that the unbiased estimator is never "too" large and that $P(z;x)$ and $P(z;y)$ differ enough for an appropriate choice of z, they can estimate this quantity using random traces to distinguish between x and y. In our case, applying the same estimator will give us an unbiased estimator for $P'(z;x) := \mathbb{E}_i[P(z;x^{(i)})]$, where $x^{(i)}$ is the ith cyclic shift of x. Unfortunately, it turns out that $P'(z;x) = P'(z;y)$ as polynomials in z as long as x, y have the same number of 1's, even if x and y are vastly different as circular strings. Our goal will then be to establish some other polynomial $Q(z;x)$ such that we can construct a good unbiased estimator, but at the same time $Q'(z;x) := \mathbb{E}_i[Q(z;x^{(i)})]$ and $Q'(z;y) := \mathbb{E}_i[Q(z;y^{(i)})]$ are distinct polynomials for any distinct cyclic strings x, y. We show that the polynomial $Q(z;x) := z^{kn}P(z;x)^k P(z^{-k};x)$ will do the job, for some some small integer k. We provide a (significantly more complicated) unbiased estimator of $Q(z;x)$ using a random trace: the construction is similar to that of [31], which shows how to estimate $P(z;x)^k$ for some integer k. To show that $Q(z;x) \neq Q(z;y)$ as polynomials, we first show that $Q(z;x)$ has the special property that if z is a cyclotomic nth root of unity, this polynomial is invariant under cyclic shifts of x! Thus, it just suffices to show that if $x, y \in \{0,1\}^n$ are not cyclic shifts of each other, there is some nth root of unity $z = e^{2\pi i r/n}$ for some $0 \leq r \leq n-1$ such that $P(z;x)^k P(z^{-k};x) \neq P(z;y)^k P(z^{-k};y)$. This will require significant number theoretic computation, and will be true as long as n is a prime or a product of two primes.

The bulk of the proof of Theorem 4 will be proving Lemma 5, which reconstructs all consecutive substrings of length $100 \log n$ in the unknown circular string x. For a random string x, these substrings will all be sufficiently different, so once we know the substrings, we can reconstruct the full string because there will only be one way to "glue" together the substrings. Therefore, we focus on explaining the ideas for Lemma 5. Our goal will be to determine how many times a string s appears consecutively in x for each string s of length $100 \log n$. For an unknown string x and i between 0 and $n - 100 \log n$, we let c_i be the number of times s appears (possibly non-contiguously) in some contiguous block of length $i + 100 \log n$ in x. Then, a basic enumerative argument shows that for a random (cyclically shifted) trace $\tilde{x} = \tilde{x}_1 \tilde{x}_2 \cdots \tilde{x}_m$, the probability that $\tilde{x}_1 \cdots \tilde{x}_{100 \log n} = s$ can be written as $\sum_{i \geq 0} c_i (1-q)^{100 \log n} q^i$, and we wish to recover c_0. The $(1-q)^{100 \log n}$ term is a constant that equals $1/\mathrm{poly}(n)$, so it is easy to recover an approximation to $\sum_{i \geq 0} c_i q^i$. We truncate this polynomial at an appropriate degree (approximately $C \log n$ for some large C) and show that the truncated polynomial $\sum_{i=0}^{C \log n} c_i x^i$ is very close to the original polynomial, but differs from $\sum_{i=0}^{C \log n} c'_i x^i$ for some $x \in [q, (1-q)/2]$ by a significant amount, if $c'_0 \neq c_0$, using ideas based on [10]. We can also simulate a trace with deletion probability $x > q$ by taking a "trace of the trace." This will be sufficient in determining c_0, and therefore, the (multi)-set of all consecutive substrings of length $100 \log n$.

The proof of Theorem 6 proceeds by showing that the laws of the traces of $x = 10^n 10^{n+1} 10^{n+k}$ and $y = 10^n 10^{n+k} 10^{n+1}$ are close to each other in the sense of Hellinger distance and concluding by a lemma in [21] that was used in a similar fashion to show a lower bound for linear trace reconstruction. It is first shown that conditioned on a 1 being deleted, a trace from x is equidistributed as a trace from y. Then explicit expressions for the probabilities that the trace is $10^a 10^b 10^c$ are computed and compared, yielding an upper bound on the Hellinger distance. The difference between the probabilities for x and y is proportional to the product of $(a-b)(b-c)(a-c)$ and a symmetric polynomial in a, b, c. Both x and y consist of three 1's separated by runs of 0's of approximate length n, so with high probability we have that a, b, c are approximately np, with square root fluctuations. The contribution of the $(a-b)(b-c)(a-c)$ term allows us to recover a $\tilde{\Omega}(n^3)$ bound.

1.4 Outline

In Section 2, we go over some preliminary definitions and results. In Section 3, we prove Theorem 2. In Section 4, we prove Theorem 4. In Section 5, we prove Theorem 6. Finally, in Section 6, we conclude by discussing open problems and avenues for further research. Some proofs, such as the proof of Proposition 1 and the full proof of Theorem 3, are deferred to Appendix A in the full version of this paper on arXiv.

2 Preliminaries

First, we explain a basic definition we will use involving complex numbers.

▶ **Definition 7.** *For $z \in \mathbb{C}$, let $|z|$ be the* magnitude *of z, and if $z \neq 0$, let $\arg z$ be the* argument *of z, which is the value of $\theta \in (-\pi, \pi]$ such that $\frac{z}{|z|} = e^{i\theta}$.*

Next, we state a Littlewood-type result about bounding polynomials on arcs of the unit circle.

▶ **Theorem 8** ([9])**.** *Let $f(z) = \sum_{j=0}^{n} a_j z^j$ be a nonzero polynomial of degree n with complex coefficients. Suppose there is some positive integer M such that $|a_0| \geq 1$ and $|a_j| \leq M$ for all $0 \leq j \leq n$. Then, if A is an arc of the unit circle $\{z \in \mathbb{C} : |z| = 1\}$ with length $0 < a < 2\pi$, there exists some absolute constant $c_1 > 0$ such that*

$$\sup_{z \in A} |f(z)| \geq \exp\left(\frac{-c_1(1 + \log M)}{a}\right).$$

Next, we state two well known results about roots of unity in cyclotomic fields.

▶ **Lemma 9** ([29])**.** *Let $\omega = e^{2\pi i/n}$. Then, the set of $\{\omega^k\}$ for $k \in \mathbb{Z}, \gcd(k, n) = 1$ are all* Galois conjugates*. This means that if $P(x)$ is an integer polynomial, then $P(\omega^k) = 0$ if and only if $P(\omega) = 0$ for any $k \in \mathbb{Z}$ with $\gcd(k, n) = 1$. Moreover, $P(\omega) = 0$ if and only if P is a multiple of the nth* Cyclotomic polynomial*.*

▶ **Lemma 10** ([29])**.** *Let $\omega = e^{2\pi i/n}$ be an nth root of unity, and let $\mathbb{Q}[\omega]$ be the nth degree cyclotomic field. Then, if $z \in \mathbb{Q}[\omega]$ is such that $z^r = 1$ for some integer $r \geq 1$, z must equal ω^k or $-\omega^k$ for some integer k.*

Finally, we define the Hellinger distance between two probability measures and state a folklore bound on distinguishing between distributions based on samples in terms of the Hellinger distance.

▶ **Definition 11.** *Let μ and ν be discrete probability measures over some set Ω. In other words, for $x \in \Omega$, $\mu(x)$ is the probability of selecting x when drawing from the measure μ. Then, the Hellinger distance is defined as*

$$d_H(\mu, \nu) = \left(\sum_{x \in \Omega} \left(\sqrt{\mu(x)} - \sqrt{\nu(x)}\right)^2\right)^{1/2}.$$

The following proposition is quite well-known (see for instance, [21, Lemma A.5]).

▶ **Proposition 12.** *If μ, ν are discrete probability measures, then if given i.i.d. samples from either μ or ν, one must see at least $\Omega(d_H(\mu, \nu)^{-2})$ i.i.d. samples to determine whether the distribution is μ or ν with at least $2/3$ success probability.*

3 Worst Case: Upper Bound

In this section, we prove Theorem 2, i.e., we provide an $\exp\left(\tilde{O}(n^{1/3})\right)$-sample algorithm for circular trace reconstruction when the length n is a prime or product of two primes.

For a (linear) string $x \in \{0,1\}^n$ and $z \in \mathbb{C}$, we define $P(z;x) := \sum_{i=1}^{n} x_i z^i$. The first lemma we require creates an unbiased estimator for $\prod_{i=1}^{m} P(z_i;x)$ for some complex numbers $z_1, \ldots, z_m$, using only random traces of x. The proof of the following lemma greatly resembles the proof of [31, Lemma 4.1], so we defer the proof to Appendix A in the full version of this paper on arXiv.

▶ **Lemma 13.** *Let x be a linear string of length n. Fix q as the deletion probability and $p = 1-q$ as the retention probability. Then, for any integer $m \geq 1$ and any $Z = (z_1, \ldots, z_m)$ for $z_1, \ldots, z_m \in \mathbb{C}$, there exists some function $g_m(\tilde{x}, Z)$ such that*

$$\mathbb{E}_{\tilde{x}}[g_m(\tilde{x}, Z)] = \prod_{k=1}^{m} \left(\sum_{i=1}^{n} x_i z_k^i \right),$$

where the expectation is over traces drawn from x. Moreover, for any $L \geq 1$, and for all $\tilde{x} \in \{0,1\}^n$ and all Z such that $|z_1|, \ldots, |z_m| = 1$ and $|\arg z_i| \leq \frac{1}{L}$ for all $1 \leq i \leq m$,

$$|g_m(\tilde{x}, Z)| \leq (p^{-1}mn)^{O(m)} \cdot e^{O(m^2 n/(p^2 L^2))}.$$

For $x \in \{0,1\}^n$ and $z \in \mathbb{C}$, let $P(z;x) := \sum_{i=1}^{n} x_i z^i$. Our main goal will be to determine the value of $f_t(z;x) := P(z;x)^t \cdot P(z^{-t};x)$ for some integer t, where z is an nth root of unity. Importantly, we note that $f_t(z;x)$ is invariant under rotations of x, since for $z = e^{2\pi i k/n}$,

$$\sum_{i=1}^{n} x_{(i+1) \pmod n} z^i = \sum x_i z^{i-1} = P(z;x) \cdot z^{-1}$$

whereas

$$\sum_{i=1}^{n} x_{(i+1) \pmod n} z^{-t \cdot i} = \sum x_i z^{-t(i-1)} = P(z^{-t};x) \cdot z^t.$$

Therefore, if we define $x^{(j)}$ as the string x rotated by j places (so $x_i^{(j)} = x_{(i+j) \pmod n}$), then $f(z;x) = f(z;x^{(j)})$ for all $z = e^{2\pi i k/n}$ and $0 \leq j \leq n-1$.

Now, choose some z with $|z| = 1$ and $|\arg z| \leq \frac{1}{L}$. Also, fix some integer t, let $m = t+1$, and let $Z = (\underbrace{z, \ldots, z}_{t \text{ times}}, z^{-t})$. Then, if j is randomly chosen in $\{0, 1, \ldots, n-1\}$ and $\tilde{x}$ is a random trace,

$$\begin{aligned}
\mathbb{E}_{\tilde{x}}[nz^{tn} \cdot g_m(\tilde{x}, Z)] &= (n \cdot z^{tn}) \cdot \left(\frac{1}{n} \cdot \sum_{j=0}^{n-1} P(z;x^{(j)})^t \cdot P(z^{-t};x^{(j)}) \right) \\
&= \sum_{j=0}^{n-1} z^{tn} \cdot P(z;x^{(j)})^t \cdot P(z^{-t};x^{(j)}),
\end{aligned}$$

where $g_m(\tilde{x}, Z)$ is defined in Lemma 13. Note that $\sum_{j=0}^{n-1} z^{tn} \cdot P(z;x^{(j)})^t \cdot P(z^{-t};x^{(j)})$ is a polynomial of z of degree at most $3tn$ and all coefficients bounded by n^{t+1}. We write this polynomial as $Q_t(z;x)$. Thus, if we define $h_t(\tilde{x}, z) := nz^{tn} g_m(\tilde{x}, Z)$, we have that $\mathbb{E}_{\tilde{x}}[h_t(\tilde{x}, z)] = Q(z;x)$ for $\tilde{x}$ a trace of a randomly shifted x, and that $|h_t(\tilde{x};z)| \leq (p^{-1}tn)^{O(t)} \cdot e^{O(t^2 n/(p^2 L^2))}$ whenever $|z| = 1$ and $|\arg z| \leq \frac{1}{L}$ for $L \geq 2$ and $m = t+1$, by Lemma 13.

Now, we will state two important results that will lead to the proof of the main result.

▶ **Lemma 14.** *Let $n \geq 2$, and suppose that x, x' are strings in $\{0,1\}^n$ such that $Q_t(z;x) \neq Q_t(z;x')$ as polynomials in z. Then, there is an absolute constant c_2 such that for any $L \geq 2$, there exists z such that $|z| = 1$, $|\arg z| \leq \frac{1}{L}$, and*

$$|Q_t(z;x) - Q_t(z;x')| \geq n^{-c_2 tL}.$$

Proof. Note that $Q_t(z;x) - Q_t(z;x')$ is a nonzero polynomial in z of degree at most $(t+1)n$ and with all coefficients bounded by $2n^{t+1}$. Therefore, by Theorem 8,

$$\begin{aligned}\sup_{|z|=1, |\arg z| \leq 1/L} |Q_t(z;x) - Q_t(z;x')| &\geq \exp\left(-\frac{c_1(1+\log(2n^{t+1}))}{2/L}\right) \\ &\geq \exp\left(-c_2 \cdot L \cdot t \cdot \log n\right) \\ &= n^{-c_2 tL},\end{aligned}$$

where we note that the arc $\{z : |z| = 1, |\arg z| \leq \frac{1}{L}\}$ has length $\frac{2}{L}$. ◀

The next important result we need will be Theorem 3. We defer the full proof of Theorem 3 to the full version of this paper on arXiv, but as the proof of the case where n is prime is simpler, we prove this special case here. Using this, we can get an $\exp\left(\tilde{O}(n^{1/3})\right)$ sample upper bound at least for n prime. As a note, we will define $\omega = e^{2\pi i/n}$ from now on, where n will be clear from context.

▶ **Proposition 15.** *Suppose that $n = p$ is prime, and $a_0, \ldots, a_{p-1}, b_0, \ldots, b_{p-1} \in \{0,1\}$ are such that for all $0 \leq k < p$, there is some integer c_k such that $\sum_{i=0}^{p-1} a_i = \omega^{c_k} \cdot \sum_{i=0}^{p-1} b_i$. Then, the sequences $\{a_1, \ldots, a_p\}$ and $\{b_1, \ldots, b_p\}$ are equivalent up to a cyclic permutation.*

Proof. First, $\sum_{i=0}^{p-1} a_i = \omega^{c_0} \cdot \sum_{i=0}^{p} b_i$. Since $\sum_{i=0}^{p-1} a_i$ and $\sum_{i=0}^{p-1} b_i$ are both nonnegative real numbers, and since ω^{c_0} is a root of unity, we must have that $\sum_{i=0}^{p-1} a_i = \sum_{i=0}^{p-1} b_i$.

Next, we have that $\sum_{i=0}^{p-1} a_i \omega^i = \omega^{c_1} \cdot \sum_{i=0}^{p-1} b_i \omega^i$. Letting $b'_i = b_{(i-c_1) \pmod p}$, we have that b' is a cyclic shift of b, and $\sum_{i=0}^{p-1} a_i = \sum_{i=0}^{p-1} b'_i$ and $\sum_{i=0}^{p-1} a_i \omega^i = \sum_{i=0}^{p-1} b'_i \omega^i$. Letting $Q(x) = \sum_{i=0}^{p-1} (a_i - b'_i) x^i$, we have that ω and 1 are both roots of $Q(x)$. Since $Q(x)$ is an integer-valued polynomial, this implies that all Galois conjugates of ω are roots, so $1, \omega, \omega^2, \ldots, \omega^{p-1}$ are roots of $Q(x)$. Thus, $x^p - 1$ divides $Q(x)$. But since $Q(x)$ has degree at most $p-1$, $Q(x)$ must equal 0, so $a_i = b'_i$ for all i. Since the sequence b' is just a shift of b, we are done. ◀

By using Theorem 3 (or Proposition 15 in the case of n prime), we obtain the following number theoretic result.

▶ **Lemma 16.** *Let n be a prime or a product of two primes, and let $a = a_1 a_2 \cdots a_n$ and $b = b_1 b_2 \cdots b_n$ be distinct n-bit strings (even up to cyclic shift). Then, for some $0 \leq \ell \leq n-1$ with $z = \omega^\ell$, and for some $2 \leq t \leq 5$, we have that $P(z;a)^t P(z^{-t};a) \neq P(z;b)^t P(z^{-t};b)$.*

Proof. First choose k such that $\sum_{i=1}^{n} a_i \omega^{i \cdot k} \neq \omega^{c_k} \cdot \sum_{i=1}^{n} b_i \omega^{i \cdot k}$ for all integers c_k, which exists by Theorem 3. If $k = 0$, then $P(\omega^k;a) = P(1;a)$ and $P(\omega^k;b) = P(1;b)$ are distinct nonnegative integers, so we trivially have $P(1;a)^t P(1;a) \neq P(1;b)^t P(1;b)$. Otherwise, let t be the smallest prime that doesn't divide $\frac{n}{\gcd(n,k)}$ (so $t \leq 5$ as n has at most 2 prime factors). If $\sum_{i=1}^{n} a_i \omega^{i \cdot k} = 0$, then $\sum_{i=1}^{n} b_i \omega^{i \cdot k} \neq 0$. Now, since ω^{-tk} is a Galois conjugate of ω^k (since $t \nmid n$), we also have that $\sum_{i=1}^{n} b_i \omega^{-ti \cdot k} \neq 0$. This means that $P(\omega^k;a) = 0$ so $P(\omega^k;a)^t P((\omega^k)^{-t};a) = 0$, but $P(\omega^k;b)^t P((\omega^k)^{-t};b) \neq 0$. Likewise, if $\sum_{i=1}^{n} b_i \omega^{i \cdot k} = 0$, we'll have $P(\omega^k;a)^t P((\omega^k)^{-t};a) \neq 0$, but $P(\omega^k;b)^t P((\omega^k)^{-t};b) = 0$. This means the result follows if either $P(\omega^k;a) = 0$ or $P(\omega^k, b) = 0$.

Otherwise, $P(\omega^k; a) = \sum_{i=1}^{n} a_i \omega^{i \cdot k}$ and $P(\omega^k; b) = \sum_{i=1}^{n} b_i \omega^{i \cdot k}$ are both nonzero. This means that for all $r \geq 0$, $P(\omega^{(-t)^r \cdot k}; a)$ and $P(\omega^{(-t)^r \cdot k}; b)$ are both nonzero, since $\omega^{(-t)^r \cdot k}$ and ω^k are Galois conjugates. This means that if $P(z; a)^t P(z^{-t}; a) = P(t; b)^t P(z^{-t}; b)$ for all $z = \omega^{(-t)^r \cdot k}$, then

$$\frac{P(\omega^{(-t)^{r+1} \cdot k}; a)}{P(\omega^{(-t)^r \cdot k}; a)^{-t}} = \frac{P(z^{-t}; a)}{P(z; a)^{-t}} = \frac{P(z^{-t}; b)}{P(z; b)^{-t}} = \frac{P(\omega^{(-t)^{r+1} \cdot k}; b)}{P(\omega^{(-t)^r \cdot k}; b)^{-t}}$$

for all $r \geq 0$, so we inductively have that

$$\frac{P(\omega^{(-t)^r \cdot k}; a)}{P(\omega^k; a)^{(-t)^r}} = \frac{P(\omega^{(-t)^r \cdot k}; b)}{P(\omega^k; b)^{(-t)^r}}.$$

Now, letting $r = \varphi\left(\frac{n}{\gcd(n,k)}\right)$, we know that $k \cdot (-t)^r \equiv k \pmod{n}$ by Euler's theorem, which means that $\omega^{(-t)^r \cdot k} = \omega^k$. Thus,

$$P(\omega^k; a)^{1-(-t)^r} = P(\omega^k; b)^{1-(-t)^r}.$$

Since $k \neq 0$, we have that $\frac{n}{\gcd(n,k)} > 1$ so $r \geq 1$. Thus, since $t \geq 2$, $1 - (-t)^r \neq 0$. Now, since $P(\omega^k; a), P(\omega^k; b)$ are nonzero, we have that $\frac{P(\omega^k; a)}{P(\omega^k; b)}$ is a $|1 - (-t)^r|^{\text{th}}$ root of unity. Also, $P(\omega^k; a), P(\omega^k; b) \in \mathbb{Q}[\omega]$, which means $\frac{P(\omega^k; a)}{P(\omega^k; b)} \in \mathbb{Q}[\omega]$. However, all roots of unity in $\mathbb{Q}[\omega]$ are of the form $\pm\omega^i$ for some i, and since $(-t)^r - 1$ is odd if n is odd (since $t = 2$), we must have that $\frac{P(\omega^k; a)}{P(\omega^k; b)} = \omega^{c_k}$ for some integer c_k. This is a contradiction, so we must have that $P(z; a)^t P(z^{-t}; a) \neq P(z; b)^t P(z^{-t}; b)$, for some $z = \omega^{(-t)^r \cdot k}$, $r \geq 0$. ◀

Finally, we are ready to prove Theorem 2.

Proof of Theorem 2. Suppose that we are trying to distinguish between the original circular string being $a = a_1 a_2 \cdots a_n$ or $b = b_1 b_2 \cdots b_n$, where a, b are distinct, even up to cyclic shifts.

We choose ℓ, t based on Lemma 16, so that $P(\omega^\ell; a)^t P((\omega^\ell)^{-t}; a) \neq P(\omega^\ell; b)^t P((\omega^\ell)^{-t}; b)$. As we have already noted, if z is an n^{th} root of unity, then $P(z; a)^t P(z^{-t}; a)$ is invariant under rotation of a, and $P(z; b)^t P(z^{-t}; b)$ is invariant under rotation of b. By our definition of $Q_t(z; x)$, we have that $Q_t(\omega^\ell; a) \neq Q_t(\omega^\ell; b)$, so $Q_t(z; a) \neq Q_t(z; b)$ as polynomials in z. Therefore, by Lemma 14, there is some z such that $|z| = 1, |\arg z| \leq \frac{1}{L}$, and

$$|Q_t(z; a) - Q_t(z; b)| \geq n^{-c_2 t L} \geq n^{-5 c_2 L}.$$

So, for $L = \left\lceil n^{1/3} (\log n)^{-1/3} p^{-2/3} \right\rceil$, there exists z with $|z| = 1$ and $|\arg z| \leq \frac{1}{L}$ and some $2 \leq t \leq 5$ such that

$$|Q_t(z; a) - Q_t(z; b)| \geq n^{-5 c_2 L} \geq \exp\left(-c_3 \cdot n^{1/3} (\log n)^{2/3} p^{-2/3}\right),$$

but

$$|h_t(\tilde{x}, z)| \leq (p^{-1} n)^{O(1)} \cdot \exp\left(O\left(\frac{n}{p^2 L^2}\right)\right) \leq \exp\left(c_4 \cdot n^{1/3} (\log n)^{2/3} p^{-2/3}\right)$$

for any trace $\tilde{x}$ of either a or b. Sample $R = \exp\left(O\left(n^{1/3} (\log n)^{2/3} p^{-2/3}\right)\right)$ traces $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(R)}$ and choose z and t based on Lemma 16. Then, if we define $h_t(z)$ to be the average of $h_t(\tilde{x}^{(i)}, z)$ over i from 1 to R, the Chernoff bound tells us that with failure probability at most 10^{-n}, $|h_t(z) - Q_t(z; a)| \leq \frac{1}{3} \cdot \exp\left(c_4 \cdot n^{1/3} (\log n)^{2/3} p^{-2/3}\right)$ if the original string were a,

and $|h(z) - Q_t(z;b)| \leq \frac{1}{3} \cdot \exp\left(c_4 \cdot n^{1/3}(\log n)^{2/3}p^{-2/3}\right)$ if the original string were b. Thus, by returning a if $h(z)$ is closer to $Q_t(z;a)$ and returning b otherwise, we can distinguish between the original string being a or b using $\exp\left(O\left(n^{1/3}(\log n)^{2/3}p^{-2/3}\right)\right)$ traces, with $1 - 10^{-n}$ failure probability.

Thus, to reconstruct the original string x, we simply run the distinguishing algorithm for all pairs $a, b \in \{0,1\}^n$ such that $a \neq b$, using the same R traces $\tilde{x}^1, \ldots, \tilde{x}^R$. With probability at least $1 - (4/10)^n \geq 1 - 2^{-n}$, the true string x will be the only string such that the distinguishing algorithm will successfully choose x over all other strings. Thus, for n a prime or a product of two primes, the circular trace reconstruction problem can be solved using $\exp\left(O\left(n^{1/3}(\log n)^{2/3}p^{-2/3}\right)\right)$ traces. ◀

4 Average Case: Upper Bound

We now consider the situation in which the unknown circular string x is random. For the sake of simplicity, we will assume that we may sample x by sampling a uniform random linear binary string of length n and applying a uniform random cyclic shift. Note that the resulting distribution on x is not uniform over all possible circular strings, as strings with nontrivial cyclic symmetries are more likely to appear. However, such strings form a negligible fraction of all strings, and our arguments can easily be modified to handle the situation in which the distribution is uniform over all possible circular strings. We use the randomness to rule out certain problematic strings with high probability, and this can be done for uniform random circular strings as well as other distributions, for example if in the sampling procedure each bit is independently biased towards 0 or 1.

▶ **Theorem 17.** *Let x be a random (in the sense described above) unknown circular string of length n and let q be the deletion probability of each element. Then there exists a constant C_q depending only on q such that we can determine x with failure probability at most n^{-10} using $O(n^{C_q})$ traces.*

In what follows, we will let $x = x_1 \cdots x_n$ and take indices of bits in x modulo n. Let $k = 100 \log n$. We first note that with high probability, all of the consecutive substrings of x of length k and $k-1$ are pairwise distinct. We will refer to such strings x as *regular* strings. Indeed, the probability that $x_i \cdots x_{i+k-1} = x_j \cdots x_{j+k-1}$ for $i \neq j$ is 2^{-k} (where indices are taken modulo n), and union bounding over all i, j as well as both k and $k-1$ gives a failure probability of at most $O(n^2 2^{-k}) \ll n^{-10}$.

If we assume that x is regular, the length k consecutive substrings of x uniquely determine x. Indeed, given $x_i \cdots x_{i+k-1}$, we can uniquely determine x_{i+k} as there is a unique length k consecutive substring of x that begins with $x_{i+1} \cdots x_{i+k-1}$. Iteratively applying this allows us to recover the entire string x. Thus, to prove Theorem 17, it suffices to prove Lemma 5, i.e., to determine how many times each length k substring appears consecutively in any string x using $O(n^{C_q})$ traces, which will allow us to recover x if x is regular.

Proof of Lemma 5. For a circular string x, let $S_x = \{x_i x_{i+1} \cdots x_{i+k-1}\}_{i=1}^n$ be the *k-deck* of x, which is the multiset of contiguous substrings of x of length k. Let S and T be the k-decks of some circular strings of length n such that $S \neq T$. Suppose that given $O(n^{C_q})$ traces of a circular string, we are able to distinguish between whether its k-deck is S or T correctly with failure probability 10^{-n}. Then, by union bounding over all possible pairs of distinct k-decks (note that there are at most 2^n different k-decks), we can with high probability correctly determine the k-deck of the string, showing the lemma.

The key property we will use is that given two distinct k-decks S and T that come from circular strings x and y, there exists some string s of length k such that the number of consecutive occurrences of s in x and in y are different. We will show the existence of C_q so that for any string s of length k satisfying this property, we can distinguish between x and y correctly using $O(n^{C_q})$ samples with failure probability 10^{-n}, from which the result follows.

Let α denote a sufficiently large positive integer only depending on q that we will determine later. For $0 \le i \le n-k$, let c_i denote the number of (not necessarily consecutive) occurrences of s in x contained in a consecutive substring of x of length at most $i+k$. Similarly, let d_i denote the number of (not necessarily consecutive) occurrences of s in y contained in a consecutive substring of y of length at most $i+k$. By assumption, we have that $c_0 \neq d_0$. By casework on the last bit of the occurrence of s, we have that $c_i, d_i \le n\binom{i+k}{k}$. Let $P(t) = \sum_{i=0}^{\alpha k} c_i t^i$ and $Q(t) = \sum_{i=0}^{\alpha k} d_i t^i$. Moreover, the following is true:

▶ **Lemma 18.** *The probability that a trace of x starts with s (where a random bit in the string is chosen as the beginning before bits are deleted) is $\frac{1}{n}(1-q)^k P(q) + O(q^{\alpha k}(\alpha+1)^k e^k)$. Similarly, the probability that a trace of y starts with s is $\frac{1}{n}(1-q)^k Q(q) + O(q^{\alpha k}(\alpha+1)^k e^k)$.*

Proof. To compute the probability that a trace of x starts with s, we do casework on how many bits are deleted before the last bit in the occurrence of s. If i bits are deleted, then note that there are c_i ways for it to be done by definition. Each such way has a probability of $\frac{1}{n}(1-q)^k q^i$ to occur. Indeed, for each way there is a $\frac{1}{n}$ probability that the correct starting bit is chosen, and the probability that only the bits corresponding to the specific instance of s are kept is $(1-q)^k q^i$. It follows that the probability is exactly $\frac{1}{n}(1-q)^k \sum_{i=0}^{n-k} c_i q^i$.

It remains to show that $\frac{1}{n}(1-q)^k \sum_{\alpha k+1}^{n-k} c_i q^i = O(q^{\alpha k}(\alpha+1)^k e^k)$. As mentioned before, we have that $c_i \le n\binom{i+k}{k}$. Thus, this sum is at most $\sum_{i>\alpha k} \binom{i+k}{k} q^i \le \binom{\alpha k+k}{k} q^{\alpha k} \sum_{i \ge 0} \left(\frac{q(\alpha+1)}{\alpha}\right)^i$. Indeed, the ratio of consecutive terms in the sequence $\binom{i+k}{k} q^i$ is equal to $q\frac{i+k}{i} \le \frac{q(\alpha+1)}{\alpha}$. For a sufficiently large choice of α, $\frac{q(\alpha+1)}{\alpha} < 1$, so $\sum_{i>\alpha k} \binom{i+k}{k} q^i = O(\binom{\alpha k+k}{k} q^{\alpha k}) = O(q^{\alpha k}(\alpha+1)^k e^k)$ by Stirling's approximation.

The argument for y is analogous. ◀

Lemma 18 allows us to estimate $P(q)$ and $Q(q)$ up to an $O(n(1-q)^{-k} q^{\alpha k}(\alpha+1)^k e^k)$ error by looking at how often traces of x or y begin with s, and then dividing by $\frac{1}{n}(1-q)^k$. So long as $P(q)$ and $Q(q)$ are sufficiently far apart, a Chernoff bound allows us to determine with high probability if the traces came from x or y. However, it may be the case that $P(q)$ and $Q(q)$ are quite close. To remedy this, we observe that it is possible to *simulate higher deletion probabilities* $q' > q$. Indeed, this can be achieved by deleting each bit in traces received independently with probability $\frac{q'-q}{1-q}$. Thus, it suffices to find $q' \in [q, r]$ with $P(q')$ and $Q(q')$ far apart for some $q < r < 1$. The existence of such a q' is proven by the following Littlewood-type result of Borwein, Erdélyi, and Kós.

▶ **Theorem 19** ([10], Theorem 5.1)**.** *There exist absolute constants $c_1 > 0$ and $c_2 > 0$ such that if f is a polynomial with coefficients in $[-1, 1]$ and $a \in (0, 1]$, then*

$$|f(0)|^{c_1/a} \le \exp\left(\frac{c_2}{a}\right) \sup_{z \in [1-a,1]} |f(z)|.$$

Let $r = \frac{q+1}{2}$. We first apply Theorem 19 to $\binom{\alpha k+k}{k}^{-1}(P(rx) - Q(rx))$ and $a = 1 - q/r$. Here, we are using the fact that the coefficients of P and Q are bounded in magnitude by $\binom{\alpha k+k}{k}$ by previous observations, and that $|P(0) - Q(0)| \ge 1$. Theorem 19 tells us that

$$\binom{\alpha k+k}{k}^{-c_1/a} \leq \exp\left(\frac{c_2}{a}\right)\binom{\alpha k+k}{k}^{-1} \sup_{z\in[1-a,1]} |P(rz)-Q(rz)|$$
$$= \exp\left(\frac{c_2}{a}\right)\binom{\alpha k+k}{k}^{-1} \sup_{q'\in[q,r]} |P(q')-Q(q')|,$$

or

$$\sup_{q'\in[q,r]} |P(q')-Q(q')| \geq c_3\binom{\alpha k+k}{k}^{-c_4}$$

for some constants c_3 and c_4 that only depend on q.

In particular, this is much larger than $10^k n(1-r)^{-k}r^{\alpha k}(\alpha+1)^k e^k$ for sufficiently large values of α (α may depend on q). Indeed, after taking kth roots and using Stirling's approximation this reduces to showing that $(e(\alpha+1))^{-c_5} > 10n^{1/k}(1-r)^{-1}r^{\alpha}(\alpha+1)e$ for sufficiently large α where c_5 is some constant that only depends on q, which is clear (since $0<r<1$ is fixed and $n^{1/k}<2$). Thus, for any $q'\in[q,r]$, the error term $\frac{1}{n}(1-q')^k\sum_{\alpha k+1}^{n-k} c_i(q')^i = O((q')^{\alpha k}(\alpha+1)^k e^k)$ is at most 10^{-k} times $\frac{1}{n}(1-q')^k\cdot\sup_{q'\in[q,r]}|P(q')-Q(q')|$.

Hence, for some $q'\in[q,r]$, the probability that a trace begins with s under bit deletion with probability q' differs between x and y by $\Omega(10^k n(1-r)^{-k}r^{\alpha k}(\alpha+1)^k e^k) = \Omega(n^{-c_6})$ for some constant c_6 that only depends on q. By a standard Chernoff bound, for some constant C_q only depending on q, we can distinguish between x and y using $O(n^{C_q})$ traces with failure probability at most $\exp(-\Omega(n))$, so Lemma 5 follows. ◀

By the previous discussions, Theorem 4 is also proven.

As mentioned before, Chen et. al. independently proved the analogue of Lemma 5 for linear strings in [15, Theorem 2]. The ideas behind their result and ours are similar: both are proved by considering a polynomial encoding the k-deck and using complex analysis to bound the corresponding polynomials for different k-decks away from each other on $[q,1]$. Here, we directly apply the Littlewood-type result from [10], while Chen et. al. prove their own result for this bound. This idea of reconstructing the k-deck of the unknown string may be useful in other variants of trace reconstruction, though the sample complexity it achieves is only polynomial and in order to apply it, we must be able to reconstruct the string from its k-deck. For worst case trace reconstruction, we must be able to reconstruct strings not uniquely determined by their k-decks, and for average case trace reconstruction, a subpolynomial sample complexity has already been achieved for constant deletion probabilities. Thus, these ideas are not directly applicable to worst case and average case trace reconstruction, though they may be helpful for variants in which it suffices to construct k-decks and a polynomial sample complexity is not known, such as the ones considered in this paper and in [15].

One difference between our result and that of Chen et. al. is that while we only addressed the sample complexity, they also give a polynomial time algorithm for reconstructing the string via a linear program. Their approach can be modified to give a polynomial time algorithm for average case circular trace reconstruction as well. For more details on their algorithm, see [15, Section 6]. Moreover, while we do not address the smoothed complexity model as in Chen et. al., our proof of Theorem 4 easily generalizes to a polynomial sample (or time) algorithm for circular trace reconstruction in the smoothed complexity model. This is because one can show that a circular string drawn from Chen et. al.'s smoothed model is regular with very high probability, in a similar way to how we showed an average string is regular. For more details, see [15, Section 3].

5 Worst Case: Lower Bound

In this section, we prove Theorem 6 and demonstrate that worst-case circular trace reconstruction requires $\tilde{\Omega}(n^3)$ traces. We first record the following lemma from [21] expressing the number of independent samples required to distinguish between two probability measures μ and ν in terms of their Hellinger distance $d_H(\mu,\nu)$, defined to be $\left(\sum_{x\in X}\left(\sqrt{\mu(\{x\})}-\sqrt{\nu(\{x\})}\right)^2\right)^{1/2}$ where the sum is over all events in some discrete sample space X. Let $d_{TV}(\mu,\nu)$ denote the total variation distance between μ and ν and μ^n denote the law of n independent samples from μ.

▶ **Lemma 20** ([21], Lemma A.5). *If μ and ν are probability measures satisfying $d_H(\mu,\nu)\le 1/2$, then $1-d_{TV}(\mu^m,\nu^m)\ge\varepsilon$ if $m\le\frac{\log(1/\varepsilon)}{9d_H^2(\mu,\nu)}$.*

Given m traces, we cannot determine if they came from μ^m or ν^m with probability higher than $\frac{1}{2}\left(1+d_{TV}(\mu^m,\nu^m)\right)$. Thus, it requires $\Omega(d_H^{-2}(\mu,\nu))$ samples to distinguish between two probability measures μ and ν with probability greater than $\frac{3}{4}$.

Proof of Theorem 6. We specialize to the case of distinguishing between $x=10^n10^{n+1}10^{n+k}$ and $y=10^n10^{n+k}10^{n+1}$ from independent traces. Let μ and ν respectively denote the laws of traces from x and y. We will show that $d_H^2(\mu,\nu)=O((\log n/n)^3)$, which establishes the result by Lemma 20.

First, we note that conditional on the first 1 in x being deleted, the resulting trace is equidistributed as a trace from y conditioned on the second 1 being deleted, as in both cases we obtain a trace from the circular string $10^{n+1}10^{2n+k}$. Similar arguments for other cases show that conditioned on any 1 being deleted, traces from x and y are equal in law. Thus, the resulting string must have three 1's to contribute to the Hellinger distance. We will henceforth assume that the resulting trace is of the form $10^a10^b10^c$ for some nonnegative integers a,b,c.

We now compute the ratio $\frac{\mu(\{10^a10^b10^c\})}{\nu(\{10^a10^b10^c\})}$ and show that it is typically $1+O((\log n/n)^{3/2})$. We have that

$$\frac{\mu(\{10^a10^b10^c\})}{q^{3n+k+1-a-b-c}(1-q)^{a+b+c}}=\binom{n}{a}\binom{n+1}{b}\binom{n+k}{c}+\binom{n}{b}\binom{n+1}{c}\binom{n+k}{a}+\binom{n}{c}\binom{n+1}{a}\binom{n+k}{b},$$

$$\frac{\nu(\{10^a10^b10^c\})}{q^{3n+k+1-a-b-c}(1-q)^{a+b+c}}=\binom{n}{a}\binom{n+k}{b}\binom{n+1}{c}+\binom{n}{b}\binom{n+k}{c}\binom{n+1}{a}+\binom{n}{c}\binom{n+k}{a}\binom{n+1}{b}.$$

It follows that

$$\frac{\mu(\{10^a10^b10^c\})}{\nu(\{10^a10^b10^c\})}=\frac{\frac{1}{(n+1-b)(n+1-c)\cdots(n+k-c)}+\frac{1}{(n+1-c)(n+1-a)\cdots(n+k-a)}+\frac{1}{(n+1-a)(n+1-b)\cdots(n+k-b)}}{\frac{1}{(n+1-c)(n+1-b)\cdots(n+k-b)}+\frac{1}{(n+1-a)(n+1-c)\cdots(n+k-c)}+\frac{1}{(n+1-b)(n+1-a)\cdots(n+k-a)}}.$$

Multiplying the numerator and denominator by $\prod_{i=1}^{k}(n+i-a)(n+i-b)(n+i-c)$ results in

$$S_1=\prod_{i=1}^{k}(n+i-a)\prod_{i=2}^{k}(n+i-b)+\prod_{i=1}^{k}(n+i-b)\prod_{i=2}^{k}(n+i-c)+\prod_{i=1}^{k}(n+i-c)\prod_{i=2}^{k}(n+i-a)$$

and

$$S_2=\prod_{i=1}^{k}(n+i-b)\prod_{i=2}^{k}(n+i-a)+\prod_{i=1}^{k}(n+i-c)\prod_{i=2}^{k}(n+i-b)+\prod_{i=1}^{k}(n+i-a)\prod_{i=2}^{k}(n+i-c),$$

respectively. We have that $S_1 - S_2 = (a-b)\prod_{i=2}^{k}(n+i-a)(n+i-b) + (b-c)\prod_{i=2}^{k}(n+i-b)(n+i-c) + (c-a)\prod_{i=2}^{k}(n+i-c)(n+i-a)$. This is an alternating polynomial in a, b, c, i.e., applying a permutation σ to a, b, c changes the sign of the polynomial by the sign of σ. Hence, it can be written in the form $(a-b)(b-c)(a-c)P_k(n, a, b, c)$, where P_k is a polynomial in n, a, b, c of degree $2k-4$ since S_1 and S_2 have degree $2k-1$.

By a standard Chernoff bound, there exists a constant C such that with probability at least $1 - n^{-100}$, $a, b, c \in [np - C\sqrt{n\log n}, np + C\sqrt{n\log n}]$. When this occurs, we have that $S_2 = \Omega(n^{2k-1})$ and $|S_1 - S_2| = O((n\log n)^{3/2}n^{2k-4})$, so $\frac{\mu(\{10^a10^b10^c\})}{\nu(\{10^a10^b10^c\})} \in [1-(c\log n/n)^{3/2}, 1+(c\log n/n)^{3/2}]$ for some constant c. We thus have that

$$d_H^2(\mu,\nu) = \sum_{a,b,c\geq 0}\left(\sqrt{\mu(\{10^a10^b10^c\})} - \sqrt{\nu(\{10^a10^b10^c\})}\right)^2$$

$$\leq 2n^{-100} + \sum_{a,b,c\in[np-C\sqrt{n\log n}, np+C\sqrt{n\log n}]} \nu(\{10^a10^b10^c\})\left(1 - \sqrt{\frac{\mu(\{10^a10^b10^c\})}{\nu(\{10^a10^b10^c\})}}\right)^2$$

$$= O((\log n/n)^3).$$

It follows by Lemma 20 that it requires $\Omega(n^3/\log^3 n)$ samples to distinguish between traces from x and y, as desired. ◀

6 Conclusion and Future Work

We note that our work leaves several open problems, including the following:

1. As noted in the introduction, Chase [14] very recently improved the worst-case linear trace reconstruction bound to $\exp\left(\tilde{O}(n^{1/5})\right)$. Is it possible to get a matching circular trace reconstruction bound, even just for certain lengths of strings?
2. For worst-case strings, can one get an upper bound of $\exp\left(\tilde{O}(n^{1/3})\right)$ or even a subexponential bound for n with an arbitrary prime factorization?
3. Can one get a subpolynomial (i.e., $n^{o(1)}$) upper bound for the average case?
4. Can one improve our current lower bound for worst-case strings, perhaps even to $n^{\omega(1)}$?
5. All of the work we have done in this paper has primarily focused on traces with constant deletion probability q. However, if $q = o(1)$, the implied results are no better than the bounds we get for fixed $0 < q < 1$. Can better bounds be obtained for circular trace reconstruction with small deletion probability (for instance, $q = 1/(\log n)^2$, or even $q = n^{-2/3}$)? In the linear case, there exist much better trace reconstruction algorithms in the low deletion probability regime (e.g., [7, 24]), so perhaps these results can be extended to circular strings.

For answering open problem 2, one method we attempted for getting a $\exp\left(\tilde{O}(n^{1/2})\right)$ upper bound was to look at the polynomial $P(y)P(z)P(y^{-1}z^{-1})$ for cyclotomic nth roots of unity y, z. One can establish a "Bivariate Littlewood"-type result and the same argument as ours to show the following. Suppose that for any $a, b \in \{0,1\}^n$, $P(y;a)P(z;a)P(y^{-1}z^{-1};a) = P(y;b)P(z;b)P(y^{-1}z^{-1};b)$ for all cyclotomic nth roots of unity y, z implies that a, b are equivalent up to cyclic rotation. Then, one can solve circular trace reconstruction using $\exp\left(\tilde{O}(n^{1/2})\right)$ traces. This result may in fact look obvious, as if $P(\omega_n;a) = \alpha \cdot P(\omega_n;b)$, one should expect via a simple induction argument that $P(\omega_n^k;a) = \alpha^k \cdot P(\omega_n^k;b)$ which implies that $\alpha = e^{2\pi ir/n}$ for some r (by looking at $k = n$). Thus, by rotating a by r elements, we will get that $P(\omega_n^k;a) = P(\omega_n^k;b)$ for all k, which implies $a = b$. Unfortunately, one can have

cases where $P(z;a) = P(z;b) = 0$ for several choices of $z = e^{2\pi ik/n}$, which can cause this induction argument to fail. Indeed, we believe that if such a result were true, proving it would again be a challenging number theoretic task.

We already noted some reasons in the introduction for why modifying the results of [35, 22] to answer open problem 3 is difficult. The main reason was that one cannot efficiently find the "start" of the string.

Finally, we note that a potential way of answering open problem 4 is via strings based on [21, 13]. One cannot use their strings directly, as their strings are equivalent up to a cyclic rotation, but perhaps an appropriate modification may improve upon our $\tilde{\Omega}(n^3)$ lower bound.

References

1 Circular DNA. Available at https://en.wikipedia.org/wiki/Circular_DNA.

2 Alexandr Andoni, Assaf Goldberger, Andrew McGregor, and Ely Porat. Homomorphic fingerprints under misalignments: sketching edit and shift distances. In *Symposium on Theory of Computing (STOC)*, pages 931–940, 2013. doi:10.1145/2488608.2488726.

3 Frank Ban, Xi Chen, Adam Freilich, Rocco A. Servedio, and Sandip Sinha. Beyond trace reconstruction: Population recovery from the deletion channel. In *Foundations of Computer Science (FOCS)*, pages 745–768, 2019. doi:10.1109/FOCS.2019.00050.

4 Frank Ban, Xi Chen, Rocco A. Servedio, and Sandip Sinha. Efficient average-case population recovery in the presence of insertions and deletions. In *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques*, pages 44:1–44:18, 2019. doi:10.4230/LIPIcs.APPROX-RANDOM.2019.44.

5 Afonso S. Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *Innovations in Theoretical Computer Science (ITCS)*, pages 745–768, 2019. doi:10.1145/2554797.2554839.

6 Afonso S. Bandeira, Jonathan Niles-Weed, and Philippe Rigollet. Optimal rates of estimation for multi-reference alignment. *Mathematical Statistics and Learning*, 2(1):25–75, 2019. doi:10.4171/MSL/11.

7 Tugkan Batu, Sampath Kannan, Sanjeev Khanna, and Andrew McGregor. Reconstructing strings from random traces. In *Symposium on Discrete Algorithms (SODA)*, pages 910–918, 2004. URL: http://dl.acm.org/citation.cfm?id=982792.982929.

8 Vinnu Bhardwaj, Pavel A. Pevzner, Cyrus Rashtchian, and Yana Safonova. Trace reconstruction problems in computational biology. *IEEE Transactions on Information Theory*, pages 1–1, 2020. doi:10.1109/TIT.2020.3030569.

9 Peter Borwein and Tamás Erdélyi. Littlewood-type polynomials on subarcs of the unit circle. *Indiana University Mathematics Journal*, 46(4):1323–1346, 1997. doi:10.1512/IUMJ.1997.46.1435.

10 Peter Borwein, Tamás Erdélyi, and Géza Kós. Littlewood-type problems on $[0, 1]$. *Proceedings of the London Mathematical Society*, 3(79):22–46, 1999. doi:10.1112/S0024611599011831.

11 Joshua Brakensiek, Ray Li, and Bruce Spang. Coded trace reconstruction in a constant number of traces. In *Foundations of Computer Science (FOCS)*, 2020. arXiv:1908.03996.

12 Panagiotis Charalampopoulos, Tomasz Kociumaka, Solon P. Pissis, Jakub Radoszewski, Wojciech Rytter, Juliusz Straszyński, Tomasz Waleń, and Wiktor Zuba. Circular pattern matching with k mismatches. *Journal of Computer and System Sciences*, 115:73–85, 2021. doi:10.1016/j.jcss.2020.07.003.

13 Zachary Chase. New lower bounds for trace reconstruction. *CoRR*, abs/1905.03031, 2019. arXiv:1905.03031.

14 Zachary Chase. New upper bounds for trace reconstruction. *CoRR*, abs/2009.03296, 2020. arXiv:2009.03296.

15 Xi Chen, Anindya De, Chin Ho Lee, Rocco A. Servedio, and Sandip Sinha. Polynomial-time trace reconstruction in the smoothed complexity model. *CoRR*, abs/2008.12386, 2020. To appear in Symposium on Discrete Algorithms (SODA), 2021. arXiv:2008.12386.

16 Mahdi Cheraghchi, Ryan Gabrys, Olgica Milenkovic, and João Ribeiro. Coded trace reconstruction. *IEEE Trans. Inf. Theory*, 66(10):6084–6103, 2020. doi:10.1109/TIT.2020.2996377.

17 George M. Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in DNA. *Science*, 337(6102):1628, 2012. doi:10.1126/science.1226355.

18 Sami Davies, Miklos Racz, and Cyrus Rashtchian. Reconstructing trees from traces. In *Conference On Learning Theory (COLT)*, pages 961–978, 2019. URL: http://proceedings.mlr.press/v99/davies19a.html.

19 Anindya De, Ryan O'Donnell, and Rocco A. Servedio. Optimal mean-based algorithms for trace reconstruction. *Annals of Applied Probability*, 29(2):851–874, 2019. doi:10.1214/18-AAP1394.

20 Lisa Hartung, Nina Holden, and Yuval Peres. Trace reconstruction with varying deletion probabilities. In *Analytic Algorithmics and Combinatorics (ANALCO)*, pages 54–61, 2018. doi:10.1137/1.9781611975062.6.

21 Nina Holden and Russell Lyons. Lower bounds for trace reconstruction. *Annals of Applied Probability*, 30(2):503–525, 2020. doi:10.1214/19-AAP1506.

22 Nina Holden, Robin Pemantle, and Yuval Peres. Subpolynomial trace reconstruction for random strings and arbitrary deletion probability. In *Conference On Learning Theory (COLT)*, pages 1799–1840, 2018. URL: http://proceedings.mlr.press/v75/holden18a.html.

23 Thomas Holenstein, Michael Mitzenmacher, Rina Panigrahy, and Udi Wieder. Trace reconstruction with constant deletion probability and related results. In *Symposium on Discrete Algorithms (SODA)*, pages 389–398, 2008. doi:10.1145/1347082.1347125.

24 Sampath Kannan and Andrew McGregor. More on reconstructing strings from random traces: insertions and deletions. In *International Symposium on Information Theory (ISIT)*, pages 297–301, 2005. doi:10.1109/ISIT.2005.1523342.

25 Akshay Krishnamurthy, Arya Mazumdar, Andrew McGregor, and Soumyabrata Pal. Trace reconstruction: Generalized and parameterized. In *European Symposium on Algorithms (ESA)*, pages 68:1–68:25, 2019. doi:10.4230/LIPIcs.ESA.2019.68.

26 Vladimir I. Levenshtein. Efficient reconstruction of sequences. *IEEE Trans. Information Theory*, 47(1):2–22, 2001. doi:10.1109/18.904499.

27 Vladimir I. Levenshtein. Efficient reconstruction of sequences from their subsequences or supersequences. *J. Comb. Theory, Ser. A*, 93(2):310–332, 2001. doi:10.1006/jcta.2000.3081.

28 Maurice Maes. On a cyclic string-to-string correction problem. *Information Processing Letters*, 35(2):73–78, 1990. doi:10.1016/0020-0190(90)90109-B.

29 Daniel A. Marcus. *Number Fields*. Springer International Publishing, 1977.

30 Andrew McGregor, Eric Price, and Sofya Vorotnikova. Trace reconstruction revisited. In *European Symposium on Algorithms (ESA)*, pages 689–700, 2014.

31 Shyam Narayanan. Population recovery from the deletion channel: Nearly matching trace reconstruction bounds. *CoRR*, abs/2004.06828, 2020. To appear in Symposium on Discrete Algorithms (SODA), 2021. arXiv:2004.06828.

32 Fedor Nazarov and Yuval Peres. Trace reconstruction with $\exp(o(n^{1/3}))$ samples. In *Symposium on Theory of Computing (STOC)*, pages 1042–1046, 2017. doi:10.1145/3055399.3055494.

33 Hoang Hiep Nguyen, Jeho Park, Seon Joo Park, Chang-Soo Lee, Seungwoo Hwang, Yong-Beom Shin, Tai Hwan Ha, and Moonil Kim. Long-term stability and integrity of plasmid-based dna data storage. *Polymers*, 10(1):28, 2018. doi:10.3390/polym10010028.

34 Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, , and Karin Strauss. Random access in large-scale dna data storage. *Nature Biotechnology*, 36:242–248, 2018. doi:10.1038/nbt.4079.

35 Yuval Peres and Alex Zhai. Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice. In *Foundations of Computer Science (FOCS)*, pages 228–239, 2017. doi:10.1109/FOCS.2017.29.

36 Amelia Perry, Jonathan Weed, Afonso S. Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multireference alignment. *SIAM Journal on Mathematics of Data Science*, 1(3):497–517, 2019. doi:10.1137/18M1214317.

37 Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Robert B. Jackson. *Campbell Biology*. Pearson, 9th edition, 2011.

38 Krishnamurthy Viswanathan and Ram Swaminathan. Improved string reconstruction over insertion-deletion channels. In *Symposium on Discrete Algorithms (SODA)*, pages 399–408, 2008. doi:10.1145/1347082.1347126.