Subpolynomial trace reconstruction for random strings and arbitrary deletion probability

Nina Holden* MIT
Robin Pemantle† University of Pennsylvania
Yuval Peres‡ Microsoft Research

Abstract

The deletion-insertion channel takes as input a bit string $\mathbf{x} \in \{0,1\}^n$, and outputs a string where bits have been deleted and inserted independently at random. The trace reconstruction problem is to recover $\mathbf{x}$ from many independent outputs (called "traces") of the deletion-insertion channel applied to $\mathbf{x}$. We show that if $\mathbf{x}$ is chosen uniformly at random, then $\exp(O(\log^{1/3} n))$ traces suffice to reconstruct $\mathbf{x}$ with high probability. The earlier upper bounds were $\exp(O(\log^{1/2} n))$ for the deletion channel with deletion probability less than $1/2$, and $\exp(O(n^{1/3}))$ for the general case.

A key ingredient in our proof is a two-step alignment procedure where we estimate the location in each trace corresponding to a given bit of $\mathbf{x}$. The alignment is done by viewing the strings as random walks, and comparing the increments in the walk associated with the input string and the trace, respectively.

1 Introduction

The **deletion-insertion channel** takes as input a bit string $\mathbf{x} = (x_0, x_1, \dots, x_{n-1}) \in \{0,1\}^n$ and outputs a noisy version of it, where bits have been randomly inserted and deleted. Let $q \in [0,1)$ be the deletion probability and let $q' \in [0,1)$ be the insertion probability. First, for each j, before the jth bit of $\mathbf{x}$ we insert $G_j - 1$ uniform and independent bits, where the independent geometric random variables $G_j \geq 1$ have parameter $1 - q'$. Then we delete each bit of the resulting string independently with probability q. The output string $\widetilde{\mathbf{x}}$ is called a **trace**. An example is shown in Figure 1.

Suppose that the input string $\mathbf{x}$ is unknown. The **trace reconstruction problem** asks the following: How many i.i.d. copies of the trace $\widetilde{\mathbf{x}}$ do we need in order to determine $\mathbf{x}$ with high probability? (See Sections 1.2 and 2 for more formal problem descriptions.)

There are two variants of this problem: the "worst case" and the "average case" (also referred to as the "random case"). In the worst case variant, we want to obtain bounds which hold

*ninah@math.mit.edu
†pemantle@math.upenn.edu
‡peres@microsoft.com

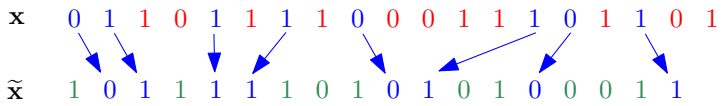

Figure 1: We obtain a trace $\widetilde{\mathbf{x}}$ by sending $\mathbf{x}$ through the deletion-insertion channel. Bits in $\mathbf{x}$ are deleted (red) or retained (blue), and new bits are inserted (green) in $\widetilde{\mathbf{x}}$.

uniformly over all possible input strings $\mathbf{x}$. In the average case variant, the input string is chosen uniformly at random. In this paper, we study the average case.

Holenstein, Mitzenmacher, Panigrahy, and Wieder [HMPW08] gave an algorithm for reconstructing random strings from the deletion channel using polynomially many traces, assuming the deletion probability q is sufficiently small. Peres and Zhai [PZ17] proved that $\exp(O(\log^{1/2} n))$ many traces suffice for the deletion channel when the deletion probability q is below $1/2$. Before the current work, the upper bound on the number of traces required for $q \geq 1/2$ was the same as for worst case strings, i.e., $\exp(O(n^{1/3}))$ [DOS17, NP17]. We improve the upper bound for all $q \in [0, 1)$, and prove a result which also holds when we allow insertions.

Theorem 1. *For $n \in \mathbb{N}$ let $\mathbf{x} \in \{0,1\}^n$ be a bit string where the bits are chosen uniformly and independently at random. Given $q, q' \in [0, 1)$ there exists $M > 0$ such that for all n we can reconstruct $\mathbf{x}$ with probability $1 - o_n(1)$ using $\lceil \exp(M \log^{1/3} n) \rceil$ traces from the deletion-insertion channel with parameters q and q'.*

We remark that the upper bound $\exp(O(\log^{1/3} n))$ in the main theorem is the best one can obtain without also improving the upper bound $\exp(O(n^{1/3}))$ for worst case strings. This holds because, given an arbitrary string of length $m = \log_{2+\varepsilon} n$ for $\varepsilon > 0$, this string will appear in a random length n string with probability converging to 1 as $n \to \infty$. In particular, a given worst case string of length m is likely to appear in our random string, and the best known algorithm for reconstructing this string requires $\exp(\Omega(m^{1/3})) = \exp(\Omega(\log^{1/3} n))$ traces.

We remark that our methods can be adapted easily to certain other reconstruction problems, e.g., to the case where one allows substitutions in addition to deletions and insertions, and the case where the bits in the input $\mathbf{x}$ are independent Bernoulli(r) random variables for arbitrary $r \in (0, 1)$, instead of $r = 1/2$.

1.1 Related work

The trace reconstruction problem dates back to the early 2000's [Lev01a, Lev01b, BKKM04]. Batu, Kannan, Khanna and McGregor, who were partially motivated by the study of mutations, considered the case where the deletion probability q is decreasing in n. They proved that if the original string $\mathbf{x}$ is random and the deletion probability $q = O(1/\log n)$, then $\mathbf{x}$ can be constructed with high probability using $O(\log n)$ samples. Furthermore, they proved that if $q = O(n^{-(1/2+\varepsilon)})$, then every string $\mathbf{x}$ can be reconstructed with high probability with $O(n \log n)$ samples.

Holenstein, Mitzenmacher, Panigrahy and Wieder [HMPW08] considered the case of random strings and constant deletion probability. They gave an algorithm for reconstruction with polynomially many traces when the deletion probability q is less than some small threshold c. The threshold c is not given explicitly in [HMPW08], but was estimated in [PZ17] to be at most 0.07.

The result of [HMPW08] was improved by Peres and Zhai [PZ17]. They showed that a subpolynomial number of traces $\exp(O(\log^{1/2} n))$ is sufficient for reconstruction, and they extended the range of allowed q to the interval $[0, 1/2)$.

Our work improves the above results in three ways. First, we improve the upper bound to $\exp(O(\log^{1/3} n))$. Second, we allow for any deletion and insertion probabilities in $[0, 1)$. Third, unlike [PZ17], our method works not only for the deletion channel, but also for the case where we allow insertions, substitutions, etc.

In [HMPW08] it is shown that $\exp(O(n^{1/2} \log n))$ traces suffice for reconstruction with high probability with worst case input. This was improved to $\exp(O(n^{1/3}))$ independently by De, O'Donnell, and Servedio [DOS17] and by Nazarov and Peres [NP17]. Until the current work, the average case upper bound was equal to the worst case upper bound for $q \geq 1/2$. The techniques developed in [DOS17, NP17] are applied in the current work and in [PZ17] to certain shorter substrings of our random string.

The best lower bounds for the number of required traces are $\Omega(\log^2 n)$ (McGregor, Price and Vorotnikova [MPV14]) in the average case and $\Omega(n)$ in the worst case ([BKKM04]). Trace reconstruction for the setting which allows insertions and substitution in addition to deletions was considered in [KM05], [VS08], [DOS17], and [NP17]. We refer to the introduction of [DOS17] and the survey [Mit09] for further background on the deletion channel.

1.2 Construction of the channel

To simplify notation we will consider bit strings of infinite length (rather than length $n \in \{1, 2, \dots\}$) throughout the paper. Observe that if we can reconstruct the first n bits of an infinite string, then we can also reconstruct length n strings. Let $\mathbb{N} = \{0, 1, \dots\}$ and let $\mathcal{S} := \{0, 1\}^{\mathbb{N}}$ denote the space of infinite sequences of zeros and ones. We denote elements of $\mathcal{S}$ by $\mathbf{x} := (x_0, x_1, \dots)$.

Fix a deletion probability q and an insertion probability q' in $[0, 1)$, and let $p = 1 - q$ and $p' = 1 - q'$. We construct $\widetilde{\mathbf{x}}$ from $\mathbf{x}$ by the procedure described above, i.e., first, for each $j \in \mathbb{N}$ we insert $G_j - 1$ uniform and independent bits before the jth bit of $\mathbf{x}$. The geometric random variables G_j are independent and satisfy

$$\mathbb{P}[G_j = v] = (q')^{v-1}(1 - q'), \qquad \forall v \in \{1, 2, \dots\}.$$

Then we delete each bit of the resulting string independently with probability q.

Let μ be the law of i.i.d. Bernoulli random variables with parameter $1/2$. We denote $\mathbb{P}_{\mathbf{x}} := \mathbb{P}_{\delta_{\mathbf{x}}}$ the law of $\widetilde{\mathbf{x}}$ when $\mathbf{x}$ is fixed; write $\mathbb{P} := \mathbb{P}_{\mu}$ for the law of $\widetilde{\mathbf{x}}$ when $\mathbf{x}$ is picked according to μ,

We call the string $\widetilde{\mathbf{x}}$ a trace. An example is given in Figure 2.

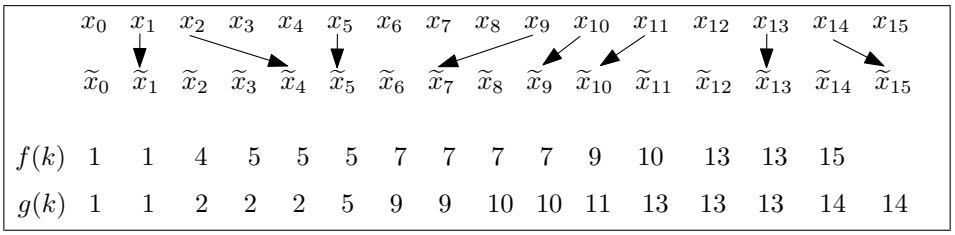

Figure 2: Illustration of the functions f and g. The arrows indicate bits which are copied from $\mathbf{x}$ to $\widetilde{\mathbf{x}}$.

For $\mathbf{x} \in \mathcal{S}$ and $0 \le i \le j < \infty$ let $\mathbf{x}(i:j) \in \{0,1\}^{j-i+1}$ denote the subsequence of $\mathbf{x}$ from position i to j. Let $\mathbf{x}(i:\infty) \in \{0,1\}^{\mathbb{N}}$ denote the substring of $\mathbf{x}$ corresponding to the bits in position i or later. Define f such that $f(k)$ is the location in $\widetilde{\mathbf{x}}$ of the first bit of $\mathbf{x}(k:\infty)$ that is preserved by the channel. Let g be the approximate inverse of f, defined such that $g(k')$ is the location in $\mathbf{x}$ of the first bit of $\widetilde{\mathbf{x}}(k':\infty)$ which was copied from $\mathbf{x}$. Observe that $g(f(k)) = k$ if and only if bit k of $\mathbf{x}$ was copied to $\widetilde{\mathbf{x}}$.

If $\mathbf{x} \in \{0,1\}^n$ is a string of finite length, then we construct the trace $\widetilde{\mathbf{x}}$ similarly: Let $q \in [0,1)$ be the deletion probability and let $q' \in [0,1)$ be the insertion probability. First, for each j, before the jth bit of $\mathbf{x}$ we insert $G_j - 1$ uniform and independent bits, where the independent geometric random variables $G_j \ge 1$ have parameter $1 - q'$. Then we delete each bit of the resulting string independently with probability q.

Worst case reconstruction problem

Let $q, q' \in [0,1)$. For any $N \in \mathbb{N}$ let $\mathbb{P}_{\mathbf{x}}^N$ denote the probability measure associated with N independent outputs of the deletion-insertion channel $\mathbb{P}_{\mathbf{x}}$ with deletion (resp. insertion) probability q (resp. q'). For $n \in \mathbb{N}$ and $\mathbf{x} \in \{0,1\}^n$ let $\mathfrak{X}$ denote a collection of $N_n \in \mathbb{N}$ traces sampled independently at random. We say that worst case strings of length n can be reconstructed with probability $1 - o_n(1)$ from N_n traces, if there is a function $G\colon \mathcal{S}^{N_n} \to \{0,1\}^n$, such that for all $\mathbf{x} \in \mathcal{S}$,

$$\mathbb{P}_{\mathbf{x}}^{N_n}[G(\mathfrak{X}) = \mathbf{x}(0:n-1)] = 1 - o_n(1).$$

Average case reconstruction problem

Let μ_n denote uniform measure on $\{0,1\}^n$. We say that uniformly random strings of length n can be reconstructed with probability $1 - o_n(1)$ from N_n traces if we can find a set $\mathcal{S}_n \subset \{0,1\}^n$ with $\mu_n(\mathcal{S}_n) = 1 - o_n(1)$, and a function $G\colon \mathcal{S}^{N_n} \to \{0,1\}^n$, such that for all $\mathbf{x} \in \mathcal{S}$ for which $\mathbf{x}(0:n-1) \in \mathcal{S}_n$, we have

$$\mathbb{P}_{\mathbf{x}}^{N_n}[G(\mathfrak{X}) = \mathbf{x}(0:n-1)] = 1 - o_n(1).$$

In particular, Theorem 1 says that uniformly random strings can be reconstructed from $N_n := \lceil \exp(M \log^{1/3} n) \rceil$ traces with probability $1 - o_n(1)$.

1.3 Outline of proof

We reconstruct the bits of $\mathbf{x}$ one by one. For any $k, n \in \mathbb{N}$ with $k < n$ we assume $\mathbf{x}(0:k)$ is given, and we want to show that with probability $1 - O(n^{-2})$ we can use $\lceil \exp(M \log^{1/3} n) \rceil$ traces to determine the subsequent bit x_{k+1}. When $\mathbf{x}$ is chosen from μ and the traces are then generated i.i.d. from $\mathbb{P}_{\mathbf{x}}$, we show that the algorithm will fail at step k with probability $O(n^{-2})$, producing an incorrect guess or no guess at all. Inductively, we see that the probability after $k+1$ steps that the algorithm has failed to correctly identify $\mathbf{x}(0:k)$ is $O(kn^{-2})$; setting $k = n$, the probability of not correctly identifying $\mathbf{x}(0:(n-1))$ is $O(n^{-1})$, as desired. For most of the paper we assume, in order to simplify notation, that $q = q'$. We explain at the end of Section 2 how to treat the case of general q, q'.

Three ingredients are required, as follows.

(*i*) A Boolean test $T(\mathbf{w}, \widetilde{\mathbf{w}})$ on pairs $(\mathbf{w}, \widetilde{\mathbf{w}})$ of bit strings of finite equal length, indicating whether $\widetilde{\mathbf{w}}$ is a plausible match for the string $\mathbf{w}$ sent through the deletion-insertion channel.

(*ii*) An alignment procedure that uses the test T repeatedly to produce for each of the independent traces $\widetilde{\mathbf{x}}$ an estimate τ for a carefully chosen position $f(k_*)$ in $\widetilde{\mathbf{x}}$ nearby $f(k)$.

(*iii*) A bit recovery procedure based on a method of [PZ17, DOS17, NP17] to produce from the approximately aligned traces an estimate of the subsequent bit or bits.

The argument is [PZ17] follows the same overall structure, with an alignment step followed by a reconstruction step for each bit in the original string. However, the greedy alignment step in [PZ17] relies crucially on the assumption that the deletion probability $q < 1/2$, and that no insertions are allowed. We overcome this problem by introducing a new kind of test for the alignment, which is based on studying correlations between blocks in the input string and in the trace.

We end the introduction by providing more details on the ingredients $(i) - (iii)$ above (in a slightly different order: (ii), (i), (iii)). Section 2 contains a proof of the main theorem modulo two key results, Theorems 2 and 3 below. Section 3 constructs the test T. Section 4 uses this to construct a good position k_* in the input to try to align. Section 5 constructs an approximate alignment τ_1 and a good alignment τ_2, and establishes properties of these culminating in one of the two key results (Theorem 2). Finally, Section 6 finishes the proof of the main theorem by proving the other key result (Theorem 3).

Alignment: Finding $f(k_*)$ in the trace

The following is our key alignment result. Assume $\mathbf{x}(0:k)$ is known for some $k \in \mathbb{N}$. We find a position $k_* < k$ satisfying $|k - k_*| = \Theta(\log n)$, and we have an algorithm which finds an estimate τ_2 for $f(k_*)$ in the trace $\widetilde{\mathbf{x}}$. For all $\mathbf{x}$ outside some exceptional set Ξ_{bad}, the

theorem gives a lower bound for the probability of true positives (meaning $|f(k_*) - \tau_2| \leq C_{\text{align}} \log^{1/3} n$, with C_{align} as below), an upper bound for the probability of false positives (meaning $|f(k_*) - \tau_2| > C_{\text{align}} \log^{1/3} n$), and an upper bound for the average discrepancy in the case of true positives.

Theorem 2 (proved in Section 5)**.** *Given any $C_{\text{sep}} > 0$ there are constants $C_{\text{back}}, C_{\text{align}}, C_{\text{false}}, C_{\text{true}}, C_{\text{avg}} \geq 1$ such that the following hold for any fixed an integer $k \in [2C_{\text{back}} \log n, n]$, and we let k_* and τ_2 be the message position and alignment pointer, respectively, produced by the alignment algorithm in Section 5 when the algorithm assumes correctly the value of $\mathbf{x}(0:k)$.*

(*i*) *τ_2 is bounded above by $2n$ when finite and (given $\mathbf{x}(0:k)$), is a stopping time on the filtration determined by $\widetilde{\mathbf{x}}$, i.e., for each i,*

$$\text{the event } \{\tau_2 \leq i\} \text{ is a function of } \widetilde{\mathbf{x}}(0:i) \text{ and } \mathbf{x}(0:k).$$

Furthermore, there is a set Ξ_{bad}, determined by the first $2n$ bits of $\mathbf{x}$, such that $\mu(\Xi_{\text{bad}}) = O(n^{-2})$ and if $\mathbf{x} \notin \Xi_{\text{bad}}$, then the following four properties hold for sufficiently large n.

(*ii*) *k_* is order $\log n$ from the end of the reconstructed input:*

$$k - C_{\text{back}} \log n \leq k_* \leq k - \frac{C_{\text{back}}}{2} \log n.$$

(*iii*) *The true positive rate is not too tiny:*

$$\mathbb{P}_{\mathbf{x}} \left[|g(\tau_2) - k_*| \leq C_{\text{align}} \log^{1/3} n \right] \geq \exp(-C_{\text{true}} \log^{1/3} n).$$

(*iv*) *The false positive rate is much smaller than the true positive rate:*

$$\mathbb{P}_{\mathbf{x}} \left[\infty > |g(\tau_2) - k_*| > C_{\text{align}} \log^{1/3} n \right] \leq \exp(-C_{\text{false}} \log^{1/3} n).$$

with $C_{\text{false}} - C_{\text{true}} > \kappa := C_{\text{sep}}(8C_{\text{avg}} + C_{\text{back}}^{1/3})$.

(*v*) *The average discrepancy when there is a true positive test is at most a small constant multiple of the threshold:*

$$\mathbb{E}_{\mathbf{x}} \left[|g(\tau_2) - k_*| \mathbf{1}_{|g(\tau_2) - k_*| \leq C_{\text{align}} \log^{1/3} n} \,\middle|\, \tau_2 < \infty \right] \leq C_{\text{avg}} \log^{1/3} n.$$

In our proof of the theorem, we find our estimate τ_2 for $f(k_*)$ in two steps. In the first step we tolerate that our initial estimate τ_1 has error $O(\log n)$, and in the second step we tolerate that our estimate τ_2 has error $O(\log^{1/3} n)$. Both steps are based on defining a function T which takes as input two intervals $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ (from the string and from the trace, respectively)

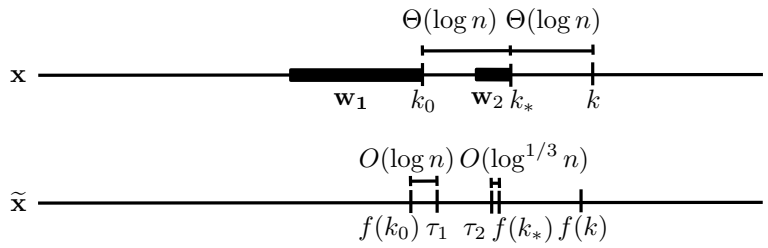

Figure 3: Illustration of indices considered in our two alignment steps. In the first alignment step we find an approximation τ_1 to $f(k_0)$, and in the second alignment step we find an approximation τ_2 to $f(k_*)$.

of the same length $\ell \in \mathbb{N}$, and returns the value 1 (resp. 0) if it seems likely (resp. unlikely) that $\widetilde{\mathbf{w}}$ was obtained by sending $\mathbf{w}$ through the deletion-insertion channel. Each string $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ can be viewed as a length ℓ random walk by looking at the partial sums (where 1 gives an increment of $+1$ and 0 gives an increment of -1), and our test is based on comparing the increments of these two walks in certain subintervals (see details below).

Assume k is at least a large constant multiple of $\log^{5/3} n$. In the first step we let $\mathbf{w}_1$ be an interval of length $\ell_1 = O(\log^{5/3} n)$ in $\mathbf{x}$, such that the right end-point k_0 of $\mathbf{w}_1$ satisfies $k_0 < k$ and $|k - k_0| = \Theta(\log n)$. For each trace we evaluate $T(\mathbf{w}_1, \widetilde{\mathbf{w}})$ for each length ℓ_1 substring $\widetilde{\mathbf{w}}$ of $\widetilde{\mathbf{x}}$, going from left to right in $\widetilde{\mathbf{x}}$. Our estimate τ_1 for $f(k_0)$ is the right end-point of the first interval $\widetilde{\mathbf{w}}$ for which $T(\mathbf{w}_1, \widetilde{\mathbf{w}}) = 1$. If we find no such interval $\widetilde{\mathbf{w}}$ in $\widetilde{\mathbf{x}}(0 : 2k_0)$, then we set $\tau_1 = \infty$, and we do not use this trace when we estimate x_{k+1}. We say that we have a true (resp. false) positive if $\tau_1 < \infty$, and if $|\tau_1 - f(k_0)|$ is smaller (resp. larger) than a constant multiple of $\log n$. We prove that, except for $\mathbf{x} \in \Xi_{\text{bad}}$, the probability of true and false positives satisfy similar bounds as in (*iii*) and (*iv*) of Theorem 2 (but with other constants than C_{true} and C_{false}).

In the second step we let $\mathbf{w}_2$ be an interval of length $\ell_2 = O(\log^{1/3} n)$ in $\mathbf{x}$, such that the right end-point k_* of $\mathbf{w}_2$ satisfies $k_0 < k_* < k$ and $|k - k_*|, |k_0 - k_*| = O(\log n)$. As above, we go through a substring of $\widetilde{\mathbf{x}}$ from left to right, and we let τ_2 be the right end-point of the first interval $\widetilde{\mathbf{w}}$ of length ℓ_2 for which $T(\mathbf{w}_2, \widetilde{\mathbf{w}}) = 1$. Using the estimate τ_1 to $f(k_0)$ from the first step, it is sufficient to only search through an interval of length $O(\log n)$ near τ_1.

Since we are using a shorter interval to align than in the first step, the test is less robust. For example, there may be several intervals $\mathbf{w}_2$ and $\widehat{\mathbf{w}}_2$ in the input string which are close to each other (distance $O(\log n)$) and similar in the sense that the associated walks have similar increments. If this is the case, we risk to consistently choose τ_2 such that $g(\tau_2)$ is near the right end-point of $\widehat{\mathbf{w}}_2$ instead of the right end-point of $\mathbf{w}_2$. In order to avoid this, we choose $\mathbf{w}_2$ carefully, such that there are no other nearby intervals $\widehat{\mathbf{w}}_2$ with this property.

The test

We will describe a simplified version of the test T. Given strings $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ of the same length $\ell \in \mathbb{N}$ and some $\lambda \in \{1, \dots, \lfloor \ell^{1/2} \rfloor\}$, divide each string $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ into $\lceil \ell/\lambda \rceil$ blocks of length

approximately λ. For $i = 1, \ldots, \lceil \ell/\lambda \rceil$ let s_i (resp. $\widetilde{s}_i$) denote the sum of $(2x_j - 1)$ as j ranges over positions in the ith block of $\mathbf{w}$ (resp. $\widetilde{\mathbf{w}}$), and we enumerate the blocks from left to right. Observe that s_i and $\widetilde{s}_i$ both have expectation 0, and that $s_i > 0$ (resp. $\widetilde{s}_i > 0$) exactly when more than half of the bits in the ith block of the string (resp. trace) are equal to 1. For some appropriately chosen $c_1 \in (0, 1)$ define

$$T(\mathbf{w}, \widetilde{\mathbf{w}}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{\lceil \ell/\lambda \rceil} \operatorname{sign}(s_i) \cdot \operatorname{sign}(\widetilde{s}_i) > c_1 \ell/\lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Observe that if $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ are sampled independently and uniformly from $\{0,1\}^\ell$, then $T(\mathbf{w}, \widetilde{\mathbf{w}}) = 0$ except on an event of probability $\exp(-\Theta(\ell/\lambda))$. On the other hand, if $\widetilde{\mathbf{w}}$ was obtained by sending $\mathbf{w}$ through the deletion-insertion channel, one can show that with probability $\exp(-\Theta(\ell/\lambda^2))$, for most blocks i a constant fraction of the bits in the trace were copied from the corresponding block in the input string. On this event, by choosing c_1 sufficiently small, we have $T(\mathbf{w}, \widetilde{\mathbf{w}}) = 1$ with uniformly positive probability. We will deduce from this that the probability of a false positive is $\exp(-\Omega(\ell/\lambda))$, while the probability of a true positive is $\exp(-O(\ell/\lambda^2))$. In the first alignment step we use the test with (ℓ, λ) of order $(\log^{5/3} n, \log^{2/3} n)$, and in the second alignment step we use the test with (ℓ, λ) of order $(\log^{1/3} n, 1)$.

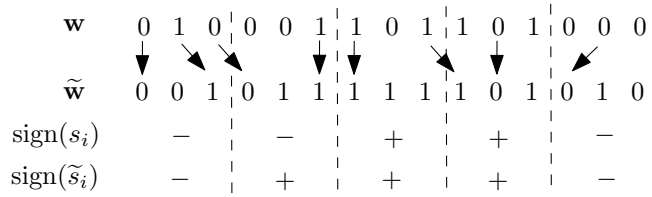

Figure 4: Illustration of our test T. We divide the length $\ell = 15$ substrings $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ of $\mathbf{x}$ and $\widetilde{\mathbf{x}}$, respectively, into blocks of length $\lambda = 3$, and find the sign of the sum of the bits in each block (replacing 0 by -1). If many bits in block i of $\widetilde{\mathbf{x}}$ were copied from the corresponding block of $\mathbf{x}$, then the sums in block i are positively correlated, and their signs are the same with probability strictly larger than $1/2$. Our test T counts how many blocks for which the signs match, and use this to predict whether the right end-points of the two substrings are likely to correspond to each other as described by the functions f and g.

The actual function T we use differs from this simplified test in two ways: First, we need to prove (v) of Theorem 2, with C_{avg} sufficiently small as compared to the other constants, in order for our reconstruction algorithm to work. For the test described above it is not clear that this holds, due to the effect described in Figure 5. To resolve this, we define a second test similarly as in (1), but where, for some $0 < c \ll 1$, we use $\mathbf{w}((k - c\ell) : \ell)$ and $\widetilde{\mathbf{w}}((k - c\ell) : \ell)$ instead of $\mathbf{w}$ and $\widetilde{\mathbf{w}}$. We require that both tests are positive when defining τ_2. The first test, which uses the full strings $\mathbf{w}$ and $\widetilde{\mathbf{w}}$, ensures that the test gives false positives with very small probability, while the second test, which uses the shorter substrings, ensures that the constant C_{avg} above is sufficiently small. The second way in which our test differs from the simplified test above, is that we choose to not sum over all the blocks $i = 1, \ldots, \lceil \ell/\lambda \rceil$ when defining T. Instead, we choose some $\theta \in (0, 1)$ and use only the $\lfloor \theta\ell/\lambda \rfloor$ blocks for which $|s_i|$ is largest. This simplifies the proof of our lower bound for the probability of having true positives.

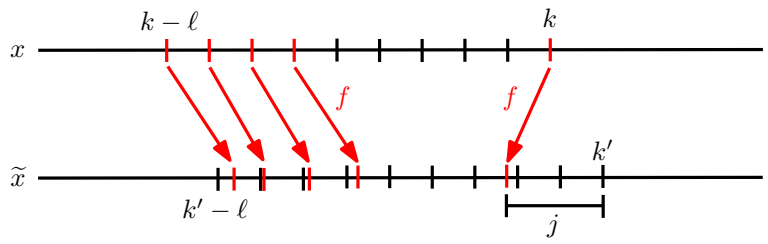

Figure 5: The red arrows to the left indicate that bits in first few block of $\mathbf{x}$ are copied to the corresponding blocks of $\widetilde{\mathbf{x}}$. If this happens for many blocks, the test T is likely to be positive, even if $j := k' - f(k)$ is large.

Consider the setting of the second alignment step above, where we evaluate $T(\mathbf{w}_2, \widetilde{\mathbf{w}})$ for all $\widetilde{\mathbf{w}}$ in a certain interval and k_* is the right end-point of $\mathbf{w}_2$. With the above test, there are three sources of false positives. First, as described above, $\mathbf{w}_2$ might be similar to some nearby substring $\widehat{\mathbf{w}}_2$ of $\mathbf{x}$, such that we often get a positive test when considering the part of the trace corresponding to $\widehat{\mathbf{w}}_2$. Second, we could have unusually many deletions or insertions right before $f(k_*)$, which could make bits in the ith block of $\widetilde{\mathbf{w}}$ be copied from the ith block of $\mathbf{w}_2$, even if the right end-point of $\widetilde{\mathbf{w}}$ is far from $f(k_*)$ (see Figure 5). Third, even if none of the above scenarios occur, there is a small chance that $T(\mathbf{w}_2, \widetilde{\mathbf{w}}) = 1$, due to the randomness of the deletions and insertions. By choosing $\mathbf{w}_2$ appropriately, we can ensure that only the two latter sources of error are relevant. Then the errors are mainly caused by the randomness of the deletions and insertions, and not the randomness of $\mathbf{x}$, so the errors happen approximately independently for each trace. Both errors have probability $\exp(-\Omega(\log^{1/3} n))$ with the parameter values used above.

From alignment to reconstruction

Finally, we explain how we can use our estimate τ_2 for $f(k_*)$ in each trace to determine x_{k+1}. In [DOS17, NP17] the authors proved that strings of length m can be reconstructed with $\lceil \exp(O(m^{1/3})) \rceil$ traces by using single bit statistics for the traces. We prove the following variant of this result (following [PZ17]) for the case where the input string has been randomly shifted before being sent through the deletion-insertion channel. The theorem implies that for different input strings $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ there exists some j, such that we can distinguish between the two strings by studying the average of the jth bit for the $\lceil \exp(M \log^{1/3} n) \rceil$ traces.

Theorem 3 (proved in Section 6)**.** *There are positive constants C_{sep} and C_{fwd} depending only on q and q', not on m, d or σ below, such that the following separation criterion holds. Let d and m satisfy $d \leq m^{2/3}$ and let $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ be any two infinite strings of bits such that $\mathbf{x}^{(1)}(0:d) = \mathbf{x}^{(2)}(0:d)$ but $\mathbf{x}^{(1)}(0:m) \neq \mathbf{x}^{(2)}(0:m)$. Let θ^s denote the shift by s on infinite bit strings and for $i = 1, 2$ and $j \geq 0$ let $q_{s,j}^{(i)} := \mathbb{P}_{\theta^s \mathbf{x}^{(i)}}[\widetilde{x}_j = 1]$ be the probability of a 1 in position j when $\mathbf{x}^{(i)}$ is shifted by s and then run through the deletion-insertion channel. Let σ be any probability measure on $\{0, \ldots, d\}$ with expected absolute deviation from its mean γ*

satisfying $\sum_{s=0}^{d} \sigma(s)|s-\gamma| \le m^{1/3}$. *Denote the averages of* $q_{s,j}^{(i)}$ *under* $s \sim \sigma$ *by*

$$q_{\sigma,j}^{(i)} := \sum_{s=0}^{d} \sigma(s) q_{s,j}^{(i)} .$$

Then there is some $j = j(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, m, d, \sigma) < C_{\mathrm{fwd}} m$, *such that*

$$\left| q_{\sigma,j}^{(1)} - q_{\sigma,j}^{(2)} \right| \ge \exp(-C_{\mathrm{sep}} m^{1/3}) . \tag{2}$$

The proof uses complex analysis techniques similar to those in [DOS17, NP17, PZ17]. We first derive an exact formula where the bit statistics are expressed as the coefficients of a particular polynomial. Then we deduce the theorem by applying a result of Borwein and Erdélyi [BE97], which says that the modulus of certain polynomials cannot be too small everywhere on a small boundary arc of the unit disk.

Applying Theorem 3 with $m = O(\log n)$ allows us to determine x_{k+1}, using our estimate τ_2 to $f(k_*)$. We apply the theorem repeatedly with all possible pairs of strings $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, such that the initial part of the strings are given by $\mathbf{x}(k_* : k)$, and we consider the traces $\widetilde{\mathbf{x}}(\tau_2 + \lfloor \log^{4/9} n \rfloor : \infty)$. Using the alignment result of Theorem 2 we can show that the random shift satisfies the assumptions of Theorem 3 with high probability. If one of the strings $\mathbf{x}^{(i)}$ is equal to our input string $\mathbf{x}$, then we can use (2) to determine from our $\lceil \exp(M \log^{1/3} n) \rceil$ traces which of the two input strings is correct. It is sufficient to consider finitely many candidate strings $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, since strings which differ only for bits very far out are unlikely to affect the part of the trace we consider.

2 Proof of main theorem modulo two key results

In this section we prove Theorem 1 modulo the two key results (Theorems 2 and 3) presented above. A set Ξ_{bad} of bad input strings depending on n will be specified such that $\mu(\Xi_{\mathrm{bad}}) = O(n^{-2})$. Fix n and k and the number $N = N_n = \lceil \exp(M \log^{1/3} n) \rceil$ of traces for some constant M to be determined later. As previously seen, it suffices to give an algorithm, which is allowed to see N traces and the correct values of $\mathbf{x}(0 : k)$, that produces a guess for x_{k+1} which is wrong with probability $O(n^{-3})$ when $\mathbf{x} \notin \Xi_{\mathrm{bad}}$ and the traces are drawn from $\mathbb{P}_{\mathbf{x}}^{N}$. First we give a alternative definition of the deletion-insertion channel.

2.1 More formal construction of the channel

Recall that $\mathbb{N} = \{0, 1, \dots\}$ and that $\mathcal{S} := \{0,1\}^{\mathbb{N}}$ denotes the space of infinite sequences of zeros and ones. Let $\Omega = \mathcal{S} \times [0,1]^{\mathbb{N}}$. We denote the first coordinate function on Ω by $\mathbf{x} := (x_0, x_1, \dots)$ and the second by $\omega := (\omega_0, \omega_1, \dots)$. Let U be the product uniform measure on $[0,1]^{\mathbb{N}}$. If ρ is any measure on $\{0,1\}^{\mathbb{N}}$, let $\mathbb{P}_\rho := \rho \times U$. We denote $\mathbb{P}_{\mathbf{x}} := \mathbb{P}_{\delta_{\mathbf{x}}}$ and $\mathbb{P} := \mathbb{P}_\mu$, where μ is the law of i.i.d. Bernoulli random variables with parameter $1/2$.

Fix a deletion probability q and an insertion probability q' in $[0,1)$, and recall that $p = 1-q$ and $p' = 1-q'$. We can construct the output $\widetilde{\mathbf{x}} = (\widetilde{x}_0, \widetilde{x}_1, \dots)$ of the deletion-insertion channel as a function of $\mathbf{x}$ and ω follows, where we view the trace $\widetilde{\mathbf{x}}$ as the string $\mathbf{x}$ run through the deletion-insertion channel with randomness ω. Temporarily denote $a := q(1-q')/(1-qq')$ and $b := q'(1-q)/(1-qq')$. For each $m \in \mathbb{N}$ we define $s(m), s'(m) \in \mathbb{N}$ inductively as follows, with $s(0) = s'(0) = 0$.

- If $\omega(m) \in [0, a]$, then define $s(m+1) = s(m) + 1$ and $s'(m+1) = s'(m)$ (deletion).
- If $\omega(m) \in (a, a + b/2]$, then set $s(m+1) = s(m)$, $s'(m+1) = s'(m) + 1$, and $\widetilde{x}_{s'(m)} = 0$ (insertion of 0).
- If $\omega(m) \in (a+b/2, a+b]$, then set $s(m+1) = s(m)$, $s'(m+1) = s'(m)+1$, and $\widetilde{x}_{s'(m)} = 1$ (insertion of 1).
- If $\omega(m) \in (a+b, 1]$, then set $s(m+1) = s(m)+1$, $s'(m+1) = s'(m)+1$, and $\widetilde{x}_{s'(m)} = x_{s(m)}$ (copy).

We justify in Lemma 4 that this version of the deletion-insertion channel is equivalent to the one given in the introduction of the paper. We remark that these two variants of the deletion-insertion channel are *not* equivalent to the variant where we first delete bits and then insert a geometric (minus 1) number of bits in the reduced string: Let $\widetilde{x}_i$ and $\widetilde{x}_j$ be the first and second, respectively, bits of $\widetilde{\mathbf{x}}$ which were copied from $\mathbf{x}$. For the deletion-insertion channel defined in the introduction, the law of $j - i$ depends on the distance between $\widetilde{x}_i$ and $\widetilde{x}_j$ in the original string; if $\widetilde{x}_i$ and $\widetilde{x}_j$ were d bits apart in the original string, then $j - i - 1$ is the sum of d independent geometric random variables (minus 1). For the variant of the channel where we first delete bits and then insert bits, the law of $j - i$ is independent of d.

Lemma 4. *Given $\mathbf{x} \in \mathcal{S}$, $q \in [0,1)$, and $q' \in [0,1)$, the following two procedures to produce the trace $\widetilde{\mathbf{x}}$ are equivalent:*

(a) First, for each $j \geq 0$, before the jth bit of $\mathbf{x}$ we insert $G_j - 1$ uniform and independent bits, where the independent geometric random variables $G_j \geq 1$ have parameter $1 - q'$. Then delete each bit of the resulting string independently with probability q.

(b) Construct $\widetilde{\mathbf{x}}$ by the inductive procedure described right above, by first sampling ω.

Proof. First observe that the procedure (b) is equivalent to the following: First mark each bit in the original string $\mathbf{x}$ independently by either D (delete) or C (copy), and then insert a geometric number minus one i.i.d. bits before each bit of the original string. The probability of D (resp. C) is equal to $q = a/(1-b)$ (resp. $1-q$) in the first step, and the geometric random variables in the second step have parameter $1 - b = 1 - q'(1-q)/(1-qq')$.

The procedure (a) can be described as follows: First insert a geometric number of bits before each bit of $\mathbf{x}$, and then mark all bits in the new string independently by either D (delete) or C (copy). The geometric random variables have parameter $1 - q'$, and the probability of D

and C is q and $1-q$, respectively. The difference from (b) is that the inserted bits may also be deleted in the second stage of the process.

To conclude that (a) and (b) are equivalent it is sufficient to show that $Z \sim \mathrm{Bin}(G-1, 1-q)$ has the law of a geometric random variable of parameter $1-b = 1-q'(1-q)/(1-qq')$ minus one, where G is a geometric random variable of parameter $1-q'$. We verify this by direct calculation, by considering the moment generating function of Z

$$\mathbb{E}[e^{tZ}] = \mathbb{E}[(q-(1-q)e^t)^{G-1}] = \frac{\frac{1-q'}{1-qq'}}{1-\frac{(1-q)q'}{1-qq'}e^t} = \frac{1-b}{1-be^t}.$$

□

Define $\psi(j) := \sup\{t \geq 0 : s'(t) = j\}$; in other words, $\psi(j)$ is the bit of ω that determines $\widetilde{x}_j$.

Now we define some σ-fields and record a strong Markov property. Define $\mathcal{G}_j^k$ to be the σ-field on Ω generated by $\mathbf{x}(0:k)$ and $\{\omega(t) : t \leq \psi(j)\}$. The differences between the σ-fields $\mathcal{G}_j^k$ for different k are irrelevant for $\mathbb{P}_{\mathbf{x}}$. We use $\mathcal{G}_j'$ for $\mathcal{G}_j^\infty$. The σ-field $\mathcal{G}_j^k$ contains $\widetilde{\mathcal{G}}_j^k := \sigma(\mathbf{x}(0:k), \widetilde{\mathbf{x}}(0:j))$ but is strictly larger because it contains information about alignment. For events in $\sigma(\omega)$, we use $\mathbb{P}_\omega$ for the common value of $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}$ for all $\mathbf{x}$.

Let θ be the shift operator on bit strings. Let $h(j)$ be the last bit of $\mathbf{x}$ examined by the time $\widetilde{x}_j$ is produced. Observe that applying $\theta^{h(j)+1}$ to $\mathbf{x}$ and $\theta^{\psi(j)+1}$ to ω induces the shift θ^{j+1} in $\widetilde{\mathbf{x}}$. As usual, if τ is a stopping time on a filtration $\{\mathcal{G}_j\}$, then $\mathcal{G}_\tau$ denotes the σ-field of events A such that $A \cap \{\tau \leq i\} \in \mathcal{G}_i$ for $i < \infty$.

2.2 Back to the proof of the main theorem

The first key result is Theorem 2, which provides an alignment algorithm. See Sections 3, 4, and 5 for a proof. Given $k < n$ and presumed values of $\mathbf{x}(0:k)$, the algorithm first defines a position $k_* < k$. Then the algorithm scans each trace and either declares failure (for that trace) or produces an alignment pointer τ_2.

The next lemma follows from the construction of $\widetilde{\mathbf{x}}$ on the canonical space $\Omega = \mathcal{S} \times [0,1]^{\mathbb{N}}$.

Lemma 5 (tails of the alignment)**.** *Let τ be any stopping time with respect to the filtration $\{\mathcal{G}_t'\}$ and suppose $q = q'$. Then there exists a constant $C_{\mathrm{RW}} > 0$ depending only on q such that for all $\mathbf{x} \in \mathcal{S}$, $a \geq 1$, and $j \in \{1, 2, \dots\}$,*

$$\mathbb{P}_{\mathbf{x}}\big[|g(\tau+j) - g(\tau) - j| \geq a \,|\, \tau < \infty\big] \quad \leq \quad \exp(-C_{\mathrm{RW}} a^2/j). \tag{3}$$

Proof. First note that since $\{g(j) - j : j \geq 0\}$ is a mean zero random walk with exponential tails, we have $\mathbb{P}_{\mathbf{x}}[|g(j) - j| \geq a] \leq \exp(-C_{\mathrm{RW}} a^2/j)$. The inequality (3) follows from a form of the strong Markov property, where $\mathbf{y} = \theta^{h(\tau)+1}\mathbf{x}$:

$$\mathbb{P}_{\mathbf{x}}\left[\theta^{h(\tau)+1}\mathbf{x} \in A, \theta^{\psi(\tau)+1}\omega \in B | \mathcal{G}_\tau'\right] = \mathbb{P}_{\mathbf{y}}\left[\mathbf{y} \in A, \omega \in B\right]$$

on the event $\{\tau < \infty\}$. This shift induces a shift of $\tau + 1$ on $\widetilde{\mathbf{x}}$, consequently, conditional on $\mathcal{G}'_\tau$, on the event $\{\tau < \infty\}$, $\{g(\tau + j) - g(\tau) - j : j \geq 1\}$ is a mean zero random walk with exponential tails. Since $g(\tau + 1) - g(\tau)$ also has exponential tails, removing the conditioning on $\mathcal{G}'_\tau$ proves (3). □

Corollary 6 (alignment farther out)**.** *Fix any $\varepsilon > 0$. Let $s := \lfloor \log^{4/9} n \rfloor$ and let E_1 be the event $\{k_* \leq g(\tau_2 + s) \leq k_* + \varepsilon \log^{2/3} n\}$. Then for sufficiently large n,*

$$\mathbb{P}_{\mathbf{x}}[\{\tau_2 < \infty\} \cap E_1^c] \leq 2 \exp(-C_{\text{false}} \log^{1/3} n)$$

provided that $\mathbf{x} \notin \Xi_{\text{bad}}$.

Proof. If $\tau_2 < \infty$ and E_1 fails then at least one of the following four events must occur:

(*i*) $g(\tau_2) - k_* \leq -C_{\text{align}} \log^{1/3} n$,

(*ii*) $g(\tau_2 + s) \leq g(\tau_2) + C_{\text{align}} \log^{1/3} n$,

(*iii*) $g(\tau_2) - k_* \geq C_{\text{align}} \log^{1/3} n$, or

(*iv*) $g(\tau_2 + s) \geq g(\tau_2) + \varepsilon \log^{2/3} n - C_{\text{align}} \log^{1/3} n$.

The first and third of these events combined have probability at most $\exp(-C_{\text{false}} \log^{1/3} n)$ by (*iv*) of Theorem 2. By Lemma 5, the second and fourth of these together have probability at most $2 \exp(-C_{\text{RW}}(\log^{4/9} n)/2)$. □

The other key result is Theorem 3, which provides a complex analytic estimate and is proved in Section 6. It concerns the result of the deletion-insertion channel after the input is randomly shifted. Under certain conditions, it is possible to conclude that for some j, the jth bit of the resulting trace will be a good test for the hypothesis $\mathbf{x} = \mathbf{x}^{(1)}$ versus $\mathbf{x} = \mathbf{x}^{(2)}$. The result in its original form is fashioned after results of [DOS17, NP17, PZ17]. Under hypotheses on the distribution of the shift, the probabilities under $\mathbb{P}_{\mathbf{x}^{(1)}}$ and $\mathbb{P}_{\mathbf{x}^{(2)}}$ of seeing a one in location j differ by at least $\exp(-C_{\text{sep}} m^{1/3})$.

To transfer Theorem 3 to the recovery setting, we require a modified result that finds the separating shift j using only the trace.

Corollary 7 (random shift in the trace)**.** *Let C_{sep} and C_{fwd} be as in Theorem 3 and let d and m satisfy $d \leq m^{2/3}$. Fix $k_* \in \mathbb{N}$ and let $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ be any two infinite strings of bits such that $\mathbf{x}^{(1)}(k_* : k_* + d) = \mathbf{x}^{(2)}(k_* : k_* + d)$ but $\mathbf{x}^{(1)}(k_* : k_* + m) \neq \mathbf{x}^{(2)}(k_* : k_* + m)$. Let τ be a (possibly infinite) stopping time on the filtration $\{\mathcal{G}'_j\}$. Let $E \subseteq \{\tau < \infty\}$ be any event measurable with respect to $\mathcal{G}'_\tau$ for which*

$$\mathbb{P}_{\mathbf{x}^{(i)}}[E^c | \tau < \infty] \leq \exp(-cm^{1/3}) \tag{4}$$

for some constant $c > C_{\text{sep}}$ and $i = 1, 2$. Suppose also that under $\mathbb{P}_{\mathbf{x}^{(i)}}$, the conditional law of $h(\tau)$ given E is supported on $\{k_*, \ldots, k_* + d - 1\}$ and has expected absolute deviation of no

more than $m^{1/3}$ from its mean. Then there is some non-random $j \le C_{\mathrm{fwd}} m$ such that for any $\varepsilon > 0$ and sufficiently large $m = m(\varepsilon)$,

$$\left|\mathbb{P}_{\mathbf{x}^{(1)}}[\widetilde{x}_{\tau+j} = 1 \,|\, \tau < \infty] - \mathbb{P}_{\mathbf{x}^{(2)}}[\widetilde{x}_{\tau+j} = 1 \,|\, \tau < \infty]\right| \ge (1-\varepsilon)\exp\left(-C_{\mathrm{sep}} m^{1/3}\right). \tag{5}$$

Proof. First we state an elementary reduction. Because

$$|\mathbb{P}_{\mathbf{x}^{(i)}}[\widetilde{x}_{\tau+j} | \tau < \infty] - \mathbb{P}_{\mathbf{x}^{(i)}}[\widetilde{x}_{\tau+j} | E]| \le \mathbb{P}_{\mathbf{x}^{(i)}}[E^c | \tau < \infty],$$

it follows from (4) and the triangle inequality that a sufficient condition for (5) is

$$|\mathbb{P}_{\mathbf{x}^{(1)}}[\widetilde{x}_{\tau+j} = 1 | E] - \mathbb{P}_{\mathbf{x}^{(2)}}[\widetilde{x}_{\tau+j} = 1 | E]| \ge \exp\left(-C_{\mathrm{sep}} m^{1/3}\right). \tag{6}$$

Next, observe that the deletion-insertion channel is Markovian with respect to the filtration $\{\mathcal{G}'_\tau\}$. In general, this means that the $\mathbb{P}_{\mathbf{x}}$-law of the pair $(\theta^{h(\tau)+1}\mathbf{x}, \theta^{\tau+1}\widetilde{\mathbf{x}})$ is the same as the $\mathbb{P}_{\theta^{h(\tau)+1}}$-law of $(\mathbf{x}, \widetilde{\mathbf{x}})$. Specifically,

$$\mathbb{P}_{\mathbf{x}}[\widetilde{x}_{\tau+j+1} = 1 \,|\, \mathcal{G}'_\tau] = \mathbb{P}_{\theta^{h(\tau)+1}\mathbf{x}}[\widetilde{x}_j = 1]. \tag{7}$$

Because $E \in \mathcal{G}'_\tau$, this implies that $\mathbb{P}_{\mathbf{x}}[\widetilde{x}_{\tau+j+1} = 1 | E]$ is a mixture of values $\mathbb{P}_{\theta^s\mathbf{x}}[\widetilde{x}_j = 1]$ where the mixing measure on s is the conditional law of $h(\tau)+1$ given E, which we denote by σ. Observe that σ is supported on $\{k_*+1, \dots, k_*+d\}$ and has absolute deviation at most $m^{1/3}$ from its mean, i.e., it satisfies the hypotheses on σ from Theorem 3 with the string $\mathbf{x}^{(i)}(k_*+1:\infty)$ in place of $\mathbf{x}^{(i)}$. The conclusion (5) of Theorem 3 then implies (6), finishing the proof of the corollary. □

We use this corollary to show that traces of strings differing somewhere before position $k_* + m$ must have distinguishable marginals in some shifted position $\tau + j$ where $j \le C_{\mathrm{fwd}} m$. Let N_1 count successful alignments, that is, $N_1 := \#\{i \le N : \tau^{(i)} < \infty\}$. Without loss of generality, renumber the traces so that the ones for which $\tau^{(i)} < \infty$ come first, that is, $\tau^{(i)} < \infty$ iff $i \le N_1$. Define

$$Y_j := \frac{1}{N_1}\sum_{i=1}^{N_1} \widetilde{x}^{(i)}_{\tau^{(i)}+j} \tag{8}$$

$$y_j(\mathbf{x}) := \mathbb{E}_{\mathbf{x}}[\widetilde{x}_{\tau+j} | \tau < \infty], \tag{9}$$

so that Y_j gives the empirical frequency of ones that occurred in (shifted) position j and $y_j(\mathbf{x})$ gives the expected value of Y_j when the input string is $\mathbf{x}$.

Lemma 8. *Suppose $\mathbf{x}, \mathbf{x}' \notin \Xi_{\mathrm{bad}}$, with $\mathbf{x}(0:k) = \mathbf{x}'(0:k)$ and $\mathbf{x}(0:k_*+m) \ne \mathbf{x}'(0:k_*+m)$. Let*

$$m := \lfloor \left((8C_{\mathrm{avg}})^3 + C_{\mathrm{back}}\right) \log n \rfloor, \tag{10}$$

$$d := \lfloor m^{2/3} \rfloor,$$

and recall that

$$\kappa \;:=\; C_{\rm sep}(8C_{\rm avg} + C_{\rm back}^{1/3}),$$

so that $C_{\rm sep}m^{1/3} = \kappa \log^{1/3} n + o_n(1)$. *For sufficiently large* n *there exists some* $j \le C_{\rm fwd}m$ *such that*

$$\left|y_j(\mathbf{x}) - y_j(\mathbf{x}')\right| \ge \frac{9}{10}\exp\left(-C_{\rm sep}m^{1/3}\right). \tag{11}$$

Proof. This is just a matter of verifying the hypotheses for Corollary 7. The quantities d and m were chosen to satisfy $d \le m^{2/3}$. The hypotheses that $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ agree to d bits but not to m bits follows from $k_* + d \le k$ and $k_* + m > k$, which follows from (i) of Theorem 2 and the definition of m. Let $\tau = \tau_2 + s$ and define E be the event $E := \{k_* < h(\tau) < k_* + d\}$. By definition of E, the conditional law of $h(\tau)$ given E is supported on $\{k_*, \ldots, k_* + d - 1\}$.

To see why inequality (4) holds, use Corollary 6 with $\varepsilon = \min\{1, (8C_{\rm avg})^2\}$ to see that $\mathbb{P}[E^c|\tau < \infty] \le 2\exp(-C_{\rm false}\log^{1/3} n)\mathbb{P}[\tau < \infty]^{-1}$ which is at most $2\exp(-(C_{\rm false} - C_{\rm true})\log^{1/3} n)$ by part (iii) of Theorem 2 and therefore by (iv) of Theorem 2 is at most $(1/2)\exp(-C_{\rm sep}(8C_{\rm avg} + C_{\rm back}^{1/3})(1+\epsilon')\log^{1/3} n) \le (1/2)\exp(-C_{\rm sep}(1+\epsilon')m^{1/3})$ for some $\epsilon' > 0$ and sufficiently large n, proving (4).

Let ν denote the conditional law $(\mathbb{P}_{\mathbf{x}}|E)$ and let $\overline{h}$ denote the ν-mean of $h(\tau)$. We will show that $\int |h(\tau) - \overline{h}|\, d\nu \le m^{1/3}$. Applying the conclusion of Corollary 7 with $\varepsilon = 1/10$ will then finish the proof by establishing (11) for $m \ge m(\varepsilon)$. Let $\overline{g}$ denote the ν-mean of $g(\tau)$. Then

$$\mathbb{E}_\nu|h(\tau) - \overline{h}| \le \mathbb{E}_\nu|h(\tau) - g(\tau)| + \mathbb{E}_\nu|g(\tau) - \overline{g}| + |\overline{g} - \overline{h}|. \tag{12}$$

If $\widetilde{x}_\tau$ was copied from $\mathbf{x}$ then $h(\tau) = g(\tau)$. Otherwise, by the strong Markov property, the law of $g(\tau) - h(\tau)$ is independent of τ. Therefore the first and third term on the right hand of (12) are independent of τ and bounded by a constant $C' > 0$ depending only on q and q'. To show that $\mathbb{E}_\nu|h(\tau) - \overline{h}| \le m^{1/3}$, it is therefore sufficient to bound $\mathbb{E}_\nu|g(\tau) - \overline{g}|$. Observe that for any random variable Y and $y \in \mathbb{R}$, Jensen's inequality gives $\mathbb{E}|y - \mathbb{E}Y| \le \mathbb{E}|y - Y|$ and hence

$$\mathbb{E}|Y - \mathbb{E}Y| \le \mathbb{E}|Y - y| + \mathbb{E}|y - \mathbb{E}Y| \le 2\mathbb{E}|Y - y|.$$

Applying this with $Y := g(\tau)$, $y = k_* + s$ and $\mathbb{P} = \nu$ gives

$$\begin{aligned}
\mathbb{E}_\nu|g(\tau) - \overline{g}| &\le 2\mathbb{E}_\nu|g(\tau) - s - k_*| \\
&\le 2\mathbb{E}_\nu|g(\tau) - g(\tau_2) - s| + 2\mathbb{E}_\nu|g(\tau_2) - k_*| \\
&= 2\mathbb{E}_\nu|g(\tau) - g(\tau_2) - s| + 2\mathbb{E}_\nu|g(\tau_2) - k_*|\mathbf{1}_{|g(\tau_2)-k_*|\le C_{\rm align}\log^{1/3} n}.
\end{aligned} \tag{13}$$

By Lemma 5 the first term on the right side is $O(s^{1/2}) = O(\log^{2/9} n)$. Using the fact that $\mathbb{P}_{\mathbf{x}}[E|\tau_2 < \infty]^{-1} \le 3/2$, the second term is bounded above by

$$3\mathbb{E}_{\mathbf{x}}\left[(g(\tau_2) - k_*)\mathbf{1}_{|g(\tau_2)-k_*|\le C_{\rm align}\log^{1/3} n}|\tau < \infty\right]$$

which is at most $3C_{\rm avg}\log^{1/3} n$ by (v) of Theorem 2. This verifies $\mathbb{E}_\nu|h(\tau) - \overline{h}| \le m^{1/3}$ whenever $4C_{\rm avg}\log^{1/3} n + 2C' \le m^{1/3}$, which holds by the definition of m for sufficiently large n, finishing the proof. □

Lemma 9. *For all $\mathbf{x}, \mathbf{x}'$ with $\mathbf{x} \notin \Xi_{\text{bad}}$ and $\mathbf{x}(0 : k_* + 4n) = \mathbf{x}'(0 : k_* + 4n)$, and for all $j \le C_{\text{fwd}} m$, if n is sufficiently large then*

$$\left| y_j(\mathbf{x}) - y_j(\mathbf{x}') \right| \le \frac{1}{100} \exp\left(-C_{\text{sep}} m^{1/3} \right) . \tag{14}$$

Proof. The conclusion of the lemma follows directly from three easily established facts:

(*i*) The $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{x}'}$ laws of τ_2 differ in total variation by at most $e^{-C_{\text{RW}} n}$.

(*ii*) $\mathbb{P}_{\mathbf{x}}[\tau_2 < \infty] \ge \exp(-C_{\text{true}} \log^{1/3} n)$.

(*iii*) $\mathbb{P}_{\mathbf{x}'}[\tau_2 < \infty] \ge \frac{1}{2} \exp(-C_{\text{true}} \log^{1/3} n)$.

Observe from Lemma 5 that the $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{x}'}$ laws of $\widetilde{\mathbf{x}}(0 : 2n)$ differ in total variation by at most $e^{-C_{\text{RW}} n}$. Because τ_2 is a stopping time and $\{\tau_2 > 2n\} = \{\tau_2 = \infty\}$, the $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{x}'}$ laws of τ_2 differ by at most $e^{-C_{\text{RW}} n}$. These two estimates yield (*i*). Fact (*ii*) follows from $\mathbf{x} \notin \Xi_{\text{bad}}$ and (*iii*) of Theorem 2, and fact (*iii*) follows for sufficiently large n by comparing the $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{x}'}$ laws of τ_2. □

Proof of Theorem 1 when $q = q'$. Fix k and assume for induction we have identified $\mathbf{x}(0 : k)$. Choose $M = 4\kappa + C_{\text{true}}$ and generate a collection of traces $\widetilde{\mathbf{x}}^{(i)}$, for $1 \le i \le N := \lceil \exp(M \log^{1/3} n) \rceil$. Let k_* and $\{\tau_2^{(i)} : 1 \le i \le N\}$ denote the result of the alignment algorithm run on the traces $\widetilde{\mathbf{x}}^{(i)}$. Let m, d, and κ be as in Lemma 8. Recall that $s := \lfloor \log^{4/9} n \rfloor$ and denote $\tau^{(i)} := \tau_2^{(i)} + s$. Clearly $\tau^{(i)}$ is a stopping time on $\{\widetilde{\mathcal{G}}_j^{k,(i)}\}$, which denotes the σ-algebra $\{\widetilde{\mathcal{G}}_j^k\}$ defined above the statement of Theorem 2 for the string $\mathbf{x}^{(i)}$.

Assume for now that $k \ge 2C_{\text{back}} \log n$ so that we may apply Theorem 2. The case $k \le 2C_{\text{back}} \log n$ will be handled separately at the end of the proof. By (*ii*) of Theorem 2, we have $k_* + d < k$, once n is sufficiently large so that $(C_{\text{back}}/2) \log n > d$. Therefore, we may assume we have identified the first d bits of $\mathbf{x}(k_* : n)$.

Lemma 10. *For each $j < C_{\text{fwd}} m$, the supremum over $\mathbf{x} \notin \Xi_{\text{bad}}$ of*

$$\mathbb{P}_{\mathbf{x}}[|Y_j - y_j(\mathbf{x})| \ge \frac{1}{100} \exp(-C_{\text{sep}} m^{1/3})]$$

decreases faster than any power of n. Consequently,

$$\mathbb{P}_{\mathbf{x}} \left[\sup_{j \le C_{\text{fwd}} m} |Y_j - y_j(\mathbf{x})| \ge \frac{1}{100} \exp(-C_{\text{sep}} m^{1/3}) \right]$$

also decreases faster than any power of n.

Proof. The event $\{|Y_j - y_j(\mathbf{x})| \geq \frac{1}{100}\exp(-C_{\text{sep}}m^{1/3})\}$ is in the union of two events $A \cup B$ where

$$\begin{aligned} A &:= \{N_1 < \exp(3\kappa \log^{1/3} n)\} \\ B &:= \{N_1 \geq \exp(3\kappa \log^{1/3} n)\} \cap \{|Y_j - y_j(\mathbf{x})| \geq \frac{1}{3}\exp(-C_{\text{sep}}m^{1/3})\}. \end{aligned}$$

The random variable N_1 is binomial with parameters $\lceil \exp(M \log^{1/3} n)\rceil$ and $p(\mathbf{x})$, the latter which is at least $\exp(-C_{\text{true}} \log^{1/3} n)$ by (iii) of Theorem 2. By choice of M, the mean of N_1 is at least $\exp(4\kappa \log^{1/3} n)$. The probability of $\text{Bin}\,(n, \lambda/n) \leq (3/4)\lambda$ decreases exponentially in λ, uniformly in n. Thus $\mathbb{P}_{\mathbf{x}}(A)$ decreases exponentially in $\exp(4\kappa \log^{1/3} n)$, hence faster than any power of n.

On the other hand, $\mathbb{P}_{\mathbf{x}}[B]$ is a mixture over values of N_1 and p of probabilities for a $\text{Bin}\,(N_1, p)$ variable to be at least $(1/100)\exp(-C_{\text{sep}}m^{1/3})N_1 \geq (1/200)\exp(-\kappa \log^{1/3} n)N_1$ away from its mean. Each of these binomials has variance at most $N_1 \geq \exp(3\kappa \log^{1/3} n)$. The probability for a binomial of variance V to be at least $\lambda V^{1/2}$ away from its mean decays exponentially in λ^2, uniformly in V. Therefore $\mathbb{P}_{\mathbf{x}}[B]$ is exponentially small in $\big((1/100)\exp(-C_{\text{sep}}m^{1/3})N_1\big)^2/N_1 \geq \big((1/100)\exp(-C_{\text{sep}}m^{1/3})\big)^2 \exp(3\kappa \log^{1/3} n) \geq \exp(\kappa \log^{1/3} n)$, hence also decaying faster than any power of n. □

Continuing the proof of Theorem 1, in the case that $q = q'$ and $k > 2C_{\text{back}} \log n$, we are now ready to reconstruct x_{k+1}. Let $\mathbf{x}|_{4n} \in \mathcal{S}$ denote the string $\mathbf{x}(0:4n)$ padded with infinitely many zeros. Observe that $\mathbf{x}|_{4n} \in \Xi_{\text{bad}}$ if and only if $\mathbf{x} \in \Xi_{\text{bad}}$. Let $\mathbf{x}_*$ denote the true input string. Let $\mathfrak{S}$ denote the set of strings $\mathbf{x}|_{4n} \notin \Xi_{\text{bad}}$ such that $\mathbf{x}(0:k) = \mathbf{x}_*(0:k)$ is the part of the message already recovered. For each $\mathbf{x} \in \mathfrak{S}$ we check whether the values $\{y_j(\mathbf{x}) : 0 \leq j \leq C_{\text{fwd}}m\}$ all agree with the corresponding observed variables $\{Y_j : 0 \leq j \leq C_{\text{fwd}}m\}$ to within $0.45\exp(-C_{\text{sep}}m^{1/3})$. Let $\mathfrak{S}'$ be the random set of all strings $\mathbf{x} \subseteq \mathfrak{S}$ that pass this test. If $\mathfrak{S}'$ is nonempty and $\mathbf{x}(0:k_*+m)$ has a common value for all $\mathbf{x} \in \mathfrak{S}'$, then we declare that this common value reconstructs of all bits up to position $k_* + m$, and in particular, reconstructs x_{k+1}.

Reconstruction of x_{k+1} fails if either $\mathfrak{S}'$ is empty or $\mathbf{x}(0:k_*+m) \neq \mathbf{x}'(0:k_*+m)$ for some $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$. If $\mathbf{x}_* \notin \Xi_{\text{bad}}$, then $\mathbf{x}_*|_{4n} \notin \Xi_{\text{bad}}$. On this event, Lemma 9, Lemma 10, and the triangle inequality imply that if we have reconstructed the first k bits correctly, then the following holds for all $j \leq C_{\text{fwd}}m$ except on an event with probability decaying faster than any polynomial in n

$$|Y_j - y_j(\mathbf{x}_*|_{4n})| < 0.44\exp(-C_{\text{sep}}m^{1/3}). \tag{15}$$

Hence when $\mathbf{x}_* \notin \Xi_{\text{bad}}$, reconstruction fails due to empty $\mathfrak{S}'$ with probability smaller than any power of n. But also, if $\mathbf{x}' \notin \Xi_{\text{bad}}$ and $\mathbf{x}'(0:k_*+m) \neq \mathbf{x}_*(0:k_*+m)$, then applying Lemma 8 to $\mathbf{x}'|_{4n}$ produces a j such that (11) holds. Together with Lemmas 9 and 10, this implies $|Y_j - y_j(\mathbf{x}'|_{4n})| > 0.45\exp(-C_{\text{sep}} \log^{1/3} m)$ except on an event whose probability decays faster than any power on n. Thus $\mathbf{x}' \notin \mathfrak{S}'$. This shows that the probability of failure at step k and success up until step k is bounded from above by the probability of $\mathbf{x} \in \Xi_{\text{bad}}$ plus a quantity decreasing faster than any polynomial in n, finishing the proof in the case $k \geq 2C_{\text{back}} \log n$.

Finally, if $k \le 2C_{\text{back}} \log n$, we work directly with the first $m' := \lfloor 2C_{\text{back}} \log n \rfloor$ bits. Applying Theorem 3, with m' in place of m and no shift ($d = 0$), any two distinct strings $\mathbf{x}$ and $\mathbf{x}'$ of length m' lead to bit statistics differing in some bit, $j \le C_{\text{fwd}} m'$, by at least $\varepsilon := \exp(-C_{\text{sep}}(m')^{1/3}) = (1 + o_n(1)) \exp(-2^{1/3} C_{\text{sep}} C_{\text{back}}^{1/3} \log^{1/3} n)$. Increasing M if necessary, $\exp(M \log^{1/3} n) > \varepsilon^{-2}$ and therefore this many traces suffice to pick out the correct initial string except with probability exponentially small in ε^{-1}, hence $o(n^{-2})$. □

Proof of Theorem 1 when $q \neq q'$. The proof of the theorem proceeds in exactly the same manner $q \neq q'$. The main difference is that our test T takes as input strings $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ which satisfy $|\widetilde{\mathbf{w}}| = \lceil |\mathbf{w}| q/q' \rceil$, instead of strings of equal length. The factor q/q' is chosen since the trace obtained from a string of length $\ell \in \mathbb{N}$ has expected length $\ell q/q'$. The test is defined exactly as before, except that the length of the blocks in the trace is scaled by q/q'. □

3 The test

In the remainder of the paper we assume that $q = q'$. In this section we give the formal definition of the test T. We also prove some estimates related to the problem of finding appropriate intervals for the alignment, and some estimates for the probability of getting false positives and true positives with our test.

3.1 Simplified test

This section is expository, describing a simplified version T_0 of the test T so that the main ideas can be outlined and motivated. The test is designed to answer whether a block $\widetilde{\mathbf{w}}$ of length ℓ in a trace is likely to have come from a block $\mathbf{w}$ of the same length in the already recovered part of the input. The test involves subdivision into blocks of size approximately $\lambda \le \sqrt{\ell}$. We will use the term **window** to denote an interval of positions of size ℓ on which a test is being run and the term **block** to denote the sub-intervals of size approximately λ within an ℓ-window. For specificity we define the right endpoints of the blocks $\{u_i\}$ given the values of k, ℓ and $\lambda \le \sqrt{\ell}$. Let $d_1 := \lceil \ell/\lambda \rceil$ denote the number of blocks and for $0 \le i \le d_1$ define $u_i := k - \ell + \lceil i\ell/d_1 \rceil$. Because $\lambda \le \sqrt{\ell} \le d_1$, this definition makes $\{(u_{i-1}, u_i] : 1 \le i \le d\}$ a partition of $(k - \ell + 1, \ldots, \ell]$ into consecutive intervals of length λ or $\lambda + 1$.

We will need to run tests for pairs (ℓ, λ) on several different scales, namely of order $(\log^{5/3} n, \log^{2/3} n)$ and $(\log^{1/3} n, 1)$, in addition to scales where the first parameter ℓ has been multiplied by a small constant. For this reason ℓ and λ remain parameters instead of being defined as fixed quantities in terms of n. Given strings $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ of length ℓ, for $1 \le i \le d_1$ we let

$$\begin{aligned} s_i &:= \sum_{j=u_{i-1}+1}^{u_i} (2x_j - 1) \\ \widetilde{s}_i &:= \sum_{j=u_{i-1}+1}^{u_i} (2\widetilde{x}_j - 1)\,. \end{aligned} \tag{16}$$

Thus, $s_i > 0$ (resp. $s_i < 0$) if and only if the majority of the bits in message block i are ones (resp. zeros), and $\widetilde{s}_i$ is the analogous majority for trace block i. For some appropriately chosen $c \in (0,1)$ define

$$T_0(\mathbf{w}, \widetilde{\mathbf{w}}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{\lceil \ell/\lambda \rceil} \operatorname{sign}(s_i) \cdot \operatorname{sign}(\widetilde{s}_i) > c\ell/\lambda, \\ 0 & \text{otherwise.} \end{cases}$$

The game plan is roughly this. Pick a window length ℓ and let $\mathbf{w}$ be the assumed already recovered bits in the interval $[k-\ell+1, \ell]$. Let $\widetilde{\mathbf{w}}$ be the trace bits in a window that slides from left to right. Wait until the test T produces a positive result, and declare the right endpoint location to be the estimate τ of $f(k)$. If the window slides to the end with no match, set $\tau = \infty$ and call this a negative test result; when $\tau < \infty$, a true positive is defined to be the event that τ estimates $f(k)$ to within a certain constant multiple of $\log^{1/3} n$; a false positive is when $\tau < \infty$ but τ does not estimate $f(k)$ to the desired accuracy; the algorithm knows when $\tau < \infty$ but does not know whether a positive is true or false. All of the work occurs in getting separation between the false and true positive rates.

The way we bound the true positive rate from below is via those traces in which the λ-blocks stay unusually well aligned. By Lemma 12 below, this occurs with probability $\exp(-\Theta(\ell/\lambda^2))$ in each case. When this occurs, for most blocks i, a constant fraction of the bits in the trace were copied from the corresponding block in the input string. On this event, by and by choosing c sufficiently small, we will have $T_0(\mathbf{w}, \widetilde{\mathbf{w}}) = 1$ with high probability.

To bound the false positive rate from above is more work because there are more ways that this can happen. One is that the right end of the window is off by more than the desired tolerance, but not too much more, and due to random fluctuations, most of the ℓ-window is actually well aligned (see Figure 5). We bound this probability from above in Lemma 14. Another way this can happen is that $\widetilde{\mathbf{w}}$ comes from a different substring $\widehat{\mathbf{w}}$ of the input but $\mathbf{w}$ and $\widehat{\mathbf{w}}$ happen to be very similar. This is the hardest aspect to deal with because in fact there will be pairs of identical substrings of length $\Theta(\log n)$ in the input. When k is the right endpoint of an ℓ-window that is too similar to another nearby ℓ-window, we have no choice but to try to align at a slightly different location, k_*. Much of the work in the previous section was the adaptation of results of [PZ17, NP17, DOS17] to show that aligning at k_* is good enough to complete the argument. Section 3.5 formulates a criterion for a position k_* in the message string to mark the right end of an ℓ-window sufficiently dissimilar from all other ℓ-windows whose right endpoint is near and left of k. Section 4 then shows that one can find a $k_* = k - \Theta(\log n)$ satisfying this criterion. The last way for a false positive to occur is by pure chance: $\widetilde{\mathbf{w}}$ comes from an input segment looking nothing like $\mathbf{w}$ but the number of sign matches is great enough so that $T_0(\mathbf{w}, \widetilde{\mathbf{w}}) = 1$. The probability of this is bounded from above by an elementary large deviation computation.

3.2 Estimates involving ω but not $\mathbf{x}$

The values of $f(t) - t$ form a mean zero random walk, as do the values of $g(s) - s$. This random walk depends only on ω, not $\mathbf{x}$. It may be helpful when discussing alignment to keep

in mind that plugging in $g(i)$ for j in $f(j)-j$ yields $f(g(j))-g(j)$ which is nearly identical to $j-g(j)$, so there are multiple ways of defining the closeness of alignment; we will use the most convenient at the time. The following is a useful formulation for whether the various blocks within a window remain aligned roughly the way they are aligned at the right endpoint.

Definition 11. *Fix a position k, a window length ℓ and a block length λ and let d_1 and $\{u_i : 0 \le i \le d_1\}$ be as in the beginning of Section 3.1. Say that a trace is* **λ-aligned** *in the interval $[k-\ell+1, k]$ if for all $j \in \{1, \dots, \ell\}$ such that*

$$f(g(j+f(k)-\ell)) = j+f(k)-\ell \qquad \text{and} \qquad k-f(k)+u_{i-1} < j \le k-f(k)+u_i,$$

it holds that

$$g(j+f(k)-\ell) \in \Big[u_{i-1} - \frac{\lambda}{100}, u_i + \frac{\lambda}{100}\Big].$$

Informally, if some bit in the ith block of the trace was copied from the input string, then this bit has distance at most $\lambda/100$ from the ith block of the input string.

Lemma 12. *There exists a $c \in (0,1)$ depending only on q, such that for all $\ell \in \mathbb{N}$, $k \in \{\ell, \ell+1, \dots\}$, and $\lambda \in \{\lceil c^{-1} \rceil, \dots, \lfloor \ell^{1/2} \rfloor\}$, the probability that the trace is λ-aligned in $[k-\ell+1, k]$ is at least $\exp(-\ell/(10c\lambda^2))$.*

Proof. We will prove a stronger result, namely that the random walk $(X_j)_{1 \le j \le \ell}$ satisfies the following with probability at least $\exp(-\ell/(2c\lambda^2))$

$$|X_j| < \lambda/200, \ \forall j \in \{1, \dots, \ell\} \qquad X_j := g(j+f(k)-\ell) - (j+k-\ell). \tag{17}$$

Divide the interval $[0, \ell]$ into $\lceil \ell/\lambda^2 \rceil$ intervals of length λ^2 or $\lambda^2 - 1$. We say that X is well-aligned in one of these intervals $[t_1, t_2]$ if $|X_j| < \frac{\lambda}{200}$ for all $j \in [t_1, t_2]$ and $|X_{t_2}| < \frac{\lambda}{400}$. By Donsker's theorem, on each of the $\lceil \ell/\lambda^2 \rceil$ intervals the process $(m\alpha)^{-1/2} X_{\lceil mt \rceil}$ converges in law to a standard Brownian motion as $m \to \infty$. Therefore, given that X is well-aligned in the first $m-1$ intervals, it is well-aligned in the mth interval with uniformly positive probability. If X is well aligned in all intervals, then the desired property (17) holds, finishing the proof. □

The interval in our second alignment step (recalling the description at the end of Section 1.3) will be chosen in order to minimize the probability of false positives. The following event $E_{\ell;k,k'}$ will help us to distinguish false positives caused by the deletions and insertions ω, from false positives caused by particular patterns in the input string $\mathbf{x}$. See Figure 5 for an illustration of the *complement* $E^c_{\ell;k,k'}$ of the non-overlapping event.

Definition 13. *Given $\ell, k \in \mathbb{N}$ and $k' \in \mathbb{Z}$ we say that the non-overlapping event $E_{\ell;k,k'}$ occurs if $k, k' \ge \ell - 1$, and if either*

(i) $f(k-i) - (k'-i) > \sqrt{\ell}$ *for* $i = 0, \dots, \ell$, *or*

(ii) $f(k-i) - (k'-i) < -\sqrt{\ell}$ *for* $i = 0, \dots, \ell$.

Lemma 14. *For $c \in (0,1)$ sufficiently small depending only on q,*

$$\mathbb{P}_\omega[E^c_{\ell;k,f(k)+j}] \le c^{-1}\exp(-10cj^2/\ell). \tag{18}$$

Proof. The values of $\{f(k-i)-(f(k)-i) : 0 \le i \le \ell\}$ rescale to a Brownian bridge, therefore we may bound the probability by considering $\sup_{0\le t\le 1}|B_t|$, which has Gaussian tails. □

3.3 Clear robust bias and the real test

Let $\ell \in \mathbb{N}$, $C_1 \ge 1$, $c_1 > 0$, and $\lambda \in \{2,\dots,\lfloor \ell^{1/2}\rfloor\}$. Given a string $\mathbf{x} \in \mathcal{S}$ and a trace $\widetilde{\mathbf{x}} \in \mathcal{S}$ we will define a function $T = T^{\lambda,C_1,\ell,c_1}_{\mathbf{x},\widetilde{\mathbf{x}}} : \{\ell,\ell+1,\dots\}^2 \to \{0,1\}$, which indicates for each pair $k,k' \ge \ell$ whether we are likely to have $f(k) = k'$. Recall $d_1 = \lceil \ell/\lambda\rceil$ and the intervals $(u_{i-1},u_i]$, $1 \le i \le d_1$. The **robust bias** of the block $\mathbf{x}(u_{i-1}+1:u_i)$ is defined by

$$\lambda^{-1/2} \inf_{\substack{t_1,t_2\in\mathbb{N}\,:\\ |t_1-u_{i-1}|<\lambda/100\\ |t_2-u_i|<\lambda/100}} \Big|\sum_{j=t_1}^{t_2}(2x_j-1)\Big|. \tag{19}$$

We say that a block has a **clear robust bias** if its robust bias is at least 1. See Figure 6 for an illustration. For some $\theta \in (0,1/10)$ let $\mathcal{I}_1 \subset \{1,\dots,d_1\}$ be the $\lceil \theta d_1\rceil$ blocks for which the robust bias is largest (with draws resolved in some arbitrary way). By Donsker's Theorem, for θ sufficiently small and λ sufficiently large compared to θ, it holds with high probability for large ℓ that all blocks in $\mathcal{I}_1$ have a clear robust bias. We fix such a choice of θ as follows for B a standard Brownian motion

$$\theta := \frac{1}{10}\mathbb{P}\left[\inf_{t_1\in[0,1/50],t_2\in[1,1+1/50]}|B_{t_2}-B_{t_1}|>1\right] > 0\,. \tag{20}$$

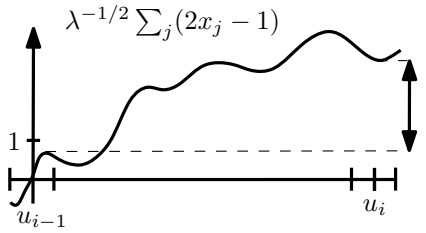

Figure 6: The length of the vertical arrow describes the robust bias associated with the block $\mathbf{x}(u_{i-1}+1:u_i)$. The curve represents the partial sums $\lambda^{-1/2}\sum_j(2x_j-1)$, renormalized to equal 0 at u_{i-1}. We say that the robust bias is clear if it is at least 1, such as shown in the given example.

Define $T_1 = T^{\lambda,\ell,c_1}_{1;\mathbf{x},\widetilde{\mathbf{x}}} : \mathbb{Z}^2 \to \{0,1\}$ by

$$T_1(k,k') = \begin{cases} 1 & \text{if } k',k \ge \ell-1 \text{ and } \displaystyle\sum_{i\in\mathcal{I}_1}\mathrm{sign}(s_i)\cdot\mathrm{sign}(\widetilde{s}_i) > c_1|\mathcal{I}_1|\,,\\ 0 & \text{otherwise.}\end{cases} \tag{21}$$

Define $d_2 := \lceil \ell/(\lambda C_1) \rceil$, and let $\mathcal{I}_2 \subset \{1, \dots, d_1\}$ be the set consisting of the $\lceil \theta d_2 \rceil$ blocks which are contained in $\mathbf{x}(k - \lceil \ell/C_1 \rceil + 1 : k)$ and which have the largest robust bias. Define $T_2 = T^{\lambda, C_1, \ell, c_1}_{2;\mathbf{x},\widetilde{\mathbf{x}}} : \mathbb{Z}^2 \to \{0, 1\}$ by

$$T_2(k, k') = \begin{cases} 1 & \text{if } k', k \geq \ell - 1 \text{ and } \sum_{i \in \mathcal{I}_2} \operatorname{sign}(s_i) \cdot \operatorname{sign}(\widetilde{s}_i) > c_1 |\mathcal{I}_2|, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 15. *Define* $T = T^{\lambda, C_1, \ell, c_1}_{\mathbf{x},\widetilde{\mathbf{x}}} : \mathbb{Z}^2 \to \{0, 1\}$ *by*

$$T(k, k') := T_1(k, k') \wedge T_2(k, k').$$

Note that the value of $T^{\lambda, C_1, \ell, c_1}_{\mathbf{x},\widetilde{\mathbf{x}}}(k, k')$ *depends on* $\mathbf{x}$ *and* $\widetilde{\mathbf{x}}$ *only on the windows* $\mathbf{x}(k - \ell + 1 : k)$ *and* $\widetilde{\mathbf{x}}(k' - \ell + 1 : k')$*, respectively.*

The purpose of requiring both $T_1(k, k') = 1$ and $T_2(k, k') = 1$ is the following. We require $T_1(k, k') = 1$ since this condition makes it very unlikely that the test gives false positives, due to the large number of blocks used in the test T_1. However, if we had only required $T_1(k, k') = 1$ (not also $T_2(k, k') = 1$) the long length ℓ of the string would have made it rather likely that $|f(k_*) - \tau_2|$ equals a large multiple of $\log^{1/3}(n)$ in the second alignment step, due to the effect described in Figure 5. We require $T_2(k, k') = 1$ to make sure that $|f(k_*) - \tau_2|$ is typically not more than a *small* constant multiple of $\log^{1/3}(n)$, on the event that we have a true positive test.

Lemma 16. *Let* $K_1 \in \sigma(\omega)$ *be the event that the trace is* λ*-aligned in* $[k - \ell + 1, k]$*. There are* $\delta, \lambda_0, c_1 > 0$ *depending only on* q *such that for all* $C_1 \geq 1$*, we have* $\mathbb{P}_{\mathbf{x}}\big[T^{\lambda, C_1, \ell, c_1}(k, f(k)) = 1 | K_1\big] \geq \delta$ *when* $\lambda > \lambda_0$ *and* $\mathbf{x}$ *is such that all the blocks in* $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *have clear robust bias.*

Proof. Let $j_1(i)$ (resp. $j_2(i)$) denote the position of the first (resp. last) bit in $\mathbf{x}$ that was copied to a position in block i of $\widetilde{\mathbf{x}}$. Let m_i denote the number of bits in block i of $\widetilde{\mathbf{x}}$ that were copied from $\widetilde{\mathbf{x}}$. Let $\mathcal{H}$ be the σ-field containing each $j_1(i), j_2(i)$ and m_i for all $i \leq d_1$. Observe that $K_1 \in \mathcal{H}$. Under $\mathbb{P}_{\mathbf{x}}$, conditional on $\mathcal{H}$, the $\lambda + 1 - m_i$ or $\lambda - m_i$ non-copied bits in the ith block of $\widetilde{\mathbf{x}}$ are all inserted bits, therefore are i.i.d. mean zero Rademacher variables (and independent as i varies as well). Their sum is well approximated by a normal with mean zero and variance $\lambda - m_i$.

Fix i and without loss of generality assume $s_i > 0$, which if the ith block has clear robust bias implies $s_i > \lambda^{1/2}$. Conditioned on $\mathcal{H}$, the copied bits in the ith block are $j_1(i)$ and $j_2(i)$ together with $m_i - 2$ locations uniformly chosen from among all size $(m_i - 2)$ subsets of $\mathbb{Z} \cap [j_1(i) + 1, j_2(i) - 1]$. On K_1, we know $j_1(i)$ and $j_2(i)$ are within $\lambda/100$ of u_{i-1} and u_i respectively, and when the ith block of $\mathbf{x}$ has clear robust bias, the number of positive bits of $\mathbf{x}$ in this range is at least $(j_2(i) - j_1(i) + 1 + \lambda^{1/2})/2$. Therefore the sum of the copied values of $2x_j - 1$ is approximately normal with mean at least $m_i \lambda^{-1/2}$ and variance m_i, independently of the sum of the non-copied bits.

Adding the copied and non copied bits gives approximately a normal with mean at least $m_i \lambda^{-1/2}$ and variance λ. We know that $m_i > \lambda p/2$ with probability tending to 1 as $\lambda \to \infty$.

Therefore, when all the blocks in $\mathcal{I}_1$ and $\mathcal{I}_2$ have clear robust bias, there is a λ_0 and a $c_1 > 0$ such that for all $\lambda > \lambda_0$, $\mathbb{P}_{\mathbf{x}}[\text{sign}(s_i) = \text{sign}(\widetilde{s}_i) > 0] \geq 1/2 + c_1$. Furthermore under $\mathbb{P}_{\mathbf{x}}$, this holds independently over all the blocks. In this case the expected sum over $i \in \mathcal{I}_1$ of $\text{sign}(s_i) \cdot \text{sign}(\widetilde{s}_i)$ is at least $2c_1|\mathcal{I}_1|$, conditional on $\mathcal{H}$, when K_1 holds. By Markov's inequality, we obtain a positive lower bound δ_0 on $\mathbb{P}_{\mathbf{x}}[T_1(k, f(k)] = 1|\mathcal{H})$ when K_1 holds. The same holds for T_2, and furthermore, by the Harris inequality [Har60, Gri99],

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}\Big[T_1(k, f(k)) = 1 \text{ and } T_2(k, f(k)) = 1|\mathcal{H}\Big] \quad &\geq \quad \mathbb{P}_{\mathbf{x}}[T_1(k, f(k)) = 1|\mathcal{H}] \\ &\quad \cdot \mathbb{P}_{\mathbf{x}}[T_2(k, f(k)) = 1|\mathcal{H}],\end{aligned}$$

because we consider increasing events in conditionally independent variables $\text{sign}(s_i) \cdot \text{sign}(\widetilde{s}_i)$. Taking $\delta = \delta_0^2$ and removing the conditioning on $\mathcal{H}$ finishes the proof. □

3.4 Choice of constants

The test T and a number of events used in its analysis depend on the parameter C_1. Throughout the remainder of Section 3, the constant C_1 remains as a free parameter. To make the arguments in the rest of the paper more transparent, we discuss in advance what relationship is needed between C_1 and the constants in Theorem 2, and what choices will be made to ensure the necessary inequalities. A constant $\widetilde{c} \ll 1$ will be small enough to be a witness for c in Lemmas 12 and 14 as well as ensuring some properties of the the random walks $\{g(j)\}$ and $\{f(j)\}$. The constant C_1 will be sufficiently large to ensure that a number of other asymptotic behaviors have kicked in. In particular, we will take $C_1 = 64\widetilde{c}^{-12}$. A constant C_{BIG} will then be chosen that is large compared to C_1. The constants in Theorem 2 will be chosen as follows.

$$\begin{aligned}C_{\text{back}} &= 5\, C_{\text{BIG}} \\ C_{\text{align}} &= \frac{1}{10}\, C_{\text{BIG}} \\ C_{\text{true}} &= \frac{2\widetilde{c}^{-1}}{C_1^{10/3}}\, C_{\text{BIG}} \\ C_{\text{false}} &= \frac{\widetilde{c}}{2C_1^{5/3}}\, C_{\text{BIG}} \\ C_{\text{avg}} &= \frac{2}{C_1^2}\, C_{\text{BIG}}\end{aligned}$$

With $C_{\text{BIG}} \gg C_1 \gg \widetilde{c}^{-1} \gg 1$, the necessary inequalities will then result from the relation between the powers of C_1 in $C_{\text{true}}, C_{\text{false}}$, and C_{avg}. We conclude this subsection by specifying C_1 and $\widetilde{c}(q)$.

Definition 17. *In the remainder of the paper let $\theta \in (0, 1/10)$ be given by (20), let $c_1 \in (0, 1/10)$ be as in Lemma 16, let $C_1 = 64\widetilde{c}^{-12}$ with $\widetilde{c}$ as we define next, and let $\widetilde{c} > 0$ be a constant depending only on q and which is sufficiently small such that following properties hold.*

(*i*) $\widetilde{c}$ *is smaller than the values of* c *in the conclusion of Lemmas 12 and 14,*

(*ii*) $\widetilde{c} \le c_1^2\theta/10$,

(*iii*) *recalling* (19)*, the probability that the robust bias of the block* $\mathbf{x}(u_{i-1}+1 : u_i)$ *is clear is at least* 4θ *when* $\lambda \ge (10\widetilde{c})^{-1}$*, and the constant* λ_0 *from Lemma 16 satisfies* $\lambda_0 < (10\widetilde{c})^{-1}$,

(*iv*) $\mathrm{Var}(g(1) - g(0)) \le (10\widetilde{c})^{-1}$,

(*v*) $\mathbb{P}_\omega[f(1) - f(0) > \ell] \le \exp(-10\widetilde{c}\ell)$ *and* $\mathbb{P}_\omega[|f(k+u) - f(k) - u| \ge a] \le \exp\left(-\dfrac{10\widetilde{c}a^2}{u}\right)$ *for any* $u \in \{1, 2 \dots\}$*, and*

(*vi*) $2/\widetilde{c}^3 - \widetilde{c}^{15}/8 > 7C_{\mathrm{sep}}$.

3.5 Estimates related to the test T

In the remainder of the section we prove various estimates related to the test T. The following lemma will be used to lower bound the probability that the test gives true positives, i.e., it will be used to lower bound $\mathbb{P}_{\mathbf{x}}[T(k, f(k)) = 1]$.

In Definitions 18 and 20 below we define two events $\mathcal{Q}(k, \widehat{k})$ and $\mathcal{H}(k)$. These are measurable with respect to $\mathbf{x}$, that is, they depend only on the input; their definitions contain probabilities with respect to the channel but do not depend on the sample ω. They depend on the parameters λ and ℓ; additionally $\mathcal{H}(k)$ depends on C_1. When we choose an interval for alignment, we will require that the right end-point k_* satisfies $\mathcal{H}(k_*)$ and $\mathcal{Q}(k_*, \widehat{k})$ for all $\widehat{k}$ in a given interval. From this we will deduce an upper bound for the probability of false positives and a lower bound for the probability of true positives. Occurrence of the event $\mathcal{Q}(k, \widehat{k})$ ensures that the ℓ-windows in $\widetilde{\mathbf{x}}$ ending between $f(\widehat{k})$ and $f(\widehat{k}+1)$ are sufficiently different from $\mathbf{x}(k_* - \ell + 1 : k_*)$ that the test $T(k, i)$ is unlikely to give a positive result when i is close to $f(\widehat{k})$ instead being close to $f(k)$, as desired; see Figure 7.

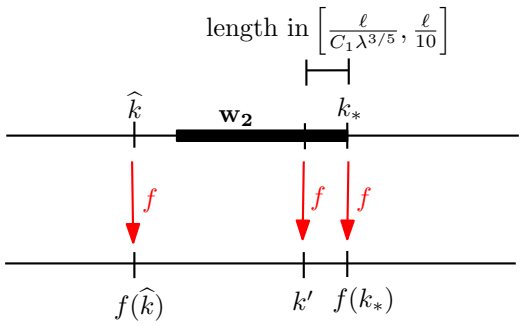

Figure 7: The right end-point k_* of the interval used in the second alignment step is (roughly speaking) chosen such that $\mathbb{P}_{\mathbf{x}}[T_1(k_*, f(\widehat{k})) = 1] \le \exp(-\widetilde{c}\ell/\lambda)$, $\mathbb{P}_{\mathbf{x}}[T_2(k_*, k') = 1] \le \exp(-\widetilde{c}\ell/(C_1\lambda^{6/5}))$, and $\mathbb{P}_{\mathbf{x}}[T(k_*, f(k_*)) = 1] \ge \exp(-\ell/(\widetilde{c}\lambda^2))$. See Definitions 18 and 20 for the precise requirements.

Definition 18. *For a fixed string $\mathbf{x} \in \mathcal{S}$, $\ell \in \mathbb{N}$, $\lambda \in \{1, \dots, \lfloor \ell^{1/2} \rfloor\}$, and $k, \widehat{k} \geq 2\ell$, define the event $\mathcal{Q}(k, \widehat{k}) = \mathcal{Q}_{\mathbf{x}}^{\lambda,\ell}(k, \widehat{k})$ to hold when*

$$\mathbb{P}_{\mathbf{x}}[E] < e^{-\widetilde{c}\ell/\lambda},$$

where

$$E := \left\{ \bigcup_{i=f(\widehat{k})}^{f(\widehat{k}+1)} \{T_1(k,i) = 1\} \cap E_{\ell;k,i} \, ; |g(i) - g(i-\ell)| < 11\ell/10 \right\}.$$

Remark. *The event $\mathcal{Q}(k, \widehat{k})$ is measurable with respect to $\mathbf{x}(k-\ell+1 : k)$ and $\mathbf{x}(\widehat{k} - \lfloor 11\ell/10 \rfloor : \widehat{k}+2)$. This ensures independence of the events $\mathcal{Q}(k_1, \widehat{k}_1)$ and $\mathcal{Q}(k_2, \widehat{k}_2)$ when k_1 and $\widehat{k}_1$ are sufficiently far from k_2 and $\widehat{k}_2$. We will use this independence property and the following lemma to argue that with high probability we can find a k_* in a given interval such that $\mathcal{Q}(k_*, \widehat{k})$ occurs for all $\widehat{k}$ in a given larger interval.*

Lemma 19. *Define the σ-field $\mathcal{G}_{k,\ell}$ by*

$$\mathcal{G}_{k,\ell} = \sigma(x_i \, : \, i \notin \{k-\ell+1, \dots, k\}) \, .$$

Then for all $\lambda > 3$ and ℓ sufficiently large,

$$\mu \left[\mathcal{Q}_{\mathbf{x}}^{\lambda,\ell}(k, \widehat{k})^c \, | \, \mathcal{G}_{k,\ell} \right] \leq \exp(-\widetilde{c}\ell/\lambda). \tag{22}$$

Proof. Observe that $\mathcal{Q}(k, \widehat{k})^c = \{\mathbb{P}_{\mathbf{x}}[E] \geq \exp(-\widetilde{c}\ell/\lambda)\}$. We will show that for ℓ sufficiently large,

$$\mathbb{P}[E \, | \, \mathcal{G}_{k,\ell}] \leq \exp(-2\widetilde{c}\ell/\lambda) \, . \tag{23}$$

This is sufficient to complete the proof, because Markov's inequality and the identity $\mathbb{E}[\mathbb{P}_{\mathbf{x}}[E] \, | \, \mathcal{G}_{k,\ell}] = \mathbb{P}[E \, | \, \mathcal{G}_{k,\ell}]$ give

$$\mu \left[\mathcal{Q}(k, \widehat{k})^c \, | \, \mathcal{G}_{k,\ell} \right] \leq \frac{\mathbb{P}[E \, | \, \mathcal{G}_{k,\ell}]}{\exp(-\widetilde{c}\ell/\lambda)} \leq \exp(-\widetilde{c}\ell/\lambda) \, .$$

Assume $\widehat{k} < k$; the case $\widehat{k} \geq k$ can be treated similarly. Let $i_1, \dots, i_{|\mathcal{I}_1|}$ be an enumeration of the elements of $\mathcal{I}_1$ in increasing order. Let $i_0 = k - \ell$. Define the filtration $\mathcal{F}_j$, $j = 0, \dots, |\mathcal{I}_1|$, by

$$\mathcal{F}_j = \sigma(\omega, \mathbf{x}(0 : i_j), \mathbf{x}(k+1 : \infty), \mathcal{I}_1) \, .$$

Observe that $\mathcal{F}_0 = \sigma(\omega, \mathcal{G}_{k,\ell})$. For fixed $i \in \mathbb{N}$ let $M = (M_j)_{j \in \{0, \dots, |\mathcal{I}_1|\}}$ be the stochastic process defined by the partial sums in (21), with $f(\widehat{k}) + i$ being the right end-point of the considered interval of the trace, i.e., $M_0 = 0$, and for $j = 1, \dots, |\mathcal{I}_1|$,

$$M_j = \sum_{j'=1}^{j} \operatorname{sign}\left(s_{i_{j'}}(k, \ell, \lambda)\right) \cdot \operatorname{sign}\left(\widetilde{s}_{i_{j'}}(f(\widehat{k}) + i, \ell, \lambda)\right) ,$$

where $s_{i_{j'}}(k,\ell,\lambda)$ and $\widetilde{s}_{i_{j'}}(f(\widehat{k})+i,\ell,\lambda)$ are defined as in (16). Observe that for each fixed i, the sequence $\{M_j \mathbf{1}_{E_{\ell;k,f(\widehat{k})+i}} : 0 \le j \le |\mathcal{I}_1|\}$ is a martingale on the filtration $\mathcal{F}_j$ with increments bounded in magnitude by 1. The martingale property follows from the fact that $E_{\ell;k,f(\widehat{k})+i} \in \sigma(\omega)$, that $\widetilde{s}_{i_j} \in \mathcal{F}_{j-1}$ on the non-overlapping event $E_{\ell;k,f(\widehat{k})+i}$, and that $\mathrm{sign}(s_{i_j})$ has expectation zero given $\mathcal{F}_{j-1}$. The Azuma-Hoeffding inequality [Hoe63, Azu67] therefore gives $\mathbb{P}[M_j \mathbf{1}_{E_{\ell;k,f(\widehat{k})+i}} \ge t] \le \exp(-t^2/(2j))$. With $t = c_1|\mathcal{I}_1|$ and $j = |\mathcal{I}_1|$ we may compute $t^2/(2j) = c_1^2|\mathcal{I}_1|/2 = c_1^2\theta\ell/(2\lambda)$, whence

$$\begin{aligned}\mathbb{P}\left[T_1(k,f(\widehat{k})+i)=1;\, E_{\ell;k,f(\widehat{k})+i} \,|\, \mathcal{G}_{k,\ell},\omega\right] &= \mathbb{P}\left[M_{|\mathcal{I}_1|} > c_1|\mathcal{I}_1|;\, E_{\ell;k,f(\widehat{k})+i} \,|\, \mathcal{G}_{k,\ell},\omega\right] \quad (24)\\ &\le \exp(-c_1^2\theta\ell/(2\lambda)) < \exp(-3\widetilde{c}\ell/\lambda),\end{aligned}$$

where the last inequality follows from part (ii) of Definition 17. Define $\mathcal{I}' = |\{f(\widehat{k}),\ldots,f(\widehat{k}+1)\}|$. The inequality (23) now follows from a union bound, using part (v) of Definition 17 and (24) in the third inequality, with the last inequality following when ℓ is sufficiently large from $\lambda \le \ell^{1/2}$.

$$\begin{aligned}\mathbb{P}[E \,|\, \mathcal{G}_{k,\ell}] &\le \mathbb{P}\left[\bigcup_{i=f(\widehat{k})}^{f(\widehat{k}+1)} \{T_1(k,i)=1\} \cap E_{\ell;k,i} \,\middle|\, \mathcal{G}_{k,\ell}\right]\\ &\le \mathbb{P}[|\mathcal{I}'| > \ell \,|\, \mathcal{G}_{k,\ell}] + \sum_{i=0}^{\ell-1} \mathbb{P}\left[T_1(k,f(k)+i)=1\,;E_{\ell;k,f(\widehat{k})+i} \,|\, \mathcal{G}_{k,\ell}\right]\\ &\le \exp(-\widetilde{c}\ell) + \ell\exp(-3\widetilde{c}\ell/\lambda) < \exp(-2\widetilde{c}\ell/\lambda)\,.\end{aligned}$$

□

Part (i) of the event $\mathcal{H}(k)$ defined just below will ensure that the probability of true positives is sufficiently large. Part (ii) of the event will be used to bound from above the probability that $|\tau_2 - f(k_*)| > c\log^{1/3} n$ for some small constant c.

Definition 20. *Let $\mathbf{x} \in \mathcal{S}$, $\ell \in \mathbb{N}$, $\lambda \in \{1,\ldots,\lfloor \ell^{1/2}\rfloor\}$, $C_1 \ge 1$, and $k \ge 2\ell$. Define the event $\mathcal{H}(k) = \mathcal{H}_{\mathbf{x}}^{\lambda,C_1,\ell}(k) \in \sigma(\mathbf{x})$ to occur if*

- *(i) $\mathbb{P}_{\mathbf{x}}[T(k,f(k))=1;\, |f(k)-f(k-\lfloor 11\ell/10\rfloor)| > \ell] > \exp(-\ell/(2\widetilde{c}\lambda^2))$, and*
- *(ii) for $\mathcal{I}' = \mathcal{I}'_1 \cup \mathcal{I}'_2$ with*

$$\begin{aligned}\mathcal{I}'_1 &:= \left\{k' \in \mathbb{N} \,:\, g(k') \in \{k-\lfloor \ell/10\rfloor,\ldots,k-\lfloor \ell/(C_1\lambda^{3/5})\rfloor\}\right\},\\ \mathcal{I}'_2 &:= \left\{k' \in \mathbb{N} \,:\, g(k') \in \left\{k+\lfloor \ell/(C_1\lambda^{3/5})\rfloor,\ldots,k+\lfloor \ell/10\rfloor\right\}\right\},\end{aligned}$$

we have

$$\mathbb{P}_{\mathbf{x}}\left[\bigcup_{k'\in\mathcal{I}'} \{T_2(k,k')=1\} \cap \{|f(k-\lfloor \ell/10\rfloor-1) - f(k-\ell-2\lfloor \ell/10\rfloor)| > \ell\}\right] < \exp(-\widetilde{c}\ell/(C_1\lambda^{6/5})).$$

Remark. *In fact $\mathcal{H}(k)$ is measurable with respect to $\mathbf{x}(k-\ell-2\lfloor \ell/10\rfloor : k+\lfloor \ell/10\rfloor)$.*

Lemma 21. *For $\ell \in \mathbb{N}$ sufficiently large, $\lambda \in \{10\lceil \widetilde{c}^{-1}\rceil, \ldots, \lfloor \ell^{1/2}\rfloor\}$, and $C_1 \geq 1$,*

$$\mu\left[\mathcal{H}_{\mathbf{x}}^{\lambda,C_1,\ell}(k)^c\right] < \exp(-\widetilde{c}\ell/(C_1\lambda)). \tag{25}$$

Proof. We verify separately that each property (i) and (ii) of Definition 20 fail on a set of μ-measure at most half the quantity on the right-hand side of (25). We claim that (i) holds whenever all blocks in $\mathcal{I}_1$ and $\mathcal{I}_2$ have clear robust bias. To see this let K_1 be the event that the trace is λ-aligned in $[k-\ell+1, k]$, and let K_2 be the event that $|f(k)-f(k-\lfloor 11\ell/10\rfloor)| > \ell$. Then

$$\mathbb{P}_{\mathbf{x}}[T(k,f(k))=1] \geq P_{\mathbf{x}}(K_1)\mathbb{P}_{\mathbf{x}}[T(k,f(k))=1|K_1].$$

Lemma 12 gives us $\mathbb{P}_{\mathbf{x}}[K_1] \geq \exp(-\ell/(10\widetilde{c}\lambda^2))$ and Lemma 16 gives us $\mathbb{P}_{\mathbf{x}}[T(k,f(k)) = 1|K_1] \geq \delta$ when all the blocks in $\mathcal{I}_1$ and $\mathcal{I}_2$ have clear robust bias. Thus, when all the blocks have clear robust bias,

$$\mathbb{P}_{\mathbf{x}}[T(k,f(k))=1; K_2] \geq \delta \exp\left(-\frac{\ell}{10\widetilde{c}\lambda^2}\right) - \mathbb{P}_{\mathbf{x}}[K_2^c]\,.$$

By part (v) of Definition 17 and the hypothesis $\lambda \geq 10\widetilde{c}^{-1}$ gives

$$\begin{aligned}
\mathbb{P}\left[|f(k)-f(k-\lfloor 11\ell/10\rfloor)| \leq \ell\right] &\leq \mathbb{P}\left[|f(k)-f(k-\lfloor 11\ell/10\rfloor)-(11\ell/10)| \geq \ell/10\right] \\
&\leq \exp\left(-\frac{10\widetilde{c}(\ell/10)^2}{12\ell/10}\right) \\
&= \exp\left(-\frac{\widetilde{c}\ell}{12}\right) \\
&\leq \frac{1}{2}\exp\left(-\frac{\ell}{\widetilde{c}\lambda^2}\right),
\end{aligned}$$

proving the claim.

To see that all blocks in $\mathcal{I}_1$ and $\mathcal{I}_2$ have clear robust bias except on a set of μ-measure at most half the right side of (25), we require the following large deviation bound for binomial variables:

$$\mathbb{P}[\mathrm{Bin}\,(n,4\theta) \leq 2\theta n] \leq \exp(-2(1-\ln 2)n\theta) < \exp(-n\theta/2)\,. \tag{26}$$

To derive this, begin with the well-known Kullback-Leibler bound for $Z \sim \mathrm{Bin}\,(n,p)$ and $r \leq p$,

$$\frac{1}{n}\log \mathbb{P}[Z \leq rn] \leq r\log\frac{p}{r} + (1-r)\log\frac{1-p}{1-r}\,,$$

which may be obtained, for instance, by applying Markov's inequality with $\mathbb{E}e^{\lambda Z} = (1-p+pe^{\lambda})^n$ and $\lambda = \log(r/p) - \log((1-r)/(1-p))$. Plugging in $p = 4\theta$ and $r = 2\theta$ yields, for $\theta < 1/4$,

$$\frac{1}{n}\log \mathbb{P}[\mathrm{Bin}\,(n,4\theta) \leq 2\theta n] \leq h(\theta) := \log\frac{1-4\theta}{1-2\theta} + 2\theta\log\frac{2-4\theta}{1-4\theta}\,.$$

Observing that $h'(0) = -2(1-\ln 2) < -1/2$ and $h'' < 0$ on $(0, 1/4)$ then establishes (26).

Each block has a clear robust bias with probability at least 4θ by part (iii) of Definition 17 and the requirement $\lambda > 10\widetilde{c}^{-1}$. Furthermore, the event that this holds is independent for any pair of blocks which are not adjacent. Applying (26) to the $\ell/(2\lambda C_1)$ even numbered blocks from among which $\mathcal{I}_2$ was chosen shows that at least $\theta\ell/(\lambda C_1)$ of these, hence all blocks in $\mathcal{I}_2$, have clear robust bias except on an event of probability at most $\exp(-\theta\ell/(4\lambda C_1)) < \exp(-2\widetilde{c}\ell/(C_1\lambda))$. For $\mathcal{I}_1$ the same bound holds without the factor of C_1 in the denominator. These negative exponents are both at least twice the negative exponent on the right side of (25), therefore sum to at most half the right side of (25) once ℓ is sufficiently large.

Now we consider (ii) of Definition 20. Let k_1' (resp. k_2') be the largest (resp. smallest) element of $\mathcal{I}_1'$ (resp. $\mathcal{I}_2'$), and define $d = \lfloor \ell/(2C_1\lambda^{3/5})\rfloor$. Recalling Definition 13 and Lemma 14, we have for ℓ sufficiently large,

$$\begin{aligned}\mathbb{P}[E^c_{\lfloor \ell/C_1\rfloor,k,k_1'}] &\le \mathbb{P}\left[E^c_{\lfloor \ell/C_1\rfloor,k,f(k)-d}\right] + \mathbb{P}[|f(k)-k_1'|<d] \\ &\le \widetilde{c}^{-1}\exp(-10\widetilde{c}d^2C_1/\ell) + \exp(-\widetilde{c}d) \\ &< \widetilde{c}^{-1}\exp(-2\widetilde{c}\ell/(C_1\lambda^{6/5}))\,.\end{aligned}$$

Similarly, $\mathbb{P}[E^c_{\lfloor \ell/C_1\rfloor,k,k_2'}] \le \widetilde{c}^{-1}\exp(-2\widetilde{c}\ell/(C_1\lambda^{6/5}))$. We also have $\mathbb{P}[|\mathcal{I}'|\ge\ell] < \exp(-\widetilde{c}\ell)$ and $\mathbb{P}[f(k)\ge k_2'] = \mathbb{P}[f(k) = f(k+d)] \le \exp(-10\widetilde{c}d)$, where the last inequality follows from part (v) of Definition 17. Define the event $E \in \sigma(\omega)$ by

$$E = E_{\lfloor \ell/C_1\rfloor,k,k_1'} \cap E_{\lfloor \ell/C_1\rfloor,k,k_2'} \cap \left\{\left|\mathcal{I}'\right| < \ell\right\} \cap \{f(k) < k_2'\},$$

and observe that

$$\mathbb{P}[E^c] = \mathbb{P}_{\mathbf{x}}[E^c] < 3\widetilde{c}^{-1}\exp(-2\widetilde{c}\ell/(C_1\lambda^{6/5})). \tag{27}$$

By a union bound,

$$\begin{aligned}&\mathbb{P}_{\mathbf{x}}\left[\bigcup_{k'\in\mathcal{I}'}\{T_2(k,k')=1\}\cap\{|f(k-\lfloor\ell/10\rfloor)-f(k-\ell-2\lfloor\ell/10\rfloor)|>\ell\}\right] \\ &\qquad\le \mathbb{P}_{\mathbf{x}}\left[\bigcup_{k'\in\mathcal{I}'}\{T_2(k,k')=1\}\right] \\ &\qquad\le \mathbb{P}_{\mathbf{x}}[E^c] + \sum_{j=0}^{\ell-1}\mathbb{P}_{\mathbf{x}}\left[T_2(k,k_2'+j)=1;\,E\right] + \mathbb{P}_{\mathbf{x}}\left[T_2(k,k_1'-j)=1;\,E\right].\end{aligned} \tag{28}$$

We bound the first term on the right side of (28) by (27). To bound the other terms on the right side of (28), observe that the event $E_{\lfloor \ell/C_1\rfloor,k,k_2'} \cap \{f(k) < k_2'\}$ can occur only via (i), not (ii), in Definition 13. Thus $E_{\lfloor \ell/C_1\rfloor,k,k_2'}$ implies $E_{\lfloor \ell/C_1\rfloor,k,k_2'+j}$ for $j \in \{0,\dots,\ell-1\}$. This implies that no bit of the original string was copied to the respective block of the trace. Repeating the martingale argument in the proof of Lemma 19, Azuma's inequality gives

$$\begin{aligned}\mathbb{P}\left[T_2(k,k_2'+j)=1;\,E\right] &\le \exp\left(-(c_1|\mathcal{I}_2|)^2/(2|\mathcal{I}_2|)\right) \\ &\le \exp\left(-c_1^2\ell\theta/(2.1\lambda C_1)\right) \le \exp(-4\widetilde{c}\ell/(C_1\lambda)).\end{aligned} \tag{29}$$

Markov's inequality gives further that except on a set of μ-measure $\exp(-2\widetilde{c}\ell/(C_1\lambda))$ we have

$$\mathbb{P}_{\mathbf{x}}\left[T_2(k,k_2'+j)=1;\, E\right] \le \exp(-2\widetilde{c}\ell/(C_1\lambda)).$$

A similar result holds for the terms on the form $\mathbb{P}_{\mathbf{x}}[T_2(k,k_1'-j)=1;\, E]$ on the right side of (28). Therefore, except on an event of μ-measure $2\ell\exp(-2\widetilde{c}\ell/(C_1\lambda)) < \exp(-\widetilde{c}\ell/(C_1\lambda))$ for ℓ sufficiently large, we have

$$\mathbb{P}_{\mathbf{x}}\left[\bigcup_{k'\in\mathcal{I}'}\{T_2(k,k')=1\}\cap\{|f(k-\lfloor\ell/10\rfloor)-f(k-\ell-2\lfloor\ell/10\rfloor)|>\ell\}\right]$$
$$\le 3\widetilde{c}^{-1}\exp(-2\widetilde{c}\ell/(C_1\lambda^{6/5}))+2\ell\exp(-2\widetilde{c}\ell/(C_1\lambda)) \le \exp(-\widetilde{c}\ell/(C_1\lambda^{6/5})).$$

□

4 Existence of good positions

Several more technical lemmas and a somewhat intricate definition are needed to finish proving Theorem 2. To motivate these, we first describe the rest of the proof. The determination of (k_*,τ_2) begins with construction of the rough approximation, τ_1.

Set $\ell=\Theta(\log^{5/3}n)$, $\lambda=\lfloor\ell^{2/5}\rfloor=\Theta(\log^{2/3}n)$, and the constant C_1 to the value chosen in Section 3.4. Slide an ℓ-window from left to right in the trace until the test $T(k-9C_{\mathrm{BIG}}\log n,k')$ produces a value of 1, and let τ_1 be the value of k' at which this first occurs. Define τ_1 to be ∞ if this fails to occur for all $k'\le 2n$.

What we need from τ_1 is that the true positive rate for aligning within $9C_{\mathrm{BIG}}\log n$ is at least $\exp(-c\log^{1/3}n)$ and that the false positive rate is at most a similar but smaller function of the same form, $\exp(-C\log^{1/3}n)$ for $C\gg c$. However, and this is crucial, we need this to hold not only in the space $\mathbb{P}_\mu$ but in the space $\mathbb{P}_{\mathbf{x}}$ for all strings $\mathbf{x}$ other than those in a "bad" set which must have measure $o(1/n)$ and therefore be much smaller than $\exp(-\Theta(\log^{1/3}n))$.

Conditioning on a positive alignment $\tau_1\approx f(k-9C_{\mathrm{BIG}}\log n)$, we now retest to find an alignment of a carefully chosen position k_*, accurate to within $\Theta(\log^{1/3}n)$. Set $\ell=\Theta(\log^{1/3}n)$ and $\lambda=\Theta(1)$, with the same value of C_1 as before. With ℓ this small, we can no longer expect most strings $\mathbf{x}$ to be free of bad spots where the proportion of traces producing false positive alignments is intolerably high. For this reason, we will find a $k_*\in[k-5C_{\mathrm{BIG}}\log n,k-4C_{\mathrm{BIG}}\log n]$ such that we can run the test $T(k_*,k')$ over a window of length $\Theta(\log n)$ sitting between τ_1 and $\tau_1+9C_{\mathrm{BIG}}\log n$.

The argument used to construct (k_*,τ_2) is more elaborate but similar to the argument used to construct τ_1. For this reason, we use the same lemmas in both constructions. Some elements of the argument appear unnecessarily complicated in the case of τ_1, but it still saves on space and ideas not to duplicate the sequence of lemmas. With this in mind, we outline the sequence of lemmas.

The key definition is that of the "good" position, $k_3 = k_*$. This definition takes as input the test parameter ℓ, λ and C_1, as well as an interval $\mathcal{I}$ of values of k' over which the test $T(k_3, k')$ is performed. The position k_3 is deemed good if the quenched probabilities $\mathbb{P}_{\mathbf{x}}$ for true and false positives and the quenched expectation $\mathbb{E}_{\mathbf{x}}$ of the truncated discrepancy satisfy inequalities that will be used to prove $(iii) - (v)$ of Theorem 2. The corresponding key lemma, Lemma 23 below, gives a lower bound on the probability of finding a good k_3 such that the ℓ-window $[k_2, k_3]$ lies within a specified interval $[k_1, k_4]$. The bound will improve as the ratio b of the length of $[k_1, k_4]$ to the length of $[k_2, k_3]$ grows. In the case of τ_1 it is applied in the somewhat degenerate situation that $b = 1$ to prove that with high probability we may take $k_3 = k_4$.

One further complication is that the restrictions on where to search take place in the trace, not the message. We would like to restrict the search to positions $f(j)$ corresponding to j in some interval $[k_0, k_5]$. Values of f are not known to the algorithm, therefore $\mathcal{I}$ is never guaranteed to be a subset of $[f(k_0), f(k_5)]$ and the probability estimates must include a fudge term accounting for failure of this inclusion. The lemma is proved for all choices of $\mathcal{I}$ satisfying some desired inclusion property with sufficiently high probability but in fact only two choices are required, one when constructing τ_1 and one when constructing τ_2.

Definition 22 (good alignment position)**.** *Fix a string* $\mathbf{x}$ *and positive integers* $k_0 \le k_2 < k_3 \le k_5$*. Let* $\ell := k_3 - k_2 + 1$ *and* $L := k_5 - k_0 + 1$*, and assume* $k_2 - k_0 > \ell$*. Fix a constant* $C \ge 1$ *and a positive integer* $\lambda \le \ell^{1/2}$*. For every (possibly random) set* $\mathcal{I} \subseteq \mathbb{N}$ *we define an event* $A_{\mathcal{I}}$ *by*

$$A_{\mathcal{I}} := \Big\{ f(k_3) \in \mathcal{I} \subset \{f(k_0), f(k_0)+1, \dots, f(k_5)\} \Big\} \cap \{ f(k_3) - \ell > f(k_0) + \lceil L/9 \rceil \}, \quad (30)$$

and a random variable $\tau_{\mathcal{I}}$ *by*

$$\tau_{\mathcal{I}} = \inf\{ k' \in \mathcal{I} \,:\, T(k_3, k') = 1 \}, \quad (31)$$

where $T = T^{\lambda, C, \ell, c_1}_{\mathbf{x}, \widetilde{\mathbf{x}}}$*. The infimum of the empty set is considered to be* $+\infty$*. Define the event* $\mathbf{GOOD}(k_3) = \mathbf{GOOD}(k_0, k_2, k_3, k_5, \lambda, C, \mathbf{x})$ *to hold if the following three properties* $(i) - (iii)$ *are satisfied for all (possibly random)* $\mathcal{I}$*, for all positive integers* $a \le \ell$*, and for all events* $A^{(1)}_{\mathcal{I}}$ *and* $A^{(2)}_{\mathcal{I}}$ *which satisfy*

$$A^{(1)}_{\mathcal{I}} \cap A^{(2)}_{\mathcal{I}} \subset A_{\mathcal{I}}, \qquad \mathbb{P}_{\mathbf{x}}\big[\big(A^{(2)}_{\mathcal{I}}\big)^c \big] \le \exp(-\ell^2),$$
$$A^{(1)}_{\mathcal{I}} \in \mathcal{G}^{k_5}_{f(k_0)+\lceil L/9 \rceil}, \qquad \mathbb{P}_{\mathbf{x}}\big[A^{(1)}_{\mathcal{I}} \big] > \exp\big(-\ell/(2\widetilde{c}\lambda^2) \big).$$

(*i*) $\mathbb{P}_{\mathbf{x}}[|k_3 - g(\tau_{\mathcal{I}})| > a; \tau_{\mathcal{I}} < \infty; A_{\mathcal{I}}] < 2L \exp(-\widetilde{c}\ell/\lambda) + 3\widetilde{c}^{-1} \exp(-\widetilde{c}a^2/\ell)$,

(*ii*) $\mathbb{P}_{\mathbf{x}}[\tau_{\mathcal{I}} < \infty; A_{\mathcal{I}}] > \frac{1}{2} \exp(-\ell/(\widetilde{c}\lambda^2))$, *and*

(*iii*) $\mathbb{E}_{\mathbf{x}}[|k_3 - g(\tau_{\mathcal{I}})| \mathbf{1}_{A_{\mathcal{I}}} \mathbf{1}_{|k_3 - g(\tau_{\mathcal{I}})| \le \ell/10} \,|\, \tau_{\mathcal{I}} < \infty] \le \dfrac{3\ell}{2C\lambda^{3/5}}$.

Remarks.

1. The random variable $\tau_{\mathcal{I}}$ *is a stopping time with respect to the filtration* $\{\mathcal{G}^{k_5}_j\}$.

2. *The event* $\mathbf{GOOD}(k_3)$ *is measurable with respect to* $\mathbf{x}(0:k_5)$*, since* $\tau_{\mathcal{I}}$ *is bounded above by the greatest element of* $\mathcal{I}$ *on the event that* $\tau_{\mathcal{I}} < \infty$*, which implies that on the event* $A_{\mathcal{I}}$*, the probabilities* $\mathbb{P}_{\mathbf{x}}$ *and expectation* $\mathbb{E}_{\mathbf{x}}$ *in* $(i)-(iii)$ *depend on* $\mathbf{x}$ *only via* $\mathbf{x}(0:k_5)$*.*

3. *The events* $A_{\mathcal{I}}^{(1)}$ *and* $A_{\mathcal{I}}^{(2)}$ *will be used in the second alignment step. If the event* $A_{\mathcal{I}}^{(1)}$ *occurs then the first alignment was successful. We will need to bound from below the probability of a successful second alignment, given that the first alignment step was successful. When we do this it will be useful to know that* $A_{\mathcal{I}}^{(1)}$ *depends mainly on the deletions and insertions far away from the alignment position* k_3 *in the second alignment step. The event* $A_{\mathcal{I}}^{(2)}$ *has very high probability, so* $\mathbb{P}_{\mathbf{x}}\big[\big(A_{\mathcal{I}}^{(2)}\big)^c\big]$ *is negligible, and we therefore allow this event to depend on the deletions and insertions close to* k_3*.*

Lemma 23. *Let*

$$1 \le \ell \le L, \qquad 1 \le b \le L/\ell, \qquad 2\ell \le k_0 \le k_1 < k_4 \le k_5,$$
$$|k_4 - k_1| \ge b\ell - 1, \qquad |k_5 - k_0| = L - 1$$

be constants with values in $\mathbb{N}$*. Assume* ℓ *is sufficiently large, and that* $C \ge 1$ *and* $\lambda \in \{1, \dots, \lfloor \ell^{1/2} \rfloor\}$ *satisfy*

$$2\frac{1}{\widetilde{c}\lambda^2} < \frac{\widetilde{c}}{C\lambda^{6/5}}\,. \tag{32}$$

Then the μ*-measure of* $\mathbf{x}$ *such that* $\mathbf{GOOD}(k_0, k_2, k_3, k_5, \lambda, C, \mathbf{x})$ *holds for some* $k_1 \le k_2 < k_2 + \ell - 1 = k_3 \le k_4$ *is at least least* $1 - L^b \exp(-\widetilde{c}b\ell/(10C\lambda))$*.*

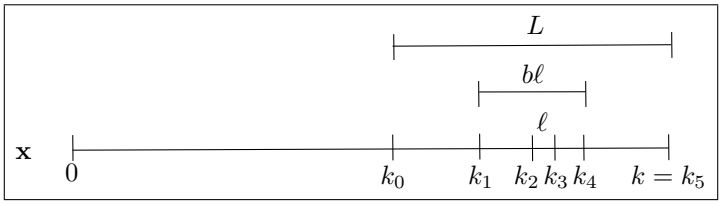

Figure 8: Illustration of indices and intervals defined in Definition 22 and Lemma 23.

To prove this we identify an event $R(k)$, which is an intersection of events of the form $\mathcal{Q}(k, k')$ and $\mathcal{H}(k)$, that implies $\mathbf{GOOD}(k)$ (Lemma 24 below), and that can be shown to happen with high probability (Lemma 25 below). For $k \in \{k_0, \dots, k_5\}$, define

$$R(k) = R_{\mathbf{x}}^{\lambda,C,\ell,k_0,k_5}(k) := \mathcal{H}_{\mathbf{x}}^{\lambda,C,\ell}(k) \cap \left(\bigcap_{\widehat{k}=k_0}^{k_5} \mathcal{Q}_{\mathbf{x}}^{\lambda,\ell}(k,\widehat{k}) \right). \tag{33}$$

Lemma 24. *Let* $\ell, L, b, k_0, k_1, k_4, k_5, C$ *and* λ *be as in Lemma 23. For* $k_3 \in [k_1, k_4]$*, the event* $R(k_3)$ *implies* $\mathbf{GOOD}(k_3)$*.*

Proof. We verify the properties (i)-(iii) of Definition 22 separately.

(i) By a union bound,

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}\left[|k_3 - g(\tau_{\mathcal{I}})| > a; \tau_{\mathcal{I}} < \infty; A_{\mathcal{I}}\right] \;\leq\; & \mathbb{P}_{\mathbf{x}}\left[\bigcup_{i=f(k_0)}^{f(k_3)-\lceil a/2\rceil} \{T(k_3,i)=1\} \cap E_{\ell;k_3,i}\right] \\ & + \mathbb{P}_{\mathbf{x}}\left[\bigcup_{i=f(k_3)+\lceil a/2\rceil}^{f(k_5)} \{T(k_3,i)=1\} \cap E_{\ell;k_3,i}\right] \\ & + \mathbb{P}_{\mathbf{x}}\left[E^c_{\ell;k_3,f(k_3)-\lceil a/2\rceil}\right] \\ & + \mathbb{P}_{\mathbf{x}}\left[E^c_{\ell;k_3,f(k_3)+\lceil a/2\rceil}\right] \\ & + \mathbb{P}_{\mathbf{x}}\left[g(f(k_3)-\lceil a/2\rceil) < k_3 - a\right] \\ & + \mathbb{P}_{\mathbf{x}}\left[g(f(k_3)+\lceil a/2\rceil) > k_3 + a\right] \\ & + \mathbb{P}_{\mathbf{x}}\left[|g(f(k_3)) - k_3| > a\right] .\end{aligned}$$

Using the definition of $R(k)$ and Lemma 14 we see that this is at most

$$2L\exp\left(-\frac{\widetilde{c}\ell}{\lambda}\right) + 2\widetilde{c}^{-1}\exp\left(-\frac{\widetilde{c}a^2}{\ell}\right) + 3\exp(-\widetilde{c}a)\,,$$

which is at most $2L\exp(-\widetilde{c}\ell/\lambda) + 3\widetilde{c}^{-1}\exp(-\widetilde{c}a^2/\ell)$.

(ii) By definition, $R(k_3)$ implies $\mathcal{H}(k_3)$, which implies that $T(k_3, f(k_3)) = 1$ with probability at least $\exp(-\ell/(2\widetilde{c}\lambda^2))$ by part (i) of Definition 20. Also observe that the lower bound on the latter probability holds (up to multiplication by $(1-o_n(1))$) even if we condition on $A_{\mathcal{I}}$, due to the requirement that $A^{(1)}_{\mathcal{I}} \in \mathcal{G}^{k_5}_{f(k_0)+\lceil L/9\rceil}$, $A_{\mathcal{I}} \subset \{f(k_0)+\lceil L/9\rceil < f(k_3) - \ell\}$ that $A^{(2)}_{\mathcal{I}}$ is very unlikely, and the last condition in the definition of $A_{\mathcal{I}}$. Therefore

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}[\{\tau_{\mathcal{I}} < \infty\} \cap A_{\mathcal{I}}] &= \mathbb{P}_{\mathbf{x}}[\tau_{\mathcal{I}} < \infty \,|\, A_{\mathcal{I}}] \cdot \mathbb{P}_{\mathbf{x}}[A_{\mathcal{I}}] \\ &\geq \exp(-\ell/(2\widetilde{c}\lambda^2))\exp(-\ell/(2\widetilde{c}\lambda^2))(1-o_n(1)).\end{aligned}$$

(iii) Let $Z := |g(\tau_{\mathcal{I}}) - k_3|$. Then,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}\left[Z\mathbf{1}_{A_{\mathcal{I}}}\mathbf{1}_{Z\leq \ell/10} \,|\, \tau_{\mathcal{I}} < \infty\right] \;<\; & \frac{\mathbb{E}_{\mathbf{x}}\left[Z\mathbf{1}_{A_{\mathcal{I}}}\mathbf{1}_{Z<\ell/(C\lambda^{3/5})}\right]}{\mathbb{P}_{\mathbf{x}}[\tau_{\mathcal{I}} < \infty]} \\ & + \frac{\mathbb{E}_{\mathbf{x}}\left[Z\mathbf{1}_{A_{\mathcal{I}}}\mathbf{1}_{\ell/(C\lambda^{3/5})\leq Z\leq \ell/10}\right]}{\mathbb{P}_{\mathbf{x}}[\tau_{\mathcal{I}} < \infty]} \\ \leq\; & \frac{\ell}{C\lambda^{3/5}} + \frac{\ell\mathbb{P}_{\mathbf{x}}[\ell/(C\lambda^{3/5}) \leq Z \leq \ell/10; A_{\mathcal{I}}]}{\mathbb{P}_{\mathbf{x}}[\tau_{\mathcal{I}} < \infty; A_{\mathcal{I}}]} \\ \leq\; & \frac{\ell}{C\lambda^{3/5}} + \ell\frac{\exp\left(-\frac{\widetilde{c}\ell}{C\lambda^{6/5}}\right) + \exp(-\widetilde{c}\ell)}{\frac{1}{2}\exp\left(-\frac{\ell}{\widetilde{c}\lambda^2}\right)},\end{aligned}$$

where the last inequality uses part (ii) of Definition 20 (the definition of $\mathcal{H}$) for the numerator of the second term and part (ii) of Definition 22 (that we just proved) for the denominator. By (32), this last quantity is at most the following for all large values of n

$$\frac{3\ell}{2C\lambda^{3/5}}\,.$$

□

Lemma 25. *Let $\ell, L, b, k_0, k_1, k_4, k_5, C$ and λ be as in Lemma 23. Then for ℓ sufficiently large,*

$$\mu\left[\bigcup_{k=k_1+\ell-1}^{k_4} R(k)\right] \geq 1 - L^b \exp\left(-\frac{\widetilde{c}\ell b}{10C\lambda}\right)\,.$$

Proof. Define

$$\mathcal{J} := \left\{k_1+\ell-1, k_1+3\ell-1, \ldots, k_1+2\left\lfloor\frac{b+1}{2}\right\rfloor\ell-\ell-1\right\} \subset [k_1+\ell-1, k_4]\,.$$

To conclude, it is sufficient to show that

$$\mu\left[\bigcup_{k\in\mathcal{J}} R(k)\right] \geq 1 - L^b \exp\left(-\frac{\widetilde{c}\ell b}{10C\lambda}\right)\,. \tag{34}$$

If the event on the left side of (34) does not occur, then for each $k \in \mathcal{J}$ at least one of the events $\mathcal{H}(k)$ and $\mathcal{Q}(k,k')$ for some $k' \in \{k_0, \ldots, k_5\}$ does not occur. Therefore we can write $\mathcal{J}$ as the union of two disjoint sets $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$, and we can find a function $h : \mathcal{J} \to \{k_0, \ldots, k_5\}$, such that the following event $\check{E}$ occurs

$$\check{E} := \left(\bigcap_{k\in\mathcal{J}_1} \check{R}(k)^c\right) \cap \left(\bigcap_{k\in\mathcal{J}_2} \mathcal{Q}(k,h(k))^c\right)\,. \tag{35}$$

There are $2^{|\mathcal{J}|}$ ways to choose the sets $\mathcal{J}_1$ and $\mathcal{J}_2$, and, given $\mathcal{J}_1$ and $\mathcal{J}_2$, there are at most $L^{|\mathcal{J}|}$ ways to define the function h. Since $|\mathcal{J}| = \lfloor (b+1)/2 \rfloor$ and $2^{|\mathcal{J}|} \cdot L^{|\mathcal{J}|} \leq L^b$ for $L \geq 2$, in order prove (34), it is sufficient to show that for any fixed choice of $\mathcal{J}_1$, $\mathcal{J}_2$, and h we have

$$\mu\left[\check{E}\right] \leq \exp\left(-\frac{\widetilde{c}\ell b}{10C\lambda}\right)\,. \tag{36}$$

Define $M_0 = \lceil b/10 \rceil$, and fix $\mathcal{J}_1$, $\mathcal{J}_2$, and h as above. For $m \in \mathbb{N}$ define $\mathcal{N}(m) := \{k - \ell - 2\lfloor \ell/10 \rfloor : k + \lfloor \ell/10 \rfloor\}$. We will first argue that for $i = 1, \ldots, M_0$ we can define $m_i \in \mathcal{J}$ iteratively, such that

$$m_i \notin \bigcup_{j\in\{1,\ldots,i-1\}} \mathcal{N}(m_j) \cup \mathcal{N}(h(m_j))\,. \tag{37}$$

Observe that each interval $\mathcal{N}(m_j)$ intersects one interval $\mathcal{N}(m)$ for $m \in \mathcal{J}$, and that each interval $\mathcal{N}(h(m_j))$ intersects at most two intervals $\mathcal{N}(m)$ for $m \in \mathcal{J}$. Therefore the set on the right side of (37) intersects at most $3(i-1)$ of the intervals $\mathcal{N}(m)$ for $m \in \mathcal{J}$. Because $3(i-1) \le 3(M_0-1) < \lfloor \frac{b+1}{2} \rfloor = |\mathcal{J}|$, the pigeonhole principle gives the existence of m_i satisfying (37).

Define the filtration $\widehat{\mathcal{G}}_i$, $i = 0, \dots, M_0$, by

$$\widehat{\mathcal{G}}_i := \sigma\Big(x(j) \,:\, j \in \mathcal{N}(k) \cup \mathcal{N}(h(k)),\ k \in \{m_1, \dots, m_i\} \Big) .$$

For $i = 1, \dots, M_0$ let $\check{E}_i$ be the event $\check{E}$, except that we only consider the indices $m_1, \dots, m_i$, that is,

$$\check{E}_i = \left(\bigcap_{k \in \mathcal{J}_1 \cap \{m_1, \dots, m_i\}} \mathcal{H}(k)^c \right) \cap \left(\bigcap_{k \in \mathcal{J}_2 \cap \{m_1, \dots, m_i\}} \mathcal{Q}(k, h(k))^c \right) .$$

By the remarks following Definitions 18 and 20, we see that $\check{E}_i \in \widehat{\mathcal{G}}_i$ for all i. For $i \in \mathcal{J}_1$, the event $\mathcal{H}(m_i)$ is independent of $\widehat{\mathcal{G}}_{i-1}$, so by Lemma 21,

$$\mu[\check{E}_i \,|\, \widehat{\mathcal{G}}_{i-1}] = \mu[\check{E}_{i-1} \cap \mathcal{H}(m_i)^c \,|\, \widehat{\mathcal{G}}_{i-1}] = \mu[\mathcal{H}(m_i)^c] \mathbf{1}_{\check{E}_{i-1}} < \exp\left(-\frac{\widetilde{c}\ell}{C_1 \lambda} \right) .$$

Similarly, when $i \in \mathcal{J}_2$ and letting $\mathcal{G}_{m_i,\ell}$ be as in Lemma 19, an application of Lemma 19 and the observation $\widehat{\mathcal{G}}_{i-1} \subset \mathcal{G}_{m_i,\ell}$ give

$$\begin{aligned} \mu[\check{E}_i \,|\, \widehat{\mathcal{G}}_{i-1}] &= \mu[\check{E}_{i-1} \cap \mathcal{Q}(m_i, h(m_i))^c \,|\, \widehat{\mathcal{G}}_{i-1}] \\ &= \mu[\mu[\mathcal{Q}(m_i, h(m_i))^c \,|\, \mathcal{G}_{m_i,\ell}] \,|\, \widehat{\mathcal{G}}_{i-1}] \mathbf{1}_{\check{E}_{i-1}} \\ &\le \exp\left(-\frac{\widetilde{c}\ell}{\lambda} \right) \\ &< \exp\left(-\frac{\widetilde{c}\ell}{C_1 \lambda} \right) . \end{aligned}$$

Using the above and $\check{E}_j \in \widehat{\mathcal{G}}_{i-1}$ for $j < i$, we get (36) via

$$\begin{aligned} \mu[\check{E}] \le \mu[\check{E}_{M_0}] = \mu\left[\bigcap_{i=1}^{M_0} \check{E}_i \right] &= \prod_{i=1}^{M_0} \mu\left[\check{E}_i \,\Big|\, \bigcap_{j=1}^{i-1} \check{E}_j \right] \\ &\le \exp(-\widetilde{c}\ell M_0/(C_1\lambda)) \le \exp(-\widetilde{c}\ell b/(10 C_1 \lambda)). \end{aligned}$$

□

5 Two stages of alignment and the proof of Theorem 2

Definition 26 (the rough alignment τ_1)**.** *Let $k \in \mathbb{N}$ and $\mathbf{x}(0:k)$ be given, and assume $k \ge \lceil 9 C_{\mathrm{BIG}} \log n \rceil$. Set $C = 1$, $C_0 := 600 \widetilde{c}^{-1}$, $\ell := \lceil (C_0 \log n)^{5/3} \rceil$, and $\lambda := \lfloor \ell^{2/5} \rfloor$. Let*

$\rho := k - \lceil 9C_{\rm BIG}\log n\rceil$. *Recalling Definition 17 and that* $T = T^{\lambda,C,\ell,c_1}$, *define the rough alignment for* ρ *in the trace by*

$$\tau_1 := \inf\{k' \in [2\ell, 1.5n]\cap\mathbb{Z} \,:\, T(\rho,k') = 1\}\,.$$

Lemma 27. *There is a set* $\Xi_{\rm bad}(1,k)\subset\mathcal{S}$ *of* μ*-measure at most* n^{-3} *such that the following two inequalities hold for* $\mathbf{x}\notin\Xi_{\rm bad}(1,k)$ *in the setting of Definition 26, with* n *sufficiently large and* $10\lceil C_{\rm BIG}\log n\rceil \le k\le n$.

$$(i)\ \ \mathbb{P}_{\mathbf{x}}\left[|\rho - g(\tau_1)| > C_{\rm BIG}\log n;\tau_1<\infty\right] < \exp\left(-\frac{\widetilde{c}C_{\rm BIG}^2}{3C_0^{5/3}}\log^{1/3} n\right);$$

$$(ii)\ \ \mathbb{P}_{\mathbf{x}}[\tau_1<\infty] > \exp\left(-\frac{2C_0^{1/3}}{\widetilde{c}}\log^{1/3} n\right);$$

$$(iii)\ \ \mathbb{P}_{\mathbf{x}}[\{f(k_0-\lceil C_{\rm BIG}\log^{1/3} n\rceil)>\tau_1\}\cup\{f(k_0+\lceil C_{\rm BIG}\log^{1/3} n\rceil)<\tau_1\};\tau_1<\infty]$$
$$< \exp\left(-\frac{\widetilde{c}C_{\rm BIG}^2}{4C_0^{5/3}}\log^{1/3} n\right).$$

Proof. First assume $k > 3C_0^{5/3}\log^{5/3} n + 9C_{\rm BIG}\log n$. Use Lemma 23 with the values of ℓ,λ and C in Definition 26, $L=\ell$, $b=1$, $[k_1,k_4]=[\rho-\ell+1,\rho]$ and $[k_0,k_5]=[\ell,2n]$. Take the set $\mathcal{I}$ to be $[2\ell,1.5n]\cap\mathbb{Z}$, so that $\tau_1=\tau_{\mathcal{I}}$. The last sentence of Lemma 23 in this case requires that $[k_2,k_3]=[k_1,k_4]$. The conclusion of the lemma is that **GOOD** holds for $k_3=\rho$ for all $\mathbf{x}$ except in a set $\Xi_{\rm bad}(k,1)$ of μ-measure at most $1-2n\exp(-\widetilde{c}\ell/(10\lambda))$. Observing that

$$\frac{\widetilde{c}\ell}{10\lambda} \ge \frac{\widetilde{c}}{10}C_0\log n = 60\log n,$$

we see that for sufficiently large n,

$$\mu(\Xi_{\rm bad}(k,1)) < n^{-3}\,. \tag{38}$$

When $\mathbf{x}\notin\Xi_{\rm bad}(k,1)$, we deduce from the definition of $A_{\mathcal{I}}$ and from clause (i) of the definition of $\mathbf{GOOD}(\rho)$ with $a := C_{\rm BIG}\log n$, that

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}\left[|\rho-g(\tau_1)|>C_{\rm BIG}\log n;\tau_1<\infty\right] \ &< \ 4n\exp\left(-\frac{\widetilde{c}\ell}{\lambda}\right) + 3\widetilde{c}^{-1}\exp\left(-\frac{\widetilde{c}C_{\rm BIG}^2\log^2 n}{\ell}\right)\\ &\quad +\mathbb{P}_\omega[f(2\ell)<\ell] \ + \ \mathbb{P}_\omega[f(\lfloor 1.5n\rfloor)>2n]\,.\end{aligned}$$

The first term is bounded above by $4n\exp(-600\log n) = \exp(-\Theta(\log n))$, the third term is bounded above by $\exp(-\Theta(\log^{5/3} n))$ and last term is bounded above by $\exp(-\Theta(n))$. All of these are asymptotically negligible compared to the second term, which is

$$\exp\left(-(1-o_n(1))\frac{\widetilde{c}C_{\rm BIG}^2}{C_0^{5/3}}\log^{1/3} n\right).$$

Comparing this to what is needed, the right-hand side in conclusion (i) of the lemma has an extra factor of 3 in the denominator, therefore conclusion (i) holds for sufficiently large n.

Clause (ii) of the definition of **GOOD** yields

$$\mathbb{P}_{\mathbf{x}}[\tau_1 < \infty] > \frac{1}{2}\exp\left(-\frac{\ell}{\widetilde{c}\lambda^2}\right) = \exp(-(1+o(1))(C_0^{1/3}/\widetilde{c})\log^{1/3} n),$$

and the extra factor of 2 on the right-hand side of conclusion (ii) of the lemma ensures it holds for sufficiently large n.

Finally, if $10C_{\mathrm{BIG}}\log n < k \le 3C_0^{5/3}\log^{5/3} n + 9C_{\mathrm{BIG}}\log n$, we can take $\tau_1 = \rho$. The second conclusion of the lemma is automatically satisfied. For the first,

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}\left[|\rho - g(\tau_1)| > C_{\mathrm{BIG}}\log n; \tau_1 < \infty\right] &= \mathbb{P}_{\omega}[|g(\rho)-\rho| > C_{\mathrm{BIG}}\log n] \\ &\le \exp\left(-\frac{10\widetilde{c}C_{\mathrm{BIG}}^2}{3C_0^{5/3}}(1-o_n(1))\log^{1/3} n,\right),\end{aligned}$$

where the second inequality holds by part (v) of Definition 17 with $a = C_{\mathrm{BIG}}\log n$.

To prove (iii), we set $\mathcal{C} = \lceil C_{\mathrm{BIG}}\log n\rceil$ and apply a union bound to get

$$\begin{aligned}&\mathbb{P}_{\mathbf{x}}[\{f(\rho-\mathcal{C}) > \tau_1\}\cup\{f(\rho+\mathcal{C}) < \tau_1\}; \tau_1 < \infty] \\ &\qquad < \mathbb{P}_{\mathbf{x}}\left[|\rho - g(\tau_1)| > 99\mathcal{C}/100; \tau_1 < \infty\right] + \mathbb{P}_{\mathbf{x}}[f(\rho-\mathcal{C}) = f(\rho - \lceil 99\mathcal{C}/100\rceil)].\end{aligned}$$

The first term on the right side, which dominates asymptotically, can be bounded as in our proof of (i). □

Finally, we are able to define k_* and then τ_2 from Theorem 2. Recall from the remarks following Definition 22 that the event **GOOD** is measurable with respect to $\mathbf{x}(0:k_5)$, hence with $k_5 = k$, the algorithm knows whether **GOOD** has occurred.

Definition 28 (the good alignment location k_*)**.** *Let $k \in \mathbb{N}$ satisfy $k \ge 9\lceil C_{\mathrm{BIG}}\log n\rceil$, and let $\mathbf{x}(0:k)$ be given. Set*

$$\begin{aligned}\ell &= \lceil C_{\mathrm{BIG}}\log^{1/3} n\rceil, \\ \lambda &= C_1^{5/3}, \\ C &= C_1, \\ k_0 &:= k - 9\lceil C_{\mathrm{BIG}}\log n\rceil, \\ k_5 &:= k.\end{aligned} \tag{39}$$

Let k_* be the least $k_3 \in [k-5\lceil C_{\mathrm{BIG}}\log n\rceil + \ell, k-4\lceil C_{\mathrm{BIG}}\log n\rceil]\cap\mathbb{Z}$ satisfying $\mathbf{GOOD}(k_0, k_3-\ell+1, k_3, k_5, \lambda, C, \mathbf{x}')$ *where $\mathbf{x}'(0:k) = \mathbf{x}(0:k)$ and $x'_j = 0$ for $j > k$. If the set is empty we set $k_* = \infty$. Let $\Xi_{\mathrm{bad}}(k,2)$ denote the set of $\mathbf{x}(0:k)$ for which $k_* = \infty$. For $k < 9\lceil C_{\mathrm{BIG}}\log n\rceil$ let $\Xi_{\mathrm{bad}}(k,2)$ be empty.*

We remark that $k_* \in \sigma(\mathbf{x})$ is a function of the message only, not the trace. Therefore, when aligning the $\lceil \exp(-M \log^{1/3} n) \rceil$ conditionally independent traces, the position k_* will be the same for all of them.

Lemma 29. *For sufficiently large n, the inequality $C_{\rm BIG} \geq 80\widetilde{c}^{-1} C_1^{8/3}$ implies*

$$\mu(\Xi_{\rm bad}(k,2)) \leq n^{-3}\,.$$

Proof. The definition is built for applying Lemma 23. Define

$$\begin{aligned} k_1 &:= k - 5\lceil C_{\rm BIG} \log n \rceil, \\ k_4 &:= k - 4\lceil C_{\rm BIG} \log n \rceil, \\ L &= 9\lceil C_{\rm BIG} \log n \rceil, \\ b &= \lceil (\log^{2/3} n)/2 \rceil, \end{aligned}$$

and observe that $b \leq L/\ell$. Applying Lemma 23, it follows that for all sufficiently large n,

$$\begin{aligned} \mu(\Xi_{\rm bad}(k,2)) &\leq L^b \exp\left(-\frac{\widetilde{c} b \ell}{10 C_1 \lambda}\right) \\ &\leq \exp\left(\left(\frac{\log^{2/3} n \log(9 C_{\rm BIG} \log n)}{2} - \frac{\widetilde{c} C_{\rm BIG} \log n}{20 C_1^{8/3}}\right)(1 - o_1(1))\right) \\ &\leq \exp\left(\frac{\log n}{2} - 4 \log n\right), \end{aligned}$$

under the hypothesis that $\widetilde{c} C_{\rm BIG}/(20 C_1^{8/3}) \geq 4$. □

Definition 30 (the true alignment τ_2)**.** *Let τ_1 be as in Definition 26, let $[k_2, k_3] := [k_* - \ell + 1, k_*]$ and let k_0, k_5, ℓ, λ and C be as in* (39)*. If $\tau_1 < \infty$ define the set*

$$\mathcal{I} := [j_1, j_2] \cap \mathbb{Z} := [\tau_1 + 2\lceil C_{\rm BIG} \log n \rceil\,,\, \tau_1 + 7\lceil C_{\rm BIG} \log n \rceil] \cap \mathbb{Z},$$

and if $\tau_1 = \infty$ set $\mathcal{I} = \emptyset$. Define τ_2 to be $\tau_{\mathcal{I}}$ in (31)*, that is,*

$$\tau_2 = \inf\{k' \in \mathcal{I} \,:\, T(k_3, k') = 1\}\,.$$

Proof of Theorem 2. Choose $C_{\rm BIG} \geq 80\widetilde{c}^{-1} C_1^{8/3}$ so that Lemma 29 may be applied. Let $\Xi_{\rm bad} := \bigcup_{k=1}^{n} (\Xi_{\rm bad}(k,1) \cup \Xi_{\rm bad}(k,2))$. If $A_{\mathcal{I}}$ fails, then because $k_3 \in [k_1, k_4]$, one of the following must occur: $f(k_0) > j_1$ or $f(k_5) < j_2$ or $f(k_1) > j_1$ or $f(k_4) < j_2$ or $f(k_3) - \ell \leq f(k_0) + \mathcal{C}$. Let $\mathcal{C} = \lceil C_{\rm BIG} \log n \rceil$. Let $A_{\mathcal{I}}^{(1)}$ be the event that $f(k_0 - \mathcal{C}) \leq \tau_1$ and $f(k_0 + \mathcal{C}) \geq \tau_1$, and let $A_{\mathcal{I}}^{(2)}$ be the event that none of the following inequalities are satisfied

$$\begin{aligned} f(k_0) - f(k_0 - \mathcal{C}) &\geq 2\mathcal{C}, \\ f(k_0 + 9\mathcal{C}) - f(k_0 + \mathcal{C}) &\leq 7\mathcal{C}, \\ f(k_0 + 4\mathcal{C}) - f(k_0 + \mathcal{C}) &\leq 2\mathcal{C}, \\ f(k_0 + 5\mathcal{C}) - f(k_0 - \mathcal{C}) &\geq 7\mathcal{C}. \end{aligned} \tag{40}$$

We will first verify that $A_{\mathcal{I}}^{(1)}$ and $A_{\mathcal{I}}^{(2)}$ satisfy the assumptions of Definition 22. It is immediate by definition that $A_{\mathcal{I}}^{(1)} \cap A_{\mathcal{I}}^{(2)} \subset A_{\mathcal{I}}$. Each of the events in (40) have probability bounded above by $\exp(-10\widetilde{c}(2\mathcal{C})^2/(8\mathcal{C})) = n^{-5\,\widetilde{c}\,C_{\rm BIG}}$, which implies that $\mathbb{P}_{\mathbf{x}}[(A_{\mathcal{I}}^{(2)})^c]$ decays at least polynomially in $\exp(-\ell^3)$. Observe that $A_{\mathcal{I}}^{(1)} \in \mathcal{G}^{k_5}_{f(k_0)+\lceil L/9\rceil} = \mathcal{G}^{k_5}_{f(k_0)+\mathcal{C}}$. We see from Lemma 27$(ii)-(iii)$ that

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}[A_{\mathcal{I}}^{(1)}] &\geq \mathbb{P}_{\mathbf{x}}[\tau_1 < \infty] - \mathbb{P}_{\mathbf{x}}\left[\tau_1 < \infty;\, \left(A_{\mathcal{I}}^{(1)}\right)^c\right] \\ &\geq \exp\left(-\frac{2C_0^{1/3}}{\widetilde{c}}\log^{1/3} n\right) - \exp\left(-\frac{\widetilde{c}C_{\rm BIG}^2}{4C_0^{5/3}}\log^{1/3} n\right) \geq \exp\left(-\frac{\ell}{2\widetilde{c}\lambda^2}\right). \end{aligned} \tag{41}$$

We now check, slightly out of order, that the conclusions $(i)-(v)$ of Theorem 2 and the inequality in conclusion (iv) hold when the constants are as described in Section 3.4.

(i) By construction τ_2 is bounded above by $2n$ when finite and is a stopping time on $\{\widetilde{\mathcal{G}}_i^k\}$. Also by construction the set $\Xi_{\rm bad}$ depends only on the first $2n$ bits of $\mathbf{x}$. From (38) and Lemma 29, we see that for n sufficiently large, $\mu(\Xi_{\rm bad}) \leq \sum_{k=1}^n \mu(\Xi_{\rm bad}(k,1)) + \mu(\Xi_{\rm bad}(k,2)) \leq 2n^{-2}$.

(ii) Recall $C_{\rm back} = 5C_{\rm BIG}$. By construction $k_* \in [k-5\mathcal{C}, k-4\mathcal{C}]$, therefore (ii) is satisfied.

(iv) Recall $C_{\rm align} = C_{\rm BIG}/10$ and $C_{\rm false} = \widetilde{c}\,C_{\rm BIG}/(2C_1^{5/3})$. Applying clause (i) in the definition of **GOOD** gives

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}[\infty > |g(\tau_2) - k_*| > C_{\rm align}\log^{1/3} n] \;\leq\; & \mathbb{P}_{\mathbf{x}}[A_{\mathcal{I}}^c; \tau_1 < \infty] + 18\mathcal{C}\exp\left(-\frac{\widetilde{c}C_{\rm BIG}}{C_1^{5/3}}\log^{1/3} n\right) \\ & + 3\widetilde{c}^{-1}\exp\left(-\frac{\widetilde{c}C_{\rm BIG}^2\log^{2/3} n}{100C_{\rm BIG}\log^{1/3} n}\right).\end{aligned}$$

We observe that the second term on the right side dominates, and for n sufficiently large, the right side is bounded above by $\exp(-C_{\rm false}\log^{1/3} n)$.

(iii) Recall $C_{\rm avg} = 2C_{\rm BIG}/C_1^2$ and $C_{\rm true} = 2C_{\rm BIG}/(\widetilde{c}C_1^{10/3})$. Applying clause (ii) of the definition of **GOOD** gives

$$\begin{aligned}\mathbb{P}_{\mathbf{x}}[|g(\tau_2) - k_*| \leq C_{\rm align}\log^{1/3} n] \;&\geq\; \mathbb{P}_{\mathbf{x}}[\tau_2 < \infty; A_{\mathcal{I}}] - \mathbb{P}_{\mathbf{x}}[\infty > |g(\tau_2) - k_*| > C_{\rm align}\log^{1/3} n] \\ &\geq\; \frac{1}{2}\exp\left(-\frac{C_{\rm BIG}\log^{1/3} n}{\widetilde{c}C_1^{10/3}}\right) - \exp(-C_{\rm false}\log^{1/3} n) \\ &\geq\; \frac{1}{2}\exp\left(-\frac{1}{2}C_{\rm true}\log^{1/3} n\right) - \exp(-C_{\rm false}\log^{1/3} n) \\ &\geq\; \exp(-C_{\rm true}\log^{1/3} n)\end{aligned}$$

for n sufficiently large, once we verify that $C_{\rm false} > C_{\rm true}$.

(v) Recall $C_{\rm avg} = 3C_{\rm BIG}/C_1^2$. Applying clause (iii) of the definition of **GOOD** gives, with $Z := |k_3 - g(\tau_2)|$,

$$\mathbb{E}_{\mathbf{x}}\left[Z\mathbf{1}_{Z\leq \ell/10}\mathbf{1}_{A_{\mathcal{I}}} \,|\, \tau_2 < \infty\right] \leq \frac{3\ell}{2C_1\lambda^{3/5}}.$$

Removing the restriction to $A_{\mathcal{I}}$ adds at most $(\ell/10)\mathbb{P}_{\mathbf{x}}[Z > \ell/10 \,|\, \tau_2 < \infty]$. Therefore, substituting $\ell = \lceil 10 C_{\text{align}} \log^{1/3} n \rceil = \lceil C_{\text{BIG}} \log^{1/3} n \rceil$ gives

$$\mathbb{E}_{\mathbf{x}}\left[Z\mathbf{1}_{Z \le C_{\text{align}} \log^{1/3} n} \,|\, \tau_2 < \infty\right] \le \frac{3\lceil C_{\text{BIG}} \log^{1/3} n \rceil}{2C_1^2} + \frac{\lceil C_{\text{BIG}} \log^{1/3} n \rceil}{10} \frac{\mathbb{P}_{\mathbf{x}}[Z > C_{\text{align}} \log^{1/3} n]}{\mathbb{P}_{\mathbf{x}}[\tau_2 < \infty]}.$$

Conclusions (iii) and (iv), along with $C_{\text{false}} > C_{\text{true}}$, imply that $\dfrac{\mathbb{P}_{\mathbf{x}}[Z > C_{\text{align}} \log^{1/3} n]}{\mathbb{P}_{\mathbf{x}}[\tau_2 < \infty]} \le \exp(-\Theta(\log^{1/3} n))$, therefore, for n sufficiently large,

$$\mathbb{E}_{\mathbf{x}}\left[Z\mathbf{1}_{Z \le C_{\text{align}} \log^{1/3} n} \,|\, \tau_2 < \infty\right] \le \frac{3.1 C_{\text{BIG}} \log^{1/3} n}{2C_1^2}$$

as required.

Finally, we check the inequality in conclusion (iv),

$$\begin{aligned} \frac{C_{\text{false}} - C_{\text{true}}}{C_{\text{BIG}}} &= \frac{\widetilde{c}}{2C_1^{5/3}} - \frac{2}{\widetilde{c} C_1^{10/3}}, \\ \frac{C_{\text{sep}}(8C_{\text{avg}} + C_{\text{back}}^{1/3})}{C_{\text{BIG}}} &= C_{\text{sep}}\left(\frac{2}{C_1^2} + 5C_{\text{BIG}}^{-2/3}\right). \end{aligned}$$

Multiplying through by C_1^2, we require

$$\frac{\widetilde{c}}{2} C_1^{1/3} - \frac{2}{\widetilde{c}} C_1^{-4/3} > 2C_{\text{sep}} + 5C_{\text{sep}} C_{\text{BIG}}^{-2/3} C_1^2.$$

Having chosen $C_1 = 64\widetilde{c}^{-12}$, and with C_{BIG} sufficiently large, this is satisfied when

$$\frac{2}{\widetilde{c}^3} - \frac{\widetilde{c}^{15}}{8} > 7C_{\text{sep}},$$

which is assumption (vi) in Definition 17. □

6 Reconstruction from approximately aligned strings: Proof of Theorem 3

Lemma 31. *Let $\mathbf{a} = (a_0, a_1, \dots) \in [-1, 1]^{\mathbb{N}}$, and let $\widetilde{\mathbf{a}}$ be the output from the deletion-insertion channel with deletion (resp. insertion) probability q (resp. q'), applied to the randomly shifted string $\theta^S \mathbf{a}$, where the shift S is as in Theorem 3. Let $\phi_1(w) = pw + q$, $\phi_2(w) = \frac{p'w}{1-q'w}$, and $\beta_s = \mathbb{P}[S = s]$ for $s \in \mathbb{N}$. Define*

$$P(z) := \sum_{s=0}^{d} \beta_s z^s, \qquad Q(z) := \sum_{j=0}^{\infty} a_j z^j.$$

Then, for any $|w| < 1$,

$$\mathbb{E}\left[\sum_{j \ge 0} \widetilde{a}_j w^j\right] = p \cdot P\left(\frac{1}{\phi_2 \circ \phi_1(w)}\right) \cdot Q(\phi_2 \circ \phi_1(w)). \tag{42}$$

Proof. Recall the construction of $\widetilde{\mathbf{x}}$ from $\mathbf{x}$ mentioned in Section 1.2, where we first insert a geometric number (minus one) bits before each bit of $\mathbf{x}$ and then delete each bit independently with probability q. From this description we see that we can sample $\widetilde{\mathbf{a}}$ by first setting $\widetilde{\mathbf{a}}^{(2)} = \theta^S \mathbf{a}$, then letting $\mathbf{a}^{(3)}$ be the string we get when sending $\mathbf{a}^{(2)}$ through the insertion channel with insertion probability q' (and no deletions), and finally obtain $\widetilde{\mathbf{a}}$ by sending $\widetilde{\mathbf{a}}^{(3)}$ through the deletion channel with deletion probability q (and no insertions). Three elementary generating function manipulations (see, respectively, ([PZ17, Lemma 4.2], [NP17, Lemma 5.2], and [NP17, Lemma 2.1]) give

$$\mathbb{E}\left[\sum_{j\geq 0} a_j^{(2)} w^j\right] = P(w^{-1})Q(w), \qquad \mathbb{E}\left[\sum_{j\geq 0} a_j^{(3)} w^j \;\middle|\; \mathbf{a}^{(2)}\right] = \sum_{j\geq 0} a_j^{(2)} \phi_2(w)^j,$$
$$\mathbb{E}\left[\sum_{j\geq 0} \widetilde{a}_j w^j \;\middle|\; \mathbf{a}^{(3)}\right] = \sum_{j\geq 0} a_j^{(3)} \phi_1(w)^j.$$

Combining these identifies we get (42):

$$\begin{aligned}\mathbb{E}\left[\sum_{j\geq 0} \widetilde{a}_j w^j\right] &= \mathbb{E}\left[\sum_{j\geq 0} a_j^{(3)} \phi_1(w)^j\right] = \mathbb{E}\left[p \sum_{j\geq 0} a_j^{(2)} \big(\phi_2 \circ \phi_1(w)\big)^j\right] \\ &= pP\bigg(\frac{1}{\phi_2 \circ \phi_1(w)}\bigg) Q(\phi_2 \circ \phi_1(w)).\end{aligned}$$

□

The following result is Corollary 3.2 of [BE97] with $M = 1$, $a = \ell$ and $c_1 = C_{\mathrm{BE}}$, observing that the class of polynomials whose coefficients have modulus at most 1 are in their class $\mathcal{K}_1^1$ and that that their statement $\mathcal{K}_M := \mathcal{K}_M^0$ after their definition of $\mathcal{K}_M^\mu$ should be ignored in favor of the correct statement $\mathcal{K}_M := \mathcal{K}_M^1$ occurring in their Corollary 3.2.

Lemma 32 (Borwein and Erdélyi 1997)**.** *There is a universal constant C_{BE} such that for any polynomial f satisfying $|f(0)| = 1$ and whose coefficients have modulus at most 1, and for any arc α of the unit circle whose angular length is denoted $s \in (0, 2\pi)$,*

$$\sup_{z\in\alpha} |f(z)| \geq e^{-C_{\mathrm{BE}}/s}.$$

□

Proof Theorem 3. Let $\mathbf{a} = \mathbf{x}^{(1)} - \mathbf{x}^{(2)} \in \{-1, 0, 1\}^{\mathbb{N}}$, $L = m^{1/3}$, and $\rho = 1 - 1/L^2$. Define $j_0 := \inf\{j \in \mathbb{N} : a_j \neq 0\} \in \{d, d+1, \ldots, m\}$ and $\widetilde{Q}(z) := z^{-j_0} Q(z)$ with $|\widetilde{Q}(0)| = 1$.

> Claim: There is a $c_2 \in (0, 1/20)$ depending only on q, q' such that if $|\arg(z)| \leq c_2/L$, $|z| = 1$, and $w = \phi_1^{-1}(\phi_2^{-1}(\rho \cdot z))$, then $|w| \leq 1 - c_2/L^2$.

Proof: Observe that ϕ_2 (resp. ϕ_1) is a Möbius transformation mapping $\mathbb{D}$ to a smaller disk which is contained in $\overline{\mathbb{D}}$, which is tangent to $\partial\mathbb{D}$ at 1, and which maps $\mathbb{R}$ to $\mathbb{R}$. In particular, defining $\Psi := \phi_1^{-1} \circ \phi_2^{-1}$, we get by linearizing the map around $z = 1$ that $\Psi(1 + \widetilde{z}) = 1 + a\widetilde{z} + O(|\widetilde{z}|^2)$ for $a > 1$ depending only on q, q'. Writing $z = e^{i\theta}$, we have

$$\begin{aligned} w &= \Psi(\rho e^{i\theta}) \\ &= 1 + a(\rho e^{i\theta} - 1) + O(|\rho e^{i\theta} - 1|^2) \\ &= 1 + a\big((1 - L^{-2})(1 + i\theta) - 1\big) + O(\theta^2 + L^{-4}) \\ &= 1 + a(-L^{-2} + i\theta) + O(\theta^2 + L^{-4}), \end{aligned}$$

so $|w| < 1 - c_2/L^2$ when $c_2 = c_2(q, q')$ is sufficiently small, and the claim is proved.

Observe that $z \mapsto \widetilde{Q}(\rho \cdot z)$ has coefficients of modulus at most 1, hence we may apply Lemma 32 to find $z_0 = e^{i\theta}$ with $|\theta| \le c_2/L$ such that $|\widetilde{Q}(\rho z_0)| \ge e^{-C_{\mathrm{BE}} L/c_2}$. By definition of c_2, we see that $w_0 := \Psi(\rho \cdot z_0)$ satisfies $|w_0| \le 1 - c_2/L^2$. An illustration of the points z_0 and w_0 is given in Figure 9. We show next that

$$\left| P\left(\frac{1}{\rho z_0}\right) \right| \ge \frac{1}{2}\,. \tag{43}$$

To see this, define $\widetilde{P}(z) = z^{-\mathbb{E}S} P(z)$, which is an analytic function in the right half-plane. For all z in the right half-plane satisfying $1 \le |z| \le \rho^{-1}$, differentiating $\widetilde{P}$ and using $\mathbb{E}[|S - \mathbb{E}S|] \le L$

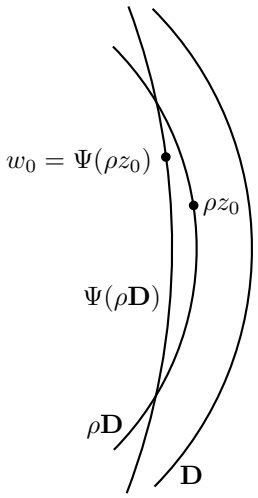

Figure 9: Illustration of the points $z_0, w_0 \in \mathbb{C}$ defined in the proof of Theorem 3. We first choose $z_0 = e^{i\theta}$ for $|\theta| \le c_2/L$, such that $|\widetilde{Q}(\rho\cdot)|$ is bounded from below. Then we observe that $|w_0| < 1 - c_2/L^2$, which helps us to bound the modulus of $\mathbb{E}[\sum_{j\ge 0} \widetilde{a}_j w_0^j]$ from below.

and $d \le L^2$ gives

$$|\widetilde{P}'(z)| = \left|\sum_{j=0}^{d}(j-\mathbb{E}S)\beta_j z^{j-\mathbb{E}S-1}\right| \le \sum_{j=0}^{d}|j-\mathbb{E}S|\cdot\beta_j\cdot|z|^{j-\mathbb{E}S-1}$$
$$\le \rho^{-d}\cdot\mathbb{E}[|S-\mathbb{E}S|] \le \rho^{-d}L \le e^{\frac{1.1d}{L^2}}\cdot L \le 4L,$$

We also have

$$|\rho^{-1}z_0^{-1}-1| = \rho^{-1}|1-\rho z_0| \le |z_0-1| + \rho^{-1}(1-\rho) \le \frac{c_2}{L}+\frac{2}{L^2}.$$

Therefore, for all sufficiently large m,

$$\begin{aligned}\left|P(\rho^{-1}z_0^{-1})\right| &= \rho^{-\mathbb{E}S}\left|\widetilde{P}(\rho^{-1}z_0^{-1})\right| \ge 1-|\widetilde{P}(\rho^{-1}z_0^{-1})-1| \\ &= 1-\left|\int_1^{\rho^{-1}z_0^{-1}}\widetilde{P}'(z)\,dz\right| \ge 1-|\rho^{-1}z_0^{-1}-1|\cdot 4L \\ &\ge 1-\left(\frac{c_2}{L}+\frac{2}{L^2}\right)\cdot 4L \ge \frac{1}{2},\end{aligned}$$

proving (43).

Using Lemma 31 and the above estimates, it follows that

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{j\ge 0}\widetilde{a}_j w_0^j\right]\right| &\ge p\cdot\left|P\left(\frac{1}{\rho z_0}\right)\right|\cdot\rho^{j_0}\cdot|\widetilde{Q}(\rho z_0)| \\ &\ge p\frac{1}{2}\left(1-\frac{1}{L^2}\right)^m e^{-C_{\mathrm{BE}}L/c_2} \\ &\ge e^{-C_{\mathrm{sep}}L}\end{aligned} \tag{44}$$

for a constant $C_{\mathrm{sep}} > 1$ depending only on q, q'. Therefore, for any $C_{\mathrm{fwd}} > 1$,

$$\left|\sum_{j\ge C_{\mathrm{fwd}}m}\widetilde{a}_j w_0^j\right| \le \left|\sum_{j\ge C_{\mathrm{fwd}}m}\left(1-\frac{c_2}{L^2}\right)^j\right| \le L^2 c_2^{-1}e^{-C_{\mathrm{fwd}}L/c_2}\,. \tag{45}$$

Combining (44) and (45), for C_{fwd} a sufficiently large constant multiple of C_{sep},

$$\mathbb{E}\left[\sum_{j\ge 0}\left|\widetilde{a}_j w_0^j\right|\right] \ge \left|\mathbb{E}\left[\sum_{j\ge 0}\widetilde{a}_j w_0^j\right]\right| \ge \frac{1}{2}\exp(-C_{\mathrm{sep}}L). \tag{46}$$

It follows that there is a $j < C_{\mathrm{fwd}}m$ for which

$$|\widetilde{a}_j| \ge |\widetilde{a}_j w_0^j| \ge (2C_{\mathrm{fwd}}m)^{-1}\exp(-C_{\mathrm{sep}}L)\,.$$

Increasing C_{sep} if necessary finishes the proof. □

References

[Azu67] K. Azuma. Weighted sums of certain dependent random variables. *Tôhoku Math. J. (2)*, 19:357–367, 1967. MR0221571

[BE97] P. Borwein and T. Erdélyi. Littlewood-type problems on subarcs of the unit circle. *Indiana Univ. Math. J.*, 46(4):1323–1346, 1997. MR1631600

[BKKM04] 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*, pages 910–918. ACM, New York, 2004. MR2290981

[DOS17] A. De, R. O'Donnell, and R. Servedio. Optimal mean-based algorithms for trace reconstruction. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 17, pages 1047–1056, New York, NY, USA, 2017. ACM.

[Gri99] G. Grimmett. *Percolation*, volume 321 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 1999. MR1707339

[Har60] T. E. Harris. A lower bound for the critical probability in a certain percolation process. *Proc. Cambridge Philos. Soc.*, 56:13–20, 1960. MR0115221

[HMPW08] 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*, pages 389–398. ACM, New York, 2008. MR2487606

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. MR0144363

[KM05] 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)*, pages 297–301. IEEE, 2005.

[Lev01a] V. I. Levenshtein. Efficient reconstruction of sequences. *IEEE Trans. Inform. Theory*, 47(1):2–22, 2001. MR1819952

[Lev01b] V. I. Levenshtein. Efficient reconstruction of sequences from their subsequences or supersequences. *J. Combin. Theory Ser. A*, 93(2):310–332, 2001. MR1805300

[Mit09] M. Mitzenmacher. A survey of results for deletion channels and related synchronization channels. *Probab. Surv.*, 6:1–33, 2009. MR2525669

[MPV14] A. McGregor, E. Price, and S. Vorotnikova. Trace reconstruction revisited. In *Algorithms—ESA 2014*, volume 8737 of *Lecture Notes in Comput. Sci.*, pages 689–700. Springer, Heidelberg, 2014. MR3253172

[NP17] 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 17, pages 1042–1046, New York, NY, USA, 2017. ACM.

[PZ17] Y. Peres and A. Zhai. Average-case reconstruction for the deletion channel: subpolynomially many traces suffice, 2017. To appear in FOCS.

[VS08] K. Viswanathan and R. Swaminathan. Improved string reconstruction over insertion-deletion channels. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 399–408. ACM, New York, 2008. MR2487607