Algorithmica (2006) 46: 469–491
DOI: 10.1007/s00453-006-0126-4

Algorithmica

Average-Case Analysis of Approximate Trie Search[1]

Moritz G. Maaß[2]

Abstract. For the exact search of a pattern of length m in a database of n strings the trie data structure allows an optimal lookup time of $O(m)$. If mismatches are allowed between the pattern and the database strings, no such structure with reasonable size is known. Some work can be saved using a trie and running times superior to the comparison with every string in the database can be achieved. We investigate a comparison-based model where matches and mismatches are defined between pairs of characters. When comparing two characters, let q be the probability of an error. Between any two strings we bound the number of errors by d, which we consider a function of n. We study the average-case complexity of the number of comparisons for searching in a trie in dependence of the parameters q and d. Our analysis yields the asymptotic behavior for memoryless sources with uniform probabilities. It turns out that there is a jump in the average-case complexity at certain thresholds for q and d. Our results can be applied for any comparison-based error model, for instance, Hamming distance, don't cares, or geometric character distances.

Key Words. Pattern matching, Trie, Rice's integral, Hamming distance, Don't cares.

1. Introduction. We study the average-case behavior of the simple problem of finding a given pattern in a set of patterns subject to two conditions. The set of patterns is given in advance and may be preprocessed with linear space. We also seek occurrences where the pattern is found with a given number of mismatches.

This research was triggered by a project for the control of animal breeding via SNPs [33]. A single nucleotide polymorphism (SNP) is a location on the DNA sequence of a species that varies among the population but is stable through inheritance. Without going into detail, a sequence of SNPs can be encoded as a string that is an identifier for an individual. For instance, in the above project the SNPs had one of the types "heterozygous," "homozygous 1," "homozygous 2," "assay failure" encoded in an alphabet of size four. Because of errors, the search in the dataset of all individuals needs to be able to deal with mismatches and don't cares ("assay failure"). The nature of the data allowed a very efficient binary encoding with two bits per SNP. This yielded a reasonably fast algorithm that just compares a pattern with each string in the set. On the other hand, a trie was used as an index for the data. A trie has the same worst-case lookup time, but we expect to save some work because fewer comparisons are necessary. As a drawback, the constants in the algorithm are a bit higher due to the tree structure involved. Thus, we are interested in which of the above methods is faster.

[1] A preliminary version appeared at CPM '04 [18]. This research was supported in part by DFG, Grant Ma 870/5-1 (Leibnizpreis Ernst W. Mayr).

[2] Fakultät für Informatik, TU München, Boltzmannstr. 3, D-85748 Garching, Germany. maass@informatik.tu-muenchen.de.

Received December 20, 2004; revised February 5, 2006. Communicated by P. Jacquet, D. Panario, and W. Szpankowski. Online publication October 28, 2006.

In the following we call the first variant "Linear Search" (LS) and the second variant "Trie Search" (TS).

An abstract formulation of the problem is the following. Given n strings $X_1, \ldots, X_n$ of the same length m, a pattern P of length m, a mismatch probability q, and a bound d for the maximal number of mismatches allowed, let L_n^d be the number of comparisons made by the LS algorithm and let T_n^d be the number of comparisons made by the TS algorithm using a trie as an index. The parameter q depends on the definition of matches and mismatches, it gives the relative number of character pairs inducing a mismatch. What is the threshold $d = d(n)$, up to where the average $\mathbf{E}[T_n^d]$ is asymptotically better than the average $\mathbf{E}[L_n^d]$? What is the effect of different mismatch probabilities? The answers to these questions give hints at choosing the faster method for our original problem.

Let Σ be an arbitrary finite alphabet of size $\sigma := |\Sigma|$. Let $t = t_1 t_2 t_3 \cdots t_n$ be a string with characters $t_i \in \Sigma$, we define $|t| = n$ to be its length. For the average-case analysis we assume that we deal with strings of infinite length, each string $X = \{x_k\}_{k=1}^{\infty}$ is generated independently at random by a memoryless source with uniform probabilities $\Pr\{x_j = s_i\} = 1/\sigma$ for all $s_i \in \Sigma$. We assume that all strings $X_1, \ldots, X_n$ used in the search are different (i.e., $X_i \neq X_j$ for $i \neq j$). Since the strings are generated randomly the probability that two identical strings of infinite length occur is indeed 0. We further assume that a search pattern $P = \{p_k\}_{k=1}^{\infty}$ is generated by a source with the same probabilities and similarly of infinite length. Indeed, the last assumption is only necessary for certain error models. What is needed for the analysis is the fact that at all stages the probability that a randomly chosen character matches the character at the current position in the pattern is the same (which is true, e.g., for the Hamming distance). When two arbitrary, randomly generated characters are compared, let q denote the probability of a mismatch and let $p := 1 - q$ be the probability of a match. For example, we have a mismatch probability of $q = 1 - 1/\sigma$ and a match probability of $p = 1/\sigma$ for the Hamming distance.

It is easy to prove that the average number of comparisons made by the LS algorithm is $\mathbf{E}[L_n^d] = (d + 1)n/q$. Indeed, one can prove almost sure convergence to this value. The LS algorithm has a linear or quasi-linear average-case behavior for small d (see Section 4). The more interesting part is the analysis of the TS algorithm. In the trie, each edge represents the same character in a number of strings (the number of leaves in the subtree). Let there be k leaves below an edge. If the TS algorithm compares a character from the pattern to the character at the edge, the LS algorithm needs to make k comparisons. In essence, this can be seen as the trie "compressing" the set of strings. It can be proven that the average number of characters thus "compressed" is asymptotically $n \log_\sigma n + O(n)$. Hence, for $d \geq (1 + \varepsilon) \log_\sigma n$ there can be no asymptotic gain when using a trie.

We will show that the TS algorithm performs sublinear for $d < q \log_\sigma n$ and superlinear for $d > q \log_\sigma n$. When d is a constant the asymptotic running time can be calculated exactly and is $O((\log n)^{d+1})$, for $p = 1/\sigma$, $O((\log n)^d n^{\log_\sigma p + 1})$, for $p > 1/\sigma$, and $O(1)$, otherwise.

2. Related Work. Digital tries have a wide range of applications and belong to the most studied data structures in computer science. They have been around for years; for their usefulness and beauty of analysis they have received much attention. Tries were

introduced by Briandais [4] and Fredkin [12]. A useful extension are PATRICIA trees introduced by Morrison [19].

Using tree (especially trie or suffix tree) traversals for indexing problems is a common technique. For instance, in computer linguistics one often needs to correct misspelled input. Schulz and Mihov [26] pursue the idea of correcting misspelled words by finding correctly spelled candidates from a dictionary implemented as a trie or automaton. They build an automaton for the input word and traverse the trie with it in a depth-first search. The search automaton is linear in the size of the pattern if only a constant number of errors is allowed. A similar approach has been investigated by Oflazer [23], except that he directly calculates the edit distance instead of using an automaton.

Flajolet and Puech [10] analyze the average-case behavior of partial matching in k-d-tries. A pattern in k domains with s specified values and $k - s$ don't cares is searched. Each entry in a k-dimensional data set is represented by the binary string constructed by concatenating the first bits of the k domain values, the second bits, the third bits, and so forth. Using the Mellin transform it is proven that the average search time is $O(n^{1-s/k})$ under the assumption of an independent uniform distribution of the bits. In terms of ordinary strings this corresponds to matching with a fixed mask of don't cares that is iterated through the pattern.

Baeza-Yates and Gonnet [2] study the problem of searching regular expressions in a trie. The deterministic finite-state automaton for the regular expression is built, its size depending only upon the query size (although possibly exponential in the size of the query). The automaton is simulated on the trie and a hit is reported every time a final state is reached. Extending the average-case analysis of Flajolet and Puech [10], the authors are able to show that the average search time depends upon the largest eigenvalue (and its multiplicity) of the incidence matrix of the automaton. As a result, they find that a sublinear number of nodes of the trie is visited. Apostolico and Szpankowski [1] note that suffix trees and tries for independent strings asymptotically do not differ too much, which is an argument for transferring the results on tries to suffix trees. In another article Baeza-Yates and Gonnet [3] study the average cost of calculating an all-against-all sequence matching. Here, for all strings the substrings that match each other with a certain (fixed) number of errors are sought. With the use of tries the average time is shown to be subquadratic.

For approximate indexing (with edit distance) Navarro and Baeza-Yates [21] have proposed a method that flexibly partitions the pattern in pieces that can be searched in sublinear time in the suffix tree for a text. For an error rate $\alpha = k/m$, where m is the pattern length and k the allowed number of errors, they show that a sublinear search is possible if $\alpha < 1 - e/\sqrt{\sigma}$, thereby partitioning the pattern into $j = (m + k)/\log_\sigma n$ pieces. The threshold plays two roles, it gives a bound on the search depth in a suffix tree and it gives a bound on the number of verifications needed. In [20] the bound is investigated more closely. It is conjectured that the real threshold, where the number of matches of a pattern in a text decreases exponentially fast in the pattern length, is $\alpha = 1 - c/\sqrt{\sigma}$ with $c \approx 1.09$. Higher error rates make a filtration algorithm useless because of too many verifications.

More careful tree traversal techniques can lower the number of nodes that need to be visited. This idea is pursued by Jokinen and Ukkonen [14] (on a DAWG), Ukkonen [32], and Cobbs [7]. No exact average-case analysis is available for these algorithms.

The start of precise analysis of algorithms is contributed to Knuth (i.e., [25]). Especially the analysis of digital trees has yielded a vast amount of results. The Mellin transform and Rice's integrals have been the methods of choice for many results dating back as early as the analysis of radix exchange sort in Knuth's famous books [17]. See [29] for a recent book with a rich bibliography.

Our analysis of the average search time in the trie leads to an alternating sum of the type

$$\sum_{k=m}^{n} \binom{n}{k} (-1)^k f(n,k). \tag{1}$$

The above sum is intimately connected to tries and appears very often in their analysis (see, e.g., [17]). Similar sums, where $f(n,k)$ only depends on k, have also been considered by Szpankowski [27] and Kirschenhofer [16]. The asymptotic analysis can often be done through Rice's integrals (a technique that already appears in Section 1 in Chapter 8 of [22]). It transfers the sum to a complex integral, which is evaluated by the Cauchy residue theorem.

Our main contribution is the general analysis of error models that depend only on the comparison of two characters and limit the number of errors allowed before regarding two strings as different. Unless the pattern length is very short, the asymptotic order of the running time depends on an error-threshold relative to the number of strings in the database and independent of the pattern length. The methods applied here can be used to determine exact asymptotics for each concrete error bound. It also allows us to estimate the effect of certain error models in limiting the search time. Furthermore, for constant error bounds we find thresholds with respect to the error probability which reveal an interesting behavior hidden for the most used model, the Hamming distance.

3. Main Results. The trie T for a set of strings $X_1, \ldots, X_n$ is a rooted, directed tree where each edge is labeled with a character from Σ, all outgoing edges of any node are labeled with different characters, and the strings spelled out by the edges leading from the root to the leaves are exactly $X_1, \ldots, X_n$. We store $\mathsf{value}(v) = i$ at leaf v, if the path to v spells out the string X_i. The paths from the last branching nodes to the leaves are often compressed to a single edge.

When searching for a pattern P we want to know all strings X_i such that P is a prefix of X_i or vice versa (with the special case of all strings having the same length). The assumption that all strings have infinite length is not severe. Indeed, this reflects the situation that the pattern is not found. Otherwise, the search would be ended earlier, so our analysis gives an upper bound. Pseudocode for the analyzed algorithms is given in Figure 1.

We focus mainly on the TS algorithm, the LS algorithm is used as a benchmark. Our main result is the following theorem. For fixed d the constants in the Landau symbols and further terms of the asymptotic can also be computed (or at least bounded).

LS Algorithm

Input: Strings $X_1, \ldots, X_n$ and pattern P, bound d.
for i **from** 1 **to** n **do**
  $j := 1$
  $c := 0$
  $l := \min\{\mathsf{length}(P), \mathsf{length}(X_i)\}$
  while $c \le d$ **do**
    while $j \le l$ **and** $\mathsf{match}(P[j], X_i[j])$ **do**
      $j := j + 1$
    $c := c + 1$
    $j := j + 1$
  if $j - 2 = l$ **then**
    report match for X_i

TS Algorithm : $\mathsf{rfind}(v, P, pos, d)$

if $d \ge 0$ **then**
  if v is a leaf **then**
    report match for $X_{\mathsf{value}(v)}$
  else if $pos > \mathsf{length}(P)$ **then**
    for all leaves u in the subtree of v **do**
      report match for $X_{\mathsf{value}(u)}$
  else
    for each child u of v **do**
      let c be the edge label of (u, v)
      if $\mathsf{match}(P[pos], c)$ **then**
        $\mathsf{rfind}(u, P, pos + 1, d)$
      else
        $\mathsf{rfind}(u, P, pos + 1, d - 1)$

Fig. 1. Pseudocode of the LS and the TS algorithms. The recursive TS algorithm is started with $\mathsf{rfind}(r, P, 0, d)$, where r is the root of the trie for the strings $X_1, \ldots, X_n$.

THEOREM 1 (Average Complexity of the TS algorithm).

$$\mathbf{E}[T_n^d] = \begin{cases} O((\log n)^{d+1}), & \text{for } d = O(1) \text{ and } p = \sigma^{-1}, \\ O((\log_\sigma n)^d n^{\log_\sigma p + 1}), & \text{for } d = O(1) \text{ and } p > \sigma^{-1}, \\ O(1), & \text{for } d = O(1) \text{ and } p < \sigma^{-1}, \\ o(n), & \text{for } d + 1 < q \log_\sigma n, \\ \Omega(n \log_\sigma n), & \text{for } d + 1 > q \log_\sigma n. \end{cases} \tag{2}$$

More exact bounds are possible through (27), (35), and (36) for $d = O(1)$ and the cases $p < \sigma^{-1}$, $p = \sigma^{-1}$, and $p > \sigma^{-1}$. These results can be applied to different models. For instance, for the Hamming distance model with alphabet size 4 we get the exact first-order term $((4 \cdot 3^d)/(d+1)!)(\log_4 n)^{d+1}$. For $d \sim c \log_\sigma n$ we give some more exact estimates in Remark 13.

It is well known that the average depth of a trie is asymptotically equal to $\log_\sigma n$ (see, e.g., [24] and [28]). When no more branching takes place the TS and LS algorithms behave the same; both algorithms perform a constant number of comparisons on average. If we allow enough errors to go beyond the depth of the trie, they should perform similarly. With an error probability of q we expect to make qm errors on m characters. Thus, it comes as no surprise that the threshold is $q \log_\sigma n$.

With respect to the matching probability p we have a different behavior for the three cases $p < \sigma^{-1}$, $p = \sigma^{-1}$, and $p > \sigma^{-1}$. To explain this phenomena we take a look at the conditional probability of a match for an already chosen character. If $p < \sigma^{-1}$, then the conditional probability must be smaller than 1, i.e., with some probability independent of the pattern, we have a mismatch and thus restrict the search independently of the pattern. If $p > \sigma^{-1}$, the conditional probability must be greater than 1. Hence, with some probability independent of the pattern, we have a match and thereby extend our search. This restriction or extension is independent of the number of errors allowed and, hence, the additional factor in the complexity.

For the model where we bound the number of don't cares we have $q = 2/\sigma - 1/\sigma^2$ and $p = 1 - 2/\sigma + 1/\sigma^2$. In the SNP database problem mentioned above, the alphabet size is $\sigma = 4$, including the don't care character. We find that the average-case behavior, bounding only the don't cares, is approximately $O((\log n)^d n^{0.585})$ when allowing d don't cares. For the number of mismatches we could resort to the Hamming distance case mentioned above, but in this application a don't care cannot induce a mismatch. Therefore, the average-case complexity is approximately $O((\log n)^d n^{0.661})$ when allowing d mismatches. This is significantly worse than the Hamming distance alone, which is $O((\log n)^{d+1})$. It also dominates the bound on the number of don't cares. When deciding whether the LS or the TS algorithm should be used in this problem, we find that for $d > \frac{3}{8}\log_4 n - 1$ the LS algorithm will outperform the TS algorithm. As another application we apply our results to the model used by Buchner et al. [5], [6] for searching protein structures. Here the angles of a protein folding are used for approximate search of protein substructures. The full range of 360° is discretized into an alphabet $\Sigma = \{[0, 15), \ldots, [345, 360)\}$. The algorithm then searches a protein substructure by considering all angles within a number of intervals to the left and right, i.e., for $i = 2$ intervals to both sides, the interval $[0, 15)$ matches $[330, 345)$, $[345, 360)$, $[0, 15)$, $[15, 30)$, and $[30, 45)$. If i intervals to the left or right are allowed, then the probability of a match is $(2i + 1)/\sigma$. In their application Buchner et al. [5], [6] allow no mismatch, i.e., the search is stopped if the angle is not within the specified range. The asymptotic running time is thus $O(n^{\log_\sigma (2i+1)})$ if i intervals to the left or right are considered. Although a suffix tree is used and we do not expect the underlying distribution of angles to be uniform and memoryless, this result can be used as a (rough) estimate, especially of the effect of different choices of i.

4. Basic Analysis. For completeness we give a quick derivation of the expected value of L_n^d, the number of comparisons made by the **LS algorithm**. The probability of k comparisons is

$$\Pr\{L_n^d = k\} = \sum_{i_1+\cdots+i_n=k} \prod_{j=1}^{n} \binom{i_j - 1}{d} q^{d+1} p^{i_j - d - 1}. \tag{3}$$

From it we can derive the probability generating function

$$g_{L_n^d}(z) = \mathbf{E}[z^{L_n^d}] = \sum_{k=0}^{\infty} \Pr\{L_n^d = k\} z^k = \left(\frac{zq}{1 - zp}\right)^{n(d+1)}, \tag{4}$$

which yields the expected value $\mathbf{E}[L_n^d] = ((d + 1)/q)n$. The stochastic process is very stable. We can uses Chebyshev's inequality to derive convergence in probability of L_n^d:

$$\Pr\left\{\left|\frac{L_n^d}{n(d+1)} - \frac{1}{q}\right| > \varepsilon\right\} = \Pr\left\{\left|L_n^d - \frac{n(d+1)}{q}\right| > \varepsilon n(d+1)\right\} < \frac{p}{q^2\varepsilon^2 n(d+1)}.$$

Hence, we have

$$\lim_{n\to\infty} \frac{L_n^d}{n(d+1)} = \frac{1}{q} \quad \text{(pr.)}.$$

Note that if $d = d(n)$ is a function of n, we already have almost sure convergence if $d(n) = \omega(\log^{1+\varepsilon} n)$ by a simple application of the Borel–Cantelli lemma and the fact that $\sum_n 1/(n \log^{1+\varepsilon} n)$ is convergent. Using a method of Kesten and Kingman (see, for instance, [29] or [15]) this can be extended to almost sure convergence. Note that $L_n^d < L_{n+1}^d$, so L_n^d is non-decreasing. For any positive constant s let $r = r(n)$ be the largest integer with $r^s \le n$, then $L_{r^s}^d \le L_n^d \le L_{(r+1)^s}^d$ and thus

$$\limsup_{n\to\infty} \frac{L_n^d}{n(d+1)} \le \limsup_{r\to\infty} \frac{L_{r^s}^d}{(r+1)^s(d+1)} = \limsup_{r\to\infty} \frac{L_{r^s}^d}{r^s(d+1)} \frac{r^s}{(r+1)^s}$$

and equally

$$\liminf_{n\to\infty} \frac{L_n^d}{n(d+1)} \ge \liminf_{r\to\infty} \frac{L_{(r+1)^s}^d}{r^s(d+1)} = \liminf_{r\to\infty} \frac{L_{(r+1)^s}^d}{(r+1)^s(d+1)} \frac{(r+1)^s}{r^s}.$$

By the Borel–Cantelli lemma, $L_{r^s}^d/r^s(d+1)$ converges to $1/q$ almost surely, since for all $s \ge 1+\varepsilon$ we have

$$\sum_{r=0}^{\infty} \Pr\left\{ \left| \frac{L_{r^s}^d}{r^s(d+1)} - \frac{1}{q} \right| \ge \varepsilon \right\} \le \sum_{r=0}^{\infty} \frac{p}{q^2\varepsilon^2} \cdot \frac{1}{d+1} \cdot \frac{1}{r^s} = \frac{p}{q^2\varepsilon^2} \cdot \frac{1}{d+1} \cdot \zeta(s) < \infty$$

and thus

$$\lim_{r\to\infty} \frac{L_{r^s}^d}{r^s(d+1)} = \frac{1}{q} \quad \text{(a.s.)}.$$

Since $(r+1)^s \sim r^s$ we have

$$\lim_{n\to\infty} \frac{L_n^d}{n(d+1)} = \frac{1}{q} \quad \text{(a.s.)}.$$

Note that, interpreting $d = d(n)$ as a function of n, this also holds for $d(n) = \Omega(1)$.

For the **TS algorithm** it is easier to examine the number of nodes visited. Observe that the number of nodes visited is larger by one than the number of character comparisons. Each time a node is visited the remaining leaves split up by a random choice of the next character. Depending on the next character of the pattern the number of allowed mismatches in the subtree may stay the same (with probability p) or may decrease (with probability q). For the average number we can set up the following equation:

$$\text{(5)} \quad \mathbf{E}[T_n^d] = 1 + \sum_{i_1+\cdots+i_\sigma=n} \binom{n}{i_1, \ldots, i_\sigma} \sigma^{-n} \left(\sum_{j=1}^{\sigma} q\mathbf{E}[T_{i_j}^{d-1}] + \sum_{j=1}^{\sigma} p\mathbf{E}[T_{i_j}^d] \right).$$

The boundary conditions are $\mathbf{E}[T_n^{-1}] = 1$, counting the character comparison that induced the last mismatch, and $\mathbf{E}[T_0^d] = 0$. For $n = 1$ we have $\mathbf{E}[T_1^d] = 1 + (d+1)/q$, which is the same as $\mathbf{E}[L_1^d]$, except that additionally the root is counted.

From (5) we can derive the exponential generating function of $\mathbf{E}[T_n^d]$:

$$\text{(6)} \qquad t^d(z) = e^z + \sum_{j=1}^{\sigma} q t^{d-1}\left(\frac{z}{\sigma}\right) e^{(1-1/\sigma)z} + \sum_{j=1}^{\sigma} p t^d\left(\frac{z}{\sigma}\right) e^{(1-1/\sigma)z} - 1.$$

We multiply with $\exp(-z)$ (the so-called Poisson transform, see, e.g., Section 7.6.1 of [29]) and define $\tilde{t}^d(z) = t^d(z)e^{-z}$. We have

$$\tilde{t}^d(z) = 1 - \exp(-z) + \sigma q \tilde{t}^{d-1}\left(\frac{z}{\sigma}\right) + \sigma p \tilde{t}^d\left(\frac{z}{\sigma}\right). \tag{7}$$

Let y_n^d be the coefficients of $\tilde{t}^d(z)$. Then we have

$$y_n^d = (-1)^n \sum_{k=0}^{n} \binom{n}{k} (-1)^k \mathbf{E}[T_k^d] \quad \text{and} \quad \mathbf{E}[T_n^d] = \sum_{k=0}^{n} \binom{n}{k} y_k^d.$$

We get the boundary conditions $y_1^d = 1 + (d+1)/q$, $y_0^d = 0$, $y_n^{-1} = (-1)^{n-1}$ for $n > 0$, and $y_0^{-1} = 0$. Comparing coefficients in (7) we find that for $n > 1$,

$$y_n^d = \frac{(-1)^{n-1} + y_n^{d-1}\sigma^{1-n}q}{1 - \sigma^{1-n}p}, \tag{8}$$

which by iteration leads to

$$y_n^d = \frac{(-1)^n}{1 - \sigma^{1-n}} \left(\sigma^{1-n} \left(\frac{\sigma^{1-n}q}{1 - \sigma^{1-n}p} \right)^{d+1} - 1 \right). \tag{9}$$

Finally, we translate this back to

$$\mathbf{E}[T_n^d] = n\left(1 + \frac{d+1}{q}\right) + \sum_{k=2}^{n} \binom{n}{k} \frac{(-1)^k}{\sigma^{k-1} - 1} \left(\frac{q\sigma^{1-k}}{1 - p\sigma^{1-k}} \right)^{d+1} - \sum_{k=2}^{n} \binom{n}{k} \frac{(-1)^k}{1 - \sigma^{1-k}}. \tag{10}$$

Let $A_n := \sum_{k=2}^{n} \binom{n}{k} ((-1)^k/(1 - \sigma^{1-k}))$. A similar derivation to the above shows that the sum is the solution to

$$A_n = n - 1 + \sum_{i_1 + \cdots + i_\sigma = n} \binom{n}{i_1, \ldots, i_\sigma} \sigma^{-n} \sum_{j=1}^{\sigma} A_{i_j}, \tag{11}$$

which we call the average "compression number." It gives the average sum of the number of characters "hidden" by all edges, i.e., an edge with n leaves in its subtree "hides" $n-1$ characters (which would be examined by the LS but not by the TS algorithm). Hence, $n(d+1)/q - A_n$ is an upper bound for the average performance of the TS algorithm. The average case can easily be evaluated using Rice's integrals (see Theorem 5).

LEMMA 2 (Asymptotic Behavior of the Compression Number). *The asymptotic behavior of A_n is*

$$A_n = n \log_\sigma n + n \left(\frac{1}{2} - \frac{1 - \gamma}{\ln \sigma} + \frac{\sum_{k \in \mathbb{Z} \setminus \{0\}} n^{-2\pi \imath k / \ln \sigma} \Gamma(-1 + 2\pi \imath k / \ln \sigma)}{\ln \sigma} \right) + O(1).$$

PROOF. Already, in Section 5.2.2 of [17], the asymptotic behavior of

$$U_n = \sum_{k=2}^{n} \binom{n}{k} \frac{(-1)^k}{2^{k-1} - 1}$$

is determined to be

$$U_n = n\log_2 n + n\left(\frac{\gamma-1}{\ln 2} - \frac{1}{2} + \frac{\sum_{k\in\mathbb{Z}\setminus\{0\}} n^{-2\pi\imath k/\ln 2}\Gamma\left(-1-2\pi\imath k/\ln 2\right)}{\ln 2}\right) + O(1).$$

If we replace σ by $\frac{1}{2}$ in A_n we have the same formula as for $-U_n$. The formula also serves as an example in many other works (see, e.g., [11]); therefore, we omit a detailed proof. □

One can show that $\sum_{k=1}^{\infty} |\Gamma\left(-1+2\pi k\imath/\ln\sigma\right)|$ is very small (below one for $\sigma \le 10^6$), but growing in σ. We now turn to the evaluation of the sum

$$\mathfrak{S}_n^{(d)} := \sum_{k=2}^{n} \binom{n}{k} \frac{(-1)^k}{\sigma^{k-1}-1} \left(\frac{q}{\sigma^{k-1}-p}\right)^{d+1}. \tag{12}$$

Note that if $\mathfrak{S}_n^{(d)}$ is sublinear the main term of the asymptotic growth of $\mathbf{E}[T_n^d]$ is determined by (10) and Lemma 2 to be $n\left((d+1)/q - \log_\sigma n\right)$, i.e., the the TS and LS algorithms have the same asymptotic complexity in this case.

5. Asymptotic Analysis. We can prove two theorems regarding the growth of $\mathfrak{S}_n^{(d)}$ for different bounds d. For constant d we can give a very precise answer.

THEOREM 3 (Searching with a Constant Bound). *Let $d = O(1)$, then*

$$\mathfrak{S}_n^{(d)} = -n\left(1+\frac{d+1}{q}\right) + A_n + \begin{cases} O((\log n)^{d+1}), & \text{for} \quad p = \sigma^{-1}, \\ O((\log_\sigma n)^d n^{\log_\sigma p+1}), & \text{for} \quad p > \sigma^{-1}, \\ x\,O\left(1\right), & \text{otherwise}. \end{cases} \tag{13}$$

For logarithmic d we give a less exact answer, which yields a threshold where the complexity jumps from sublinear to linear logarithmic.

THEOREM 4 (Searching with a Logarithmic Bound). *If $d+1 = c\log_\sigma n$, then we have*

$$\mathfrak{S}_n^{(d)} = \begin{cases} -n\left(1+(d+1)/q\right) + A_n + o(n), & \text{for} \quad c < q, \\ o(n), & \text{for} \quad c > q. \end{cases} \tag{14}$$

The two theorems immediately yield Theorem 1. Both proofs rely on transferring the sum to a complex integral by Rice's integrals (see for instance Section 8.1 in Chapter 8 of [22]).

THEOREM 5 (Rice's Formula). *Let $f(z)$ be an analytic continuation of $f(k) = f_k$ that contains the half-line $[m,\infty)$. Then*

$$\sum_{k=m}^{n} (-1)^k \binom{n}{k} f_k = \frac{(-1)^n}{2\pi\imath} \int_{\mathcal{C}} f(z) \frac{n!}{z(z-1)\cdots(z-n)}\, dz, \tag{15}$$

where $\mathcal{C}$ is a positively oriented curve that encircles $[m, n]$ and does not include any of the integers $0, 1, \ldots, m-1$ or other singularities of $f(z)$.

For details on this method we refer the reader to the literature [17], [27], [11], [16], and [29]. The proof of the theorem stems from the Cauchy residue theorem. The function

$$z \to \frac{1}{\sigma^{z-1}-1}\left(\frac{q}{\sigma^{z-1}-p}\right)^{d+1}$$

is meromorphic; therefore, the value of the sum $\mathfrak{S}_n^{(d)}$ can be expressed as the sum of the residues (see, e.g., Theorem 2 of [11]). We do not take this approach for two reasons: The poles in consideration are of order d and they involve the Beta function. For the proof of Theorem 4, we let $d \sim c \log n$, which makes the evaluation of the residues very hard. Therefore, we first approximate the Beta function in the integral by a ratio of Gamma functions. Along the way we express the sum as a simple integral that we can later bound for the case $d \sim c \log n$ to prove Theorem 4. Finally, we can evaluate the remaining residues involving the simpler Gamma function.

We now apply Theorem 5 and then begin with the approximation of the resulting integral.

LEMMA 6 (from Sum to Integral). *For $1 < \xi < 2$ we have*

$$\mathfrak{S}_n^{(d)} = \frac{1}{2\pi\imath}\int_{-\xi-\imath\infty}^{-\xi+\imath\infty}\frac{1}{\sigma^{-1-z}-1}\left(\frac{q}{\sigma^{-1-z}-p}\right)^{d+1}B(n+1,z)\,dz + O(1). \tag{16}$$

PROOF. By Rice's theorem we can write $\mathfrak{S}_n^{(d)}$ as

$$\mathfrak{S}_n^{(d)} = \frac{(-1)^n}{2\pi\imath}\int_{\mathcal{C}}\frac{1}{\sigma^{-1-z}-1}\left(\frac{q}{\sigma^{-1-z}-p}\right)^{d+1}B(n+1,z)\,dz,$$

where $\mathcal{C}$ is a positively oriented curve that encircles $[2, m]$ and does not include the integers 0 and 1 or other singularities. Let

$$f(z) = \frac{1}{\sigma^{-1-z}-1}\left(\frac{q}{\sigma^{-1-z}-p}\right)^{d+1}.$$

The function $f(z)$ is periodic with respect to the imaginary part of z and bounded (or decreasing) with respect to the real part. We only need to avoid its singularities. Thus, the growth of $f(z)$ for $z \to \infty$ is $O(1)$ if the singularities at $z = 1$, $z = 0$, $z = 1 \pm 2\pi\imath k/\ln\sigma$, and $z = -\log_\sigma p - 1 \pm 2\pi\imath k/\ln\sigma$, $k \in \mathbb{Z}$, are avoided, whereas $B(n+1,z) = n!/(z(z+1)\cdots(z+n)) = O(z^{-n-1})$. Let $\mathcal{K}$ be a circle of radius M, carefully chosen to avoid any singularities. Then for some constant C we have

$$\left|\frac{1}{2\pi\imath}\int_{\mathcal{K}}\frac{1}{\sigma^{-1-z}-1}\left(\frac{q}{\sigma^{-1-z}-p}\right)^{d+1}B(n+1,z)\,dz\right| \le MCn!\,(M-n)^{-n} \xrightarrow[M\to\infty]{} 0.$$

The same holds for any partial path on $\mathcal{K}$. Splitting $\mathcal{K}$ into two half circles divided by the line $\Re(z) = -\xi$, we find that the negative sum of the residues in the right half must

be the same as $\mathfrak{S}_n^{(d)}$, except for some small error $O(1)$ dependent on the choice of M. This proves (16). □

The integral in (16) can be extended to a half-circle to the right because the contribution of the bounding path is very small. Hence, the integral is equal to the sum of the negative of the residues right to the line $\Re(z) = -\xi$. These residues are located at $z = 1$, $z = 0$, $z = 1 \pm 2\pi\imath k/\ln\sigma$, and $z = -\log_\sigma p - 1 \pm 2\pi\imath k/\ln\sigma$, $k \in \mathbb{Z}$. The real part of the last ranges from 1 to $1 - \log_\sigma(\sigma^2 - 1) > -1$ under the assumption that $\sigma^{-2} \le p \le 1 - \sigma^{-2}$. The last is fulfilled if there can be at least one mismatch and at least one match (otherwise, everything becomes trivial).

The evaluation of the residues proves tricky for the Beta function. We approximate the Beta function using an asymptotic expansion by Tricomi and Erdélyi [31] with the help of a result of Fields [9]. This approach was already used by Szpankowski [27]. In the following we lay a rigorous basis for it.

LEMMA 7 (Asymptotic Approximation of a Beta Function Integral). *For real constant* $x \notin \{0, -1, -2, \ldots\}$ *we have*

$$\int_{-\infty}^{\infty} |B(n, x + \imath y)|\, dy = O(n^{-x}). \tag{17}$$

PROOF. We begin by applying Stirling's formula. Then we divide the range of integration into parts handled separately.

$$\begin{aligned}\int_{-\infty}^{\infty} |B(n, x + \imath y)|\, dy &= 2\int_0^{\infty} |B(n, x + \imath y)|\, dy = 2\int_0^{\infty} \frac{\Gamma(n)\,|\Gamma(x + \imath y)|}{|\Gamma(n + x + \imath y)|}\, dy \\ &= 2\int_0^{\infty} \sqrt{2\pi}\left(\frac{n}{|n + x + \imath y|}\right)^{n-1/2}\left(\frac{|x + \imath y|}{|n + x + \imath y|}\right)^{x} \\ &\qquad \times \frac{1}{\sqrt{|x + \imath y|}} e^{-y\varphi(n,x,y)}\, dy(1 + O(n^{-1})),\end{aligned}$$

where $0 < \varphi(n, x, y) = \arg(x + \imath y) - \arg(x + n + \imath y) < \pi/2$. Since

$$\cos(\varphi(n, x, y)) = \frac{nx + x^2 + y^2}{\sqrt{x^2 + y^2}\sqrt{n^2 + 2nx + x^2 + y^2}}. \tag{18}$$

We analyze $\varphi(n, x, y)$ as a function $\varphi(y)$ of $y \ge 0$. Note that $\cos(\varphi(y))$ grows inversely to $\varphi(y)$ on $[0, \pi/2]$. If $x < 0$, we have

$$\frac{nx + x^2 + y^2}{\sqrt{x^2 + y^2}\sqrt{n^2 + 2nx + x^2 + y^2}} = 0 \quad \Leftrightarrow \quad y = \sqrt{-nx - x^2}.$$

The derivative of (18) is

$$\frac{d}{dy}\cos(\varphi(n, x, y)) = \frac{yn^2(-nx - x^2 + y^2)}{(x^2 + y^2)^{3/2}(n^2 + 2nx + x^2 + y^2)^{3/2}},$$

which, for $x > 0$, is 0 if $y = \sqrt{nx + x^2}$ or $y = 0$. There is a minimum at $y = \sqrt{nx + x^2}$. Hence, $\varphi(y)$ goes from 0 to a maximum at $y = \sqrt{nx + x^2}$ and then decreases back to 0. For $x < 0$ the derivative is always positive, and $\varphi(y)$ decreases monotonically from $\pi - \varepsilon$ to 0.

The term $n/|n + x + \imath y|$ is monotonically decreasing in y and tends to 0. The term $|x + \imath y|/|n + x + \imath y|$ is monotonically increasing in y and tends to 1.

We make a case distinction for the integration interval. For $x > 0$ we get

$$\begin{aligned}
&\int_0^x \left(\frac{n}{|n + x + \imath y|}\right)^{n-1/2} \left(\frac{|x + \imath y|}{|n + x + \imath y|}\right)^x \frac{1}{\sqrt{|x + \imath y|}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le \left(\frac{n}{n + x}\right)^{n-1/2} \left(\sqrt{\frac{x^2 + x^2}{n^2 + 2nx + 2x^2}}\right)^x \frac{1}{\sqrt{x}} \int_0^x e^0\, dy \\
&\qquad \le (\sqrt{2}x)^x n^{-x} \sqrt{x} = O(n^{-x}).
\end{aligned}$$

For n large enough we have

$$\begin{aligned}
&\int_x^n \left(\frac{n}{|n + x + \imath y|}\right)^{n-1/2} \left(\frac{|x + \imath y|}{|n + x + \imath y|}\right)^x \frac{1}{\sqrt{|x + \imath y|}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le \left(\sqrt{\frac{n^2}{(n + x)^2 + x^2}}\right)^{n-1/2} \frac{1}{\sqrt{2x}} \int_x^n \left(\sqrt{\frac{x^2 + y^2}{n^2}}\right)^x e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le \frac{n^{-x}}{\sqrt{2x}} \int_x^n \left(\sqrt{2}y\right)^x e^{-y(\pi/5)}\, dy \le \frac{n^{-x}}{\sqrt{2x}} \frac{5}{\pi} \left(\sqrt{2}\frac{5}{\pi}\right)^x \Gamma(x + 1) \\
&\qquad = O(n^{-x}).
\end{aligned}$$

Since $\varphi(n, x, x) \xrightarrow[n\to\infty]{} \arccos(1/\sqrt{2}) = \pi/4$ and $\varphi(n, x, n) \xrightarrow[n\to\infty]{} \arccos(1/\sqrt{2}) = \pi/4$ we have $\varphi(n, x, y) > \pi/5$ for some n large enough.

The third part is

$$\begin{aligned}
&\int_n^{n^2} \left(\frac{n}{|n + x + \imath y|}\right)^{n-1/2} \left(\frac{|x + \imath y|}{|n + x + \imath y|}\right)^x \frac{1}{\sqrt{|x + \imath y|}} e^{-y\varphi(n,x,y)} dy \\
&\qquad \le \left(\sqrt{\frac{n^2}{(n + x)^2 + n^2}}\right)^{n-1/2} \left(\sqrt{\frac{x^2 + n^4}{(n + x)^2 + n^4}}\right)^x (x^2 + n^2)^{-1/4} n^2 \\
&\qquad = O(2^{-n/2} n^2).
\end{aligned}$$

For the last part we have

$$\begin{aligned}
&\int_{n^2}^\infty \left(\frac{n}{|n + x + \imath y|}\right)^{n-1/2} \left(\frac{|x + \imath y|}{|n + x + \imath y|}\right)^x \frac{1}{\sqrt{|x + \imath y|}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le \int_{n^2}^\infty \left(\sqrt{\frac{n^2}{(n + x)^2 + yn^2}}\right)^{n-1/2} \frac{1}{\sqrt{y}}\, dy \le \int_{n^2}^\infty y^{-n/2-1/4}\, dy \\
&\qquad = O(n^{-n-1/2}).
\end{aligned}$$

Thus the whole integral is $O(n^{-x})$.

For $x < 0$ ($-x \notin \mathbb{N}$), the derivation is almost the same. We start with

$$\begin{aligned}
&\int_0^n \left(\frac{n}{|n+x+\imath y|}\right)^{n-1/2} \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^{x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le \sqrt{\frac{n}{n+x}} \left(\frac{n}{n+x}\right)^{n} \int_0^n \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^{x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le 4e^{-x}x^x\,(2n)^{-x}\,\frac{1}{\sqrt{x}} \int_0^n e^{-y(\pi/5)}\, dy = n^{-x}\frac{4(2e)^{-x}x^x}{\sqrt{x}}\,\frac{-5}{\pi}(e^{-(1/5)n\pi}-1) \\
&\qquad = O(n^{-x}).
\end{aligned}$$

Since $(n/(n+x))^n < 2e^{-x}$, $\sqrt{n/(n+x)} < 2$, and $\varphi(n,x,y) > \pi/5$ for some n large enough (with $\varphi(n,x,n) \xrightarrow[n\to\infty]{} = \pi/4$ and $\varphi(n,x,0) = \pi$).

The second part is

$$\begin{aligned}
&\int_n^{n^2} \left(\frac{n}{|n+x+\imath y|}\right)^{n-1/2} \left(\frac{|n+x+\imath y|}{|x+\imath y|}\right)^{-x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le \left(\sqrt{\frac{n^2}{(n+x)^2+n^2}}\right)^{n-1/2} \left(\sqrt{\frac{(n+x)^2+n^2}{x^2+n^2}}\right)^{-x} (x^2+n^2)^{-1/4} n^2 \\
&\qquad \le \left(\sqrt{\frac{n^2}{\frac{3}{2}n^2}}\right)^{n-1/2} 2^{-x/2} n^2 = O((\tfrac{3}{2})^{-n/2} n^2).
\end{aligned}$$

The final part is

$$\begin{aligned}
&\int_{n^2}^{\infty} \left(\frac{n}{|n+x+\imath y|}\right)^{n-1/2} \left(\frac{|n+x+\imath y|}{|x+\imath y|}\right)^{-x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le \int_{n^2}^{\infty} \left(\frac{n}{\sqrt{(n+x)^2+y^2}}\right)^{n-1/2} \left(\sqrt{\frac{2y^2}{y^2}}\right)^{-x} \frac{1}{\sqrt{y}} e^{-y\varphi(n,x,y)}\, dy \\
&\qquad \le 2^{-x/2} \int_{n^2}^{\infty} y^{-n/2-1/4}\, dy \\
&\qquad = O(n^{-n-1/2}).
\end{aligned}$$

This finishes the proof. □

LEMMA 8 (Tail of a Beta Function Integral). *For $x < 0$ and any strictly positive function $f(n) \in \omega(1)$ we have*

$$\int_{f(n)\ln n}^{\infty} |B(n, x+\imath y)|\, dy = O(n^{-f(n)(\pi/4-\varepsilon)-x}). \tag{19}$$

PROOF. We use the same derivations as in Lemma 7 together with $\varphi(n,x,n) \xrightarrow[n\to\infty]{} = \pi/4$, and

$$\begin{aligned}
&\int_{f(n)\ln n}^{n} \left(\frac{n}{|n+x+\imath y|}\right)^{n-1/2} \left(\frac{|n+x+\imath y|}{|x+\imath y|}\right)^{-x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\varphi(n,x,y)}\,dy \\
&\quad \le\ e^{-x}\left(\frac{f(n)\ln n}{\sqrt{2}n}\right)^{x} (\ln n)^{-(1+\varepsilon)/2} \int_{f(n)\ln n}^{n} e^{-y(\pi/4-\varepsilon)} dy \\
&\quad \le\ e^{-x}\left(\frac{f(n)\ln n}{\sqrt{2}n}\right)^{x} (f(n)\ln n)^{-1/2} \frac{1}{(\pi/4-\varepsilon)} e^{-f(n)\ln n(\pi/4-\varepsilon)} \\
&\quad =\ O(n^{-f(n)(\pi/4-\varepsilon)-x}).
\end{aligned}$$

□

THEOREM 9 (Approximation of Beta Integrals). *Let $f(n,z)$ be a function such that $|f(n,z)| = O(n^k)$ for a constant k. Let $z = x+\imath y$. We can approximate for some constant c and fixed N,*

$$\begin{aligned}
\int_{x-\imath\infty}^{x+\imath\infty} f(n,z)B(n,z)\,dz =& \sum_{k=0}^{N-1} \int_{x-\imath\infty}^{x+\imath\infty} f(n,z)\frac{(-1)^k B_k^{(1-z)}(1)}{k!}\Gamma(z+k)n^{-k-z}\,dz \\
&+ c\int_{x-\imath\infty}^{x+\imath\infty} f(n,z)n^{-N-z}\Gamma(z)\,dz + O(n^{-N/3-x}e^{-(\pi/3)n^{1/3}}).
\end{aligned}$$

PROOF. The proof relies on a standard expansion of the ratio of two Gamma functions by Tricomi and Erdélyi [31], the proof that the expansion is uniform by Fields [9], and the exponentially small tails of the integrands. The expansion is

$$\begin{aligned}
\frac{\Gamma(z+\alpha)}{\Gamma(z+\beta)} \ =\ & z^{\alpha-\beta}\sum_{k}\frac{1}{k!}\frac{\Gamma(1+\alpha-\beta)}{\Gamma(1+\alpha-\beta-k)}B_k^{(1+\alpha-\beta)}(\alpha)z^{-k} \\
&+ O(z^{\alpha-\beta-m}(1+|\alpha-\beta|^m)(1+|\alpha|+|\alpha-\beta|)^m),
\end{aligned} \tag{20}$$

which is uniformly valid for $|\arg(z+\alpha)| < \pi$, and $(1+|\alpha-\beta|)(1+|\alpha|+|\alpha-\beta|) = o(z)$, $\beta-\alpha \notin \mathbb{N}$, $z\to\infty$. The $B_n^{(a)}(x)$ are the generalized Bernoulli polynomials (see [30] or Section 5 of Chapter 6 in [22], which are multivariate polynomials in a and x of degree n. The first polynomials are $B_0^{(a)}(x) = 1$, $B_1^{(a)}(x) = -a/2+x$, and $B_2^{(a)}(x) = (3a^2+12x^2-a(1+12x))/12$.

Assume $x < 0$. By (20), we can approximate $B(n,x+\imath y) \le g_N(n,x+\imath y)$ on the interval $0 \le y \le n^{1/3}$ by

$$\begin{aligned}
g_N(n,x+\imath y) \ =\ & \sum_{k=0}^{N-1}\frac{(-1)^k B_k^{(1-x-\imath y)}(1)}{k!}\Gamma(x+\imath y+k)n^{-k-x-\imath y} \\
&+ cn^{-N-x-\imath y}|y|^{2N}\Gamma(x+\imath y).
\end{aligned} \tag{21}$$

For $y > n^{1/3}$ we use Lemma 8 and get

$$\int_{f(n)\ln n}^{\infty} |B(n,x+\imath y)|\,dy = O(e^{-n^{1/3}(\pi/4-\varepsilon)}n^{-x}).$$

The terms of $g_N(n, x+\imath y)$ are of the type $n^{-k-x-\imath y}\Gamma(x+\imath y+k)(x+\imath y)^l$ $(l \le k)$. Integrating over a term like this yields

$$\begin{aligned}
&\int_{f(n)\ln n}^{\infty} |n^{-k-x-\imath y}\Gamma(x+\imath y+k)(x+\imath y)^k|\,dy \\
&\quad\le 2^k\sqrt{2\pi}n^{-k-x} \\
&\qquad\times\int_{n^{1/3}}^{\infty} y^k|x+k+\imath y|^{x+k-1/2}\exp\left(-y\arg(x+k+\imath y)-x-k+\frac{x+k}{12((x+k)^2+y^2)}\right)dy \\
&\quad\le 2^{x+2k}\sqrt{2\pi}n^{-k-x}\int_{n^{1/3}}^{\infty} y^{2k}e^{-y(\pi/3)}\,dy \le 2^{x+2k}\sqrt{2\pi}n^{-k-x}\left(\frac{3}{\pi}\right)^{2k+1} \\
&\qquad\times\int_{(\pi/3)n^{1/3}}^{\infty} u^{2k}e^{-u}\,du \\
&\quad\le 2^{x+2k}\sqrt{2\pi}n^{-k-x}\left(\frac{3}{\pi}\right)^{2k+1} 2\left(\frac{\pi}{3}n^{1/3}\right)^{2k+1} e^{-(\pi/3)n^{1/3}} \\
&\quad= O(n^{-k/3-x}e^{-(\pi/3)n^{1/3}}).
\end{aligned}$$

Since we have $\arg(x+k+\imath) \ge (\pi/3)\operatorname{sgn}(y)$, and $\int_{(\pi/3)n^{1/3}}^{\infty} u^{2k}e^{-u}\,du \le 2((\pi/3)n^{1/3})^{2k}e^{-(\pi/3)n^{1/3}}$ by iterated integration (for n large enough), hence, the error on the tail $y > n^{1/3}$ is exponentially small. □

We concentrate on the first term of the expansion since each further term is by an order of magnitude smaller than the previous. This is due to the fact, that each new term introduces a factor n^{-1} and possibly a factor z which reduces the order of singularities (in particular those of $\Gamma(z)$). As a result, this leads to

$$\mathfrak{I}_{\xi,n}^{(d)} := \frac{1}{2\pi\imath}\int_{-\xi-\imath\infty}^{-\xi+\imath\infty} \frac{1}{\sigma^{-1-z}-1}\left(\frac{q}{\sigma^{-z-1}-p}\right)^{d+1}\Gamma(z)n^{-z}\,dz, \tag{22}$$

with $\mathfrak{S}_n^{(d)} = \mathfrak{I}_{\xi,n}^{(d)} + O(1)$ for $\xi \in (1,2)$. In Figure 2 we visualize the behavior of the function under the integral in the complex plane. Depending on the parameter $p = 1-q$ the right line (dotted with circles) of residues moves left or right. The real value $-\xi^*$,

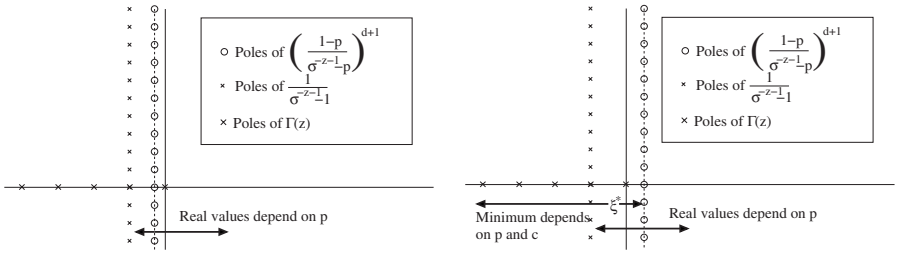

Fig. 2. Behavior of the studied function in the complex plane.

where the absolute value of the function under the integral is minimal, also depends upon c for the case $d+1 = c\log_\sigma n$. If $-\xi^* > -1$ is right of the left line of residues, we move the line of integration over the residues at $\Re(z) = -1$ and get a sublinear behavior. We might need to account for the residue of $\Gamma(z)$ at $z = 0$ if the dotted line is right of $\Re(z) = 0$. Otherwise, if $-\xi^* < -1$, the whole integral is too small to compensate A_n. For small values of d we take all residues into account.

The most important residues are those where the singularity is at a point with real value -1.

LEMMA 10 (Residues at $z = -1 \pm 2\pi\imath k/\ln\sigma$). *Let* $g(z) := (1/(\sigma^{-1-z}-1))(q/(\sigma^{-z-1}-p))^{d+1}\Gamma(z)n^{-z}$:

$$\operatorname{res}[g(z), z=-1] = n\left(\frac{1-\gamma}{\ln\sigma} - \log_\sigma n + \frac{1}{2} + \frac{(d+1)}{q}\right), \tag{23}$$

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} \operatorname{res}\left[g(z), z=-1+\frac{2\pi\imath k}{\ln\sigma}\right] = \frac{1}{\ln\sigma}\sum_{k\in\mathbb{Z}\setminus\{0\}} -\Gamma\left(-1+\frac{2\pi k\imath}{\ln\sigma}\right) n^{1-2\pi k\imath/\ln\sigma} \tag{24}$$

PROOF. The residue can be derived easiest from the series decompositions. These are, at $z = -1 + 2\pi\imath k/\ln\sigma$,

$$\begin{aligned}
\frac{1}{\sigma^{-1-z}-1} &= \sum_{l=-1}^{\infty} \frac{B_{l+1}(-\ln\sigma)^l}{(l+1)!}\left(z+1-\frac{2\pi\imath k}{\ln\sigma}\right)^l \\
&= \frac{-1}{\ln\sigma(z+1-2\pi\imath k/\ln\sigma)} - \frac{1}{2} + O\left(\left(z+1-\frac{2\pi\imath k}{\ln\sigma}\right)\right), \\
\left(\frac{q}{\sigma^{-z-1}-p}\right)^{d+1} &= 1 + \frac{(d+1)\ln\sigma}{q}\left(z+1-\frac{2\pi\imath k}{\ln\sigma}\right) + O\left(\left(z+1-\frac{2\pi\imath k}{\ln\sigma}\right)^2\right), \\
n^{-z} &= n\sum_{l=0}^{\infty} n^{-2\pi\imath k/\ln\sigma}\frac{(-\ln n)^l}{l!}\left(z+1-\frac{2\pi\imath k}{\ln\sigma}\right)^l, \\
\Gamma(z) &= \begin{cases} \sum_{l=-1}^{\infty}\gamma_l^{(-1)}(z+1)^l \\ \quad = -1/(z+1) + (\gamma-1) + O((z+1)), & \text{for } k\neq 0, \\ \sum_{l=0}^{\infty}\gamma_l^{(-1+2\pi\imath k/\ln\sigma)}(z+1-2\pi\imath k/\ln\sigma)^l, & \text{otherwise.} \end{cases}
\end{aligned}$$

The B_k are the Bernoulli numbers (see, e.g., [30] or (34)). □

If we expanded the next term by Theorem 9, we would find that the residues at $\Re(z) = -1$ of $g_1(z) := (1/(\sigma^{-1-z}-1))(q/(\sigma^{-z-1}-p))^{d+1}\Gamma(z+1)((-1-z)/2)n^{-z+1}$ are $O(1)$. As a result, we have

$$-\left(\operatorname{res}[g(z), z=-1] + \sum_{k\in\mathbb{Z}\setminus\{0\}} \operatorname{res}\left[g(z), z=-1+\frac{2\pi\imath k}{\ln\sigma}\right]\right) + n\left(1+\frac{d+1}{q}\right) - A_n = O(1).$$

Moving the line of integration to $-1+\varepsilon$ we get

$$\mathfrak{S}_n^{(d)} = A_n - n\left(1+\frac{d+1}{q}\right) + \mathfrak{I}_{1-\varepsilon,n}^{(d)} + O(1). \tag{25}$$

If we keep the line right of -1 we get

$$\mathfrak{S}_n^{(d)} = \mathfrak{I}_{1+\varepsilon,n}^{(d)} + O(1). \tag{26}$$

By (17), we can bound the integral for some constants c, C and for $d+1 = c\log_\sigma n$ as

$$\mathfrak{I}_{\xi,n}^{(d)} \le \frac{C}{\sigma^{\xi-1}-1} n^{c\log_\sigma (q/(\sigma^{\xi-1}-p))+\xi}.$$

Let $\mathfrak{E}_{c,p,\xi} := c\log_\sigma ((1-p)/(\sigma^{\xi-1}-p)) + \xi$ be the exponent. We can bound the exponent as follows.

LEMMA 11. *For $0 < p < 1$, $c \ge 0$, and $c \ne 1-p$ there exists a $\xi < 2$ such that $\mathfrak{E}_{c,p,\xi} < 1$. If $c < 1$, then $\mathfrak{E}_{c,p,\xi}$ has a minimum at $\xi^* = -\log_\sigma(1-c) + \log_\sigma p + 1$. If $c \ge 1$ or $\xi^* \ge 2$, then some value $\xi \in (1,2)$ satisfies $\mathfrak{E}_{c,p,\xi} < 1$.*

PROOF. The exponent $\mathfrak{E}_{c,p,\xi}$ has at most one extreme value for real ξ at $\xi^* = -\log_\sigma (1-c) + \log_\sigma p + 1$.

Assume $c < 1$ and let $c = 1-\sigma^{-x}$, then the exponent has a minimum at $\xi^* = x + \log_\sigma p + 1$, where it takes the value

$$(1-\sigma^{-x})\left(\log_\sigma\left(\frac{1}{p}-1\right) - \log_\sigma(\sigma^x - 1)\right) + x + 1 + \log_\sigma p.$$

With respect to x there is a single extreme value, a maximum, at $x^* = -\log_\sigma p$, where the exponent takes the value 1. So for all other values of x the exponent is smaller than 1. This takes care of the interval $\xi^* < 2$.

For $\xi^* \ge 2$ we derive that $c > 1 - p/\sigma$ and that the minimum is taken for some $\varepsilon > 0$ at $\xi = 2-\varepsilon$. Since $c\log_\sigma((1-p)/(\sigma^{1-\varepsilon}-p)) < c\log_\sigma 1 = 0$ we find that

$$\mathfrak{E}_{c,p,1-\varepsilon} < \left(1-\frac{p}{\sigma}\right)\log_\sigma\left(\frac{1-p}{\sigma^{1-\varepsilon}-p}\right) + 2 - \varepsilon.$$

The derivative of $(1-p/\sigma)\log_\sigma((1-p)/(\sigma^{1-\varepsilon}-p)) + 2 - \varepsilon$ with respect to p is

$$\frac{-1}{\sigma}\log_\sigma\left(\frac{1-p}{\sigma^{1-\varepsilon}-p}\right) + \left(1-\frac{p}{\sigma}\right)\left(-\frac{\sigma^{1-\varepsilon}-1}{(\sigma^{1-\varepsilon}-p)(1-p)\ln\sigma}\right).$$

The second derivative is

$$\frac{2}{\sigma}\left(\frac{\sigma^{1-\varepsilon}-1}{(\sigma^{1-\varepsilon}-p)(1-p)\ln\sigma}\right) - \left(1-\frac{p}{\sigma}\right)\left(\frac{(\sigma^{1-\varepsilon}-1)((1-p)+(\sigma^{1-\varepsilon}-p))}{(\sigma^{1-\varepsilon}-p)^2(1-p)^2\ln\sigma}\right),$$

which is smaller than 0 for

$$p < \frac{\sigma^{2-\varepsilon}+\sigma-2\sigma^{1-\varepsilon}}{1+2\sigma-\sigma^{1-\varepsilon}} \xrightarrow[\varepsilon\to 0]{} \sigma.$$

Therefore, the first derivative is decreasing for $p < \sigma - \varepsilon$, it is maximal at $p = 0$. Here we have a value of

$$\frac{(1-\varepsilon)\ln\sigma - \sigma + \sigma^{\varepsilon}}{\sigma \ln\sigma} \xrightarrow[\varepsilon\to 0]{} \frac{\ln\sigma - \sigma + 1}{\sigma\ln\sigma} < 0 \qquad \text{for} \quad \sigma \geq 2.$$

The term is decreasing in p and the maximal value is attained at $p = 0$, where we have a value of 1. As a result we have for some small ε,

$$\mathfrak{E}_{c,p,1-\varepsilon} < 1.$$

Finally, for $c \geq 1$ the derivative of $\mathfrak{E}_{c,p,\xi}$ is

$$\frac{-c\sigma^{\xi-1}}{\sigma^{\xi-1} - p} + 1 < \frac{-\sigma^{\xi-1}}{\sigma^{\xi-1} - p} + 1 \leq 0, \qquad \text{for} \quad \xi > 1 + \log_{\sigma} p.$$

Thus, we again have a minimum at $\xi = 2 - \varepsilon$. We find that

$$\mathfrak{E}_{c,p,1-\varepsilon} \leq \log_{\sigma}\left(\frac{1-p}{\sigma^{1-\varepsilon} - p}\right) + 2 - \varepsilon,$$

where the right term's derivative with respect to p is $-(\sigma^{1-\varepsilon}-1)/(\sigma^{1-\varepsilon}-p)(1-p)\ln\sigma < 0$. This, by the same arguments as above, shows that $\mathfrak{E}_{c,p,1-\varepsilon} < 1$ for $p > 0$. □

We can now prove Theorem 4. If $\xi^* > 1$ or $c \geq 1$ we find a $\xi \in (1, 2)$ such that $\mathfrak{E}_{c,p,\xi} < 1$, thus $\mathfrak{I}^{(d)}_{\xi,n} = o(n)$ and $\mathfrak{S}^{(d)}_{n} = o(n)$ by (26).

If $\xi^* < 1$, we move the line of integration to $-\xi^* = -1+\varepsilon$ and find that $\mathfrak{I}^{(d)}_{1-\varepsilon,n} = o(n)$, and by (25) we have $\mathfrak{S}^{(d)}_{n} = A_n - n(1 + (d+1)/q) + o(n)$. A special situation occurs only for $\xi^* < 0$ because we have to take the singularity of the Gamma function at $z = 0$ into account (the singularities at $z = -\log_{\sigma} p - 1 \pm 2\pi\imath k/\ln\sigma$ are always to the right of ξ^*).

LEMMA 12 (Singularity at $z = 0$ for $\xi^* < 0$). *For $\xi^* < 0$ we have*

$$-\operatorname{res}[g(z), z{=}0] = \frac{\sigma}{\sigma - 1}\left(\frac{q\sigma}{1 - p\sigma}\right)^{d+1} \qquad (= o(n),\ \textit{for } d{+}1{=}c\log_{\sigma} n). \tag{27}$$

PROOF. The singularities at $z = -\log_{\sigma} p - 1 \pm 2\pi\imath k/\ln\sigma$ are always to the right of ξ^* since $-\log_{\sigma}(1-c)$ is positive. The only other singularity is the singularity at $z = 0$. Under the assumption $\xi^* < 0$ we have $c < 1 - p\sigma$ (thus $p < \sigma^{-1}$) and

$$\begin{aligned} -\operatorname{res}[g(z), z = 0] \;&=\; \frac{\sigma}{\sigma - 1}\left(\frac{q\sigma}{1 - p\sigma}\right)^{d+1} = \frac{\sigma}{\sigma - 1} n^{c\log_{\sigma}((1-p)/(\sigma^{-1}-p))} \\ &<\; \frac{\sigma}{\sigma - 1} n^{(1-p\sigma)\log_{\sigma}((1-p)/(\sigma^{-1}-p))}. \end{aligned} \tag{28}$$

The exponent $(1 - p\sigma)\log_{\sigma}((1-p)/\sigma^{-1} - p))$ has derivative (with respect to p)

$$-\sigma\log_{\sigma}\left(\frac{1-p}{\sigma^{-1} - p}\right) + \sigma\frac{(1-\sigma^{-1})}{(1-p)\ln\sigma} < 0,$$

since $\log_\sigma((1-p)/(\sigma^{-1}-p)) > 1$ and $(1-\sigma^{-1})/(1-p)\ln\sigma < 1$. The exponent has value 1 at $p = 0$, hence it is smaller than 1 for $p > 0$. Thus, the singularities contribution is $o(n)$, which does not affect the result. □

Thus, the complexity has two cases depending on whether ξ^* is left or right of 1. This translates to $\xi^* < 1$ if and only if $c < 1 - p = q$, which proves Theorem 4.

REMARK 13 (Explicit Computation of the Exponent). It is possible to determine concrete values for the exponent. The interesting case is $0 < c < 1 - p$, where we have a sublinear behavior of the search time. By either Lemma 11 or Lemma 12, we get a minimal value of the exponent at

$$\mathfrak{E}_{c,p,\min(\xi^*,0)} = \begin{cases} \log_\sigma((1-c)^{-c+1}c^{-c}\sigma p^{1-c}(1-p)^c), & \text{for } 1-p\sigma \le c < 1-p, \\ \log_\sigma((1-p)^c\sigma^c(1-p\sigma)^{-c}), & \text{for } 0 \le c < 1-p\sigma. \end{cases} \tag{29}$$

For $p = \sigma^{-1}$ we find that the exponent is $H(c)/\ln\sigma + c\log_\sigma(\sigma-1)$, where $H(x) = -x\ln(x) - (1-x)\ln(1-x)$ is the entropy function. The exponent is 0 at $c = 0$ and 1 at $c = 1 - p$. For $\sigma = 2$ we have a scaled entropy function. With growing σ, the linear part dominates more and more while the entropy part diminishes, i.e., the behavior of the exponent converges to a linear function. As a result, we can estimate that the running time can be bounded by a function between $O(n^{c(\sigma/(\sigma-1))})$ and $O(n^{H((c/2)(\sigma/(\sigma-1)))/\ln 2})$.

REMARK 14 (Relation between the Error Probability and the Number of Errors). For the LS algorithm we can double the number of allowed errors and the error probability and get the same asymptotic running time. For the TS algorithm, in the parameter range with sublinear behavior, this is different. Here, a higher variance in the depth that a string is matched by the pattern increases the expected running time. This behavior is explained by the difference in savings or expenses of an early or late end in matching a pattern against a string. Because of the trie structure, characters at the beginning of the strings are subsumed while characters at the end are not. Thus, not comparing some characters early in the string in exchange for comparing some characters more towards the end of another string is disadvantageous. The savings in the earlier characters correspond to less nodes than the extra cost in the characters towards the end. In our formula for the exponent given by (29), this can be seen when replacing c by ct and q by qt:

$$\mathfrak{E}_{ct,1-tq,\min(\xi^*,0)} = \begin{cases} \log_\sigma\left(\dfrac{\sigma(1-tq)^{1-tc}(tq)^{tc}}{(1-tc)^{1-tc}(tc)^{tc}}\right), & \text{for } 1-(1-tq)\sigma \le c < tq, \\ tc\log_\sigma\left(\dfrac{tq\sigma}{1-(1-tq)\sigma}\right), & \text{for } 0 \le c < 1-(1-tq)\sigma. \end{cases} \tag{30}$$

The derivative (with respect to t) of the above is always negative when t, c, and q result in valid parameters: For the second line we have

$$\underbrace{c\log_\sigma\left(\frac{tq\sigma}{1-\sigma+tq\sigma}\right)}_{\geq 0}+\underbrace{\frac{tc}{\ln\sigma}}_{\geq 0}\underbrace{\frac{1-\sigma+tq\sigma}{tq\sigma}}_{\geq 1}\underbrace{\frac{\overbrace{q\sigma}^{\geq 0}\overbrace{(1-\sigma)}^{<0}}{(1-\sigma+tq\sigma)^2}}_{\geq 0}<0.$$

For the first case we find that the derivative is

$$c\log_\sigma\left(\frac{q-tqc}{c-tqc}\right)+\frac{c-q}{(1-tq)\ln\sigma},$$

which is negative if

$$c\log_\sigma(q-tqc)-\frac{q}{(1-tq)\ln\sigma}<c\log_\sigma(c-tqc)-\frac{c}{(1-tq)\ln\sigma}.$$

However, the function $x\to c\log_\sigma(x-tqc)-x/(1-tq)\ln\sigma$ decreases for $c\leq x\leq q$ because the derivative with respect to x is

$$\frac{c}{(x-tqc)\ln\sigma}-\frac{1}{(1-tq)\ln\sigma},$$

which is negative or zero because $c\leq x$ and $1-tq>0$.

To prove Theorem 3 we calculate the remaining residues at $z=0$ and $z=-\log_\sigma p-1\pm 2\pi\imath k/\ln\sigma$.

LEMMA 15 (Residues at $z=0$, $z=-\log_\sigma p-1+2\pi\imath k/\ln\sigma$). *Let* $g(z):=(1/(\sigma^{-1-z}-1))(q/(\sigma^{-z-1}-p))^{d+1}\Gamma(z)n^{-z}$. *If* $p=\sigma^{-1}$ *we have*

$$\begin{aligned}(31)\quad \operatorname{res}[g(z),z=0]\;=\;&\sum_{l=0}^{d+1}\frac{(\sigma-1)^{d+1}(-1)^{-l}B_{-l+d+1}^{(d+1)}}{(-l+d+1)!}\\&\times\sum_{i=0}^{l}\left[\sum_{j=0}^{i}\frac{A_j(\sigma^{-1})}{j!\,(i-j)!}\left(\frac{\sigma}{1-\sigma}\right)^{j+1}\left(-\frac{\ln n}{\ln\sigma}\right)^{i-j}\right]\\&\times\frac{\gamma_{l-i-1}^{(0)}}{(\ln\sigma)^{l-i}},\end{aligned}$$

otherwise, *the residue is given by* (28).

For $k \in \mathbb{Z}$ and $k \neq 0$ or $p \neq \sigma^{-1}$ we have

$$\begin{aligned}(32)\qquad \operatorname{res}\left[g(z), z = -\log_\sigma p - 1 + \frac{2\pi\imath k}{\ln\sigma}\right] &= \sum_{l=0}^{d}\left(\frac{1-p}{p}\right)^{d+1}\frac{(-\ln\sigma)^{-l}B_{-l+d}^{(d+1)}}{(-l+d)!}\\ &\quad\times\sum_{i=0}^{l}\left[\sum_{j=0}^{i}\frac{-A_j(p)(-\ln\sigma)^j}{(1-p)^{j+1}j!}n^{\log_\sigma p+1-2\pi\imath k/\ln\sigma}\frac{(-\ln n)^{i-j}}{(i-j)!}\right]\\ &\quad\times(\gamma_{l-i}^{(-\log_\sigma p-1+2\pi\imath k/\ln\sigma)}).\end{aligned}$$

Here $B_k^{(a)}$ are generalized Bernoulli numbers, $A_l(x)$ are Eulerian polynomials, and $\gamma_l^{(z_0)}$ are the coefficients of the series for $\Gamma(z)$ at $z = z_0$.

PROOF. We compute the residues at $z = 0$ and $z = -\log_\sigma p - 1 + 2\pi\imath k/\ln\sigma$. If $p \neq \sigma^{-1}$, then the residue at $z = 0$ is given by (28). Otherwise, we have a higher-order singularity at $z = 0$, i.e., we need to use a different series expansion for the Gamma function. We need the series representations for the factors of

$$g(z) = \frac{1}{\sigma^{-1-z}-1}\left(\frac{q}{\sigma^{-z-1}-p}\right)^{d+1}\Gamma(z)n^{-z} \qquad \text{at} \quad z = -\log_\sigma p - 1 + \frac{2\pi\imath k}{\ln\sigma}.$$

These can be derived best in terms of Eulerian polynomials $A_n(u)$ defined by

$$(33)\qquad \frac{1-u}{1-ue^{t(1-u)}} = \sum_{n=0}^{\infty}A_n(u)\frac{t^n}{n!} = \sum_{n=0}^{\infty}\frac{t^n}{n!}\sum_{k=0}^{n}A_{n,k}u^k, \qquad |t| < \frac{\ln u}{u-1}$$

(see [8] or [13] for Eulerian numbers $A_{n,k}$). We also need generalized Bernoulli numbers $B_k^{(a)}$, defined by

$$(34)\qquad \left(\frac{t}{e^t-1}\right)^a = \sum_{k=0}^{\infty}B_k^{(a)}\frac{t^k}{k!}, \qquad |t| < 2\pi$$

(see [30] or [22]). We then have the following series representations:

$$\begin{aligned}\frac{1}{\sigma^{-1-z}-1} &= \sum_{l=0}^{\infty}\frac{-A_l(p)(-\ln\sigma)^l}{(1-p)^{l+1}l!}\left(z+\log_\sigma p+1-\frac{2\pi\imath k}{\ln\sigma}\right)^l,\\ \left(\frac{q}{\sigma^{-z-1}-p}\right)^{d+1} &= \sum_{l=-d-1}^{\infty}\left(\frac{1-p}{p}\right)^{d+1}\frac{(-\ln\sigma)^l B_{l+d+1}^{(d+1)}}{(l+d+1)!}\left(z+\log_\sigma p+1-\frac{2\pi\imath k}{\ln\sigma}\right)^l,\\ n^{-z} &= \sum_{l=0}^{\infty}n^{\log_\sigma p+1}n^{-\frac{2\pi\imath k}{\ln\sigma}}\frac{(-\ln n)^l}{l!}\left(z+\log_\sigma p+1-\frac{2\pi\imath k}{\ln\sigma}\right)^l,\end{aligned}$$

$$\Gamma(z) = \begin{cases} \sum_{l=-1}^{\infty} \gamma_l^{(0)} z^l, & \text{for} \quad p = \sigma^{-1} \\ & \text{and} \quad k = 0, \\ \sum_{l=-1}^{\infty} \gamma_l^{(-\log_\sigma p - 1 + 2\pi\imath k/\ln\sigma)} & \\ \quad \times (z + \log_\sigma p + 1 - 2\pi\imath k/\ln\sigma)^l, & \text{otherwise.} \end{cases}$$

One can show that $|\gamma_l^{(0)} - (-1)^l| < (\frac{1}{2} + \varepsilon)^l$. The $\gamma_l^{(-\log_\sigma p - 1 + 2\pi\imath k/\ln\sigma)}$ are just the result of a simple Taylor expansion. The relevant singularities are of order $d+1$ (or $d+2$ for $p = \sigma^{-1}$). The residue is the sum of all combinations of coefficients such that the power of $\left(z + \log_\sigma p + 1 - 2\pi\imath k/\ln\sigma\right)$ is -1. The series lead directly to the residues. □

We consider d, p, q, σ constant, so we look for the largest term in n. If $p = \sigma^{-1}$, this term is

$$-\frac{\sigma(\sigma-1)^d}{(d+1)!}\left(\log_\sigma n\right)^{d+1}, \tag{35}$$

otherwise, this term is

$$-\frac{(1-p)^d}{d!\; p^{d+1}}\left(\log_\sigma n\right)^d n^{\log_\sigma p + 1} \sum_{k\in\mathbb{Z}} n^{-2\pi\imath k/\ln\sigma}\,\Gamma\left(-\log_\sigma p - 1 + \frac{2\pi\imath k}{\ln\sigma}\right). \tag{36}$$

For $\log_\sigma p + 1 < 0$ this is $o(1)$. In this case the residue at $z = 0$ yields $O(1)$, see (27). Note also that for real values $\Gamma(x)$ has different signs left and right of $x = 0$. For the calculation of $\mathfrak{I}_{\xi,n}^{(d)}$ we sum up the negative of the residues. There are infinitely many residues, but due to the behavior of the Gamma function for large imaginary values we have

$$\left|\sum_{k\in\mathbb{Z}} n^{-2\pi\imath k/\ln\sigma}\,\Gamma\left(-\log_\sigma p - 1 + \frac{2\pi\imath k}{\ln\sigma}\right)\right| = O(1). \tag{37}$$

Hence, the growth for constant d is

$$\mathfrak{S}_n^{(d)} = A_n - n\left(1 + \frac{d+1}{q}\right) + \begin{cases} O((\log n)^{d+1}), & \text{for} \quad p = \sigma^{-1}, \\ O((\log_\sigma n)^d n^{\log_\sigma p + 1}), & \text{for} \quad p > \sigma^{-1}, \\ O(1), & \text{otherwise.} \end{cases} \tag{38}$$

Thus, we have proven Theorem 3.

References

[1] A. Apostolico and W. Szpankowski. Self-alignments in words and their applications. *J. Algorithms*, 13:446–467, 1992.

[2] R. A. Baeza-Yates and G. H. Gonnet. Fast text searching for regular expressions or automaton searching on tries. *J. ACM*, 43(6):915–936, 1996.

[3] R. A. Baeza-Yates and G. H. Gonnet. A fast algorithm on average for all-against-all sequence matching. In *Proc. 6th Int. Symp. on String Processing and Information Retrieval* (*SPIRE*), pages 16–23. IEEE, New York, 1999.

[4] R. D. L. Briandais. File searching using variable length keys. In *Proc. Western Joint Computer Conference*, pages 295–298, March 1959.

[5] A. Buchner and H. Täubig. A fast method for motif detection and searching in a protein structure database. Technical Report TUM-I0314, Fakultät für Informatik, TU München, Sept. 2003.

[6] A. Buchner, H. Täubig, and J. Griebsch. A fast method for motif detection and searching in a protein structure database. In *Proc. German Conference on Bioinformatics* (*GCB*), volume 2, pages 186–188, Oct. 2003.
[7] A. L. Cobbs. Fast approximate matching using suffix trees. In *Proc. 6th Symp. on Combinatorial Pattern Matching* (*CPM*), pages 41–54. Volume 937 of Lecture Notes in Computer Science. Springer, Berlin, 1995.
[8] L. Comtet. *Advanced Combinatorics*. Reidel, Dordrecht, 1974.
[9] J. L. Fields. The uniform asymptotic expansion of a ratio of Gamma functions. In *Proc. Int. Conf. on Constructive Function Theory*, pages 171–176, Varna, May 1970.
[10] P. Flajolet and C. Puech. Partial match retrieval of multidimensional data. *J. ACM*, 33(2):371–407, 1986.
[11] P. Flajolet and R. Sedgewick. Mellin transforms and asymptotics: finite differences and Rice's integral. *Theoret. Comput. Sci.*, 144(1–2):101–124, 1995.
[12] E. Fredkin. Trie memory. *Comm. ACM*, 3(9):490–499, 1960.
[13] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*, 2nd edition. Addison-Wesley, Reading, MA, 1994.
[14] P. Jokinen and E. Ukkonen. Two algorithms for approximate string matching in static texts. In *Proc. 16th Int. Symp. on Mathematical Foundations of Computer Science* (*MFCS*), pages 240–248. Volume 520 of Lecture Notes in Computer Science. Springer, Berlin, 1991.
[15] J. F. C. Kingman. Subadditive ergodic theory. *Ann. Probab.*, 1(6):883–909, 1973.
[16] P. Kirschenhofer. A note on alternating sums. *Electron. J. Combin.*, 3(2), R7, 1996.
[17] D. E. Knuth. *The Art of Computer Programming—Sorting and Searching*, volume 3, 2nd edition. Addison-Wesley, Reading, MA, Feb. 1998.
[18] M. G. Maaß. Average-case analysis of approximate trie search. In *Proc. 15th Symp. on Combinatorial Pattern Matching* (*CPM*), pages 472–484. Volume 3109 of Lecture Notes in Computer Science. Springer, Berlin, July 2004.
[19] D. R. Morrison. PATRICIA—practical algorithm to retrieve information coded in alphanumeric. *J. ACM*, 15(4):514–534, Oct. 1968.
[20] G. Navarro. Approximate Text Searching. Ph.D. thesis, Dept. of Computer Science, University of Chile, Santiago, Chile, 1998.
[21] G. Navarro and R. Baeza-Yates. A hybrid indexing method for approximate string matching. *J. Discrete Algorithms*, 1(1):205–209, 2000. Special issue on Matching Patterns.
[22] N. E. Nörlund. *Vorlesungen über Differenzenrechnung*. Springer, Berlin, 1924.
[23] K. Oflazer. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Comput. Linguist.*, 22(1):73–89, 1996.
[24] B. Pittel. Paths in a random digital tree: limiting distributions. *Adv. in Appl. Probab.*, 18:139–155, 1986.
[25] H. Prodinger and W. Szpankowski (Guest Editors). *Theoret. Comput. Sci.*, 144(1–2) (special issue), 1995.
[26] K. U. Schulz and S. Mihov. Fast string correction with Levenshtein automata. *Int. J. Doc. Anal. Recog.* (*IJDAR*), 5:67–85, 2002.
[27] W. Szpankowski. The evaluation of an alternative sum with applications to the analysis of some data structures. *Inform. Process. Lett.* (*IPL*), 28:13–19, 1988.
[28] W. Szpankowski. Some results on v-ary asymmetric tries. *J. Algorithms*, 9:224–244, 1988.
[29] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*, 1st edition. Wiley-Interscience, New York, 2000.
[30] N. M. Temme. *An Introduction to Classical Functions of Mathematical Physics*. Wiley, New York, 1996.
[31] F. G. Tricomi and A. Erdélyi. The asymptotic expansion of a ratio of Gamma functions. *Pacific J. Math.*, 1:133–142, 1951.
[32] E. Ukkonen. Approximate string-matching over suffix trees. In *Proc. 4th Symp. on Combinatorial Pattern Matching* (*CPM*), pages 228–242. Volume 684 of Lecture Notes in Computer Science. Springer, Berlin, 1993.
[33] F. Werner, G. Durstewitz, F. Habermann, G. Thaller, W. Krämer, S. Kollers, J. Buitkamp, M. Georges, G. Brem, J. Mosner, and R. Fries. Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Animal Genetics*, 35(1):44–49, Feb. 2004.