# Random suffix search trees

Luc Devroye[1] and Ralph Neininger[2]
School of Computer Science
McGill University
3480 University Street
Montreal, H3A 2K6
Canada

February 4, 2006

### Abstract

A random suffix search tree is a binary search tree constructed for the suffixes $X_i = 0.B_i B_{i+1} B_{i+2} \ldots$ of a sequence $B_1, B_2, B_3., \ldots$ of independent identically distributed random $b$-ary digits $B_j$. Let $D_n$ denote the depth of the node for $X_n$ in this tree when $B_1$ is uniform on $\mathbb{Z}_b$. We show that for any value of $b > 1$, $\mathbb{E} D_n = 2 \log n + O(\log^2 \log n)$, just as for the random binary search tree. We also show that $D_n / \mathbb{E} D_n \to 1$ in probability.

**AMS subject classifications.** Primary: 60D05; secondary: 68U05.

**Key words** Random binary search tree. Suffix tree. Lacunary sequences. Random spacings. Probabilistic analysis of algorithms.

## 1 Introduction

Current research in data structures and algorithms is focused on the efficient processing of large bodies of text (encyclopedia, search engines) and strings of data (DNA strings, encrypted bit strings). For storing the data such that string searching is facilitated, various data structures have been proposed. The most popular among these are the suffix tries and suffix trees (Weiner, 1973; McCreight, 1976),

[1] Research of both authors supported by NSERC grant A3450.

[2] Research supported by the Deutsche Forschungsgemeinschaft.

and suffix arrays (Manber and Myers, 1990). Related intermediate structures such as the suffix cactus (Karkkainen, 1995) have been proposed as well. Apostolico (1985), Crochemore and Rytter (1994), and Stephen (1994) cover most aspects of these data structures, including their applications and efficient construction algorithms (Ukkonen 1995, Weiner 1973, Giegerich and Kurtz, 1997, and Kosaraju, 1994). If the data are thought of as strings $B_1, B_2, \ldots$ of symbols taking values in an alphabet $\mathbb{Z}_b = \{0, 1, \ldots, b-1\}$ for fixed finite $b$, then the suffix trie is an ordinary $b$-ary trie for the strings $X_i = (B_i, B_{i+1}, \ldots)$, $1 \leq i \leq n$. The suffix tree is a compacted suffix trie. The suffix array is an array of lexicographically ordered strings $X_i$ on which binary search can be performed. Additional information on suffix trees is given in Farach (1997), Farach and Muthukrishnan (1996, 1997), Giancarlo (1993, 1995), Giegerich and Kurtz (1995), Gusfield (1997), Sahinalp and Vishkin (1994), Szpankowski (1993). The suffix search tree we are studying in this paper, first suggested by Manber and Myers (1993), is the search tree obtained for $X_1, \ldots, X_n$, where again lexicographical ordering is used. Care must be taken to store with each node the position in the text, so that the storage comprises nothing but pointers to the text. Suffix search trees permit dynamic operations, including the deletion, insertion, and alteration of parts of the string. Suffix arrays on the other hand are clearly only suited for off-line applications.

The analysis of random tries has a long history (see Szpankowski, 2001, for references). Random suffix tries were studied by Jacquet, Rais and Szpankowski (1995) and Devroye, Szpankowski and Rais (1992). The main model used in these studies is the independent model: the $B_i$'s are independent and identically distributed. Markovian dependence has also been considered. If $p_j = \mathbb{P}\{B_1 = j\}$, $0 \leq j < b$, then it is known that the expected depth of a typical node in an $n$-node suffix trie is close in probability to $(1/\mathcal{E}) \log n$, where $\mathcal{E} = \sum_j p_j \log(1/p_j)$ is the entropy of $B_1$. The height is in probability close to $(b/\xi) \log n$, where $\xi = \log(1/\sum_j p_j^b)$. If $\xi$ or $\mathcal{E}$ are small, then the performance of these structures deteriorates to the point that perhaps more classical structures such as the binary search tree are preferable.

In this paper, we prove that for first order asymptotics, random suffix search trees behave roughly as random binary search trees. If $D_n$ is the depth of $X_n$, then

$$\mathbb{E} D_n = 2 \log n + O(\log^2 \log n)$$

and $D_n / \log n \to 2$ in probability, just as for the random binary search tree constructed as if the $X_i$'s were independent identically distributed strings (Knuth, 1973, and Mahmoud, 1992, have references and accounts). We prove this for $b = 2$ and $p_0 = p_1 = 1/2$. The generalization to $b > 2$ is straightforward as long as $B_1$ is uniform on $\mathbb{Z}_b$.

The second application area of our analysis is related directly to random binary search trees. We may consider the $X_i$'s as real numbers on $[0, 1]$ by considering the $b$-ary expansions

$$X_i = 0.B_i B_{i+1} \ldots , \quad 1 \leq i \leq n \ .$$

In that case, we note that $X_{i+1} = \{bX_i\} := (bX_i) \bmod 1$. If we start with $X_1$ uniform on $[0, 1]$, then every $X_i$ is uniform on $[0, 1]$, but there is some dependence in the sequence $X_1, X_2, \ldots$. The sequence generated by applying the map $X_{i+1} = \{bX_i\}$ resembles the way in which linear congruential sequences are generated on a computer, as an approximation of random number sequences. In fact, all major numerical packages in use today use linear congruential sequences of the form $x_{n+1} = (bx_n + a) \bmod M$, where $a, b, x_n, x_{n+1}, M$ are integers. The sequence $x_n/M$ is then used as an approximation of a truly random sequence. Thus, our study reveals what happens when we replace i.i.d. random variables with the multiplicative sequence. It is reassuring to note that the first order behavior of binary search trees is identical to that for the independent sequence.

The study of the behavior of random binary search trees for dependent sequences in general is quite interesting. For the sequence $X_n = (nU) \mod 1$, with $U$ uniform on $[0, 1]$, a detailed study by Devroye and Goudjil (1998) shows that the height of the tree is in probability $\Theta(\log n \log \log n)$. The behavior of less dependent sequences $X_n = (n^\alpha U) \mod 1$, $\alpha > 1$, is largely unknown. The present paper shows of course that $X_n = (2^n U) \mod 1$ is sufficiently independent to ensure behavior as for an i.i.d. sequence. Antos and Devroye (2000) looked at the sequence $X_n = \sum_{i=1}^n Y_i$, where the $Y_i$'s are i.i.d. random variables and showed that the height is in probability $\Theta(\sqrt{n})$. Cartesian trees (Devroye 1994) provide yet another model of dependence with heights of the order $\Theta(\sqrt{n})$.

The paper is organized as follows: in sections 2 through 5, we develop the basic tools for our analysis. In section 6, we show that

$$\mathbb{E} D_n = 2 \log n + O(\log^2 \log n).$$

In section 7, a more general refined analysis leads to a weak law of large numbers: $D_n / \mathbb{E} D_n \to 1$ in probability. These are our main results — they rest on a key circular symmetrization argument used in the proof of Lemma 5.1. There is another avenue, based on the observation that if $S_0^n, \dots, S_n^n$ are the lengths of the spacings defined on $[0,1]$ by $X_1, \dots, X_n$, then the expected depth of $X_{n+1}$ in the tree for $X_1, \dots, X_n$ is roughly given by

$$\sum_{j=1}^{n-1} \sum_{i=0}^{j} \mathbb{E}\left[(S_i^j)^2\right].$$

The study of the spacings is also important for the analysis of the size of the subtree rooted at $X_j$, as this has expected value roughly given by

$$(n-j)\, \mathbb{E}\, S_j^*(j-1),$$

where $S_n^*(i)$ is the length of the unique spacing among $S_0^i, \dots, S_i^i$ that covers $X_n$. Thus we embark on the study of the spacings in section 8 and 9, where we show first that a randomly picked spacing in the $n$-th partition is asymptotically of size $E/n$, where $E$ is an exponential random variable (just as for the case of spacings defined by i.i.d. uniform $[0,1]$ random variables). Although this result can be obtained from the number theoretical work of Rudnick and Zaharescu (2002), a self-contained probabilistic proof is included in this paper. In section 10 and 11, the spacings argument is fleshed out to show, for example that the size of a subtree rooted at $X_j$ times $j/n$ tends in distribution to a Gamma(2) random variable, whenever $j/\log^5 n \to \infty$ and $(j \log^2 n)/n \to 0$.

## 2 Notation

Denote the uniform distribution on $[0,1]$ by $U[0,1]$ and the Bernoulli($p$) distribution by $Be[p]$. We have given a $U[0,1]$ distributed random variable $X_1$ and define $X_k := T(X_{k-1})$ for $k \geq 2$, with the map $T : [0,1] \to [0,1], x \mapsto \{2x\} = 2x \mod 1$.

In the binary representation $X_1 = 0.B_1B_2\dots$, the $B_k$ are independent $Be[1/2]$ bits. Then we have

$$X_k = 0.B_kB_{k+1}B_{k+2}\cdots$$

for all $k \geq 1$. For $m \geq 1$ we introduce the corresponding perturbed random variates

$$Y_k^{\langle m\rangle} := 0.B_kB_{k+1}\dots B_{k+m-1}B_1^{(k)}B_2^{(k)}\dots, \quad k = 1, \dots, n,$$

where $\{B_j^{(k)} : k, j \geq 1\}$ is a family of independent $Be[1/2]$ distributed bits, independent of $X_1$. Then we have for all $k \geq 1$,

$$|X_k - Y_k^{\langle m\rangle}| \leq \frac{1}{2^m},$$

and $Y_i^{\langle m\rangle}, Y_j^{\langle m\rangle}$ are independent if $|i-j| \geq m$.

## 3 The perturbed tree

In this section we control the probability that the random suffix search tree built from $X_1, \dots, X_n$ and the perturbed tree generated by $Y_1^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle}$ coincide. We denote by $\|\lfloor x \rfloor\| := 2\lfloor x/2 \rfloor$ the largest even integer not exceeding $x$. For a vector $(a_1, \dots, a_n)$ of distinct real numbers, let $\pi(a_1, \dots, a_n)$ be the permutation given by the vector.

**Lemma 3.1** *If $m := 18\|\lfloor \log_2 n \rfloor\|$, then for all $n \geq 16$,*

$$\mathbb{P}\left(\pi(X_1, \dots X_n) \neq \pi(Y_1^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle})\right) \leq \frac{8}{n^2}.$$

**Proof:** We introduce the truncated random variables $Y_k$ by their binary representations $Y_k := 0.B_k B_{k+1} \cdots B_{k+m-1}$ for $k \geq 1$. Then we have

$$\left\{\pi(X_1, \dots X_n) \neq \pi(Y_1^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle})\right\} \subseteq \bigcup_{1 \leq i < j \leq n} \{Y_i = Y_j\}.$$

This implies

$$\begin{aligned} & \mathbb{P}\left(\pi(X_1, \dots X_n) \neq \pi(Y_1^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle})\right) \\ \leq & \sum_{1 \leq i < j \leq n} \mathbb{P}(Y_i = Y_j) \\ \leq & \frac{n^2}{2^m} + n \sum_{j=2}^{m} \mathbb{P}(B_1 \cdots B_m = B_j \cdots B_{m+j-1}). \end{aligned}$$

For $1 < j \leq m$ we have $\mathbb{P}(B_1 \cdots B_{j-1} = B_j \cdots B_{2j-2}) = 1/2^{j-1}$. We split the bit vector $B_1 \cdots B_m$ into $b := \lfloor m/(j-1) \rfloor$ blocks of length $j-1$. Then we obtain

$$\begin{aligned} & \mathbb{P}(B_1 \cdots B_m = B_j \cdots B_{m+j-1}) \\ \leq & \mathbb{P}(B_1 \cdots B_{j-1} = B_j \cdots B_{2j-2} = \cdots = B_{(b-1)(j-1)+1} \cdots B_{b(j-1)}) \\ \leq & \frac{1}{2^{(b-1)(j-1)}} \\ \leq & \frac{1}{2^{m-2j+2}}. \end{aligned}$$

Altogether we have

$$\begin{aligned}
&\mathbb{P}\left(\pi(X_1,\dots X_n) \neq \pi(Y_1^{\langle m\rangle},\dots,Y_n^{\langle m\rangle})\right)\\
&\quad\leq \frac{n^2}{2^m} + n\left(\sum_{j=2}^{\lceil m/3\rceil} \frac{1}{2^{m-2j+2}} + \sum_{j=\lceil m/3\rceil+1}^{m} \frac{1}{2^{j-1}}\right)\\
&\quad\leq \frac{n^2}{2^m} + n\left(\frac{4}{3}\frac{1}{2^{m/3}} + \frac{2}{2^{m/3}}\right).
\end{aligned}$$

With $m = 18\|\lfloor \log_2 n\rfloor\|$ we obtain $n^2/2^m \leq 4/n^2$ for all $n \geq 8$ and $n((4/3+2)/2^{m/3}) \leq 4/n^2$ for all $n \geq 16$. The assertion follows. ■

The perturbed tree and the original tree are thus identical with high probability. In the perturbed tree, note that $Y_i^{\langle m\rangle}$ and $Y_j^{\langle m\rangle}$ are independent whenever $|i-j| \geq m$. Unfortunately, it is not true that random binary search trees constructed on the basis of identically distributed $m$-dependent sequences behave as those for i.i.d. sequences, even when $m$ is as small as 1. For example, the depth of a typical node and the height may increase by a factor of $m$ when $m$ is small and positive.

For later use we provide a technical lemma on the distances of the quantities $X_i, X_j$ and $Y_i^{\langle t\rangle}, Y_j^{\langle t\rangle}$ respectively.

**Lemma 3.2** *For all integer* $1 \leq i < j$, $t \geq 1$ *and real* $\varepsilon > 0$,

$$\mathbb{P}(|X_i - X_j| \leq \varepsilon) \leq 2\varepsilon, \quad \mathbb{P}(|Y_i^{\langle t\rangle} - Y_j^{\langle t\rangle}| \leq \varepsilon) \leq 8\varepsilon.$$

**Proof:** Define $k := j - i$. We have $\{|X_i - X_j| \leq \varepsilon\} = \{|X_i - T^k(X_i)| \leq \varepsilon\}$, where $T^k$ is the $k$-th iteration of the map $T$ defined in section 2. With the representation

$$X_i = \frac{\ell}{2^k} + \frac{\xi}{2^k}, \quad \ell \in \{0,\dots,2^k-1\},\ \xi \in [0,1], \tag{1}$$

we obtain $T^k(X_i) = \xi$. Thus we have $|X_i - X_j| = |\ell/2^k + \xi/2^k - \xi| \leq \varepsilon$ if and only if

$$\xi \in \left[\frac{\ell - 2^k\varepsilon}{2^k-1}, \frac{\ell + 2^k\varepsilon}{2^k-1}\right] \cap [0,1].$$

Plugging this into (1) we obtain

$$\{|X_i - X_j| \leq \varepsilon\} = \left\{X_i \in \bigcup_{\ell=0}^{2^k-1}\left(\left[\frac{\ell-\varepsilon}{2^k-1}, \frac{\ell+\varepsilon}{2^k-1}\right] \cap \left[\frac{\ell}{2^k}, \frac{\ell+1}{2^k}\right]\right)\right\}, \tag{2}$$

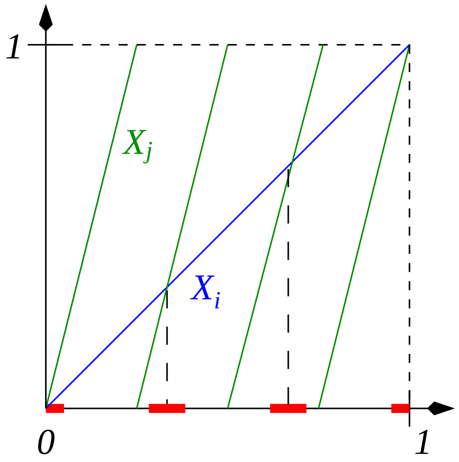

Figure 1: *Shown is the set* $\{|X_i - X_j| \le \varepsilon\}$ *(in red) for* $k = j - i = 2$ *and* $\varepsilon = 3/20$ *together with* $X_i$ *and* $X_j$*, where* $X_i$ *is modeled as the identity on* $[0, 1]$.

see Figure 1. Since $X_i$ is $U[0, 1]$ distributed, we obtain $\mathbb{P}(|X_i - X_j| \le \varepsilon) \le 2\varepsilon$.

For the second statement note that for $k > t$ there is nothing to prove, since $Y_i^{\langle t\rangle}, Y_j^{\langle t\rangle}$ are independent in this case. Hence, we assume $k \le t$ and denote $\mathcal{J}_{\ell m} := [(\ell - 1)/2^k + (m - 1)/2^t, (\ell - 1)/2^k + m/2^t]$ for $\ell = 1, \dots, 2^k$, $m = 1, \dots, 2^{t-k}$. Conditioned on $\{Y_i^{\langle t\rangle} \in \mathcal{J}_{\ell m}\}$ the variables $Y_i^{\langle t\rangle}, Y_j^{\langle t\rangle}$ are independent with uniform distributions on $\mathcal{J}_{\ell m}$ and $\mathcal{J}_m := [(m-1)/2^{t-k}, m/2^{t-k}]$ respectively. We abbreviate these conditioned variates by $V_{\ell m}$ an $W_m$. Then we have

$$\mathbb{P}(|Y_i^{\langle t\rangle} - Y_j^{\langle t\rangle}| \le \varepsilon) = \frac{1}{2^t} \sum_{\ell=1}^{2^k} \sum_{m=1}^{2^{t-k}} \mathbb{P}(|V_{\ell m} - W_m| \le \varepsilon). \tag{3}$$

Note that conditioning on $V_{\ell m}$ we obtain the estimate $\mathbb{P}(|V_{\ell m} - W_m| \le \varepsilon) \le 2\varepsilon 2^{t-k}$, valid for all $\ell, m$. We fix $\ell$ in (3) and distinguish two cases:
<u>Case $\varepsilon \le 2^{-(t-k)}$:</u> We have $\{|V_{\ell m} - W_m| \le \varepsilon\} \ne \emptyset$ for at most three of the $m \in \{1, \dots, 2^{t-k}\}$. Thus, we obtain

$$\mathbb{P}(|Y_i^{\langle t\rangle} - Y_j^{\langle t\rangle}| \le \varepsilon) \le \frac{1}{2^t} \sum_{\ell=1}^{2^k} 6\varepsilon 2^{t-k} = 6\varepsilon.$$

<u>Case $\varepsilon \ge 2^{-(t-k)}$:</u> Since incrementing $m$ by one changes the distance between the centers of $\mathcal{J}_{\ell m}$ and $\mathcal{J}_m$ by $2^{-(t-k)} - 2^{-t} \ge 2^{-(t-k+1)}$, at most $2 + 2\lceil \varepsilon 2^{t-k+1}\rceil$ of the

events $\{|V_{\ell m} - W_m| \leq \varepsilon\}$ are nonempty. Thus

$$\mathbb{P}(|Y_i^{\langle t\rangle} - Y_j^{\langle t\rangle}| \leq \varepsilon) \leq \frac{1}{2^t}\sum_{\ell=1}^{2^k}(2\varepsilon 2^{t-k+1} + 4) = 4\varepsilon + 4(2^{-(t-k)}) \leq 8\varepsilon,$$

which completes the proof. ■

## 4 A rough bound for the height

We will need a rough upper bound for the mean of the height of the random suffix search tree.

**Lemma 4.1** *Let a binary search tree $\mathcal{T}$ be built up from distinct numbers $x_1, \dots, x_n$ and denote its height by $H$. We assume that the set of indices $\{1, \dots, n\}$ is decomposed into $k$ nonempty subsets $\mathcal{I}_1, \dots, \mathcal{I}_k$ of cardinalities $|\mathcal{I}_j| = n_j$. Assume that $\mathcal{I}_j$ consists of the indices $n(j,1) < \cdots < n(j,n_j)$ and denote the height of the binary search tree $\mathcal{T}_j$ built up from $x_{n(j,1)}, \dots, x_{n(j,n_j)}$ by $H_j$ for $j = 1, \dots, k$. Then we have*

$$H \leq k - 1 + \sum_{j=1}^{k} H_j. \tag{4}$$

**Proof:** A basic property of the binary search tree is that a pair of keys $x < y$ is inserted in nodes on a common path form the root if and only if no key $s$ with $x < s < y$ has been inserted before $x$ and $y$. For an arbitrary node $u$ in $\mathcal{T}$ we consider two keys on its path to the root such that their indices $i_1 < i_2$ belong to the same set $\mathcal{I}_j$ for some $j \in \{1, \dots, k\}$. It follows that no key $x_i$ exists with index $i < i_1$ and $x_{i_1} < x_i < x_{i_2}$. In particular there is no such $x_i$ with $i \in \mathcal{I}_j$. Therefore, in $\mathcal{T}_j$ the keys $x_{i_1}, x_{i_2}$ are inserted on a common path from the root as well. This implies that the number of nodes in $\mathcal{T}$ on the path from the root to $u$ having indices in $\mathcal{I}_j$ is at most $H_j + 1$. The assertion follows. ■

**Lemma 4.2** *Let $H_n$ denote the height of the random suffix search tree with $n$ nodes. Then $\mathbb{E}H_n = O(\log^2 n)$.*

**Proof:** For $j = 1, \dots, m := 18\|\lfloor \log_2 n \rfloor\|$ we define $\mathcal{I}_j := \{bm + j : b \in \mathbb{N}_0, bm + j \leq n\}$. The families $(Y_i^{\langle m\rangle})_{i \in \mathcal{I}_j}$ consist each of independent random variables being $U[0,1]$ distributed. Thus these families form random equiprobable permutations.

The trees $\mathcal{T}_j$ built from $(Y_i^{\langle m\rangle})_{i\in\mathcal{I}_j}$ are *random* binary search trees, where *random* refers to the random permutation model. With $\hat{H}_j$ denoting the height of $\mathcal{T}_j$, and $\bar{H}_n$ denoting the height of the tree built from $Y_1^{\langle m\rangle},\ldots,Y_n^{\langle m\rangle}$, by Lemma 4.1,

$$\bar{H}_n \le m + \sum_{j=1}^{m} \hat{H}_j. \tag{5}$$

From the analysis of random binary search trees we know $\mathbb{E}\hat{H}_j \sim \gamma\log(n/m)$ with $\gamma > 0$ (see Devroye 1987). Thus (5) implies $\mathbb{E}\bar{H}_n = O(\log^2 n)$. Finally, we have

$$\begin{aligned} H_n &= \bar{H}_n + \mathbf{1}_{\{\pi(X_1,\ldots X_n)\neq\pi(Y_1^{\langle m\rangle},\ldots,Y_n^{\langle m\rangle})\}}(H_n - \bar{H}_n) \\ &\le \bar{H}_n + \mathbf{1}_{\{\pi(X_1,\ldots X_n)\neq\pi(Y_1^{\langle m\rangle},\ldots,Y_n^{\langle m\rangle})\}}n. \end{aligned}$$

Here, $\mathbf{1}_A$ denotes the indicator function of an event $A$. Lemma 3.1, for $n \ge 16$, implies $\mathbb{E}H_n \le \mathbb{E}\bar{H}_n + 8/n = O(\log^2 n)$. ■

Lemma 4.2 is valid for our model, but also for any random binary search tree constructed on the basis of $U[0,1]$ random variables that are $m$-dependent, with $m = O(\log n)$.

## 5 A key lemma

We introduce the events $A_j = \{X_j \text{ is ancestor of } X_n \text{ in the tree}\}$. Then we have the representations

$$D_n = \sum_{j=1}^{n-1} \mathbf{1}_{A_j}, \qquad \mathbb{E}D_n = \sum_{j=1}^{n-1} \mathbb{P}(A_j).$$

We use the notation $\alpha,\beta \triangleright \gamma_1,\ldots,\gamma_n$, if there does not exist $k$ with $1 \le k \le n$ for which $\alpha < \gamma_k < \beta$ or $\beta < \gamma_k < \alpha$, i.e., $\alpha,\beta$ are neighbors in $\{\gamma_1,\ldots,\gamma_n\}$. Note that $A_j = \{X_j, X_n \triangleright X_1,\ldots,X_{j-1}\}$. We use $A_j^{\langle m\rangle}$ for the corresponding event involving the $Y_k^{\langle m\rangle}$: $A_j^{\langle m\rangle} = \{Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle},\ldots,Y_{j-1}^{\langle m\rangle}\}$. Thoughout we abbreviate $m = 18\|\lfloor\log_2 n\rfloor\|$.

Our key lemma consists of an analysis of the depth of the $n$-th inserted node $X_n$ conditioned on its location. For $x \in [0,1]$ and $1 \le i \le n-1$, define

$$p_i(x) := \mathbb{P}\Big(Y_i^{\langle m\rangle}, x \triangleright Y_1^{\langle m\rangle},\ldots,Y_{i-1}^{\langle m\rangle}\Big).$$

We use the following *bad set*:

$$B_n(\xi) := \bigcup_{k=1}^{m} \{x \in [0,1] : |x - T^k(x)| < \xi\}, \quad \xi > 0,$$

where $T$ is the map $T(x) := \{2x\}$ and $T^k$ its $k$-th iteration, see Figure 2.

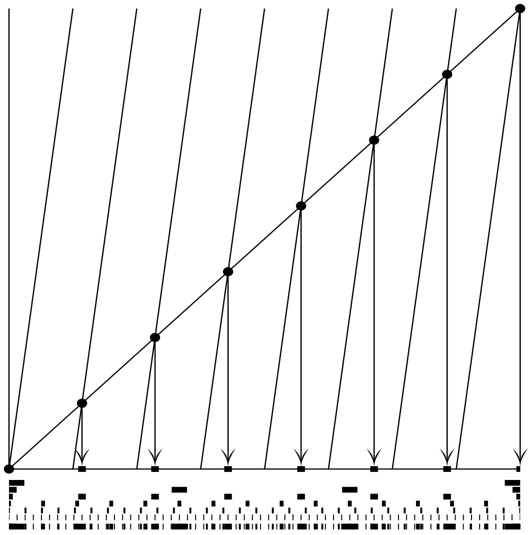

Figure 2: *The last line shows the bad set $B_n(\xi)$ for $m = 6$ and $\xi = 3/50$. The six lines above show the sets $\{|x - T^k(x)| \le \xi\}$ for $k = 1, \dots, 6$. In the square, for the case $k = 3$, it is shown how these sets emerge.*

**Lemma 5.1** *For all $n$ sufficiently large, all $x \in [0,1]$, and $1 \le i < n$, we have*

$$\begin{aligned} p_i(x) &= \mathbf{1}_{[m^2/i, 1-m^2/i]}(x) \left( \frac{2}{i} + R_1(n,i) + \mathbf{1}_{B_n(2m^2/\sqrt{i})}(x) R_2(n,i) \right) \\ &\quad + \left(1 - \mathbf{1}_{[m^2/i,1-m^2/i]}(x)\right) R_3(n,i), \end{aligned}$$

*where for appropriate constants $C_1, C_2, C_3 > 0$,*

$$\begin{aligned} |R_1(n,i)| &\le C_1 \frac{\log^6 n}{i^{3/2}}, \\ |R_2(n,i)| &\le C_2 \frac{\log^3 n}{i}, \\ |R_3(n,i)| &\le C_3 \frac{\log n}{i}. \end{aligned}$$

**Proof:** Let $X_1, \ldots, X_i$ be given. Recall the notation from section 2:

$$\begin{aligned} Y_1^{\langle m\rangle} &= 0.B_1B_2\ldots B_mB_1^{(1)}B_2^{(1)}\ldots \\ Y_i^{\langle m\rangle} &= 0.B_iB_{i+1}\ldots B_{i+m-1}B_1^{(i)}B_2^{(i)}\ldots . \end{aligned}$$

We rename these variates by $Z_k := Y_k^{\langle m\rangle}$ for $k = 1, \ldots, i$, and circularly complete the $Z_k$ as follows:

$$\begin{aligned} Z_{i+1} &:= 0.B_{i+1}B_{i+2}\ldots B_{i+m-1}B_1B_1^{(i+1)}B_2^{(i+1)}\ldots \\ Z_{i+2} &:= 0.B_{i+2}B_{i+3}\ldots B_{i+m-1}B_1B_2B_1^{(i+2)}B_2^{(i+2)}\ldots \\ &\vdots \\ Z_{i+m-1} &:= 0.B_{i+m-1}B_1B_2\ldots B_{m-1}B_1^{(i+m-1)}B_2^{(i+m-1)}\ldots \end{aligned}$$

Define $Z_k := Z_{k-i-m+1}$ for $k \geq i+m$, and let $S$ be a random index uniformly distributed on $\{1, \ldots, i+m-1\}$, and independent of the other quantities. Subsequently we will repeatedly use the fact that, by the cyclic nature of the sequence $(Z_k)$, the vectors $(Z_S, Z_{S+1}, \ldots, Z_{S+i+m-2})$ and $(Z_1, \ldots, Z_{i+m-1})$ are identically distributed. We write

$$\begin{aligned} p_i(x) &= \mathbb{P}(Y_i^{\langle m\rangle}, x \triangleright Y_1^{\langle m\rangle}, \ldots, Y_{i-1}^{\langle m\rangle}) \\ &= \mathbb{P}(\{Y_i^{\langle m\rangle}, x \triangleright Y_1^{\langle m\rangle}, \ldots, Y_{i-1}^{\langle m\rangle}\} \cap \{Y_i^{\langle m\rangle} < x\}) \\ &\quad + \mathbb{P}(\{Y_i^{\langle m\rangle}, x \triangleright Y_1^{\langle m\rangle}, \ldots, Y_{i-1}^{\langle m\rangle}\} \cap \{Y_i^{\langle m\rangle} \geq x\}). \end{aligned} \tag{6}$$

We bound the first summand in the latter expression. The second one can be treated similarly. We have

$$\{Y_i^{\langle m\rangle}, x \triangleright Y_1^{\langle m\rangle}, \ldots, Y_{i-1}^{\langle m\rangle}\} \cap \{Y_i^{\langle m\rangle} < x\} = \{Z_i = \max\{Z_k : k \leq i, Z_k \leq x\}\}.$$

Note that $\{Z_i = \max\{Z_k : k \leq i, Z_k \leq x\}\}$ implies that $Z_i$ is one of the $m$ largest among all $Z_1, \ldots, Z_{i+m-1}$ with $Z_k \leq x$ for $k = 1, \ldots, i+m-1$. Since $(Z_S, Z_{S+1}, \ldots, Z_{S+i+m-2})$ and $(Z_1, \ldots, Z_{m+i-1})$ are identically distributed, the probability for that is the same as for $Z_{i-1+S}$ being one of the $m$ largest among $Z_S, \ldots, Z_{S+i+m-2}$. Conditioned on $Z_1, \ldots, Z_{i+m-1}$, which is the same as conditioning on the whole sequence $(Z_k)$, this probability is at most $m/(i+m-1)$ since $S$ is uniformly distributed on $\{1, \ldots, i+m-1\}$ and has at most $m$ choices. Note

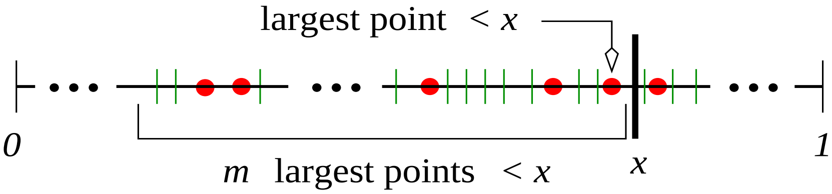

Figure 3: *The interval* $[0,1]$ *is shown with the* $m$ *largest of the points* $Z_1,\dots,Z_{i+m-1}$ *less than* $x$, *where the (green) lines mark those of these* $m$ *points belonging to* $\{Z_1,\dots,Z_i\}$ *and the (red) dots mark the corresponding points from* $\{Z_{i+1},\dots,Z_{i+m-1}\}$.

that $S$ has $m$ choices if at least $m$ of the points $Z_1,\dots,Z_{i+m-1}$ are $\le x$ and less than $m$ choices otherwise. Thus we have

$$\mathbb{P}(Z_i = \max\{Z_k : k \le i, Z_k \le x\}) \le \frac{m}{i+m} \le \frac{m}{i}.$$

Since the second term in (6) can be estimated similarly we obtain the assertion of the Lemma for $x \notin [m^2/i, 1-m^2/i]$.

Subsequently, we assume $m^2/i \le 1/2$ and $x \in [m^2/i, 1-m^2/i]$. We have the disjoint decomposition

$$\begin{aligned}
&\{Z_i = \max\{Z_k : k \le i, Z_k \le x\}\} \\
&= \quad \{Z_i = \max\{Z_k : k \le i+m-1, Z_k \le x\}\} \\
&\qquad \cup \Big(\{Z_i = \max\{Z_k : k \le i, Z_k \le x\}\} \\
&\qquad\qquad \cap \{Z_i \ne \max\{Z_k : k \le i+m-1, Z_k \le x\}\}\Big) \\
&=: \quad E_1 \cup E_1',
\end{aligned}$$

hence $\mathbb{P}(Z_i = \max\{Z_k : k \le i, Z_k \le x\}) = \mathbb{P}(E_1) + \mathbb{P}(E_1')$.

Using the fact that $(Z_S, Z_{S+1},\dots,Z_{S+i+m-2})$ and $(Z_1,\dots,Z_{i+m-1})$ are identically distributed we argue, by conditioning on the sequence $(Z_k)$, as above as follows: Conditioned on $(Z_k)$ and that there is at least one of the $Z_k$ with $Z_k \le x$ we have one possible choice for $S$ and thus in this case the conditional probability of $E_1$ is $1/(i+m-1)$. Clearly the conditional probability of $E_1$ is zero if there is no $Z_k$ with $Z_k \leq x$. Hence, we have

$$\begin{aligned}\mathbb{P}(E_1) &= \mathbb{P}(Z_{S+i-1} = \max\{Z_{S+k-1} : 1 \leq k \leq i+m-1, Z_{S+k-1} \leq x\}) \\ &= \frac{1}{i+m-1}\mathbb{P}\biggl(\bigcup_{k=1}^{i+m-1}\{Z_k \leq x\}\biggr).\end{aligned}$$

Since $x \geq m^2/i$ and $m^2/i \leq 1/2$ we obtain, denoting $b = \lfloor i/m \rfloor - 1$,

$$\begin{aligned}\mathbb{P}\biggl(\bigcap_{k=1}^{i+m-1}\{Z_k > x\}\biggr) &\leq \mathbb{P}\biggl(\bigcap_{k=0}^{b}\{Z_{1+km} > x\}\biggr) \\ &= (1-x)^{b+1} \\ &\leq \Bigl(1 - \frac{m^2}{i}\Bigr)^{i/m-1} \\ &\leq 2\exp(-m) \\ &= O\Bigl(\frac{1}{n^{18}}\Bigr).\end{aligned}$$

Together we obtain $\mathbb{P}(E_1) = 1/(i+m-1) + O(n^{-17})$. This term will lead to the main term $2/i$ in the representation of $p_i(x)$. The contribution of $\mathbb{P}(E_1')$ gives the error terms and thus can be bounded from above.

For this we define $\Delta := m^2/i$. For $x \geq \Delta$ and with $I = [x-\Delta, x]$ we have

$$\begin{aligned}E_1' \quad &\subseteq \quad \{\exists\, 1 \leq k \leq i+m-1 : Z_k, Z_{k+1}, \dots, Z_{k+i-1} \notin I\} \\ &\qquad \cup \Bigl(\{Z_i = \max\{Z_k : k \leq i, Z_k \leq x\}\} \\ &\qquad\qquad \cap \{Z_i \neq \max\{Z_k : k \leq i+m-1, Z_k \leq x\}\} \cap \{Z_i \in I\}\Bigr) \\ &=: \quad E_2 \cup E_3.\end{aligned}$$

Using the fact that $Z_1, Z_{1+m}, Z_{1+2m}, \dots$ are independent and that $1 - \Delta \geq 1/2$, we obtain

$$\begin{aligned}\mathbb{P}(E_2) &\leq (i+m-1)\mathbb{P}(Z_1, \dots, Z_i \notin I) \\ &\leq (i+m-1)\mathbb{P}(Z_1, Z_{1+m}, Z_{1+2m}, \dots \notin I) \\ &\leq (i+m-1)(1-\Delta)^{i/m-1} \\ &\leq 2(i+m-1)\exp(-\Delta i/m) \\ &\leq 2(i+m-1)\exp(-m) \\ &= O(n^{-17}) \\ &= O(i^{-3/2}).\end{aligned}$$

For the estimate of $\mathbb{P}(E_3)$, we first associate an event $E_3(S)$ similarly as for the analysis of $E_1$,

$$\begin{aligned} E_3(S) &:= \{Z_{S+i-1} = \max\{Z_{S+k-1} : 1 \le k \le i, Z_{S+k-1} \le x\}\} \\ &\quad \cap \{Z_{S+i-1} \neq \max\{Z_{S+k-1} : 1 \le k \le i+m-1, Z_{S+k-1} \le x\}\} \\ &\quad \cap \{Z_{S+i-1} \in I\}. \end{aligned}$$

We have $\mathbb{P}(E_3) = \mathbb{P}(E_3(S))$ since $(Z_S, Z_{S+1}, \dots, Z_{S+i+m-2})$ and $(Z_1, \dots, Z_{i+m-1})$ are identically distributed. Note that the probability of $E_3(S)$ conditioned on any event involving only the sequence $(Z_k)$ is at most $m/(n+m-1)$ since $S$ has at most $m$ choices out of $n+m-1$ equally likely indices. These are the choices such that $Z_{S+i-1}$ is among the $m$ largest of the points $Z_1, \dots, Z_{i+m-1}$ less or equal than $x$, cf. Figure 3. We condition on

$$F := \bigcup_{i=1}^{n+m-1} \left( \{Z_i \in I\} \cap \bigcup_{k=1}^{m-1} \{|Z_i - Z_{i+k}| \le \Delta\} \right).$$

Clearly, $\mathbb{P}(E_3(S) \,|\, F^c) = 0$, hence we obtain

$$\begin{aligned} \mathbb{P}(E_3) &= \mathbb{P}(E_3(S)) \\ &= \mathbb{P}(E_3(S) \,|\, F)\mathbb{P}(F) \\ &\le \frac{m}{n+m-1}\mathbb{P}\left( \bigcup_{i=1}^{n+m-1} \left( \{Z_i \in I\} \cap \bigcup_{k=1}^{m-1} \{|Z_i - Z_{i+k}| \le \Delta\} \right)\right) \\ &\le m\mathbb{P}\left( \{Z_1 \in I\} \cap \bigcup_{k=1}^{m-1} \{|Z_1 - Z_{1+k}| \le \Delta\} \right) \\ &\le m\mathbb{P}(X_1 \in I^+ \cap B_n(\Delta + 2^{37}/n^{18}\})), \end{aligned} \tag{7}$$

where, for $I^+$, we use the notation $[a,b]^+ := [a - 2^{36}/n^{18}, b + 2^{36}/n^{18}]$ for intervals $[a,b]$. Note that we have $|Z_k - X_k| \le 1/2^m$ for $k = 1, \dots, m$ and $m \ge 18\log_2 n - 36$. For $n$ sufficiently large we have $2^{36}/n^{18} \le \Delta$ and thus $I^+ \subseteq \bar{I} := [x - 2\Delta, x + \Delta]$. With the representation given in (2) for $\{|X_1 - X_{1+k}| \le 3\Delta\}$ we find that $B_n(3\Delta)$ intersects $\bar{I}$ at most in $3\Delta(2^k - 1) + 2$ intervals of lengths at most $6\Delta/(2^k - 1)$. This implies the bound

$$\begin{aligned} \mathbb{P}(E_3) &\le m \sum_{k=1}^{m-1} \left( 18\Delta^2 + \frac{12\Delta}{2^k - 1} \right) \\ &\le 18m^2\Delta^2 + 24m\Delta \\ &= \frac{18m^6}{i^2} + \frac{24m^3}{i}. \end{aligned} \tag{8}$$

Together with $\mathbb{P}(E_2)$ this yields an error term of the order of $R_2(n,i)$.

Finally, we consider $x \notin B_n(\Delta^*)$ with $\Delta^* := 2\sqrt{n}\Delta$ and refine the estimate in (8). For $x \notin B_n(\Delta^*)$ we get a contribution of $\{|X_1 - X_{1+k}| \leq 3\Delta\} \cap \bar{I}$ in (7) only if $3\Delta/(2^k-1) + 2\Delta > \Delta^*/(2^k-1)$, see Figure 4, which holds exactly for $k > \log_2(\sqrt{n} - 1/2)$. Therefore for $x \notin B_n(\Delta^*)$ the summation in (8) can be refined

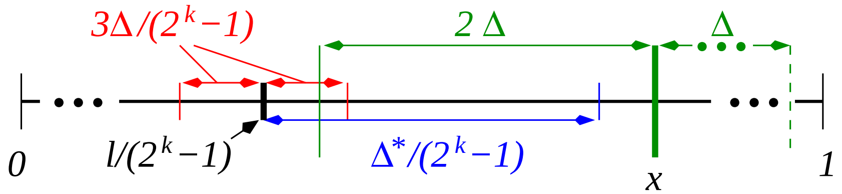

Figure 4: *Shown is a case, where the part* $\{|T^k(x) - x| < 3\Delta\}$ *of the bad set* $B_n(3\Delta)$ *(in red) intersects the interval* $[x - 2\Delta, x + \Delta]$ *(in green), while* $x$ *is outside the part* $\{|T^k(x) - x| < \Delta^*\}$ *of the bad set* $B_n(\Delta^*)$ *(in blue).*

to

$$\begin{aligned}\mathbb{P}(E_3) &\leq m \sum_{k=\lceil \log_2(\sqrt{n}-1/2)\rceil}^{m-1} \left(18\Delta^2 + \frac{12\Delta}{2^k-1}\right) \\ &\leq 18m^2\Delta^2 + \frac{48m\Delta}{\sqrt{n}} \\ &\leq \frac{18m^6}{i^2} + \frac{48m^3}{i^{3/2}}.\end{aligned}$$

We have estimated the first summand in (6) for all different ranges of $x$ appearing in Lemma 5.1. Since the second summand in (6) can be estimated similarly we obtain for all $x \in [0,1]$ and $1 \leq i \leq n-1$,

$$\begin{aligned}p_i(x) &= \mathbf{1}_{[m^2/i, 1-m^2/i]}(x)\left(\frac{2}{i+m-1} + R_1(n,i) + \mathbf{1}_{B_n(2m^2/\sqrt{i})}(x)R_2(n,i)\right) \\ &\quad + \left(1 - \mathbf{1}_{[m^2/i,1-m^2/i]}(x)\right)R_3(n,i),\end{aligned}$$

with orders for $R_k(n,i)$, $k = 1,2,3$, as in the Lemma. Since we have $|2/i - 2/(i + m - 1)| \leq C(\log n)/i^2$ for some constant $C > 0$ the assertion follows. ■

## 6 Expansion of the mean of the depth

In this section we find the mean of $D_n$:

**Theorem 6.1** *The depth $D_n$ of the $n$-th node inserted into a random suffix search tree satisfies*

$$\mathbb{E}D_n = 2\log n + O(\log^2 \log n).$$

**Proof:** We recall the events $A_j = \{X_j \text{ is ancestor of } X_n \text{ in the tree}\}$ and the representations

$$D_n = \sum_{j=1}^{n-1} \mathbf{1}_{A_j}, \quad \mathbb{E}D_n = \sum_{j=1}^{n-1} \mathbb{P}(A_j).$$

For the estimate of $\mathbb{P}(A_j)$ we distinguish three ranges for the index $j$, namely $1 \le j \le \lceil \log_2^{12} n \rceil$, $\lceil \log_2^{12} n \rceil < j \le n - m$, and $n - m < j < n$, where we choose $m = 18\|\lfloor \log_2 n \rfloor\|$.

The range $1 \le j \le \lceil \log_2^{12} n \rceil$: Note that $\sum_{j=1}^{\lceil \log_2^{12} n \rceil} \mathbf{1}_{A_j}$ is bounded from above by the height of the random suffix search tree with $\lceil \log_2^{12} n \rceil$ nodes. Thus, by Lemma 4.2, we obtain

$$\sum_{j=1}^{\lceil \log_2^{12} n \rceil} \mathbb{P}(A_j) \le \mathbb{E}H_{\lceil \log_2^{12} n \rceil} = O(\log_2^2 \log_2^{12} n)) = O(\log^2 \log n).$$

The range $\lceil \log_2^{12} n \rceil < j \le n - m$: We start, using Lemma 3.1, with the representation

$$\begin{aligned} \mathbb{P}(A_j) &= \mathbb{P}(X_j, X_n \rhd X_1, \ldots, X_{j-1}) \\ &= \mathbb{P}(Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \rhd Y_1^{\langle m \rangle}, \ldots, Y_{j-1}^{\langle m \rangle}) + O(1/n^2) \\ &= \mathbb{P}(A_j^{\langle m \rangle}) + O(1/n^2). \end{aligned}$$

Note that $Y_n^{\langle m \rangle}$ is independent of $Y_1^{\langle m \rangle}, \ldots, Y_j^{\langle m \rangle}$, since $j \le n - m$. Thus for the calculation of $\mathbb{P}(A_j^{\langle m \rangle})$ we may condition on $Y_n^{\langle m \rangle}$. With the notation of Lemma 5.1 and using the fact that $Y_n^{\langle m \rangle}$ is $U[0,1]$ distributed this yields for all $1 \le j \le n - m$,

$$\mathbb{P}(A_j^{\langle m \rangle}) = \mathbb{E}[p_j(Y_n^{\langle m \rangle})] = \frac{2}{j} + R_{n,j}, \quad |R_{n,j}| \le C\frac{\log^6 n}{j^{3/2}},$$

for some constant $C > 0$. When summing note that

$$\sum_{j=\lceil \log_2^{12} n \rceil}^{\infty} \frac{\log^6 n}{j^{3/2}} \le \log^6 n \int_{\lceil \log_2^{12} n \rceil - 1}^{\infty} \frac{1}{x^{3/2}}\, dx = O(1).$$

We obtain

$$\sum_{j=\lceil \log_2^{12} n\rceil}^{n-m} \mathbb{P}(A_j) = \sum_{j=\lceil \log_2^{12} n\rceil}^{n-m} \Big(\frac{2}{j} + R_{n,j} + O\Big(\frac{1}{n^2}\Big)\Big) = 2\log n + O(\log\log n).$$

Hence, this range gives the main contribution.

The range $n - m < j < n - 1$: With $q := \lfloor j/m \rfloor - 1$ we have

$$\begin{aligned}
\mathbb{P}(A_j) &= \mathbb{P}(X_j, X_n \triangleright X_1, \dots, X_{j-1}) \\
&\leq \mathbb{P}(X_j, X_n \triangleright X_{j-m}, \dots, X_{j-qm}) \\
&= \mathbb{P}(Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_{j-m}^{\langle m\rangle}, \dots, Y_{j-qm}^{\langle m\rangle}) + O(1/n^2).
\end{aligned}$$

We have, using Lemma 3.2, for $n$ sufficiently large,

$$\begin{aligned}
&\mathbb{P}(Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_{j-m}^{\langle m\rangle}, \dots, Y_{j-qm}^{\langle m\rangle}) \qquad (9)\\
&\leq \mathbb{P}(|Y_j^{\langle m\rangle} - Y_n^{\langle m\rangle}| < m^2/j) \\
&\qquad + \mathbb{P}\Big(\{|Y_j^{\langle m\rangle} - Y_n^{\langle m\rangle}| \geq m^2/j\} \cap \{Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_{j-m}^{\langle m\rangle}, \dots, Y_{j-qm}^{\langle m\rangle}\}\Big) \\
&\leq 8\frac{m^2}{j} + \Big(1 - \frac{m^2}{j}\Big)^{j/m-2} \\
&\leq 8\frac{m^2}{j} + 4\exp(-m) \\
&\leq 8\frac{m^2}{j} + O\left(\frac{1}{n^{18}}\right).
\end{aligned}$$

The summation yields

$$\sum_{j=n-m}^{n-1} \mathbb{P}(A_j) = O(1),$$

so that the third range makes an asymptotically negligible contribution. Collecting the estimates of the three ranges, we obtain the assertion. ■

## 7 A weak law of large numbers

In this section we prove a weak law of large numbers for the depth $D_n$.

**Theorem 7.1** *We have $D_n / \mathbb{E} D_n \to 1$ in probability as $n \to \infty$.*

**Proof:** Let $\varepsilon, \varepsilon' > 0$ be given. We have to show

$$\mathbb{P}\left(\left|\frac{D_n}{\mathbb{E}D_n} - 1\right| > \varepsilon\right) < \varepsilon'$$

for all $n$ sufficiently large. We define the decomposition $D_n = D_n^* + D_n^{**}$, where

$$D_n^* := \sum_{j=\lfloor \log^{68} n\rfloor}^{\lfloor n/2\rfloor} \mathbf{1}_{A_j}, \qquad D_n^{**} := \sum_{j=1}^{\lfloor \log^{68} n\rfloor - 1} \mathbf{1}_{A_j} + \sum_{j=\lfloor n/2\rfloor+1}^{n-1} \mathbf{1}_{A_j}.$$

From Theorem 6.1 we have $\mathbb{E}D_n^* \sim 2\log n$ and $\mathbb{E}D_n^{**} = O(\log\log n)$. We bound the summands in the estimate

$$\mathbb{P}\left(\left|\frac{D_n}{\mathbb{E}D_n} - 1\right| > \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{D_n^*}{\mathbb{E}D_n} - 1\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\frac{D_n^{**}}{\mathbb{E}D_n} > \frac{\varepsilon}{2}\right) \tag{10}$$

separately. By Markov's inequality we have

$$\mathbb{P}\left(\frac{D_n^{**}}{\mathbb{E}D_n} > \frac{\varepsilon}{2}\right) \leq \frac{\mathbb{E}D_n^{**}}{(\varepsilon/2)\,\mathbb{E}D_n} = O\left(\frac{\log\log n}{\log n}\right) \leq \varepsilon'/2$$

for all $n$ suffitiently large. Thus we only need the first summand in (10) to be at most $\varepsilon'/2$. By Chebychev's inequality this is implied by

$$\frac{\mathrm{Var}(D_n^*)}{(\mathbb{E}D_n)^2} \to 0,$$

as $n \to \infty$. Since we have $(\mathbb{E}D_n)^2 = \Omega(\log^2 n)$ and $(\mathbb{E}D_n^*)^2 \sim 4\log^2 n$, it is sufficient for completing the proof of Theorem 7.1 to establish

$$\mathbb{E}\left[(D_n^*)^2\right] = \sum_{\lfloor \log^{68} n\rfloor \leq i \leq j \leq \lfloor n/2\rfloor} \mathbb{P}(A_i \cap A_j) \sim 4\log^2 n.$$

Since the contribution of the summands with $i = j$ is of the order $O(\log n)$ we may additionally assume $i < j$. We distinguish the cases where $j - i > \log^{32} n$ and $j - i \leq \log^{32} n$.

The case $j - i \leq \log^{32} n$: We have

$$\begin{aligned}\mathbb{P}(A_i \cap A_j) \;&=\; \mathbb{P}\Big(A_i \cap A_j \cap \{|X_i - X_j| \geq (2\log^2 n)/i\}\Big) \\ &\quad + \mathbb{P}\Big(A_i \cap A_j \cap \{|X_i - X_j| < (2\log^2 n)/i\}\Big).\end{aligned} \tag{11}$$

For all $n$ large enough we obtain with $b := \lfloor i/m \rfloor - 1$, Lemma 3.1, and $(\log^2 n)/i \leq 1/2$,

$$\begin{aligned}
&\mathbb{P}\Big(A_i \cap A_j \cap \{|X_i - X_j| \geq (2\log^2 n)/i\}\Big) \\
&\quad \leq \mathbb{P}\Big(A_i^{\langle m \rangle} \cap A_j^{\langle m \rangle} \cap \{|Y_i^{\langle m \rangle} - Y_j^{\langle m \rangle}| \geq (\log^2 n)/i\}\Big) + \frac{8}{n^2} \\
&\quad \leq \mathbb{P}\Big(\{Y_i^{\langle m \rangle}, Y_j^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, \dots , Y_{i-1}^{\langle m \rangle}\} \\
&\qquad\qquad \cap \{|Y_i^{\langle m \rangle} - Y_j^{\langle m \rangle}| \geq (\log^2 n)/i\}\Big) + \frac{8}{n^2} \\
&\quad \leq \mathbb{P}\Big(\{Y_i^{\langle m \rangle}, Y_j^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, Y_{1+m}^{\langle m \rangle} \dots , Y_{1+bm}^{\langle m \rangle}\} \\
&\qquad\qquad \cap \{|Y_i^{\langle m \rangle} - Y_j^{\langle m \rangle}| \geq (\log^2 n)/i\}\Big) + \frac{8}{n^2} \\
&\quad \leq \left(1 - \frac{\log^2 n}{i}\right)^{i/m-2} + \frac{8}{n^2} \\
&\quad \leq 4\exp\left(-\frac{\log^2 n}{m}\right) + \frac{8}{n^2} \\
&\quad \leq \frac{12}{n^2}.
\end{aligned}$$

For the last summand in (11) we introduce the lengths of the spacings formed by $X_1, \dots , X_n$ on $[0, 1]$ by $S_j^n := X_{(j+1)} - X_{(j)}$ for $j = 1, \dots , n-1$ and $S_0^n := X_{(1)}$, $S_n^n := 1 - X_{(n)}$, where $X_{(j)}$ denotes the $j$-th order statistic of $X_1, \dots , X_n$. Furthermore we denote the maximal spacing $M_i := \max_{0 \leq k \leq i-m} S_k^{i-m}$ of the $X_1, \dots , X_{i-m}$ and correspondingly, $M_i^{\langle m \rangle}$ for the maximum spacing of the perturbed variates. For $n$ sufficiently large, and with $n - i > m$, we have

$$\begin{aligned}
&\mathbb{P}\Big(A_i \cap A_j \cap \{|X_i - X_j| < (2\log^2 n)/i\}\Big) \\
&\quad \leq \mathbb{P}\Big(A_i \cap A_j \cap \{|X_i - X_j| < (2\log^2 n)/i\} \cap \{M_i \leq 1/\sqrt{i}\}\Big) \\
&\qquad + \mathbb{P}\Big(A_i \cap A_j \cap \{|X_i - X_j| < (2\log^2 n)/i\} \cap \{M_i > 1/\sqrt{i}\}\Big) \\
&\quad \leq \mathbb{P}\Big(A_i^{\langle m \rangle} \cap A_j^{\langle m \rangle} \cap \{|Y_i^{\langle m \rangle} - Y_j^{\langle m \rangle}| < (4\log^2 n)/i\} \cap \{M_i^{\langle m \rangle} \leq 2/\sqrt{i}\}\Big) \qquad (12) \\
&\qquad + \mathbb{P}\Big(\{M_i^{\langle m \rangle} > 1/(2\sqrt{i})\}\Big) + \frac{16}{n^2}.
\end{aligned}$$

For the estimate of $\mathbb{P}(\{M_i^{\langle m\rangle} > 1/(2\sqrt{i})\})$ note that for any $0 \le x \le 1/2$ we have, with $b = \lfloor i/m \rfloor - 1$,

$$\begin{aligned}
&\mathbb{P}\Big(\{M_i^{\langle m\rangle} > 2x\}\Big) && (13)\\
&\le \mathbb{P}\left(\bigcup_{\ell=1}^{\lceil 1/x\rceil - 1} \{Y_1^{\langle m\rangle}, \dots, Y_{i-m}^{\langle m\rangle} \notin [(\ell-1)x, \ell x]\} \cup \{Y_1^{\langle m\rangle}, \dots, Y_{i-m}^{\langle m\rangle} \notin [1-x, x]\}\right)\\
&\le \lceil 1/x \rceil \sup_{y\in[0,1-x]} \mathbb{P}\left(Y_1^{\langle m\rangle}, Y_{1+m}^{\langle m\rangle}, \dots, Y_{1+bm}^{\langle m\rangle} \notin [y, y+x]\right)\\
&\le \lceil 1/x \rceil (1-x)^{i/m-2}\\
&\le 4\lceil 1/x \rceil \exp\left(-\frac{xi}{m}\right).
\end{aligned}$$

Using this with $x = 1/(4\sqrt{i})$ we obtain $\mathbb{P}(\{M_i^{\langle m\rangle} > 1/(2\sqrt{i})\}) \le (16\sqrt{i} + 4)\exp(-\sqrt{i}/(4m)) \le 1/n^2$ for $n$ sufficiently large, since we have $i \ge \log^{68} n$.

It remains to bound the term in (12). Note that $A_i^{\langle m\rangle}$ in particular implies that $\{Y_i^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle}, \dots, Y_{i-m}^{\langle m\rangle}\}$. Under $\{M_i^{\langle m\rangle} \le 2/\sqrt{i}\}$ this implies $\{|Y_i^{\langle m\rangle} - Y_n^{\langle m\rangle}| \le 2/\sqrt{i}\}$. Hence, using Lemma 3.2 and that $Y_n^{\langle m\rangle}$ is independent of $(Y_i^{\langle m\rangle}, Y_j^{\langle m\rangle})$ we obtain

$$\begin{aligned}
&\mathbb{P}\Big(A_i^{\langle m\rangle} \cap A_j^{\langle m\rangle} \cap \{|Y_i^{\langle m\rangle} - Y_j^{\langle m\rangle}| < (4\log^2 n)/i\} \cap \{M_i^{\langle m\rangle} \le 2/\sqrt{i}\}\Big)\\
&\quad\le \mathbb{P}\Big(\{|Y_i^{\langle m\rangle} - Y_j^{\langle m\rangle}| < (4\log^2 n)/i\} \cap \{|Y_i^{\langle m\rangle} - Y_n^{\langle m\rangle}| \le 2/\sqrt{i}\}\Big)\\
&\quad\le \frac{128\log^2 n}{i^{3/2}}.
\end{aligned}$$

Combining all this yields $\mathbb{P}(A_i \cap A_j) \le (C\log^2 n)/i^{3/2}$ for an appropriate constant $C > 0$. Therefore, the contribution of this range is

$$\begin{aligned}
\sum_{\substack{i \ge \lfloor \log^{68} n\rfloor \\ i<j\le i+\lfloor \log^{32} n\rfloor}} \mathbb{P}(A_i \cap A_j) &\le C\log^{34} n \sum_{i \ge \lfloor \log^{68} n\rfloor} \frac{1}{i^{3/2}}\\
&\le C\log^{34} n \int_{\lfloor \log^{68} n\rfloor - 1}^{\infty} x^{-3/2}\,dx\\
&= O(1).
\end{aligned}$$

The case $j-i>\log^{32} n$: We have

$$\begin{aligned}
\mathbb{P}(A_i\cap A_j) &= \mathbb{P}\Big(\{X_i,X_n \triangleright X_1,\dots,X_{i-1}\}\cap\{X_j,X_n\triangleright X_1,\dots,X_{j-1}\}\Big)\\
&\le \mathbb{P}\Big(\{Y_i^{\langle m\rangle},Y_n^{\langle m\rangle}\triangleright Y_1^{\langle m\rangle},\dots,Y_{i-1}^{\langle m\rangle}\}\\
&\qquad \cap\{Y_j^{\langle m\rangle},Y_n^{\langle m\rangle}\triangleright Y_1^{\langle m\rangle},\dots,Y_{j-1}^{\langle m\rangle}\}\Big)+\frac{8}{n^2}\\
&\le \mathbb{P}\Big(\{Y_i^{\langle m\rangle},Y_n^{\langle m\rangle}\triangleright Y_1^{\langle m\rangle},\dots,Y_{i-1}^{\langle m\rangle}\}\\
&\qquad \cap\{Y_j^{\langle m\rangle},Y_n^{\langle m\rangle}\triangleright Y_{i+m+1}^{\langle m\rangle},\dots,Y_{j-1}^{\langle m\rangle}\}\Big)+\frac{8}{n^2},
\end{aligned}$$

where we assume that $n$ is sufficiently large such that $\log^{32} n>2m$. Conditioned on $Y_n^{\langle m\rangle}$ these two events are independent. This implies

$$\mathbb{P}(A_i\cap A_j)\le \mathbb{E}[p_i(Y_n^{\langle m\rangle})p_{j-i-m}(Y_n^{\langle m\rangle})]+\frac{8}{n^2}.$$

We abbreviate $\ell:=j-i-m$ and $s:=i\wedge\ell$. Thus from Lemma 5.1, for appropriate constants $C,C'>0$,

$$\begin{aligned}
\mathbb{E}\,\Big[\big(1-\mathbf{1}_{[m^2/s,1-m^2/s]}(Y_n^{\langle m\rangle})\big)p_i(Y_n^{\langle m\rangle})p_\ell(Y_n^{\langle m\rangle})\Big] &\le \frac{Cm^4}{si\ell},\\
\mathbb{E}\,\Big[\mathbf{1}_{B_n(2m^2/\sqrt{s})}(Y_n^{\langle m\rangle})p_i(Y_n^{\langle m\rangle})p_\ell(Y_n^{\langle m\rangle})\Big] &\le \frac{C'm^{15}}{\sqrt{s}i\ell}.
\end{aligned}$$

Note that in the last estimate we used Lemma 3.2 to obtain $\lambda(B_n(2m^2/\sqrt{s}))\le 4m^3/\sqrt{s}$, where $\lambda$ denotes Lebesgue measure. Therefore, with an appropriate constant $C''>0$ and $C_1$ as in Lemma 5.1, we have

$$\mathbb{P}(A_i\cap A_j)\le\left(\frac{2}{i}+\frac{C_1m^6}{i^{3/2}}\right)\left(\frac{2}{\ell}+\frac{C_1m^6}{\ell^{3/2}}\right)+\frac{C''m^{15}}{\sqrt{s}i\ell}+\frac{8}{n^2}$$

Thus we obtain

$$\begin{aligned}
&\sum_{\substack{\lfloor \log^{68} n\rfloor \le i \le \lfloor n/2\rfloor \\ i+\lfloor \log^{32} n\rfloor \le j \le \lfloor n/2\rfloor}} \mathbb{P}(A_i \cap A_j) \\
\le\ & \sum_{\substack{\lfloor \log^{68} n\rfloor \le i \le \lfloor n/2\rfloor \\ \lfloor \log^{32} n\rfloor - m \le \ell \le \lfloor n/2\rfloor}} \left(\left(\frac{2}{i} + \frac{C_1 m^6}{i^{3/2}}\right)\left(\frac{2}{\ell} + \frac{C_1 m^6}{\ell^{3/2}}\right) + \frac{C'' m^{15}}{\sqrt{s i \ell}} + \frac{8}{n^2}\right) \\
\le\ & 4\log^2 n + O(\log n) + 2C'' m^{15} \sum_{\lfloor \log^{32} n\rfloor - m \le s \le r \le \lfloor n/2\rfloor} \left(\frac{1}{s^{3/2} r}\right) + O(1) \\
\le\ & 4\log^2 n + O(\log n) + 2C'' m^{16} \sum_{s \ge \lfloor \log^{32} n\rfloor - m} \frac{1}{s^{3/2}} \\
=\ & 4\log^2 n + O(\log n) + O(1).
\end{aligned}$$

The assertion follows. ■

## 8 Further analysis of the model

In the remaining sections, we analyze the random suffix search tree from another perspective, based on the spacings defined by $X_1, \dots, X_n$ on $[0,1]$. This approach provides some new insight, and bears many fruits, as it permits us to analyze the size of the subtrees at the nodes. We begin with four auxiliary lemmas in the present section, and obtain the fundamental limit theorem for the size of a random spacing in the next section. The implications for the random suffix search tree are explained in section 11.

**Lemma 8.1** *Let $I$ be an interval in $[0,1]$ of length $|I|$. Then for all $1 \le i \le -\log_2 |I|$ we have*

$$\mathbb{P}(X_1, X_{1+i} \in I) \le \frac{|I|}{2^i}.$$

**Proof:** With the map $T(x) := \{2x\}$ we have $X_{1+i} = T^i(X_1)$ and $\{X_{1+i} \in I\} = \{X_1 \in T^{-i}(I)\}$, where $T^{-i}$ is the $i$-th iterate of the inverse image of $T$. With $I = [x, x+\Delta]$ we obtain the representation

$$T^{-i}(I) = \bigcup_{k=1}^{2^i} I_k^i, \quad I_k^i := \left[\frac{k-1}{2^i} + \frac{x}{2^i}, \frac{k-1}{2^i} + \frac{x+\Delta}{2^i}\right].$$

Since $i \leq -\log_2 |I|$, the interval $I$ can only cover either one of the intervals $I_k^i$ or parts of two consecutive $I_k^i$s with total length covered being at most $|I|/2^i$. This implies

$$\mathbb{P}(X_1, X_{1+i} \in I) = |I \cap T^{-i}(I)| \leq \frac{|I|}{2^i}. \;\blacksquare$$

**Lemma 8.2** *For all integer $1 \leq i < j$, $t \geq 1$ and real $\varepsilon > 0$, and $U$ being $U[0,1]$ distributed and independent of $X_1, Y_i^{\langle t \rangle}, Y_j^{\langle t \rangle}$ we have*

$$\mathbb{P}(X_i, X_j \in [U, U+\varepsilon]) \leq 2\varepsilon^2, \quad \mathbb{P}(Y_i^{\langle t \rangle}, Y_j^{\langle t \rangle} \in [U, U+\varepsilon]) \leq 8\varepsilon^2.$$

**Proof:** With Lemma 3.2 we have

$$\begin{aligned} &\mathbb{P}\Big(X_i, X_j \in [U, U+\varepsilon]\Big) \\ &= \mathbb{P}\Big(|X_i - X_j| \leq \varepsilon\Big)\mathbb{P}\Big(X_i, X_j \in [U, U+\varepsilon] \,\Big|\, |X_i - X_j| \leq \varepsilon\Big) \\ &\leq 2\varepsilon\varepsilon = 2\varepsilon^2. \end{aligned}$$

The second statement follows analogously. ■

**Lemma 8.3** *For any Borel set $A \subseteq [0,1]$, real $\varepsilon, \delta > 0$, integer $i \geq 0$, and $U$ being $U[0,1]$ distributed we have*

$$\mathbb{P}(\lambda(T^{-i}((U, U+\varepsilon)) \cap A) \geq \delta) \leq \frac{\varepsilon\lambda(A)}{\delta},$$

*where $\lambda(\,\cdot\,)$ denotes Lebesgue measure.*

**Proof:** Applying Markov's inequality and Fubini's theorem we obtain

$$\begin{aligned} \mathbb{P}(\lambda(T^{-i}((U, U+\varepsilon)) \cap A) \geq \delta) &\leq \frac{1}{\delta}\mathbb{E}[\lambda(T^{-i}((U, U+\varepsilon)) \cap A)] \\ &= \frac{1}{\delta}\int_0^1 \int_0^1 \mathbf{1}_A(y)\mathbf{1}_{T^{-i}((x, x+\varepsilon))}(y)\, dy\, dx \\ &\leq \frac{1}{\delta}\int_0^1 \mathbf{1}_A(y) \sum_{j=0}^{2^i-1} \int_0^1 \mathbf{1}_{[2^i y - j - \varepsilon, 2^i y - j]}(x)\, dx\, dy. \end{aligned}$$

A nonzero contribution of the inner integrals may happen at most for two values of $j$. We obtain

$$\sum_{j=0}^{2^i-1} \int_0^1 \mathbf{1}_{[2^i y - j - \varepsilon, 2^i y - j]}(x)\, dx \leq \varepsilon,$$

uniformly over all $y \in [0,1]$. The assertion follows. ■

**Lemma 8.4** *For all $n \geq 1$, $a \in [0,1)$, and $\Delta \in (0, 1/\sqrt{n})$ with $a + \Delta \leq 1$, we have*

$$\mathbb{P}\Big(Y_1^{\langle m\rangle}, \dots, Y_{L/2}^{\langle m\rangle} \notin [a, a+\Delta]\Big) \leq 1 - \frac{L\Delta}{4} + \frac{2L}{n},$$

*where $L = \|\lfloor \log_2 n \rfloor\|$ and $m = 18L$.*

**Proof:** Since $|Y_j^{\langle m\rangle} - X_j| \leq 4/n^{18} \leq 4/n$ for $j = 1, \dots, L/2$ we have

$$\begin{aligned} &\mathbb{P}\Big(Y_1^{\langle m\rangle}, \dots, Y_{L/2}^{\langle m\rangle} \notin [a, a+\Delta]\Big) \\ &\leq \ \mathbb{P}\Big(X_1, \dots, X_{L/2} \notin [a + 4/n, a + \Delta - 4/n]\Big). \end{aligned} \tag{14}$$

Applying the Chung-Erdös inequality (Chung and Erdös 1952) and denoting $I := [a + 4/n, a + \Delta - 4/n]$, we obtain

$$\begin{aligned} &\mathbb{P}(X_1, \dots, X_{L/2} \notin I) \\ &= \ 1 - \mathbb{P}\left(\bigcup_{j=1}^{L/2} \{X_j \in I\}\right) \\ &\leq \ 1 - \frac{\left(\sum_{j=1}^{L/2} \mathbb{P}(X_j \in I)\right)^2}{\sum_{j=1}^{L/2} \mathbb{P}(X_j \in I) + \sum_{1 \leq i < j \leq L/2} \mathbb{P}(X_i, X_j \in I)}. \end{aligned} \tag{15}$$

We have

$$\sum_{1 \leq i < j \leq L/2} \mathbb{P}(X_i, X_j \in I) = \sum_{k=1}^{L/2-1} (L/2 - k) \mathbb{P}(X_1, X_{1+k} \in I),$$

and since $k \leq L/2 \leq \log_2 \sqrt{n} \leq -\log_2 \Delta \leq -\log_2 |I|$, we may apply Lemma 8.1 to obtain $\mathbb{P}(X_1, X_{1+k} \in I) \leq (\Delta - 8/n)/2^k$. Thus,

$$\sum_{1 \leq i < j \leq L/2} \mathbb{P}(X_i, X_j \in I) \leq \frac{L}{2}\left(\Delta - \frac{8}{n}\right).$$

Plugging this estimate into (15) and (14) we obtain

$$\begin{aligned} \mathbb{P}\Big(Y_1^{\langle m\rangle}, \dots, Y_{L/2}^{\langle m\rangle} \notin [a, a+\Delta]\Big) \ &\leq \ 1 - \frac{((L/2)(\Delta - 8/n))^2}{(L/2)(\Delta - 8/n) + (L/2)(\Delta - 8/n)} \\ &= \ 1 - \frac{L\Delta}{4} + \frac{2L}{n}. \ \blacksquare \end{aligned}$$

## 9 Weak convergence of a random spacing

The lengths of the spacings formed by $X_1, \dots, X_n$ on $[0,1]$ are denoted by $S_j^n := X_{(j+1)} - X_{(j)}$ for $j = 1, \dots, n-1$ and $S_0^n := X_{(1)}$, $S_n^n := 1 - X_{(n)}$, where $X_{(j)}$ denotes the $j$-th order statistic of $X_1, \dots, X_n$. In this section we provide a limit law for the rescaled length of a spacing chosen uniformly from $S_0^n, \dots, S_n^n$, where by uniform we mean that we choose one of the indices $j = 0, \dots, n$ uniformly at random. In the next section we will choose an index by into which spacing an $U[0,1]$ random variable, independent of $X_1$, falls.

**Lemma 9.1** *We have*

$$nS_{I_n}^n \xrightarrow{\mathcal{L}} E, \quad (n \to \infty), \tag{16}$$

*where $E$ is exp$(1)$-distributed, i.e., has Lebesgue-density $e^{-x}$ on $[0,\infty)$ and $I_n$ is uniformly distributed on $\{0,\dots,n\}$ and independent of $X_1$.*

This can be reduced to the following result on the spacings between fractional parts of lacunary sequences due to Rudnick and Zaharescu (2002). A *lacunary sequence* is a sequence $(a_j)_{j\geq 1}$ of integers with $\liminf_{j\to\infty} a_{j+1}/a_j > 1$. The primary example is $a_j = 2^j$. Now, for an $\alpha \in \mathbb{R}$ we define $S_j^n(\alpha)$ for $j = 0, \dots, n$ as the spacings between the fractional parts of $\alpha a_j$, $j = 1, \dots, n$, in the unit interval $[0,1]$. More precisely, for $\vartheta_j^n := \{\alpha a_j\}$ we define $S_j^n(\alpha) := \vartheta_{(j+1)} - \vartheta_{(j)}$ for $j = 1, \dots, n-1$ as well as $S_0^n(\alpha) := \vartheta_{(1)}$ and $S_n^n(\alpha) := 1 - \vartheta_{(n)}$. Then Rudnick and Zaharescu (2002) prove:

**Theorem 9.2** *Let $(a_j)$ be a lacunary sequence. Then we have for almost all $\alpha \in \mathbb{R}$ and all $0 \leq a < b$,*

$$\lim_{n\to\infty} \frac{1}{n+1} \#\{0 \leq j \leq n : nS_j^n(\alpha) \in [a,b]\} = \int_a^b e^{-x} dx. \tag{17}$$

For background, see also Kurlberg and Rudnick (1999, Appendix A). This can directly be turned into a proof of Lemma 9.1:

**Proof of Lemma 9.1**: The result of Rudnick and Zaharescu says that $nS_{I_n}^n(\alpha) \to E$ in distribution for almost all $\alpha \in \mathbb{R}$. Note that in this notation the variate $S_{I_n}^n(U)$, where $U$ is a unif$[0,1]$ distributed random variable being independent of $I_n$, coincides in distribution with the $S_{I_n}^n$ defined previously. Let $F_n^\alpha, F_n, F_E$ denote the distribution functions of $nS^n_{I_n}(\alpha)$, $nS^n_{I_n}$, and $E$ respectively. Then, by dominated convergence, for all $x \in \mathbb{R}$,

$$F_n(x) = \int_0^1 F_n^\alpha(x)\,d\alpha \longrightarrow \int_0^1 F_E(x)\,d\alpha = F_E(x), \quad (n \to \infty), \tag{18}$$

thus $nS^n_{I_n} \to E$ in distribution. ■

However, since the proof of Theorem 9.2 is rather involved, it is of interest to give a direct probabilistic proof of Lemma 9.1.

**Probabilistic proof of Lemma 9.1**: We have to show $\mathbb{P}(nS^n_{I_n} \le t) \to 1 - e^{-t}$ for all $t > 0$. Define $\varepsilon := t/n$. Using the convention $X_0 := 0$, we have

$$\begin{aligned}\mathbb{P}(S^n_{I_n} \ge \varepsilon) \quad = \quad & \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{P}\Big(\{X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n \notin [X_k, X_k + \varepsilon]\} \\ & \qquad\qquad \cap \{X_k \le 1 - \varepsilon\}\Big).\end{aligned}$$

It suffices to show that each summand satisfies

$$\lim_{n\to\infty} \mathbb{P}(\{X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n \notin [X_k, X_k + \varepsilon]\} \cap \{X_k \le 1 - \varepsilon\}) = e^{-t}. \tag{19}$$

We derive an upper and a lower bound. Throughout we set $m = 18\|\lfloor \log_2 n \rfloor\|$.

<u>Upper bound for (19):</u> For intervals $[a, b]$, we use the notation $[a, b]^- := [a + 2^{36}/n^{18}, b - 2^{36}/n^{18}]$. This yields with $\varepsilon' = \varepsilon - 2^{36}/n^{18}$

$$\begin{aligned}&\mathbb{P}\Big(\{X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n \notin [X_k, X_k + \varepsilon]\} \cap \{X_k \le 1 - \varepsilon\}\Big) \qquad (20)\\ &\quad \le \quad \mathbb{P}\Big(\{Y_1^{\langle m\rangle}, \dots, Y_{k-1}^{\langle m\rangle}, Y_{k+1}^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle} \notin [Y_k^{\langle m\rangle}, Y_k^{\langle m\rangle} + \varepsilon]^-\} \cap \{Y_k^{\langle m\rangle} \le 1 - \varepsilon'\}\Big) \\ &\quad \le \quad \mathbb{P}\Big(\{Y_1^{\langle m\rangle}, \dots, Y_{k-m}^{\langle m\rangle}, Y_{k+m}^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle} \notin [Y_k^{\langle m\rangle}, Y_k^{\langle m\rangle} + \varepsilon]^-\} \cap \{Y_k^{\langle m\rangle} \le 1 - \varepsilon'\}\Big).\end{aligned}$$

In the last expression, $Y_k^{\langle m\rangle}$ is independent of the other random variables. Now condition on $Y_k^{\langle m\rangle} = x$ for $0 \le x \le 1-\varepsilon'$: We have

$$\begin{aligned}
&\mathbb{P}\Big(Y_1^{\langle m\rangle}, \dots, Y_{k-m}^{\langle m\rangle}, Y_{k+m}^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big) \qquad (21)\\
&\quad = \quad \mathbb{P}\Big(Y_1^{\langle m\rangle} \notin [x, x+\varepsilon]^-\big)\,\mathbb{P}\Big(Y_{m+1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big)\\
&\qquad\qquad \times \cdots \times \mathbb{P}\Big(Y_{(\lceil n/m\rceil-1)m+1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big)\\
&\qquad\quad \times \mathbb{P}\Big(Y_2^{\langle m\rangle} \notin [x, x+\varepsilon]^- \,|\, Y_1^{\langle m\rangle}, Y_{m+1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big)\\
&\qquad\qquad \times \mathbb{P}\Big(Y_{m+2}^{\langle m\rangle} \notin [x, x+\varepsilon]^- \,|\, Y_{m+1}^{\langle m\rangle}, Y_{2m+1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big)\\
&\qquad\qquad \times \mathbb{P}\Big(Y_{2m+2}^{\langle m\rangle} \notin [x, x+\varepsilon]^- \,|\, Y_{2m+1}^{\langle m\rangle}, Y_{3m+1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big)\\
&\qquad\qquad \times \cdots\\
&\qquad\quad \times \cdots\\
&\qquad\quad \times \mathbb{P}\Big(Y_m^{\langle m\rangle} \notin [x, x+\varepsilon]^- \,|\, Y_1^{\langle m\rangle}, \dots, Y_{m-1}^{\langle m\rangle}, Y_{m+1}^{\langle m\rangle}, \dots\\
&\qquad\qquad\qquad\qquad\qquad \dots, Y_{2m-1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big)\\
&\qquad\qquad \times \cdots\\
&\quad \le \quad (1-p_{1,\varepsilon})^{n/m-3} \cdots (1-p_{m,\varepsilon})^{n/m-3},
\end{aligned}$$

where, for $i = 1, \dots, m$,

$$p_{i,\varepsilon} = \mathbb{P}\Big(Y_i^{\langle m\rangle} \in [x, x+\varepsilon]^- \,|\, Y_1^{\langle m\rangle}, \dots, Y_{i-1}^{\langle m\rangle}, Y_{m+1}^{\langle m\rangle}, \dots, Y_{m+i-1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-\Big).$$

We introduce the *bad set*

$$\widehat{B}_n(\xi) := \bigcup_{0\le i<j\le 2m} \{x \in [0,1] : |T^i(x) - T^j(x)| \le \xi\}, \quad \xi > 0. \qquad (22)$$

Note that, by Lemma 3.2, we obtain

$$\lambda(\{x \in [0,1] : |T^i(x) - T^j(x)| \le \xi\}) = \mathbb{P}(|X_{i+1} - X_{j+1}| \le \xi) \le 2\xi, \qquad (23)$$

so that

$$\lambda(\widehat{B}_n(\xi)) \le 8\xi m^2 = O(\xi \log^2 n).$$

With the notion of the bad set and $[a,b]^{--} := ([a,b]^-)^-$ for intervals $[a,b]$ we have

$$\begin{aligned}
p_{i,\varepsilon} \quad &\ge \quad \mathbb{P}(Y_i^{\langle m\rangle} \in [x, x+\varepsilon]^-, Y_1^{\langle m\rangle}, \dots, Y_{i-1}^{\langle m\rangle}, Y_{m+1}^{\langle m\rangle}, \dots, Y_{m+i-1}^{\langle m\rangle} \notin [x, x+\varepsilon]^-)\\
&\ge \quad \mathbb{P}(X_i \in [x, x+\varepsilon]^{--}, X_1, \dots, X_{i-1}, X_{m+1}, \dots, X_{m+i-1} \notin [x, x+\varepsilon])\\
&\ge \quad \mathbb{P}(X_1 \in T^{-i}([x, x+\varepsilon]^{--}) \cap \widehat{B}_n^c(\varepsilon)).
\end{aligned}$$

In order to estimate this term we define for $i \geq 1$ the sets

$$D_i := \{x \in [0,1] : \mathbb{P}(X_1 \in T^{-i}([x, x+\varepsilon]^{--}) \cap \widehat{B}_n^c(\varepsilon)) \geq (1 - 1/\log n)\varepsilon\}$$

and

$$\bar{D}_n := \bigcap_{i=1}^{m} D_i.$$

With $U$ being $U[0,1]$ distributed, Lemma 8.3 and the estimate (23) imply

$$\begin{aligned} \lambda(D_i) &= \mathbb{P}(\lambda(T^{-i}([U, U+\varepsilon]^{--}) \cap \widehat{B}_n^c(\varepsilon)) \geq (1 - 1/\log n)\varepsilon) \\ &\geq \mathbb{P}(\lambda(T^{-i}([U, U+\varepsilon]^{--}) \cap \widehat{B}_n(\varepsilon)) \leq \varepsilon/\log n - 2^{38}/n^{18}) - \varepsilon \\ &= 1 - \mathbb{P}(\lambda(T^{-i}([U, U+\varepsilon]^{--}) \cap \widehat{B}_n(\varepsilon)) > \varepsilon/\log n - 2^{38}/n^{18}) - \varepsilon \\ &\geq 1 - \mathbb{P}(\lambda(T^{-i}([U, U+\varepsilon]) \cap \widehat{B}_n(\varepsilon)) > \varepsilon/\log n - 2^{38}/n^{18}) - \varepsilon \\ &\geq 1 - \frac{\varepsilon\lambda(\widehat{B}_n(\varepsilon))}{\varepsilon/\log n - 32/n^{18}} - \varepsilon \\ &\geq 1 - Ct\frac{\log^3 n}{n}, \end{aligned}$$

for $n$ sufficiently large, where $C > 0$ is an appropriate constant. Hence, we have

$$\lambda(\bar{D}_n) \geq 1 - O\left(t\frac{\log^4 n}{n}\right) \to 1, \quad n \to \infty.$$

For $x \in \bar{D}_n$ we have $p_{i,\varepsilon} \geq (1 - 1/\log n)\varepsilon$ for all $i = 1, \ldots, m$. Thus we obtain for the term (20) splitting into the set $\{Y_k^{\langle m\rangle} \in \bar{D}_n\}$ and its complement

$$\begin{aligned} &\mathbb{P}(Y_1^{\langle m\rangle}, \ldots, Y_{k-m}^{\langle m\rangle}, Y_{k+m}^{\langle m\rangle}, \ldots, Y_n^{\langle m\rangle} \notin [Y_k^{\langle m\rangle}, Y_k^{\langle m\rangle} + \varepsilon]^-) \\ &\leq \lambda(\bar{D}_n)\left(1 - \left(1 - \frac{1}{\log n}\right)\frac{t}{n}\right)^{n-3m} + \lambda(\bar{D}_n^c) \\ &\to e^{-t}, \quad n \to \infty. \end{aligned}$$

Lower bound for (19): We have the basic estimate

$$\begin{aligned} &\mathbb{P}\Big(\{X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n \notin [X_k, X_k + \varepsilon]\} \cap \{X_k \leq 1 - \varepsilon\}\Big) \\ &\geq \mathbb{P}\Big(\{X_1, \ldots, X_{k-m}, X_{k+m}, \ldots, X_n \notin [X_k, X_k + \varepsilon]\} \cap \{X_k \leq 1 - \varepsilon\}\Big) \quad (24) \\ &\quad - \mathbb{P}\left(\{X_k \leq 1 - \varepsilon\} \cap \bigcup_{\{j:1 \leq |j-k| < m\}} \{X_j \in [X_k, X_k + \varepsilon]\}\right). \end{aligned}$$

The second summand in (24) has a negligible asymptotic contribution: We have, using Lemma 3.2,

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{\{j:1\le|j-k|<m\}}\{X_j\in[X_k,X_k+\varepsilon]\}\Big) &\le \mathbb{P}\Big(\bigcup_{\{j:1\le|j-k|<m\}}\{|X_j-X_k|\le\varepsilon\}\Big)\\ &\le 2m2\varepsilon\\ &\le 72t\frac{\log_2 n}{n}.\end{aligned}$$

For the lower bound of the first summand in (24) we use for intervals $[a,b]$ the notation $[a,b]^+ := [a-2^{36}/n^{18}, b+2^{36}/n^{18}]$. We have with $\varepsilon' = \varepsilon + 2^{36}/n^{18}$,

$$\begin{aligned}&\mathbb{P}\Big(\{X_1,\dots,X_{k-m},X_{k+m},\dots,X_n\notin[X_k,X_k+\varepsilon]\}\cap\{X_k\le 1-\varepsilon\}\Big)\\ &\quad\ge\ \mathbb{P}\Big(\{Y_1^{\langle m\rangle},\dots,Y_{k-m}^{\langle m\rangle},Y_{k+m}^{\langle m\rangle},\dots,Y_n^{\langle m\rangle}\notin[Y_k^{\langle m\rangle},Y_k^{\langle m\rangle}+\varepsilon]^+\}\cap\{Y_k^{\langle m\rangle}\le 1-\varepsilon'\}\Big).\end{aligned}$$

A decomposition as in (21) gives

$$\begin{aligned}&\mathbb{P}(Y_1^{\langle m\rangle},\dots,Y_{k-m}^{\langle m\rangle},Y_{k+m}^{\langle m\rangle},\dots,Y_n^{\langle m\rangle}\notin[x,x+\varepsilon]^+)\\ &\quad=\ (1-p'_{1,\varepsilon})^{\lfloor n/m\rfloor-1}\cdots(1-p'_{m,\varepsilon})^{\lfloor n/m\rfloor-1}\times\bar{p},\end{aligned}$$

where analogously to (22) we have

$$p'_{i,\varepsilon}=\mathbb{P}(Y_i^{\langle m\rangle}\in[x,x+\varepsilon]^+\,|\,Y_1^{\langle m\rangle},\dots,Y_{i-1}^{\langle m\rangle},Y_{m+1}^{\langle m\rangle},\dots,Y_{m+i-1}^{\langle m\rangle}\notin[x,x+\varepsilon]^+),$$

and, with $s_1=n-m(\lfloor n/m\rfloor-1)$ and $s_2=n-m\lfloor n/m\rfloor$,

$$\begin{aligned}\bar{p} &= \prod_{i=1}^{m}\mathbb{P}\Big(Y_{s_1+i}^{\langle m\rangle}\notin[x,x+\varepsilon]^+\,|\,Y_{s_1+1}^{\langle m\rangle},\dots,Y_{s_1+i-1}^{\langle m\rangle},Y_{s_2+1}^{\langle m\rangle},\dots,\\ &\qquad\qquad Y_{(s_2+i-1)\wedge n}^{\langle m\rangle}\notin[x,x+\varepsilon]^+\Big)\\ &\quad\times\prod_{i=1}^{n-s_2}\mathbb{P}\Big(Y_{s_2+i}^{\langle m\rangle}\notin[x,x+\varepsilon]^+\,|\,Y_{s_2+1}^{\langle m\rangle},\dots,Y_{s_2+i-1}^{\langle m\rangle}\notin[x,x+\varepsilon]^+\Big).\end{aligned}$$

For $x\le 1-\varepsilon'$ we obtain the estimate

$$\begin{aligned}p'_{i,\varepsilon} &\le \frac{\mathbb{P}(Y_i^{\langle m\rangle}\in[x,x+\varepsilon]^+)}{\mathbb{P}(Y_1^{\langle m\rangle},\dots,Y_{i-1}^{\langle m\rangle},Y_{m+1}^{\langle m\rangle},\dots,Y_{m+i-1}^{\langle m\rangle}\notin[x,x+\varepsilon]^+)}\\ &\le \frac{\varepsilon+2^{37}/n^{18}}{1-2m(\varepsilon+2^{37}/n^{18})},\end{aligned}$$

which is independent of $i$. Analogously we have

$$\bar{p} \;\geq\; \Big(1 - \frac{\varepsilon + 2^{37}/n^{18}}{1 - 2m(\varepsilon + 2^{36}/n^{18})}\Big)^{n-s_1},$$

so that

$$\begin{aligned}
&\mathbb{P}(Y_1^{\langle m\rangle}, \dots, Y_{k-m}^{\langle m\rangle}, Y_{k+m}^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle} \notin [x, x+\varepsilon]^+) \\
&\geq \;\; \left(1 - \frac{\varepsilon + 2^{37}/n^{18}}{1 - 2m(\varepsilon + 2^{36}/n^{18})}\right)^n \\
&= \;\; \left(1 - (1 + o(1))\frac{t}{n}\right)^n \\
&\to \;\; e^{-t}, \quad n \to \infty.
\end{aligned}$$

This implies the remaining lower bound. ■

## 10 Uniform integrability

In this section we show that the convergence in Corollary 9.1 holds for all moments. Analogously to the notation $S_j^n$ we introduce the lengths $S_j^{\langle m\rangle, n}$ for $j = 1, \dots, n$ of the spacing formed by $Y_1^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle}$. In this section we denote $m := 18\|\lfloor \log_2 n\rfloor\|$ and define intervals of indices as follows. For $j = 1, \dots, s := \lfloor n/(18.5\|\lfloor \log_2 n\rfloor\|)\rfloor$ we define

$$\begin{aligned}
\mathcal{G}_j \;&:=\; \{18.5(j-1)L + 1, \dots, (18.5(j-1) + 1/2)L\}, \\
\mathcal{D}_j \;&:=\; \{(18.5(j-1) + 1/2)L + 1, \dots, 18.5jL\}.
\end{aligned}$$

Then, by construction, the random vectors $(Y_k^{\langle m\rangle})_{k\in\mathcal{G}_1}, \dots, (Y_k^{\langle m\rangle})_{k\in\mathcal{G}_s}$ are independent.

**Lemma 10.1** *For all fixed $p > 0$*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\,(nS_{I_n}^n)^p < \infty,$$

*where the random index $I_n$ is unif$\{0, \dots, n\}$ distributed and independent of $X_1$.*

**Proof:** From $|X_i - Y_i^{\langle m\rangle}| \leq 2^{36}/n^{18}$ we obtain for $n$ sufficiently large $|S_i^n - S_i^{\langle m\rangle, n}| \leq 1/n^2$ and therefore

$$\left(nS_{I_n}^n\right)^p \leq \left(nS_{I_n}^{\langle m\rangle, n} + 1/n\right)^p \leq 2^p\left(nS_{I_n}^n\right)^p + (2/n)^p.$$

Hence it is sufficient to prove $\sup_{n\in\mathbb{N}} \mathbb{E}\,(nS_{I_n}^{\langle m\rangle,n})^p < \infty$. We have the basic estimate

$$\begin{aligned}\mathbb{E}\,(nS_{I_n}^{\langle m\rangle,n})^p &= \int_0^\infty \mathbb{P}\left((nS_{I_n}^{\langle m\rangle,n})^p \geq x\right) dx \\ &\leq \int_0^{n^{p/2}} \mathbb{P}\left(S_{I_n}^{\langle m\rangle,n} \geq \frac{x^{1/p}}{n}\right) dx + \int_{n^{p/2}}^{n^p} \mathbb{P}\left(S_{I_n}^{\langle m\rangle,n} \geq \frac{x^{1/p}}{n}\right) dx \\ &=: J_1 + J_2.\end{aligned}$$

With $y := x^{1/p}$ the integrands can be rewritten as

$$\begin{aligned}&\mathbb{P}\left(S_{I_n}^{\langle m\rangle,n} \geq y/n\right) \\ &\quad= \frac{1}{n+1}\sum_{i=0}^{n} \mathbb{P}\Big(\{Y_1^{\langle m\rangle},\dots,Y_{i-1}^{\langle m\rangle},Y_{i+1}^{\langle m\rangle},\dots,Y_n^{\langle m\rangle} \notin [Y_i^{\langle m\rangle},Y_i^{\langle m\rangle}+y/n)\} \\ &\qquad\qquad \cap \{Y_i^{\langle m\rangle} \leq 1-y/n\}\Big),\end{aligned}$$

since a random interval drawn equally likely among all $n+1$ intervals can be drawn by choosing one of the left endpoints $Y_0^{\langle m\rangle} := 0, Y_1^{\langle m\rangle},\dots,Y_n^{\langle m\rangle}$ equally likely.

Estimate of $J_1$: Note that $Y_i^{\langle m\rangle}$ is independent of at least $s-2$ of the families $(Y_k^{\langle m\rangle})_{k\in\mathcal{G}_j}$, $j=1,\dots,s$, say the first $s-2$ families. This implies with Lemma 8.4, $y \leq \sqrt{n}$, noting that $Y_i^{\langle m\rangle}$ is uniformly distributed on $[0,1]$, and for $n$ sufficiently large such that $L(y-4)/(2n) \leq 1/2$,

$$\begin{aligned}&\mathbb{P}\Big(\{Y_1^{\langle m\rangle},\dots,Y_{i-1}^{\langle m\rangle},Y_{i+1}^{\langle m\rangle},\dots,Y_n^{\langle m\rangle} \notin [Y_i^{\langle m\rangle},Y_i^{\langle m\rangle}+y/n]\} \cap \{Y_i^{\langle m\rangle} \leq 1-y/n\}\Big) \\ &\quad\leq \int_0^{1-y/n} \mathbb{P}\Bigg(\bigcap_{l=1}^{s}\bigcap_{k\in\mathcal{G}_l}\Big\{Y_k^{\langle m\rangle} \notin [a,a+y/n]\Big\}\Bigg) da \\ &\quad\leq \int_0^{1-y/n} \Bigg(\mathbb{P}\Bigg(\bigcap_{k\in\mathcal{G}_l}\Big\{Y_k^{\langle m\rangle} \notin [a,a+y/n]\Big\}\Bigg)\Bigg)^{s-2} da \\ &\quad\leq \left(1-\frac{Ly}{2n}+\frac{2L}{n}\right)^{s-2} \\ &\quad\leq \left(1-\frac{L(y-4)}{2n}\right)^{n/(18.5L)-3} \\ &\quad\leq 8\exp\left(-\frac{y-4}{37}\right).\end{aligned}$$

Thus, we obtain

$$J_1 \quad \leq \quad \int_0^{n^{p/2}} 8\exp\left(-\frac{x^{1/p}-4}{37}\right)dx \leq 8\int_0^{\infty} \exp\left(-\frac{x^{1/p}-4}{37}\right)dx < \infty,$$

uniformly in $n \in \mathbb{N}$ for all $p > 0$.

Estimate of $J_2$: For all $i \in \{0, \ldots, n\}$ at least $s-2$ of the random variables $Y^{\langle m\rangle}_{18.5(j-1)L+1}$, $j = 1, \ldots, s$ are independent of $Y_i^{\langle m\rangle}$. For $y \geq \sqrt{n}$ and $n$ sufficiently large we obtain

$$\begin{aligned}
&\mathbb{P}\Big(\{Y_1^{\langle m\rangle}, \ldots, Y_{i-1}^{\langle m\rangle}, Y_{i+1}^{\langle m\rangle}, \ldots, Y_n^{\langle m\rangle} \notin [Y_i^{\langle m\rangle}, Y_i^{\langle m\rangle} + y/n)\} \cap \{Y_i^{\langle m\rangle} \leq 1 - y/n)\}\Big) \\
&\quad \leq \quad \mathbb{P}\left(\bigcap_{j=1}^{s} \left\{Y^{\langle m\rangle}_{18.5(j-1)L+1} \notin \left[Y_i^{\langle m\rangle}, Y_i^{\langle m\rangle} + \frac{y}{n}\right)\right\} \cap \left\{Y_i^{\langle m\rangle} \leq 1 - \frac{y}{n}\right\}\right) \\
&\quad \leq \quad \left(1 - \frac{y}{n}\right)^{s-2} \\
&\quad \leq \quad \left(1 - \frac{n^{1/2}}{n}\right)^{n/(18.5L)-3} \\
&\quad \leq \quad 8\exp\left(-\frac{n^{1/2}}{18.5L}\right) \\
&\quad \leq \quad C\exp(-n^{1/3})
\end{aligned}$$

with an appropriate constant $C > 0$. Hence, we obtain

$$J_2 \quad \leq \quad \int_{n^{p/2}}^{n^p} C\exp(-n^{1/3})dx \leq Cn^p \exp(-n^{1/3}) \to 0, \quad n \to \infty,$$

for any $p > 0$. ■

The limit law of Theorem 9.1 together with the uniform integrability of Lemma 10.1 implies convergence of all moments (Billingsley 1979, Theorem 25.12). Thus we have

$$\lim_{n\to\infty} \mathbb{E}\,(nS^n_{I_n})^\ell = \int_0^\infty x^\ell e^{-x}\,dx = \ell!, \quad \ell = 0, 1, 2, \ldots . \tag{25}$$

In particular we have:

**Corollary 10.2** *We have* $\mathbb{E}\,(nS^n_{I_n})^2 \to 2$ *for* $n \to \infty$.

We turn to the analysis of the rescaled length of a spacing chosen according to into which spacing an indepedendent $U[0,1]$ random variable falls. For this we define the conditional distribution of the index $J_n$ chosen, by

$$\mathbb{P}(J_n = k \,|\, S_0^n, \dots, S_n^n) = S_k^n, \quad k = 0, \dots, n.$$

Then we have the following limit law:

**Lemma 10.3** *We have*

$$nS_{J_n}^n \xrightarrow{\mathcal{L}} G_2, \quad (n \to \infty),$$

*where $G_2$ is Gamma$(2)$-distributed, i.e., has Lebesgue density $xe^{-x}$ on $[0,\infty)$.*

**Proof:** We use the method of moments. With $I_n$ uniformly distributed on $\{0,\dots,n\}$ and independent of $X_1$ we have for all $\ell \geq 0$,

$$\begin{aligned}
\mathbb{E}(nS_{J_n}^n)^\ell &= n^\ell \,\mathbb{E}\left[\,\mathbb{E}[(S_{J_n}^n)^\ell \,|\, S_0^n, \dots, S_n^n]\right] \\
&= n^\ell \,\mathbb{E}\Big[\sum_{k=0}^n \mathbb{P}(J_n = k \,|\, S_0^n, \dots, S_n^n)\, \mathbb{E}[(S_k^n)^\ell \,|\, S_0^n, \dots, S_n^n]\Big] \\
&= n^\ell \,\mathbb{E}\Big[\sum_{k=0}^n (S_k^n)^{\ell+1}\Big] \\
&= n^\ell (n+1)\, \mathbb{E}\,(S_{I_n}^n)^{\ell+1} \\
&= \frac{n+1}{n}\, \mathbb{E}\,(nS_{I_n}^n)^{\ell+1} \\
&\to \int_0^\infty x^{\ell+1} e^{-x}\, dx,
\end{aligned}$$

where we used the representation (25). The last integral is the $\ell$th moment of the Gamma(2)-distribution. Hence we have convergence of all moments of $nS_{J_n}^n$ to the corresponding moments of the Gamma(2)-distribution. Since these moments characterize the Gamma(2)-distribution uniquely we obtain the assertion. ■

## 11 Applications of spacings

In this section we show how the analysis of the random spacings generated by $X_1, \dots, X_n$ can be used for the asymptotic analysis of parameters of the random suffix search tree. First we show how the leading order term of $\mathbb{E}D_n$ can be obtained. This provides an alternative path to that followed in Theorem 6.1. Afterwards we obtain a limit law for the size of the subtree rooted at $X_j$ for a large range of values $j$. This result is rooted in the lemmas of section 10. We provide two lemmas.

**Lemma 11.1** *Let $(a_n)$ be a sequence of real numbers with $a_n \to a \neq 0$ and $(\tau_n)$, $(\xi_n)$ be sequences of integers with $\log(\tau_n) = o(\log(n))$ and $n - \xi_n = \Omega(n)$. Then we have*

$$\sum_{j=\tau_n}^{n-\xi_n} \frac{a_j}{j} \sim a \log(n), \quad (n \to \infty).$$

**Lemma 11.2** *For all $q > 0$ we have*

$$\sum_{i=1}^{\infty} \left(\frac{1}{2^i}\right)^2 \left(1 - \frac{1}{2^i}\right)^q \leq \frac{6}{q^2}.$$

**Proof:** The function $f : [0,1] \to \mathbb{R}_0^+$, $y \mapsto y^2(1-y)^q$ is unimodal with the maximum at $2/(2+q)$, thus $\|f\|_\infty \leq 4/q^2$. Therefore we may estimate the series by the corresponding integral where we have to estimate the summand being maximal separately. This implies

$$\begin{aligned} \sum_{i=1}^{\infty} \left(\frac{1}{2^i}\right)^2 \left(1 - \frac{1}{2^i}\right)^q &\leq \frac{4}{q^2} + \int_0^\infty \left(\frac{1}{2^y}\right)^2 \left(1 - \frac{1}{2^y}\right)^q dy \\ &= \frac{4}{q^2} + \frac{1}{\log 2} \int_0^1 x(1-x)^q \, dx \\ &= \frac{4}{q^2} + \frac{1}{\log 2} \frac{1}{(q+1)(q+2)}, \end{aligned}$$

which implies the assertion. ■

**Proof of $\mathbb{E}\mathbf{D_n} \sim \mathbf{2} \log \mathbf{n}$:** We introduce the events $A_j = \{X_j$ is ancestor of $X_n$ in the tree$\}$ and recall the representations

$$D_n = \sum_{j=1}^{n-1} \mathbf{1}_{A_j}, \quad \mathbb{E}D_n = \sum_{j=1}^{n-1} \mathbb{P}(A_j).$$

For the estimate of $\mathbb{P}(A_j)$ we distinguish three ranges for the index $j$, namely $1 \leq j \leq \lceil \log_2^6 n \rceil$, $\lceil \log_2^6 n \rceil < j \leq n-m$, and $n-m < j < n$, where we choose $m = 18\|\lfloor \log_2 n \rfloor\|$. <u>Ranges $1 \leq j \leq \lceil \log_2^6 n \rceil$ and $n - m < j < n$:</u> We refer to the corresponding range $1 \leq j \leq \lceil \log_2^{12} n \rceil$ and $n - m < j < n$ in the proof of Theorem 6.1. We only treat the critical middle range:

The range $\lceil \log_2^6 n \rceil < j \leq n - m$: We use the notation $\alpha, \beta \triangleright \gamma_1, \dots, \gamma_n$, if there does not exist $k$ with $1 \leq k \leq n$ for which $\alpha < \gamma_k < \beta$ or $\beta < \gamma_k < \alpha$, i.e., $\alpha, \beta$ are neighbors in $\{\gamma_1, \dots, \gamma_n\}$. In this notation we have, by construction of a binary search tree, $A_j = \{X_j, X_n \triangleright X_1, \dots, X_{j-1}\}$. Denoting $t := j - m$, we start, using Lemma 3.1, with the representation

$$\begin{aligned}
\mathbb{P}(A_j) &= \mathbb{P}(X_j, X_n \triangleright X_1, \dots, X_{j-1}) && (26)\\
&= \mathbb{P}(Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, \dots, Y_{j-1}^{\langle m \rangle}) + O(1/n^2)\\
&= \mathbb{P}(Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, \dots, Y_t^{\langle m \rangle}) + O(1/n^2)\\
&\quad - \mathbb{P}\Big(\{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, \dots, Y_t^{\langle m \rangle}\} && (27)\\
&\qquad\qquad \cap \{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, \dots, Y_{j-1}^{\langle m \rangle}\}^c\Big).
\end{aligned}$$

With $q := \lfloor j/m \rfloor - 1$ we estimate

$$\begin{aligned}
&\mathbb{P}(\{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, \dots, Y_t^{\langle m \rangle}\} \cap \{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_1^{\langle m \rangle}, \dots, Y_{j-1}^{\langle m \rangle}\}^c) && (28)\\
&\leq \sum_{i=t}^{j-1} \mathbb{P}(\{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_{j-m}^{\langle m \rangle}, Y_{j-2m}^{\langle m \rangle}, \dots, Y_{j-qm}^{\langle m \rangle}\} \cap \{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_i^{\langle m \rangle}\}^c).
\end{aligned}$$

In order to estimate the latter summands we introduce $\varepsilon_k := 1/2^k$ and for $a \in [0,1]$ the intervals $a[\varepsilon_k^-] := [a - \varepsilon_k, a]$ and $a[\varepsilon_k^+] := [a, a + \varepsilon_k]$. Then we have

$$\begin{aligned}
&\mathbb{P}(\{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_{j-m}^{\langle m \rangle}, Y_{j-2m}^{\langle m \rangle}, \dots, Y_{j-qm}^{\langle m \rangle}\} \cap \{Y_j^{\langle m \rangle}, Y_n^{\langle m \rangle} \triangleright Y_i^{\langle m \rangle}\}^c) && (29)\\
&\leq \mathbb{P}\Bigg(\bigcup_{k \geq 1} \Big(\{\varepsilon_k \leq Y_n^{\langle m \rangle} \leq 1\} \cap \{Y_i^{\langle m \rangle}, Y_j^{\langle m \rangle} \in Y_n^{\langle m \rangle}[\varepsilon_{k-1}^-]\}\\
&\qquad\qquad \cap \{Y_{j-m}^{\langle m \rangle}, Y_{j-2m}^{\langle m \rangle}, \dots, Y_{j-qm}^{\langle m \rangle} \notin Y_n^{\langle m \rangle}[\varepsilon_k^-]\}\Big)\Bigg)\\
&\quad + \mathbb{P}\Bigg(\bigcup_{k \geq 1} \Big(\{0 \leq Y_n^{\langle m \rangle} \leq 1 - \varepsilon_k\} \cap \{Y_i^{\langle m \rangle}, Y_j^{\langle m \rangle} \in Y_n^{\langle m \rangle}[\varepsilon_{k-1}^+]\}\\
&\qquad\qquad \cap \{Y_{j-m}^{\langle m \rangle}, Y_{j-2m}^{\langle m \rangle}, \dots, Y_{j-qm}^{\langle m \rangle} \notin Y_n^{\langle m \rangle}[\varepsilon_k^+]\}\Big)\Bigg).
\end{aligned}$$

The last two summands are the same. We will consider the first summand. Note that all random variates appearing there are $U[0,1]$ distributed and that we have dependency only between $Y_i^{\langle m \rangle}$ and $Y_j^{\langle m \rangle}$. Therefore with Lemma 8.2 and Lemma

11.2, the first summand in the latter display is bounded from above by

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbb{P}\Big(Y_i^{\langle m\rangle}, Y_j^{\langle m\rangle} \in Y_n^{\langle m\rangle}[\varepsilon_{k-1}^{-}]\Big)(1-\varepsilon_k)^q &\le \sum_{k=1}^{\infty} 16\varepsilon_k^2(1-\varepsilon_k)^q \\ &\le 96\left(\frac{1}{\lfloor j/m\rfloor - 1}\right)^2 \\ &= O\left(\frac{\log^2(n)}{j^2}\right).\end{aligned}$$

Note that the big $O$ term is independent of $i$. Plugging this into (28) we obtain

$$\begin{aligned}&\mathbb{P}\Big(\{Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle}, \dots, Y_t^{\langle m\rangle}\} \cap \{Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle}, \dots, Y_{j-1}^{\langle m\rangle}\}^c\Big) \\ &\quad= O\left(\frac{\log^3 n}{j^2}\right) = O\left(j^{-3/2}\right),\end{aligned}$$

since $j = \Omega(\log^6(n))$ in the range under consideration. Substituting this into (26) we obtain

$$\mathbb{P}(A_j) = \mathbb{P}(Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle}, \dots, Y_t^{\langle m\rangle}) + O(j^{-3/2}). \tag{30}$$

Now, we note that $Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle}$ are independent and $U[0,1]$ distributed, and independent of $Y_1^{\langle m\rangle}, \dots, Y_t^{\langle m\rangle}$. Therefore, with $I_t$ uniformly distributed on $\{0, \dots, t\}$ and independent of all other quantities, we obtain, by Corollary 10.2, and using $|X_i - Y_i^{\langle m\rangle}| \le 2^{36}/n^{18}$,

$$\begin{aligned}\mathbb{P}(Y_j^{\langle m\rangle}, Y_n^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle}, \dots, Y_t^{\langle m\rangle}) &= \mathbb{E}\sum_{i=0}^{t}(S_i^{\langle m\rangle,t})^2 \\ &= \mathbb{E}\sum_{i=0}^{t}((S_i^t)^2) + O((\log n)/n^{18}) \\ &= (t+1)\,\mathbb{E}\,(S_{I_t}^t)^2 + O(1/n^2) \\ &= \frac{t+1}{t^2}\,\mathbb{E}\,(tS_{I_t}^t)^2 + O(1/n^2) \\ &\sim \frac{2}{j-m}, \quad (j \to \infty).\end{aligned} \tag{31}$$

Putting (30) and (31) together we finally obtain for the second range

$$\sum_{j=\lceil \log_2^6 n\rceil + 1}^{n-m} \mathbb{P}(A_j) = \sum_{j=\lceil \log_2^6 n\rceil + 1 - m}^{n-2m} \left(\frac{a_j}{j}\right) + O(1),$$

with $a_j \to 2$ for $j \to \infty$ which, by Lemma 11.1, implies

$$\sum_{j=\lceil \log_2^6 n\rceil+1}^{n-m} \mathbb{P}(A_j) = 2\log n + o(\log n). \quad \blacksquare$$

We turn to the analysis of the size $N_{n,j}$ of the subtree rooted at $X_j$.

**Theorem 11.3** *The size $N_{n,j}$ of the subtree of the random suffix search tree of size $n$ rooted at $X_j$ satisfies for $j = j(n)$ with $j = o(n/\log^2 n)$ and $j/\log^5 n \to \infty$,*

$$\mathbb{E} N_{n,j} \sim \frac{2n}{j}, \quad \frac{j}{n} N_{n,j} \xrightarrow{\mathcal{L}} G_2,$$

*as $n \to \infty$, where $G_2$ denotes the Gamma$(2)$-distribution.*

**Proof:** Recall the notation $S_n^*(j)$ for the length of the unique spacing among the spacings formed by $X_1, \dots, X_j$ on $[0,1]$ which covers $X_n$. We denote by $\widehat{S}_n^*(j)$ the corresponding length for the quantities $Y_1^{\langle m\rangle}, \dots, Y_n^{\langle m\rangle}$ and by $S_n(j)$ and $\widehat{S}_n(j)$ these chosen spacings respectively. We show first that for the $j = j(n)$ under consideration we have $\mathbb{E}\widehat{S}_j^*(j-1) \sim 2/j$ and $j\widehat{S}_j^*(j-1) \to G_2$ in distribution. From this we will then obtain the assertions.

Claim: $\mathbb{E}\widehat{S}_j^*(j-1) \sim 2/j$. With the notation $M_j^{\langle m\rangle}$ for the maximal spacing formed by $Y_1^{\langle m\rangle}, \dots, Y_{j-m}^{\langle m\rangle}$ as introduced in the proof of Theorem 7.1 we define the sets

$$\begin{aligned} V &:= \bigcup_{k=1}^{m-1} \{Y_j^{\langle m\rangle}, Y_{j-k}^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle}, \dots Y_{j-m}^{\langle m\rangle}\}, \\ W &:= \left\{ M_j^{\langle m\rangle} \le \frac{m^2}{j} \right\}. \end{aligned}$$

Then we have

$$\begin{aligned} \widehat{S}_j^*(j-1) &= \widehat{S}_j^*(j-m) - \mathbf{1}_V \Big( \widehat{S}_j^*(j-m) - \widehat{S}_j^*(j-1) \Big) \qquad (32) \\ &= \widehat{S}_j^*(j-m) - \big( \mathbf{1}_W + \mathbf{1}_{W^c} \big) \mathbf{1}_V \Big( \widehat{S}_j^*(j-m) - \widehat{S}_j^*(j-1) \Big). \end{aligned}$$

Using the the estimate (13) we obtain $\mathbb{P}(W^c) = O(1/n^2)$, thus together with $\widehat{S}_j^*(j-m) - \widehat{S}_j^*(j-1) \le 1$ we obtain

$$\mathbb{E}\Big[ \mathbf{1}_{W^c} \mathbf{1}_V \Big( \widehat{S}_j^*(j-m) - \widehat{S}_j^*(j-1) \Big) \Big] = O\left( \frac{1}{n^2} \right).$$

On the set $W$ we have $\widehat{S}^*_j(j-m) - \widehat{S}^*_j(j-1) \leq m^2/j$ thus we obtain

$$\begin{aligned} \mathbb{E}\Big[\mathbf{1}_W \mathbf{1}_V \Big(\widehat{S}^*_j(j-m) - \widehat{S}^*_j(j-1)\Big)\Big] &\leq \frac{m^2}{j}\mathbb{P}(V) \qquad (33) \\ &\leq \frac{m^2}{j}\sum_{k=1}^{m} \mathbb{P}\Big(Y_j^{\langle m\rangle}, Y_{j-k}^{\langle m\rangle} \triangleright Y_1^{\langle m\rangle}, \dots Y_{j-m}^{\langle m\rangle}\Big) \\ &= O\left(\frac{\log^5 n}{j^2}\right), \end{aligned}$$

where we estimate the last summands as shown in (9).

For the estimate of $\mathbb{E}\,\widehat{S}^*_j(j-m)$ note that this is now the length of the spacing among the $S_0^{\langle m\rangle, j-m}, \dots, S_{m-j}^{\langle m\rangle, j-m}$, which are generated by $Y_1^{\langle m\rangle}, \dots Y_{j-m}^{\langle m\rangle}$ on $[0,1]$, where $Y_j^{\langle m\rangle}$ falls into. Moreover $Y_j^{\langle m\rangle}$ is independent of the generating points. Thus applying Corollary 10.2 we obtain similarly to the estimate (31),

$$\begin{aligned} \mathbb{E}\,\widehat{S}^*_j(j-m) &= \mathbb{E}\sum_{k=0}^{j-m}\left(S_i^{\langle m\rangle, j-m}\right)^2 \\ &= \mathbb{E}\sum_{k=0}^{j-m}\left(S_i^{j-m}\right)^2 + O\left(\frac{1}{n^2}\right) \\ &= (j-m+1)\,\mathbb{E}\left(S_{I_{m-j}}^{j-m}\right)^2 + O\left(\frac{1}{n^2}\right) \\ &= \frac{2}{j-m}(1+o(1)), \end{aligned}$$

as $j-m \to \infty$, where $I_{m-j}$ is uniformly distributed on $\{0, \dots, m-j\}$ and independent of $X_1$. Collecting all the estimates we obtain

$$\mathbb{E}\,\widehat{S}^*_j(j-1) = \frac{2}{j}(1+o(1)) + O\left(\frac{\log^5 n}{j^2}\right) \sim \frac{2}{j},$$

as $n \to \infty$, when $\log^5 n = o(j)$.

Claim: $j\widehat{S}^*_j(j-1) \xrightarrow{\mathcal{L}} G_2$. Note that $\widehat{S}^*_j(j-m)$ is in distribution equal to the quantity $S_{J_{j-m}}^{j-m}$ appearing in Lemma 10.3, thus $(j-m)\widehat{S}^*_j(j-m) \to G_2$ in distribution. Now note that for $\widehat{S}^*_j(j-1)$ we have the representaion (32) and that $\mathbb{P}(V) = O(\log^3 /j) = o(1)$ as shown in (33) for our $j$ under consideration. Thus the second summand in (32) tends to zero in probability. Since $j/(j-m) \to 1$, the first summand there tends to $G_2$ in distribution.

Claim: $\mathbb{E} N_{n,j} \sim 2n/j$. Applying Lemma 3.1, we obtain

$$\begin{aligned}
\mathbb{E} N_{n,j} &= \mathbb{E} \sum_{k=j}^{n} \mathbf{1}_{\{X_k \in S_j(j-1)\}} \\
&= \mathbb{E} \sum_{k=j}^{n} \mathbf{1}_{\{Y_k^{\langle m \rangle} \in \widehat{S}_j(j-1)\}} + O\left(\frac{1}{n^2}\right) \\
&= \mathbb{E} \sum_{k=j+m}^{n} \mathbf{1}_{\{Y_k^{\langle m \rangle} \in \widehat{S}_j(j-1)\}} + O(\log n) \\
&= (n-j-m+1)\mathbb{P}(\{Y_n^{\langle m \rangle} \in \widehat{S}_j(j-1)\}) + O(\log n) \\
&= (n-j-m+1)\, \mathbb{E} \widehat{S}_j^*(j-1) + O(\log n) \\
&\sim \frac{2n}{j}
\end{aligned}$$

as $n \to \infty$, $j = o(n/\log n)$ and $\log^5 n = o(j)$, where we used that for $k \geq j+m$ we have independence between $Y_k^{\langle m \rangle}$ and $\widehat{S}_j(j-1)$ and $\mathbb{E}\widehat{S}_j^*(j-1) \sim 2n/j$.

Claim: $(j/n)N_{n,j} \to G_2$. We denote the number of nodes of the subtree rooted at $Y_j^{\langle m \rangle}$ in the tree built from $Y_1^{\langle m \rangle}, \dots, Y_n^{\langle m \rangle}$ by $N_{n,j}^{\langle m \rangle}$. Then we have

$$N_{n,j} = N_{n,j}^{\langle m \rangle} + \mathbf{1}_A \big(N_{n,j} - N_{n,j}^{\langle m \rangle}\big),$$

where $A$ denotes the event that $X_1, \dots, X_n$ and $Y_1^{\langle m \rangle}, \dots Y_n^{\langle m \rangle}$ do not give the same permutation. By Lemma 3.1 we have $\mathbb{P}(A) \to 0$ as $n \to \infty$ thus the second summand in the last display tends to zero in probability. Hence it is sufficient to show $N_{n,j}^{\langle m \rangle} \to G_2$ in distribution for the choices of $j$ under consideration.

The number $N_{n,j}^{\langle m \rangle}$ is given as the sum $P + \sum_{k=1}^{m} P_k$, where $P$ denotes the number of points among $Y_{j+1}^{\langle m \rangle}, \dots, Y_{j+m}^{\langle m \rangle}$ which contribute to the subtree rooted at $Y_j^{\langle m \rangle}$ and $P_k$ denotes the corresponding number for the points $Y_{j+k+m}^{\langle m \rangle}, Y_{j+k+2m}^{\langle m \rangle}, Y_{j+k+q_k m}^{\langle m \rangle}$, where $\lceil q_k = (n-j-k-m+1)/m \rceil$. Thus we have $\sum_{k=1}^{m} q_k = n-j-m$. Note that given $\widehat{S}_j^*(j-1) = T$, by indepedence, $P_k$ is binomial $B(q_k, T)$ distributed for $k = 1, \dots, m$. Thus by Chebyshev's inequality, noting that $0 \leq P \leq m$, and denoting $T_j = \widehat{S}_j^*(j-1)$ we obtain for all $\delta > 0$, almost surely

$$\begin{aligned}
\mathbb{P}\Big(\Big|N_{n,j}^{\langle m\rangle} - (n-j-m)T_j\Big| \geq m\delta + m \Big| T_j\Big) &\leq \sum_{k=1}^{m} \mathbb{P}\Big(\Big|P_k - q_k T_j\Big| \geq \delta \Big| T_j\Big) \\
&\leq \sum_{k=1}^{m} \frac{q_k T_j(1-T_j)}{\delta^2} \\
&\leq \frac{n}{\delta^2} T_j.
\end{aligned}$$

Thus for arbitrary $\varepsilon > 0$, we have, choosing $\delta = \varepsilon n/((m+1)j)$, almost surely

$$\mathbb{P}\Big(\Big|\frac{j}{n} N_{n,j}^{\langle m\rangle} - \frac{(n-j-m)j}{n} T_j\Big| \geq \varepsilon \Big| T_j\Big) \leq \frac{(m+1)^2 j^2}{\varepsilon^2 n} T_j.$$

Now, for all $x \geq 0$, denoting by $F_{G_2}$ the distribution function of the Gamma(2) distribution and using the last estimate, $j\widehat{S}_j^*(j-1) \to G_2$ in distribution, and $\mathbb{E}\widehat{S}_j^*(j-1) \sim 2/j$, we obtain, as $n \to \infty$, $j = j(n) = o(n/\log^2 n)$ and $j$ tending to infinity,

$$\begin{aligned}
\mathbb{P}\Big(\frac{j}{n} N_{n,j}^{\langle m\rangle} \leq x\Big) &= \mathbb{P}\Big(\frac{j}{n} N_{n,j}^{\langle m\rangle} \leq x, \frac{(n-j-m)j}{n} T_j > x + \varepsilon\Big) \\
&\quad + \mathbb{P}\Big(\frac{j}{n} N_{n,j}^{\langle m\rangle} \leq x, \frac{(n-j-m)j}{n} T_j \leq x + \varepsilon\Big) \\
&\leq \mathbb{P}\Big(\Big|\frac{j}{n} N_{n,j}^{\langle m\rangle} - \frac{(n-j-m)j}{n} T_j\Big| \geq \varepsilon\Big) \\
&\quad + \mathbb{P}\Big(\frac{(n-j-m)j}{n} \widehat{S}_j^*(j-1) \leq x + \varepsilon\Big) \\
&\leq \frac{(m+1)^2 j^2}{\varepsilon^2 n} \mathbb{E}\Big[\widehat{S}_j^*(j-1)\Big] + F_{G_2}(x+\varepsilon) + o(1) \\
&\sim \frac{2(m+1)^2 j}{\varepsilon^2 n} + F_{G_2}(x+\varepsilon) \\
&\to F_{G_2}(x+\varepsilon).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathbb{P}\Big(\frac{j}{n}N_{n,j}^{\langle m\rangle} \le x\Big) &= 1-\mathbb{P}\Big(\frac{j}{n}N_{n,j}^{\langle m\rangle} > x\Big)\\
&= 1-\mathbb{P}\Big(\frac{j}{n}N_{n,j}^{\langle m\rangle} > x, \frac{(n-j-m)j}{n}T_j < x-\varepsilon\Big)\\
&\quad -\mathbb{P}\Big(\frac{j}{n}N_{n,j}^{\langle m\rangle} > x, \frac{(n-j-m)j}{n}T_j \ge x-\varepsilon\Big)\\
&\ge 1-\mathbb{P}\Big(\Big|\frac{j}{n}N_{n,j}^{\langle m\rangle} - \frac{(n-j-m)j}{n}T_j\Big| \ge \varepsilon\Big)\\
&\quad -\mathbb{P}\Big(\frac{(n-j-m)j}{n}\widehat{S}_j^*(j-1) \ge x-\varepsilon\Big)\\
&\ge 1-\frac{(m+1)^2j^2}{\varepsilon^2 n}\,\mathbb{E}\left[\widehat{S}_j^*(j-1)\right] - (1-F_{G_2}(x-\varepsilon)) + o(1)\\
&\to F_{G_2}(x-\varepsilon).
\end{aligned}$$

Since $F_{G_2}$ is continuous and $\varepsilon > 0$ arbitrary we obtain $(j/n)N_{n,j}^{\langle m\rangle} \to G_2$ in distribution and thus $(j/n)N_{n,j} \to G_2$ in distribution. ■

Using similiar arguments it can be shown that in the case $j \sim \alpha n$ with $\alpha \in (0,1)$ the size $N_{n,j}$ tends in distribution to the negative binomial distibution with parameters $(2,\alpha)$, given by its generating function $s \mapsto (\alpha/(1-(1-\alpha)s))^2$.

# References

[1] Antos, A. and Devroye, L. (2000) Rawa Trees. *Mathematics and Computer Science (Versailles, 2000)*, 3–15, Birkhäuser, Basel.

[2] Apostolico, A. (1985) The myriad virtues of suffix trees. *Combinatorial Algorithms on Words*, 85–96, Springer-Verlag.

[3] Billingsley, P. (1979) *Probability and Measure.* John Wiley, New York-Chichester-Brisbane.

[4] Chung, K. L. and Erdös, P. (1952) On the application of the Borel-Cantelli lemma. *Trans. Amer. Math. Soc.* **72**, 179–186.

[5] Crochemore, M. and Rytter, W. (1994) *Text Algorithms.* Oxford University Press, New York,

[6] Devroye, L. (1986) A note on the height of binary search trees. *Journal of the ACM* **33**, 489–498.

[7] Devroye, L. (1987) Branching processes in the analysis of the heights of trees. *Acta Inform.* **24**, 277–298.

[8] Devroye, L. (1994) On random Cartesian trees. *Random Structures Algorithms* **5**, 305–327.

[9] Devroye, L. and Goudjil, A. (1998) A study of random Weyl trees. *Random Structures Algorithms* **12**, 271–295.

[10] Devroye, L., Szpankowski, W. and Rais, B. (1992) A note on the height of suffix trees. *SIAM Journal on Computing* **21**, 48–53.

[11] Farach, M. (1997) Optimal suffix tree construction with large alphabets. *IEEE Symp. Found. Computer Science*, 137—143.

[12] Farach, M. and Muthukrishnan, S. (1996) Optimal logarithmic time randomized suffix tree construction. *Proc. 23rd ICALP*, 550–561.

[13] Farach, M. and Muthukrishnan, S. (1997) An optimal, logarithmic time, randomized parallel suffix tree contruction algorithm. *Algorithmica* **19**, 331–353.

[14] Giancarlo, R. (1993) The suffix tree of a square matrix, with applications. *Proc. of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, 402–411.

[15] Giancarlo, R. (1995) A generalization of the suffix tree to square matrices, with applications. *SIAM Journal on Computing*, 520–562.

[16] Giegerich, R. and Kurtz, S. (1995) A comparison of imperative and purely functional suffix tree constructions. *Science of Computer Programming* **25**, 187–218.

[17] Giegerich, R. and Kurtz, S. (1997) From Ukkonen to McCreight and Weiner: a unifying view of linear-time suffix tree construction. *Algorithmica* **19**, 331–353.

[18] Gusfield, D. (1997) *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology.* Cambridge University Press, Cambridge.

[19] Jacquet, P., Rais, B. and Szpankowski, B. (1995) Compact suffix trees resemble PATRICIA tries: limiting distribution of depth. Technical Report RR-1995, Department of Computer Science, Purdue University.

[20] Karkkainen, J. (1995) Suffix cactus : a cross between suffix tree and suffix array. *Combinatorial Pattern Matching, Proc. 6th Symposium on Combinatorial Pattern Matching, CPM 95* **937**, 191–204.

[21] Knuth, D. E. (1973) *The Art of Computer Programming, Vol. 1: Fundamental Algorithms.* Addison-Wesley, Reading, Mass., 2nd Ed.

[22] Knuth, D. E. (1973) *The Art of Computer Programming, Vol. 3: Sorting and Searching.* Addison-Wesley, Reading, Mass.

[23] Kosaraju, S. (1994) Real-time pattern matching and quasi-real-time construction of suffix trees. *Proc. of the 26th Ann. ACM Symp. on Theory of Computing*, 310–316, ACM.

[24] Kurlberg, P. and Rudnick, Z. (1999) The distribution of spacings between quadratic residues. *Duke Jour. of Math.* **100**, 211–242.

[25] Mahmoud, H. M. (1992) *Evolution of Random Search Trees.* John Wiley, New York.

[26] Manber, U. and Myers, G. (1990) Suffix arrays: a new method for on-line string searches. *Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms*, 319–327. SIAM, Philadelphia.

[27] McCreight, E. M. (1976) A space-economical suffix tree construction algorithm. *Journal of the ACM* **23**, 262–272.

[28] Rudnick, Z. and Zaharescu, A. (2002) The distribution of spacings between fractional parts of lacunary sequences. *Forum Math.*, to appear.

[29] Sahinalp, S. C. and Vishkin, U. (1994) Symmetry breaking for suffix tree construction. *Proc. 26th Symp. on Theory of Computing*, 300–309.

[30] Stephen, G. A. (1994) *String Searching Algorithms.* World Scientific, Singapore.

[31] Szpankowski, W. (1993) A Generalized Suffix Tree and its (Un)Expected Asymptotic Behaviors. *SIAM Journal on Computing* **22**, 1176–1198.

[32] Szpankowski, W. (2001) *Average-Case Analysis of Algorithms on Sequences.* John Wiley, New York.

[33] Ukkonen, E. (1995) On-line construction of suffix trees. *Algorithmica* **14**, 249–260.

[34] Weiner, P. (1973) Linear pattern matching algorithms. *Proceedings 14th Annual Symposium on Switching and Automata Theory*, 1–11. IEEE Press, New York.