T U M

I N S T I T U T F Ü R I N F O R M A T I K

Average-Case Analysis of Approximate Trie Search

Moritz G. Maaß

TUM-I0405
März 04

T E C H N I S C H E U N I V E R S I T Ä T M Ü N C H E N

TUM-INFO-03-I0405-0/1.-FI

Druck: Institut für Informatik der
Technischen Universität München

Average-Case Analysis of Approximate Trie Search

Moritz G. Maaß*

maass@informatik.tu-muenchen.de

Institut für Informatik, Technische Universität München
Boltzmannstr. 3, D-85748 Garching, Germany

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 errors are allowed between the pattern and the database strings, no such structure with reasonable size is known. Using a trie some work can be saved and running times superior to the comparison with every string in the database can be achieved. We investigate a comparison-based model where "errors" and "matches" are defined between pairs of characters. When comparing two characters, let p 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 p 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 p and D. Our results can be applied for any comparison-based error model, for instance, mismatches (Hamming distance), don't cares, or geometric character distances.

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 have been preprocessed with linear space, and we are also interested in occurrences where the pattern is found with a given number of errors.

This research was triggered by a project for the control of animal breeding via SNPs [31]. Without going into details, each SNP can be encoded as a binary string that is an identifier for an individual. Because of errors the search in the dataset of all individuals needs to be able to deal with mismatches and don't cares. The nature of the data allows a very efficient binary encoding, which yields a reasonable fast algorithm that just compares a pattern to each string in the set. On the other hand, a trie can be used as an index for the data. It 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. We call the first variant "Linear Search" (LS) and the second variant "Trie Search" (TS).

*Research supported in part by DFG, grant Ma 870/5-1 (Leibnizpreis Ernst W. Mayr).

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, an error probability p, and a bound D for the maximal number of errors allowed, let L_n^D be the number of comparisons made by the LS algorithm and T_n^D be the number of comparisons made by the TS algorithm using a trie as an index. The parameters p depends on the definition of errors, it gives the relative number of character pairs inducing an error (or mismatch).What is the threshold $D = D(n)$, up to where the average $\mathbf{E}\left[T_n^D\right]$ is asymptotically better than the average $\mathbf{E}\left[L_n^D\right]$? What is the effect of different error 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 $t = t_1 t_2 t_3 \ldots t_n$ 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.

When two arbitrary, randomly generated characters are compared, let p denote the probability of an error and $q := 1 - p$ the probability of a match. For example, we have a mismatch probability of $p = 1 - 1/\sigma$ and a match probability of $q = 1/\sigma$ for Hamming distance.

It is easy to prove that the average number of comparisons made by the LS algorithm is $\mathbf{E}\left[L_n^D\right] = (D+1)n/p$. 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.

The 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 $D \geq (1+\epsilon)\log_\sigma n$ there can be no asymptotic gain when using a trie.

We will show that the TS algorithm performs sublinear for $D < p \log_\sigma n$ and superlinear for $D > p \log_\sigma n$. When D is a constant the asymptotic running time can be calculated very exactly and is $O\left((\log n)^{D+1}\right)$, for $q = 1/\sigma$, $O\left((\log n)^D n^{\log_\sigma q + 1}\right)$, for $q > 1/\sigma$, 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 a lot of attention. Tries were introduced by Brandais [4] and Fredkin [11]. A useful extension are Patricia trees introduced by Morrison [17].

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 [24] 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 small number of errors is allowed. A similar approach has been investigated by Oflazer [21], except that he directly calculates 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 $k - s$ don't cares and s specified values 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 $O\left(n^{1-s/k}\right)$ 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 [19] 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 Navarro [18] the bound is investigated more closely. It is conjectured that the real threshold, where the number of matches of a pattern in a text grows subexponentially 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 [13] (on a DAWG), Ukkonen [30], 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., [23]). 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 [16]. See [27] 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., [16]). Similar sums, where $f(n,k)$ only depends on k, have also been considered by Szpankowski [25] and Kirschenhofer [15]. The asymptotic analysis can be done through Rice's integrals (a technique that already appears in Nörlund [20], chap. 8, §1). It transfers the sum to a complex integral, which is evaluated by the Cauchy residue theorem.

Our contribution is the general analysis of error models that depend only one 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 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. Pseudo code for the analyzed algorithms is given in Figure 1.

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 $c \leq D$ **do**
      while $j \leq l$ **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 $D \geq 0$ **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
          $\mathsf{rfind}(u, P, pos + 1, D - 1)$

Figure 1: Pseudo code of the LS and the TS algorithm. The recursive TS algorithm is started with $\mathsf{rfind}(r, P, 0, D)$, where r is the root of the trie for the Strings $X_1, \ldots, X_n$.

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).

Theorem 1 (Average Complexity of the TS algorithm).

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

Exacter bounds are possible through equations (27), (33), and (34) for the cases $q < \sigma^{-1}$, $q = \sigma^{-1}$, and $q > \sigma^{-1}$. For instance, equation (27) tells us that the number of nodes visited grows by $\frac{p\sigma}{1-q\sigma}$ for each additional error allowed. 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 $\frac{4\cdot 3^D}{(D+1)!}(\log_4 n)^{D+1}$.

It is well known that the average depth of a trie is asymptotically equal to $\log_\sigma n$ (see, e.g., [22, 26]). When no more branching takes place the TS and LS algorithm 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 similar. With an error probability of p we expect to make pm errors on m characters. Thus, it comes as no surprise that the threshold is $p\log_\sigma n$.

With respect to the matching probability q we have a different behavior for the three cases $q < \sigma^{-1}$, $q = \sigma^{-1}$, and $q > \sigma^{-1}$. To explain this phenomena we take a look at the conditional probability of a match for an already chosen character. If $q < \sigma^{-1}$, 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 $q > \sigma^{-1}$, 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 $p = 2/\sigma - 1/\sigma^2$ and $q = 1-2/\sigma+1/\sigma^2$. 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 $O\left((\log n)^D n^{0.585}\right)$ 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 $O\left((\log n)^D n^{0.292}\right)$ when allowing D mismatches. This is significantly worse than Hamming distance only, which is $O\left((\log n)^{D+1}\right)$. 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 $D > (5/8)\log_4 n$ 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 degrees 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 $O\left(n^{\log_\sigma(2i+1)}\right)$ if i intervals to the left or right are considered. Although a suffix tree is used and the underlying distribution of angles is probably not 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

$$\Pr\left\{L_n^D = k\right\} = \sum_{i_1+\ldots+i_n=k} \prod_{j=1}^{n} \binom{i_j-1}{D} p^{D+1} q^{i_j-D-1} \quad . \tag{3}$$

From it we can derive the probability generating function

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

which yields the expected value $\mathbf{E}\left[L_n^D\right] = \frac{D+1}{p}n$. The stochastic process is very stable. We can uses Chebyshevs inequality to derive convergence in probability of L_n^D.

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

Hence, we have

$$\lim_{n\to\infty} \frac{L_n^D}{n(D+1)} = \frac{1}{p} \qquad \text{(pr.)} \ .$$

Note that if $D = D(n)$ is a function of n than we already have almost sure convergence if $D(n) = \omega(\log^{1+\epsilon} n)$ by a simply application of the Borel-Cantelli Lemma and the fact that $\sum_n 1/(n\log^{1+\epsilon} n)$ is convergent. Using a method of Kesten and Kingman (see, for instance, Szpankowski [27] or Kingman [14]) 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, $\frac{L_{r^s}^D}{r^s(D+1)}$ converges to $\frac{1}{p}$ almost sure, since for all $s \ge 1+\epsilon$ we have

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

and thus

$$\lim_{r\to\infty} \frac{L^D_{r^s}}{r^s(D+1)} = \frac{1}{p} \qquad \text{(a.s.)}$$

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

$$\lim_{n\to\infty} \frac{L^D_{n}}{n(D+1)} = \frac{1}{p} \qquad \text{(a.s.)}$$

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 by one larger 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 q) or may decrease (with probability p). For the average number we can set up the following equation.

$$\mathbf{E}\left[T^D_n\right] = 1 + \sum_{i_1+\cdots+i_\sigma=n} \binom{n}{i_1,\ldots,i_\sigma} \sigma^{-n} \left(\sum_{j=1}^{\sigma} p\mathbf{E}\left[T^{D-1}_{i_j}\right] + \sum_{j=1}^{\sigma} q\mathbf{E}\left[T^{D}_{i_j}\right] \right) \quad . \tag{5}$$

The boundary conditions are $\mathbf{E}\left[T^{-1}_n\right] = 1$, counting the character comparison that induced the last mismatch, and $\mathbf{E}\left[T^D_0\right] = 0$. For $n = 1$ we have $\mathbf{E}\left[T^D_1\right] = 1 + \frac{D+1}{p}$, which is the same as $\mathbf{E}\left[L^D_1\right]$, except that additionally the root is counted.

From equation (5) we can derive the exponential generating function of $\mathbf{E}\left[T^D_n\right]$.

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

We multiply with $\exp(-z)$ (which corresponds to applying some kind of binomial inversion) and define $\tilde{t}^D(z) = t^D(z)e^{-z}$. We have

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

Let y^D_n be the coefficients of $\tilde{t}^D(z)$. The we have

$$y^D_n = (-1)^n \sum_{k=0}^{n} \binom{n}{k} (-1)^k \mathbf{E}\left[T^D_k\right] \quad \text{and} \quad \mathbf{E}\left[T^D_n\right] = \sum_{k=0}^{n} \binom{n}{k} y^D_k \ .$$

We get the boundary conditions $y^D_1 = 1 + (D+1)/p$, $y^D_0 = 0$, $y^{-1}_n = (-1)^{n-1}$ for $n > 0$, and $y^{-1}_0 = 0$. Comparing coefficients in equation (7) we find that for $n > 1$

$$y^D_n = \frac{(-1)^{n-1} + y^{D-1}_n \sigma^{1-n} p}{1 - \sigma^{1-n} q} \ , \tag{8}$$

which by iteration leads to

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

Finally, we translate this back to

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

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

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

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, $n(D+1)/p - A_n$ is an upper bound for the average performance of the TS algorithm. To examine the average-case let us recall Rice's integrals (see for instance Nörlund [20], chap. 8, §1).

Theorem 2 (Rice's Formula). *Let $f(z)$ be an analytic continuation of $f(k) = f_k$ that contains the half line $[m,\infty)$. Then*

$$\sum_{k=m}^{n}(-1)^k\binom{n}{k}f_k = \frac{(-1)^n}{2\pi\imath}\int_{\mathcal{C}} f(z)\frac{n!}{z(z-1)\cdots(z-n)}dz \ , \quad (12)$$

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 $f(z)$.

See the liturature [16, 27, 25, 15] for details on this method (called Gamma-method by Knuth). The proof of the theorem stems from the Cauchy residue theorem. We use this transformation to examine the average-case behavior.

Lemma 3 (Asymptotic Behavior of the Compression Number). *The asymptotic behavior of A_n is*

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

Proof. Recall that A_n is defined as

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

with $A_0 = 0$ and $A_1 = 0$. The exponential generating function is

$$A(z) = ze^z - e^z + \sum_{i=1}^{\sigma}A\left(\frac{z}{\sigma}\right)e^{\left(1-\frac{1}{\sigma}\right)z} + 1 \ .$$

Multiplying by e^{-z} we get (with $\tilde{A}(z) = A(z)e^{-z}$)

$$\tilde{A}(z) = z - 1 + \sum_{i=1}^{\sigma} \tilde{A}\left(\frac{z}{\sigma}\right) + e^{-z} \ .$$

Hence, $\tilde{A}_0 = 0$ and $\tilde{A}_1 = 0$. For $n > 1$ this, by comparing coefficients, yields

$$\tilde{A}_n = \frac{(-1)^n}{1 - \sigma^{1-n}} \ .$$

Translating back we get

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

Using Rice's theorem (see Theorem 2) we show that

$$A_n = \frac{(-1)^n}{2\pi\imath} \int_{\mathcal{C}} \frac{1}{1 - \sigma^{1-z}} \frac{n!}{z(z-1)\cdots(z-n)} dz \ .$$

The growth of $\frac{1}{1-\sigma^{1-z}}$ for $z \to \infty$ is $O(1)$, whereas $\frac{n!}{z(z-1)\cdots(z-n)} = O\left(z^{-n-1}\right)$. Let $\mathcal{C}$ be a circle of radius M, carefully chosen as to avoid any singularities. Then $\frac{1}{1-\sigma^{1-z}} \leq C$ on the circle and we have

$$\left| (-1)^n MCn! \int_0^1 \frac{1}{M\exp(2\pi\imath t)(M\exp(2\pi\imath t) - 1)\cdots(M\exp(2\pi\imath t) - n)} dt \right| \leq MCn!(M-n)^{-n} \xrightarrow[M\to\infty]{} 0 \ .$$

Thus, we extend $\mathcal{C}$ to an infinite radius and find that A_n equals the negative sum of the residues of $\frac{1}{1-\sigma^{1-z}} \frac{(-1)^n n!}{z(z-1)\cdots(z-n)} = \mathrm{B}(n+1, -z)$ at $z = 1$, $z = 0$ and $z = 1 + \frac{2\pi\imath k}{\ln\sigma}$, $k \in \mathbb{Z} \setminus \{0\}$. These residues are separated by the line $\Re(z) = 3/2$ from the residues at $\{2, \ldots, n\}$. Integrating along this line yields the value of the residues:

$$A_n = \frac{1}{2\pi\imath} \int_{\frac{3}{2}-\imath\infty}^{\frac{3}{2}+\imath\infty} \frac{1}{1 - \sigma^{1-z}} \mathrm{B}(n+1, -z) dz \ .$$

By the same argument as in Theorem 9, we can approximate the Beta function by $\Gamma(-z)n^z$, thus, we can calculate the residues of $\frac{1}{1-\sigma^{1-z}}\Gamma(-z)n^z$, which are

$$\operatorname{res}\left[\frac{1}{1 - \sigma^{1-z}} \Gamma(-z) n^z, z = 0\right] = -\frac{1}{\sigma - 1} \ ,$$

$$\operatorname{res}\left[\frac{1}{1 - \sigma^{1-z}} \Gamma(-z) n^z, z = 1\right] = -\frac{n}{2} - \frac{\gamma - 1}{\ln\sigma} n - n\log_\sigma n \ ,$$

and

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

The second term in the approximation of the Beta function is $\frac{1}{1-\sigma^{1-z}}\Gamma(-z+1)n^{z-1}\frac{1-z}{2}$. Here we have the residues

$$\operatorname{res}\left[\frac{1}{1-\sigma^{1-z}}\Gamma(-z+1)n^{z-1}\frac{1-z}{2}, z=1\right] = \frac{1}{2\ln\sigma} \quad ,$$

and

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

These residues are $O(1)$. Hence, we find that

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

□

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

$$\mathfrak{S}_n^{(D)} := \sum_{k=2}^{n}\binom{n}{k}\frac{(-1)^k}{\sigma^{k-1}-1}\left(\frac{p}{\sigma^{k-1}-q}\right)^{D+1} \quad . \tag{13}$$

Note that if $\mathfrak{S}_n^{(D)}$ is sublinear the main term of the asymptotic growth of $\mathbf{E}\left[T_n^D\right]$ is determined by equation (10) and Lemma 3 to be $n\left(\frac{D+1}{p} - \log_\sigma n\right)$.

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 4 (Searching with a Constant Bound). *Let $D = O(1)$, then*

$$\mathfrak{S}_n^{(D)} = -n\left(1+\frac{D+1}{p}\right) + A_n + \begin{cases} O\left((\log n)^{D+1}\right), & \text{for } q = \sigma^{-1} \\ O\left((\log_\sigma n)^D n^{\log_\sigma q+1}\right), & \text{for } q > \sigma^{-1} \\ O(1), & \text{otherwise.} \end{cases} \tag{14}$$

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

Theorem 5 (Searching with a Logarithmic Bound). *If $D+1 = c\log_\sigma n$, then we have*

$$\mathfrak{S}_n^{(D)} = \begin{cases} -n\left(1+\frac{D+1}{p}\right) + A_n + o(n), & \text{for } c < p \\ o(n), & \text{for } c > p. \end{cases} \tag{15}$$

The two theorems immediately yield Theorem 1. Both proofs rely on transferring the sum to a complex integral by Theorem 2:

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

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

Proof. By Rice' theorem we can write $\mathfrak{S}_n^{(D)}$ as

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

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

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

The function $f(z)$ is periodic with respect to the complex part 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 $O\left(1\right)$ if the singularities at $z = 1$, $z = 0$, $z = 1 \pm 2\pi\imath k/\ln\sigma$, and $z = -\log_\sigma q - 1 \pm 2\pi\imath k/\ln\sigma$, $k \in \mathbb{Z}$, are avoided, whereas $B(n+1,z) = \frac{n!}{z(z+1)\cdots(z+n)} = O\left(z^{-n-1}\right)$. Let $\mathcal{K}$ to be a circle of radius M, carefully chosen as to avoid any singularities. Then for some constant C we have

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

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 $O\left(1\right)$ dependent on the choice of M. This shows that

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

$\square$

The integral in equation (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 $z = -\log_\sigma q - 1 \pm 2\pi\imath k/\ln\sigma$, $k \in \mathbb{Z}$. The real part of the last ranges from 1 to $1 - \log_\sigma\left(\sigma^2 - 1\right) > -1$ under the assumption that $\sigma^{-2} \leq q \leq 1 - \sigma^{-2}$.

The evaluation of the residues proves tricky for the Beta function. We approximate the Beta function using an asymptotic expansion by Tricomi and Erdélyi [29] with help of a result of Fields [9]. This approach was already used by Szpankowski [25]. In the following we will lay a rigorous basis for this.

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

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

Proof.

$$\int_{-\infty}^{\infty} |\mathrm{B}(n, x+\imath y)|\, dy = 2\int_{0}^{\infty} |\mathrm{B}(n, x+\imath y)|\, dy = 2\int_{0}^{\infty} \frac{\Gamma(n)\, |\Gamma(x+\imath y)|}{|\Gamma(n+x+\imath y)|} dy$$
$$= 2\int_{0}^{\infty} \sqrt{2\pi} \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^{x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)} dy \left(1 + O\left(n^{-1}\right)\right) \quad ,$$

where $0 < \phi(n, x, y) = \arg(x + \imath y) - \arg(x + n + \imath y) < \frac{\pi}{2}$. Since

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

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

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

The derivative of (18) is

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

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, $\phi(y)$ 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 $\phi(y)$ decreases monotonically from $\pi - \epsilon$ to 0.

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

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

$$\int_{0}^{x} \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^{x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)} dy$$
$$\leq \left(\frac{n}{n+x}\right)^{n-\frac{1}{2}} \left(\sqrt{\frac{x^2+x^2}{n^2+2nx+2x^2}}\right)^{x} \frac{1}{\sqrt{x}} \int_{0}^{x} e^{0} dy \leq \left(\sqrt{2}x\right)^{x} n^{-x}\sqrt{x} = O\left(n^{-x}\right) \quad .$$

And for n large enough we have

$$\int_x^n \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^x \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)}dy$$
$$\leq \left(\sqrt{\frac{n^2}{(n+x)^2+x^2}}\right)^{n-\frac{1}{2}} \frac{1}{\sqrt{2x}} \int_x^n \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^x e^{-y\phi(n,x,y)}dy$$
$$\leq \int_x^n \left(\sqrt{\frac{x^2+y^2}{(n+x)^2+y^2}}\right)^x e^{-y\phi(n,x,y)}dy \leq \int_x^n \left(\sqrt{\frac{x^2+y^2}{n^2}}\right)^x e^{-y\phi(n,x,y)}dy$$
$$\leq n^{-x}\int_x^n \left(\sqrt{2}y\right)^x e^{-y\frac{\pi}{5}}dy \leq n^{-x}\frac{5}{\pi}\left(\sqrt{2}\frac{5}{\pi}\right)^x \Gamma(x+1) = O\left(n^{-x}\right) \quad .$$

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

The third part is

$$\int_n^{n^2} \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^x \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)}dy$$
$$\leq \left(\sqrt{\frac{n^2}{(n+x)^2+n^2}}\right)^{n-\frac{1}{2}} \left(\sqrt{\frac{x^2+n^4}{(n+x)^2+n^4}}\right)^x (x^2+n^2)^{-\frac{1}{4}}n^2 = O\left(2^{-\frac{n}{2}}n^2\right) \quad .$$

For the last part we have

$$\int_{n^2}^{\infty} \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^x \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)}dy$$
$$\leq \int_{n^2}^{\infty} \left(\sqrt{\frac{n^2}{(n+x)^2+yn^2}}\right)^{n-\frac{1}{2}} \frac{1}{\sqrt{y}}dy \leq \int_{n^2}^{\infty} y^{-\frac{n}{2}-\frac{1}{4}}dy = O\left(n^{-n-\frac{1}{2}}\right) \quad .$$

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

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

$$\int_0^n \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^x \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)}dy$$
$$\leq \sqrt{\frac{n}{n+x}}\left(\frac{n}{n+x}\right)^n \int_0^n \left(\frac{|x+\imath y|}{|n+x+\imath y|}\right)^x \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)}dy$$
$$\leq 4e^{-x}x^x\,(2n)^{-x}\frac{1}{\sqrt{x}}\int_0^n e^{-y\frac{\pi}{5}}dy = n^{-x}\frac{4(2e)^{-x}x^x}{\sqrt{x}}\frac{-5}{\pi}(e^{-\frac{1}{5}n\pi}-1) = O\left(n^{-x}\right) \quad .$$

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

The second part is

$$\int_n^{n^2} \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|n+x+\imath y|}{|x+\imath y|}\right)^{-x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)} dy$$
$$\leq \left(\sqrt{\frac{n^2}{(n+x)^2+n^2}}\right)^{n-\frac{1}{2}} \left(\sqrt{\frac{(n+x)^2+n^2}{x^2+n^2}}\right)^{-x} (x^2+n^2)^{-\frac{1}{4}} n^2$$
$$\leq \left(\sqrt{\frac{n^2}{\frac{3}{2}n^2}}\right)^{n-\frac{1}{2}} 2^{-\frac{x}{2}} n^2 = O\left(\left(\frac{3}{2}\right)^{-\frac{n}{2}} n^2\right) \quad .$$

The final part is

$$\int_{n^2}^{\infty} \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|n+x+\imath y|}{|x+\imath y|}\right)^{-x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)} dy$$
$$\leq \int_{n^2}^{\infty} \left(\frac{n}{\sqrt{(n+x)^2+y^2}}\right)^{n-\frac{1}{2}} \left(\sqrt{\frac{2y^2}{y^2}}\right)^{-x} \frac{1}{\sqrt{y}} e^{-y\phi(n,x,y)} dy \leq 2^{-\frac{x}{2}} \int_{n^2}^{\infty} y^{-\frac{n}{2}-\frac{1}{4}} dy$$
$$= O\left(n^{-n-\frac{1}{2}}\right) \quad .$$

This finishes the proof. $\square$

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

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

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

$$\int_{f(n)\ln n}^{n} \left(\frac{n}{|n+x+\imath y|}\right)^{n-\frac{1}{2}} \left(\frac{|n+x+\imath y|}{|x+\imath y|}\right)^{-x} \frac{1}{\sqrt{|x+\imath y|}} e^{-y\phi(n,x,y)} dy$$
$$\leq e^{-x} \left(\frac{f(n)\ln n}{\sqrt{2}n}\right)^{x} (\ln n)^{-\frac{1+\epsilon}{2}} \int_{f(n)\ln n}^{n} e^{-y\left(\frac{\pi}{4}-\epsilon\right)} dy$$
$$\leq e^{-x} \left(\frac{f(n)\ln n}{\sqrt{2}n}\right)^{x} (f(n)\ln n)^{-\frac{1}{2}} \frac{1}{\left(\frac{\pi}{4}-\epsilon\right)} e^{-f(n)\ln n\left(\frac{\pi}{4}-\epsilon\right)} = O\left(n^{-f(n)\left(\frac{\pi}{4}-\epsilon\right)-x}\right) \quad .$$

$\square$

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

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

Proof. The proof relies on a standard expansion of the ratio of two Gamma functions by Tricomi and Erdélyi [29], the proof that the expansion is uniform by Fields [9], and the exponentially small tails of the integrands. The expansion is

$$\frac{\Gamma(z+\alpha)}{\Gamma(z+\beta)} = z^{\alpha-\beta}\sum_k \frac{1}{k!}\frac{\Gamma(1+\alpha-\beta)}{\Gamma(1+\alpha-\beta-k)}B_k^{(1+\alpha-\beta)}(\alpha)z^{-k} + O\left(z^{\alpha-\beta-m}\left(1+|\alpha-\beta|^m\right)\left(1+|\alpha|+|\alpha-\beta|\right)^m\right) \quad , \tag{20}$$

which is uniformly valid for $|\arg(z+\alpha)| < \pi$, and $\left(1+|\alpha-\beta|\right)\left(1+|\alpha|+|\alpha-\beta|\right) = o(z)$, $\beta-\alpha \notin \mathbb{N}$, $z \to \infty$. The $B_n^{(a)}(x)$ are the generalized Bernoulli polynomials (See Temme [28] or Nörlund [20], chap. 6, §5), which are multivariant 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 $B_2^{(a)}(x) = \left(3a^2 + 12x^2 - a(1+12x)\right)/12$.

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

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

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

$$\int_{f(n)\ln n}^{\infty} |\mathrm{B}(n,x+\imath y)|dy = O\left(e^{-n^{\frac{1}{3}}\left(\frac{\pi}{4}-\epsilon\right)}n^{-x}\right) \quad .$$

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

$$
\begin{aligned}
&\int_{f(n)\ln n}^{\infty} \left|n^{-k-x-\imath y}\Gamma(x+\imath y+k)(x+\imath y)^k\right| dy \\
&\quad\le 2^k\sqrt{2\pi}n^{-k-x}\int_{n^{\frac13}}^{\infty} y^k|x+k+\imath y|^{x+k-\frac12}e^{-y\arg(x+k+\imath y)-x-k+\frac{x+k}{12((x+k)^2+y^2)}}dy \\
&\quad\le 2^{x+2k}\sqrt{2\pi}n^{-k-x}\int_{n^{\frac13}}^{\infty} y^{x+2k}e^{-y\arg(x+k+\imath y)}dy \\
&\quad\le 2^{x+2k}\sqrt{2\pi}n^{-k-x}\int_{n^{\frac13}}^{\infty} y^{2k}e^{-y\frac{\pi}{3}}dy \le 2^{x+2k}\sqrt{2\pi}n^{-k-x}\left(\frac{3}{\pi}\right)^{2k+1}\int_{\frac{\pi}{3}n^{\frac13}}^{\infty} u^{2k}e^{-u}du \\
&\quad\le 2^{x+2k}\sqrt{2\pi}n^{-k-x}\left(\frac{3}{\pi}\right)^{2k+1} 2\left(\frac{\pi}{3}n^{\frac13}\right)^{2k+1} e^{-\frac{\pi}{3}n^{\frac13}} = O\left(n^{-\frac{k}{3}-x}e^{-\frac{\pi}{3}n^{\frac13}}\right) \quad .
\end{aligned}
$$

Since we have $\arg(x+k+\imath) \ge \frac{\pi}{3}\mathrm{sgn}(y)$, and $\int_{\frac{\pi}{3}n^{\frac13}}^{\infty} u^{2k}e^{-u}du \le 2\left(\frac{\pi}{3}n^{\frac13}\right)^{2k} e^{-\frac{\pi}{3}n^{\frac13}}$ by iterated integration (for n large enough). Hence, the error on the tail $y > n^{\frac13}$ is exponentially small. $\square$

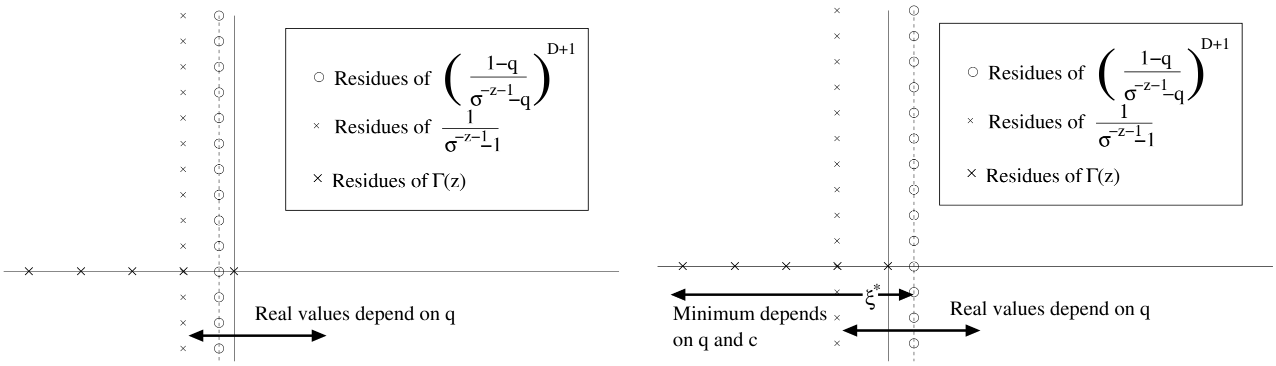

Figure 2: Behavior of the studied function in the complex plane.

We concentrate on the first term of the expansion since each further term is by an order of magnitude smaller then 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

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

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 $q = 1-p$ the right line (dotted with circles) of residues moves left or right. The real value $-\xi^*$, where the absolute value of the function under the integral is minimal, also depends upon c for the case of $D+1 = c\log_\sigma n$. 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 the $-\xi^* < -1$, the whole integral is too small to compensate for 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. We find that

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

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

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

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

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

The B_k are the Bernoulli numbers (see, e.g., Temme [28] or equation (32)). ☐

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

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

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

$$\mathfrak{S}_n^{(D)} = A_n - n\left(1+\frac{D+1}{p}\right) + \mathfrak{I}_{1-\epsilon,n}^{(D)} + O(1) \quad . \tag{25}$$

If we keep the line right of -1 we get

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

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

$$\mathfrak{I}_{\xi,n}^{(D)} \leq \frac{C}{\sigma^{\xi-1}-1} n^{c\log_\sigma\left(\frac{p}{\sigma^{\xi-1}-q}\right)+\xi} \quad .$$

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

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

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

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

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

With respect to x there is a single extreme value, a maximum, at $x^* = -\log_\sigma q$, 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 $\xi^* \geq 2$ we derive that $c > 1-\frac{q}{\sigma}$ and that the minimum is taken for some $\epsilon > 0$ at $\xi = 2-\epsilon$. Since $c\log_\sigma\left(\frac{1-q}{\sigma^{1-\epsilon}-q}\right) < c\log_\sigma 1 = 0$ we find that

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

The derivative of $\left(1-\frac{q}{\sigma}\right)\log_\sigma\left(\frac{1-q}{\sigma^{1-\epsilon}-q}\right)+2-\epsilon$ w.r.t. q is

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

The second derivative is

$$\frac{2}{\sigma}\left(\frac{\sigma^{1-\epsilon}-1}{(\sigma^{1-\epsilon}-q)(1-q)\ln\sigma}\right)-\left(1-\frac{q}{\sigma}\right)\left(\frac{(\sigma^{1-\epsilon}-1)\left((1-q)+(\sigma^{1-\epsilon}-q)\right)}{(\sigma^{1-\epsilon}-q)^2(1-q)^2\ln\sigma}\right) \quad ,$$

which is smaller 0 for $q < \frac{\sigma^{2-\epsilon}+\sigma-2\sigma^{1-\epsilon}}{1+2\sigma-\sigma^{1-\epsilon}} \xrightarrow[\epsilon\to 0]{} \sigma$. Therefore, the first derivative is decreasing for $q < \sigma-\epsilon$, it is maximal at $q=0$. Here we have a value of $\frac{(1-\epsilon)\ln\sigma-\sigma+\sigma^\epsilon}{\sigma\ln\sigma} \xrightarrow[\epsilon\to 0]{} \frac{\ln\sigma-\sigma+1}{\sigma\ln\sigma} < 0$ for $\sigma \geq 2$. The term is decreasing in q and the maximal value is attained at $q=0$, where we have a value of 1. As a result we have for some small ϵ

$$\mathfrak{E}_{c,q,1-\epsilon} < 1 \quad .$$

Finally, for $c \geq 0$ the derivative of $\mathfrak{E}_{c,q,\xi}$ is

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

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

$$\mathfrak{E}_{c,q,1-\epsilon} \le \log_\sigma\left(\frac{1-q}{\sigma^{1-\epsilon}-q}\right) + 2 - \epsilon \ ,$$

where the right terms derivative w.r.t. q is $-\frac{\sigma^{1-\epsilon}-1}{(\sigma^{1-\epsilon}-q)(1-q)\ln\sigma} < 0$. This, by the same arguments as above, shows that $\mathfrak{E}_{c,q,1-\epsilon} < 1$ for $q > 0$. $\square$

Note that the above formulas could be used to determine the exponent exactly for concrete values of q and c.

We can now prove Theorem 5. If $\xi^* > 1$ or $c \ge 1$ we find a $\xi \in (1,2)$ such that $\mathfrak{E}_{c,q,\xi} < 1$, thus $\mathfrak{I}^{(D)}_{\xi,n} = o(n)$ and $\mathfrak{S}^{(D)}_n = o(n)$ by equation (26).

If $\xi^* < 1$, we move the line of integration to $-\xi^* = -1+\epsilon$ and find that $\mathfrak{I}^{(D)}_{1-\epsilon,n} = o(n)$, and by equation (25) we have $\mathfrak{S}^{(D)}_n = A_n - n\left(1+\frac{D+1}{p}\right) + o(n)$. 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 $z = -\log_\sigma q - 1 \pm 2\pi\imath k/\ln\sigma$ are always to the right of ξ^*).

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

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

Proof. The singularities at $z = -\log_\sigma q - 1 \pm 2\pi\imath k/\ln\sigma$ 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 $c < 1 - q\sigma$ (thus $q < \sigma^{-1}$) and

$$-\mathrm{res}\left[g(z), z=0\right] = \frac{\sigma}{\sigma-1}\left(\frac{p\sigma}{1-q\sigma}\right)^{D+1} = \frac{\sigma}{\sigma-1} n^{c\log_\sigma\left(\frac{1-q}{\sigma^{-1}-q}\right)} < \frac{\sigma}{\sigma-1} n^{(1-q\sigma)\log_\sigma\left(\frac{1-q}{\sigma^{-1}-q}\right)} \ . \tag{28}$$

The exponent $(1-q\sigma)\log_\sigma\left(\frac{1-q}{\sigma^{-1}-q}\right)$ has derivative (w.r.t. q)

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

since $\log_\sigma\left(\frac{1-q}{\sigma^{-1}-q}\right) > 1$ and

$$\frac{(1-q\sigma)}{\ln\sigma}\left[\frac{1}{\sigma^{-1}-q} - \frac{1}{1-q}\right] = \frac{(1-q\sigma)(1-\sigma^{-1})}{(\sigma^{-1}-q)(1-q)\ln\sigma} = \frac{\sigma(1-q\sigma)(1-\sigma^{-1})}{(1-q\sigma)(1-q)\ln\sigma} = \sigma\frac{(1-\sigma^{-1})}{(1-q)\ln\sigma} < \sigma \ .$$

The exponent has value 1 at $q=0$, hence it is smaller than 1 for $q > 0$. Thus, the singularities contribution is $o(n)$, which does not affect the result. $\square$

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 $c < 1 - q = p$, which proves Theorem 5.

To prove Theorem 4 we calculate the remaining residues at $z = 0$ and $z = -\log_\sigma q - 1 \pm 2\pi\imath k/\ln\sigma$.

Lemma 13 (Residues at $z = 0$, $z = -\log_\sigma q - 1 + \frac{2\pi\imath k}{\ln\sigma}$).

Let $g(z) := \frac{1}{\sigma^{-1-z}-1}\left(\frac{p}{\sigma^{-z-1}-q}\right)^{D+1}\Gamma(z)n^{-z}$. If $q = \sigma^{-1}$ we have

$$\operatorname{res}\left[g(z), z=0\right] =$$
$$\sum_{l=0}^{D+1}\frac{(\sigma-1)^{D+1}(-1)^{-l}B_{-l+D+1}^{(D+1)}}{(-l+D+1)!}\sum_{i=0}^{l}\left[\sum_{j=0}^{i}\frac{A_j(\sigma^{-1})}{j!(i-j)!}\left(\frac{\sigma}{1-\sigma}\right)^{j+1}\left(-\frac{\ln n}{\ln\sigma}\right)^{i-j}\right]\frac{\gamma_{l-i-1}^{(0)}}{(\ln\sigma)^{l-i}}\ , \tag{29}$$

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

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

$$\operatorname{res}\left[g(z), z = -\log_\sigma q - 1 + \frac{2\pi\imath k}{\ln\sigma}\right] =$$
$$\sum_{l=0}^{D}\left(\frac{1-q}{q}\right)^{D+1}\frac{(-\ln\sigma)^{-l}B_{-l+D}^{(D+1)}}{(-l+D)!}\sum_{i=0}^{l}\left[\sum_{j=0}^{i}\frac{-A_j(q)(-\ln\sigma)^j}{(1-q)^{j+1}j!}n^{\log_\sigma q+1-\frac{2\pi\imath k}{\ln\sigma}}\frac{(-\ln n)^{i-j}}{(i-j)!}\right]$$
$$\left(\gamma_{l-i}^{\left(-\log_\sigma q-1+\frac{2\pi\imath k}{\ln\sigma}\right)}\right)\ . \tag{30}$$

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 $z = -\log_\sigma q - 1 + \frac{2\pi\imath k}{\ln\sigma}$. If $q \neq \sigma^{-1}$, then the residue at $z = 0$ is given by equation (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 $g(z) = \frac{1}{\sigma^{-1-z}-1}\left(\frac{p}{\sigma^{-z-1}-q}\right)^{D+1}\Gamma(z)n^{-z}$ at $z = -\log_\sigma q - 1 + \frac{2\pi\imath k}{\ln\sigma}$. These can be derived best in terms of Eulerian polynomials $A_n(u)$ defined by

$$\frac{1-u}{1-ue^{t(1-u)}} = \sum_{n=0}^{\infty}A_n(u)\frac{t^n}{n!} = \sum_{n=0}^{\infty}\frac{t^n}{n!}\sum_{k=0}^{n}A_{n,k}u^k \qquad |t| < \frac{\ln u}{u-1} \tag{31}$$

(see Comtet [8] or Graham et al. [12] for Eulerian numbers $A_{n,k}$). We also need generalized Bernoulli numbers $B_k^{(a)}$, defined by

$$\left(\frac{t}{e^t-1}\right)^a = \sum_{k=0}^{\infty}B_k^{(a)}\frac{t^k}{k!} \qquad |t| < 2\pi \tag{32}$$

(see Temme [28] or Nörlund [20]). We then have the following series representations:

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

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

□

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

$$
-\frac{\sigma(\sigma-1)^D}{(D+1)!}\left(\log_\sigma n\right)^{D+1} \quad , \tag{33}
$$

otherwise, this term is

$$
-\frac{(1-q)^D}{D!q^{D+1}}\left(\log_\sigma n\right)^D n^{\log_\sigma q+1} n^{-\frac{2\pi\imath k}{\ln\sigma}}\Gamma\left(-\log_\sigma q-1+\frac{2\pi\imath k}{\ln\sigma}\right) \quad . \tag{34}
$$

For $\log_\sigma q+1<0$ this is $o(1)$. In this case the residue at $z=0$ yields $O(1)$, see equation (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

$$
\left|\sum_{k\in\mathbb{Z}} n^{-\frac{2\pi\imath k}{\ln\sigma}}\gamma_{l-i}^{\left(-\log_\sigma q-1+\frac{2\pi\imath k}{\ln\sigma}\right)}\right| = O(1) \quad . \tag{35}
$$

Hence, the growth for constant D is

$$
\mathfrak{S}_n^{(D)} = A_n - n\left(1+\frac{D+1}{p}\right) + \begin{cases} O\left(\left(\log n\right)^{D+1}\right), & \text{for } q=\sigma^{-1}\\ O\left(\left(\log_\sigma n\right)^D n^{\log_\sigma q+1}\right), & \text{for } q>\sigma^{-1}\\ O(1), & \text{otherwise.}\end{cases} \tag{36}
$$

Thus, we have proven Theorem 4.

References

[1] A. Apostolico and W. Szpankowski. Self-alignments in words and their applications. *Journal of 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. *Journal of the 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 *String Processing and Information Retrieval Symp. SPIRE*, pages 16–23. IEEE, 1999.

[4] R. D. L. Briandais. File searching using variable length keys. In *Proc. of the 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, September 2003.

[6] A. Buchner, H. Täubig, and J. Griebsch. A fast method for motif detection and searching in a protein structure database. In *Proceedings of the German Conference on Bioinformatics (GCB'03)*, volume 2, pages 186–188, October 2003.

[7] A. L. Cobbs. Fast approximate matching using suffix trees. In *Proc. of the 6th Sym. on Combinatorial Pattern Matching (CPM)*, volume 937 of *LNCS*, pages 41–54. Springer, 1995.

[8] L. Comtet. *Advanced Combinatorics.* D. Reidel Publishing, Dortrecht-Holland, 1974.

[9] J. L. Fields. The uniform asymptotic expansion of a ratio of Gamma functions. In *Proc. of the 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] E. Fredkin. Trie memory. *Communications of the ACM*, 3(9):490–499, 1960.

[12] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley, 2nd edition, 1994.

[13] P. Jokinen and E. Ukkonen. Two algorithms for approximate string matching in static texts. In *Proc. of the 16th Int'l Symp. on Mathematical Foundations of Computer Science (MFCS)*, volume 520 of *LNCS*, pages 240–248. Springer, 1991.

[14] J. Kingman. Subadditive ergodic theory. *The Annals of Probability*, 1(6):883–909, 1973.

[15] P. Kirschenhofer. A note on alternating sums. *Electronic Journal of Combinatorics*, 3(2), 1996.

[16] D. E. Knuth. *The Art of Computer Programming – Sorting and Searching*, volume 3. Addison Wesley, 2nd edition, feb 1998.

[17] D. R. Morrison. PATRICIA – practical algorithm to retrieve information coded in alphanumeric. *J. of the ACM*, 15(4):514–534, oct 1968.

[18] G. Navarro. *Approximate Text Searching*. PhD thesis, University of Chile, Dept. of Computer Science, University of Chile, Santiago, Chile, 1998.

[19] G. Navarro and R. Baeza-Yates. A hybrid indexing method for approximate string matching. *Journal of Discrete Algorithms (JDA)*, 1(1):205–209, 2000. Special issue on Matching Patterns.

[20] N. E. Nörlund. *Vorlesungen über Differenzenrechnung*. Springer, Berlin, 1924.

[21] K. Oflazer. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computer Linguist*, 22(1):73–89, 1996.

[22] B. Pittel. Paths in a random digital tree: Limiting distributions. *Adv. Appl. Prob.*, 18:139–155, 1986.

[23] H. Prodinger and W. Szpankowski (Guest Editors). *Theoretical Computer Science*. Elsevier, 144(1–2) (Special Issue), 1995.

[24] K. U. Schulz and S. Mihov. Fast string correction with Levenshtein automata. *Int. J. on Document Analysis and Recognition (IJDAR)*, 5:67–85, 2002.

[25] W. Szpankowski. The evaluation of an alternative sum with applications to the analysis of some data structures. *Information Processing Letters*, 28:13–19, 1988.

[26] W. Szpankowski. Some results on v-ary asymmetric tries. *J. of Algorithms*, 9:224–244, 1988.

[27] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley-Interscience, 1 edition, 2000.

[28] N. M. Temme. *An Introduction to Classical Functions of Mathematical Physics*. Wiley, New York, 1996.

[29] F. G. Tricomi and A. Erdélyi. The asymptotic expansion of a ratio of Gamma functions. *Pacific J. of Mathematics*, 1:133–142, 1951.

[30] E. Ukkonen. Approximate string-matching over suffix trees. In *Proc. of the 4th Sym. on Combinatorial Pattern Matching (CPM)*, volume 684 of *LNCS*, pages 228–242. Springer, 1993.

[31] F. A. O. Werner, G. Durstewitz, F. A. 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*, to appear, 2003.