Mean-Based Trace Reconstruction over Practically any Replication-Insertion Channel

Mahdi Cheraghchi, Joseph Downs, João Ribeiro, and Alexandra Veliche

***Abstract*—Mean-based reconstruction is a fundamental, natural approach to worst-case trace reconstruction over channels with synchronization errors. It is known that $\exp(O(n^{1/3}))$ traces are necessary and sufficient for mean-based worst-case trace reconstruction over the deletion channel, and this result was also extended to certain channels combining deletions and geometric insertions of uniformly random bits. In this work, we use a simple extension of the original complex-analytic approach to show that these results are examples of a much more general phenomenon: $\exp(O(n^{1/3}))$ traces suffice for mean-based worst-case trace reconstruction over any memoryless channel that maps each input bit to an arbitrarily distributed sequence of replications and insertions of random bits, provided the length of this sequence follows a sub-exponential distribution.**

I. Introduction

When any length-n message $x \in \{-1,1\}^n$ is sent through a noisy channel Ch, the channel modifies the input x in some way to produce a distorted copy of x, called a *trace*. The goal of worst-case trace reconstruction over Ch is to design a reconstruction algorithm which recovers any input string $x \in \{-1,1\}^n$ with high probability from as few independent and identically distributed (i.i.d.) traces as possible. This problem was first introduced by Levenshtein [1], [2], who studied it over combinatorial channels causing synchronization errors, such as deletions and insertions of symbols and certain discrete memoryless channels. Trace reconstruction over the *deletion channel*, which independently deletes each input symbol with some probability, was first considered by Batu, Kannan, Khanna, and McGregor [3]. Some of their results were quickly generalized to what we call the *geometric insertion-deletion channel* [4], [5], which prepends a geometric number of independent, uniformly random symbols to each input symbol and then deletes it with a given probability. Both the deletion and geometric insertion-deletion channels are examples of discrete memoryless synchronization channels [6], [7].

Holenstein, Mitzenmacher, Panigrahy, and Wieder [8] were the first to obtain non-trivial worst-case trace reconstruction algorithms for the deletion channel with constant deletion probability. They showed that $\exp(\widetilde{O}(\sqrt{n}))$ traces suffice for *mean-based* reconstruction of any input string with high probability. By mean-based reconstruction, we mean that the reconstruction algorithm only requires knowledge of the expected value of each trace coordinate. In general, this procedure works as follows: Let $Y_x = (Y_{x,1}, Y_{x,2}, \dots)$ denote the *trace distribution* on input $x \in \{-1,1\}^n$ and Y'_x denote the infinite string obtained by padding Y_x with zeros on the right. The *mean trace* μ_x is given by

$$\mu_x = (\mathbb{E}[Y'_{x,1}], \mathbb{E}[Y'_{x,2}], \dots).$$

As the first step, the algorithm estimates μ_x from t traces $T^{(1)}, T^{(2)}, \dots, T^{(t)}$ sampled i.i.d. according to Y'_x via the empirical means

$$\widehat{\mu}_i = \frac{1}{t}\sum_{j=1}^{t} T_i^{(j)}, \quad i = 1, 2, \dots. \tag{1}$$

Subsequently, it outputs the string $\widehat{x} \in \{-1,1\}^n$ that minimizes $\|\mu_{\widehat{x}} - \widehat{\mu}\|_1$. If $t = t(n)$ is large enough, we have $\widehat{x} = x$ with high probability over the randomness of the traces. Because of their structure, pinpointing the number of traces required for mean-based reconstruction over any channel reduces to bounding $\|\mu_x - \mu_{x'}\|_1$ for any pair of distinct strings x and x'. Overall, mean-based reconstruction is a natural paradigm, and it is not only useful over channels with synchronization errors. For example, $O(\log n)$ traces suffice for mean-based reconstruction over the binary symmetric channel, which is optimal.

More recently, an elegant complex-analytic approach was employed concurrently by De, O'Donnell, and Servedio [9] and by Nazarov and Peres [10] to show that $\exp(O(n^{1/3}))$ traces suffice for mean-based worst-case trace reconstruction not only over the deletion channel with constant deletion probability, but also over the more general geometric insertion-deletion channel we described previously.[1] Remarkably, $\exp(\Omega(n^{1/3}))$ traces were shown to also be necessary for mean-based reconstruction over the deletion channel.

Given the fundamental nature of mean-based reconstruction and this state of affairs, the following question arises naturally: *Are these results examples of a much more general phenomenon?* In particular, is it true that $\exp(O(n^{1/3}))$ traces suffice for mean-based trace reconstruction over *any* discrete memoryless synchronization channel? We make significant progress in this direction by showing that a simple extension of the analysis from [9], [10] yields that result for a much broader

M. Cheraghchi, J. Downs, and A. Veliche are with the EECS Department, University of Michigan – Ann Arbor, MI, USA. Emails: {mahdich, josdowns, aveliche}@umich.edu.

J. Ribeiro is with the Department of Computing, Imperial College London, UK. Email: j.lourenco-ribeiro17@imperial.ac.uk.

This material is based upon work supported by the National Science Foundation under Grant No. CCF-2006455.

[1] Nazarov and Peres [10] consider a slightly modified geometric insertion-deletion channel: First, a geometric number of independent, uniformly random symbols is added independently before each input symbol. Then, the resulting string is sent through a deletion channel. The analysis is similar to that of the geometric-insertion channel.

class of such channels which map each input symbol to an *arbitrarily distributed* sequence of noisy symbol replications and insertions of random symbols, under a mild assumption.

Research in this direction has other practical and theoretical implications. First, studying trace reconstruction over channels introducing more complex synchronization errors than simple i.i.d. deletions is fundamental for the design of reliable DNA-based data storage systems with nanopore-based sequencing [11], [12], [13]. Second, understanding the structure of the mean trace of a string is by itself a natural information-theoretic problem which may lead to improved capacity bounds and coding techniques for channels with synchronization errors, both notoriously difficult problems (see the extensive surveys [14], [15], [7], [16]).

A. Related Work

Besides the works mentioned above, there has been significant recent interest in various notions of trace reconstruction.

The mean-based approach of [9], [10] has proven useful to some problems incomparable to our general setting: the deletion channel with position- and symbol-dependent deletion probabilities satisfying strong monotonicity and periodicity assumptions [17]; a combination of the geometric insertion-deletion channel and random shifts of the output string as an intermediate step in the design of *average-case* trace reconstruction algorithms (which are only required to have low *average* reconstruction error probability) [18], [19]; trace reconstruction of trees with i.i.d. deletions of vertices [20]; trace reconstruction of matrices with i.i.d. deletions of rows and columns [21]; trace reconstruction of circular strings over the deletion channel [**?**]. In another direction, Grigorescu, Sudan, and Zhu [22] studied the performance of mean-based reconstruction for distinguishing between strings at low Hamming or edit distance from each other over the deletion channel.

Different complex-analytic methods have been used to, for example, obtain the current best upper bound of $\exp(\widetilde{O}(n^{1/5}))$ traces on the trace complexity of the deletion channel [23], as well as upper bounds for trace reconstruction of "smoothed" worst-case strings over the deletion channel [24]. We note, however, that mean-based reconstruction remains the state-of-the-art approach for the geometric insertion-deletion channel.

Other related problems considered include the already-mentioned average-case trace reconstruction problem over the deletion and geometric insertion-deletion channels [3], [8], [25], [18], [19], trace reconstruction over the deletion and geometric insertion-deletion channels with vanishing deletion probabilities [3], [4], [5], [26], trace complexity lower bounds for the deletion channel [3], [25], [27], [**?**], trace reconstruction of coded strings over the deletion channel [28], [29], approximate trace reconstruction [30], alternative trace reconstruction models motivated by immunology [31], and population recovery over the deletion and geometric insertion-deletion channels [32], [33], [34]. Moreover, related complex analytic techniques have been applied to population recovery over the binary symmetric and erasure channels [35], [36] and parameter learning in mixture models [37].

B. Notation

For convenience, we denote discrete random variables and their corresponding distributions by uppercase letters, such as X, Y, and Z. The expected value of X is denoted by $\mathbb{E}[X]$. Sets are denoted by calligraphic uppercase letters such as $\mathcal{S}$ and $\mathcal{T}$, and we write $[n] := \{1, 2, \dots, n\}$. The open disk of radius r centered at $z \in \mathbb{C}$ is $\mathcal{D}_r(z) = \{z' \in \mathbb{C} : |z - z'| < r\}$. The 1-norm of vector x is denoted by $\|x\|_1$. The concatenation of strings x and y is denoted by $x\|y$.

C. Channel Model

We consider a general replication-insertion channel model that, in particular, captures the models studied in [4], [5], [9], [10]. A *replication-insertion channel* $\mathsf{Ch}_{(M,\mathcal{R},p_{\mathsf{flip}})}$ is characterized by a constant $p_{\mathsf{flip}} \in [0, 1/2)$ and a joint probability distribution $(M, \mathcal{R})$ with $M \in \{0, 1, 2, \dots\}$ and $\mathcal{R} \subseteq [M]$. To avoid trivial settings where trace reconstruction is impossible, we require that $\Pr[M > 0] > 0$ and $\Pr[\mathcal{R} \neq \emptyset] > 0$. Given an input string $x \in \{-1, 1\}^n$, the channel $\mathsf{Ch}_{(M,\mathcal{R},p_{\mathsf{flip}})}$ behaves independently on each input x_i as follows:

1) Sample a pair $(m_i, \mathcal{R}_i)$ with $m_i \in \{0, 1, 2, \dots\}$ and $\mathcal{R}_i \subseteq [m_i]$ according to the distribution of $(M, \mathcal{R})$;
2) Construct an output string $Y_{x_i} \in \{-1, 1\}^{m_i}$. If $j \in \mathcal{R}_i$, then let $Y_{x_i,j} = -x_i$ with probability p_{flip} and $Y_{x_i,j} = x_i$ with probability $1 - p_{\mathsf{flip}}$. If $j \notin \mathcal{R}_i$, then let $Y_{x_i,j}$ be a uniformly random bit.

The overall output of $\mathsf{Ch}_{(M,\mathcal{R},p_{\mathsf{flip}})}$ on input x is

$$Y_x = Y_{x_1} \| Y_{x_2} \| \cdots \| Y_{x_n}.$$

For example, the geometric insertion-deletion channel from [4], [5], [9] can be easily instantiated in this general framework by sampling $(M, \mathcal{R})$ as follows: First, sample G following a geometric distribution with support in $\{0, 1, 2, \dots\}$ and success probability σ and B following a Bernoulli distribution with success probability δ. Then, set $M = B + G$ and let $\mathcal{R} = \emptyset$ if $B = 1$ and $\mathcal{R} = \{1\}$ otherwise. Likewise, the alternative geometric insertion-deletion channel from [10], [18], [19] can also be easily instantiated in our framework.

D. Our Contributions

Our main theorem shows that previous results on mean-based trace reconstruction over the deletion and geometric insertion-deletion channels are examples of a much more general phenomenon.

Theorem 1: Worst-case mean-based trace reconstruction with success probability at least $1 - e^{-\Omega(n)}$ over any channel $\mathsf{Ch}_{(M,\mathcal{R},p_{\mathsf{flip}})}$ with sub-exponential[2] random variable M is achievable with $\exp(O(n^{1/3}))$ traces.

Note that many common distributions satisfy the requirement that M is sub-exponential, including geometric, Poisson, and all finitely-supported distributions.

[2]A random variable M is *sub-exponential* if there exists a constant $\alpha > 0$ such that $\Pr[|M| \geq \tau] \leq 2e^{-\alpha\tau}$ for all $\tau \geq 0$.

II. Proof of Theorem 1

Fix a replication-insertion channel $\mathsf{Ch}_{(M,\mathcal{R},p_{\mathsf{flip}})}$, where M is a sub-exponential random variable. To every string $x \in \{-1,1\}^n$, we can associate a polynomial P_x over $\mathbb{C}$ defined as

$$P_x(z) := \sum_{i=1}^{n} x_i z^{i-1}.$$

Then, using the definition of mean trace above, we define the mean trace power series $\overline{P}_x$ as

$$\overline{P}_x(z) := \sum_{i=1}^{\infty} \mu_{x,i} z^{i-1}.$$

Let $N > 0$ and denote the mean trace truncated at N by

$$\mu_x^N := (\mu_{x,1}, \dots, \mu_{x,N}).$$

To prove Theorem 1, we will show that there exists a constant $C > 0$ such that for a large enough n, appropriate N, and *any* distinct input strings $x, x' \in \{-1,1\}^n$, their truncated mean traces satisfy

$$\left\|\mu_x^N - \mu_{x'}^N\right\|_1 = \sum_{i=1}^{N} \left|\mu_{x,i} - \mu_{x',i}\right| \geq \delta(n) := e^{-Cn^{1/3}}. \quad (2)$$

This implies that $\exp(O(n^{1/3}))$ traces suffice for mean-based worst-case trace reconstruction as follows: Let x be the true input and suppose that we have access to $t := n/\delta(n)^2 = \exp(O(n^{1/3}))$ traces. Then a direct application of the Chernoff bound and a union bound over all coordinates $i = 1, \dots, N$ shows that the empirical mean trace $\widehat{\mu}^N = (\widehat{\mu}_1, \dots, \widehat{\mu}_N)$ defined in (1) satisfies

$$\left\|\widehat{\mu}^N - \mu_x^N\right\|_1 \leq \frac{\delta(n)}{4} \quad (3)$$

with probability at least $1 - e^{-\Omega(n)}$ over the randomness of the traces. On the other hand, if (3) holds, we also have that

$$\left\|\widehat{\mu}^N - \mu_{x'}^N\right\|_1 \geq \frac{3\delta(n)}{4}$$

for all $x' \neq x$ as a result of (2). This allows us to recover x naively from $\widehat{\mu}$ by computing $\mu_{\widehat{x}}^N$ for every $\widehat{x} \in \{-1,1\}^n$ and outputting the $\widehat{x}$ minimizing $\left\|\widehat{\mu}^N - \mu_{\widehat{x}}^N\right\|$.

We prove (2) by relating $\left\|\widehat{\mu}^N - \mu_x^N\right\|_1$ to $|\overline{P}_x(z) - \overline{P}_{x'}(z)|$ for an appropriate choice of $z \in \mathbb{C}$. By the triangle inequality, we have

$$\begin{aligned}
&\left|\overline{P}_x(z) - \overline{P}_{x'}(z)\right| \\
&\leq \sum_{i=1}^{\infty} \left|\mu_{x,i} - \mu_{x',i}\right| |z|^{i-1} \\
&= \sum_{i=1}^{N} \left|\mu_{x,i} - \mu_{x',i}\right| |z|^{i-1} + \sum_{i=N+1}^{\infty} \left|\mu_{x,i} - \mu_{x',i}\right| |z|^{i-1} \\
&\leq |z|^N \left\|\mu_x^N - \mu_{x'}^N\right\|_1 + \sum_{i=N+1}^{\infty} \left|\mu_{x,i} - \mu_{x',i}\right| |z|^{i-1}
\end{aligned}$$

for every $z \in \mathbb{C}$ such that $|z| \geq 1$. Rearranging, it follows that $\left\|\mu_x^N - \mu_{x'}^N\right\|_1$ is lower bounded by

$$|z|^{-N} \left(\left|\overline{P}_x(z) - \overline{P}_{x'}(z)\right| - \sum_{i=N+1}^{\infty} \left|\mu_{x,i} - \mu_{x',i}\right| |z|^{i-1} \right) \quad (4)$$

for any such z. The lower bound in (2), and thus Theorem 1, follows by combining (4) with the next two lemmas, each bounding a different term in the right-hand side of (4).

Lemma 2: There exist constants $c_1, c_2 > 0$ such that for n large enough and any distinct strings $x, x' \in \{-1,1\}^n$, it holds that $\left|\overline{P}_x(z) - \overline{P}_{x'}(z)\right| \geq e^{-c_1 n^{1/3}}$ for some z satisfying $1 \leq |z| \leq e^{c_2 n^{-2/3}}$.

Lemma 3: If $1 \leq |z| \leq e^{c_3 n^{-2/3}}$ for some constant $c_3 > 0$, there exist constants $c_4, c_5 > 0$ such that if $N = c_4 n$ then

$$\sum_{i=N+1}^{\infty} \left|\mu_{x,i} - \mu_{x',i}\right| |z|^{i-1} \leq e^{-c_5 n}$$

for all distinct $x, x' \in \{-1,1\}^n$ when n is large enough.

Invoking Lemmas 2 and 3, we have that for n large enough and any distinct $x, x' \in \{-1,1\}^n$, there exists an appropriate choice $z^\star \in \mathbb{C}$ possibly depending on x and x' which satisfies $1 \leq |z^\star| \leq e^{c_2 n^{-2/3}}$ and by setting $z = z^\star$ and $N = c_4 n$ in (4) yields

$$\begin{aligned}
\left\|\mu_x^N - \mu_{x'}^N\right\|_1 &\geq e^{-c_4 \cdot c_2 n^{1/3}} \left(e^{-c_1 n^{1/3}} - e^{-c_5 n}\right) \\
&\geq e^{-Cn^{1/3}}
\end{aligned}$$

for some constant $C > 0$, implying (2).

We prove Lemmas 2 and 3 in Sections III and IV, respectively, which completes the argument.

III. Proof of Lemma 2

Our proof of Lemma 2 follows the blueprint of [9, Sections 4 and 5] and [10, Sections 2 and 3]. The key differences lie in Lemmas 4 and 6 below. Lemma 4 generalizes [9, Section 4 and Appendix A.3] and [10, Lemmas 2.1 and 5.2] to arbitrary replication-insertion channels well beyond the deletion and geometric insertion-deletion channels. Lemma 6 requires analyzing the local behavior of the inverse of an *arbitrary* probability generating function (PGF) in the complex plane around $z = 1$. Remarkably, the desired behavior follows by combining the standard inverse function theorem for analytic functions with basic properties of PGFs. In contrast, the PGFs associated to the deletion and geometric insertion-deletion channels treated in [9], [10], [18], [19] are all Möbius transformations, meaning that their inverses have simple explicit expressions and could be easily analyzed directly.

As a first step, we show that the mean trace power series $\overline{P}_x$ is related to the input polynomial P_x through a change of variable. This allows us to bound $\left|\overline{P}_x(z) - \overline{P}_{x'}(z)\right|$ in terms of $\left|P_x(w) - P_{x'}(w)\right|$ for some w related to z.

Lemma 4: Let $\mathsf{Ch}_{(M,\mathcal{R},p_{\mathsf{flip}})}$ be a replication-insertion channel. Suppose $\mathbb{E}\llbracket\mathcal{R}|\rrbracket > 0$ is finite. Let W be the distribution given by

$$W(j) = \frac{\Pr[j+1 \in \mathcal{R}]}{\mathbb{E}\llbracket\mathcal{R}|\rrbracket}, \quad j = 0,1,2\ldots,$$

and g_M and g_W be the probability generating functions of M and W, respectively. Then for every $x \in \{-1,1\}^n$ and $z \in \mathbb{C}$ such that z is in the disk of convergence of both g_M and g_W, we have $\overline{P}_x(z) = (1-2p_{\mathsf{flip}}) \cdot \mathbb{E}\llbracket\mathcal{R}|\rrbracket \cdot g_W(z) \cdot P_x(g_M(z))$.

Let $C_1 := (1-2p_{\mathsf{flip}}) \cdot \mathbb{E}\llbracket\mathcal{R}|\rrbracket$. Then, Lemma 4 yields

$$\begin{aligned}&\left|\overline{P}_x(z) - \overline{P}_{x'}(z)\right| \\ &\quad = |C_1| \cdot \left|g_W(z)\right| \cdot \left|P_x(g_M(z)) - P_{x'}(g_M(z))\right|. \quad (5)\end{aligned}$$

Analogously to [9], [10], we use the following lemma due to Borwein and Erdélyi [38] to lower bound

$$\left|P_x(g_M(z)) - P_{x'}(g_M(z))\right|.$$

Lemma 5 ([38]): There is a universal constant $c > 0$ for which the following holds: Let $\mathbf{a} = (a_0, ..., a_{\ell-1}) \in \{-1,0,1\}^\ell$ be non-zero and define $A(w) := \sum_{j=0}^{\ell-1} a_j w^j$. Let γ_L denote the arc $\left\{e^{i\varphi} : \varphi \in [-\frac{\pi}{L}, \frac{\pi}{L}]\right\}$. Then, we have $\max_{w\in\gamma_L} |A(w)| \geq e^{-cL}$ for every $L > 0$.

This lemma implies that there is a constant $C_2 > 0$ such that for every $L > 0$ there exists $w_L = e^{i\varphi_L}$ with $|\varphi_L| \leq \frac{\pi}{L}$ satisfying

$$\left|P_x(w_L) - P_{x'}(w_L)\right| \geq e^{-C_2 L}. \quad (6)$$

We can use (6) to lower bound (5), provided there exists z_L such that $g_M(z_L) = w_L$ with good properties. The following lemma ensures this.

Lemma 6: For L large enough there is a constant $c' > 0$ such that for any $\varphi \in \left[-\frac{\pi}{L}, \frac{\pi}{L}\right]$ there exists z_φ satisfying $g_M(z_\varphi) = e^{i\varphi}$, $\left|g_W(z_\varphi)\right| \geq 1/2$, and $1 \leq \left|z_\varphi\right| \leq 1 + c'\varphi^2$.

As a result of Lemma 6, we can choose a z_L that satisfies $g_M(z_L) = w_L$, $|g_W(z_L)| \geq 1/2$, and $1 \leq |z_L| \leq 1 + \frac{C_2}{L^2} \leq e^{C_2/L^2}$ for large enough L. Using this together with (6), we obtain

$$\begin{aligned}\left|P_x(g_M(z_L)) - P_{x'}(g_M(z_L))\right| &= \left|P_x(w_L) - P_{x'}(w_L)\right| \\ &\geq e^{-C_2 L}. \quad (7)\end{aligned}$$

Set $L = n^{1/3}$. Combining (5), (7), and the fact that $|g_W(z_L)| \geq 1/2$ for large enough L, we obtain

$$\begin{aligned}\left|\overline{P}_x(z_L) - \overline{P}_{x'}(z_L)\right| &\geq |C_1| \cdot \left|g_W(z_L)\right| \cdot e^{-C_2 L} \\ &\geq e^{-C_3 n^{1/3}}\end{aligned}$$

for some constant $C_3 > 0$ when n is large enough. This concludes the proof of Lemma 2 assuming Lemmas 4 and 6.

A. Proofs of Lemmas 4 and 6

In this section, we prove the remaining lemmas.

Proof of Lemma 4: Fix an input string $x \in \{-1,1\}^n$. For each $i \in [n]$, let $\mathcal{R}_i$ denote the indices of Y_x that correspond to replications of x_i and let $\mathcal{I}_i$ denote the indices of Y_x that correspond to insertions of random bits resulting from the channel's action on x_i. Then we may write

$$(Y'_x)_j = \sum_{i=1}^n [B_{i,j} x_i \cdot \mathbf{1}_{\{j\in\mathcal{R}_i\}} + \mathbf{1}_{\{j\in\mathcal{I}_i\}} \cdot U_{i,j}],$$

where the $U_{i,j}$ are uniformly distributed over $\{-1,1\}$, the $B_{i,j}$ are random variables over $\{-1,1\}$ that are -1 with probability p_{flip}, and all these are independent of each other, $\mathcal{R}_i$, and $\mathcal{I}_i$. Note that if an output bit is in $\mathcal{R}_i$, then it has expected value $(1-2p_{\mathsf{flip}})x_i$. Therefore, we have that

$$\mathbb{E}[(Y'_x)_j] = \sum_{i=1}^n (1-2p_{\mathsf{flip}}) x_i \cdot \Pr[j \in \mathcal{R}_i].$$

We can use this to show that

$$\begin{aligned}\overline{P}_x(z) &= \sum_{j=1}^\infty \mathbb{E}[(Y'_x)_j] \cdot z^{j-1} \\ &= \sum_{j=1}^\infty \sum_{i=1}^n (1-2p_{\mathsf{flip}}) x_i \cdot \Pr[j \in \mathcal{R}_i] \cdot z^{j-1} \\ &= (1-2p_{\mathsf{flip}}) \sum_{i=1}^n x_i \cdot \sum_{j=1}^\infty \Pr[j \in \mathcal{R}_i] \cdot z^{j-1}. \quad (8)\end{aligned}$$

We proceed to simplify $\sum_{j=1}^\infty \Pr[j \in \mathcal{R}_i] \cdot z^{j-1}$. Let $M^{(\ell)} := \sum_{k=1}^\ell M_k$, where the $M_k := |Y_{x_k}|$ denote the length of the channel outputs associated to each input bit x_k and are i.i.d. according to M. We have

$$\begin{aligned}&\sum_{j=1}^\infty \Pr[j \in \mathcal{R}_i] \cdot z^{j-1} \\ &= \sum_{j=1}^\infty \Pr[M^{(i-1)} < j, j \in \mathcal{R}_i] \cdot z^{j-1} \\ &= \sum_{j=1}^\infty \sum_{j'=0}^{j-1} \Pr[M^{(i-1)} = j'] \cdot \Pr[j \in \mathcal{R}_i | M^{(i-1)} = j'] \cdot z^{j-1} \\ &= \sum_{j=1}^\infty \sum_{j'=0}^{j-1} \Pr[M^{(i-1)} = j'] \cdot \Pr[j - j' \in \mathcal{R}] \cdot z^{j-1} \\ &= \sum_{j'=0}^\infty \Pr[M^{(i-1)} = j'] \cdot \sum_{j=j'+1}^\infty Pr[j - j' \in \mathcal{R}] \cdot z^{j-1} \\ &= \sum_{j'=0}^\infty \Pr[M^{(i-1)} = j'] z^{j'} \cdot \sum_{j=1}^\infty \Pr[j \in \mathcal{R}] \cdot z^{j-1} \\ &= g_M(z)^{i-1} \cdot \sum_{j=1}^\infty \Pr[j \in \mathcal{R}] \cdot z^{j-1}. \quad (9)\end{aligned}$$

We can interchange the sums above because z is in the disk of convergence of g_M and g_W. From the definition of W,

$$g_W(z) = \sum_{j=0}^{\infty} W(j) \cdot z^j$$
$$= \frac{1}{\mathbb{E}\llbracket \mathcal{R} | \rrbracket} \cdot \sum_{j=1}^{\infty} \Pr[j \in \mathcal{R}] \cdot z^{j-1}. \quad (10)$$

Combining (9) with (10) yields

$$\sum_{j=1}^{\infty} \Pr[j \in \mathcal{R}_i] \cdot z^{j-1} = \mathbb{E}\llbracket \mathcal{R} | \rrbracket \cdot g_W(z) \cdot g_M(z)^{i-1},$$

Recalling (8) concludes the proof. ■

We prove Lemma 6 using the standard inverse function theorem stated below.

Lemma 7 ([39, Section VIII.4], adapted): Let $g : \Omega \to \mathbb{C}$ be a non-constant function analytic on a connected open set $\Omega \subseteq \mathbb{C}$ such that $g'(z) \neq 0$ for a given $z \in \Omega$. Then, there exist radii $\rho, \delta > 0$ such that for every $w \in \mathcal{D}_\delta(g(z))$ there exists a unique $z_w \in \mathcal{D}_\rho(z)$ satisfying $g(z_w) = w$. Moreover, the inverse function $f : \mathcal{D}_\delta(g(z)) \to \mathcal{D}_\rho(z)$ defined as $f(w) = z_w$ is analytic on $\mathcal{D}_\delta(g(z))$.

Proof of Lemma 6: Because M is sub-exponential and non-trivial, g_M is a non-constant analytic function on some open ball $\mathcal{D}_r(0)$ of radius $r > 1$ that satisfies $g'_M(1) = \mathbb{E}[M] \neq 0$. Hence, Lemma 7 applies with $g = g_M$, so there exist $\rho, \delta > 0$ and an analytic function $f : \mathcal{D}_\delta(1) \to \mathcal{D}_\rho(1)$ such that $g_M(f(w)) = w$. In particular, there exists $\gamma \in (0, \delta)$ such that for every $w \in \mathcal{D}_\gamma(1)$ we can write

$$f(w) = 1 + f'(1)(w-1) + \sum_{i=2}^{\infty} \frac{f^{(i)}(1)}{i!}(w-1)^i. \quad (11)$$

This is because $f(1) = 1$ since $g(1) = 1$. Note that $|f(w)| \leq 1 + \rho$ for all w such that $|w - 1| \leq \gamma$ since $f(z) \in D_\rho(1)$. Then by standard estimates [39, Section V.4], $\left|\frac{f^{(i)}(1)}{i!}\right| \leq \frac{1+\rho}{\gamma^i}$ for all i. Choosing L large enough so that $|w - 1| \leq \gamma/2$ yields the bound

$$\sum_{i=2}^{\infty} \left|\frac{f^{(i)}(1)}{i!}\right| \cdot |w-1|^i \leq \frac{c''}{2} \cdot |w-1|^2 \cdot \sum_{i=2}^{\infty} 2^{-(i-2)} = c'' \cdot |w-1|^2 \quad (12)$$

for the constant $c'' := \frac{2(1+\rho)}{\gamma^2}$ and all such w.

Assume that L is large enough so that $e^{i\varphi} \in \mathcal{D}_{\gamma/2}(1)$ for all $\varphi \in \left[-\frac{\pi}{L}, \frac{\pi}{L}\right]$. Then, we set $z_\varphi = f(e^{i\varphi})$. Note that $g_M(z_\varphi) = e^{i\varphi}$ by the definition of f, as required.

Combining (11) with $w = e^{i\varphi}$ and the triangle inequality, we have

$$|z_\varphi - 1| = O\left(\left|e^{i\varphi} - 1\right|\right) \to 0$$

as $L \to \infty$. Since g_W is a continuous function on a neighborhood of 1 and $g_W(1) = 1$, it follows that $|g_W(z_\varphi)| \geq 1/2$ if L is large enough. On the other hand, combining (11) with the fact that

$$f'(1) = \frac{1}{g'(f(1))} = \frac{1}{g'(1)} = \frac{1}{\mathbb{E}[M]} \in \mathbb{R},$$

by the chain rule, we obtain

$$\begin{aligned}
|z_\varphi| &\leq \left|1 + \frac{e^{i\varphi} - 1}{\mathbb{E}[M]}\right| + c''\left|e^{i\varphi} - 1\right|^2 \\
&\leq \sqrt{\left(1 - \frac{1 - \cos(\varphi)}{\mathbb{E}[M]}\right)^2 + \frac{\sin(\varphi)^2}{\mathbb{E}[M]^2}} + c''\varphi^2 \\
&\leq \sqrt{1 + \frac{2(1 - \cos(\varphi))}{\mathbb{E}[M]^2}} + c''\varphi^2 \\
&\leq 1 + \left(\frac{1}{\mathbb{E}[M]^2} + c''\right)\varphi^2.
\end{aligned}$$

The second inequality holds because $\left|e^{i\varphi} - 1\right| \leq \varphi$. The last inequality follows by noting that $1 - \cos(\varphi) \leq \varphi^2/2$ and $\sqrt{1+x} \leq 1 + x$ for $x \geq 0$. ■

IV. Proof of Lemma 3

To conclude the argument, we prove Lemma 3 using an argument analogous to [9, Appendix A.2] and the fact that M is sub-exponential.

Let $M_1, M_2, \ldots, M_n$ be i.i.d. according to M, and set $M^{(n)} = \sum_{i=1}^{n} M_i$. Then, we have

$$|\mu_{x,i} - \mu_{x',i}| \leq \Pr[M^{(n)} \geq i]$$

for every i. Since M is sub-exponential, a direct application of Bernstein's inequality [40, Theorem 2.8.1] guarantees the existence of constants $c_4, c_6 > 0$ such that for $N = c_4 n$ and any $j \geq 1$ we have

$$\Pr[M^{(n)} \geq N + j] \leq 2e^{-c_6(N+j)}.$$

Combining these observations with the assumption that $|z| \leq e^{c_3 n^{-2/3}}$ yields

$$\begin{aligned}
\sum_{i=N+1}^{\infty} |\mu_{x,i} - \mu_{x',i}| \, |z|^{i-1} &\leq \sum_{j=1}^{\infty} 2e^{-c_6(N+j)} \cdot e^{c_3 n^{-2/3}} \\
&\leq e^{-c_5 n}
\end{aligned}$$

for some constant $c_5 > 0$ and n large enough.

V. Future Work

We have shown that $\exp(O(n^{1/3}))$ traces suffice for mean-based worst-case trace reconstruction over a broad class of replication-insertion channels. However, our channel model does not cover all discrete memoryless synchronization channels as defined by Dobrushin [6], [7]. It would be interesting to extend the result in some form to all such non-trivial channels. On the other hand, to complement the above, it would be interesting to prove trace complexity lower bounds for mean-based reconstruction over all these channels.

Furthermore, it is unclear whether the assumption that M be sub-exponential is necessary for our result. A clear extension of this work would be to either remove this condition or prove that it is necessary for mean-based trace reconstruction from $\exp(O(n^{1/3}))$ traces.

REFERENCES

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

[2] ——, "Efficient reconstruction of sequences from their subsequences or supersequences," *Journal of Combinatorial Theory, Series A*, vol. 93, no. 2, pp. 310–332, 2001.

[3] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2004, pp. 910–918.

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

[5] K. Viswanathan and R. Swaminathan, "Improved string reconstruction over insertion-deletion channels," in *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2008, pp. 399–408.

[6] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 18–36, 1967.

[7] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Transactions on Information Theory*, 2020, to appear. DOI: 10.1109/TIT.2020.2997329.

[8] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2008, pp. 389–398.

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

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

[11] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, 2015.

[12] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific reports*, vol. 7, no. 1, p. 5011, 2017.

[13] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen *et al.*, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

[14] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, vol. 6, pp. 1–33, 2009.

[15] H. Mercier, V. K. Bhargava, and V. Tarokh, "A survey of error-correcting codes for channels with symbol synchronization errors," *IEEE Communications Surveys Tutorials*, vol. 12, no. 1, pp. 87–96, First Quarter 2010.

[16] B. Haeupler and A. Shahrasbi, "Synchronization strings and codes for insertions and deletions – a survey," *IEEE Transactions on Information Theory*, 2021, to appear. Available at https://arxiv.org/abs/2101.00711.

[17] L. Hartung, N. Holden, and Y. Peres, "Trace reconstruction with varying deletion probabilities," in *Proceedings of the 15th Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, 2018, pp. 54–61.

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

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

[20] S. Davies, M. Z. Racz, and C. Rashtchian, "Reconstructing trees from traces," in *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 2019, pp. 961–978.

[22] E. Grigorescu, M. Sudan, and M. Zhu, "Limitations of mean-based algorithms for trace reconstruction at small distance," *arXiv e-prints*, p. arXiv:2011.13737, Nov. 2020.

[21] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Trace reconstruction: Generalized and parameterized," in *27th Annual European Symposium on Algorithms (ESA)*, 2019, pp. 68:1–68:25.

[23] Z. Chase, "New lower bounds for trace reconstruction," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 57, no. 2, pp. 627 – 643, 2021. [Online]. Available: https://doi.org/10.1214/20-AIHP1089

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

[25] A. McGregor, E. Price, and S. Vorotnikova, "Trace reconstruction revisited," in *22nd Annual European Symposium on Algorithms (ESA)*, 2014, pp. 689–700.

[26] X. Chen, A. De, C. H. Lee, R. A. Servedio, and S. Sinha, "Polynomial-time trace reconstruction in the low deletion rate regime," *arXiv e-prints*, p. arXiv:2012.02844, Dec. 2020, to appear in ITCS 2021.

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

[28] 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.

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

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

[31] V. Bhardwaj, P. A. Pevzner, C. Rashtchian, and Y. Safonova, "Trace reconstruction problems in computational biology," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3295–3314, 2021.

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

[33] F. Ban, X. Chen, R. A. Servedio, and S. Sinha, "Efficient average-case population recovery in the presence of insertions and deletions," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*, 2019, pp. 44:1–44:18.

[34] S. Narayanan, "Improved algorithms for population recovery from the deletion channel," in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2021, pp. 1259–1278.

[35] Y. Polyanskiy, A. T. Suresh, and Y. Wu, "Sample complexity of population recovery," in *Proceedings of the 2017 Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, S. Kale and O. Shamir, Eds., vol. 65. PMLR, 07–10 Jul 2017, pp. 1589–1618. [Online]. Available: http://proceedings.mlr.press/v65/polyanskiy17a.html

[36] A. De, R. O'Donnell, and R. A. Servedio, "Sharp bounds for population recovery," *Theory of Computing*, vol. 16, no. 6, pp. 1–20, 2020. [Online]. Available: http://www.theoryofcomputing.org/articles/v016a006

[37] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Algebraic and analytic approaches for parameter learning in mixture models," in *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, A. Kontorovich and G. Neu, Eds., vol. 117. San Diego, California, USA: PMLR, 08 Feb–11 Feb 2020, pp. 468–489. [Online]. Available: http://proceedings.mlr.press/v117/krishnamurthy20a.html

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

[39] T. Gamelin, *Complex Analysis*, ser. Undergraduate Texts in Mathematics. Springer New York, 2001.

[40] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.