On k-Mer-Based and Maximum Likelihood Estimation Algorithms for Trace Reconstruction

Kuan Cheng, * Elena Grigorescu, † Xin Li, ‡ Madhu Sudan, § and Minshen Zhu, ¶
*Peking University, Haidian, Beijing, China, ckkcdh@pku.edu.cn
†Purdue University, West Lafayette, IN, USA, elena-g@purdue.edu
‡ Johns Hopkins University, Baltimore, MD, USA, lixints@cs.jhu.edu
§ Harvard University, Cambridge, Massachusetts, USA, madhu@cs.harvard.edu
¶ minshen.zh@gmail.com

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619392

***Abstract*—The goal of the trace reconstruction problem is to recover a string $\mathbf{x} \in \{0,1\}^n$ given many independent *traces* of $\mathbf{x}$, where a trace is a subsequence obtained from deleting bits of $\mathbf{x}$ independently with some given probability.**

In this paper we consider two kinds of algorithms for the trace reconstruction problem.

We first observe that the state-of-the-art result of Chase (STOC 2021), which is based on statistics of arbitrary length-k subsequences, can also be obtained by considering the "k-mer statistics", i.e., statistics regarding occurrences of *contiguous* k-bit strings (a.k.a, *k-mers*) in the initial string $\mathbf{x}$, for $k = 2n^{1/5}$. Mazooji and Shomorony (ISIT 2023) show that such statistics (called k-mer density map) can be estimated within ε accuracy from $\mathrm{poly}(n, 2^k, 1/\varepsilon)$ traces. We call an algorithm to be *k-mer-based* if it reconstructs $\mathbf{x}$ given estimates of the k-mer density map. Such algorithms essentially capture all the analyses in the worst-case and smoothed-complexity models of the trace reconstruction problem we know of so far.

Our first, and technically more involved, result shows that any k-mer-based algorithm for trace reconstruction must use $\exp(\Omega(n^{1/5}\sqrt{\log n}))$ traces, under the assumption that the estimator requires $\mathrm{poly}(2^k, 1/\varepsilon)$ traces, thus establishing the optimality of this number of traces. Our analysis also shows that the analysis technique used by Chase is essentially tight, and hence new techniques are needed in order to improve the worst-case upper bound.

Our second, simple, result considers the performance of the Maximum Likelihood Estimator (MLE), which specifically picks the source string that has the maximum likelihood to generate the samples (traces). We show that the MLE algorithm uses a nearly optimal number of traces, i.e., up to a factor of n in the number of samples needed for an optimal algorithm, and show that this factor of n loss may be necessary under general "model estimation" settings.

I. Introduction

The trace reconstruction problem is an infamous question introduced by Batu, Kannan, Khanna and McGregor [1] in the context of computational biology. It asks to design algorithms that recover a string $\mathbf{x} \in \{0,1\}^n$ given access to *traces* $\tilde{\mathbf{x}}$ of $\mathbf{x}$, obtained by deleting each bit independently with some given probability $p \in [0,1)$. The best current upper and lower bounds are exponentially apart, namely $\exp(\widetilde{O}(n^{1/5}))$ traces are sufficient for reconstruction [2] (improving upon the $\exp(O(n^{1/3}))$ of [3], [4]) and $\widetilde{\Omega}(n^{3/2})$ [5], [6] are necessary.

The problem has been recently studied in several variants so far [1]–[5], [7]–[24] and it continues to elicit interest due to its deceptively simple formulation, as well as its motivating applications to DNA computing [25].

In this paper, we focus on the worst-case formulation of the problem, which is equivalent from an information-theoretic point of view to the *distinguishing* variant. In this variant, the goal is to distinguish whether the received traces come from string $\mathbf{x} \in \{0,1\}^n$ or from $\mathbf{y} \in \{0,1\}^n$, for some known $\mathbf{x} \neq \mathbf{y}$.

a) Algorithms based on k-bit statistics : A very natural kind of algorithms [3], [4], [9] operates using the mean of the received traces at each location $i \in [n]$ (one may assume that traces of smaller length than n are padded with 0's at the end). Indeed, let $\mathcal{D}_{\mathbf{x}}$ be the distribution of the traces induced by the deletion channel on input $\mathbf{x}$. A mean/1-bit-statistics -based algorithm first estimates from the received traces the mean vector $\mathbf{E}(\mathbf{x}) = \big(E_0(\mathbf{x}), \cdots, E_{n-1}(\mathbf{x})\big) \in [0,1]^n$, where the j-th coordinate is defined as $E_j(\mathbf{x}) = \mathop{\mathbb{E}}_{\tilde{\mathbf{x}} \sim \mathcal{D}_{\mathbf{x}}} \left[\tilde{x}_j\right]$. It then may perform further post-processing without further inspection of the traces. Solving the distinguishing problem then

reduces by standard arguments to understanding the ℓ_1-norm between the mean traces of $\mathbf{x}$ and $\mathbf{y}$, namely the number T of traces satisfies $\Omega\left(1/\left\|\mathbf{E}(\mathbf{x})-\mathbf{E}(\mathbf{y})\right\|_{\ell_1}\right) = T = O\left(1/\left\|\mathbf{E}(\mathbf{x})-\mathbf{E}(\mathbf{y})\right\|_{\ell_1}^2\right)$. Further, [3], [4] related the ℓ_1-norm above with the supremum of a certain real univariate polynomial over the complex plane. Using techniques from complex analysis they proved that mean-based algorithms using $\exp(O(n^{1/3}))$ traces and outputting the string $\mathbf{s} \in \{\mathbf{x}, \mathbf{y}\}$ whose $\mathbf{E}(\mathbf{s})$ is closer in ℓ_1-distance to the estimate is a successful reconstruction algorithm. Furthermore, any mean-based algorithm needs $\exp(\Omega(n^{1/3}))$ traces to succeed with high probability [3], [4].

A general class of algorithms may operate by using k-bit statistics [2], for $k \geq 1$. Specifically, for $w \in \{0,1\}^k$, the algorithm estimates from the given traces, for tuples $0 \leq i_0 < i_1 < \cdots < i_{k-1} \leq n-1$, the quantity $\mathop{\mathbb{E}}_{\widetilde{\mathbf{x}}\sim\mathcal{D}_{\mathbf{x}}}\left[\prod_{0\leq j<k} \mathbf{1}\left\{\, \widetilde{x}_{i_j} = w_j \,\right\}\right]$. After the estimation step, whose accuracy can be argued via standard Chernoff bounds, the algorithm does not need the traces anymore and may perform further post-processing in order to output the correct string. The result of Chase follows from showing that for $k = 2n^{1/5}$ there is a string $w \in \{0,1\}^k$ for which the ℓ_1-distance between the corresponding k-bit statistics between $\mathbf{x}$ and $\mathbf{y}$ is large.

b) Algorithms based on k-mer statistics: Another variant proposed by Mazooji and Shomorony [26] considers algorithms which operate using estimates of statistics regarding occurrences of *contiguous* k-bit strings (a.k.a, *k-mers*) in the *initial* string $\mathbf{x}$. We denote by $\mathbf{1}\left\{\, \mathbf{x}[j\colon j+k-1] = w \,\right\}$ the indicator bit of whether $w \in \{0,1\}^k$ occurs as a subword in $\mathbf{x}$ from position j. The following definition which is central to our paper.

Definition 1 ([26]). *Given $\mathbf{x} \in \{0,1\}^n$ and a k-mer $w \in \{0,1\}^k$, for $i = 0, 1, \ldots, n-1$ denote $K_{w,\mathbf{x}}[i] \coloneqq \sum_{j=0}^{n-k} \binom{j}{i} p^{j-i}(1-p)^i \cdot \mathbf{1}\left\{\, \mathbf{x}[j\colon j+k-1] = w \,\right\}$. The vector $K_{\mathbf{x}} \coloneqq \left(K_{w,\mathbf{x}}[i]\colon w \in \{0,1\}^k, i \in [n]\right)$ is called the k-mer density map of $\mathbf{x}$.*

Note that the mean vector $\mathbf{E}(\mathbf{x})$ is, up to a factor of $1-p$, equivalent to the 1-mer density map. Indeed, for $k = 1$ and $w = 1$ we have $E_i(\mathbf{x}) = \mathop{\mathbb{E}}_{\widetilde{\mathbf{x}}\sim\mathcal{D}_{\mathbf{x}}}[\widetilde{x}_i] = \sum_{j=0}^{n-1} \Pr\left[\widetilde{x}_i \text{ comes from } x_j\right] \cdot x_j = \sum_{j=0}^{n-1} \binom{j}{i} p^{j-i}(1-p)^{i+1} \cdot x_j = (1-p)\cdot\sum_{j=0}^{n-1} \binom{j}{i} p^{j-i}(1-p)^i \cdot \mathbf{1}\left\{\, \mathbf{x}[j:j] = 1 \,\right\} = (1-p)K_{1,\mathbf{x}}[i]$.

As noted in [26], the techniques of [19] in the smoothed complexity model of trace reconstruction can also be viewed as based on k-mer density maps. Indeed, for a fixed $w \in \{0,1\}^k$, the number of its occurrences as a subword in $\mathbf{x}$ is $\sum_{j=0}^{n-1} \mathbf{1}\left\{\, \mathbf{x}[j\colon j+k-1] = w \,\right\} = \sum_{i=0}^{n-1} K_{w,\mathbf{x}}[i]$. They show that for $k = O(\log n)$, the subword vector (indexed by $w \in \{0,1\}^k$) uniquely determines the source string, with high probability [19, Lemma 1.1].

The main result of [26] is that given access to $T = \varepsilon^{-2} \cdot 2^{O(k)}\mathrm{poly}(n)$ traces of $\mathbf{x}$, one can recover an estimation $\hat{K}_{\mathbf{x}}$ of the k-mer density map $K_{\mathbf{x}}$ which is entry-wise ε-accurate, i.e., $\left\|\hat{K}_{\mathbf{x}} - K_{\mathbf{x}}\right\|_{\ell_\infty} \leq \varepsilon$. We remark that by replacing ε with $\varepsilon/(2^k n)$, one gets an estimate which is ε-accurate in ℓ_1-norm, while using asymptotically the same number of traces.

We make the following definition generalizing mean-based algorithms ([3], [4]).

Definition 2. *(Algorithms based on k-mer statistics) A trace reconstruction algorithm based on k-mer statistics works in two steps as follows:*

1) *Once the unknown source string $\mathbf{x} \in \{0,1\}^n$ is picked, it chooses an accuracy parameter $\varepsilon \in (0,1]$. It then receives an ϵ-accurate estimate (in ℓ_1-norm) of the k-mer density map $K_{\mathbf{x}}$ based on the traces. From here on the algorithm has no more access to the traces themselves. We define the cost of this part to be $2^k/\varepsilon$.*
2) *The algorithm may perform further post-processing and finish by outputting the source string.*

Since there is an algorithm to ε-estimate the k-mer density map with $\varepsilon^{-2} \cdot 2^{O(k)}\mathrm{poly}(n)$ many traces [26], it follows that an algorithm defined as in Definition 2 with cost T can be turned into a trace reconstruction algorithm with $\mathrm{poly}(T)$ samples.

We note that the k-mer density map estimators of [26] only use k-bit statistics of the traces, in fact statistics about contiguous k bits in the traces, and hence k-mer-based algorithms are a subclass of algorithms based on k-bit statistics.

In this work, we first observe that the upper bounds of Chase [2] can be in fact obtained via k-mer-based algorithms (see the formal statement in Theorem 1), and hence by only using statistics of contiguous subwords of the traces. Our main result says that k-mer-based algorithms require $\exp(\Omega(n^{1/5})\sqrt{n})$ many traces (see Theorem 2). In addition, the analysis of this result implies that the proof technique in Chase [2] cannot lead to a better analysis of the sample complexity (up to $\log^{4.5} n$

factors in the exponent), and hence new techniques are needed to significantly improve the current upper bound.

c) The Maximum Likelihood Estimator: In model estimation settings, a common tool for picking a "model" that best explains the observed data is the Maximum Likelihood Estimator (MLE). In the setting of trace reconstruction, it is natural to ask: What is the most likely trace distribution $\mathcal{D}_{\mathbf{x}}$ (and hence $\mathbf{x}$) to have produced the given sample/trace(s)? We formalize MLE next.

Definition 3 (Maximum Likelihood Estimation)**.** *Let $\mathcal{D} = \{ D_1, D_2, \ldots, D_m \}$ be a finite set of probability distributions over a common domain Ω. Given a sample $x \in \Omega$, the output of the Maximum Likelihood Estimation is (ties are broken arbitrarily)* $\mathrm{MLE}(x; \mathcal{D}) \coloneqq \arg\max_{i \in [m]} D_i(x)$. *For independently and identically distributed samples $x_1, x_2, \ldots, x_k \in \Omega$ the output of the Maximum Likelihood Estimation is (ties are broken arbitrarily) is* $\mathrm{MLE}(x_1, x_2, \ldots x_k; \mathcal{D}) \coloneqq \arg\max_{i \in [m]} \prod_{j \in [k]} D_i(x_j)$.

We present a simple proof that this algorithm (which takes exponential time, as it searches through all $\mathbf{x} \in \{0,1\}^n$) is in fact optimal in the number of *traces* used, up to an $O(n)$ factor blowup.

We also observe that in the average-case setting, where the source string is a uniformly random string from $\{0,1\}^n$, MLE is indeed optimal – without the $O(n)$ factor blowup (see Remark 1.)

A. Our Contributions

a) The power of k-mer-based algorithms: Our first result shows that algorithms based on k-mer statistics can reconstruct a source string using $\exp(\widetilde{O}(n^{1/5}))$ many traces. This follows from the following theorem.

Theorem 1 (Implied by [2])**.** *Let $\mathbf{x}, \mathbf{y} \in \{ 0,1 \}^n$ be two arbitrary distinct strings, and let $K_{\mathbf{x}}, K_{\mathbf{y}}$ be their k-mer density maps, respectively. Assuming $k = 2n^{1/5}$, it holds that* $\left\| K_{\mathbf{x}} - K_{\mathbf{y}} \right\|_{\ell_1} \geq \exp\left(-O(n^{1/5} \log^5 n) \right)$.

Based on Theorem 1, the algorithm estimates $\hat{K}$ within an accuracy of $\varepsilon = \exp(-O(n^{1/5} \log^5 n))$ and outputs the $\mathbf{x}$ that minimizes $\left\| \hat{K} - K_{\mathbf{x}} \right\|_{\ell_1}$. The cost of this k-mer-based algorithm is $\exp(O(n^{1/5} \log^5 n))$.

Our main result regarding k-mer-based algorithms is the following theorem which shows the tightness of the bound in Theorem 1.

Theorem 2. *Fix any $k \leq n^{1/5}$. Suppose $K_{\mathbf{x}}$ stands for the k-mer density map of $\mathbf{x}$. There exist distinct strings $\mathbf{x}, \mathbf{y} \in \{ 0,1 \}^n$ such that* $\left\| K_{\mathbf{x}} - K_{\mathbf{y}} \right\|_{\ell_1} \leq \exp\left(-\Omega(n^{1/5} \sqrt{\log n}) \right)$.

Hence, Theorem 2 implies that the cost of any k-mer-based algorithm for worst-case trace reconstruction is $\exp(\Omega(n^{1/5} \sqrt{\log n}))$.We remark that the proof of Theorem 2 further implies that the analysis technique of [2] is essentially tight, in the sense that no better upper bound (up to $\log^{4.5} n$ factors in the exponent) can be obtained via his analysis.

b) Maximum Likelihood Estimator: an optimal algorithm: We next turn to analyzing the performance of the MLE algorithm in the setting of trace reconstruction. Our main result essentially shows that if there is an algorithm for trace reconstruction that uses T traces and succeeds with probability $3/4$ then the MLE algorithm using $O(nT)$ traces succeeds with probability $3/4$. Hence, given that the current upper bounds for the worst-case reconstruction problem are exponential in n, we may view the MLE as an optimal algorithm for trace reconstruction.

Theorem 3. *Suppose $\mathcal{D} = \{ D_0, D_1, \ldots, D_m \}$ is such that $d_{\mathrm{TV}}(D_0, D_i) \geq 1 - \varepsilon$ for any $1 \leq i \leq m$. Then we have* $\Pr_{x \sim D_0}\left[\mathrm{MLE}(x; \mathcal{D}) = 0 \right] \geq 1 - m\varepsilon$.

We remark that the loss of a factor of m in Theorem 3 is generally inevitable. Here is a simple example: let D_0 be the uniform distribution over $[m]$, and for $i = 1, 2, \ldots, m$, let D_i be the point distribution supported on $\{ i \}$. We have $d_{\mathrm{TV}}(D_0, D_i) = ((m-1)/m + (1 - 1/m))/2 = 1 - 1/m$. However, $\Pr_{x \sim D_0}\left[\mathrm{MLE}(x; \mathcal{D}) = 0 \right] = 0$.

For a string $\mathbf{x} \in \{ 0,1 \}^n$, let $D_{\mathbf{x}}$ denote the trace distribution of $\mathbf{x}$. Theorem 3 implies the following corollary, which implies that in some sense the Maximum Likelihood Estimation is a universal algorithm for trace reconstruction.

Corollary I.1. *Suppose T traces are sufficient for worst-case trace reconstruction with a success rate $3/4$. Then for any $\varepsilon > 0$, Maximum Likelihood Estimation with $8 \ln(1/\varepsilon) \cdot nT$ traces solves worst-case trace reconstruction with success rate $1 - \varepsilon$.*

Corollary I.1 incurs a factor of $O(n)$ to the sample complexity. While we currently do not know whether this blowup is necessary for trace reconstruction, the next result shows that it is inevitable for the more general "model estimation" problem.

Theorem 4. *For any integer $n \geq 1$, there is a set of distributions $\mathcal{D} = \{ D_0, D_1, D_2, \ldots, D_m \}$ over a*

common domain Ω *of size* $|\Omega| = m + n$*, where* $m = \binom{n}{\lfloor n/4 \rfloor} = 2^{\Theta(n)}$*, satisfying the following conditions.*

1) *There is a distinguisher* A *which given* one *sample* $x \sim A_j$ *for an unknown* $j \in \{0, 1, \ldots, m\}$*, recovers* j *with probability at least* $2/3$*. In other words, for all* $j = 0, 1, \ldots, m$*,*

$$\Pr_{x \sim D_j}[A(x) = j] \geq 2/3.$$

2) MLE *fails to distinguish* D_0 *from other distributions with probability 1, even with* $T = n/4$ *samples. In other words,*

$$\Pr_{x_1, \ldots, x_T \sim D_0}[\mathrm{MLE}(x_1, \ldots, x_T; \mathcal{D}) = 0] = 0.$$

Remark 1. Finally, we remark that in the average-case setting MLE is indeed optimal (with no factor of $O(n)$ factor blowup in the number of traces). This is because maximizing the likelihood is equivalent to maximizing the posterior probability under the uniform prior distribution (which is optimal), as can be seen via the Bayes rule $\mathcal{D}_{\mathbf{x}}(\widetilde{x}_1, \ldots, \widetilde{x}_T) = p(\mathbf{x} \mid \widetilde{x}_1, \ldots, \widetilde{x}_T) \cdot \frac{\sum_{\mathbf{x}' \in \{0,1\}^n} p(\mathbf{x}') \cdot \mathcal{D}_{\mathbf{x}'}(\widetilde{x}_1, \ldots, \widetilde{x}_T)}{p(\mathbf{x})} = p(\mathbf{x} \mid \widetilde{x}_1, \ldots, \widetilde{x}_T) \cdot \sum_{\mathbf{x}' \in \{0,1\}^n} \mathcal{D}_{\mathbf{x}'}(\widetilde{x}_1, \ldots, \widetilde{x}_T) = p(\mathbf{x} \mid \widetilde{x}_1, \ldots, \widetilde{x}_T) \cdot f(\widetilde{x}_1, \ldots, \widetilde{x}_T)$. Therefore maximizing both sides with respect to $\mathbf{x}$ yields the same result.

B. Overview of the techniques

a) Lower bounds for k*-mer-based algorithms:* In recent development of the trace reconstruction problem, the connection to various real and complex polynomials has been a recurring and intriguing theme [2]–[4], [9], [11], [13], [19], [22]–[24], [27]. The starting point of these techniques is to design a set of statistics that can be easily estimated from the traces (e.g., mean traces), with the property that for different source strings the corresponding statistics are somewhat "far apart". To establish this property, one key idea is to associate each source string $\mathbf{x}$ with a generating polynomial $P_{\mathbf{x}}$ where the coefficients are exactly the statistics of $\mathbf{x}$.

Definition 4 (k-mer generating polynomial)**.** *Let* $\mathbf{x} \in \{0,1\}^n$ *and* $w \in \{0,1\}^k$*. The* k-mer generating polynomial $P_{w,\mathbf{x}}$ *for string* $\mathbf{x}$ *and* k*-mer* w *is the following degree-*$(n-1)$ *polynomial in* z*:* $P_{w,\mathbf{x}}(z) := \sum_{\ell=0}^{n-1} K_{w,\mathbf{x}}[\ell] \cdot z^{\ell}$.

Due to the structure of the deletion channel, in many cases, this generating polynomial (under a change of variables) is identical to another polynomial $Q_{\mathbf{x}}$ that is much easier to get a handle on. Formally, it can be shown that $P_{w,\mathbf{x}}(z) = \sum_{\ell=0}^{n-1} K_{w,\mathbf{x}}[\ell] \cdot z^{\ell} = \sum_{j=0}^{n-k} \mathbf{1}\{\mathbf{x}[j\colon j+k-1] = w\} \cdot \big(p + (1-p)z\big)^j$, which, under change of variable $z_0 = p + (1-p)z$ gives exactly the polynomial $Q_{\mathbf{x}}$ studied in [2].

For example, the coefficients of $Q_{\mathbf{x}}$ are usually 0/1, and they are easily determined from $\mathbf{x}$. To show that the statistics corresponding to $\mathbf{x}$ and $\mathbf{y}$ are far apart (say, in ℓ_1-distance), it is sufficient to show that $\big|Q_{\mathbf{x}}(w) - Q_{\mathbf{y}}(w)\big|$ is large for an appropriate choice of w. This is the point where all sorts of analytical tools are ready to shine. For instance, the main technical result in [2] is a complex analytical result that says that a certain family of polynomials cannot be uniformly small over a sub-arc of the complex unit circle, which has applications beyond the trace reconstruction problem.

This analytical view of trace reconstruction can lead to a tight analysis of certain algorithms/statistics. The best example would be mean-based algorithms, for which a tight bound of $\exp(\Theta(n^{1/3}))$ traces is known to be sufficient and necessary for worst-case trace reconstruction [3], [4]. The tightness of the sample complexity is exactly due to the tightness of a complex analytical result by Borwein and Erdélyi [28]. Our lower bound for k-mer-based algorithms is obtained in a similar fashion, via establishing a complex analytical result complementary to that of [2] (See Lemma II.1).

On the other hand, our argument takes a different approach than that of [28]. At a high level, both results use a Pigeonhole argument to show the existence of two univariate polynomials which are uniformly close over a sub-arc Γ of the complex unit circle. The difference lies in the objects playing the role of "pigeons". [28]'s argument can be viewed as two steps: (1) apply the Pigeonhole Principle to obtain two polynomials that have close evaluations over a *discrete* set of points in Γ, and (2) use a continuity argument to extend the closeness to the entire sub-arc. Here the roles of pigeons and holes are played by evaluation vectors, and Cartesian products of small intervals. Our approach considers the coordinates of a related polynomial in the *Chebyshev basis*, which play the roles of pigeons in place of the evaluation vector. The properties of Chebyshev polynomials allow us to get rid of the continuity argument. Instead, we complete the proof by leveraging rather standard tools from complex analysis (e.g., bounds on the Chebyshev coefficients and Hadamard Three Circle Theorem) We believe this approach has the advantage of being generalizable to multivariate polynomials over the product of sub-arcs $\Gamma = \Gamma_1 \times \cdots \times \Gamma_m$ via multivariate Chebyshev series (see, e.g., [29], [30]), whereas the same generalization

seems to be tricky for the continuity argument.

Finally, the counting argument considers a special set of strings for which effectively only one k-mer contains meaningful information about the initial string. Since previous arguments did not exploit structural properties of the strings, this is another technical novelty of our proof.

b) Maximum Likelihood Estimation: Most of our results regarding Maximum Likelihood Estimation hold under the more general "model estimation" setting, where one is given a sample x drawn from an unknown distribution $D \in \mathcal{D}$ and tries to recover D. Our main observation is that if such a distinguisher works in worst-case, then the distributions in $\mathcal{D}$ have large pairwise statistical distances. The maximization characterization of statistical distance, in conjunction with a union bound, implies that for a sample $x \sim D$ its likelihood is maximized by D except with a small probability. The $O(n)$ factor loss in the sample complexity is essentially due to the union bound, and we show that this loss is tight in general by constructing a set of distributions which attains equality in the union bound.

C. Organization

In Section II we prove our main result Theorem 2. The missing proofs and further related work appear in the full version of the paper https://arxiv.org/abs/2308.14993.

II. A lower bound for k-mer based algorithms: Proof of Theorem 2

The proof of Theorem 2 relies on the next lemma.

Lemma II.1. *There exists* $\mathbf{x}, \mathbf{y} \{0,1\}^n$ *such that for any* k*-mer* w*, it holds that* $\sup_{z\colon |z|=1} \left|P_{w,\mathbf{x}}(z) - P_{w,\mathbf{y}}(z)\right| \le 2^{-cn^{1/5}\sqrt{\log n}}$.

Proof of Theorem 2 using Lemma II.1. We can extract $K_{w,\mathbf{x}}[\ell] - K_{w,\mathbf{y}}[\ell]$ by the contour integral (cf. [31, §4, Theorem 2.1]) $K_{w,\mathbf{x}}[\ell] - K_{w,\mathbf{y}}[\ell] = \frac{1}{2\pi i}\int_{|z|=1}\left(P_{w,\mathbf{x}}(z) - P_{w,\mathbf{y}}(z)\right)\cdot z^{-\ell-1}\,\mathrm{d}z$. Therefore $\left|K_{w,\mathbf{x}}[\ell] - K_{w,\mathbf{y}}[\ell]\right| \le \frac{1}{2\pi}\int_{|z|=1}\left|P_{w,\mathbf{x}}(z) - P_{w,\mathbf{y}}(z)\right| \cdot |z|^{-\ell-1}\cdot|\mathrm{d}z| \le 2^{-cn^{1/5}\sqrt{\log n}}$. We stress that the bound holds for any $\ell \in [n]$ and k-mer w. Note that for any fixed ℓ, there are at most $n-k+1$ different k-mers w for which $K_{w,\mathbf{x}}[\ell] > 0$. Namely, if $w \notin \{x[j\colon j+k-1]\colon 0 \le j \le n-k\}$ then $K_{w,\mathbf{x}}[\ell] = 0$. It follows that $\left\|K_{\mathbf{x}} - K_{\mathbf{y}}\right\|_{\ell_1} = \sum_{\ell=0}^{n-1}\sum_{w}\left|K_{w,\mathbf{x}}[\ell] - K_{w,\mathbf{y}}[\ell]\right| \le n\cdot 2(n-k+1)\cdot 2^{-cn^{1/5}\log^{2/5} n} \le 2^{-c'n^{1/5}\sqrt{\log n}}$. □

Next, we prove Lemma II.1 assuming the following result, which is our main technical lemma.

Lemma II.2. *Fix any* $k \le L^{1/3}$*. There exist distinct* $\mathbf{x}, \mathbf{y} \in \{0,1\}^L$ *both starting with a run of 0s of length* $L^{1/3}-1$*, such that for any* k*-mer* w*, it holds that* $\sup_{\theta\colon |\theta| \le L^{-2/3}\log^{1/4} L}\left|P_{w,\mathbf{x}}(e^{i\theta}) - P_{w,\mathbf{y}}(e^{i\theta})\right| \le 2^{-L^{1/3}\sqrt{\log L}/20}$.

Proof of Lemma II.1 using Lemma II.2. Let $\beta \ge 3/5$ be a parameter to be decided later. Denote $L := n^{\beta}$. We have $k \le n^{1/5} = L^{1/(5\beta)} \le L^{1/3}$, so that the premise of Lemma II.2 is satisfied. Therefore, there exist distinct $\mathbf{x}', \mathbf{y}' \in \{0,1\}^L$ both starting with a run of 0s of length $L^{1/3}-1$, such that for any k-mer w, it holds that $\sup_{\theta\colon |\theta| \le L^{-2/3}\log^{1/4} L}\left|P_{w,\mathbf{x}'}(e^{i\theta}) - P_{w,\mathbf{y}'}(e^{i\theta})\right| \le 2^{-L^{1/3}\sqrt{\log L}/20}$.

Let $\mathbf{x} = 0^{n-L}\mathbf{x}'$ and $\mathbf{y} = 0^{n-L}\mathbf{y}'$. Since $k \le L^{1/3}$, by construction we have $\mathbf{x}[j\colon j+k-1] = \mathbf{y}[j\colon j+k-1]$ for all $0 \le j \le n-L$. Therefore, for any k-mer w we have

$$P_{w,\mathbf{x}}(e^{i\theta}) - P_{w,\mathbf{y}}(e^{i\theta}) = \sum_{j=0}^{n-k}\left(\mathbf{1}\{\mathbf{x}[j\colon j-k+1] = w\} - \mathbf{1}\{\mathbf{y}[j\colon j-k+1] = w\}\right)\cdot$$

$$(p+qe^{i\theta})^j = \left(p+qe^{i\theta}\right)^{n-L}\cdot$$
$$\cdot\sum_{j=n-L}^{n-k}\left(\mathbf{1}\{\mathbf{x}[j\colon j-k+1] = w\} - \mathbf{1}\{\mathbf{y}[j\colon j-k+1] = w\}\right)\cdot$$

$$(p+qe^{i\theta})^{j-(n-L)} = \left(p+qe^{i\theta}\right)^{n-L}\cdot$$
$$\sum_{j=0}^{L-k}\left(\mathbf{1}\{\mathbf{x}'[j\colon j-k+1] = w\} - \mathbf{1}\{\mathbf{y}'[j\colon j-k+1] = w\}\right)\cdot$$

$$(p+qe^{i\theta})^j = \left(p+qe^{i\theta}\right)^{n-L}\left(P_{w,\mathbf{x}'}(e^{i\theta}) - P_{w,\mathbf{y}'}(e^{i\theta})\right).$$

Here $q = 1-p$. When $|\theta|$ is large, we can upper bound the supremum as $\sup_{\theta\colon |\theta| > L^{-2/3}\log^{1/4} L}\left|P_{w,\mathbf{x}}(e^{i\theta}) - P_{w,\mathbf{y}}(e^{i\theta})\right| = \sup_{\theta\colon |\theta| > L^{-2/3}\log^{1/4} L}\left|p+qe^{i\theta}\right|^{n-L}\left|P_{w,\mathbf{x}'}(e^{i\theta}) - P_{w,\mathbf{y}'}(e^{i\theta})\right| \le \left(1 - c_1 L^{-4/3}\log^{1/2} L\right)^{n-L}\cdot$
$\cdot\sup_{\theta\colon |\theta| > L^{-2/3}\log^{1/4} L}\left|P_{w,\mathbf{x}'}(e^{i\theta}) - P_{w,\mathbf{y}'}(e^{i\theta})\right| \le \exp\left(-c_1(n-L)L^{-4/3}\log^{1/2} L\right)\cdot(L-k+1) \le \exp_2\left(-c_2 n^{1-4\beta/3}\log^{1/2} n\right)$. Here the first inequality is due to $\left|p+qe^{i\theta}\right| \le 1 - c_1 a^2$ for some constant c_1 (depending on p) when $|\theta| \ge a$. When $|\theta|$ is small, as above $\sup_{\theta\colon |\theta| \le L^{-2/3}\log^{1/4} L}\left|P_{w,\mathbf{x}}(e^{i\theta}) - P_{w,\mathbf{y}}(e^{i\theta})\right| \le \sup_{\theta\colon |\theta| \le L^{-2/3}\log^{1/4} L}\left|P_{w,\mathbf{x}'}(e^{i\theta}) - P_{w,\mathbf{y}'}(e^{i\theta})\right| \le \exp_2\left(-L^{1/3}\sqrt{\log L}/20\right) \le \exp_2\left(-c_3 n^{\beta/3}\log^{1/2} n\right)$. Finally, the value of β is determined by balancing the two cases. Namely, we let $1-4\beta/3 = \beta/3$, or $\beta = 3/5$, which gives the bound $2^{-cn^{1/5}\sqrt{\log n}}$ for both cases. Here $c = \min\{c_2, c_3\}$. □

REFERENCES

[1] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11-14, 2004*, J. I. Munro, Ed. SIAM, 2004, pp. 910–918.

[2] Z. Chase, "Separating words and trace reconstruction," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2021*. ACM, 2021, pp. 21–31.

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

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

[5] N. Holden and R. Lyons, "Lower bounds for trace reconstruction," *The Annals of Applied Probability*, vol. 30, no. 2, pp. 503–525, 2020.

[6] Z. Chase, "New lower bounds for trace reconstruction," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 57, no. 2. Institut Henri Poincaré, 2021, pp. 627–643.

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

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

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

[10] A. McGregor, E. Price, and S. Vorotnikova, "Trace reconstruction revisited," in *22th Annual European Symposium on Algorithms, ESA 2014*, ser. Lecture Notes in Computer Science, vol. 8737. Springer, 2014, pp. 689–700.

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

[12] R. Gabrys and O. Milenkovic, "The hybrid k-deck problem: Reconstructing sequences from short and long traces," in *IEEE International Symposium on Information Theory, ISIT 2017*. IEEE, 2017, pp. 1306–1310.

[13] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability," in *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet, and P. Rigollet, Eds., vol. 75. PMLR, 2018, pp. 1799–1840.

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

[15] R. Gabrys and O. Milenkovic, "Unique reconstruction of coded strings from multiset substring spectra," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 7682–7696, 2019.

[16] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6084–6103, 2020.

[17] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Trace reconstruction: Generalized and parameterized," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3233–3250, 2021.

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

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

[20] Z. Chase and Y. Peres, "Approximate trace reconstruction of random strings from a constant number of traces," *arXiv preprint arXiv:2107.06454*, 2021.

[21] S. Narayanan and M. Ren, "Circular trace reconstruction," in *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[22] J. Sima and J. Bruck, "Trace reconstruction with bounded edit distance," in *IEEE International Symposium on Information Theory, ISIT 2021*. IEEE, 2021, pp. 2519–2524.

[23] E. Grigorescu, M. Sudan, and M. Zhu, "Limitations of mean-based algorithms for trace reconstruction at small edit distance," *IEEE Trans. Inf. Theory*, vol. 68, no. 10, pp. 6790–6801, 2022. [Online]. Available: https://doi.org/10.1109/TIT.2022.3168624

[24] I. Rubinstein, "Average-case to (shifted) worst-case reduction for the trace reconstruction problem," in *50th International Colloquium on Automata, Languages, and Programming, ICALP 2023, July 10-14, 2023, Paderborn, Germany*, ser. LIPIcs, K. Etessami, U. Feige, and G. Puppis, Eds., vol. 261. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023, pp. 102:1–102:20. [Online]. Available: https://doi.org/10.4230/LIPIcs.ICALP.2023.102

[25] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free dna-based data storage," *Scientific Reports*, vol. 7, pp. 2045–2322, 2017.

[26] K. Mazooji and I. Shomorony, "Substring density estimation from traces," *CoRR*, vol. abs/2210.10917, 2022.

[27] M. Cheraghchi, J. Downs, J. L. Ribeiro, and A. Veliche, "Mean-based trace reconstruction over practically any replication-insertion channel," in *IEEE International Symposium on Information Theory, ISIT 2021*. IEEE, 2021, pp. 2459–2464.

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

[29] J. C. Mason, "Near-best multivariate approximation by fourier series, chebyshev series and chebyshev interpolation," *Journal of Approximation Theory*, vol. 28, no. 4, pp. 349–358, 1980.

[30] L. Trefethen, "Multivariate polynomial approximation in the hypercube," *Proceedings of the American Mathematical Society*, vol. 145, no. 11, pp. 4837–4844, 2017.

[31] S. Lang, *Complex analysis*. Springer Science & Business Media, 2013, vol. 103.