The Annals of Applied Probability
2022, Vol. 32, No. 4, 3201–3203
https://doi.org/10.1214/22-AAP1827
Main article: https://doi.org/10.1214/19-AAP1506

ERRATUM TO "LOWER BOUNDS FOR TRACE RECONSTRUCTION"

BY NINA HOLDEN[1,a] AND RUSSELL LYONS[2,b]

[1]*Department of Mathematics, ETH Zürich,* [a]*ninahold@gmail.com*
[2]*Department of Mathematics, Indiana University,* [b]*rdlyons@indiana.edu*

We correct the proof of Lemma 3.1 of our paper *Ann. Appl. Probab.* **30** (2020) 503–525.

Lemma 3.1 asserts that $\mathbf{E}_{\mathbf{y}_n}[Z(\widetilde{\mathbf{y}}_n)] - \mathbf{E}_{\mathbf{x}_n}[Z(\widetilde{\mathbf{x}}_n)] = \Theta(n^{-1/2})$ and $\mathbf{E}_{\mathbf{y}_n}[Z(\widetilde{\mathbf{y}}_n)] > \mathbf{E}_{\mathbf{x}_n}[Z(\widetilde{\mathbf{x}}_n)]$ 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 $\mathbf{y}_n$ is shifted to the right as compared to the defect in $\mathbf{x}_n$, the defect of $\mathbf{y}_n$ is slightly more likely to "fall into" the test window $\{\lceil 2np+1\rceil, \dots, \lfloor 2np+\sqrt{npq}\rfloor\}$ of the trace than is the defect of $\mathbf{x}_n$. 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 $k \in \{\lceil 2np+1\rceil, \dots, \lfloor 2np+\sqrt{npq}\rfloor\}$. Let $E(m,k)$ denote the event that bit m in the input string is copied to position k in the trace. First observe that

$$\mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1] = \sum_{m=k}^{4n} \mathbf{P}_{\mathbf{x}_n}\big[E(m,k)\big]\mathbf{P}_{\mathbf{x}_n}\big[\widetilde{x}_k = \widetilde{x}_{k+1} = 1 | E(m,k)\big],$$

$$\mathbf{P}_{\mathbf{y}_n}[\widetilde{y}_k = \widetilde{y}_{k+1} = 1] = \sum_{m=k}^{4n} \mathbf{P}_{\mathbf{y}_n}\big[E(m,k)\big]\mathbf{P}_{\mathbf{y}_n}\big[\widetilde{y}_k = \widetilde{y}_{k+1} = 1 | E(m,k)\big],$$

and

$$\mathbf{P}_{\mathbf{x}_n}\big[E(m,k)\big] = \mathbf{P}_{\mathbf{y}_n}\big[E(m,k)\big] = (1-q)^k q^{m-k}\binom{m-1}{k-1}, \quad m \in \{k, \dots, 4n\}.$$

Note that the string $\mathbf{x}_n$ centered at m is identical to the string $\mathbf{y}_n$ centered at $m+2$, except for two bits at the ends. Therefore, for every $m \in \{k, \dots, 3n\}$, we have

$$\mathbf{P}_{\mathbf{x}_n}\big[\widetilde{x}_k = \widetilde{x}_{k+1} = 1 | E(m,k)\big] = \mathbf{P}_{\mathbf{y}_n}\big[\widetilde{y}_k = \widetilde{y}_{k+1} = 1 | E(m+2,k)\big] \pm o^{\infty}(n),$$

where $o^{\infty}(n)$ denotes something nonnegative that decays at least exponentially fast in n. Combining this with $\mathbf{P}_{\mathbf{x}_n}[E(m,k)] = o^{\infty}(n)$ for $m < k+2$ or $m > 3n$ yields

$$\begin{aligned}&\mathbf{P}_{\mathbf{y}_n}[\widetilde{y}_k = \widetilde{y}_{k+1} = 1] - \mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1] \\ &\quad = \sum_{m=k}^{3n} \big(\mathbf{P}_{\mathbf{x}_n}\big[E(m+2,k)\big] - \mathbf{P}_{\mathbf{x}_n}\big[E(m,k)\big]\big)\mathbf{P}_{\mathbf{x}_n}\big[\widetilde{x}_k = \widetilde{x}_{k+1} = 1 | E(m,k)\big] \pm o^{\infty}(n).\end{aligned}$$

Received January 2022; revised March 2022.
MSC2020 subject classifications. Primary 62C20, 68Q25, 51K99; secondary 68W40, 68Q87, 60K30.
Key words and phrases. Strings, deletion channel, sample complexity.

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

$$\sum_{m=k}^{3n} \left(\mathbf{P}_{\mathbf{x}_n}[E(m+2,k)] - \mathbf{P}_{\mathbf{x}_n}[E(m,k)]\right) a_m = \pm o^{\infty}(n).$$

Subtracting this from the previous display gives

$$\begin{aligned} &\mathbf{P}_{\mathbf{y}_n}[\widetilde{y}_k = \widetilde{y}_{k+1} = 1] - \mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1] \\ &\quad = \sum_{m=k}^{3n} \left(\mathbf{P}_{\mathbf{x}_n}[E(m+2,k)] - \mathbf{P}_{\mathbf{x}_n}[E(m,k)]\right) \\ &\qquad \cdot \left(\mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1 | E(m,k)] - a_m\right) \pm o^{\infty}(n). \end{aligned} \tag{E.1}$$

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

$$\mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1 | E(m,k)] - a_m \approx \begin{cases} 0 & \text{if } m \le 2n-3 \text{ is odd}, \\ q^{2n-m-2}(1-q)^2 & \text{if } m \le 2n-2 \text{ is even}, \\ \dfrac{q^2}{1+q} & \text{if } m = 2n-1, \\ -\dfrac{q}{1+q} & \text{if } m = 2n, \\ 0 & \text{if } 2n+1 \le m \le 3n, \end{cases} \tag{E.2}$$

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

$$1 - q - q^{2j_0} > 0. \tag{E.3}$$

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),

$$\begin{aligned} &\mathbf{P}_{\mathbf{y}_n}[\widetilde{y}_k = \widetilde{y}_{k+1} = 1] - \mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1] \\ &\quad \ge \sum_{m=2n-2j_0}^{2n} \left(\mathbf{P}_{\mathbf{x}_n}[E(m+2,k)] - \mathbf{P}_{\mathbf{x}_n}[E(m,k)]\right) \\ &\qquad \cdot \left(\mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1 | E(m,k)] - a_m\right) - o^{\infty}(n). \end{aligned} \tag{E.4}$$

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

$$\mathbf{P}_{\mathbf{x}_n}[E(m+2,k)] - \mathbf{P}_{\mathbf{x}_n}[E(m,k)] = \mathbf{P}_{\mathbf{x}_n}[E(2n,k)]\left(\frac{\xi}{nq} \pm O\left(\frac{1}{n}\right)\right), \tag{E.5}$$

because for $m \in \{2n-2j_0, \dots, 2n\}$,

$$\begin{aligned} \frac{\mathbf{P}_{\mathbf{x}_n}[E(m,k)]}{\mathbf{P}_{\mathbf{x}_n}[E(2n,k)]} &= \frac{(m-k+1)(m-k+2)\cdots(2n-k)}{m(m+1)\cdots(2n-1)\cdot q^{2n-m}} \\ &= \frac{(\frac{m-2np+1}{2nq} - \frac{\xi}{2nq})(\frac{m-2np+2}{2nq} - \frac{\xi}{2nq})\cdots(1 - \frac{\xi}{2nq})}{\frac{m}{2n}\cdot\frac{m+1}{2n}\cdots(1-\frac{1}{2n})} \\ &= 1 - \xi(2n-m)/(2nq) \pm O(1/n) \pm O(\xi^2/n^2) \\ &= 1 - \xi(2n-m)/(2nq) \pm O(1/n); \end{aligned}$$

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

$$\mathbf{P}_{\mathbf{x}_n}\big[E(2n,k)\big]\left(\frac{\xi}{nq} \pm O\left(\frac{1}{n}\right)\right) \cdot \frac{(1-q)(1-q-q^{2j_0})}{1+q}. \tag{E.6}$$

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

It remains to prove the upper bound, namely, $\mathbf{E}_{\mathbf{y}_n}[Z(\widetilde{\mathbf{y}}_n)] - \mathbf{E}_{\mathbf{x}_n}[Z(\widetilde{\mathbf{x}}_n)] = O(n^{-1/2})$. Let $b_{m,n}$ denote the absolute value of the right-hand side of (E.2). By (E.1) and (E.2), we have

$$\begin{aligned}&\big|\mathbf{P}_{\mathbf{y}_n}[\widetilde{y}_k = \widetilde{y}_{k+1} = 1] - \mathbf{P}_{\mathbf{x}_n}[\widetilde{x}_k = \widetilde{x}_{k+1} = 1]\big| \\ &\quad \le \sum_{m=\lceil 2np-1\rceil}^{3n} \big|\mathbf{P}_{\mathbf{x}_n}\big[E(m+2,k)\big] - \mathbf{P}_{\mathbf{x}_n}\big[E(m,k)\big]\big| \cdot b_{m,n} + o^{\infty}(n).\end{aligned}$$

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

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

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

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

Acknowledgments. We are grateful to Benjamin Gunby and Xiaoyu He for noticing the error and to the referees of the erratum for a 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.