58th Annual IEEE Symposium on Foundations of Computer Science

Average-case reconstruction for the deletion channel: subpolynomially many traces suffice

Yuval Peres
Microsoft Research
Redmond, WA, USA
Email: peres@microsoft.com

Alex Zhai
Department of Mathematics
Stanford University
Stanford, CA, USA
Email: azhai@stanford.edu

***Abstract*—The *deletion channel* takes as input a bit string $\mathbf{x} \in \{0,1\}^n$, and deletes each bit independently with probability q, yielding a shorter string. The *trace reconstruction problem* is to recover an unknown string $\mathbf{x}$ from many independent outputs (called "traces") of the deletion channel applied to $\mathbf{x}$.**

We show that if $\mathbf{x}$ is drawn uniformly at random and $q < 1/2$, then $e^{O(\log^{1/2} n)}$ traces suffice to reconstruct $\mathbf{x}$ with high probability. The previous best bound, established in 2008 by Holenstein, Mitzenmacher, Panigrahy, and Wieder [1], uses $n^{O(1)}$ traces and only applies for q less than a smaller threshold (it seems that $q < 0.07$ is needed).

Our algorithm combines several ideas: 1) an alignment scheme for "greedily" fitting the output of the deletion channel as a subsequence of the input; 2) a version of the idea of "anchoring" used in [1]; and 3) complex analysis techniques from recent work of Nazarov and Peres [2] and De, O'Donnell, and Servedio [3].

***Keywords*-deletion channel; trace reconstruction; sequence alignment**

I. Introduction

The *deletion channel* takes as input a bit string $\mathbf{x} \in \{0,1\}^n$. Each bit of $\mathbf{x}$ is (independently of other bits) retained with probability p and deleted with probability $q := 1-p$. The channel then outputs the concatenation of the retained bits; such an output is called a *trace*. Suppose that the input $\mathbf{x}$ is unknown. The *trace reconstruction problem* asks the following: how many i.i.d. traces from the deletion channel do we need to observe in order to determine $\mathbf{x}$ with high probability?

There are two basic variants of this problem, which we will call the "worst case" and "average case". In the worst case variant, the problem is to provide bounds that hold uniformly over all possible input strings $\mathbf{x}$. The average case variant supposes that the input is chosen uniformly at random. In particular, we are allowed to ignore some "hard-to-reconstruct" inputs, as long as they comprise a small fraction of all 2^n possible inputs. In this paper, we study the average case. Our main result is the following.

Theorem 1. *Suppose $q < \frac{1}{2}$, and let $\mathbf{X} \in \{0,1\}^n$ be an unknown bit string of length n chosen uniformly at random. There is a constant C_q depending only on q such that it is possible to reconstruct $\mathbf{X}$ with probability at least $1 - \frac{C_q}{n}$ using at most $\exp\left(C_q\sqrt{\log n}\right)$ independent samples from the deletion channel with deletion probability q applied to $\mathbf{X}$.*

A. Related work

The study of trace reconstruction for the deletion channel seems to have been initiated by Batu, Kannan, Khanna and McGregor [4], who were motivated by multiple sequence alignment problems in computational biology. We focus on the regime where the deletion probability q is held constant as n grows.

Previously, the best bound in the average case was due to Holenstein, Mitzenmacher, Panigrahy and Wieder [1], who gave an algorithm for reconstructing random inputs using polynomially many traces when q is less than some small threshold c.[1] Theorem 1 improves on this result in two ways: the number of traces is subpolynomial, and we extend the range of allowed q to the interval $(0, 1/2)$.

In [1] it is also shown that $e^{O(n^{1/2}\log n)}$ traces suffice for reconstruction with high probability with worst case input. This was recently improved by Nazarov-Peres [2] and De-O'Donnell-Servedio [3] (simultaneously and independently) to $e^{O(n^{1/3})}$. Their techniques, which we use in Section IV, play an important role in our proofs.

The question of whether the above bounds are optimal remains open. The best lower bounds known are of order $\log^2 n$ (McGregor, Price and Vorotnikova [5]) in the average case and order n in the worst case ([4]).

Other settings for trace reconstruction include the case when $q \to 0$ ([4]), when insertions and substitutions are allowed as well as deletions ([6], [7]), or when the strings are taken over an alphabet whose size grows with n ([5]). For a more comprehensive review of the literature, see the introduction of [3] or the survey of Mitzenmacher [8].

B. Outline of approach

Let us give a high-level description of the algorithm used to prove Theorem 1. Suppose that we have already reconstructed the first k bits of $\mathbf{X}$, and we consider a new trace $\widetilde{\mathbf{X}}$. Roughly speaking, our goal is to do the following:

[1]The threshold c is not given explicitly in [1]. It seems that by optimizing their methods we cannot achieve $c > 0.07$.

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.29

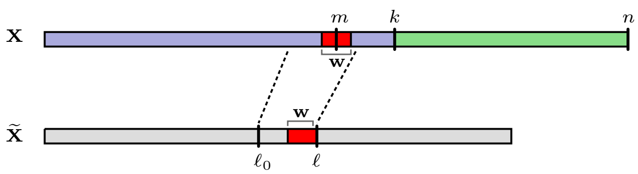

Figure I.1: Illustration of the alignment strategy. Dotted lines indicate correspondences between positions in $\widetilde{\mathbf{X}}$ and positions in $\mathbf{X}$.

Alignment: Find some suitable index m slightly less than k, and try to (approximately) identify the position ℓ in $\widetilde{\mathbf{X}}$ that corresponds to the m-th position of $\mathbf{X}$. This occurs in two stages (see Figure I.1):

- **Initial alignment:** Find a position ℓ_0 in $\widetilde{\mathbf{X}}$ whose corresponding position in $\mathbf{X}$ is known to be about $O(\log n)$ places ahead of m.
- **Refined alignment:** Consider a specific substring $\mathbf{w}$ of $\mathbf{X}$ located at m and having length $O(\log^{1/2} n)$. Look for $\mathbf{w}$ to occur in $\widetilde{\mathbf{X}}$ within $O(\log n)$ characters following position ℓ_0, and take ℓ to be the last position of this occurrence of $\mathbf{w}$.

Reconstruction: Use the bits of $\widetilde{\mathbf{X}}$ after ℓ as a trace of the bits of $\mathbf{X}$ after m. From these "traces", we reconstruct at least $k+1-m$ bits of $\mathbf{X}$ starting from position m, which in particular includes the $(k+1)$-th bit of $\mathbf{X}$.

We can repeat the above procedure for each k. In each iteration, the number of traces needed will be $e^{O(\sqrt{\log n})}$. Moreover, these traces may be reused for each iteration, because we will ultimately bound the probability of failure by a union bound.

1) Initial alignment step: The initial alignment step is based on fitting $\widetilde{\mathbf{X}}$ as a subsequence of $\mathbf{X}$ following a "greedy algorithm". Let X_i and $\widetilde{X}_i$ denote the i-th bits of $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, respectively. We associate $\widetilde{X}_1$ to the first bit in $\mathbf{X}$ that matches $\widetilde{X}_1$, then associate $\widetilde{X}_2$ to the next bit in $\mathbf{X}$ that matches $\widetilde{X}_2$, and so on (see Figure I.2). This gives the "first possible" occurrence of $\widetilde{\mathbf{X}}$ as a subsequence of $\mathbf{X}$, but does not necessarily reflect the true alignment of $\widetilde{\mathbf{X}}$ to $\mathbf{X}$. However, when $q < 1/2$ and $\mathbf{X}$ is random, it turns out that this greedy alignment actually matches the true one to within $O(\log n)$ (stated precisely in Lemma 1).

Let us briefly describe why this is so. Suppose the position assigned by our greedy algorithm lags behind the true position. Looking at the next bit in the trace, the true position should advance by $\frac{1}{1-q} < 2$ places in expectation. However, since the bits of $\mathbf{X}$ are uniformly random, the position for the greedy algorithm should advance like a geometric random variable with mean 2, thereby "catching up".

The same greedy matching idea was also considered by Mitzenmacher (see Section 3 of [8]) in the context of decoding for the deletion channel. Lemma 1 is a variant of Theorem 3.2 in [8]. However, many details are omitted in [8], so we provide a self-contained proof in Section II.

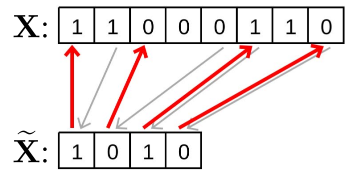

Figure I.2: Illustration of the greedy algorithm used in the initial alignment step. Here, $\mathbf{X} = 11000110$ and $\widetilde{\mathbf{X}} = 1010$. Gray arrows point from the positions in $\mathbf{X}$ that were retained to their corresponding positions in $\widetilde{\mathbf{X}}$. Red arrows indicate the associations produced by our algorithm (i.e. $\widetilde{X}_1$ goes to X_1, $\widetilde{X}_2$ goes to X_3, $\widetilde{X}_3$ goes to X_6, $\widetilde{X}_4$ goes to X_8).

2) Refined alignment step: For the refined alignment, we take an approach similar to the use of "anchors" in [1]. We again rely on the randomness of $\mathbf{X}$ and the assumption $q < 1/2$. Consider a substring $\mathbf{w}$ of length $a \approx \log^{1/2} n$ which contains the m-th bit of $\mathbf{X}$. (In the language of [1], $\mathbf{w}$ is our "anchor".)

With probability p^a, the string $\mathbf{w}$ appears in our trace because none of its bits were deleted. There is also a chance that this exact sequence just happens to appear after deletions to another part of the input. However, because $\mathbf{X}$ is random, the latter scenario only happens with probability $2^{-a} \ll p^a$. Thus, when we see $\mathbf{w}$ in our trace, it most likely came from near position m of $\mathbf{X}$ (we discard traces if we do not see $\mathbf{w}$), thereby aligning to within $O(\log^{1/2} n)$.

We remark here that the above discussion sweeps under the rug a few considerations about how to avoid accumulation of many small probabilities of error. In particular, note that the error probabilities involved during the refined alignment step are like $e^{-O(\log^{1/2} n)}$, which is not small enough to union bound over the whole string.

For example, a problem may arise if we have another copy of $\mathbf{w}$ appearing in $\mathbf{X}$ that is only $O(\log n)$ positions away from m. In that case, appearances of $\mathbf{w}$ in $\widetilde{\mathbf{X}}$ might come from either copy of $\mathbf{w}$ in $\mathbf{X}$, and it would be hard to distinguish the two scenarios.

Recall, however, that we have allowed ourselves some flexibility in the choice of m. Note that the initial alignment step means that we only need to worry about what $\mathbf{X}$ looks like within distance $O(\log n)$ from the location m. We look at $O(\log^{1/2} n)$ possible locations of m which are spaced $O(\log n)$ apart, and we argue that with high probability, at least one of these locations (and the corresponding choice of $\mathbf{w}$) behaves in the desired way.

3) Reconstruction step: For the reconstruction step, we analyze bit statistics using methods based on those of [2] and [3]. However, two adaptations are needed. First, our reconstruction step only needs to recover a small number of bits, not the full string. The statement we need is roughly that $e^{O(r^{1/3})}$ traces are enough to recover the first r bits of an unknown string, which we apply with $r = O(\log^{3/2} n)$.

Second, since our alignment is not perfect, we must allow some random shifts of the input string. The amount of shifting we can tolerate is relatively small, which explains the need for accurate alignment. The issue of calculating bit statistics with random shifts also appears in [1], although our techniques for handling this are rather different from theirs.

These two adaptations can be carried out by small modifications to proofs in [2] and [3], which are based on bounds for Littlewood polynomials on arcs of the unit circle.

C. Notation

We will use boldface to denote bit strings, while the values of their bits are non-bolded and subscripted by indices; for example, $\mathbf{x} = (x_1, x_2 \ldots, x_n) \in \{0,1\}^n$. Let $|\mathbf{x}| = n$ denote the length of $\mathbf{x}$, and let $\mathbf{x}^{a:b}$ denote the substring $(x_a, x_{a+1}, \ldots, x_b)$. For brevity, we also write $\mathbf{x}^{a:} = \mathbf{x}^{a:|\mathbf{x}|}$ for the suffix of $\mathbf{x}$ starting at x_a.

Next, we introduce notation for describing the deletion channel. For a given parameter $p \in (0,1)$, let $\mathcal{D}_p^*(\mathbf{x})$ denote the distribution over pairs $(\mathbf{t}, \widetilde{\mathbf{x}})$ of sequences defined as follows: $\mathbf{t} = (t_1, t_2, \ldots, t_m)$ is the random sequence of indices of $\mathbf{x}$ which are retained by the deletion channel applied to $\mathbf{x}$ with deletion probability $q = 1-p$, and $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_m)$ is given by $\widetilde{x}_i = x_{t_i}$. Note that the length $m = |\mathbf{t}|$ is random.

In some cases, we are only interested in the final output $\widetilde{\mathbf{x}}$ and not in $\mathbf{t}$. Thus, we also introduce the notation $\mathcal{D}_p(\mathbf{x})$ for the marginal distribution of $\mathcal{D}_p^*(\mathbf{x})$ over the strings $\widetilde{\mathbf{x}}$. We will sometimes use the notation $\mathbb{P}_{\mathbf{x}}(\,\cdot\,)$ to emphasize that the string going through the deletion channel is $\mathbf{x}$.

At some point, we will want to use $\mathbf{t}$ to associate several indices at once in $\widetilde{\mathbf{x}}$ to their counterparts in $\mathbf{x}$, or vice versa. Consider sets $S \subseteq \{1, 2, \ldots, |\mathbf{x}|\}$ and $\widetilde{S} \subseteq \{1, 2, \ldots, |\widetilde{\mathbf{x}}|\}$. Then, we use the notation

$$\mathbf{t}(\widetilde{S}) := \{t_s : s \in \widetilde{S}\} \qquad \text{and} \qquad \mathbf{t}^{-1}(S) := \{s : t_s \in S\},$$

which matches the usual notation for images/preimages if $\mathbf{t}$ is regarded as a map from indices in $\widetilde{\mathbf{x}}$ to indices in $\mathbf{x}$.

Finally, in addition to the standard notation $O(\,\cdot\,)$ and $\Omega(\,\cdot\,)$, we also use $O_p(\,\cdot\,)$ and $\Omega_p(\,\cdot\,)$ in cases where the implied constant may depend on p but nothing else.

D. Organization of the paper

The rest of the paper is organized as follows. In Sections II and III, we prove the lemmas needed for the initial and refined alignment steps, respectively. In Section IV, we prove lemmas for the reconstruction step. Finally, in Section V, we pull together all the ingredients to prove Theorem 1. A few proofs of technical lemmas are omitted in this condensed version of the paper; see [9] for full details.

II. Alignment by Greedy Matching

Suppose we have a string $\mathbf{x} \in \{0,1\}^n$ and a sample $(\mathbf{t}, \widetilde{\mathbf{x}}) \sim \mathcal{D}_p^*(\mathbf{x})$. Given only $\mathbf{x}$ and $\widetilde{\mathbf{x}}$, it is not in general possible to infer uniquely what $\mathbf{t}$ is. However, we may obtain an approximation using a "greedy algorithm" as described in Section I-B. To state things precisely, consider any two bit strings $\mathbf{x}$ and $\mathbf{y}$. We define a sequence $(g_k(\mathbf{y}, \mathbf{x}))_{k=1}^{|\mathbf{y}|}$ as follows:

- Take $g_1(\mathbf{y}, \mathbf{x})$ to be the least index such that $x_{g_1(\mathbf{y},\mathbf{x})} = y_1$. If no bits in $\mathbf{x}$ are equal to y_1, we set $g_1(\mathbf{y}, \mathbf{x}) = \infty$.
- For $k < |\mathbf{y}|$, define inductively $g_{k+1}(\mathbf{y}, \mathbf{x})$ to be the least index greater than $g_k(\mathbf{y}, \mathbf{x})$ for which $x_{g_{k+1}(\mathbf{y},\mathbf{x})} = y_{k+1}$. If no bits in $\mathbf{x}$ after the $g_k(\mathbf{y}, \mathbf{x})$-th position are equal to y_{k+1}, we set $g_{k+1}(\mathbf{y}, \mathbf{x}) = \infty$. (Note that in particular if $g_k(\mathbf{y}, \mathbf{x}) = \infty$, then $g_{k+1}(\mathbf{y}, \mathbf{x}) = \infty$).

We are primarily interested in the case where $\mathbf{y} = \widetilde{\mathbf{x}}$, where $\widetilde{\mathbf{x}}$ is a trace drawn from $\mathcal{D}_p(\mathbf{x})$. In this situation, $g_k(\widetilde{\mathbf{x}}, \mathbf{x})$ represents the "earliest possible" place in $\mathbf{x}$ that the k-th bit of $\widetilde{\mathbf{x}}$ could have come from. For an illustration, we refer back to Figure I.2. In that picture, we have $g_1(\widetilde{\mathbf{x}}, \mathbf{x}) = 1$, $g_2(\widetilde{\mathbf{x}}, \mathbf{x}) = 3$, $g_3(\widetilde{\mathbf{x}}, \mathbf{x}) = 6$, and $g_4(\widetilde{\mathbf{x}}, \mathbf{x}) = 8$.

One may check by a straightforward induction that $g_k(\widetilde{\mathbf{x}}, \mathbf{x}) \le t_k$ for all $1 \le k \le |\widetilde{\mathbf{x}}|$. (This means that $g_k(\widetilde{\mathbf{x}}, \mathbf{x})$ is never ∞; the possibility of having $g_k(\mathbf{y}, \mathbf{x}) = \infty$ doesn't come into play until the proof of Lemma 11.) We will show that for retention probability $p > \frac{1}{2}$ and $\mathbf{x}$ drawn uniformly at random, $g_k(\widetilde{\mathbf{x}}, \mathbf{x})$ is usually not much less than t_k. The following definition makes this precise.

Definition 1. *Consider a sequence $\mathbf{x} \in \{0,1\}^n$, and take $(\mathbf{t}, \widetilde{\mathbf{x}}) \sim \mathcal{D}_p^*(\mathbf{x})$. We say that $\mathbf{x}$ is (α, β)-**trackable** if*

$$\mathbb{P}_{\mathbf{x}}\left(\max_{1 \le k \le |\mathbf{t}|} (t_k - g_k(\widetilde{\mathbf{x}}, \mathbf{x})) \ge \lambda\right) \le e^{-\frac{\lambda - \alpha}{\beta}}.$$

The main result of this section is the following lemma.

Lemma 1. *Suppose $p > \frac{1}{2}$, and let $\mathbf{X} \in \{0,1\}^n$ be a uniformly random string of n bits. There exists $C_p > 0$ depending only on p such that*

$$\mathbb{P}\left(\mathbf{X} \textit{ is } (C_p \log n, C_p)\textit{-trackable}\right) \ge 1 - O_p\left(\frac{1}{n}\right)$$

Lemma 1 is implied by Theorem 3.2 of [8]. However, many details are omitted there, so we devote this section to proving Lemma 1 formally. We use the same general approach, except that it is more natural for us to focus on the quantity $t_k - g_k(\widetilde{\mathbf{X}}, \mathbf{X})$ rather than a slightly different quantity considered in [8]. We start with a conditional independence property similar to Lemma 3.3 of [8].

Lemma 2. *Let $\mathbf{X} \in \{0,1\}^n$ be drawn uniformly at random, and suppose $(\mathbf{t}, \widetilde{\mathbf{X}}) \sim \mathcal{D}_p^*(\mathbf{X})$. Then, for any integer $k \ge 1$, conditioned on the event $|\mathbf{t}| \ge k$ and the values of*

$$t_1, t_2, \ldots, t_k \textit{ and } g_1(\widetilde{\mathbf{X}}, \mathbf{X}), g_2(\widetilde{\mathbf{X}}, \mathbf{X}), \ldots, g_k(\widetilde{\mathbf{X}}, \mathbf{X}),$$

the bits $X_{g_k(\widetilde{\mathbf{X}},\mathbf{X})+1}, X_{g_k(\widetilde{\mathbf{X}},\mathbf{X})+2}, \ldots, X_n$ are i.i.d. uniformly distributed.

Remark 1. *The above lemma also applies when $\mathbf{X}$ is an infinite sequence of i.i.d. uniform bits. In this case, the conclusion is that all of $(X_i)_{i=g_k(\widetilde{\mathbf{X}},\mathbf{X})+1}^{\infty}$ are i.i.d. uniform.*

Proof: We first condition on $\mathbf{t}$; this conditioning will stay in effect for the remainder of the proof. Note that all of the X_i are still i.i.d. uniform, since the t_i depend only on which bits are deleted and not on the values of the bits themselves. Since we have conditioned on $\mathbf{t}$, we may regard $g_i(\widetilde{\mathbf{X}}, \mathbf{X})$ as a deterministic function of $\mathbf{X}$. Therefore, for brevity we will write $g_i(\mathbf{X}) = g_i(\widetilde{\mathbf{X}}, \mathbf{X})$.

Next, fix any sequence S of integers $s_1, s_2, \ldots, s_k$ where $s_1 < s_2 < \cdots < s_k$ and $s_i \le t_i$ for each i. We say a bit string $\mathbf{z}$ is S-*compatible* if $g_i(\mathbf{z}) = s_i$ for each i, and let E_S be the event that $\mathbf{X}$ is S-compatible.

Consider any two strings $\mathbf{w}, \mathbf{w}' \in \{0,1\}^{n-s_k}$ which differ in a single bit. We will give a bijection between S-compatible realizations of $\mathbf{X}$ that end in $\mathbf{w}$ and those that end in $\mathbf{w}'$. This is enough to establish the lemma, since by repeated application, it shows that any two strings for $\mathbf{X}^{(s_k+1):}$ are equally likely conditioned on E_S, and this holds for arbitrary S.

To carry out the bijection, for any index j with $1 \le j \le k$, we define its *influencing set* to be the set

$$I_j = \{t : s_{j-1} < t \le s_j\},$$

with the convention $s_0 = 0$. Informally, it is the set of all indices t where the value of X_t had some effect on the value of $g_j(\mathbf{X})$ (which is equal to s_j if $\mathbf{X}$ is S-compatible).

For any two indices i and j with $1 \le i, j \le k$, we say i *directly influences* j if $t_i \in I_j$. Note that because $s_j \le t_j$, we see that if i influences j, then $i \le j$ with equality if and only if $s_i = t_i$. We say that i *influences* j if there is a chain of direct influences from i to j (i.e. there exist $c_1, c_2, \ldots, c_N$ such that $c_1 = i$, $c_N = j$, and c_α directly influences $c_{\alpha+1}$ for $\alpha = 1, 2, \ldots, N-1$).

Suppose now that we have a S-compatible sequence $\mathbf{z}$ that ends in $\mathbf{w}$. We will describe a way to modify $\mathbf{z}$ so that it remains S-compatible but ends in $\mathbf{w}'$. Let ℓ be the index at which $w_\ell \ne w'_\ell$. First, suppose that $s_k + \ell \ne t_i$ for any $i \le k$. Then, we may simply flip the $(s_k + \ell)$-th bit of $\mathbf{z}$ to obtain a S-compatible sequence ending in $\mathbf{w}'$.

Otherwise, $s_k + \ell = t_m$ for some $m \le k$. Define the sets

$$U = \{m\} \cup \{j : j \text{ influences } m\}$$
$$V = \{t_m\} \cup \Big(\bigcup_{j \in U} I_j \Big).$$

We claim that by flipping all the bits of $\mathbf{z}$ at positions in V, the resulting sequence $\mathbf{z}'$ ends in $\mathbf{w}'$ and is S-compatible. The first claim follows from the fact that $I_j \subseteq \{1, 2, \ldots, s_k\}$ for all $j \le k$, so the only bit flipped after position s_k is the bit at position $t_m = s_k + \ell$.

To show S-compatibility, we show by induction on j that $g_j(\mathbf{z}') = s_j$ for each j, where the base case $j = 0$ is established by the convention $g_0(\mathbf{z}') = s_0 = 0$. For the inductive step, suppose that $g_i(\mathbf{z}') = s_i$ for each $i < j$. We consider two cases.

Case $j \in U$. By the definition of U, either $j = m$ or there exists $j' \in U$ for which $t_j \in I_{j'}$. In either case, we see that $t_j \in V$. We also have by definition that $I_j \subseteq V$. By S-compatibility of $\mathbf{z}$, the condition $g_j(\mathbf{z}) = s_j$ says that s_j is the first position after $g_{j-1}(\mathbf{z}) = s_{j-1}$ having the same value as position t_j. In other words, s_j is the unique position in I_j with the same value as position t_j.

The bits at positions t_j and elements of I_j are all flipped for $\mathbf{z}'$, so the same property holds in $\mathbf{z}'$. Since $g_{j-1}(\mathbf{z}') = s_{j-1}$ by the inductive hypothesis, we have $g_j(\mathbf{z}') = s_j$ as well.

Case $j \notin U$. Note that $t_m > s_k$, so $t_m \notin I_j$. Since $j \notin U$, it follows that I_j is disjoint from V. Note that if $t_j \in I_{j'}$ for some $j' \in U$, then j directly influences j' and hence influences m, but this contradicts $j \notin U$. Also, clearly $t_j \ne t_m$ since $j \ne m$. Thus, $t_j \notin V$.

We see that none of the bits at positions t_j or elements of I_j are flipped for $\mathbf{z}'$, so by the same argument as in the previous case, we conclude that $g_j(\mathbf{z}') = s_j$.

This completes the induction, showing that $\mathbf{z}'$ indeed ends in $\mathbf{w}'$ and is S-compatible. Furthermore, observe that the set V depends only on S, and so we may symmetrically recover $\mathbf{z}$ from $\mathbf{z}'$ by the same transformation. Thus, this gives a bijection from S-compatible sequences ending in $\mathbf{w}$ to those ending in $\mathbf{w}'$, completing the proof. ■

The next two lemmas describe how closely g_k tracks t_k. To avoid boundary issues, we state them for infinite bit sequences.

Lemma 3. *Let $\mathbf{X}$ be an infinite sequence of i.i.d. uniform bits, and let $(\mathbf{t}, \widetilde{\mathbf{X}}) \sim \mathcal{D}_p^*(\mathbf{X})$. Define $d_k = t_k - g_k(\widetilde{\mathbf{X}}, \mathbf{X})$. Then, $d_{k+1} - d_k$ is independent of $d_1, \ldots, d_k$ and has the same law as $\max(G_p - G_{1/2}, -d_k)$, where G_p and $G_{1/2}$ are independent geometrics with parameters p and $\frac{1}{2}$, respectively.*

Proof: For brevity, write $g_k = g_k(\widetilde{\mathbf{X}}, \mathbf{X})$. We condition on t_i and g_i for $1 \le i \le k$. By Lemma 2, the bits $(X_i)_{i=g_k+1}^{\infty}$ are i.i.d. uniform even after this conditioning. Next, we sample t_{k+1}, which we may write as $t_{k+1} = t_k + G_p$ since each bit is retained independently with probability p. We then examine the bits

$$X_{g_k+1}, X_{g_k+2}, \ldots, X_{t_{k+1}},$$

which are still i.i.d. uniformly distributed. Recall that g_{k+1} is defined to be the earliest position of these bits where the value matches $\widetilde{X}_{k+1} = X_{t_{k+1}}$. Each of the above bits has a $\frac{1}{2}$ chance of being a match except for the last one, which is guaranteed to match. Thus, g_{k+1} may be written as $\min(g_k + G_{1/2}, t_{k+1})$, so that

$$\begin{aligned} d_{k+1} - d_k &= (t_{k+1} - t_k) - (g_{k+1} - g_k) \\ &= G_p - \min(G_{1/2}, t_{k+1} - g_k) \\ &= \max(G_p - G_{1/2}, G_p + g_k - t_{k+1}) \\ &= \max(G_p - G_{1/2}, -d_k), \end{aligned}$$

as desired. ∎

Lemma 4. *Suppose $p > \frac{1}{2}$. Let $\mathbf{X}$ be an infinite sequence of i.i.d. uniform bits, and let $(\mathbf{t}, \widetilde{\mathbf{X}}) \sim \mathcal{D}_p^*(\mathbf{x})$. Define $d_k = t_k - g_k(\widetilde{\mathbf{X}}, \mathbf{X})$. Then, there exist constants $c_p, C_p > 0$ depending only on p such that for each k, we have*

$$\mathbb{P}(d_k \geq \lambda) \leq C_p e^{-c_p \lambda}.$$

Proof: By Lemma 3, we see that $(d_k)_{k=1}^{\infty}$ behaves like a random walk on non-negative integers with a bias towards zero when $p > \frac{1}{2}$. Thus, we can show by standard arguments that the d_k have exponential tails; see [9] for details. ∎

Proof of Lemma 1: For a given string $\mathbf{z}$ of n bits and $(\mathbf{t_z}, \widetilde{\mathbf{z}}) \sim \mathcal{D}_p^*(\mathbf{z})$, write

$$\begin{aligned} d(\mathbf{z}) &= \max_{1 \leq k \leq |\mathbf{t_z}|} (t_{\mathbf{z},k} - g_k(\widetilde{\mathbf{z}}, \mathbf{z})) \\ r_\lambda(\mathbf{z}) &= \mathbb{P}_{\mathbf{z}}(d(\mathbf{z}) \geq \lambda). \end{aligned}$$

We apply Lemma 4 to the sequence $\mathbf{X}$, where we may think of $\mathbf{X}$ as the first n bits of an infinite sequence of i.i.d. uniform bits. Union bounding over all indices $1 \leq k \leq n$, we have

$$\mathbb{E}[r_\lambda(\mathbf{X})] \leq n \cdot C_{1,p} \cdot e^{-c_{1,p}\lambda},$$

where $C_{1,p}$ and $c_{1,p}$ are constants depending only on p. Consequently,

$$\mathbb{P}\left(r_\lambda(\mathbf{X}) \geq e^{-c_{1,p}\lambda/2}\right) \leq n \cdot C_{1,p} \cdot e^{-c_{1,p}\lambda/2}. \quad (1)$$

Define the event

$$E = \bigcap_{\lambda = 2\lceil \log n \rceil}^{\infty} \left\{ r_{2\lambda/c_{1,p}}(\mathbf{X}) \leq e^{-\lambda} \right\}.$$

Then, a union bound using (1) gives

$$\mathbb{P}(E) \geq 1 - n \cdot C_{1,p} \sum_{\lambda = 2\lceil \log n \rceil}^{\infty} e^{-\lambda} \geq 1 - \frac{C_{2,p}}{n}, \quad (2)$$

where $C_{2,p}$ is another constant depending only on p.

On the event E, consider any $t > \frac{2}{c_{1,p}}(2\lceil \log n \rceil + 1)$. Let $t' = \left\lfloor \frac{c_{1,p} t}{2} \right\rfloor$. Since $t' \geq 2\lceil \log n \rceil$, we have

$$\begin{aligned} \mathbb{P}(d(\mathbf{X}) \geq t) &\leq \mathbb{P}\left(d(\mathbf{X}) \geq \frac{2t'}{c_{1,p}}\right) = r_{2t'/c_{1,p}}(\mathbf{X}) \\ &\leq e^{-t'} \leq e^{-\frac{c_{1,p}t}{2} + 1}. \end{aligned} \quad (3)$$

Combining (2) and (3), we conclude that

$$\mathbb{P}(\mathbf{X} \text{ is } (C_p \log n, C_p)\text{-trackable}) \geq 1 - \frac{C_p}{n}$$

for a sufficiently large constant C_p. ∎

III. Alignment by seeing a particular sequence

In this section, we develop the tools for our second alignment strategy based on looking for a particular sequence of consecutive bits. The strategy follows the same main idea as the use of "anchors" in [1]. However, our analysis is more precise. We first establish some terminology and notation.

Definition 2. *For any two bit strings $\mathbf{w}$ and $\mathbf{y}$, we say that* **$\mathbf{w}$ occurs in $\mathbf{y}$** *if there is some index j such that $y_{j+i-1} = w_i$ for $i = 1, 2, \ldots, |\mathbf{w}|$. We use the following notation to describe occurrences:*

- *$\mathrm{Ind}_{\mathbf{w}}(\mathbf{y})$ denotes the first index at which $\mathbf{w}$ occurs in $\mathbf{y}$ (i.e. the smallest possible j as above), or ∞ if $\mathbf{w}$ does not occur in $\mathbf{y}$.*
- *Whenever $\mathrm{Ind}_{\mathbf{w}}(\mathbf{y}) < \infty$,*

  $$\mathrm{IndSet}_{\mathbf{w}}(\mathbf{y}) := \{j : \mathrm{Ind}_{\mathbf{w}}(\mathbf{y}) \leq j < \mathrm{Ind}_{\mathbf{w}}(\mathbf{w}) + |\mathbf{w}|\}$$

  denotes the set of all the indices in $\mathbf{y}$ corresponding to the occurrence of $\mathbf{w}$ in $\mathbf{y}$.

In later sections, we will be interested in occurrences of $\mathbf{w}$ within a particular substring $\mathbf{y}^{i:j}$ of $\mathbf{y}$. However, we still want to work with indices based on position in $\mathbf{y}$ rather than in $\mathbf{y}^{i:j}$. In these cases, we use the notation

- $\mathrm{Ind}_{\mathbf{w}}^{i:j}(\mathbf{y}) := \mathrm{Ind}_{\mathbf{w}}(\mathbf{y}^{i:j}) + i - 1.$
- $\mathrm{IndSet}_{\mathbf{w}}^{i:j}(\mathbf{y}) := \{k : \mathrm{Ind}_{\mathbf{w}}^{i:j}(\mathbf{y}) \leq k < \mathrm{Ind}_{\mathbf{w}}^{i:j}(\mathbf{w}) + |\mathbf{w}|\}.$

Suppose that $\mathbf{x}$ is a string of length $2n$, and $\mathbf{w} = \mathbf{x}^{(n-a+1):(n+a)}$ is a substring in the middle of $\mathbf{x}$. Now, suppose we observe a trace $\widetilde{\mathbf{x}} \sim \mathcal{D}_p(\mathbf{x})$, and we see that $\mathbf{w}$ occurs in $\widetilde{\mathbf{x}}$. We would like to say that in this case the bits in $\widetilde{\mathbf{x}}$ corresponding to the occurrence of $\mathbf{w}$ likely came from the occurrence of $\mathbf{w}$ in $\mathbf{x}$ (or at least, some of them did). Not all strings $\mathbf{x}$ have this property, but as we will see shortly, it turns out that typical ones do. We formalize the property in the following definition.

Definition 3. *Suppose $p > \frac{1}{2}$, let $\mathbf{x} \in \{0,1\}^{2n}$, and take $(\mathbf{t}, \widetilde{\mathbf{x}}) \sim \mathcal{D}_p^*(\mathbf{x})$. Consider a positive integer $a \leq n$ and positive real $\gamma < 1$, and write $\mathbf{w} = \mathbf{x}^{(n-a+1):(n+a)}$. We say that $\mathbf{x}$ is* **(a, γ)-distinguishable** *if*

$$\mathbb{P}_{\mathbf{x}}\begin{pmatrix} \mathrm{Ind}_{\mathbf{w}}(\widetilde{\mathbf{x}}) < \infty \text{ and} \\ \mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{x}})) \cap [n-a, n+a] = \emptyset \end{pmatrix} \leq \gamma^a \cdot p^{2a}.$$

Remark 2. *It is always possible for $\mathbf{w}$ to occur in $\widetilde{\mathbf{x}}$ if each of the positions $n-a+1$ through $n+a$ in $\mathbf{x}$ are retained. This happens with probability p^{2a}. The bound on the probability in the above definition is given in the form $\gamma^a \cdot p^{2a}$ to highlight that it should be smaller than p^{2a} by a factor that is exponential in a.*

The main result of this section is that random sequences are likely to be distinguishable.

Lemma 5. *Suppose $p > \frac{1}{2}$, and suppose $\mathbf{X} \in \{0,1\}^{2n}$ is chosen uniformly at random. Then, there exist $\gamma_p < 1$ and*

$c_p > 0$ *depending only on* p *such that*

$$\mathbb{P}\left(\mathbf{X} \text{ is } (\lceil n^{1/2} \rceil, \gamma_p)\text{-distinguishable}\right) \geq 1 - e^{-c_p n^{1/2}}.$$

Proof: Let $a = \lceil n^{1/2} \rceil$, let $\mathbf{w} = \mathbf{X}^{(n-a+1):(n+a)}$, and take $(\mathbf{t}, \widetilde{\mathbf{X}}) \sim \mathcal{D}_p^*(\mathbf{X})$. Let

$$\begin{aligned} J &= \mathbf{t}^{-1}\Big([1, 2n] \setminus [n-a, n+a]\Big) \\ &= \Big\{ j : 1 \leq j \leq |\widetilde{\mathbf{X}}|, \quad t_j \notin [n-a, n+a] \Big\} \end{aligned}$$

denote the set of indices of $\widetilde{\mathbf{X}}$ which did not come from the middle $2a$ positions of $\mathbf{X}$. Define the event

$$E = \Big\{ \mathrm{Ind}_{\mathbf{w}}(\widetilde{\mathbf{X}}) < \infty \quad \text{and} \quad \mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{X}})) \subseteq J \Big\},$$

which is the relevant event for (a, γ)-distinguishability.

Let us condition on the middle $2a$ bits of $\mathbf{X}$ (i.e. the bits that form $\mathbf{w}$) as well as on $\mathbf{t}$. The key observation is that $(\widetilde{X}_j)_{j \in J}$ are still i.i.d. uniform after our conditioning. Now, if $\mathbf{w}$ occurs in $\widetilde{\mathbf{X}}$, but $\mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{X}})) \subseteq J$, then it means that $\mathbf{w}$ occurs in the sequence $(\widetilde{X}_j)_{j \in J}$. However, since the $(\widetilde{X}_j)_{j \in J}$ are i.i.d., in each possible position this only happens with probability $2^{-|\mathbf{w}|} = 2^{-2a}$. Union bounding over at most $2n$ positions yields

$$\mathbb{P}(E) \leq 2n \cdot 2^{-2a},$$

where we have also taken the expectation over our initial conditioning on the middle $2a$ bits and $\mathbf{t}$.

The above probability is with respect to simultaneously two sources of randomness: the random choice of $\mathbf{X}$ and the random choice of the deletions. To highlight this, recall the notation $\mathbb{P}_{\mathbf{x}}$ for the probability over the randomness of the deletion channel for a given input string $\mathbf{x}$.

Take $\gamma_p = (2p)^{-1/2} < 1$. By Markov's inequality,

$$\begin{aligned} \mathbb{P}(\mathbb{P}_{\mathbf{X}}(E) \geq \gamma_p^a \cdot p^{2a}) &\leq \gamma_p^{-a} \cdot p^{-2a} \cdot \mathbb{E}(\mathbb{P}_{\mathbf{X}}(E)) \\ &= \gamma_p^{-a} \cdot p^{-2a} \cdot \mathbb{P}(E) \leq 2n \cdot \gamma_p^{3a} = e^{-\Omega_p(n^{1/2})}, \end{aligned}$$

which yields (a, γ_p)-distinguishability with the desired probability. ∎

We conclude the section by establishing a consequence of $(\lceil n^{1/2} \rceil, \gamma_p)$-distinguishability that is more convenient to work with than Definition 3.

Lemma 6. *Suppose* $p > \frac{1}{2}$, *let* $a = \lceil n^{1/2} \rceil$, *and suppose* $\mathbf{x} \in \{0,1\}^{2n}$ *is* (a, γ_p)-*distinguishable for some constant* $\gamma_p < 1$ *depending only on* p. *Consider* $(\mathbf{t}, \widetilde{\mathbf{x}}) \sim \mathcal{D}_p^*(\mathbf{x})$. *Then,*

$$\mathbb{P}_{\mathbf{x}}\begin{pmatrix} \mathrm{Ind}_{\mathbf{w}}(\widetilde{\mathbf{x}}) < \infty \text{ and} \\ \mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{x}})) \not\subseteq [n-10a, n+10a] \end{pmatrix} \leq e^{-\Omega_p(a)} \cdot p^{2a}.$$

Proof: The main idea is that if the set $\mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{x}}))$ intersects the interval $[n-a, n+a]$, then it is unlikely to stretch out very far from that interval.

Let

$$E_1 = \left\{ \begin{array}{c} \mathrm{Ind}_{\mathbf{w}}(\widetilde{\mathbf{x}}) < \infty \text{ and} \\ \mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{x}})) \cap [n-a, n+a] = \emptyset \end{array} \right\},$$

so that $\mathbb{P}_{\mathbf{x}}(E_1) \leq \gamma_p^a p^{2a}$ by (a, γ_p)-distinguishability. Let

$$E_2 = \left\{ \begin{array}{c} \text{more than } 7a \text{ deletions occurred among} \\ \text{some } 9a \text{ consecutive positions in } \mathbf{x} \end{array} \right\}.$$

By a standard Chernoff bound (see, e.g., [10]) and union bounding over all blocks of $9a$ bits in $\mathbf{x}$, we have

$$\begin{aligned} \mathbb{P}_{\mathbf{x}}(E_2) &\leq 2n \cdot \mathbb{P}\left(\mathrm{Binom}(9a, 1/2) > 7a\right) \leq 2n \cdot e^{-\frac{25a^2}{18a}} \\ &\leq 2n \cdot 4^{-a} \leq e^{-\Omega_p(a)} \cdot p^{2a}. \end{aligned}$$

Finally, let

$$E_3 = \left\{ \begin{array}{c} \mathrm{Ind}_{\mathbf{w}}(\widetilde{\mathbf{x}}) < \infty \text{ and} \\ \mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{x}})) \not\subseteq [n-10a, n+10a] \end{array} \right\},$$

which is the event of interest for the lemma. Suppose that E_1 holds but not E_3, i.e. $\mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{x}}))$ is not disjoint from $[n-a, n+a]$ but is also not contained within $[n-10a, n+10a]$. Then $\mathbf{t}(\mathrm{IndSet}_{\mathbf{w}}(\widetilde{\mathbf{x}}))$ must have two elements which are at least $9a$ apart, so that E_2 holds. Thus, we find that

$$\mathbb{P}_{\mathbf{x}}(E_3) \leq \mathbb{P}_{\mathbf{x}}(E_1) + \mathbb{P}_{\mathbf{x}}(E_2) \leq e^{-\Omega_p(a)} \cdot p^{2a}.$$

∎

IV. Reconstruction from approximate alignment

In this section, we adapt the trace reconstruction methods of [2] and [3] to a setting where the input string also undergoes a random shift. The main result of this section is the following lemma.

Lemma 7. *Let* k, n, *and* N *be positive integers with* $k < n < N$. *Let* $\mathbf{x}, \mathbf{x}' \in \{0,1\}^N$ *be two strings whose first* k *digits are identical but whose first* n *digits are not. Let* S *be a random variable taking integer values between* 0 *and* $k-1$.

Suppose the following conditions are satisfied:

$$\mathbb{E}[|S - \mathbb{E}S|] \leq n^{1/3}, \qquad k \leq n^{2/3}.$$

Then, for some constant C_p *depending only* p, *there exists an index* $j \leq C_p n$ *such that if* $\widetilde{\mathbf{x}} \sim \mathcal{D}_p(\mathbf{x}^{(S+1):})$ *and* $\widetilde{\mathbf{x}}' \sim \mathcal{D}_p((\mathbf{x}')^{(S+1):})$, *then*

$$\left|\mathbb{P}_{\mathbf{x}}(\widetilde{x}_j = 1) - \mathbb{P}_{\mathbf{x}'}(\widetilde{x}'_j = 1)\right| \geq \exp\left(-C_p n^{1/3}\right).$$

The first ingredient in the proof of this lemma is a polynomial identity, which is analogous to Lemma 2.1 in [2] or Section 4 in [3], but accounts for possible shifts to the input sequence.

Lemma 8. *Let* n *and* k *be positive integers with* $k \leq n$. *Let* $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ *be a sequence of real numbers whose first* k *elements are zero. Let* S *be a random variable taking integer values between* 0 *and* $k-1$, *with* $\mathbb{P}(S = i) = \beta_i$.

Let $\widetilde{\mathbf{a}} \sim \mathcal{D}_p(\mathbf{a}^{(S+1):})$, and pad $\widetilde{\mathbf{a}}$ with zeroes to the right. Then,

$$\mathbb{E}\sum_{j\geq 1}\widetilde{a}_j w^{j-1} = p\sum_{s=0}^{k-1}\beta_s(pw+q)^{-s}\sum_{j=1}^{n}a_j(pw+q)^{j-1}. \quad (4)$$

Proof: This identity can be verified by a straightforward calculation; see [9] for details. ∎

As in [2] and [3], we also use the following Littlewood-type estimate of Borwein and Erdélyi.

Lemma 9 (Borwein and Erdélyi, special case of Corollary 3.2 in [11])**.** *There exists a finite constant C such that the following holds. Let $A(z)$ be a polynomial with coefficients in $[-1,1]$ and $A(0)=1$. Denote by γ_L the arc $\left\{e^{i\theta} : -1/L \leq \theta \leq 1/L\right\}$. Then $\max_{z\in\gamma_L}|A(z)| \geq e^{-CL}$.*

Proof of Lemma 7: For a fixed value of p, clearly it is enough to prove the statement for sufficiently large n. We will assume implicitly throughout that n is sufficiently large.

Write $\beta_j = \mathbb{P}(S=j)$, let $a_j = x_j - x'_j$, and let $\mathbf{a} = (a_j)_{j=1}^n$. Define the polynomials

$$P(z) = \sum_{j=0}^{k-1}\beta_j z^j, \quad Q(z) = \sum_{j=0}^{n-1}a_{j+1}z^j,$$

$$A(z) = p\cdot P(z^{-1})Q(z).$$

Let ℓ be the smallest index for which $a_{\ell+1}\neq 0$; note that by our hypotheses, $\ell \leq n$. Define $\widetilde{Q}(z) = \frac{1}{z^\ell}Q(z)$, so that $|\widetilde{Q}(0)| = 1$.

For convenience, let $L = n^{1/3}$, and define $\rho = 1 - 1/L^2$. Applying Lemma 9 to the function $\widetilde{Q}(\rho z)$, there exists $z_0 = e^{i\theta}$ with $-\frac{p}{10L} \leq \theta \leq \frac{p}{10L}$ and $|\widetilde{Q}(\rho z_0)| \geq e^{-CL/p}$.

We next lower bound $|P(1/\rho z_0)|$. Let $\widetilde{P}(z) = z^{-\mathbb{E}S}P(z)$, which is analytic on the right half-plane. For z in the right half-plane satisfying $1 \leq |z| \leq \rho^{-1}$, differentiating $\widetilde{P}$ gives

$$\begin{aligned}|\widetilde{P}'(z)| &\leq \sum_{j=0}^{k-1}|j-\mathbb{E}S|\cdot\beta_j\cdot|z|^{j-\mathbb{E}S-1}\\ &\leq \rho^{-k}\cdot\mathbb{E}[|S-\mathbb{E}S|] \leq \rho^{-k}L \leq e^{\frac{1.1k}{L^2}}L \leq 4L,\end{aligned}$$

where we have used $\mathbb{E}[|S-\mathbb{E}S|] \leq L$ and $k \leq L^2$. Also,

$$\begin{aligned}|1/\rho z_0 - 1| &= \rho^{-1}|1-\rho z_0| \leq |z_0 - 1| + \rho^{-1}(1-\rho)\\ &\leq \frac{p}{10L} + \frac{2}{L^2} \leq \frac{p}{8L}.\end{aligned}$$

Consequently,

$$\begin{aligned}|P(1/\rho z_0)| &= \rho^{-\mathbb{E}S}|\widetilde{P}(1/\rho z_0)| \geq 1 - |\widetilde{P}(1/\rho z_0) - 1|\\ &= 1 - \left|\int_1^{1/\rho z_0}\widetilde{P}'(z)\,dz\right| \geq 1 - \left|\frac{1}{\rho z_0} - 1\right|\cdot 4L \geq \frac{1}{2}.\end{aligned}$$

Thus,

$$\begin{aligned}|A(\rho z_0)| &= p\cdot|P(1/\rho z_0)|\cdot\rho^\ell\cdot|\widetilde{Q}(\rho z_0)|\\ &\geq \frac{p}{2}\cdot e^{-\frac{1.1n}{L^2}-\frac{CL}{p}} \geq e^{-\frac{(C+2)L}{p}}.\end{aligned}$$

Next, define $w = 1 + \frac{\rho z_0 - 1}{p}$, so that $\rho z_0 = pw + q$. We have that

$$\begin{aligned}|w|^2 &= 1 + \frac{2}{p}(\rho\cdot\mathrm{Re}(z_0) - 1) + \frac{1}{p^2}|\rho z_0 - 1|^2\\ &\leq 1 + \frac{2}{p}(\rho - 1) + \frac{\rho^2}{p^2}|\rho^{-1}z_0^{-1} - 1|^2\\ &\leq 1 - \frac{2}{L^2} + \frac{1}{64L^2} \leq \rho.\end{aligned}$$

Let $\widetilde{\mathbf{a}} \sim \mathcal{D}_p(\mathbf{a}^{(S+1):})$. By Lemma 8,

$$\left|\mathbb{E}\left[\sum_{j\geq 1}\widetilde{a}_j w^{j-1}\right]\right| = |A(\rho z_0)| \geq e^{-\frac{(C+2)L}{p}}.$$

Now, take C_p to be an integer larger than $\frac{C+4}{p}$. Note that

$$\left|\sum_{j=C_p n}^{\infty}\mathbb{E}[\widetilde{a}_j]w^{j-1}\right| \leq \sum_{j=C_p n}^{\infty}\rho^j \leq L^2\rho^{C_p n} \leq \frac{1}{2}\cdot e^{-\frac{(C+2)L}{p}}.$$

Hence,

$$\left|\mathbb{E}\left[\sum_{j=1}^{C_p n-1}\widetilde{a}_j w^j\right]\right| \geq \frac{1}{2}\cdot e^{-\frac{(C+2)L}{p}} \geq e^{-(C_p-1)L},$$

and therefore, we must have for some j with $1 \leq j \leq C_p n - 1$ that

$$\begin{aligned}\left|\mathbb{P}(\widetilde{x}_j = 1) - \mathbb{P}(\widetilde{x}'_j = 1)\right| &= |\mathbb{E}\widetilde{a}_j| \geq |\mathbb{E}\widetilde{a}_j w^j|\\ &\geq \frac{1}{C_p n}e^{-(C_p-1)L} \geq e^{-C_p L},\end{aligned}$$

as desired. ∎

V. Proof of Theorem 1

Throughout this section, we fix a deletion probability $q < \frac{1}{2}$ (and hence a retention probability $p > \frac{1}{2}$). In addition, all of our inequalities are meant to apply for n sufficiently large (i.e. larger than a constant depending only on p).

Let C_p be the larger of the two constants in Lemmas 1 and 7, and let c_p be the constant in Lemma 5. We define the following integers:

$$\begin{aligned}M &= \lceil C_p \log n\rceil, & K_1 &= 40M,\\ K_0 &= \left\lceil K_1^{1/2}\right\rceil, & K_2 &= \left\lceil \frac{10}{c_p}K_1^{1/2}\log n\right\rceil.\end{aligned}$$

It is helpful to keep in mind that $K_0 = \Theta_p(\log^{1/2} n)$, $K_1 = \Theta_p(\log n)$, and $K_2 = \Theta_p(\log^{3/2} n)$.

Recall the high-level strategy of the proof from Section I-B: we align traces against what we have reconstructed so far, and then we use bit statistics to reconstruct additional bits. The alignment step in particular relies on the input $\mathbf{X}$ having certain special properties which don't hold for all strings but do hold for "most". We encapsulate these properties in the following definition.

Definition 4. *Let $\gamma_p < 1$ be the constant from Lemma 5. We say that a string $\mathbf{x} \in \{0,1\}^n$ is* **good** *if the following conditions are satisfied:*

(i). $\mathbf{x}$ *is (M, C_p)-trackable,*

(ii). *there is no run of M consecutive identical bits in $\mathbf{x}$,*

(iii). *among any K_2 consecutive bits of $\mathbf{x}$, there is a block of $2K_1$ of them that is (K_0, γ_p)-distinguishable.*

Lemma 10. *Let $\mathbf{X} \in \{0,1\}^n$ be drawn uniformly at random. Then,*

$$\mathbb{P}(\mathbf{X} \text{ is good}) \geq 1 - O_p\left(\frac{1}{n}\right).$$

Proof: It suffices to show that each condition in 4 holds with probability at least $1 - O_p\left(\frac{1}{n}\right)$. For condition (i), this is immediate by Lemma 1.

To establish condition (ii), note that the probability for M i.i.d. uniform bits to be identical is 2^{1-M}. Union bounding over all blocks of M consecutive bits in $\mathbf{X}$, we find that (ii) holds with probability at least $1 - n \cdot 2^{1-M} \geq 1 - O\left(\frac{1}{n}\right)$.

Finally, for condition (iii), note that any K_2 consecutive bits contain at least $\lfloor K_2/2K_1 \rfloor \geq \frac{2\log n}{c_p\sqrt{K_1}}$ disjoint blocks of size $2K_1$. By Lemma 5, the probability that a single block fails to be (K_0, γ_p)-distinguishable is at most $e^{-c_p\sqrt{K_1}}$. Thus, the probability that none of these blocks is (K_0, γ_p)-distinguishable is at most

$$\exp\left(-\frac{2\log n}{c_p\sqrt{K_1}} \cdot c_p\sqrt{K_1}\right) = \frac{1}{n^2}.$$

Union bounding over at most n possible blocks of K_2 consecutive bits shows that condition (iii) also holds with probability at least $1 - O\left(\frac{1}{n}\right)$. ∎

A. Alignment

Suppose $\mathbf{x}$ is a bit string that we know, and let $m \leq |\mathbf{x}|$ be some position in $\mathbf{x}$. Suppose that we also have a sample $\widetilde{\mathbf{x}}$ from the deletion channel applied to $\mathbf{x}$ (or some longer string having $\mathbf{x}$ as a prefix). As described in Section I-B, we would like to identify (with high probability) a bit of $\widetilde{\mathbf{x}}$ that was originally positioned near the m-th bit of $\mathbf{x}$. This motivates the following definition.

Definition 5. *An* **alignment rule** *is a function $\mathcal{L}$ which takes as input a bit string $\mathbf{x}$, an index $m \leq |\mathbf{x}|$, and another bit string $\mathbf{y}$. It outputs a value $\mathcal{L}(\mathbf{x}, m, \mathbf{y}) \in \{1, 2, \ldots, |\mathbf{y}| - 1, |\mathbf{y}|, \infty\}$.*

In addition, we require that $\mathcal{L}$ satisfy the following adaptedness property with respect to $\mathbf{y}$: whenever $\mathcal{L}(\mathbf{x}, m, \mathbf{y}) < \infty$, for any other string $\mathbf{y}'$ identical to $\mathbf{y}$ in their first $\mathcal{L}(\mathbf{x}, m, \mathbf{y})$ bits, we have $\mathcal{L}(\mathbf{x}, m, \mathbf{y}') = \mathcal{L}(\mathbf{x}, m, \mathbf{y})$.

Let us explain the conceptual meaning of this definition. We emphasize that for our purposes, $\mathbf{y}$ will be a sample from the deletion channel applied to a string whose prefix is $\mathbf{x}$. The idea is that bits near the m-th position of $\mathbf{x}$ should end up near the $\mathcal{L}(\mathbf{x}, m, \mathbf{y})$-th position in $\mathbf{y}$ after going through the deletion channel; in this way, the position m in $\mathbf{x}$ is "aligned" with position $\mathcal{L}(\mathbf{x}, m, \mathbf{y})$ in $\mathbf{y}$. When $\mathcal{L}(\mathbf{x}, m, \mathbf{y}) = \infty$, it means that the rule cannot reliably locate which bits of $\mathbf{y}$ came from around the m-th position of $\mathbf{x}$.

The adaptedness condition says that an alignment rule must proceed by examining the bits of $\mathbf{y}$ in order one by one, either outputting the current position, giving up and outputting ∞, or moving on to the next bit. In particular, we do not allow alignment rules to look ahead in the string $\mathbf{y}$ before deciding whether a previous position should be the output. The purpose of this requirement is to ensure that the deletion pattern after our alignment position is independent of the alignment itself.

The next lemma constructs a particular alignment rule that has good quantitative bounds on the quality of the alignment.

Lemma 11. *Let k be a given integer with $K_2 \leq k \leq n/2$, and let $\mathbf{x}_0 \in \{0,1\}^k$ be a string of length k. Then, there exists an index m with $k - K_2 + K_1 \leq m \leq k - K_1$ and an alignment rule $\mathcal{L}$ with the following property:*

For any good sequence $\mathbf{x} \in \{0,1\}^n$ with $\mathbf{x}_0$ as a prefix, taking $(\mathbf{t}, \widetilde{\mathbf{x}}) \sim \mathcal{D}_p^*(\mathbf{x})$, we have

(i). $\mathbb{P}_{\mathbf{x}}(\mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}}) < \infty) \geq \frac{1}{2}p^{2K_0}$

(ii). $\mathbb{P}_{\mathbf{x}}(|t_{\mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}})} - m| \geq K_1 \mid \mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}}) < \infty) \leq n^{-\Omega(1)}$

(iii). $\mathbb{P}_{\mathbf{x}}(|t_{\mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}})} - m| \geq 10K_0 \mid \mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}}) < \infty) \leq e^{-\Omega_p(K_0)}$.

Informally speaking, the properties in the above lemma should be interpreted as saying that (i) the alignment succeeds with some not-too-small probability; (ii) it is extremely likely to align within K_1 of the correct position; and (iii) it usually aligns within $10K_0$. Before giving the proof, we first establish an auxiliary lemma.

Lemma 12. *Suppose $\mathbf{x} \in \{0,1\}^n$ is a good sequence, and suppose $(\mathbf{t}, \widetilde{\mathbf{x}}) \sim \mathcal{D}_p^*(\mathbf{x})$. Consider any $k \leq n/2$, and let ℓ be the smallest index such that $g_\ell(\widetilde{\mathbf{x}}, \mathbf{x}) \geq k$. Then,*

$$\mathbb{P}_{\mathbf{x}}\Big(\ell \text{ exists and } k \leq t_\ell \leq k + 4M\Big) \geq 1 - \frac{1}{n}$$

for all sufficiently large n.

Proof: Let $n' = \left\lfloor \frac{5}{6} \cdot pn \right\rfloor$. If ℓ does not exist, it means that $g_{|\widetilde{\mathbf{x}}|}(\widetilde{\mathbf{x}}, \mathbf{x}) < k$ (or $\widetilde{\mathbf{x}}$ is empty). This can be bounded by

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}(\ell \text{ doesn't exist}) &\leq \mathbb{P}_{\mathbf{x}}(|\widetilde{\mathbf{x}}| < n') + \mathbb{P}_{\mathbf{x}}(g_{n'}(\widetilde{\mathbf{x}}, \mathbf{x}) < n/2) \\ &\leq \mathbb{P}_{\mathbf{x}}(|\widetilde{\mathbf{x}}| < n') + \mathbb{P}_{\mathbf{x}}(t_{n'} - g_{n'}(\widetilde{\mathbf{x}}, \mathbf{x}) \geq n/6) \\ &\quad + \mathbb{P}_{\mathbf{x}}(t_{n'} < 2n/3). \qquad (5)\end{aligned}$$

Note that $|\widetilde{\mathbf{x}}|$ is distributed as $\mathrm{Binom}(n, p)$, so $\mathbb{P}_{\mathbf{x}}(|\widetilde{\mathbf{x}}| < n') = e^{-\Omega_p(n)}$. The second term in (5) is at most $e^{-\Omega(n)}$

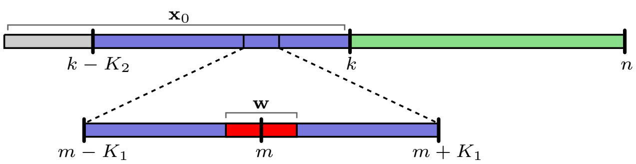

Figure V.1: Illustration of positions involved in the proof of Lemma 11.

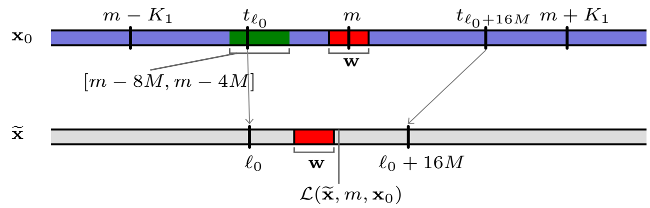

Figure V.2: A possible configuration for $\mathbf{x}_0$, m, and $\widetilde{\mathbf{x}}$. In the diagram above, events F_0, F_1, and E_1 all hold.

because $\mathbf{x}$ was assumed to be (M, C_p)-trackable. Finally, if $t_{n'} < 2n/3$, it means that at least $5pn/6$ out of the first $2n/3$ bits were retained, which also occurs with probability at most $e^{-\Omega_p(n)}$. Thus, all three probabilities in (5) are exponentially small in n, so

$$\mathbb{P}_{\mathbf{x}}(\ell \text{ does not exist}) \le \frac{1}{n^2} \tag{6}$$

for large enough n.

We now work under the assumption that ℓ exists. We always have

$$t_\ell \ge g_\ell(\widetilde{\mathbf{x}}, \mathbf{x}) \ge k \tag{7}$$

Since $\mathbf{x}$ does not have more than M consecutive identical bits, by the minimality of ℓ, we must have $g_\ell(\widetilde{\mathbf{x}}, \mathbf{x}) \le k+M$. By (M, C_p)-trackability of $\mathbf{x}$, we have

$$\mathbb{P}_{\mathbf{x}}(t_\ell > k + 4M) \le \mathbb{P}_{\mathbf{x}}(t_\ell - g_\ell(\widetilde{\mathbf{x}}, \mathbf{x}) > 3M) \le \frac{1}{n^2}. \tag{8}$$

Combining (6), (7), and (8) completes the proof. ∎

Proof of Lemma 11: If $\mathbf{x}_0$ is not a prefix of any good sequence, then there is nothing to prove. Otherwise, because $\mathbf{x}_0$ is a prefix of a good sequence, there must exist m with $k - K_2 + K_1 \le m \le k - K_1$ such that $\mathbf{x}_0^{(m-K_1+1):(m+K_1)}$ is (K_0, γ_p)-distinguishable. We choose such an m and let $\mathbf{w} = \mathbf{x}_0^{(m-K_0+1):(m+K_0)}$ (see Figure V.1).

Now, suppose $\mathbf{x}$ is any good sequence having $\mathbf{x}_0$ as a prefix, and take $(\mathbf{t}, \widetilde{\mathbf{x}}) \sim \mathcal{D}_p^*(\mathbf{x})$. Roughly speaking, our alignment rule will be to first use Lemma 12 to identify an index ℓ_0 in $\widetilde{\mathbf{x}}$ such that t_{ℓ_0} is slightly smaller than m. Then, we look for an occurrence of $\mathbf{w}$ in $\widetilde{\mathbf{x}}$ shortly after position ℓ_0. If such an occurrence exists, we output the position of the last bit of the occurrence. If not, we output ∞.

To specify the alignment rule precisely, consider any string $\mathbf{y}$, and define the statement

$$P_0(\mathbf{y}) = \begin{array}{l}\text{“there exists } \ell \text{ such that} \\ \infty > g_\ell(\mathbf{y}, \mathbf{x}_0) \ge m - 8M\text{''.}^2\end{array}$$

Whenever $P_0(\mathbf{y})$ holds, take $\ell_0(\mathbf{y})$ to be the smallest such ℓ. Then, define

$$P_1(\mathbf{y}) = P_0(\mathbf{y}) \wedge \text{“}\mathrm{Ind}_{\mathbf{w}}^{\ell_0(\mathbf{y}):(\ell_0(\mathbf{y})+16M)}(\mathbf{y}) < \infty\text{''}.$$

We then define our alignment rule to be

$$\mathcal{L}(\mathbf{x}_0, m, \mathbf{y}) = \mathrm{Ind}_{\mathbf{w}}^{\ell_0(\mathbf{y}):(\ell_0(\mathbf{y})+16M)}(\mathbf{y}) + 2K_0 - 1$$

whenever $P_1(\mathbf{y})$ holds and $\mathcal{L}(\mathbf{x}_0, m, \mathbf{y}) = \infty$ otherwise.

Note that this satisfies the adaptedness requirement for alignment rules. We will specifically apply the above definition with $\mathbf{y} = \widetilde{\mathbf{x}}$, so it is convenient to define the events

$$E_0 = \{P_0(\widetilde{\mathbf{x}}) \text{ holds}\}, \qquad E_1 = \{P_1(\widetilde{\mathbf{x}}) \text{ holds}\},$$

and we abbreviate $\ell_0 = \ell_0(\widetilde{\mathbf{x}})$.

Next, we establish properties (i), (ii), and (iii). In what follows, the reader may find it helpful to refer to Figure V.2. Define

$$F_0 = E_0 \cap \{m - 8M \le t_{\ell_0} \le m - 4M\}.$$

By Lemma 12, we have

$$\mathbb{P}_{\mathbf{x}}(F_0) \ge 1 - \frac{1}{n} \tag{9}$$

We note a subtlety in our use of the lemma: the event E_0 concerns existence of $g_\ell(\widetilde{\mathbf{x}}, \mathbf{x}_0)$, while Lemma 12 concerns existence of $g_\ell(\widetilde{\mathbf{x}}, \mathbf{x})$. However, as long as $t_{\ell_0} \le m-4M$, the relevant indices are all less than k, so there is no difference between using $\mathbf{x}_0$ and using $\mathbf{x}$, and the lemma still applies.

We can now lower bound $\mathbb{P}_{\mathbf{x}}(\mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}}) < \infty) = \mathbb{P}_{\mathbf{x}}(E_1)$. Conditioned on F_0, it is always possible for E_1 to occur by retaining all the bits in positions $m - K_0 + 1$ through $m + K_0$ in $\mathbf{x}$. Thus,

$$\mathbb{P}_{\mathbf{x}}(\mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}}) < \infty) = \mathbb{P}_{\mathbf{x}}(E_1) \ge \mathbb{P}_{\mathbf{x}}(F_0) \cdot p^{2K_0} \ge \frac{1}{2} p^{2K_0},$$

establishing property (i).

To show property (ii), consider the event

$$F_1 = \{t_{\ell_0+16M} \le m + K_1\}.$$

Note that if F_0 and F_1^c both occur, then it means that fewer than $16M$ bits were retained among the positions in $\mathbf{x}$ between $m-4M$ and $m+K_1$. There are $K_1 + 4M = 44M$ such positions, so

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}(F_0 \cap F_1^c) &\le \mathbb{P}_{\mathbf{x}}(\mathrm{Binom}(44M, p) < 16M) \\ &\le e^{-\Omega(M)} = n^{-\Omega(1)}.\end{aligned} \tag{10}$$

If F_0, F_1, and E_1 all occur, then we have

$$\begin{aligned}t_{\mathcal{L}(\mathbf{x}, m, \widetilde{\mathbf{x}})} &\ge t_{\ell_0} \ge m - 8M \ge m - K_1 \\ t_{\mathcal{L}(\mathbf{x}, m, \widetilde{\mathbf{x}})} &\le t_{\ell_0+16M} \le m + K_1.\end{aligned}$$

Thus,

$$\mathbb{P}_{\mathbf{x}}\left(|t_{\mathcal{L}(\mathbf{x},m,\widetilde{\mathbf{x}})} - m| \leq K_1 \;\middle|\; E_1\right) \geq \mathbb{P}_{\mathbf{x}}(F_0 \cap F_1 \mid E_1)$$
$$\geq 1 - \frac{\mathbb{P}_{\mathbf{x}}(F_0^c) + \mathbb{P}_{\mathbf{x}}(F_0 \cap F_1^c)}{\mathbb{P}_{\mathbf{x}}(E_1)} \geq 1 - n^{-\Omega(1)},$$

establishing property (ii).

Finally, we show property (iii). Let

$$I = \mathbf{t}^{-1}\left(\{m - K_1 + 1, m - K_1 + 2, \ldots, m + K_1\}\right)$$

be the set of indices in $\widetilde{\mathbf{x}}$ which "came from" $\mathbf{x}_0^{(m-K_1+1):(m+K_1)}$. Note that we can regard $(\widetilde{x}_i)_{i \in I}$ as being drawn from $\mathcal{D}_p\left(\mathbf{x}_0^{(m-K_1+1):(m+K_1)}\right)$. Consider the event

$$F_2 = E_1 \cap \left\{ \begin{array}{c} \mathbf{t}\left(\mathrm{IndSet}_{\mathbf{w}}^{\ell_0:(\ell_0+16M)}(\widetilde{\mathbf{x}})\right) \\ \subseteq [m - 10K_0, m + 10K_0] \end{array} \right\}.$$

Note that we have the implication

$$\mathbf{t}\left(\mathrm{IndSet}_{\mathbf{w}}^{\ell_0:(\ell_0+16M)}(\widetilde{\mathbf{x}})\right) \not\subseteq [m - 10K_0, m + 10K_0]$$
$$\text{and} \qquad [\ell_0 : (\ell_0 + 16M)] \subseteq I$$
$$\implies \mathbf{t}\left(\mathrm{IndSet}_{\mathbf{w}}^{I}(\widetilde{\mathbf{x}})\right) \not\subseteq [m - 10K_0, m + 10K_0],$$

which means

$$\begin{aligned} &\mathbb{P}_{\mathbf{x}}\left(F_2^c \cap (F_0 \cap F_1 \cap E_1)\right) \\ &\quad \leq \mathbb{P}_{\mathbf{x}}\left(\mathbf{t}\left(\mathrm{IndSet}_{\mathbf{w}}^{I}(\widetilde{\mathbf{x}})\right) \not\subseteq [m - 10K_0, m + 10K_0]\right) \\ &\quad \leq e^{-\Omega_p(K_0)} \cdot p^{2K_0}, \end{aligned} \tag{11}$$

where the last inequality follows from the fact that $\mathbf{x}_0^{(m-K_1+1):(m+K_1)}$ is (K_0, γ_p)-distinguishable combined with Lemma 6.

Recall from property (i) that $\mathbb{P}_{\mathbf{x}}(E_1) \geq \frac{1}{2}p^{2K_0}$. We conclude that

$$\mathbb{P}_{\mathbf{x}}\left(|t_{\mathcal{L}(\widetilde{\mathbf{x}})} - m| \leq 10K_0 \;\middle|\; E_1\right) \geq \mathbb{P}_{\mathbf{x}}(F_0 \cap F_2 \mid E_1)$$
$$\geq 1 - \frac{\mathbb{P}_{\mathbf{x}}(F_0^c) + \mathbb{P}_{\mathbf{x}}(E_1 \cap F_0 \cap F_1^c) + \mathbb{P}_{\mathbf{x}}(E_1 \cap F_0 \cap F_1 \cap F_2^c)}{\mathbb{P}_{\mathbf{x}}(E_1)}$$
$$\geq 1 - n^{-\Omega(1)} - n^{-\Omega(1)} - e^{-\Omega_p(K_0)} = 1 - e^{-\Omega_p(K_0)},$$

where we have used (9), (10), and (11) to bound the numerator appearing in the second line. This proves (iii). ■

B. Reconstruction

The following lemma provides a template for how we will reconstruct bits.

Lemma 13. *Consider integers k_1 and k_2 with $k_1 < k_2$, and let $\mathcal{S} \subseteq \{0,1\}^{k_2}$ be a known set of length-k_2 bit strings. Suppose that we have a number $\epsilon > 0$ and a family of statistics $b_j : \mathcal{S} \to \mathbb{R}$ for $1 \leq j \leq k_2$ which satisfies the following property: for any two strings $\mathbf{w}, \mathbf{w}' \in \mathcal{S}$ whose first k_1 bits are not identical, there exists an index $j_{\mathbf{w},\mathbf{w}'}$ such that $|b_{j_{\mathbf{w},\mathbf{w}'}}(\mathbf{w}) - b_{j_{\mathbf{w},\mathbf{w}'}}(\mathbf{w}')| > \epsilon$.*

Let $\mathbf{z} \in \mathcal{S}$ be an unknown string, and suppose that we observe estimates $(\hat{b}_j)_{j=1}^{k_2}$ such that $|\hat{b}_j - b_j(\mathbf{z})| < \epsilon/2$ for each j. Then, we can determine the first k_1 bits of $\mathbf{z}$.

Proof: For any two strings $\mathbf{w}, \mathbf{w}' \in \mathcal{S}$ whose first k_1 bits are not identical, we say that $\mathbf{w}$ *beats* $\mathbf{w}'$ if $\hat{b}_{j_{\mathbf{w},\mathbf{w}'}}$ is closer to $b_{j_{\mathbf{w},\mathbf{w}'}}(\mathbf{w})$ than to $b_{j_{\mathbf{w},\mathbf{w}'}}(\mathbf{w}')$. We say $\mathbf{w}$ is *dominant* if it beats all other strings $\mathbf{w}' \in \mathcal{S}$ that do not share its first k_1 bits.

Our hypotheses imply that $\mathbf{z}$ is dominant. Moreover, any two dominant strings must share their first k_1 bits. Thus, we may recover the first k_1 bits of $\mathbf{z}$ as the first k_1 bits of any dominant string. ■

We now apply the template in two lemmas. The first lemma reconstructs the initial K_2 bits, and the second lemma reconstructs additional bits once we have already reconstructed a long enough prefix of $\mathbf{x}$.

Lemma 14. *Let $\mathbf{x} \in \{0,1\}^n$ be a good sequence. There is a constant C_p' depending only on p such that $N = \lceil \exp\left(C_p'\sqrt{\log n}\right) \rceil$ independent samples from $\mathcal{D}_p(\mathbf{x})$ are sufficient to recover the first K_2 bits of $\mathbf{x}$ with probability at least $1 - \frac{1}{n}$ for all sufficiently large n.*

Proof: Let $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N$ be the sampled traces. For each $j \leq n$, let

$$\widetilde{x}_j^{\mathrm{avg}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{x}_{i,j}$$

be the average of the bits of the $\widetilde{\mathbf{x}}_i$ at position j, where $\widetilde{\mathbf{x}}_i$ are padded to the right with zeroes.

We will apply Lemma 13 with $k_1 = K_2$ and $k_2 = n$. We consider statistics $b_j(\mathbf{z})$ equal to the expected value of the j-th bit of a string drawn from $\mathcal{D}_p(\mathbf{z})$. By Lemma 7, we may take $\epsilon = e^{-O_p\left(K_2^{1/3}\right)} = e^{-O_p\left(\log^{1/2} n\right)}$.

Choose C_p' sufficiently large so that $\epsilon^2 N \geq e^{\sqrt{\log n}}$. Noting that $\mathbb{E}[\widetilde{x}_j^{\mathrm{avg}}] = b_j(\mathbf{x})$, by a Chernoff bound we have

$$\mathbb{P}_{\mathbf{x}}(|\widetilde{x}_j^{\mathrm{avg}} - b_j(\mathbf{x})| > \epsilon/2) \leq e^{-\frac{\epsilon^2 N}{2}} \leq \frac{1}{n^2}$$

for all large enough n. Thus, a union bound gives

$$\mathbb{P}_{\mathbf{x}}\left(\{|\widetilde{x}_j^{\mathrm{avg}} - b_j(\mathbf{x})| \leq \epsilon/2 \text{ for each } j\}\right) \geq 1 - \frac{1}{n}.$$

Using $\hat{b}_j = \widetilde{x}_j^{\mathrm{avg}}$ as our estimates, Lemma 13 asserts that we can recover the first K_2 bits of $\mathbf{x}$ when the above event holds, which proves the desired statement. ■

Lemma 15. *Let n be a positive integer, and let k be an integer with $K_2 \leq k \leq n/2$. There is a constant C_p' depending only on p such that the following holds:*

Consider a good sequence $\mathbf{x} \in \{0,1\}^n$, and suppose that $N := \lceil \exp\left(C_p'\sqrt{\log n}\right) \rceil$ i.i.d. samples $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N$ are drawn from $\mathcal{D}_p(\mathbf{x})$. Then, whenever n is sufficiently large, *seeing only the first k bits of $\mathbf{x}$ and the traces $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N$ is sufficient to recover the $(k+1)$-th bit of $\mathbf{x}$ with probability at least $1 - \frac{1}{n^2}$.*

Proof: Let m and $\mathcal{L}$ be the index and alignment rule given by Lemma 11, where we take $\mathbf{x}_0 = \mathbf{x}^{1:k}$ (which we can see). Let us consider a single trace $\widetilde{\mathbf{x}} \sim \mathcal{D}_p(\mathbf{x})$. For brevity, write $\ell = \ell(\widetilde{\mathbf{x}}) = \mathcal{L}(\mathbf{x}_0, m, \widetilde{\mathbf{x}})$.

We say that $\widetilde{\mathbf{x}}$ is a *usable* trace if $\ell < \infty$. Let E denote the event that $\widetilde{\mathbf{x}}$ is usable, and let

$$E' = E \cap \{m - K_1 \leq t_\ell \leq m + K_1\}$$

$$E'' = E \cap \{m - 10K_0 \leq t_\ell \leq m + 10K_0\}.$$

Lemma 11 ensures that $\mathbb{P}_{\mathbf{x}}(E' \mid E) \geq 1 - n^{-\Omega(1)}$ and $\mathbb{P}_{\mathbf{x}}(E'' \mid E) \geq 1 - e^{-\Omega_p(K_0)}$, which together imply that

$$\mathbb{P}_{\mathbf{x}}(E'' \mid E') \geq 1 - e^{-\Omega_p(K_0)} = 1 - e^{-\Omega_p(\log^{1/2} n)}. \quad (12)$$

Let $H = m - K_1$, and let Δ be a random variable having the same distribution as $t_\ell - H$ conditioned on E'. The reason for defining Δ in this particular way will be made clearer shortly. First, let us note several properties of Δ:

- Δ is an integer between 0 and $2K_1$.
- The distribution of Δ can be calculated just by looking at $\mathbf{x}_0$ (in particular, it does not depend on bits of $\mathbf{x}$ after the k-th one).[3]
- By (12), it is straightforward to deduce that $\mathbb{E}\Delta = m + O_p(K_0)$ and $\mathbb{E}[|\Delta - \mathbb{E}\Delta|] = O_p(K_0)$.

Define $K_3 = \left\lceil C_p'' \log^{3/2} n \right\rceil$, where C_p'' is a large enough constant to ensure that

$$\mathbb{E}[|\Delta - \mathbb{E}\Delta|] \leq K_3^{1/3}, \quad 2K_1 \leq K_3^{2/3}, \quad \text{and} \quad K_3 > K_2.$$

Our goal will be to distinguish the true suffix $\mathbf{x}^{(H+1):}$ from other possible suffixes via Lemma 13, where we take $(k_1, k_2) = (K_3, n - H)$. Here, the set $\mathcal{S}$ is taken to be all strings of length $n - H$ having $\mathbf{x}^{(H+1):k}$ as a prefix. By reconstructing the first K_3 bits of $\mathbf{x}^{(H+1):}$, we will have in particular reconstructed x_{k+1}, since

$$H + K_3 = m - K_1 + K_3 > k + 1.$$

The statistics we use are, for any $\mathbf{z} \in \mathcal{S}$,

$$b_j(\mathbf{z}) := \begin{array}{c}\text{expected value of the } j\text{-th bit of} \\ \text{a string drawn from } \mathcal{D}_p(\mathbf{z}^{(\Delta+1):}),\end{array}$$

and we take $\epsilon = e^{-C_p K_3^{1/3}}$. Note that we are able to compute these quantities $b_j(\mathbf{z})$ since we are able to compute the distribution of Δ.

We first verify the property required of the b_j and ϵ in Lemma 13. Consider any two strings $\mathbf{w}, \mathbf{w}' \in \mathcal{S}$ that do not agree in their first K_3 bits. We apply Lemma 7 to these strings with $(k, n, S) = (2K_1, K_3, \Delta)$. To check the hypotheses of the lemma, note that by the definition of $\mathcal{S}$ and the assumption $m \leq k - K_1$, $\mathbf{w}$ and $\mathbf{w}'$ agree in their first $k - H = k + K_1 - m \geq 2K_1$ bits, as required. We also recall that by the way we defined K_3, the conditions

$$\mathbb{E}[|\Delta - \mathbb{E}\Delta|] = O_p(K_0) \leq K_3^{1/3}, \qquad 2K_1 \leq K_3^{2/3}$$

are satisfied. Thus, Lemma 7 tells us that there exists an index $j_{\mathbf{w},\mathbf{w}'}$ for which

$$|b_j(\mathbf{w}) - b_j(\mathbf{w}')| \geq e^{-C_p K_3^{1/3}} = \epsilon,$$

establishing that b_j and ϵ are suitable for use in Lemma 13.

Unfortunately, we cannot directly observe samples with the law of $\mathcal{D}_p(\mathbf{x}^{(H+1+\Delta):})$ in order to estimate $b_j(\mathbf{x}^{(H+1):})$. However, a usable trace $\widetilde{\mathbf{x}}$ allows us to sample from this distribution approximately. The fact that $\mathcal{L}$ is adapted to $\widetilde{\mathbf{x}}$ (as required in Definition 5) means that if we condition on $t_{\ell(\widetilde{\mathbf{x}})} = h$ for some index h, the string $\widetilde{\mathbf{x}}^{(\ell+1):}$ has the same distribution as $\mathcal{D}_p(\mathbf{x}^{(h+1):})$. Thus, the definition of Δ ensures that, conditioned on the event E', $\widetilde{\mathbf{x}}^{(\ell+1):}$ has exactly the law of $\mathcal{D}_p(\mathbf{x}^{(H+1+\Delta):})$.

As long as $\widetilde{\mathbf{x}}$ is usable, we define $\hat{b}_j(\widetilde{\mathbf{x}}) := \widetilde{\mathbf{x}}_{\ell+j}$ and $\bar{b}_j = \mathbb{E}(\hat{b}_j(\widetilde{\mathbf{x}}) \mid E)$. The above discussion implies that

$$\left|\bar{b}_j - b_j(\mathbf{x}^{(H+1):})\right| \leq \mathbb{P}_{\mathbf{x}}(E'^c \mid E) \leq n^{-\Omega(1)} \leq \epsilon/4, \quad (13)$$

where the bound on $\mathbb{P}_{\mathbf{x}}(E'^c \mid E)$ comes from Lemma 11.

Averaging over our N traces $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N$ will then give us a fairly good estimate on $b_j(\mathbf{x}^{(H+1):})$. Choose C_p' large enough so that the following hold:

$$N \geq 64p^{-6K_0} \implies \frac{1}{2}p^{2K_0} N \geq 2N^{2/3} \quad (14)$$

$$N \geq \epsilon^{-3} e^{\sqrt{\log n}} \implies N^{2/3}\epsilon^2 = e^{\Omega(\sqrt{\log n})}. \quad (15)$$

Let M be the number of usable traces. Since our alignment rule ensures that the probability of being usable is at least $\frac{1}{2}p^{2K_0}$, it follows by a Chernoff bound and (14) that

$$\mathbb{P}_{\mathbf{x}}(M < N^{2/3}) \leq e^{-2N^{4/3}/N} = e^{-2e^{\Omega(\sqrt{\log n})}} \leq \frac{1}{n^3}.$$

Define

$$\hat{b}_j^{\text{avg}} = \frac{1}{M} \sum_{\widetilde{\mathbf{x}}_i \text{ is usable}} \hat{b}_j(\widetilde{\mathbf{x}}_i).$$

By another Chernoff bound and (15),

$$\mathbb{P}_{\mathbf{x}}\left(|\hat{b}_j^{\text{avg}} - \bar{b}_j| > \epsilon/4\right) \leq \mathbb{P}_{\mathbf{x}}(M < N^{2/3}) + e^{-\frac{-N^{2/3}\epsilon^2}{8}}$$

$$\leq \frac{1}{n^3} + e^{-e^{\Omega(\sqrt{\log n})}} \leq \frac{1}{n^2}. \quad (16)$$

Combining (13) and (16), we conclude that

$$\mathbb{P}_{\mathbf{x}}\left(|\hat{b}_j^{\text{avg}} - b_j(\mathbf{x}^{(H+1):})| < \epsilon/2 \text{ for all } j \leq n\right) \geq 1 - \frac{1}{n^2}.$$

[3] It should be noted that the probability $\mathbb{P}_{\mathbf{x}}(E)$ of having a usable trace *does* depend on later bits of $\mathbf{x}$. However, the additional constraint $m - K_1 \leq t_\ell \leq m + K_1$ combined with the adaptedness property of $\mathcal{L}$ removes this dependence.

Thus, with probability at least $1 - \frac{1}{n^2}$, the conclusion of Lemma 13 allows us to determine the first K_3 bits of $\mathbf{x}^{(H+1):}$. As noted earlier, this includes the $(k+1)$-th bit of $\mathbf{x}$, as desired. ■

C. Completing the proof

We are finally ready to prove Theorem 1, which is mostly a matter of combining Lemmas 14 and 15.

Proof of Theorem 1: We sample $N = \left\lceil e^{C'_p \sqrt{\log n}} \right\rceil$ traces, with C'_p large enough so that Lemmas 14 and 15 apply. We first condition on a realization $\mathbf{X} = \mathbf{x}$, and suppose that $\mathbf{x}$ is good. We will construct a string $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$. Let E_k denote the event that $\hat{\mathbf{x}}$ matches $\mathbf{x}$ in the first k bits. We construct the first K_2 bits of $\hat{\mathbf{x}}$ using Lemma 14, which yields

$$\mathbb{P}_{\mathbf{x}}(E_{K_2}) \geq 1 - \frac{1}{n}. \quad (17)$$

Next, consider any k with $K_2 \leq k \leq n/2$, and suppose we have constructed $\hat{x}_1, \ldots, \hat{x}_k$ already. We apply the algorithm of Lemma 15 and set $\hat{x}_{k+1}$ to its output. Although we do not have access to the first k bits of $\mathbf{x}$, we use the first k bits of $\hat{\mathbf{x}}$ instead. As long as E_k holds, this will give us the correct value for $\hat{x}_{k+1}$ with probability at least $1 - \frac{1}{n^2}$. Thus,

$$\mathbb{P}_{\mathbf{x}}(E_{k+1}) \geq \mathbb{P}_{\mathbf{x}}(E_k) - \frac{1}{n^2}. \quad (18)$$

By (17) and repeated applications of (18), we have

$$\mathbb{P}_{\mathbf{x}}(E_{\lceil n/2 \rceil}) \geq 1 - \frac{1}{n}.$$

By symmetry, we can repeat the same procedure in reverse to reconstruct the last $\lceil n/2 \rceil$ bits of $\mathbf{x}$. Thus, the total probability of failure is at most $\frac{2}{n}$.

The final possible mode of failure is if $\mathbf{x}$ is not good. However, by Lemma 10, this only happens with probability at most $\frac{1}{n}$. In total, we can reconstruct $\mathbf{X}$ with probability at least $1 - \frac{3}{n}$. Moreover, we have only used $N = e^{O_p(\sqrt{\log n})}$ traces. This completes the proof. ■

Acknowledgment

Most of this work was carried out while the second author was visiting Microsoft Research in Redmond. He thanks Microsoft for the hospitality.

References

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

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

[3] A. De, R. O'Donnell, and R. Servedio, "Optimal mean-based algorithms for trace reconstruction," in *Proceedings of the Forty-ninth Annual ACM Symposium on the Theory of Computing (STOC)*. Association for Computing Machinery, 2017, pp. 1047–1056.

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

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

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

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

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

[9] Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: subpolynomially many traces suffice," *arXiv preprint arXiv:1708.00854*, 2017.

[10] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.

[11] P. Borwein and T. Erdélyi, "Littlewood-type problems on subarcs of the unit circle," *Indiana Univ. Math. J.*, vol. 46, no. 4, pp. 1323–1346, 1997. [Online]. Available: http://dx.doi.org/10.1512/iumj.1997.46.1435