Trace reconstruction with varying deletion probabilities

Lisa Hartung Nina Holden Yuval Peres

Abstract

In the trace reconstruction problem an unknown string $\mathbf{x} = (x_0, \dots, x_{n-1}) \in \{0, 1, ..., m-1\}^n$ is observed through the deletion channel, which deletes each x_k with a certain probability, yielding a contracted string $\widetilde{\mathbf{X}}$. Earlier works have proved that if each x_k is deleted with the same probability $q \in [0, 1)$, then $\exp(O(n^{1/3}))$ independent copies of the contracted string $\widetilde{\mathbf{X}}$ suffice to reconstruct $\mathbf{x}$ with high probability. We extend this upper bound to the setting where the deletion probabilities vary, assuming certain regularity conditions. First we consider the case where x_k is deleted with some known probability q_k. Then we consider the case where each letter $\zeta \in \{0, 1, ..., m-1\}$ is associated with some possibly unknown deletion probability q_ζ.

1 Introduction

Let $n \in \mathbb{N}$, $m \in \{2, 3, \dots\}$, and $[m] = \{0, \dots, m-1\}$. In trace reconstruction the goal is to reconstruct an unknown string $\mathbf{x} = (x_0, \dots, x_{n-1}) \in [m]^n$ from noisy observations of $\mathbf{x}$. Here we study the case where the data is noisy due to a deletion channel in which each bit is deleted independently with a certain probability. In other words, instead of observing $\mathbf{x}$ we observe many independent strings $\widetilde{\mathbf{X}}$ obtained by sending $\mathbf{x}$ through the deletion channel. The probability of deleting x_k might depend on either (I) the location k of x_k in the string, or (II) the letter x_k.

In Case (I) and given $p_k \in (0, 1]$ for $k \in [n]$, the string $\widetilde{\mathbf{X}}$ is obtained in the following way for $k = 0, 1, \dots, n-1$, starting from an empty string.

- (retention) With probability p_k, copy x_k to the end of $\widetilde{\mathbf{X}}$ and increase k by one.
- (deletion) With probability $q_k = 1 - p_k$, increase k by one.

In Case (II) and given $p_\zeta \in (0, 1]$ for $\zeta \in [m]$, the string $\widetilde{\mathbf{X}}$ is obtained by performing the following steps for $k = 0, 1, \dots, n-1$.

- (retention) With probability p_{x_k}, copy x_k to the end of $\widetilde{\mathbf{X}}$ and increase k by one.
- (deletion) With probability $q_{x_k} = 1 - p_{x_k}$, increase k by one.

In Case (I) we assume the deletion probabilities are known, while we do not make this assumption in Case (II).

For $T \in \mathbb{N}$ we consider T independent outputs $\widetilde{\mathbf{X}}^{(1)}, \dots, \widetilde{\mathbf{X}}^{(T)}$ (called "traces") from the deletion channel. Our main question is the following. Given $\varepsilon > 0$, how many samples are needed, such that for some $\widehat{\mathbf{X}} \in [m]^n$ we have $\mathbb{P}[\widehat{\mathbf{X}} = \mathbf{x}] > 1 - \varepsilon$? Holenstein, Mitzenmacher, Panigrahy, and Wieder [5] proved that $\exp\left(n^{1/2} \operatorname{polylog}(n)\right)$ traces are sufficient in the case where all x_k are deleted independently with the same probability $q \in [0, 1)$. See [6] for an alternative proof. This result was recently improved by De, O'Donnell, and Servedio [3], and by Nazarov and Peres [8]. Using single bit statistics for the traces, they proved that reconstruction is possible with $\exp\left(O(n^{1/3})\right)$ traces, and that this is optimal for reconstruction techniques using only single bit statistics. The case of random strings was studied in [1] and [9].

1.1 Main result We study the following two settings in Case (I) where the deletion probabilities depend on the location in the string.

(i) (weak monotonicity) For some $\delta \in (0, 1)$ the retention probabilities $(p_k)_{k \in \mathbb{N}}$ satisfy $p_\ell > \frac{p_k}{2} + \delta$ for all $k > \ell$, and $p_k > \delta$ for all $k > 0$.

(ii) (periodicity) The deletion probabilities are 2-periodic, meaning that

$$q_k = q \text{ for } k \text{ even}, \quad q_k = \widetilde{q} \text{ for } k \text{ odd}. \tag{1.1}$$

In particular, (i) covers the case where the retention probabilities are monotonically decreasing and bounded away from zero, and the case where all retention probabilities are in some interval $[p + \delta, 2p]$ for $p \in (0, 1/2]$ and $\delta > 0$. Also observe that, by reversing the sequence, we can also study strings where the deletion probabilities satisfy $p_\ell > \frac{p_j}{2} + \delta$ for all $\ell > j$ (instead of $\ell < j$).

THEOREM 1.1. *Let $\varepsilon > 0$ and $m \in \{2, 3, \dots\}$, and let the deletion probabilities be known and satisfy Assumption (i) or (ii). There exists a constant $C > 0$ depending only on $\varepsilon, m,$ and δ (for Assumption (i)), and $\varepsilon, m, q,$ and $\widetilde{q}$ (for Assumption (ii)), such that the original string $\mathbf{x} \in [m]^n$, can be reconstructed with probability at least $1 - \varepsilon$ from $T = \lceil \exp\left(Cn^{1/3}\right) \rceil$ i.i.d. samples of the deletion channel applied to $\mathbf{x}$.*

For case (II), where the deletion probabilities vary by letter, we prove the following.

THEOREM 1.2. *For any $\delta, \varepsilon > 0$ and $m \in \{2, 3, \dots\}$ there is a constant $C > 0$ depending only on δ, m, and δ, such that if $\min_{\zeta \in [m]} p_\zeta \geq \delta$, then the string $\mathbf{x} \in [m]^n$ can be reconstructed with probability at least $1 - \varepsilon$ from $T = \lceil \exp\left(Cn^{1/3}\right) \rceil$ i.i.d. samples of the deletion channel applied to $\mathbf{x}$. This result holds also in the case when the deletion probabilities are unknown.*

The deletion channel (or variants of it which also allow insertions, substitutions, swaps, etc.) is relevant for the study of mutations and DNA sequencing errors. In the case of mutations several studies have revealed that mutation probabilities vary by location in the genome [4, 7, 10, 11].

The weak monotonicity assumption is in the context of applications a very natural one. But we strongly believe that the result and the techniques used should in principal allow to treat a far more general case. As a strong indication we consider the 2-periodic case which is in some sense the furthest away from the weak monotonicity setting.

Outline of the paper: To prove Theorem 1.1 we will first derive in Section 2.1 an exact formula expressing the single bit statistics of $\widetilde{\mathbf{X}}$ as the coefficients of a particular polynomial. In Sections 2.2 and 2.3 we prove that this polynomial cannot be too small everywhere on a particular boundary arc of the unit disk $\mathbb{D}$. In Section 3 we use this to prove that for any two input strings, at least one bit in the trace has a sufficiently different expectation so we are able distinguish the two strings. Theorem 1.2 will be proved by first using the traces to obtain good estimates for the deletion probabilities, and then use a single bit test (with the estimated probabilities) for the traces sent through a second deletion channel.

Acknowledgements: We thank Ryan O'Donnell for suggesting the problem of varying deletion probabilities. Most of this work was done while the first and second author were visiting Microsoft Research, and they want to thank Microsoft for the hospitality.

2 Preparatory Lemmas

2.1 A polynomial identity

For $\mathbf{a} \in \mathbb{R}^n$ define

$$\Psi(w) = \sum_{k=0}^{n-1} a_k p_k \prod_{\ell=0}^{k-1} (p_\ell w + q_\ell). \tag{2.2}$$

LEMMA 2.1. *Let $\mathbf{a} = (a_0, a_1, \dots, a_{n-1}) \in \mathbb{R}^n$ and let $\widetilde{\mathbf{a}}$ be the output of the deletion channel with input $\mathbf{a}$, padded with an infinite sequence of 0's on the right side. Then*

$$\mathbb{E}\left(\sum_{j \geq 0} \widetilde{a}_j w^j\right) = \Psi(w). \tag{2.3}$$

Proof. The expectation on the left side of (2.3) can be written as

$$\sum_{k=0}^{n-1} a_k p_k \sum_{j=0}^{k-1} w^j \mathbb{P}\left(\sum_{\ell=0}^{k-1} B_\ell = j\right), \tag{2.4}$$

where B_ℓ for $\ell = 0, \dots, n-1$ are independent Bernoulli(p_ℓ)-distributed random variables. Observe that

$$\begin{aligned} \sum_{j=0}^{k-1} w^j \mathbb{P}\left(\sum_{\ell=0}^{k-1} B_\ell = j\right) &= \mathbb{E}\left(w^{\sum_{\ell=0}^{k-1} B_\ell}\right) \\ &= \prod_{\ell=0}^{k-1} (q_\ell + p_\ell w), \end{aligned} \tag{2.5}$$

where we used independence in the last step. Plugging (2.5) into (2.4) concludes the proof of Lemma 2.1.

Next, we will establish that (perhaps after a change of variables) the function $|\Psi(w)|$ on a small arc of the unit disc is not to small. We use tools from standard complex analysis (inspired by [2]).

2.2 The deletion probabilities satisfy a monotonicity property

LEMMA 2.2. *Assume that $a_0 = 1$, $a_k \in [-1, 1]$ for $k \geq 1$, and that there is some $\delta \in (0, 1/10)$ such*

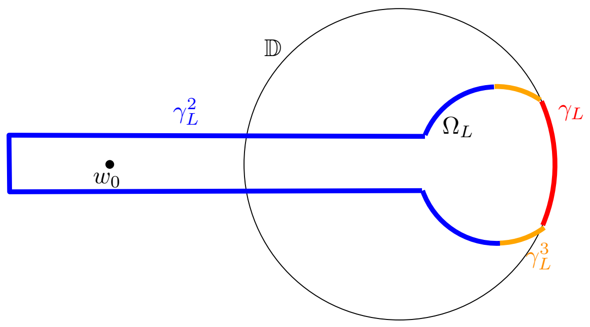

Figure 1: Illustration of objects defined in the proof of Lemma 2.2. The domain Ω_L is bounded by the colored curves.

that $p_0 > \frac{p_k}{2} + \delta$ *and* $p_k > \delta$ *for all* $k > 0$. *Define the arc* $\gamma_L = \{z = e^{i\theta}, \frac{-\pi}{L} \le \theta \le \frac{\pi}{L}\}$. *Then there is a constant* $c > 0$ *depending only on* δ *such that* $\max_{\gamma_L} |\Psi(w)| > e^{-cL}$ *for all* $L \ge 1$.

Proof. Throughout the proof c is a positive constant which may depend on δ and which may change line by line. Define $w_0 := -q_0/p_0$. Let Ω' be a rectangular domain which contains w_0 and $1/2$, has sides parallel to the coordinate axes, and which is such that $|pw+(1-p)| < 1-\delta/2$ for all $p \in \{p_0, p_1, \ldots\}$ and $w \in \Omega'$. Observe that an appropriate domain exists since the required inequality is satisfied for all $w \in [w_0, 1/2]$. Furthermore, observe that we may assume Ω' changes continuously as we vary w_0. Let $\Omega = \Omega' \cup B_{1/2}(1/2)$. For $L > 100$, let Ω'_L be the horizontal translation of Ω by some $t_L \in (0, 1/10)$, such that γ_L is the intersection of Ω'_L and the right half of $\partial\mathbb{D}$. There is an $L_0 > 0$ depending only on δ such that $w_0 \in \Omega'_L$ for all $L > L_0$, and it is sufficient to prove the result of the lemma for $L > L_0$. Let $\Omega_L = \Omega'_L \cap (\mathbb{D} \cup \{z : \mathrm{Re}(z) < 0\})$. Then $\partial\Omega_L$ is the union of the three disjoint sets γ_L, γ_L^2, and γ_L^3, where

$$\gamma_L^2 := \{z \in \partial\Omega_L \ : \ \mathrm{Re}(z) \le 1/2 + t_L\},$$
$$\gamma_L^3 := \{z \in \partial\Omega_L \setminus \gamma_L \ : \ \mathrm{Re}(z) > 1/2 + t_L\}.$$

See Figure 1 for an illustration of $\partial\Omega_L$. Since $\log|\Psi|$ is subharmonic, and letting $\mu_{\Omega_L}^{w_0}$ denote harmonic measure of Ω_L relative to w_0,

$$\log|\Psi(w_0)| \le \int_{\gamma_L} \log|\Psi(z)| d\mu_{\Omega_L}^{w_0}(z) + \int_{\gamma_L^2} \log|\Psi(z)| d\mu_{\Omega_L}^{w_0}(z) + \int_{\gamma_L^3} \log|\Psi(z)| d\mu_{\Omega_L}^{w_0}(z). \tag{2.6}$$

Without loss of generality we assume $|\Psi(z)| < 1$ everywhere on γ_L. For each fixed w_0 there is a constant c_{w_0} such that $\mu_{\Omega_L}^{w_0}(\gamma_L) \ge c_{w_0}/L$. Since we assume Ω' varies continuously as we vary $w_0 \in [1-1/\delta, 0]$, we can find a $c' > 0$ such that $c_{w_0} > c'$ for all w_0. This gives

$$\int_{\gamma_L} \log|\Psi(z)| d\mu_{\Omega_L}^{w_0}(z) \le \mu_{\Omega_L}^{w_0}(\gamma_L) \cdot \log \max_{\gamma_L} |\Psi(w)| \le \frac{c'}{L} \log \max_{\gamma_L} |\Psi(w)|.$$

Observe that $\log|\Psi(z)| < c$ for any $z \in \partial\Omega_L \cap \{z : \mathrm{Re}(z) < 1/2 + t_L\}$, so

$$\int_{\gamma_L^2} \log|\Psi(z)| d\mu_{\Omega_L}^{w_0}(z) \le c.$$

For any $w \in \mathbb{D}$,

$$|\Psi(w)| \le \sum_{k=0}^{n-1} p_k \prod_{\ell=0}^{k-1} |q_\ell + p_\ell w| \le \sum_{k=0}^{n-1} \prod_{\ell=0}^{k-1} (q_k + p_k|w|) \le \sum_{k=0}^{n-1} (1 - \delta + \delta|w|)^k \le \frac{1}{\delta(1-|w|)}. \tag{2.7}$$

It follows that

$$\int_{\gamma_L^3} \log|\Psi(z)| d\mu_{\Omega_L}^{w_0}(z) \le \log\left(\frac{1}{\delta}\right) + \int_{\gamma_L^3} \log \frac{1}{1-|z|} d\mu_{\Omega_L}^{w_0}(z). \tag{2.8}$$

A Brownian motion started at w_0 must hit the line segment $\ell_L := \{z \in \Omega_L : \mathrm{Re}(z) = 1/4 + t_L\}$ before hitting γ_L^3. Therefore, defining $D_L = B_{1/2}(1/2 + t_L)$, the integral on the right side of (2.8) is bounded above by

$$c \sup_{w \in \ell_L} \int_{\gamma_L^3} \log \frac{1}{1-|z|} d\mu_{D_L}^{w}(z). \tag{2.9}$$

Since ℓ_L has distance $> 1/4$ from γ_L^3, the density of $\mu_{D_L}^w$ on γ_L^3 is bounded above by a universal constant. For any $z \in \gamma_L^3$ in the first quadrant we have $1 - |z| \geq c^{-1}(\theta - \theta_0)^2$, where $\theta \in (-\pi, \pi]$ is the angle for polar coordinates with origin at $1/2 + t_L$ and $\theta_0 \in [0, \pi/2]$ corresponds to the point of intersection between ∂D_L and $\partial \mathbb{D}$ in the first quadrant (see [8, equation (4.4)]). Therefore we can bound (2.9) from above by

$$c \int_0^{\pi/2} \log \left| \frac{1}{\theta} \right| d\theta < \infty. \tag{2.10}$$

We conclude that

$$\int_{\gamma_L^3} \log |\Psi(z)| d\mu_{\Omega_L}^{w_0}(z) < c. \tag{2.11}$$

Inserting the above estimates into (2.6), we get

$$\log |p_0| = \log |\Psi(w_0)| \leq c + \frac{c'}{L} \log \max_{\gamma_L} |\Psi(w)|,$$

which implies the lemma.

2.3 The 2-periodic case

Assume without loss of generality that $p < \widetilde{p}$. If $z = \widetilde{p}w + \widetilde{q}$ then

$$\Psi(w) = \widetilde{\Psi}(z) := \sum_{k=0}^{n-1} a_k p_k \prod_{\ell=0}^{k-1} (p_\ell \frac{z - \widetilde{q}}{\widetilde{p}} + q_\ell). \tag{2.12}$$

LEMMA 2.3. *Assume that $a_0 = 1$, $a_k \in \{-1, 0, 1\}$ for $k \geq 1$, and that $(p_\ell)_\ell$ is 2-periodic. If $|\widetilde{\Psi}(z)| < \beta$ everywhere on the the arc $\gamma_L = \{z = e^{i\theta}, \frac{-\pi}{L} \leq \theta \leq \frac{\pi}{L}\}$ then there exists c depending on q and $\widetilde{q}$ such that $\beta > e^{-cL}$ for all $L \geq 1$.*

Proof. We consider the following cases.

Case 1) $p_0 = \widetilde{p}$. Then (2.12) gives

$$\widetilde{\Psi}(z) = a_0 \widetilde{p} + a_1 p z + a_2 \widetilde{p} z \left(p \frac{z - \widetilde{q}}{\widetilde{p}} + q \right) + \dots. \tag{2.13}$$

Setting $z = 0$ we get $\widetilde{\Psi}(0) = \widetilde{p} > 0$.

Case 2) $p_0 = p$. Then (2.12) gives

$$\begin{aligned} \widetilde{\Psi}(z) &= a_0 p + a_1 \widetilde{p} \left(p \frac{z - \widetilde{q}}{\widetilde{p}} + q \right) \\ &\quad + a_2 p z \left(p \frac{z - \widetilde{q}}{\widetilde{p}} + q \right) + \dots. \end{aligned} \tag{2.14}$$

Setting again $z = 0$ gives

$$\widetilde{\Psi}(0) = a_0 p + a_1 \widetilde{p} \left(q - \frac{p\widetilde{q}}{\widetilde{p}} \right) = p + a_1(\widetilde{p} - p). \tag{2.15}$$

The expression in (2.15) can take three values

$$\widetilde{\Psi}(0) = \begin{cases} p & \text{if } a_1 = 0, \\ \widetilde{p} & \text{if } a_1 = 1, \\ -\widetilde{p} + 2p & \text{if } a_1 = -1. \end{cases} \tag{2.16}$$

Case 2.1) If $\widetilde{p} \neq 2p$, then we see by (2.16) that $|\Psi(0)|$ is bounded from below by a constant depending only on p and $\widetilde{p}$.

Case 2.2) $p_0 = p$ and $\widetilde{p} = 2p$. Observe that in this particular case

$$p \frac{z - \widetilde{q}}{\widetilde{p}} + q = \frac{1}{2} z + \frac{1}{2}. \tag{2.17}$$

Hence inserting $z = -1$ into (2.14) gives

$$\widetilde{\Psi}(-1) = a_0 p > 0. \tag{2.18}$$

Moreover, $\widetilde{\Psi}$ is continuous in some neighborhood of $z = -1$ (uniformly in $a_0, a_1, \dots$). Hence, there exists $\delta_1 \in (0, 1)$ depending only on p such that $|\widetilde{\Psi}(1 - \delta_1)|$ is bounded away from 0.

Again $|\widetilde{\Psi}(z)|$ for $|z| \leq 1$ is bounded in terms of a geometric series. Set $p^* = \frac{p}{\widetilde{p}}$ and $q^* = 1 - p^*$. Then

$$|\widetilde{\Psi}(z)| \leq \frac{1}{p^*(1 - |z|)}. \tag{2.19}$$

Now the claim of Lemma 2.3 follows by letting Ω be a ball contained in $\mathbb{D}$ such that $\partial\Omega$ is tangent to $\partial\mathbb{D}$ at 1 and $-1 + \delta_1 \in \Omega$, and then proceeding as in Lemma 2.2 using the subharmonic function $\log |\widetilde{\Psi}|$.

3 Proof of main result: Deletion probabilities varying with position

Proof. [Proof of Theorem 1.1] Throughout the proof all constants should depend only on δ and m. Let $\mathbf{x}, \mathbf{y} \in [m]^n$ be two different strings. We first consider the case $m = 2$. Let $\mathbf{a} = \mathbf{x} - \mathbf{y}$. By Lemma 2.1,

$$\sum_{j=0}^{n-1} \mathbb{E} \left(\widetilde{X}_j - \widetilde{Y}_j \right) w^j = \sum_{k=0}^{n-1} a_k p_k \prod_{\ell=0}^{k-1} (q_\ell + p_\ell w). \tag{3.20}$$

Let $k^* = \min\{k : a_k \neq 0\}$. Then the right side of (3.20) is equal to

$$\begin{aligned} &\prod_{\ell < k^*} (q_\ell + p_\ell w) \Bigg(p_{k^*} a_{k^*} \\ &\qquad + \sum_{k=k^*+1}^{n-1} a_k p_k \prod_{\ell=k^*}^{k-1} (q_\ell + p_\ell w) \Bigg). \end{aligned} \tag{3.21}$$

Let $L \in \mathbb{N}$. We first consider deletion probabilities satisfying Assumption (i). First, we verify that for $w \in \gamma_L$

$$\left|\prod_{\ell < k^*} (q_\ell + p_\ell w)\right| \geq \exp(-nC_2/L^2), \tag{3.22}$$

for some universal constant C_2. To see that observe that by writing $w = \cos(\theta) + i\sin(\theta)$, for any $0 < p < 1$,

$$\begin{aligned} |pw + (1-p)|^2 &= 1 - 2(1-p)p\,(1 - \cos(\theta)) \\ &\geq 1 - (1-p)p\theta^2 + O(\theta^4) \\ &= \exp\left(-(1-p)p\theta^2 + O(\theta^4)\right), \end{aligned} \tag{3.23}$$

where we used the series expansion of cos. Since $\theta \in \gamma_L$, (3.22) follows. Define

$$A(w) = p_{k^*} a_{k^*} + \sum_{k=k^*+1}^{n-1} a_k p_k \prod_{\ell = k^*}^{k-1} (q_\ell + p_\ell w). \tag{3.24}$$

By Lemma 2.2 there exists $w_1 \in \ \gamma_L$ such that $|A(w_1)| \geq e^{-cL}$. Hence, taking absolute values in (3.20) gives for $w \in \gamma_L$,

$$\begin{aligned} \sum_{j \geq 0} \left|\mathbb{E}\left(\widetilde{X}_j - \widetilde{Y}_j\right)\right| \geq& \left|\prod_{\ell < k^*} (q_\ell + p_\ell w_1)\right| |A(w_1)| \\ &\geq \exp(-nC_2/L^2) e^{-cL}. \end{aligned} \tag{3.25}$$

To approximately maximize the term on the right side of (3.25) we choose L of order $n^{1/3}$ and obtain that there is a constant $C_3 > 0$ such that

$$\sum_{j \geq 0} \left|\mathbb{E}\left(\widetilde{X}_j - \widetilde{Y}_j\right)\right| \geq \exp\left(-C_3 n^{1/3}\right). \tag{3.26}$$

We now conclude the proof similarly as the proof of [8, Theorem 1.1] (see the argument starting with equation (2.4) of that paper). We first observe that for some $j \geq 0$ and $C_4 > 0$ we have $\left|\mathbb{E}\left(\widetilde{X}_j - \widetilde{Y}_j\right)\right| \geq \exp(-C_4 n^{1/3})$. By Hoeffding's inequality this allows us to test whether our bit sequence is more likely to equal $\mathbf{x}$ or $\mathbf{y}$; in case our string equals either $\mathbf{x}$ or $\mathbf{y}$ our test fails with probability at most $\exp(-C_5 n^{1/3})$. Repeating this for all possible pairs of bit sequences $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$, we can determine the original bit string with high probability as $n \to \infty$. By increasing the constant C appearing in the statement of the theorem the result holds for any n. See [8] for a more detailed argument.

For $m \neq 2$ we proceed similarly. Let $\mathbf{x}, \mathbf{y} \in [m]^n$ for $m \neq 2$. For each fixed $\zeta \in [m]$, we define $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ to be equal to $\mathbf{x}$ and $\mathbf{y}$, respectively, except that we replace x_k by 1 if $x_k = \zeta$, and we replace x_k by 0 if $x_k \neq \zeta$. Using the above procedure we can find all k such that $x_k = \zeta$ in the original string. Repeating for all $\zeta \in [m]$ we determine the original string.

For deletion probabilities satisfying Assumption (ii) let

$$\begin{aligned} A(z) =& p_{k^*} a_{k^*} \\ &+ \sum_{k=k^*+1}^{n-1} a_k p_k \prod_{\ell=k^*}^{k-1} \left(\frac{p_\ell}{\widetilde{p}} z + q_\ell - \frac{p_\ell \widetilde{q}}{\widetilde{p}}\right). \end{aligned} \tag{3.27}$$

By Lemma 2.3 there exists $z_0 \in \ \gamma_L$ such that $|A(z_0)| \geq e^{-cL}$. Similarly as above, for $z \in \gamma_L$ and $w = \frac{z - \widetilde{q}}{\widetilde{p}}$,

$$\left|\prod_{\ell < k^*} (q_\ell + p_\ell w)\right| \geq \exp(-nC_6/L^2), \tag{3.28}$$

and

$$|w| \geq \exp\left(-C_7/L^2\right). \tag{3.29}$$

Taking again absolute values in (3.20), this gives for $z \in \gamma_L$,

$$\begin{aligned} &\sum_{j \geq 0} \left|\mathbb{E}\left(\widetilde{X}_j - \widetilde{Y}_j\right)\right| |w|^j \\ &\geq \left|\prod_{\ell < k^*} (q_\ell + p_\ell w)\right| |A(z)| \\ &\geq \exp(-nC_6/L^2) e^{-cL}. \end{aligned} \tag{3.30}$$

Using (3.29), this gives that $\left|\mathbb{E}\left(\widetilde{X}_j - \widetilde{Y}_j\right)\right| \geq \exp(-C_8 n^{1/3})$ for some $j \geq 0$ and $C_8 > 0$. We conclude the proof as above.

4 Proof of main result: Deletion probabilities varying with letter

To prove Theorem 1.2, we first observe that the theorem is immediate from [8] and [3] in the case where the deletion probabilities $q_0, \ldots, q_{m-1}$ are known. This follows since we can send the traces through a second deletion channel, where each letter $\zeta \in [m]$ is retained with probability $p_\zeta^{-1}\min_{\zeta'\in[m]} p_{\zeta'}$. The traces obtained in the second deletion channel can be obtained directly from $\mathbf{x}$ sent through a single deletion channel with constant retention probability $\min_{\zeta\in[m]} p_\zeta$, and $\mathbf{x}$ can therefore be reconstructed with $\exp(O(n^{1/3}))$ traces. For the case of unknown deletion probabilities, we show in Lemma 4.1 that we can obtain good estimates for the deletion probabilities by studying the traces. Then we use Lemma 4.2 to argue that these estimates are sufficiently good, so that the single bit test still works when we use our estimated values for the deletion probabilities.

Lemma 4.1. *Consider the setting of Theorem 1.2, where we assume the deletion probabilities are unknown, and that each letter in $[m]$ appears at least once in $\mathbf{x}$. Given any $\delta, C_1 > 0$, we can find a $C_0 > 0$ depending only on δ, m, and C_1, such that if we have at least $T = \lceil \exp(C_0 n^{1/3})\rceil$ traces, then we can use the traces to find an estimate $\widehat{p}_\zeta$ for each p_ζ satisfying*

$$\mathbb{P}\left[\max_{\zeta\in[m]} |\widehat{p}_\zeta - p_\zeta| > \exp(-C_1 n^{1/3})\right] < \delta. \tag{4.31}$$

Proof. Fix $\zeta \in [m]$ and define $p := p_\zeta$, $r := |\{k \in [n] : x_k = \zeta\}|$, $\mu := rp$, and $v := rp(1-p)$. Observe that $1 - p = v/\mu$. If $Y_t = |\{k \in [|\widetilde{X}^t|] : \widetilde{X}^t_k = \zeta\}|$, then $\mathbb{E}[Y_t] = \mu$ and $\mathrm{Var}[Y_t] = v$. We use the following estimates $\widehat{\mu}, \widehat{v}$, and $\widehat{p}$ for μ, v, and p, respectively,

$$\widehat{\mu} = \frac{1}{T}\sum_{t=1}^{T} Y_t, \qquad \widehat{v} = \frac{1}{T-1}\sum_{t=1}^{T}(Y_t - \widehat{\mu})^2,$$
$$1 - \widehat{p} = \widehat{v}/\widehat{\mu}.$$

We have $\mathbb{E}[\widehat{\mu}] = \mu$, $\mathbb{E}[\widehat{v}] = v$, and $\mathrm{Var}[\widehat{\mu}] = v/T < r/T$. Also observe that for a universal constant $C > 0$, we have $\mathrm{Var}[\widehat{v}] \le Cr^4/T$. This means that for an appropriate $C_0 > 0$, with probability $1 - o_n(1)$, we have $|\widehat{\mu} - \mu|, |\widehat{v} - v| < \exp(-C_1 n^{1/3})$. Upon increasing C_0 and using the definition of $\widehat{p}$, this gives that $|\widehat{p} - p| < \exp(-C_1 n^{1/3})$ with probability $1 - o_n(1)$, which implies the lemma.

For $\mathbf{x} = (x_0, \dots, x_{n-1}) \in \mathbb{R}^n$ and $\mathbf{p} = (p_0, \dots, p_{m-1}) \in (0,1]^m$, define

$$\Phi^{\mathbf{x}}_{\mathbf{p}}(w) = \sum_{k=0}^{n-1} x_k p_{x_k} \prod_{\ell=0}^{k-1} (p_{x_\ell} w + q_{x_\ell}). \tag{4.32}$$

Lemma 4.2. *For any $C_2 > 0$, we can find a $C_1 > 0$ such that the following holds for any $p \in (0,1]$. Let $\mathbf{p} = (p, \dots, p) \in (0,1]^m$ and $\mathbf{p}' = (p'_0, \dots, p'_{m-1}) \in (0,1]^m$ satisfy*

$$|p - p'_\zeta| < \exp(-C_1 n^{1/3}), \qquad \forall \zeta \in [m]. \tag{4.33}$$

Then for each $\mathbf{x} \in [m]^n$, each coefficient in the polynomial $\Phi^{\mathbf{x}}_{\mathbf{p}}(w) - \Phi^{\mathbf{x}}_{\mathbf{p}'}(w)$ has magnitude less than $\exp(-C_2 n^{1/3})$.

Proof. To simplify notation, let $\varepsilon = \exp(-C_1 n^{1/3})$. The magnitude of the coefficient of w^j in $\Phi^{\mathbf{x}}_{\mathbf{p}}(w) - \Phi^{\mathbf{x}}_{\mathbf{p}'}(w)$, is bounded above by the following, with $B(k,p)$ denoting a binomial random variable

$$m \sum_{k=0}^{n-1} \max\Big\{ \big|\mathbb{P}[B(k, p + \widehat{\varepsilon}) = j] - \mathbb{P}[B(k,p) = j]\big| : \widehat{\varepsilon} \in [-\varepsilon, \varepsilon] \Big\} + \varepsilon,$$

which is bounded by a polynomial multiple of ε.

Proof. [Proof of Theorem 1.2] Throughout the proof we consider constants C_0, C_1, C_2 which may depend on δ, $p^* := \min_{\zeta\in[m]} p_\zeta$, and m, but which are independent of all other parameters. Let $T = \lceil \exp(C_0 n^{1/3})\rceil$ for some $C_0 > 0$ which will be determined later. Consider T traces $\widetilde{\mathbf{X}}^{(1)}, \dots, \widetilde{\mathbf{X}}^{(T)}$ obtained by sending $\mathbf{x}$ through the deletion channel considered in the statement of the theorem. Using the T traces, we find an estimate $\widehat{p}_\zeta$ for each p_ζ as described in Lemma 4.1, and let $\widehat{p}^* = \min_{\zeta\in[m]} \widehat{p}_\zeta$.

Send each trace $\widetilde{\mathbf{X}}^{(t)}$ through a second deletion channel, so that we obtain traces $\check{\mathbf{X}}^{(1)}, \dots, \check{\mathbf{X}}^{(T)}$. In the second deletion channel the letter ζ is retained with probability $\widehat{p}^*/\widehat{p}_\zeta$. Observe that each trace $\check{\mathbf{X}}^{(t)}$ can be obtained from $\mathbf{x}$ by considering a deletion channel in which the letter ζ is retained with probability $p^*_\zeta := p_\zeta \widehat{p}^*/\widehat{p}_\zeta$. In particular, if our estimate $\widehat{p}_\zeta$ for p_ζ is good for all ζ, then each letter is retained with approximately the same probability p^*. Define $\mathbf{p}^* = (p^*_0, \dots, p^*_{m-1}) \in (0,1]^m$ and $\widehat{\mathbf{p}}^* = (\widehat{p}^*, \dots, \widehat{p}^*) \in (0,1]^m$; the first string represents the actual (unknown) retention probabilities, and the second string represents our estimated retention probabilities.

Given strings $\mathbf{y}, \mathbf{y}' \in [m]^n$, let $\check{\mathbf{Y}}$ (resp. $\check{\mathbf{Y}}'$) denote a string obtained by sending $\mathbf{y}$ (resp. $\mathbf{y}'$) through the two deletion channels described above, i.e., the letter ζ is retained with probability p^*_ζ. We first assume $m = 2$. By Lemma 2.1,

$$\sum_{j=0}^{n-1} \mathbb{E}\left(\check{Y}_j - \check{Y}'_j\right) w^j = \Phi^{\mathbf{y}}_{\mathbf{p}^*}(w) - \Phi^{\mathbf{y}'}_{\mathbf{p}^*}(w) \\ = \Big(\Phi^{\mathbf{y}}_{\widehat{\mathbf{p}}^*}(w) - \Phi^{\mathbf{y}'}_{\widehat{\mathbf{p}}^*}(w)\Big) + \Big(\Phi^{\mathbf{y}}_{\mathbf{p}^*}(w) - \Phi^{\mathbf{y}}_{\widehat{\mathbf{p}}^*}(w)\Big) \\ - \Big(\Phi^{\mathbf{y}'}_{\mathbf{p}^*}(w) - \Phi^{\mathbf{y}'}_{\widehat{\mathbf{p}}^*}(w)\Big). \tag{4.34}$$

Let

$$E(C_0, C_1) := \left\{\max_{\zeta\in[m]} |\widehat{p}^* - p^*_\zeta| \geq \exp(-C_1 n^{1/3})\right\}.$$

By [8], there is a $C_2 > 0$, such that for some $j \in [n]$, the absolute value of the coefficient of w^j in $\Phi^{\mathbf{y}}_{\widehat{\mathbf{p}}^*}(w) - \Phi^{\mathbf{y}'}_{\widehat{\mathbf{p}}^*}(w)$ is at least $\exp(-C_2 n^{1/3}/2)$. From the proof in [8] we see that C_2 is universal given any lower bound on $\widehat{p}^*$, so on the event $E(C_0, C_1)^c$ and for sufficiently large n, we may assume that the constant C_2 depends only on p^*.

Given $C_2 > 0$, define $C_1 > 0$ as in Lemma 4.2. By Lemma 4.1, we can find $C_0 > 0$ such that

$$\mathbb{P}\left[E(C_0, C_1)\right] < \delta/2. \tag{4.35}$$

By Lemma 4.2 and (4.34), on $E(C_0, C_1)^c$ and for all sufficiently large n, the absolute value of the coefficient of w^j in the polynomial $\Phi^{\mathbf{y}}_{\mathbf{p}^*}(w) - \Phi^{\mathbf{y}'}_{\mathbf{p}^*}(w)$ is at least $\exp(-2C_2 n^{1/3})$ and of the same sign as the corresponding coefficient in $\Phi^{\mathbf{y}}_{\widehat{\mathbf{p}}^*}(w) - \Phi^{\mathbf{y}'}_{\widehat{\mathbf{p}}^*}(w)$. Therefore, on the event $E(C_0, C_1)^c$, Hoeffding's inequality gives that with probability at least $1 - \exp(-T\exp(-4C_2 n^{1/3})/2)$, we can determine if our unknown bit string $\mathbf{x}$ equals $\mathbf{y}$ or $\mathbf{y}'$, by considering $(\check{\mathbf{X}}^{(t)}_j)_{1\leq t\leq T}$. After possibly increasing C_0, we see by a union bound that on $E(C_0, C_1)^c$ and for $m = 2$, we can identify $\mathbf{x}$ with probability at least $1 - 2^n \exp(-T\exp(-4C_2 n^{1/3})/2)$.

Given $\mathbf{y}, \mathbf{y}' \in [m]^n$ for $m \neq 2$ we proceed similarly. For each fixed $\zeta \in [m]$ and with $\mathbf{x} = (x_0, \dots, x_{n-1})$, we first identify the set $A_\zeta := \{k \in [n] : x_k = \zeta\}$. Define $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{y}}'$ to be equal to $\mathbf{y}$ and $\mathbf{y}'$, respectively, except that all x_k such that $x_k = \zeta$ has been replaced by 1, and all x_k such that $x_k \neq \zeta$ has been replaced by 0. Using the approach above with $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{y}}'$, on the event $E(C_0, C_1)^c$ we can determine A_ζ except on an event of probability $1 - 2^n \exp(-T\exp(-4C_2 n^{1/3})/2)$. We repeat the procedure for each $\zeta \in [m]$, and see that on the event $E(C_0, C_1)^c$ we can determine the sets A_ζ, and hence $\mathbf{x}$, except on an event of probability $1 - o_n(1)$. By increasing the constant C_0 if necessary, we can reconstruct the string with probability at least $1 - \varepsilon$ for any n.

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*, pages 910–918. Society for Industrial and Applied Mathematics, 2004.

[2] P. Borwein and T. Erdélyi. Littlewood-type problems on subarcs of the unit circle. *Indiana University Mathematics Journal*, 46(4):1323, 1997.

[3] 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.

[4] H. Ellegren, N. G. Smith, and M. T. Webster. Mutation rate variation in the mammalian genome. *Current Opinion in Genetics and Development*, 13(6):562 – 568, 2003.

[5] 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 '08, pages 389–398, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[6] A. McGregor, E. Price, and S. Vorotnikova. Trace reconstruction revisited. In A. S. Schulz and D. Wagner, editors, *Algorithms - ESA 2014: 22th Annual European Symposium, Wroclaw, Poland, September 8-10, 2014. Proceedings*, pages 689–700, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[7] G. A. T. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304:581–584, 2004.

[8] 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.

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

[10] N. G. Smith, M. T. Webster, and H. Ellegren. Deterministic mutation rate variation in the human genome. *Genome Res.*, 12:1350–1356, 2002.

[11] F. Supek and B. Lehner. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550):81–84, 2015.