Optimal Versus Randomized Search of Fixed Length Binary Words

May 27, 2001

Helmut Prodinger*
School of Mathematics
University of the Witwatersrand
P. O. Wits 2050, Johannesburg
South Africa
helmut@gauss.cam.wits.ac.za

Wojciech Szpankowski†
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U. S. A.
spa@cs.purdue.edu

Abstract

We consider the search problem in which one finds a binary word among m words chosen randomly from the set of all words of fixed length n. It is well known that the optimal search is equivalent to the Huffman coding that requires on average $\log_2 m$ bits to be checked *plus* a small additional cost called the average *redundancy*. The latter is an oscillating function of m and is bounded between zero and $1-(1+\ln\ln 2)/\ln 2 \approx 0.0860713320$. As a matter of fact, it is known that finding the optimal strategy for this problem is NP-hard. We propose here several simple randomized search strategies leading respectively to the following average redundancies: 1.332746177, 0.6113986565, 0.4310617764, and 0.332746177, plus some small oscillations that we characterize quite precisely. These results should be compared to the optimal, but NP-hard, search algorithm. Our findings extend and make more precise recent results of Fedotov and Ryabko. They are obtained by analytic methods such as generating functions, Mellin transform, and analytic depoissonization.

Index Terms: Search problem, Shannon source coding theorem, Huffman code, tries, PATRICIA trie, generating functions, Mellin transform, analytic information theory.

*The work of this author was supported by The John Knopfmacher Centre for Applicable Analysis and Number Theory.

†The work of this author was supported by NSF Grant CCR-9804760 and contract 1419991431A from sponsors of CERIAS at Purdue.

1 Introduction

A combinatorial *search problem* can be defined as follows: Given a set $\mathcal{W} = \{w_1, w_2, \ldots, w_m\}$ of m words over a (binary) alphabet Σ, design a sequence of tests that successfully find the word $w^* \in \mathcal{W}$ being sought (cf. [1]). The prime goal of the *optimal* search is to find the sought word w^* with the smallest maximum or average search time. It is well known (cf. [1]) that the problem of determining a sequential strategy with the minimum average search time is equivalent to the information-theoretic problem of minimizing the average codeword length of a certain prefix code, that is, constructing the optimal Huffman code. We shall use this equivalence between the search problem and prefix codes throughout the paper. Among others this will allow us to determine the lower bound on the average search time to which we compare some randomized search algorithms.

In this paper, we deal with a randomly selected set $\mathcal{W}$ of m binary words of fixed length n. But more precisely, we assume that $w_i = b_{i1}b_{i2}\ldots b_{in}$ $(1 \le i \le m)$ is a sequence of n binary symbols $b_{ij} \in \Sigma = \{0,1\}$, and that the set $\mathcal{W} = \{w_1, \ldots, w_m\}$ is chosen with equal probability among all $\binom{2^n}{m}$ possible subsets of size m. This probabilistic model, known as *uniform model*, was suggested by Fedotov and Ryabko [3]. Let $t^*_{n,m}$ be the optimal average search time for a word in $\mathcal{W}$. From the Shannon source coding theorem (cf. [2, 11]) we know that $t^*_{n,m} \ge \log_2 m$. Due to the integer nature of the search time this lower bound is achieved only for m being a power of two. More precisely, the optimal average search time is equal to the average code length of the optimal Huffman code when words w are equally likely to be generated. From [1, 10] we know that in this case

$$\begin{aligned} t^*_{n,m} &= \lceil \log_2 m \rceil + 1 - \frac{2^{\lceil \log_2 m \rceil}}{m} \\ &= \log_2 m + 2 - \{\log_2 m\} - 2^{1-\{\log_2 m\}}, \end{aligned} \tag{1}$$

where $\{x\} = x - \lfloor x \rfloor$ is the fractional part of x. We define the *average redundancy* as the difference between the average search time (code length) and $\log_2 m$ (entropy), that is, $R_m := t_{n,m} - \log_2 m$. The average redundancy of the optimal search exhibits some oscillations as indicated by (1) and shown in Figure 1. Easy calculations reveal that the optimal redundancy R^*_m oscillates between zero and its maximum value

$$\max_m R^*_m = 1 - \frac{\ln\ln 2 + 1}{\ln 2} \approx 0.0860713320$$

which is quite negligible.

In this paper we shall follow Fedotov and Ryabko and design simple and randomized search algorithms that achieve a low redundancy. Such algorithms can be represented by digital trees such as *tries* and *PATRICIA tries* (cf. [7, 8, 11]). In tries the words of $\mathcal{W}$ are stored in leaves while internal nodes are used for branching. In *standard* tries the branching process at level k is determined by the kth symbol of the input word. If the

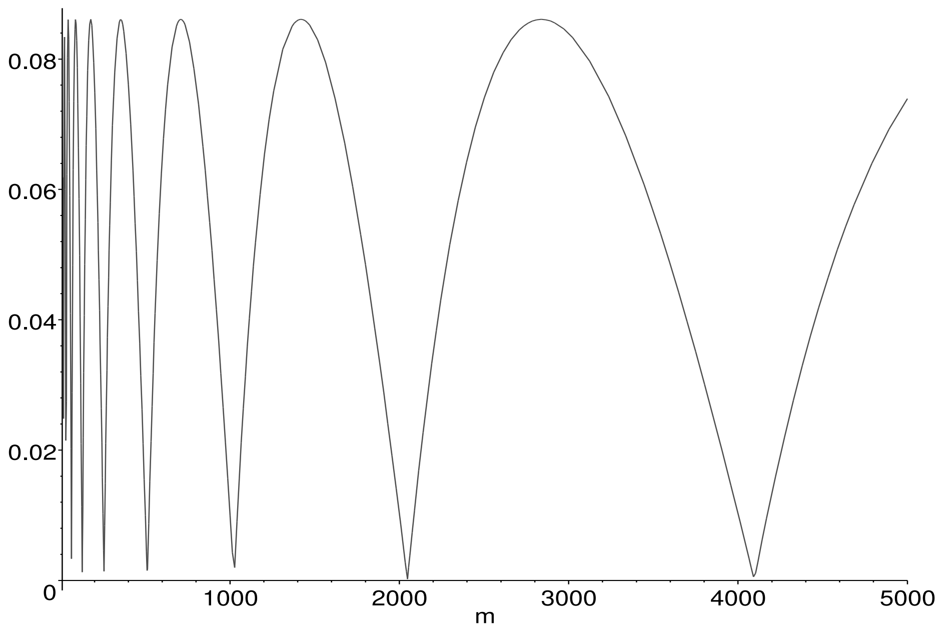

Figure 1: The average redundancy of the optimal search time.

kth symbol is 0 we move to the left, otherwise to the right. The path from the root to a leaf (external node) is the shortest prefix of a word that distinguishes it from all other words in $\mathcal{W}$. In the *optimal trie* one branches not on the consecutive symbols but rather on those that give the minimal average search cost (i. e., the minimal *typical depth* in the associated trie). In this case, internal nodes contain the index (position) according to which the branching is performed. For example, in Figure 2 we draw two tries built for the set $\mathcal{W} = \{\mathtt{aaaaa}, \mathtt{abbab}, \mathtt{baabb}, \mathtt{babaa}\}$ over the alphabet[1] $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ where the branching bit position is indicated by a number stored in the internal node. Observe that in Figure 2 the first trie is optimal and requires only 2 comparisons per word, while the average search time for second trie is 2.75.

Unfortunately, the problem of finding the optimal trie for a given set of words is known to be NP-hard [5] (basically we need to inspect all $n!$ possible permutations of bits). To circumvent this problem Fedotov and Ryabko [3] proposed a randomized algorithm in which the branching bit positions are selected randomly. They proved that the average search time is bounded by $\log_2 m + 1.04$, thus the average redundancy $R_n \leq 1.04$. In this paper, we extend and make the analysis of [3] more precise. In particular, we propose several simple randomized search strategies that lead respectively to the following redundancies: 1.332746177, 0.6113986565, 0.4310617764, and 0.332746177 plus some tiny oscillations (of amplitude $\leq 3 \cdot 10^{-5}$) that we describe precisely as Fourier series. The Fedotov and Ryabko

[1] To avoid confusions with numerical indices stored in the internal nodes the alphabet Σ in this case consists of letters $\mathtt{a}$ and $\mathtt{b}$.

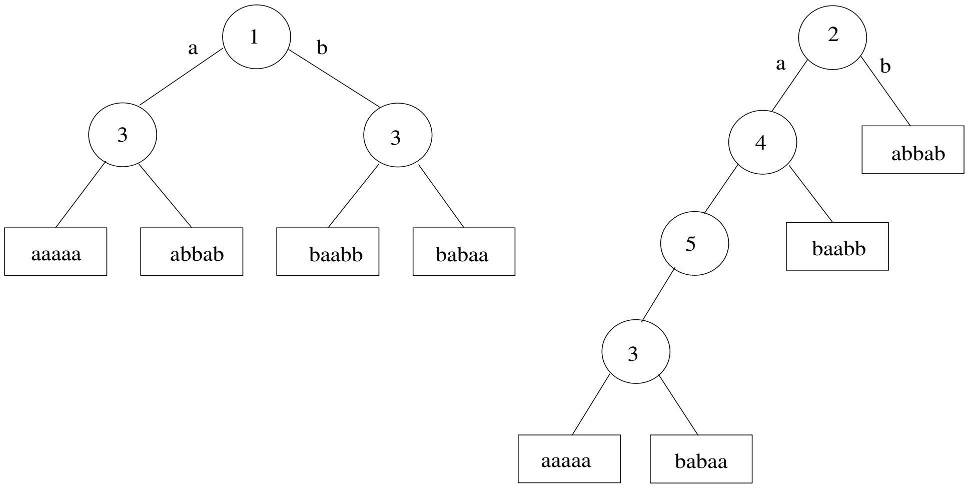

Figure 2: Two tries built over $\mathcal{W} = \{\texttt{aaaaa}, \texttt{abbab}, \texttt{baabb}, \texttt{babaa}\}$.

algorithm corresponds to our 0.4310617764 redundancy; thus we also improve their result.

Our analysis of randomized tries is based on solving exactly and asymptotically a certain recurrence on the average search cost which is equivalent to the average typical cost in a trie. Tries, and in general digital trees, were extensively analyzed over the last twenty years (cf. [7, 8, 11]) for memoryless, Markov and mixing sources. However, the probabilistic model introduced by Fedotov and Ryabko [3], which we further call the *uniform model*, differs from all the others in that m words of length n are chosen uniformly over all 2^n possible words. Although this introduces additional correlations, we are able to show that asymptotically the results are almost identical to the ones obtained for the unbiased memoryless source with m independently generated words of infinite lengths.

This paper is organized as follows. In the next section we present our main results and compare them to the optimal search. We also elaborate on the typical behavior of tries under the uniform model. The derivations of our main findings are in Section 3. We should point out that we apply analytic methods of the analysis of algorithms such as generating functions, Mellin transform, and analytic depoissonization. The reader may familiarize with these topics from the recent book [11]. In addition, in Appendix A we summarize basic properties of the Mellin transform.

2 Main Results

We recall that we consider the *uniform model* in which the set $\mathcal{W}$ of m words of fixed length n is chosen with probability $\binom{2^n}{m}$ among all subsets 2^n of size m. For a given set $\mathcal{W}$ we apply a *randomized* algorithm to construct the associated trie; that is, we select randomly one position among n possible positions and discriminate all words from $\mathcal{W}$ according to this position. All words with 1 on the selected position will move the the right in the trie, while the remaining words will move to the left. We then select another position, and continue the splitting process until all words have unique paths in the trie.

Our goal is to estimate the average *typical depth* (or equivalently the average search time). The typical depth is defined as the length of a path from the root to a *randomly* selected leaf. Observe that the typical depth is a random variable even for a given set $\mathcal{W}$. The average typical depth or the average search time we denote as $t_{n,m}$ for $m \geq 0$ and $n \geq 1$. As already observed in [3] the average depth satisfies the following recurrence

$$t_{n,m} = 1 + \sum_{k=0}^{m} \frac{\binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k}}{\binom{2^n}{m}} \left(\frac{k}{m} t_{n-1,k} + \frac{m-k}{m} t_{n-1,m-k} \right), \qquad m \geq 2. \tag{2}$$

The above recurrence holds only for $m \geq 2$, and we set for now the following initial conditions:

$$t_{n,0} \equiv t_{n,1} \equiv 0, \qquad n \geq 1. \tag{3}$$

To justify recurrence (2) we observe that with probability $c_{n,m,k} = \frac{\binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k}}{\binom{2^n}{m}}$ some k words will end up on the left to the root and the remaining words will go to the right subtree. There is a unit cost 1 associated with the path from the root to a subtree. With probability k/m the word being sought is in the left subtree, while with probability $(m-k)/m$ it will end up in the right subtree. The average depths in the subtrees are $t_{n-1,k}$ and $t_{n-1,m-k}$, respectively. This establishes the recurrence, while the initial conditions (3) are quite natural for this problem.

In the next section, we use a generating function approach to solve exactly (2)–(3), and then we apply the Mellin transform approach (cf. Appendix A) to derive the asymptotic results. We summarize our findings in the theorem below.

Theorem 1 *For all $m \geq 2$ and $n \geq 1$ recurrence (2) with the initial condition (3) possesses the following exact solution*

$$t_{n,m} = n - \binom{2^n - 1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n - 2^l}{m-1}. \tag{4}$$

For $n \to \infty$ and large m[2] *the average search time $t_{n,m}$ attains the following asymptotics*

$$t_{n,m} = \log_2 m + \frac{\gamma}{\ln 2} + \frac{1}{2} + Q_0(\log_2 m) + O\left(\frac{1}{m}\right) + O(n2^{-n}) \tag{5}$$

[2] The asymptotics are to be understood as follows. The approximation of $t_{n,m}$ by $t_{\infty,m}$ introduces an error $O(n2^{-n})$, and then the approximation of $t_{\infty,m}$ is done within an error $O(1/m)$.

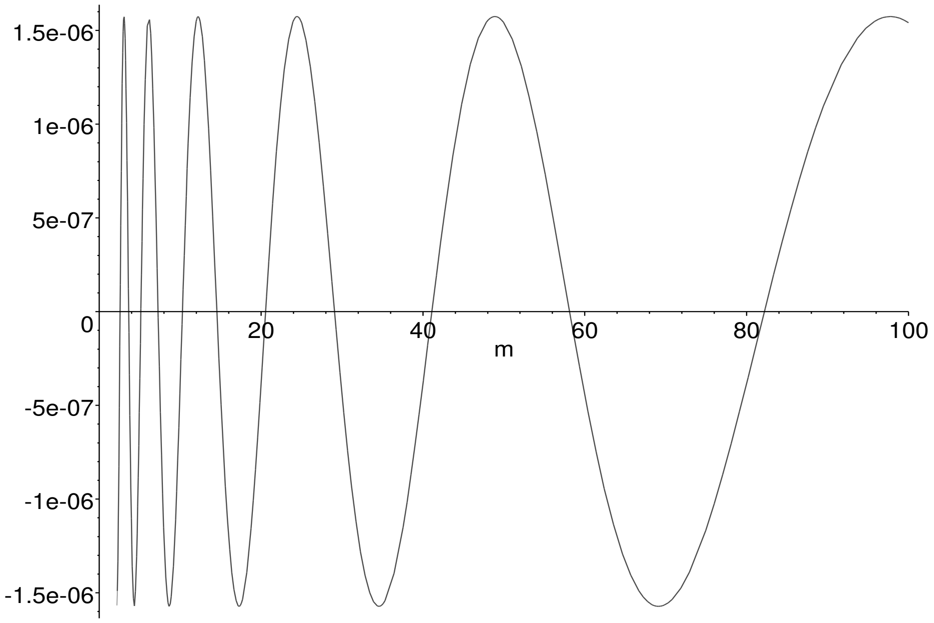

Figure 3: The periodic function $Q_0(m)$ defined in (6).

where $\gamma = 0.5772156649\ldots$ *is Euler's constant and*

$$Q_0(\log_2 m) = \frac{1}{\ln 2} \sum_{\ell \in \mathbb{Z} \setminus \{0\}} \Gamma\Big(\frac{2\pi i \ell}{\ln 2}\Big) e^{-2\pi i \ell \log_2 m} \tag{6}$$

is a periodic function of magnitude $\leq 2 \times 10^{-6}$ *(cf. Figure 3). In the above, the summation is over all non-zero integers in* $\mathbb{Z}$*, and* $\Gamma(z)$ *is the Euler gamma function (cf. [7, 11]).*

Based on the above result we conclude that the average redundancy $R_m = t_{n,m} - \log_2 m$ is equal to $\frac{\gamma}{\ln 2} + \frac{1}{2} + Q_0(\log_2 m) \approx 1.332746177$ and exceeds by at least 1.25 and at most 1.33 the optimal redundancy R_m^* computed in (1). While the cost of building such a randomized trie is only $O(n^2)$ in the worst case and $\mathcal{O}(m \log m)$ on average, the trie requires more memory than the optimal tree. The optimal tree has only $m-1$ internal nodes (and m leaves containing the words from $\mathcal{W}$). In the randomized trie some internal nodes may be unary, increasing the total number of internal nodes. By how much? This is answered in Theorem 2 below.

Let $s_{n,m}$ denote the average size of the trie, that is, the average number of internal (branching) nodes. It is easy to see that it satisfies the following recurrence

$$s_{n,m} = 1 + 2 \sum_{k=0}^{m} \frac{\binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k}}{\binom{2^n}{m}} s_{n-1,k}, \tag{7}$$

that holds for $m \geq 2$. The initial conditions are

$$s_{n,0} \equiv s_{n,1} \equiv 0.$$

To derive (7) we observe that the root contributes 1 (the additive term of the recurrence), and the two subtrees contribute $s_{n-1,k}$ and $s_{n-1,m-k}$, where k is distributed as $\frac{\binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k}}{\binom{2^n}{m}}$.

In the next section we solve the above recurrence and obtain asymptotic results,which are summarized below. They indicate that on average the randomized trie has $m/\ln 2 \approx 1.443m$ (plus a tiny oscillation) internal nodes, which should be compared to $m-1$ nodes of the optimal tree.

Theorem 2 [SIZE OF TRIES.] *For $m \geq 2$ the size $s_{n,m}$ satisfies (7) and is given by the following exact formula*

$$s_{n,m} = 2^n - 1 - \binom{2^n}{m}^{-1} \sum_{l=1}^{n} 2^{n-l} \binom{2^n - 2^l}{m} - m \binom{2^n - 1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n - 2^l}{m-1}. \tag{8}$$

For $n \to \infty$ and large m the size $s_{n,m}$ becomes

$$s_{n,m} = \frac{m}{\ln 2}\Big(1 + Q(\log_2 m)\Big) - 1 + O\Big(\frac{1}{m}\Big) + O(n2^{-n}), \tag{9}$$

where

$$Q(\log_2 m) = \frac{1}{\ln 2} \sum_{\ell \in \mathbb{Z} \setminus \{0\}} 2\pi i \ell \, \Gamma\Big(-1 + \frac{2\pi i \ell}{\ln 2}\Big) e^{-2\pi i \ell \log_2 m}$$

is a periodic function of small amplitude ($\leq 1.1 \times 10^{-6}$).

We now propose three modifications of our basic algorithm that will lead to successively improved redundancies. In the first two modifications we only change the initial conditions. Let us first consider $m = 2$. If there are only two words in $\mathcal{W}$, then they must differ on at least one position. We choose it to discriminate, and hence $t_{n,2} = 1$. Thus, in our first modification we set the initial conditions as follows:

$$t_{n,0} \equiv t_{n,1} \equiv 0, \qquad \text{and} \qquad t_{n,2} \equiv 1, \tag{10}$$

while the recursion holds for $m \geq 3$.

In the second modification we also construct an optimal tree for $m = 3$. In this case these three words must differ on at least two positions. We use this information to build an optimal tree in which one word is stored at depth 1 and two others at depth 2 giving the average typical depth $t_{n,3} = 5/3$. In summary, we adopt the following new initial conditions:

$$t_{n,0} \equiv t_{n,1} \equiv 0, \qquad \text{and} \qquad t_{n,2} \equiv 1, \qquad t_{n,3} \equiv \frac{5}{3}. \tag{11}$$

We summarize below our findings which are proved in Section 3.

Theorem 3 (i) *For all $m \geq 3$ and $n \geq 1$ recurrence (2) with the initial conditions (10) possesses the following exact solution*

$$t_{n,m} = n - \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-1} - \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-2} (2^{l-1}-1). \qquad (12)$$

For $n \to \infty$ and large m the average search time $t_{n,m}$ attains the following asymptotics

$$t_{n,m} = \log_2 m + \frac{\gamma}{\ln 2} + \frac{1}{2} - \frac{1}{2\ln 2} + Q_1(\log_2 m) + O\Big(\frac{1}{m}\Big) + O(n2^{-n}) \qquad (13)$$

where

$$Q_1(\log_2 m) = \frac{1}{\ln 2} \sum_{\ell \in \mathbb{Z}\setminus\{0\}} \left(\Gamma\Big(\frac{2\pi i\ell}{\ln 2}\Big) - \frac{1}{2}\Gamma\Big(1+\frac{2\pi i\ell}{\ln 2}\Big)\right) e^{-2\pi i\ell \log_2 m}$$

is a periodic function of small magnitude ($\leq 2\cdot 10^{-5}$).

(ii)

If the initial condition (10) is replaced by (11) we obtain

$$\begin{aligned} t_{n,m} = n &- \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-1} - \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-2} (2^{l-1}-1) \\ &- \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-3} (2^{l-1}-1)(2^{l-2}-1). \end{aligned} \qquad (14)$$

and as $n \to \infty$ for large m

$$t_{n,m} = \log_2 m + \frac{\gamma}{\ln 2} + \frac{1}{2} - \frac{1}{2\ln 2} - \frac{1}{8\ln 2} + Q_2(\log_2 m) + O\Big(\frac{1}{m}\Big) + O(n2^{-n}) \qquad (15)$$

where

$$Q_2(\log_2 m) = \frac{1}{\ln 2} \sum_{\ell \in \mathbb{Z}\setminus\{0\}} \left(\Gamma\Big(\frac{2\pi i\ell}{\ln 2}\Big) - \frac{1}{2}\Gamma\Big(1+\frac{2\pi i\ell}{\ln 2}\Big) - \frac{1}{8}\Gamma\Big(2+\frac{2\pi i\ell}{\ln 2}\Big)\right) e^{-2\pi i\ell \log_2 m}$$

is a periodic function of small magnitude ($\leq 3\cdot 10^{-5}$).

Observe that for these modifications the average redundancies are respectively equal to $R_m = \frac{\gamma}{\ln 2} + \frac{1}{2} - \frac{1}{2\ln 2} + Q_1(\log_2 m) \approx 0.6113986565$ and $R_m = \frac{\gamma}{\ln 2} + \frac{1}{2} - \frac{1}{2\ln 2} - \frac{1}{8\ln 2} + Q_2(\log_2 m) \approx 0.4310617764$. It seems that Fedotov and Ryabko [3] analyzed the second modification and obtained the 1.04 bound. Our analysis shows that in fact the average redundancy in this case is 0.4310617764 plus a small fluctuation.

Finally, we consider our last randomized search algorithm that is based on PATRICIA tries (cf. [7, 11]). In PATRICIA tries all unary branchings are compressed to a single node; thus all branching nodes are binary (cf. Figure 4). One expects that such an algorithm

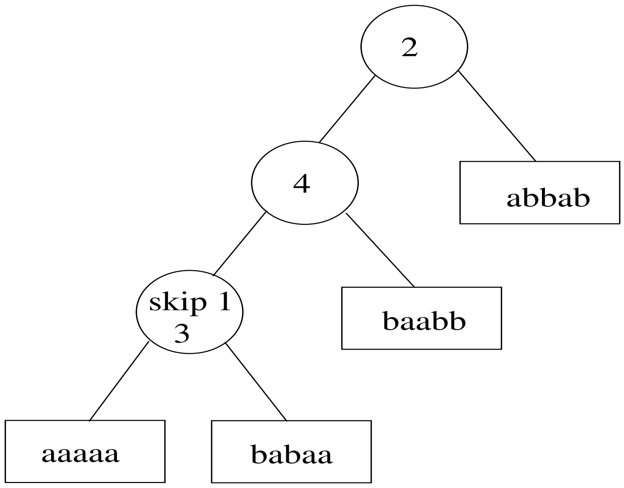

Figure 4: PATRICIA tries of the second trie in Figure 2.

will produce a low redundancy. To analyze it we first need to derive the recurrence for the uniform model. This turns out to be quite easy and we arrive at

$$t^P_{n,m} = 1 - 2\frac{\binom{2^{n-1}}{m}}{\binom{2^n}{m}} + \sum_{k=0}^{m} \frac{\binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k}}{\binom{2^n}{m}} \left(\frac{k}{m} t^P_{n-1,k} + \frac{m-k}{m} t^P_{n-1,m-k} \right), \qquad m \geq 2 \tag{16}$$

with the initial conditions

$$t^P_{n,0} \equiv t^P_{n,1} \equiv 0, \qquad n \geq 1. \tag{17}$$

The additional term $2\frac{\binom{2^{n-1}}{m}}{\binom{2^n}{m}}$ is responsible for avoiding empty left resp. right subtrees.

Using similar computations as above we prove in Section 3 our last result.

Theorem 4 [The PATRICIA trie model.] *For all $m \geq 2$ and $n \geq 1$ recurrence (16) with the initial condition (17) possesses the following exact solution*

$$t^P_{n,m} = n - \binom{2^n - 1}{m - 1}^{-1} \sum_{l=0}^{n-1} \binom{2^n - 2^l - 1}{m-1}. \tag{18}$$

For $n \to \infty$ and large m the average search time $t^P_{n,m}$ attains the following asymptotics

$$t^P_{n,m} = \log_2 m + \frac{\gamma}{\ln 2} - \frac{1}{2} + Q_0(\log_2 m) + O\Big(\frac{1}{m}\Big) + O(n2^{-n}) \tag{19}$$

where $Q_0(\log_2 m)$ is defined in (5).

We observe that the average redundancy $R^P_m = \frac{\gamma}{\ln 2} - \frac{1}{2} + Q_0(\log_2 m) \approx 0.332746177$ is the lowest among the four algorithms discussed so far. In addition, the number of internal nodes is $m - 1$ as in the optimal tree.

In summary, we proposed four randomized algorithms for finding the optimal search tree. The randomized PATRICIA algorithm is within 0.332746177 from the optimal search algorithm which is NP-hard. All proposed randomized algorithms work in $O(n^2)$ worst-case time and $\mathcal{O}(m \log m)$ average-case time.

3 Analysis

In this section we prove Theorems 1–4. We explain in depth the applied methodology when deriving Theorem 1, leaving details of the other proofs to the reader.

3.1 Regular Tries: Proof of Theorems 1 and 2

We repeat here recurrence (2) in the following form

$$t_{n,m} = 1 + 2\sum_{k=0}^{m} \frac{\binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k}}{\binom{2^n}{m}} \frac{k}{m} t_{n-1,k}.$$

This recursion holds only for $m \geq 2$, with $t_{n,0} \equiv t_{n,1} \equiv 0$ as in (3). We rewrite it as

$$m\binom{2^n}{m} t_{n,m} = m\binom{2^n}{m} + 2\sum_{k=0}^{m} \binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k} k t_{n-1,k}. \tag{20}$$

In order to solve it, we define the family of generating functions

$$T_n(x) := \sum_{m\geq 0} m\binom{2^n}{m} t_{n,m} x^m. \tag{21}$$

Multiplying (20) by x^m and summing we obtain

$$\frac{T_n(x)}{2^n(1+x)^{2^n}} = \frac{x}{1+x} - \frac{x}{(1+x)^{2^n}} + \frac{T_{n-1}(x)}{2^{n-1}(1+x)^{2^{n-1}}}.$$

This can be iterated

$$\frac{T_n(x)}{2^n(1+x)^{2^n}} = n\frac{x}{1+x} - \sum_{l=1}^{n} \frac{x}{(1+x)^{2^l}}$$

to yield the final formula

$$T_n(x) = n2^n x(1+x)^{2^n-1} - 2^n x \sum_{l=1}^{n} (1+x)^{2^n-2^l}. \tag{22}$$

Let $[x^n]f(x)$ stand for the coefficient of x^n in the series expansion of $f(x)$. Reading off the coefficients of (22) we find (cf. [9, 11])

$$\begin{aligned}
[x^m]T_n(x) &= n2^n[x^{m-1}](1+x)^{2^n-1} - 2^n[x^{m-1}]\sum_{l=1}^{n}(1+x)^{2^n-2^l} \\
&= n2^n\binom{2^n-1}{m-1} - 2^n\sum_{l=1}^{n}\binom{2^n-2^l}{m-1},
\end{aligned}$$

which finally implies

$$\begin{aligned} t_{n,m} &= n\frac{2^n\binom{2^n-1}{m-1}}{m\binom{2^n}{m}} - \frac{2^n}{m\binom{2^n}{m}}\sum_{l=1}^{n}\binom{2^n-2^l}{m-1} \\ &= n - \binom{2^n-1}{m-1}^{-1}\sum_{l=1}^{n}\binom{2^n-2^l}{m-1}; \end{aligned} \tag{23}$$

the formula holds for $m \geq 1$. This proves (4) of Theorem 1.

To derive an asymptotic expansion of $t_{n,m}$ we first observe that by the combinatorial origin of the problem the uniform model should converge as $n \to \infty$ to the unbiased memoryless model that has been studied for the last thirty years. We cite [7, 9, 11] where one finds

$$t_{\infty,m} = \sum_{k=1}^{m-1}\binom{m-1}{k}\frac{(-1)^{k-1}}{1-2^{-k}} = \sum_{l=0}^{\infty}\left[1-\left(1-2^{-l}\right)^{m-1}\right].$$

We shall provide an independent proof of the above.

We now fix m and allow $n \to \infty$. Using the asymptotics $\binom{N}{k} = \frac{N^k}{k!}(1+O(1/N))$ for fixed k and $N \to \infty$ we obtain the following chain of implications provided $m \ll 2^{n-1}$:

$$\begin{aligned} t_{n,m} &= \sum_{l=1}^{n}\left[1-\frac{\binom{2^n-2^l}{m-1}}{\binom{2^n-1}{m-1}}\right] \\ &= \sum_{l=1}^{n}\left[1-\left(\frac{2^n-2^l}{2^n-1}\right)^{m-1}\right](1+O(2^{1-n})) \\ &= \sum_{l=1}^{n}\left[1-\left(1-2^{l-n}\right)^{m-1}\right](1+O(2^{1-n})) \\ &= \sum_{l=0}^{n-1}\left[1-\left(1-2^{-l}\right)^{m-1}\right](1+O(2^{1-n})) \\ &= \sum_{l=0}^{\infty}\left[1-\left(1-2^{-l}\right)^{m-1}\right]+O(n2^{1-n}). \end{aligned}$$

In view of this we must study the asymptotics of the above sum; that we do next.

Define

$$B_m = \sum_{l=0}^{\infty}\left[1-\left(1-2^{-l}\right)^{m}\right]. \tag{24}$$

Let $y = m2^{-l}$. As long as $y < \ln m$ (or $l > \ln(m/\ln(m))$) we have $(1-y/m)^m = e^{-y}(1+O(y^2/m))$ which yields

$$\begin{aligned} B_m &= \sum_{l\leq\ln(m/\ln(m))}\left[1-\left(1-2^{-l}\right)^{m}\right] + \sum_{l>\ln(m/\ln(m))}\left[1-\left(1-2^{-l}\right)^{m}\right] \\ &= \sum_{l=0}^{\infty}\left(1-e^{-m2^{-l}}\right)+O\left(\frac{1}{m}\right). \end{aligned}$$

The last sum above is an ideal candidate for the Mellin transform. The main properties of the Mellin transform are discussed in Appendix A (cf. also [4, 11]). Let

$$B(x) = \sum_{l=0}^{\infty} \left(1 - e^{-x2^{-l}}\right).$$

We are interested in the asymptotic expansion of $S(x)$ as $x = m \to \infty$, where x is a positive real number. Observe that under property (M3) from Appendix A the function $S(x)$ falls under the *harmonic sum* paradigm and then the Mellin transform $S^*(s)$ of $S(x)$ is

$$B^*(s) = -\frac{\Gamma(s)}{1-2^s}, \qquad -1 < \Re(s) < 0,$$

where $\Gamma(s)$ is the Euler gamma function. By property (M1) we can recover $S(x)$ as

$$B(x) = \frac{1}{2\pi i} \int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} B^*(s) x^{-s} ds.$$

The evaluation of this integral is quite standard (cf. [7, 11]): we extend the line of integration to a big rectangle right to the integration line, and observe that bottom and top lines contribute negligible quantities because the gamma function decreases exponentially with the increase in the magnitude of the imaginary part, and the right side positioned at, say M, contributes x^{-M} for $M \to \infty$. Thus the main contribution to the integral comes from the residues right to the line of the integration. Observe that there is one double pole at $s = 0$ and simple poles at $\chi_k = 2\pi i k/\ln 2$. Applying property (M4) we directly obtain for $x \to \infty$ (cf. [11, pp.468])

$$B(x) = \log_2 x + \frac{\gamma}{\ln 2} + \frac{1}{2} + Q_0(\log_2 x) + O(x^{-M})$$

where $Q_0(\log_2 x)$ is defined in (6) and $M > 0$ is an arbitrary positive number. This proves (5), and completes the proof of Theorem 1.

Remark. There are many ways to derive a full asymptotic expansion of S_m. A relatively recent one, which can deal with much more complicated situations, is the analytic depoissonization technique (cf. [6]). It is described in [11], resulting in

$$\begin{aligned} B_m &= \log_2 m + \frac{\gamma}{\log 2} + \frac{1}{2} + Q_0(\log_2 m) \\ &+ \sum_{k=1}^{N} \sum_{i=1}^{k} (-1)^{k+i+1} b_{i,k+i} n^{-k} P_{k+i}(\log_2 m) + O(m^{-N-1} \log m) \end{aligned}$$

for any $N \geq 1$, where b_{ij} are the coefficients of $\exp(x \log(1+y) - xy)$ at $x^i y^j$, that is,

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_{ij} x^i y^j = \exp(x \log(1+y) - xy)$$

and

$$P_j(\log_2 m) = \frac{1}{\ln 2} \sum_{\ell=-\infty}^{\infty} \Gamma\Big(j + \frac{2\pi i\ell}{\ln 2}\Big) e^{-2\pi i\ell \log_2 m}$$

for $j \geq 1$.

Now we sketch the proof of Theorem 2. We start with recurrence (7). After introducing the generating function $S_n(x) := \sum_{m\geq 0} \binom{2^n}{m} s_{n,m} x^m$, we obtain from the recurrence

$$S_n(x) = (1+x)^{2^n} - 1 - x2^n + 2(1+x)^{2^{n-1}} S_{n-1}(x),$$

which can be solved by iteration leading to

$$S_n(x) = (2^n - 1)(1+x)^{2^n} - \sum_{l=1}^{n} 2^{n-l}(1+x)^{2^n - 2^l} - 2^n x \sum_{l=1}^{n} (1+x)^{2^n - 2^l}.$$

Computing the coefficients

$$[x^m]S_n(x) = (2^n - 1)\binom{2^n}{m} - \sum_{l=1}^{n} 2^{n-l}\binom{2^n - 2^l}{m} - 2^n \sum_{l=1}^{n} \binom{2^n - 2^l}{m-1}$$

we finally prove (8) of Theorem 2.

Next we consider the asymptotics. For $n \to \infty$ we have

$$\begin{aligned}
s_{n,m} &= \sum_{l=1}^{n} \left[2^{l-1} - 2^{n-l} \frac{\binom{2^n - 2^l}{m}}{\binom{2^n}{m}} - m \frac{\binom{2^n - 2^l}{m-1}}{\binom{2^n - 1}{m-1}} \right] \\
&\sim \lim_{n\to\infty} \sum_{l=1}^{n} \left[2^{l-1} - 2^{n-l}(1 - 2^{l-n})^m - m(1 - 2^{l-n})^{m-1} \right] \\
&\sim \lim_{n\to\infty} \sum_{l=0}^{n-1} \left[2^l - 2^l(1 - 2^{-l})^m - m(1 - 2^{-l})^{m-1} \right] \\
&\sim \sum_{l\geq 0} 2^l \left[1 - (1 - 2^{-l})^m - \frac{m}{2^l}(1 - 2^{-l})^{m-1} \right].
\end{aligned}$$

The error term is again $O(n2^{1-n})$. Let now $m \to \infty$. Then, by the same arguments as above

$$s_{n,m} = \sum_{l\geq 0} 2^l \left[1 - e^{-m/2^l} - \frac{m}{2^l} e^{-m/2^l} \right] + O\Big(\frac{1}{m}\Big).$$

We evaluate

$$S(x) = \sum_{l\geq 0} 2^l \left[1 - e^{-x/2^l} - \frac{m}{2^l} e^{-x/2^l} \right]$$

for $x \to \infty$ using the Mellin transform. However, we must be careful since both terms of the above sum have Mellin transforms with *non-overlapping* strips of convergence. We circumvent this by adding and subtracting $x2^{-l}$. Then using property (M3) of Appendix A we find

$$
\begin{aligned}
S^*(s) = \mathcal{M}(S(x);s) &= \mathcal{M}\Big(\sum_{l\geq 0} 2^l \left[1 - e^{-x/2^l} - x2^{-l} + x2^{-l} - \frac{m}{2^l} e^{-m/2^l}\right]; s\Big) \\
&= \mathcal{M}\Big(\sum_{l\geq 0} 2^l \left[1 - e^{-x/2^l} - x2^{-l}\right]; s\Big) - \mathcal{M}\Big(\sum_{l\geq 0} 2^{-l} x 2^l \left[e^{-x/2^l} - 1\right]; s\Big) \\
&= -\frac{\Gamma(s)(s+1)s}{1-2^{s+1}}, \qquad\qquad -2 < \Re(s) < -1.
\end{aligned}
$$

The rest follows the footsteps of the previous analysis. After observing that we really have only simple poles at $s_k = -1 + 2\pi i k/\ln 2$, computing the residues and using property (M4) of Appendix A, we complete the proof of Theorem 2.

3.2 Modified Tries: Proof of Theorem 3

We now provide a brief explanation how to extend our analysis from the previous section to prove Theorem 3.

We first assume that (10) is satisfied, that is, $t_{n,0} \equiv t_{n,1} \equiv 0$, $t_{n,2} \equiv 1$ while the recurrence (20) holds for $m \geq 3$. Defining the generating function $T_n(x)$ as in (21) and summing up (20) after multiplying it by x^m we finally obtain

$$
\frac{T_n(x)}{2^n(1+x)^{2^n}} = \frac{x}{1+x} - \frac{x}{(1+x)^{2^n}} - \frac{x^2(2^{n-1}-1)}{(1+x)^{2^n}} + \frac{T_{n-1}(x)}{2^{n-1}(1+x)^{2^{n-1}}}.
$$

Iterating it we arrive at

$$
T_n(x) = n2^n x(1+x)^{2^n-1} - 2^n x \sum_{l=1}^{n} (1+x)^{2^n-2^l} - 2^n x^2 \sum_{l=1}^{n} (1+x)^{2^n-2^l}(2^{l-1}-1),
$$

which yields after reading off the coefficients

$$
\begin{aligned}
[x^m]T_n(x) &= n2^n[x^{m-1}](1+x)^{2^n-1} - 2^n[x^{m-1}] \sum_{l=1}^{n} (1+x)^{2^n-2^l} \\
&\quad - 2^n[x^{m-2}] \sum_{l=1}^{n} (1+x)^{2^n-2^l}(2^{l-1}-1) \\
&= n2^n \binom{2^n-1}{m-1} - 2^n \sum_{l=1}^{n} \binom{2^n-2^l}{m-1} - 2^n \sum_{l=1}^{n} \binom{2^n-2^l}{m-2}(2^{l-1}-1)
\end{aligned}
$$

Thus the average depth $t_{n,m}$ becomes

$$
t_{n,m} = n - \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-1} - \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-2}(2^{l-1}-1);
$$

the formula holds for $m \geq 2$.

To extract asymptotics for $n \to \infty$ and large m we must study the term

$$d_{n,m} = \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-2} (2^{l-1}-1).$$

As in the previous section we find as $n \to \infty$

$$\begin{aligned}
\binom{2^n-1}{m-1}^{-1} \sum_{l=0}^{n} \binom{2^n-2^l}{m-2} (2^{l-1}-1) &\sim \frac{m-1}{2^n} \sum_{l=0}^{n} \left(1-2^{l-n}\right)^{m-2} (2^{l-1}-1) \\
&\sim \frac{m-1}{2} \sum_{l=0}^{n} 2^{l-n} \left(1-2^{l-n}\right)^{m-2} \\
&\sim \frac{m-1}{2} \sum_{l=0}^{n} 2^{l-n} \left(1-2^{l-n}\right)^{m-2} \\
&\sim \frac{m-1}{2} \sum_{l\geq 0} 2^{-l} \left(1-2^{-l}\right)^{m-2}.
\end{aligned}$$

Therefore, we must explore asymptotics of

$$\frac{m}{2} \sum_{l\geq 0} 2^{-l} e^{-m/2^l} = \frac{1}{2} \sum_{l\geq 0} \frac{m}{2^l} e^{-m/2^l}$$

as $m \to \infty$. We again apply the Mellin transform to

$$D(x) = \frac{1}{2} \sum_{l\geq 0} \frac{x}{2^l} e^{-x/2^l}.$$

Since $D(x)$ is a harmonic sum, property (M3) of Appendix A implies that the Mellin transform $D^*(s)$ of $D(x)$ is

$$D^*(s) = \frac{1}{2} \frac{\Gamma(s+1)}{1-2^s}, \qquad -1 < \Re(s) < 0.$$

This function has simple poles at $\chi_k = 2\pi i k/\ln 2$. Computing residues and applying property (M4) from Appendix A we finally find for any $M > 0$

$$D(x) = \frac{1}{2\ln 2} + \frac{1}{2} P_1(\log_2 x) + O(x^{-M}),$$

where

$$P_1(\log_2 x) = \frac{1}{\ln 2} \sum_{\ell \in \mathbb{Z}\setminus\{0\}} \Gamma\Big(1 + \frac{2\pi i \ell}{\ln 2}\Big) e^{-2\pi i \ell \log_2 x}.$$

This proves part (i) of Theorem 3.

We only sketch the proof of Theorem 3(ii). We assume that $t_{n,0} \equiv t_{n,1} \equiv 0$, $t_{n,2} \equiv 1$, $t_{n,3} \equiv \frac{5}{3}$ and the recurrence (20) holds for $m \geq 4$. A similar computation as before leads to

$$\begin{aligned}
\frac{T_n(x)}{2^n(1+x)^{2^n}} = \frac{x}{1+x} - \frac{x}{(1+x)^{2^n}} - \frac{x^2(2^{n-1}-1)}{(1+x)^{2^n}} - \frac{x^3(2^{n-1}-1)(2^{n-2}-1)}{(1+x)^{2^n}} \\
+ \frac{T_{n-1}(x)}{2^{n-1}(1+x)^{2^{n-1}}}.
\end{aligned}$$

Iterating and reading off coefficients, we find

$$t_{n,m} = n - \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-1} - \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-2} (2^{l-1}-1)$$
$$- \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-3} (2^{l-1}-1)(2^{l-2}-1);$$

the formula holds for $m \geq 3$. We are interested to see what happens to the extra term

$$h_{n,m} = \binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-3} (2^{l-1}-1)(2^{l-2}-1)$$

in the limit $n \to \infty$. We now proceed informally as follows:

$$\begin{aligned}
&\binom{2^n-1}{m-1}^{-1} \sum_{l=1}^{n} \binom{2^n-2^l}{m-3} (2^{l-1}-1)(2^{l-2}-1) \\
&\qquad \sim \frac{(m-1)(m-2)}{4^n} \sum_{l=1}^{n} \left(1-2^{l-n}\right)^{m-3} (2^{l-1}-1)(2^{l-2}-1) \\
&\qquad \sim \frac{(m-1)(m-2)}{8} \sum_{l=0}^{n} \left(1-2^{l-n}\right)^{m-3} 4^{l-n} \\
&\qquad \sim \frac{(m-1)(m-2)}{8} \sum_{l\geq 0} \left(1-2^{-l}\right)^{m-3} 4^{-l}.
\end{aligned}$$

We are interested in the behavior of the above as $m \to \infty$. It can be approximated as (with error term $O(1/m)$):

$$\frac{m^2}{8} \sum_{l\geq 0} 4^{-l} e^{-m/2^l} = \frac{1}{8} \sum_{l\geq 0} \left(\frac{m}{2^l}\right)^2 e^{-m/2^l}.$$

To evaluate the last sum asymptotically we introduce

$$H(x) = \frac{1}{8} \sum_{l\geq 0} \left(\frac{x}{2^l}\right)^2 e^{-x/2^l}$$

whose Mellin transform is

$$H^*(s) = \frac{1}{8} \frac{\Gamma(s+2)}{1-2^s}, \qquad -2 < \Re(s) < 0.$$

Following the footsteps of our previous arguments we obtain for any $M > 0$

$$H(x) = \frac{1}{8\ln 2} + \frac{1}{8} Q_2(\log_2 x) + O(x^{-M}),$$

where

$$Q_2(\log_2 x) = \frac{1}{\ln 2} \sum_{\ell \in \mathbb{Z}\setminus\{0\}} \Gamma\Big(2 + \frac{2\pi i \ell}{\ln 2}\Big) e^{-2\pi i \ell \log_2 x}.$$

This proves Theorem 3.

3.3 PATRICIA Tries: Proof of Theorem 4

We now prove Theorem 4. Let us start with recurrence (16) which we rewrite as follows:

$$m\binom{2^n}{m}t^P_{n,m} = m\binom{2^n}{m} - 2m\binom{2^{n-1}}{m} + 2\sum_{k=0}^{m}\binom{2^{n-1}}{k}\binom{2^{n-1}}{m-k}kt^P_{n-1,k}.$$

It holds now for all $m \geq 0$. The generating function $T^P_n(x) = \sum_{m\geq 0} m\binom{2^n}{m}t^P_{n,m}x^m$ becomes

$$T^P_n(x) = x2^n(1+x)^{2^n-1} - x2^n(1+x)^{2^{n-1}-1} + 2(1+x)^{2^{n-1}}T^P_{n-1}(x),$$

which can be represented as below

$$\begin{aligned}\frac{T^P_n(x)}{2^n(1+x)^{2^n}} &= \frac{x}{1+x} - \frac{x}{(1+x)^{2^{n-1}+1}} + \frac{T^P_{n-1}(x)}{2^{n-1}(1+x)^{2^{n-1}}}\\ &= n\frac{x}{1+x} - \sum_{l=0}^{n-1}\frac{x}{(1+x)^{2^l+1}}.\end{aligned}$$

This implies, after iterating the above formula,

$$T^P_n(x) = n2^nx(1+x)^{2^n-1} - 2^nx\sum_{l=0}^{n-1}(1+x)^{2^n-2^l-1}.$$

After reading off the coefficients we obtain

$$[x^m]T^P_n(x) = n2^n\binom{2^n-1}{m-1} - 2^n\sum_{l=0}^{n-1}\binom{2^n-2^l-1}{m-1}$$

and this finally yields the average $t^P_{n,m}$

$$t^P_{n,m} = n - \binom{2^n-1}{m-1}^{-1}\sum_{l=0}^{n-1}\binom{2^n-2^l-1}{m-1}.$$

Unlike for the previous cases, in order to find the asymptotics, we express $t^P_{n,m}$ in terms of the average search time $t_{n,m}$ for tries with the initial conditions (3). We proceed as follows:

$$\begin{aligned}t^P_{n,m} &= n - \binom{2^n-1}{m-1}^{-1}\sum_{l=0}^{n-1}\binom{2^n-2^l-1}{m-1}\\ &= n - \binom{2^n-1}{m-1}^{-1}\sum_{l=0}^{n-1}\frac{2^n-2^l-(m-1)}{2^n-2^l}\binom{2^n-2^l}{m-1}\\ &= n - \binom{2^n-1}{m-1}^{-1}\sum_{l=0}^{n-1}\binom{2^n-2^l}{m-1} + (m-1)\binom{2^n-1}{m-1}^{-1}\sum_{l=0}^{n-1}\binom{2^n-2^l}{m-1}\frac{1}{2^n-2^l}\\ &= \sum_{l=1}^{n}\left(1-\frac{\binom{2^n-2^l}{m-1}}{\binom{2^n-1}{m-1}}\right) - 1 + \frac{m-1}{2^{n-1}}\sum_{l=0}^{n-1}\frac{\binom{2^n-2^l}{m-1}}{\binom{2^n-1}{m-1}}\frac{1}{2-2^{l-n+1}}\\ &= t_{n,m} - 1 + O(m2^{-n}\log m).\end{aligned}$$

This proves Theorem 4.

APPENDIX: Properties of the Mellin Transform

For the readers convenience, we collected here the main properties of the Mellin transform. For details and proofs see [4, 11].

(M1) Direct and Inverse Mellin Transforms. Let c belong to the *fundamental strip* defined below.

$$f^*(s) := \mathcal{M}(f(x); s) = \int_0^\infty f(x) x^{s-1} dx \qquad \Longleftrightarrow \qquad f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} f^*(s) x^{-s} ds. \tag{A.1}$$

(M2) Fundamental Strip. The Mellin transform of $f(x)$ exists in the *fundamental strip* $\Re(s) \in (-\alpha, -\beta)$, where

$$f(x) = O(x^\alpha) \qquad (x \to 0), \qquad\qquad f(x) = O(x^\beta) \qquad (x \to \infty)$$

for $\beta < \alpha$.

(M3) Harmonic Sum Property. By linearity and the scale rule $\mathcal{M}(f(ax); s) = a^{-s}\mathcal{M}(f(x); s)$,

$$f(x) = \sum_{k \geq 0} \lambda_k g(\mu_k x) \qquad \Longleftrightarrow \qquad f^*(s) = g^*(s) \sum_{k \geq 0} \lambda_k \mu_k^{-s}. \tag{A.2}$$

(M4) Mapping Properties (Asymptotic expansion of $f(x)$ and singularities of $f^*(s)$).

$$f(x) = \sum_{(\xi,k) \in A} c_{\xi,k} x^\xi (\log x)^k + O(x^M) \qquad \Longleftrightarrow \qquad f^*(s) \asymp \sum_{(\xi,k) \in A} c_{\xi,k} \frac{(-1)^k k!}{(s+\xi)^k + 1}. \tag{A.3}$$

— (i) *Direct Mapping.* Assume that $f(x)$ admits as $x \to 0^+$ the asymptotic expansion (A.3) for some $-M < -\alpha$ and $k > 0$. Then for $\Re(s) \in (-M, -\beta)$, the transform $f^*(s)$ satisfies the singular expansion (A.3)

— (ii) *Reverse Mapping.* Assume that $f^*(s) = O(|s|^{-r})$ with $r > 1$, as $|s| \to \infty$ and that $f^*(s)$ admits the singular expansion (A.3) for $\Re(s) \in (-M, -\beta)$. Then $f(x)$ satisfies the asymptotic expansion (A.3) at $x = 0^+$.

References

[1] R. Ahlswede and I. Wegener, *Search Problems*, John Wiley & Sons, Chichester, 1987.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[3] A. Fedotov and B. Ryabko, The Estimated Cost of a Search Tree on Binary Words, *IEEE Transactions on Information Theory*, 47, 326–329, 2001.

[4] P. Flajolet, X. Gourdon, and P. Dumas, Mellin Transforms and Asymptotics: Harmonic sums, *Theoretical Computer Science*, 144, 3–58, 1995.

[5] M. Garey and D. Johnson, *Computers and Intractability*, Freeman, 1979.

[6] P. Jacquet and W. Szpankowski, Analytical Depoissonization Lemma and Its Applications, *Theoretical Computer Science*, 201, 1–62, 1998.

[7] D. Knuth, *The Art of Computer Programming. Sorting and Searching.* Vol. 3. (second edition), Addison-Wesley, 1998.

[8] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

[9] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, 1996.

[10] W. Szpankowski, Asymptotic Redundancy of Huffman (and Other) Block Codes, *IEEE Trans. Information Theory*, 46, 2434–2443, 2000.

[11] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.