The variance for partial match retrievals in k-dimensional bucket digital trees

Michael Fuchs

The variance for partial match retrievals in k-dimensional bucket digital trees

Michael Fuchs

Department of Applied Mathematics, National Chiao Tung University, Hsinchu, 300, Taiwan

The variance of partial match queries in k-dimensional tries was investigated in a couple of papers in the mid-nineties, the resulting analysis being long and complicated. In this paper, we are going to re-derive these results with a much easier approach. Moreover, our approach works for k-dimensional PATRICIA tries, k-dimensional digital search trees and bucket versions as well.

Keywords: k-dimensional digital trees, partial match retrieval, variance, JS-admissibility, Mellin transform

1 Introduction and Results

Data structures for storing and retrieving multidimensional data are of vital importance in several areas of computer science such as design of data base systems and graphics algorithms. One possible class of such data structures was introduced in [10] and is based on digital data, i.e., data which is composed of infinite 0-1 strings. We assume throughout this work that every bit in these strings is generated independently and with the same probability (symmetric Bernoulli model).

We will first describe the above data structure in more details. Therefore, assume that we have given a set of multidimensional data. Then, we apply a " regular shuffling" procedure to transform the multidimensional data into one-dimensional data. Finally, this data is stored in a digital tree. To be more precise, let $R_1, \ldots, R_n$ denote k-dimensional records, i.e.,

$$R_{i,1} = \left(R_{i,1}^{[1]}, R_{i,1}^{[2]}, R_{i,1}^{[3]}, \ldots\right),$$
$$\vdots$$
$$R_{i,k} = \left(R_{i,k}^{[1]}, R_{i,k}^{[2]}, R_{i,k}^{[3]}, \ldots\right).$$

After shuffling, we obtain the one-dimensional string $\tilde{R}_i$

$$\tilde{R}_i = \left(R_{i,1}^{[1]}, \ldots, R_{i,k}^{[1]}, R_{i,1}^{[2]}, \ldots, R_{i,k}^{[2]}, R_{i,1}^{[3]}, \ldots, R_{i,k}^{[3]}, \ldots\right).$$

Then, $\tilde{R}_1, \ldots, \tilde{R}_n$ are used to construct a digital tree.

As digital trees, we use the three standard types (see [8]). First, for the *k-dimensional trie* the underlying digital tree is the trie data structure. For the readers convenience, we recall how a trie is constructed: if we only have one record, then we place it into the root which is considered an external node; if we have more than one record, then the root becomes an (empty) internal node and the records are either directed to the left or to right subtree according to whether their first bit is 0 or 1; finally, the subtrees are build recursively by the same procedure, where the first bits are removed; see Figure 1 for an example.

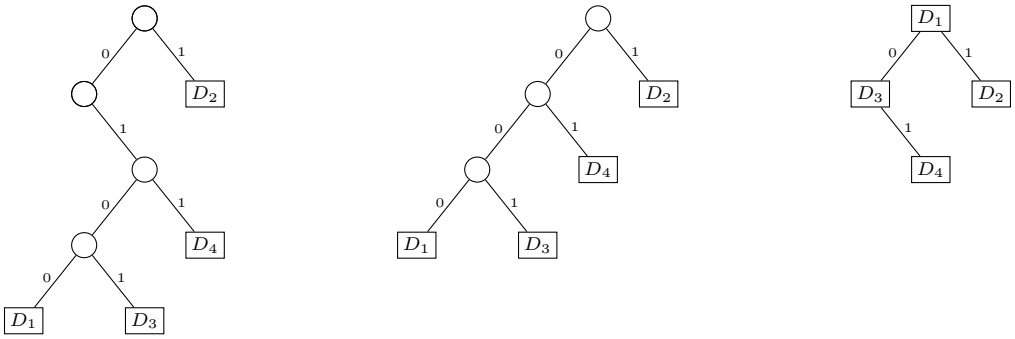

Fig. 1: A 2-dimensional trie, PATRICIA trie and digital search tree built from the data

data	D_1	D_2	D_3	D_4
$R_{i,1}$	$0010\cdots$	$1001\cdots$	$0001\cdots$	$0111\cdots$
$R_{i,2}$	$1000\cdots$	$1010\cdots$	$1101\cdots$	$1011\cdots$

A variant of k-dimensional tries are *k-dimensional PATRICIA tries*, where the shuffled data is stored in a PATRICIA trie. Recall that a PATRICIA trie is constructed as a trie only multiple one-way branching is suppressed; again see Figure 1 for an example. This yields a more balanced tree improving the overall performance of tries.

A final type is given by the *k-dimensional digital search tree* which is based on the digital search tree data structure. Recall that a digital search tree is constructed as follows: the first record is placed in the root; all other records are directed to the left or right subtree according to whether their first bit is 0 or 1; finally, the subtrees are constructed by the same principle again with the first bits removed; see Figure 1 for an example. So, in contrast to tries and PATRICIA tries, no distinction between internal and external nodes is necessary for digital search trees.

Note that all trees introduced above have the common feature that nodes only hold at most one record. If we allow nodes to hold up to $b \geq 1$ records, then the resulting trees are called *k-dimensional bucket digital trees*.

In this paper, we will study the cost of partial match retrievals in k-dimensional bucket digital trees. Here, a partial match query will ask for the retrieval of all records matching certain criteria. Formally, a partial match query is a k-dimensional vector $R = (R_1, \ldots, R_k)$ with some of its coordinates a string of

0-1 bits and others unspecified. For instance, for $k = 2$, we might have

$$R_1 = (\star, \star, \star, \star, \ldots),$$
$$R_2 = (0, 1, 1, 0, \ldots),$$

Then, a shuffled record $\tilde{R}$ is produced as before. For the above example this yields

$$\tilde{R} = (\star, 0, \star, 1, \star, 1, \star, 0, \ldots).$$

The partial match query asks now for the retrieval of all data in the tree with $\tilde{R}$ being used as search query. Here, 0 or 1 in $\tilde{R}$ means either going to the left or right subtree of the current node, whereas $\star$ means that we have to proceed with our search in both subtrees. The cost of such a partial match query will be measured by the number of nodes visited (where we only consider internal nodes for tries and PATRICIA tries); so for the query $\tilde{R}$ above, we have a cost of 2 for the trie and Patricia trie of Figure 1 and a cost of 3 for the digital search tree.

Before going on, we will fix some notation which we are going to use throughout the work. First, it should be clear that under our random model, the cost of a partial match query only depends on the *partial match pattern* q which is a k-tuple of symbols from $\{S, \star\}$, where the i-th coordinate is S if R_i is specified and $\star$ otherwise, We will fix such a q throughout this work and denote the number of $\star$ entries in q by u, where we assume that $0 < u < k$. Furthermore, we will consider cyclic shifts of the entries of q by one position to the left which will be denoted by q'; more generally, $q^{(l)}$ will denote the cyclic shift of the entries of q by l position to the left. Also, we will associate to a partial match pattern q an infinite sequence $(\delta_1, \delta_2, \delta_3, \ldots)$ with $\delta_i = 1$ if $q_{i \bmod k} = S$ and $\delta_i = 2$ otherwise. Finally, we denote the random variable describing the cost of the partial match query by $X_{q,n}$ for all three types of bucket digital trees of size n and bucket size $b \geq 1$ (for the sake of simplicity, we suppress the index b).

In this paper, we will be concerned with stochastic properties of $X_{q,n}$. Therefore, let us recall what is known about this random variable. First, the mean value of $X_{q,n}$ for k-dimensional tries was investigated in [2] where the authors proved that

$$\mathbb{E}(X_{q,n}) \sim n^{u/k} P_1(\log_2 n^{1/k})$$

with $P_1(z)$ a one-periodic function whose Fourier expansion was given in [2] as well (for a comparison of this result with other data structure for multidimensional search see [8] and [10]). Similar results were subsequently proved in [6] for k-dimensional bucket digital tries, k-dimensional PATRICIA tries, and k-dimensional digital search trees, too. As for the variance, it was conjectured in [7] that for k-dimensional tries

$$\mathrm{Var}(X_{q,n}) \sim n^{u/k} P_2(\log_2 n^{1/k}) \tag{1}$$

with $P_2(z)$ a one-periodic function. The authors proved this conjecture for $k = 2$ in [7]. The general case was then settled in [11]. Note that this result implies that $X_{q,n}/\mathbb{E}(X_{q,n})$ converges to 1 in probability. Hence, the distribution of $X_{q,n}$ is concentrated around its mean.

As for the method of proof of (1), the authors of [7] applied the analytic approach from [2] to derive asymptotic expansions of mean and second moment. Then, they used these expansions to compute the variance, where they had to cope with highly non-trivial cancellations. Here, their proof crucially rested on an identity of Ramanujan which only works in the case $k = 2$ and does not seem to have an analogue for $k > 2$. In [11], a new and mainly elementary approach was devised to settle the general case.

In this paper, we will re-derive the above result with a more simpler approach. Our approach will be analytic and use some standard tools from the analysis of algorithms (the same approach was already used in other contexts; see [4],[5],[12]). The crucial difference to the approach from [7] is that we incorporate the cancellations at a much earlier stage, making the resulting analysis more easier. Moreover, our approach will work for k-dimensional bucket tries as well.

Before explaining our approach in more details, we are going to state our result. Therefore, we need some notation. Set

$$\tilde{h}_q(z) = 2\delta_1 e_b(z)e^{-z}\tilde{L}_{q'}(z/2) + e_b(z)e^{-z} - e_b(z)^2e^{-2z},$$

where $e_b(z) = 1 + z + z^2/2! + \cdots + z^b/b!$ and $\tilde{L}_q(z) = \exp\{-z\}\sum_{n\geq 0}\mathbb{E}(X_{q,n})z^n/n!$.

Theorem 1 *The cost of a partial match query with u non-specified coordinates in a k-dimensional trie of size n satisfies*

$$\mathrm{Var}(X_{q,n}) = n^{u/k}P_2(\log_2 n^{1/k}) + \mathcal{O}(n^{2u/k-1})$$

with one-periodic function

$$P_2(z) = \sum_{r=-\infty}^{\infty} c_r e^{2\pi i r z}$$

and Fourier coefficients

$$c_r = \frac{1}{kL}\sum_{l=0}^{k-1}\delta_1\cdots\delta_l 2^{-\omega_r l}\int_0^{\infty} z^{-\omega_r-1}\tilde{h}_{q^{(l)}}(z)\mathrm{d}z,$$

where $\omega_r = u/k + 2\pi i r/(kL)$ and $L = \log 2$.

With slightly more work, the Fourier coefficients can be further simplified.

Corollary 1 *The Fourier coefficient in the above theorem can be expressed as*

$$c_r = \frac{\Gamma(-\omega_r)}{kL}\Bigg(\delta(2^{-\omega_r})\Bigg(\binom{-\omega_r+b}{b} - 2^{\omega_r}\sum_{j_1,j_2=0}^{b}\binom{j_1+j_2}{j_1}\binom{-\omega_r+j_1+j_2}{j_1+j_2}2^{-j_1-j_2}\Bigg) - \sum_{l\geq b+1}\binom{-l+b}{b}\binom{-\omega_r+l+b}{b}\binom{\omega_r}{l}\frac{2^{1-l}\sigma(2^{-\omega_r},2^{-l}))}{1-2^{-lk+u}}\Bigg), \tag{2}$$

where

$$\delta(z) = \sum_{j=0}^{k-1}\delta_1\cdots\delta_j z^j, \qquad \sigma(z_1,z_2) = \sum_{j_1,j_2=0}^{k-1}\delta_1\cdots\delta_{j_1+j_2+1}z_1^{j_1}z_2^{j_2}.$$

For instance, for $k = 2, s = 1$, and $q = (\star, S)$, the value of c_0 becomes

$$\frac{(1+\sqrt{2})\sqrt{\pi}}{2\ln 2}\left(\frac{7\sqrt{2}}{8} - 1 - 4\sqrt{2}\sum_{l\geq 2}\binom{1/2}{l}\frac{(l-1)(l+1/2)2^{-l}}{1-2^{-l+1/2}}\right) \approx 2.09184\cdots,$$

where the last approximation was computed with Maple. This value coincides with the one given in [7]. Note that the expression given in the latter paper is slightly different; we leave it as an exercise to the reader to show that they are the same.

Next, our approach can also be straightforwardly applied to k-dimensional bucket PATRICIA tries. Here, we have the same result as above only $\tilde{h}_q(z)$ replaced by

$$\begin{aligned}\tilde{h}_q(z) =& 2\delta_1(e_b(z)e^{-z} - \delta_1 e_b(z/2)e^{-z} + (\delta_1 - 1)e^{-z/2})\tilde{L}_{q'}(z/2)\\ &+ e_b(z)e^{-z} - \delta_1 e_b(z/2)e^{-z} + \delta_1 e^{-z/2} - (e_b(z)e^{-z} - \delta_1 e_b(z/2)e^{-z} + \delta_1 e^{-z/2})^2.\end{aligned}$$

Also, a similar explicit expression for the Fourier coefficients as in Corollary 1 can be given. Since, the resulting formula is more messy we do not give details.

Finally, k-dimensional bucket digital search trees are slightly more involved. Here, we will use a variant of the above approach which was introduced in [3]. In order to state our result, we again need some notation. Therefore, set

$$Q(s) = \prod_{j\geq 1}\left(1 - \frac{s}{2^j}\right), \qquad Q_l = \prod_{j=1}^{l}\left(1 - 2^{-j}\right)$$

and

$$\tilde{h}_q(z) = \left(\sum_{j=0}^{b}\binom{b}{j}\tilde{L}_q^{(j)}(z)\right)^2 - \sum_{j=0}^{b}\binom{b}{j}\left(\tilde{L}_q(z)^2\right)^{(j)}.$$

Theorem 2 *The cost of a partial match query with u non-specified coordinates in a k-dimensional digital search tree of size n satisfies*

$$\mathrm{Var}(X_{q,n}) = n^{u/k}P_2(\log_2 n^{1/k}) + \mathcal{O}(n^{2u/k-1})$$

with one-periodic function

$$P_2(z) = \sum_{r=-\infty}^{\infty} c_r e^{2\pi i r z}$$

and Fourier coefficients

$$c_r = \frac{1}{kL\Gamma(1+\omega_r)}\sum_{l=0}^{k-1}\delta_1\cdots\delta_l 2^{-\omega_r l}\int_0^\infty \frac{s^{\omega_r}}{Q(-2s)^b}\left(\int_0^\infty e^{-zs}\tilde{h}_{q^{(l)}}(z)\mathrm{d}z + p(s)\right)\mathrm{d}s, \tag{3}$$

where

$$p(s) = \frac{(1+s)^{b-1} + (-1)^b}{s+2}.$$

Moreover, the Fourier coefficients can be further simplified here, too. We will state the result for $b = 1$. Therefore, set

$$\varphi(\omega; x) = \begin{cases}\pi(x^\omega - 1)/(\sin(\pi\omega)(x-1)) & \text{if } x \neq 1;\\ \pi\omega/\sin(\pi\omega), & \text{if } x = 1.\end{cases}$$

Then, we have the following corollary.

Corollary 2 *If the bucket size equals one, then the Fourier coefficient in the above theorem can be expressed as*

$$c_r = \frac{1}{kLQ(1)\Gamma(1+\omega_r)} \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l 2^{-\omega_r l} \sum_{j_1,j_2,j_3 \geq 0} \frac{(-1)^{j_1} \bar{\delta}_{q^{(l)},j_2} \bar{\delta}_{q^{(l)},j_3} 2^{-\binom{j_1}{2}+(1-\omega_r)j_1}}{2^{j_2+j_3} Q_{j_1} Q_{j_2} Q_{j_3}} \varphi(\omega_r; 2^{j_1-j_2} + 2^{j_1-j_3}),$$

where

$$\bar{\delta}_{q,j} = \sum_{l \geq 0} \frac{(-1)^l 2^{-\binom{l+1}{2}}}{Q_l} \prod_{h=1}^{l+j} \delta_h.$$

We conclude the introduction by giving a short sketch of the paper. In the next section, we will treat k-dimensional bucket tries. Then, in Section 3, we will briefly discuss k-dimensional bucket PATRICIA tries. Finally, in Section 5, we will prove the results for k-dimensional bucket digital search trees.

2 k-dimensional Bucket Tries

Note that from the definition of k-dimensional bucket tries, we have

$$X_{q,n} \stackrel{d}{=} \begin{cases} X_{q',I_n} + X^*_{q',n-I_n} + 1, & \text{if } q = (\star, \ldots); \\ X_{q',I_n} + 1, & \text{if } q = (S, \ldots), \end{cases} \qquad (n \geq b+1),$$

where $I_n = \text{Binom}(n, 1/2)$, (X^*_n) is an independent copy of (X_n) with $X_n \stackrel{d}{=} X^*_n$, and $X_{q,0} = X_{q,1} = \cdots = X_{q,b} = 0$.

From this recurrence we will proceed as follows. First, we are going to apply the poissonization-depoissonization procedure from [5]. This will allow us to entirely focus on the Poisson model. Next, we will define a *poissonized variance* which is not really a variance, but asymptotically behaves like one (this idea was probably first used in [4]). This will be the crucial step leading to a much more simplified derivation. The remaining analysis is then carried out by using Mellin transform, a standard tool from the analysis of algorithm (for an excellent introduction see [1]).

Poissonization. Let $\tilde{P}_q(z,y)$ denote the Poisson generating function of $\mathbb{E}(\exp\{X_{q,n}y\})$, i.e.,

$$\tilde{P}_q(z,y) = e^{-z} \sum_{n \geq 0} \mathbb{E}(e^{X_n y}) \frac{z^n}{n!}.$$

Then, we obtain from the above distributional recurrence

$$\tilde{P}_q(z,y) = e^y \tilde{P}_{q'}(z/2,y)^{\delta_1} + e_b(z)e^{-z}(1-e^y).$$

Next, by taking first and second derivatives with respect to y and setting $y = 0$, we obtain the following functional equation for the Poisson generating function of the mean (denoted by $\tilde{L}_q(z)$)

$$\tilde{L}_q(z) = \delta_1 \tilde{L}_{q'}(z/2) + 1 - e_b(z)e^{-z}$$

and for the Poisson generating function of the second moment (denoted by $\tilde{M}_q(z)$)

$$\tilde{M}_q(z) = \begin{cases} 2\tilde{M}_{q'}(z/2) + 4\tilde{L}_{q'}(z/2) + 2\tilde{L}_{q'}(z/2)^2 + 1 - e_b(z)e^{-z}, & \text{if } q = (\star, \ldots); \\ \tilde{M}_{q'}(z/2) + 2\tilde{L}_{q'}(z/2) + 1 - e_b(z)e^{-z}, & \text{if } q = (S, \ldots). \end{cases}$$

Going from these Poisson generating functions back to the original quantity is done via the depoissonization tools from [5]. We will use here the language from [3], where we coined the term Jacquet-Szpankowski admissibility (or JS-admissibility for short). Recall that $\tilde{f}(z)$ is called JS-admissible if the following two conditions hold (where here and throughout this work, ϵ will denote a small constant whose value might change from one appearance to the next).

(I) There exists an $\alpha \in \mathbb{R}$ such that uniformly for $|\arg(z)| \leq \epsilon$

$$\tilde{f}(z) = \mathcal{O}\left(|z|^{\alpha}\right).$$

(O) We have, uniformly for $\epsilon \leq |\arg(z)| \leq \pi$,

$$f(z) := e^z \tilde{f}(z) = \mathcal{O}\left(e^{(1-\epsilon)|z|}\right).$$

The importance of this notation is due to the following proposition which is proved by a standard application of the saddle point method (see [5] for many more such results).

Proposition 1 *Let $\tilde{f}(z)$ be the Poisson generating function of f_n. If $\tilde{f}(z)$ is JS-admissibility, then*

$$f_n = \sum_{0 \leq j < 2l} \frac{\tilde{f}^{(j)}(n)}{j!} \tau_j(n) + \mathcal{O}(n^{\alpha - l})$$

with $\tau_j(n) = n![z^n](z-n)^j e^z$

In our context, JS-admissible is easily checked via the following result.

Proposition 2 *Assume that we have*

$$\tilde{f}_{q^{(l)}}(z) = \delta_{l+1} \tilde{f}_{q^{(l+1)}}(z/2) + \tilde{g}_{q^{(l)}}(z), \qquad (0 \leq l < k),$$

where all involved functions are entire. Moreover, assume that $\tilde{g}_{q^{(l)}}(z)$ is JS-admissible for $0 \leq l < k$. Then, $\tilde{f}_{q^{(l)}}(z)$ is JS-admissible for $0 \leq l < k$.

Proof: We only show how to prove **(I)**. Therefore, we start by iterating the recurrence. This yields

$$\tilde{f}_q(z) = 2^u \tilde{f}_q(z/2^k) + \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l \tilde{g}_{q^{(l)}}(z/2^l).$$

Now set

$$\tilde{B}_q(r) := \max_{|z|=r, |\arg(z)| \leq \epsilon} |\tilde{f}_q(z)|.$$

Then, by the assumptions, we obtain

$$\tilde{B}_q(r) \le 2^u \tilde{B}_z(r/2^k) + \mathcal{O}\left(r^\alpha\right).$$

Next, we define $\tilde{K}_q(r)$ by

$$\tilde{K}_q(r) = 2^u \tilde{K}_q(r/2^k) + \mathcal{O}\left(r^\alpha\right).$$

Then, $\tilde{B}_q(r) \le \tilde{K}_q(r)$. Moreover, we immediately obtain by iteration

$$\tilde{K}_q(r) = \begin{cases} r^\alpha, & \text{if } \alpha > u/k; \\ r^{u/k}\log r, & \text{if } \alpha = u/k; \\ r^{u/k}, & \text{if } \alpha < u/k. \end{cases}$$

This proves our claim. □

Using this result together with the closure properties from [3] proves that both $\tilde{L}_q(z)$ and $\tilde{M}_q(z)$ are JS-addmissible. Also, note that we have

$$\tilde{L}_q(z) = \mathcal{O}\left(|z|^{u/k}\right), \qquad \tilde{M}_q(z) = \mathcal{O}\left(|z|^{2u/k}\right) \tag{4}$$

uniformly as $|z| \to \infty$ and $|\arg(z)| \le \epsilon$.

Next, we define the poissonized variance as

$$\tilde{V}_q(z) = \tilde{M}_q(z) - \tilde{L}_q(z)^2.$$

Then, by a straightforward computation

$$\tilde{V}_q(z) = \delta_1 \tilde{V}_{q'}(z/2) + \tilde{h}_q(z),$$

where $\tilde{h}_q(z)$ was defined in the introduction.

Note that $\tilde{V}_q(z)$ is not the Poisson generating function of a variance but only mimicks the defintion of the variance. However, it behaves asymptotically like the variance as proved in the following proposition (see also Theorem 6 in [5]).

Proposition 3 *As $n \to \infty$,*

$$\mathrm{Var}(X_{q,n}) = \tilde{V}_q(n) + \mathcal{O}\left(n^{2u/k-1}\right).$$

Proof: From Proposition 2 and (4), we have

$$\begin{aligned} \mathrm{Var}(X_{q,n}) &= \mathbb{E}(X_{q,n}^2) - (\mathbb{E}(X_{q,n}))^2 \\ &= \tilde{M}_q(n) + \mathcal{O}\left(n^{2u/k-1}\right) - \left(\tilde{L}_q(n) + \mathcal{O}\left(n^{u/k-1}\right)\right)^2 \\ &= \tilde{V}_q(n) + \mathcal{O}\left(n^{2u/k-1}\right). \end{aligned}$$

This proves the claim. □

Asymptotic Expansion of $\tilde{L}_q(z)$. We will first look at the mean value (in the Poisson model) which is needed in the proof of Corollary 1. Therefore, by using iteration as in the proof of Proposition 2, we obtain

$$\tilde{L}_q(z) = 2^u \tilde{L}_q(z/2^k) + \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l \tilde{g}_{q^{(l)}}(z/2^l),$$

where $\tilde{g}_{q^{(l)}}(z) = 1 - e_b(z)e^{-z}$. Our goal is to derive an asymptotic expansion of $\tilde{L}_q(z)$. A standard tool for that purpose is the Mellin transform which we are going to apply next.

First, we have to clarify existence of the Mellin transform of $\tilde{L}_q(z)$. Therefore, note that by (4) and the trivial bound $\tilde{L}_q(z) = \mathcal{O}\left(z^{b+1}\right)$ as $z \to 0$. Hence, the Mellin transform of $\tilde{L}_q(z)$ exists in the strip $\langle -b-1, -u/k\rangle$. Applying Mellin transform to the above functional equation then yields

$$\mathscr{M}[\tilde{L}_q(z);\omega] = \frac{-\Gamma(w)}{1-2^{\omega k+u}} \binom{w+b}{b} \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l 2^{\omega l}, \qquad \Re(\omega) \in \langle -b-1, -u/k\rangle. \tag{5}$$

Moreover, by inverse Mellin transform and shifting the line of integration to the right (see the converse mapping theorem in [1]), we have

$$\tilde{L}_q(z) \sim z^{u/k} P_1(\log_2 z^{1/k}), \qquad (z \to \infty), \tag{6}$$

where

$$P_1(z) = \sum_{r=-\infty}^{\infty} c_r e^{2\pi i r z}, \qquad c_r = \frac{-\Gamma(-\omega_r)}{kL} \binom{-\omega_r + b}{b} \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l 2^{-\omega_r l}.$$

Note that due to the fast decay of (5) along vertical lines, (6) more generally holds uniformly for $|z| \to \infty$ and $|\arg(z)| \le \pi/2 - \epsilon$.

Asymptotic Expansion of $\tilde{V}_q(n)$. Here, we proceed as for the mean. First, by using iteration as above and applying Mellin transform to the resulting functional equation, we have

$$\mathscr{M}[\tilde{V}_q(z);\omega] = \frac{1}{1-2^{\omega k+u}} \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l 2^{\omega l} \mathscr{M}[\tilde{h}_{q^{(l)}}(z);\omega], \qquad \Re(\omega) \in \langle -b-1, -u/k\rangle.$$

Now, from (4), we have that as $z \to \infty$

$$\tilde{h}_{q^{(l)}}(z) = \mathcal{O}(z^{-\beta})$$

for any $\beta > 0$ and $0 \le l < k$. Obviously, $\tilde{h}_q^{(l)}(z) = \mathcal{O}(z^{b+1})$ as $z \to 0$ for $0 \le l < k$. Hence, the Mellin transform of $\tilde{h}_{q^{(l)}}(z)$ exists in the strip $\langle -b-1, \infty\rangle$. Our claimed result follows from this by inverse Mellin transform and shifting the line of integration to the right.

Simplification of the Fourier Coefficients. The main task is the evaluation of

$$\int_0^\infty z^{-\omega_r - 1} \left(2\delta_{l+1} e_b(z) e^{-z} \tilde{L}_{q^{(l+1)}}(z/2) + e_b(z)e^{-z} - e_b(z)^2 e^{-2z}\right) \mathrm{d}z.$$

Therefore, we concentrate on

$$\int_0^\infty z^{-\omega_r-1} e_b(z) e^{-z} \tilde{L}_{q^{(l+1)}}(z/2) \mathrm{d}z,$$

the remaining parts being easy. Note that due to (5), we have

$$\mathscr{M}[\tilde{L}_{q^{(l+1)}}(z/2);\sigma] = \frac{-2^\sigma \Gamma(\sigma)}{1-2^{\sigma k+u}} \binom{\sigma+b}{b} \delta_{q^{(l+1)}}(2^\sigma),$$

where

$$\delta_{q^{(l+1)}}(z) = \sum_{j=0}^{k-1} \delta_{l+2} \cdots \delta_{l+j+1} z^j.$$

Now, by inverse Mellin transform

$$\begin{aligned}
&\int_0^\infty z^{-\omega_r-1} e_b(z) e^{-z} \tilde{L}_{q^{(l+1)}}(z/2) \mathrm{d}z \\
&\qquad = \int_{(-b)} \frac{-2^\sigma \Gamma(\sigma)}{1-2^{\sigma k+u}} \binom{\sigma+b}{b} \delta_{q^{(l+1)}}(2^\sigma) \int_0^\infty z^{-\omega_r-\sigma-1} e_b(z) e^{-z} \mathrm{d}z \mathrm{d}\sigma \\
&\qquad = \int_{(-b)} \binom{\sigma+b}{b} \binom{-\omega_r-\sigma+b}{b} \frac{-2^\sigma \Gamma(\sigma)\Gamma(-\omega_r-\sigma)}{1-2^{\sigma k+u}} \delta_{q^{(l+1)}}(2^\sigma) \mathrm{d}\sigma,
\end{aligned}$$

where the outer integral is along the line $\Re(\sigma) = -b$. Finally, by shifting the line of integration to the left and collecting residues, we obtain the absolute convergent series

$$\begin{aligned}
&\int_0^\infty z^{-\omega_r-1} e_b(z) e^{-z} \tilde{L}_{q^{(l+1)}}(z/2) \mathrm{d}z \\
&\qquad = -\Gamma(-\omega_r) \sum_{l \ge b+1} \binom{-l+b}{b} \binom{-\omega_r+l+b}{b} \binom{\omega_r}{l} \frac{2^{-l} \delta_{q^{(l+1)}}(2^{-l})}{1-2^{-lk+u}}.
\end{aligned}$$

Collecting everything and standard computation yields the claimed result.

3 k-dimensional Bucket PATRICIA Tries

Here, from the definition of Patricia tries, we have for $q = (\star, \ldots)$

$$X_{q,n} = \begin{cases} X_{q',I_n} + X^*_{q',n-I_n}, & \text{if } I_n \in \{0,n\}, \\ X_{q',I_n} + X^*_{q',n-I_n} + 1, & \text{otherwise}, \end{cases} \qquad (n \ge b+1)$$

and for $q = (S, \ldots)$

$$X_{q,n} = \begin{cases} X_{q',I_n}, & \text{if } I_n = n, \\ X_{q',I_n} + 1, & \text{otherwise}, \end{cases} \qquad (n \ge b+1),$$

where notation and initial conditions are as in the previous section.

From this we then obtain for the Poisson generating function of the mean (with notation as before)

$$\tilde{L}_q(z) = \delta_1 \tilde{L}_{q'}(z/2) + 1 + \tilde{g}_q(z)$$

where

$$\tilde{g}_q(z) = -e_b(z)e^{-z} + \delta_1 e_b(z/2)e^{-z} - \delta_1 e^{-z/2},$$

and the Poisson generating function of the second moment

$$\tilde{M}_q(z) = \begin{cases} 2\tilde{M}_{q'}(z/2) + 2\tilde{L}_{q'}(z/2)^2 + 4(1-e^{-z/2})\tilde{L}_{q'}(z/2) + 1 + \tilde{g}_q(z), & \text{if } q = (\star, \ldots); \\ \tilde{M}_{q'}(z/2) + 2(1-e^{-z/2})\tilde{L}_{q'}(z/2) + 1 + \tilde{g}_q(z), & \text{if } q = (S, \ldots). \end{cases}$$

Moreover, we have for the poissonized variance

$$\tilde{V}_q(z) = \delta_q \tilde{V}_{q'}(z/2) + \tilde{h}_q(z),$$

where $\tilde{h}_q(z)$ was defined in the introduction. The remaining analysis now proceeds from these functional equational equations as in the previous section.

4 k-dimensional Bucket Digital Search Trees

Again, we start from a distributional recurrence for $X_{q,n}$ which for the current situation reads as follows

$$X_{q,n+b} \stackrel{d}{=} \begin{cases} X_{q',I_n} + X^*_{q',n-I_n} + 1, & \text{if } q = (\star, \ldots); \\ X_{q',I_n} + 1, & \text{if } q = (S, \ldots), \end{cases} \qquad (n \geq 0),$$

where the notation is as before and initial conditions are given by $X_{q,0} = 0$ and $X_{q,1} = \cdots = X_{q,b-1} = 1$.

From here, we can in principle proceed as before. However, we will see that the equation satisfied by the Poisson generating function is more complicated. More precisely, we have to cope with a differential-functional equation compared with the functional equation from the trie case. Here, we will first use Laplace transform to get rid of the differential operator. Then, after suitable normalization, we will be able to proceed as before. This combined use of Laplace and Mellin transform was introduced in [3] and we direct the interested reader to that paper for more details concerning technicalities.

Poissonization. We again define

$$\tilde{P}_q(z, y) = e^{-z} \sum_{n \geq 0} \mathbb{E}(e^{X_{q,n} y}) \frac{z^n}{n!}.$$

Then,

$$\sum_{j=0}^{b} \binom{b}{j} \tilde{P}_q(z, y) = e^y \tilde{P}_{q'}(z/2, y)^{\delta_1}.$$

Taking derivatives yields for the Poisson generating function of mean and second moment (denoted as before)

$$\sum_{j=0}^{b} \binom{b}{j} \tilde{L}^{(j)}_q(z) = \delta_1 \tilde{L}_{q'}(z/2) + 1 \tag{7}$$

and

$$\sum_{j=0}^{b}\binom{b}{j}\tilde{M}_q^{(j)}(z)=\begin{cases}2\tilde{M}_{q'}(z/2)+4\tilde{L}_{q'}(z/2)+2\tilde{L}_{q'}(z/2)^2+1, & \text{if } q=(\star,\ldots);\\ \tilde{M}_{q'}(z/2)+2\tilde{L}_{q'}(z/2)+1, & \text{if } q=(S,\ldots).\end{cases}$$

The first step is again to show that $\tilde{L}_q(z)$ and $\tilde{M}_q(z)$ are JS-admissible. Therefore, we need the following result which is proved by a reduction to the trie case (see [3] for similar results).

Proposition 4 *Assume that we have*

$$\sum_{j=0}^{b}\binom{b}{j}\tilde{f}_{q^{(l)}}^{(j)}(z)=\delta_{l+1}\tilde{f}_{q^{(l+1)}}(z/2)+\tilde{g}_{q^{(l)}}(z),\qquad (0\le l<k),$$

where all involved functions are entire and 0 *at* $z=0$*. Moreover, assume that* $\tilde{g}_{q^{(l)}}(z)$ *is JS-admissible for* $0\le l<k$*. Then,* $\tilde{f}_{q^{(l)}}(z)$ *is JS-admissible for* $0\le l<k$*.*

From this it then follows as in the trie case that $\tilde{L}_q(z)$ and $\tilde{M}_q(z)$ are JS-admissible.

Next, we consider the poissonized variance $\tilde{V}_q(z)=\tilde{M}_q(z)-\tilde{L}_q(z)^2$. An easy computation proves that

$$\sum_{j=0}^{b}\binom{b}{j}\tilde{V}^{(j)}(z)=\delta_1\tilde{V}_{q'}(z/2)+\tilde{h}_q(z),$$

where $\tilde{h}_q(z)$ was defined in the introduction. Then, from the JS-admissibility of $\tilde{L}_q(z)$ and $\tilde{M}_q(z)$, we obtain as for tries the following result.

Proposition 5 *As* $n\to\infty$*,*

$$\operatorname{Var}(X_{q,n})=\tilde{V}_q(n)+\mathcal{O}\left(n^{2u/k-1}\right).$$

Asymptotic Expansion of $\tilde{L}_q(z)$. Again, we first consider the mean value. Note that due to the differential operator it is not possible to iterate (7). Therefore, we first have to get rid of the differential operator which is achieved by applying Laplace transform. This yields

$$(s+1)^b\mathscr{L}[\tilde{L}_q(z);s]=2\delta_1\mathscr{L}[\tilde{L}_{q'}(z);2s]+(s+1)^{b-1}/s. \tag{8}$$

Next, we normalize with $Q(s)$ from the introduction. Therefore, set $\bar{L}_q(s)=\mathscr{L}[\tilde{L}_q(z);s]/Q(-s)^b$ and $\bar{G}(s)=(s+1)^{b-1}/(Q(-2s)^b s)$. Then,

$$\bar{L}_q(s)=2\delta_1\bar{L}_{q'}(2s)+\bar{G}(s).$$

Now, we can iterate and obtain

$$\bar{L}_q(s)=2^{k+u}\bar{L}_q(2^k s)+\sum_{l=0}^{k-1}2^l\delta_1\cdots\delta_l\bar{G}(2^l s).$$

Observe that this is a similar functional equation as in the trie case. Hence, we can proceed as before. Thus, we again apply Mellin transform. First, note that the Mellin transform of $\bar{L}_q(s)$ exists in a non-trivial strip. Moreover, due to the rapid growth of $Q(s)$ at infinity (see [3]), the Mellin transform of $\bar{G}(s)$

exists in the strip $\langle 1, \infty\rangle$. Applying Mellin transform yields

$$\mathscr{M}[\bar{L}_q(s);\omega] = \frac{\mathscr{M}[\bar{G}(s);\omega]}{1-2^{k-\omega k+u}} \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l 2^{l-\omega l}, \qquad \Re(\omega) \in \langle 1+u/k, \infty\rangle.$$

Next, by inverse Mellin transform and shifting the line of integration to the left, we obtain

$$\bar{L}_q(z) \sim \sum_{r=-\infty}^{\infty} c_r s^{-1-u/k-2\pi i r/(kL)}, \qquad (s \to 0).$$

Since, $Q(-s)^b = 1 + \mathcal{O}(|s|)$ as $s \to 0$, the same asymptotic expansion holds for $\mathscr{L}[\tilde{L}_q(z);s]$ as well. Finally, by formal inverse Laplace transform (see [3] for technical details justifying this step), we have

$$\tilde{L}_q(z) \sim z^{u/k} P_1(\log_2 z^{1/k}), \qquad (z \to \infty),$$

where P_1 is a computable, 1-periodic function. A more careful analysis shows that the above asymptotic expansion holds uniformly for $|z| \to \infty$ and $|\arg(z)| \le \pi/2 - \epsilon$.

Asymptotic Expansion of $\tilde{V}_q(z)$. Here, we proceed as above and obtain

$$\bar{V}_q(s) = 2^{k+u}\bar{V}_q(2^k s) + \sum_{l=0}^{k-1} \delta_1 \cdots \delta_l 2^l \bar{H}_{q^{(l)}}(2^l s),$$

where $\bar{V}_q(s) = \mathscr{L}[\tilde{V}_q(z);s]/Q(-s)^b$ and $\bar{H}_{q^{(l)}}(s) = (\mathscr{L}[\tilde{h}_{q^{(l)}}(z);s] + p(s))/Q(-2s)^b$ with

$$p(s) = \frac{(1+s)^{b-1} + (-1)^b}{s+2}.$$

Now, observe that

$$\tilde{h}_q(z) = \begin{cases} \mathcal{O}(z^{2u/k-2}), & \text{if } z \to \infty; \\ \mathcal{O}(1), & \text{if } z \to 0^+, \end{cases}$$

where the first bound follows from the bound of the previous paragraph (which we are allowed to differentiate due to Ritt's theorem; see [9]) and the second bound is trivial. This together with the growth properties of $Q(s)$ then in turn yields

$$\bar{H}_q(s) = \begin{cases} \mathcal{O}(1/s), & \text{if } s \to \infty; \\ \mathcal{O}(s^{-\beta}), & \text{if } s \to 0^+, \end{cases}$$

where $\beta > 0$ is an arbitrary constant. Consequently, the Mellin transform of $\bar{H}_q$ exists in the strip $\langle 1, \infty\rangle$. The remaining proof of Theorem 2 proceeds then as in the previous paragraph.

Simplification of the Fourier Coefficients for $b = 1$. First, by iteration of (8),

$$\mathscr{L}[\tilde{L}_q(z); s] = \frac{1}{s} \sum_{j \geq 0} \frac{\delta_{q,j}^*}{(s+1) \cdots (2^j s + 1)},$$

where $\delta_{q,j}^* = \prod_{l=1}^{j} \delta_l$. Next, by partial fraction expansion,

$$\mathscr{L}[\tilde{L}_q(z); s] = \frac{1}{s} \sum_{j \geq 0} \sum_{l=0}^{j} \frac{(-1)^{j-l} 2^{-\binom{j-l+1}{2}} \delta_{q,j}^*}{(2^l s + 1) Q_l Q_{j-l}} = \frac{1}{s} \sum_{l \geq 0} \frac{\bar{\delta}_{q,l}^*}{(2^l s + 1) Q_l},$$

where

$$\bar{\delta}_{q,l} = \sum_{j \geq 0} \frac{(-1)^j 2^{-\binom{j+1}{2}}}{Q_j} \delta_{q,j+l}.$$

Consequently, by inverse Laplace transform

$$\tilde{L}_q(z) = \sum_{l \geq 0} \frac{\bar{\delta}_{q,l}}{Q_l} (1 - e^{-z/2^l}).$$

This implies

$$\tilde{L}_q'(z) = \sum_{l \geq 0} \frac{\bar{\delta}_{q,l}}{2^l Q_l} e^{-z/2^l}, \qquad \tilde{L}_q'(z)^2 = \sum_{l,h \geq 0} \frac{\bar{\delta}_{q,l} \bar{\delta}_{q,h}}{2^{l+h} Q_l Q_h} e^{-z/2^l - z/2^h}.$$

Plugging this into (3) (note that for $b = 1$, we have $\tilde{h}_q(z) = \tilde{L}_q'(z)^2$) and using

$$\frac{1}{Q(-2s)} = \frac{1}{Q(1)} \sum_{j \geq 0} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j (s + 2^{-j})}$$

together with some standard computations proves the claim.

5 Conclusion

In this paper, we gave a new and simpler approach to the variance of partial match queries in k-dimensional bucket digital trees. Our method used standard tools from the analysis of algorithm such as poissonization-depoissonization and Mellin transform. The main simplification comes from the *poissonized variance* which incorporates cancellations at a much earlier stage compared to previous derivations.

Our approach allowed us to derive asymptotic expansions of the variance in k-dimensional bucket tries, k-dimensional bucket PATRICIA tries and k-dimensional bucket digital search trees. In all cases, the variance is asymptotic to $n^{u/k} P(\log_2 n^{1/k})$ where P is a 1-periodic function. Since the mean has the same order, our results show that the cost of partial match retrievals is concentrated around the mean.

We conclude by pointing out that even though we only derived the main term in the asymptotic expansions, our approach can be straightforwardly applied to derive longer asymptotic expansions, too.

Acknowledgements

We are indebted to the anonymous referees for many helpful comments. Financial support of the National Science Counsel is acknowledged as well.

References

[1] P. Flajolet, X. Gourdon, P. Dumas (1995). Mellin transforms and asymptotics: harmonic sums, *Theoret. Comput. Sci.*, **144**, 3-58.

[2] P. Flajolet and C. Puech (1986). Partial match retrival of multidimensional data, *J. ACM*, **33**, 371-407.

[3] H.-K. Hwang, M. Fuchs, V. Zacharovas (2010). Asymptotic variance of random digital search trees, *Discrete Math. Theor. Comput. Sci.*, **12**, 103-166.

[4] P. Jacquet and M. Régnier (1988). Normal limiting distribution of the size of tries, In *Performance'87* (Brussels, 1987) North-Holland Amsterdam, 209-223.

[5] P. Jacquet and W. Szpankowski (1998). Analytical de-Poissonization and its applications, *Theoret. Comput. Sci.*, **201**, 1-62.

[6] P. Kirschenhofer and H. Prodinger (1994). Multidimensional digital searching - alternative data structures, *Random Struc. Algorithms*, **5**, 123-134.

[7] P. Kirschenhofer, H. Prodinger, W. Szpankowski (1993). Multidimensional digital searching and some new parameters in tries, *Int. J. Found. Comput. Scie.*, **4**, 69-84.

[8] D. E. Knuth (1998). *The art of computer programming. Volume 3: Sorting and searching*, Addison-Wesley Publishing Co., Reading, Mass., second edition.

[9] F. W. J. Olver (1974). *Asymptotics and Special Functions*, Academic Press, Computer Science and Applied Mathematics.

[10] R. L. Rivest (1976). Partial-match retrieval algorithms, *SIAM J. Comput.*, **5**, 19-50.

[11] W. Schachinger (1995). The variance of a partial match retrieval in a multidimensional symmetric trie, *Random Struc. Algorithms*, **7**, 81-95.

[12] W. Szpankowksi (2001). *Average Case Analysis of Algorithms on Sequences*, Wiley, New York.