ERRATUM TO “LOWER BOUNDS FOR TRACE RECONSTRUCTION”

BY NINA HOLDEN¹ AND RUSSELL LYONS²

¹Department of Mathematics, Rämistrasse 101, 8092 Zurich, Switzerland
ninahold@gmail.com

²Department of Mathematics, 831 E. 3rd St., Bloomington, IN 47405-7106
rdlyons@indiana.edu


Lemma 3.1 asserts that \( \mathbb{E}_{y_n}[Z(\tilde{y}_n)] - \mathbb{E}_{x_n}[Z(\tilde{x}_n)] = \Theta(n^{-1/2}) \) and \( \mathbb{E}_{y_n}[Z(\tilde{y}_n)] > \mathbb{E}_{x_n}[Z(\tilde{x}_n)] \) for all sufficiently large \( n \). Our proof was not correct; as Benjamin Gunby and Xiaoyu He pointed out to us, we missed four terms in the computation of equation (3.3). Those terms contribute a negative amount, so the proof is more delicate. Here is a correct proof.

The intuition behind the result is that a string with a defect of the type we consider, namely, a 10 in a string of 01’s, is likely to cause more 11’s in the trace than a string without the defect. Since the defect in \( y_n \) is shifted to the right as compared to the defect in \( x_n \), the defect of \( y_n \) is slightly more likely to “fall into” the test window \( \{\lceil 2np + 1 \rceil, \ldots, \lfloor 2np + \sqrt{npq} \rfloor \} \) of the trace than is the defect of \( x_n \). More precisely, the difference in probability is of order \( n^{-1/2} \).

In the proof below, we make this intuition rigorous.

PROOF. We assume throughout the proof that \( k \in \{\lceil 2np + 1 \rceil, \ldots, \lfloor 2np + \sqrt{npq} \rfloor \} \).

First observe that

\[
\mathbb{P}_{x_n}[\tilde{x}_k = \tilde{x}_{k+1} = 1] = \sum_{m=k}^{4n} \mathbb{P}_{x_n}[E(m, k)] \mathbb{P}_{x_n}[\tilde{x}_k = \tilde{x}_{k+1} = 1 \mid E(m, k)],
\]

\[
\mathbb{P}_{y_n}[\tilde{y}_k = \tilde{y}_{k+1} = 1] = \sum_{m=k}^{4n} \mathbb{P}_{y_n}[E(m, k)] \mathbb{P}_{y_n}[\tilde{y}_k = \tilde{y}_{k+1} = 1 \mid E(m, k)],
\]

and

\[
\mathbb{P}_{x_n}[E(m, k)] = \mathbb{P}_{y_n}[E(m, k)] = (1 - q)^k q^{m-k} \binom{m-1}{k-1}, \quad m \in \{k, \ldots, 4n\}.
\]

Note that the string \( x_n \) centered at \( m \) is identical to the string \( y_n \) centered at \( m + 2 \), except for two bits at the ends. Therefore, for every \( m \in \{k, \ldots, 3n\} \), we have

\[
\mathbb{P}_{x_n}[\tilde{x}_k = \tilde{x}_{k+1} = 1 \mid E(m, k)] = \mathbb{P}_{y_n}[\tilde{y}_k = \tilde{y}_{k+1} = 1 \mid E(m + 2, k)] = o^{\infty}(n),
\]

MSC2020 subject classifications: Primary 62C20, 68Q25, 51K99; secondary 68W40, 68Q87, 60K30.

Keywords and phrases: Strings, deletion channel, sample complexity.
where $o^\infty(n)$ denotes something non-negative that decays at least exponentially fast in $n$. Combining this with $P_{x_n}[E(m,k)] = o^\infty(n)$ for $m < k + 2$ or $m > 3n$ yields

$$P_{y_n}[\bar{y}_k = \bar{y}_{k+1} = 1] - P_{x_n}[\bar{x}_k = \bar{x}_{k+1} = 1] = \sum_{m=k}^{3n} \left( P_{x_n}[E(m+2,k)] - P_{x_n}[E(m,k)] \right) P_{x_n}[\bar{x}_k = \bar{x}_{k+1} = 1 | E(m,k)]$$

$$\pm o^\infty(n).$$

Setting $a_m := q^p/(1 - q^2) = q/(1 + q)$ if $m$ is even and $a_m := 0$ otherwise, we see that

$$\sum_{m=k}^{3n} \left( P_{x_n}[E(m+2,k)] - P_{x_n}[E(m,k)] \right) a_m = \pm o^\infty(n).$$

Subtracting this from the previous display gives

$$P_{y_n}[\bar{y}_k = \bar{y}_{k+1} = 1] - P_{x_n}[\bar{x}_k = \bar{x}_{k+1} = 1] = \sum_{m=k}^{3n} \left( P_{x_n}[E(m+2,k)] - P_{x_n}[E(m,k)] \right) \cdot \left( P_{x_n}[\bar{x}_k = \bar{x}_{k+1} = 1 | E(m,k)] - a_m \right) \pm o^\infty(n).$$

(E.1)

The second factor in the above summand, modulo an additive error of $o^\infty(n)$, represents the difference in probability of the event $\bar{x}_k = \bar{x}_{k+1} = 1$ given $E(m,k)$ for the string $x_k$ as compared to a string without any defect. It takes the following explicit form:

(E.2)

$$P_{x_n}[\bar{x}_k = \bar{x}_{k+1} = 1 | E(m,k)] - a_m \approx \begin{cases} 0 & \text{if } m \leq 2n - 3 \text{ is odd,} \\ q^{2n-m-2}(1-q)^2 & \text{if } m \leq 2n - 2 \text{ is even,} \\ \frac{q^2}{1+q} & \text{if } m = 2n - 1, \\ \frac{-q}{1+q} & \text{if } m = 2n, \\ 0 & \text{if } 2n+1 \leq m \leq 3n, \end{cases}$$

where $\approx$ means that we incur an additive error of $\pm o^\infty(n)$.

Now let $j_0$ be a sufficiently large positive integer that

(E.3) $1 - q - q^{2j_0} > 0$.

Note that $j_0$ depends on $q$ but can be chosen so that it does not depend on $n$. We suppose in the rest of the proof that $n > j_0$. By (E.1) and (E.2),

$$P_{y_n}[\bar{y}_k = \bar{y}_{k+1} = 1] - P_{x_n}[\bar{x}_k = \bar{x}_{k+1} = 1] \geq \sum_{m=2n-2j_0}^{2n} \left( P_{x_n}[E(m+2,k)] - P_{x_n}[E(m,k)] \right) \cdot \left( P_{x_n}[\bar{x}_k = \bar{x}_{k+1} = 1 | E(m,k)] - a_m \right) - o^\infty(n).$$

(E.4)

For $m \in \{2n - 2j_0, \ldots, 2n + 2\}$ and with $\xi := k - 2np$, we have

(E.5) $P_{x_n}[E(m+2,k)] - P_{x_n}[E(m,k)] = P_{x_n}[E(2n,k)] \left( \frac{\xi}{nq} \pm O\left(\frac{1}{n}\right) \right),$
because for $m \in \{2n - 2j_0, \ldots, 2n\}$,

\[
\mathbb{P}_{x_n}[E(m, k)] = \frac{(m - k + 1)(m - k + 2) \cdots (2n - k)}{m(m + 1) \cdots (2n - 1) \cdot q^{2n-m}}
\]

\[
= \frac{(m-2np+1)}{2nq} - \xi \cdot \frac{(m-2np+2)}{2nq} \cdots (1 - \frac{\xi}{2nq})
\]

\[
= 1 - \xi(2n-m)/(2nq) \pm O(1/n) \pm O(\xi^2/n^2)
\]

the same result holds for $m \in \{2n + 1, 2n + 2\}$ by a similar estimate.

Combining (E.2) and (E.5), we get that the right-hand side of (E.4) is equal to

\[
(E.6) \quad \mathbb{P}_{x_n}[E(2n, k)] \left( \frac{\xi}{nq} \pm O\left(\frac{1}{n}\right) \right) \cdot \frac{(1-q)(1-q - q^{2j_0})}{1+q}.
\]

Summing the left-hand side of (E.4) over $k \in \{\lceil 2np + 1 \rceil, \ldots, \lceil 2np + \sqrt{npq} \rceil \}$ and using the last display along with $\mathbb{P}_{x_n}[E(2n, k)] = \Theta(n^{-1/2})$ and (E.3), we get the lower bounds in the lemma, namely, $\mathbb{E}_{y_n}[Z(\tilde{y}_n)] - \mathbb{E}_{x_n}[Z(\tilde{x}_n)] = \Omega(n^{-1/2})$ and $\mathbb{E}_{y_n}[Z(\tilde{y}_n)] - \mathbb{E}_{x_n}[Z(\tilde{x}_n)] = O(n^{-1/2})$.

Let $b_{m,n}$ denote the absolute value of the right-hand side of (E.2). By (E.1) and (E.2), we have

\[
\left| \mathbb{P}_{y_n}[\check{y}_k = \check{y}_{k+1} = 1] - \mathbb{P}_{x_n}[\check{x}_k = \check{x}_{k+1} = 1] \right| \leq \sum_{m=\lceil 2np-1 \rceil}^{3n} \left| \mathbb{P}_{x_n}[E(m+2, k)] - \mathbb{P}_{x_n}[E(m, k)] \right| \cdot b_{m,n} + o^\infty(n).
\]

Now sum over $k$; (2.7) of Lemma 2.2 yields $\sum_k \left| \mathbb{P}_{x_n}[E(m+2, k)] - \mathbb{P}_{x_n}[E(m, k)] \right| = O(m^{-1/2}) = O(n^{-1/2})$. In addition, $\sum b_{m,n} = O(1)$. Combining these bounds, we arrive at the upper bound of the lemma.

We remark that one can get a more precise bound in (E.6) that gives something positive for all $q \in (0, 1)$ simultaneously by not truncating the sum on the right-hand side of (E.1) and by using a more precise version of (E.5). The result, in fact, gives lower and upper bounds for the left-hand side of (E.4) of the form

\[
\mathbb{P}_{x_n}[E(2n, k)] \left( \frac{\xi}{nq} \pm O\left(\frac{1}{n}\right) \right) \cdot \frac{(1-q)^2}{1+q}.
\]

Finally, we note that in the proof of Proposition 1.4 on p. 519, the definitions of $X$ and $Y$ should be slightly modified: $c$ should be $\sqrt{c}$ both times.

**Acknowledgements.** We are grateful to Benjamin Gunby and Xiaoyu He for noticing the error and to the referees of the erratum for very careful reading and helpful suggestions.

**Funding.** N.H. is supported by grant 175505 of the Swiss National Science Foundation. R.L. is partially supported by the National Science Foundation under grant DMS-1954086 and the Simons Foundation.