EXPECTED WORST-CASE PARTIAL MATCH IN RANDOM QUADTRIES

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6
luc@cs.mcgill.ca

Carlos Zamora-Cura
Instituto de Matemáticas
Universidad Nacional Autónoma de México
México, D.F.
czamora@cs.mcgill.ca

November 18, 2002

Abstract. We consider random multivariate quadtries obtained from n points independently and uniformly distributed on the unit cube of $\mathbf{R}^d$. Let $N_n(y)$ be the complexity of the standard partial match algorithm for fixed vector y, where y is a vector in $\mathbf{R}^s$, $0 < s < d$. We study $N_n = \sup_y N_n(y)$, the worst-case time for partial match. Among other things, we show that partial match is very stable, in the sense that $\sup_y N_n(y)/\inf_y N_n(y) \to 1$ in probability.

Keywords and phrases. Multivariate trie, quadtrie, probabilistic analysis, partial match, range search, law of large numbers, concentration inequality.

CR Categories: 3.74, 5.25, 5.5.

1991 Mathematics Subject Classifications: 60D05, 68U05.

Authors' address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6. The authors' research was sponsored by NSERC Grant A3456 and FCAR Grant 90-ER-0291. The second author was sponsored by a REDII–CONACyT project. FAX number: 1–514–398-3883.

Introduction

Tries are efficient data structures that were initially developed and analyzed by Fredkin (1960) and Knuth (1973). The tries considered here are constructed from n independent strings $X_1, \ldots, X_n$, each drawn from $\prod_{j=1}^{\infty} \Omega_j$, where Ω_j, the j-th alphabet, is a countable set. By appropriate mapping, we can and do assume that for all j, $\Omega_j = \mathcal{Z}$. In practice, the alphabets are often $\{0,1\}$, but that won't even be necessary for the results in this paper. Each string $X_i = (X_{i1}, X_{i2}, \ldots)$ defines an infinite path in a tree: from the root, we take the X_{i1}-st child, then its X_{i2}-st child, and so forth. The collection of nodes and edges visited by the union of the n paths is the infinite trie.

We now introduce the **multidimensional trie**, or **quadtrie**, which is built based on n sequences of $\{0,1\}^d$-valued symbols. For brevity, we call such a symbol a **dit** (a d-dimensional bit). A string $X_i = (X_{i1}, X_{i2}, \ldots)$ is thus a sequence of dits. Let T_∞ denote the infinite complete 2^d-ary position tree. A string X_i corresponds to a path in this 2^d-ary tree, where X_{i1} denotes the child of the root, and so forth, as each possible dit is mapped in a one-to-one fashion to the index of a child. For every node u in T_∞, let $N(u)$ denote the number of strings among $X_1, \ldots, X_n$ that visit u, and let $\delta(u)$ denote its distance from the root. The **infinite multidimensional trie** is the subtree of T_∞ consisting of all nodes u with $N(u) > 0$, and the edges that connect them. The **multidimensional trie** T_n is the subtree of T_∞ consisting of all nodes u with $N(u) \geq 2$ and all nodes u with $N(u) = 1$ whose parent v has $N(u) \geq 2$. The **multidimensional trie of the second kind** (T'_n) is the subtree of T_∞ consisting of all nodes u with $N(u) \geq 2$ and all the children of such nodes. Note that T_n is smaller than T'_n and that both coincide for $d = 1$. The number of leaves in T_n is n if the data strings $X_1, \ldots, X_n$ are all different. The number of leaves of T'_n is between n and $2^{d-1}n$. Both have an identical number of internal nodes, and identical heights H_n.

There is nothing that structurally differentiates a multidimensional trie from an ordinary trie constructed on the basis of an alphabet, except that the alphabet here is explicitly determined by the 2^d possible values of the dits. However, the multidimensional trie has a natural d-dimensional interpretation, as the dits X_{ij} in the string X_i may be considered as the collection of j-th bits in an expansion of a d-dimensional string of bits. Thus, each X_i is in fact interpreted as a d-dimensional vector of binary strings. With this connection, the multidimensional trie lends itself well not only to the standard dictionary operations (search, sort, insert, delete, prefix match), but also to intrinsically multivariate queries such as partial match, in which one searches for the occurrence or absence of a given string of dits $y = (y_1, y_2, \ldots)$, but only matches in a certain number of positions of the dits are required. The indices of these positions form a set $S \subseteq \{1, 2, \ldots, d\}$. We say that X_i matches (y, S) if this is the case. With $|S| = d$, this corresponds to a classic point search (for occurrence of y) and for $|S| = 0$, regardless of y, all X_i's match any (y, S). A partial match is thus determined by a pair (y, S), and a proper partial match corresponds to $0 < |S| < d$. The collection of all strings that match (y, S) forms a subtree $T(y, S)$ of T_∞ (by the usual path interpretation of a string). All implementations of partial match are such that a node u of T_n or T'_n is visited during execution of the "algorithm" if and only if it can possible be a match for (y, S). Put differently, all nodes in $T(y, S) \cap T_n$ or $T(y, S) \cap T'_n$ are visited, and any intelligent implementation would only visit those nodes. So, let $N_n(y, S) = |T(y, S) \cap T_n|$ or $N_n(y, S) = |T(y, S) \cap T'_n|$ denote the number of nodes thus visited. The number of internal (non-leaf) nodes visited is denoted by $I_n(y, S)$, a quantity that is the same for both kinds of multidimensional tries. Clearly, for both kinds of multidimensional tries,

$$I_n(y, S) \leq N_n(y, S) \leq (1 + 2^{d-|S|}) I_n(y, S)$$

so that to study first order asymptotics, an analysis of $I_n(y,S)$ suffices. When referring to previous results in the literature, it is important to distinguish between $|T(y,S) \cap T_n|$, $|T(y,S) \cap T'_n|$ and $I_n(y,S)$.

Several models of random multidimensional tries may be considered, depending upon the distribution of the data strings X_i. In all models described below, $X_1, \ldots, X_n$ are i.i.d. Let X_1 be distributed as the generic string of dits $Z = (Z_1, Z_2, \ldots)$. The models one might consider are as follows:

A. The i.i.d. model: $Z_1, Z_2, \ldots$ are i.i.d.

  A.1. The independent bit model: the bits $(Z_{11}, \ldots, Z_{1d})$ of Z_1 are independent. These are also called Bernoulli models. In the i.i.d. bit model, all bits Z_{1i} have the same Bernoulli (p) distribution. In the literature, the case $p = 1/2$ is often called the symmetric Bernoulli model. When $p \neq 1/2$, it is called the asymmetric Bernoulli model. In the nonuniform independent bit model, Z_{1i} is Bernoulli (p_i), with $p_1, \ldots, p_d$ being the parameters that control the distribution of the trie as a whole.

  A.2. The general dit model: The dit Z_1 has a given distribution on $\{0,1\}^d$. The parameters of the distribution are the probabilities $\mathbb{P}\{Z_1 = z\}, z \in \{0,1\}^d$.

B. The Markov model: $Z_1, Z_2, \ldots$ form a Markov chain on $\{0,1\}^d$.

C. The density model: $Z_1, Z_2, \ldots$ group the first, second, etc. bits in the binary expansions of the components of a d-dimensional random vector Z' with a density f on $[0,1]^d$. Note that in this setting, we may associate with each node u a d-dimensional square of volume $1/2^{d\delta(u)}$. The multidimensional trie corresponds to a multidimensional dyadic partition of the unit cube. Note that if f is the uniform density on the unit cube, then model C coincides with the symmetric Bernoulli model. If f is not uniform, then it is still true that while the Z_i's are dependent, Z_n tends in distribution to the uniform distribution on $\{0,1\}^d$ as $n \to \infty$.

D. The general independent model: $Z_1, Z_2, \ldots$ are independent but not necessarily identically distributed.

E. Other models of dependence.

Nearly all papers on multidimensional tries deal with the i.i.d. bit model. The present paper deals with the symmetric Bernoulli model and the density model. Results for the general dit model will be reported elsewhere.

With an appropriate collection of pointers from leaf nodes to data points, this structure is useful for searching and for data base operations, including partial match. Orenstein (1982) introduced multidimensional tries for database applications. Related ideas had earlier been proposed by Bentley and Burkhard (1976). Quadtries have also been useful in the compaction of multidimensional (geometric, video) information. Puech and Yahia (1985) provide the first analytical study.

Two cases are uninteresting: if $|S| = 0$, then we may return $\{X_1, \ldots, X_n\}$ without any search, and, if the partial match algorithm is run, its complexity is $N_n(y,S) = |T'_n|$. It is known that $\mathbb{E}\{|T'_n|\} = O(n)$ for the independent bit model (Jacquet and Régnier, 1986), and therefore, this case is not interesting. If $|S| = d$, then the partial match reduces to a point search. In that case,

$$\sup_y N_n(y,S)$$

is nothing but one plus the height (H_n) of T_n and $\sup_y I_n(y,S)$ is the height of T'_n. We know that the expected value is $O(\log n)$. For the symmetric Bernoulli model, we have

$$\frac{H_n}{\log_2 n} \to \frac{2}{d}$$

in probability and in the mean (see, e.g., Régnier (1981), Mendelson (1982), Devroye (1984), Pittel (1985, 1986)). The limit law of H_n was obtained in Devroye (1984), and laws of the iterated logarithm for the difference $H_n - 2\log_2 n$ can be found in Devroye (1992). The height for other models was studied by Régnier (1981), Mendelson (1982), Flajolet and Steyaert (1982), Flajolet (1983), Devroye (1984), Pittel (1985, 1986), and Szpankowski (1988, 1989).

This leaves us with the question posed in this paper: if $0 < |S| < d$, what is the asymptotic behavior of $\sup_y N_n(y,S)$ and of $\sup_y I_n(y,S)$? We recall that if y is a string of dits, then for the symmetric Bernoulli model,

$$\mathbb{E}\{N_n(y,S)\} = \tau(\log_2 n) n^{1-|S|/d} + o\left(n^{1-|S|/d}\right)$$

(Flajolet and Puech, 1986), where τ is a continuous positive periodic function. This result gives us information about the expected time for an average query. The variance of $N_n(y,S)$ was shown, for $d=2$, $|S|=1$, to be asymptotic to $\tau(\log_2 n)\sqrt{n}$ by Kirschenhofer, Prodinger and Szpankowski (1993), where τ is again a continuous positive periodic function. This was generalized to $d > 2$ by Schachinger (1995). In 2000, Schachinger proved that

$$\frac{N_n(y,S)}{\mathbb{E}\{N_n(y,S)\}} \to 1$$

in probability, and

$$\frac{N_n(y,S) - \mathbb{E}\{N_n(y,S)\}}{\sqrt{\mathbb{V}\{N_n(y,S)\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \ .$$

For the asymmetric Bernoulli model, in all dimensions, some asymptotics for $\mathbb{E}\{N_n(y,S)\}$ are obtained by Kirschenhofer, Prodinger and Szpankowski (1993), and, with y replaced by a random Y in Schachinger (2000). In the latter paper, it is shown that $N_n(Y,S)/\mathbb{E}\{N_n(Y,S)\} \xrightarrow{\mathcal{L}} Z_p$ under an idealized partial match model that does not correspond to our definition, where the distribution of Z_p depends upon the probability p only. More recently, Schachinger has studied the asymptotic behavior of the ratio $\log N_n(y,S)/\mathbb{E}\{\log N_n(y,S)\}$.

The quantity

$$M_n = \sup_y N_n(y,S)$$

is the worst-case query time (for a random trie T_n'). It is the natural generalization of the notion of a height to the partial match setting. We show that random multidimensional tries also behave well under the worst-case query time criterion:

Theorem 1. *For a random multidimensional trie under the symmetric Bernoulli model, we have, if* $0 < |S| < d$,

$$\sup_y \mathbb{E}\{N_n(y,S)\} = O(n^{1-|S|/d}) \ .$$

Also,

$$\mathbb{E}\{M_n\} = O(n^{1-|S|/d})$$

and, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{ M_n > (1+\epsilon)\sup_y \mathbb{E}\{N_n(y,S)\}\right\} = 0 \ .$$

Finally,

$$\lim_{n\to\infty} \frac{\mathbb{E}\{\sup_y N_n(y,S)\}}{\sup_y \mathbb{E}\{N_n(y,S)\}} = 1 \ .$$

We observe thus that the expected worst-case partial match complexity is asymptotically equivalent to the worst-case of the expected partial match complexity. That implies that the complexity of a partial match query is largely independent of the position y of the query, and is rather stable. Theorem 1 follows without work from Proposition 1 and Theorem 2 developed further on.

Boucheron-Lugosi-Massart inequality

The modern concentration inequalities are mainly due to Talagrand (1996a-b) and Ledoux (1996a-b), as surveyed by McDiarmid (1998). In this section, we recall a fundamental inequality due to Boucheron, Lugosi and Massart (2000) whose proof was based on logarithmic Sobolev inequalities developed in part by Ledoux (1996a).

Let $X_1,\ldots,X_n$ be independent random variables taking values in a measurable set $\mathcal{X}$. Denote the n-vector by X. Similarly, let $x_1,\ldots,x_n$ be elements of $\mathcal{X}$, and denote the n-vector by x. Let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function, and define

$$z = f(x_1,\ldots,x_n) \ , \ Z = f(X_1,\ldots,X_n) \ ,$$

and

$$Z^{(i)} = f(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,x_n) \ .$$

Lemma 1. *Assume that there exists a positive constant c such that*

$$\mathbb{E}\left\{\sum_{i=1}^n \left(z - Z^{(i)}\right)^2 1_{\left[z>Z^{(i)}\right]}\right\} \le cz \ .$$

Then for all $t > 0$,

$$\mathbb{P}\{Z \ge \mathbb{E}\{Z\} + t\} \le \exp\left(-\frac{t^2}{4c(\mathbb{E}\{Z\}+t)}\right) \ .$$

If for each x and each i, $z \le 1 + Z^{(i)}$, then for $0 < t \le \mathbb{E}\{Z\}/2$,

$$\mathbb{P}\{Z \le \mathbb{E}\{Z\} - t\} \le \exp\left(-\frac{t^2 \log(1+(\sqrt{2}-1)/c)}{4(\mathbb{E}\{Z\}+t)}\right) \ .$$

Expected value for partial match query times

It is useful to define two new quantities. Let k denote a level in a trie, with level numbering equal to distance from the root. Define the number of internal nodes at level k visited by a partial match query algorithm by

$$I_{nk}(y,S) = \sum_{u \in T(y,S):\delta(u)=k} 1_{[N(u)\geq 2]} \ .$$

Thus,

$$I_n(y,S) = \sum_{k=0}^{\infty} I_{nk}(y,S)$$

It is important to note that for the uniform model, the distribution of both $I_{nk}(y,S)$ and $I_n(y,S)$ does not depend upon y, and depends upon S only through its size $|S|$. Thus, with S given and fixed, we write I_n and I_{nk}. The expected value of I_n and I_{nk} has been studied by Flajolet and Puech (1986). In particular, there exist positive constants $c(|S|,d)$ and $c'(|S|,d)$ such that for all n,

$$c(|S|,d) \leq \frac{\mathbb{E}\{I_n\}}{n^{1-|S|/d}} \leq c'(|S|,d) \ .$$

There is no limit for this ratio, as it oscillates periodically (in $\log_2 n$).

Proposition 1. *Assume that $0 < |S| < d$. We have*

A. For all k, $I_{nk} \leq 2^{(d-|S|)k}$.

B. For all k, $\mathbb{E}\{I_{nk}\} \leq n^2/2^{(d+|S|)k}$.

C. For all $k \geq (1/d)\log_2(2(n-1))$, $\mathbb{E}\{I_{nk}\} \geq (n-1)^2/2^{(d+|S|)k+2}$.

D. For all $k \leq (1/d)\log_2(2(n-1))$, $\mathbb{E}\{I_{nk}\} \geq 2^{(d-|S|)k-4}$.

E. For all $n \geq 2$,

$$2^{-d+|S|-4} + 2^{-d-|S|-4} \leq \frac{\mathbb{E}\{I_n\}}{(2(n-1))^{1-|S|/d}} \leq 4 \ .$$

Proof. Let B be binomial (n,p) with $p = 1/2^{dk}$. It is easy to see that I_{nk} is a sum of $2^{(d-|S|)k}$ (dependent) indicators distributed as $1_{[B\geq 2]}$. Thus,

$$\mathbb{E}\{I_{nk}\} = 2^{(d-|S|)k}\mathbb{P}\{B \geq 2\} \ .$$

Part A follows immediately. Part B follows from the union bound: $\mathbb{P}\{B \geq 2\} \leq n^2p^2$. Note that

$$\mathbb{P}\{B \geq 2\} = 1-(1-p)^n - np(1-p)^{n-1} = 1-(1-p)^{n-1}(1+p(n-1)) \geq 1 - e^{-p(n-1)}(1+p(n-1)) \ .$$

The lower bound is an increasing function of $p(n-1)$. At $p(n-1) = 1/2$, its value is more than $1/2^4$, which proves D. Using $e^{-u} \leq 1-u+u^2/2$, $u > 0$, we see that

$$\mathbb{P}\{B \geq 2\} \geq \frac{(p(n-1))^2}{2}(1-p(n-1)) \geq \frac{(p(n-1))^2}{4}$$

when $p(n-1) \leq 1/2$. This proves C. Set $a = (1/d)\log_2(2(n-1))$. We have

$$\begin{aligned}\mathbb{E}\{I_n\} &\leq \sum_{k<a} 2^{(d-|S|)k} + \sum_{k\geq a} \frac{n^2}{2^{(d+|S|)k}} \\ &\leq 2^{(d-|S|)a+1} + \frac{2n^2}{2^{(d+|S|)a}}\end{aligned}$$

$$
\begin{aligned}
&= 2(2(n-1))^{1-|S|/d} + \frac{2n^2}{(2(n-1))^{1+|S|/d}} \\
&\le 4(2(n-1))^{1-|S|/d} \ .
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbf{E}\{I_n\} &\ge \sum_{k<a} 2^{(d-|S|)k-4} + \sum_{k\ge a} \frac{(n-1)^2}{2^{(d+|S|)k+2}} \\
&\ge 2^{(d-|S|)(a-1)-4} + \frac{(n-1)^2}{2^{(d+|S|)(a+1)+2}} \\
&= 2^{-d+|S|-4}(2(n-1))^{1-|S|/d} + \frac{(n-1)^2}{2^{d+|S|+2}(2(n-1))^{1+|S|/d}} \\
&= C(2(n-1))^{1-|S|/d} \ ,
\end{aligned}
$$

with $C = 2^{-d+|S|-4} + 2^{-d-|S|-4}$. □

Concentration for partial match query times

We begin with a concentration result for $I_{nk}(y,S)$ for fixed (y,S).

LEMMA 2. *Fix $k \ge 0$, $n \ge 1$ and the query (y,S). For all $t > 0$,*

$$
\mathbf{P}\{I_{nk}(y,S) \ge \mathbf{E}\{I_{nk}(y,S)\} + t\} \le \exp\left(-\frac{t^2}{8(\mathbf{E}\{I_{nk}(y,S)\}+t)}\right) \ .
$$

Furthermore, for $0 < t \le \mathbf{E}\{I_{nk}(y,S)\}/2$,

$$
\mathbf{P}\{I_{nk}(y,S) \le \mathbf{E}\{I_{nk}(y,S)\} - t\} \le \exp\left(-\frac{t^2}{22\mathbf{E}\{I_{nk}(y,S)\}+22t}\right) \ .
$$

PROOF. $I_{nk}(y,S)$ is a function of the data $Z_1,\ldots,Z_n$. Fix data $z_1,\ldots,z_n$ (n infinite strings of d-vectors of bits). Let $I_{nk}(y,S)$ be as defined above. Replace z_i in the data by Z_i, its uniform random counterpart. Call the resulting value $I_{nk}^{(i)}(y,S)$. Note that $I_{nk}(y,S) > I_{nk}^{(i)}(y,S)$ implies that exactly one z_j, $j \ne i$, has its first k d-vectors of bits coincide with those of z_i. In any case,

$$
I_{nk}(y,S) - I_{nk}^{(i)}(y,S) \le 1 \ .
$$

Thus, to verify the condition of Lemma 1, we have

$$
\begin{aligned}
&\mathbf{E}\left\{\sum_{i=1}^n \left(I_{nk}(y,S) - I_{nk}^{(i)}(y,S)\right)^2 \mathbb{1}_{\left[I_{nk}(y,S) > I_{nk}^{(i)}(y,S)\right]}\right\} \\
&\le \sum_{i=1}^n \mathbf{P}\left\{I_{nk}(y,S) > I_{nk}^{(i)}(y,S)\right\} \\
&\le \sum_{i=1}^n \mathbb{1}_{\left[\text{exactly one } z_j, j\ne i, \text{ has its first } k \text{ dits coincide with those of } z_i\right]} \\
&\le 2 I_{nk}(y,S) \ .
\end{aligned}
$$

Thus, by Lemma 1, for all $t > 0$,

$$\mathbf{P}\{I_{nk}(y,S) \geq \mathbf{E}\{I_{nk}(y,S)\} + t\} \leq \exp\left(-\frac{t^2}{8(\mathbf{E}\{I_{nk}(y,S)\} + t)}\right) .$$

Furthermore, for $0 < t \leq \mathbf{E}\{I_{nk}(y,S)\}/2$,

$$\mathbf{P}\{I_{nk}(y,S) \leq \mathbf{E}\{I_{nk}(y,S)\} - t\} \leq \exp\left(-\frac{t^2 \log(1 + (\sqrt{2}-1)/2)}{4\mathbf{E}\{I_{nk}(y,S)\} + 4t}\right) . \square$$

We observe that Lemma 2 only needed the strings $Z_1, \ldots, Z_n$ to be independent. It is valid, however, for any distribution of Z_1. In particular, it holds for the general dit model, the Markov model, the density model, and indeed, for all models listed in the introduction.

Lemma 3. *For any* $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\{|I_n(y,S) - \mathbf{E}\{I_n(y,S)\}| > \epsilon\mathbf{E}\{I_n(y,S)\}\} = 0 .$$

Thus,

$$\frac{I_n(y,S)}{\mathbf{E}\{I_n(y,S)\}} \to 1 \text{ in probability} .$$

Proof. We first show the upper bound:

$$\lim_{n\to\infty} \mathbf{P}\{I_n(y,S) - \mathbf{E}\{I_n(y,S)\} > \epsilon\mathbf{E}\{I_n(y,S)\}\} = 0 .$$

For $t > 0$,

$$\begin{aligned}
&\mathbf{P}\{I_n(y,S) - \mathbf{E}\{I_n(y,S)\} > t\} \\
&\quad \leq \mathbf{P}\left\{\sum_{k=0}^{a} (I_{nk}(y,S) - \mathbf{E}\{I_{nk}(y,S)\}) > t/3\right\} + \mathbf{P}\left\{\sum_{k=a}^{b} (I_{nk}(y,S) - \mathbf{E}\{I_{nk}(y,S)\}) > t/3\right\} \\
&\qquad + \mathbf{P}\left\{\sum_{k>b} I_{nk}(y,S) > t/3\right\} \\
&\quad = I + II + III
\end{aligned}$$

where

$$a = \lceil 10 \log_2 \log n \rceil \ , b = \left\lfloor \frac{(2-u)\log_2 n}{d + |S|} \right\rfloor$$

and $u > 0$ is an appropriately small positive constant to be chosen later. Choose $\epsilon > 0$ and set $t = \epsilon\mathbf{E}\{I_n(y,S)\}$. Using part A of Proposition 1, we have

$$I \leq \mathbf{P}\left\{\sum_{k=0}^{a} 2^{(d-|S|)k} > t/3\right\} = 0$$

if

$$2^{(d-|S|)a+1} \leq (\epsilon/3)\gamma(2(n-1))^{1-|S|/d}$$

where γ is the lower bound in part E of Proposition 1. This is clearly the case for all $n \geq n_0$ where n_0 depends upon ϵ, d and $|S|$ only. By parts C and D of Proposition 1, we have, for all n so large that $a \leq (1/d)\log_2(2(n-1)) \leq b$,

$$\mathbf{E}\{I_{nk}(y,S)\} \geq \min\left((n-1)^2/2^{(d+|S|)b+2}, 2^{(d-|S|)a-4}\right) \geq (\log n)^9$$

for all n large enough. Consider such large n. Next, by Lemma 2, we have

$$\begin{aligned} II &\le \sum_{k=a}^{b} \mathbf{P}\left\{ I_{nk}(y,S) - \mathbf{E}\{I_{nk}(y,S)\} \ge \frac{\epsilon}{3}\mathbf{E}\{I_{nk}(y,S)\}\right\} \\ &\le \sum_{k=a}^{b} \exp\left(-\frac{\epsilon^2 \mathbf{E}\{I_{nk}(y,S)\}}{72+24\epsilon}\right) \\ &\le (1+\log_2 n)\exp\left(-\frac{\epsilon^2(\log n)^9}{72+24\epsilon}\right) \\ &= O\left(\exp\left(-(\log n)^8\right)\right) \ . \end{aligned}$$

Finally, by Markov's inequality, and parts B and E of Proposition 1,

$$\begin{aligned} III &\le \frac{3\mathbf{E}\left\{\sum_{k>b} I_{nk}(y,S)\right\}}{\epsilon \mathbf{E}\{I_n(y,S)\}} \\ &\le \frac{3\sum_{k>b} n^2/2^{(d+|S|)k}}{\epsilon\gamma n^{1-|S|/d}} \\ &\le \frac{6n^2}{\epsilon\gamma 2^{(d+|S|)b} n^{1-|S|/d}} \\ &\le \frac{6\, 2^{d+|S|} n^2}{\epsilon\gamma n^{2-u} n^{1-|S|/d}} \\ &\le \frac{2^{d+|S|+3}}{\epsilon\gamma n^{1-|S|/d-u}} \\ &\to 0 \end{aligned}$$

if $u < 1 - |S|/d$. As $I + II + III \to 0$, the upper bound follows.

For the lower bound, we argue similarly, taking t, a and b as above. For $t > 0$,

$$\begin{aligned} &\mathbf{P}\{I_n(y,S) - \mathbf{E}\{I_n(y,S)\} < -t\} \\ &\quad \le \mathbf{P}\left\{\sum_{k=0}^{a}(I_{nk}(y,S) - \mathbf{E}\{I_{nk}(y,S)\}) < -t/3\right\} + \mathbf{P}\left\{\sum_{k=a}^{b}(I_{nk}(y,S) - \mathbf{E}\{I_{nk}(y,S)\}) < -t/3\right\} \\ &\qquad + \mathbf{P}\left\{\sum_{k>b}\mathbf{E}\{I_{nk}(y,S)\} > t/3\right\} \\ &\quad = I + II + III \ . \end{aligned}$$

We have for n large enough, by part A of proposition 1,

$$I \le \mathbf{P}\left\{\sum_{k=0}^{a}\mathbf{E}\{I_{nk}(y,S)\} > t/3\right\} \le \mathbf{P}\left\{\sum_{k=0}^{a} 2^{(d-|S|)k} > t/3\right\} = 0,$$

by arguing as above. Next, by Lemma 2, we have, if $\epsilon < 1$,

$$\begin{aligned} II &\le \sum_{k=a}^{b} \mathbf{P}\{I_{nk}(y,S) - \mathbf{E}\{I_{nk}(y,S)\} < -\frac{\epsilon}{3}\mathbf{E}\{I_{nk}(y,S)\}\} \\ &\le \sum_{k=a}^{b} \exp\left(-\frac{\epsilon^2\mathbf{E}\{I_{nk}(y,S)\}}{198+66\epsilon}\right) \end{aligned}$$

$$\leq (1+\log_2 n)\exp\left(-\frac{\epsilon^2(\log n)^9}{198+66\epsilon}\right)$$
$$= O\left(\exp\left(-(\log n)^8\right)\right) \ .$$

Finally, by parts B and E of Proposition 1, $III = 0$ for all n large enough as

$$\sum_{k>b} \mathbb{E}\{I_{nk}(y,S)\} \leq \frac{\epsilon \mathbb{E}\{I_n(y,S)\}}{3} \ .$$

Indeed, the right hand side is $\Theta(n^{1-|S|/d})$, while the left hand side is $O(n^u)$, and $u < 1 - |S|/d$ by choice of u. This concludes the proof of Lemma 3. □

Expected worst-case time for a partial match query

The purpose of this section is to prove the following theorem.

Theorem 2. *Fix S such that $0 < |S| < d$. Denote by $\mu_n(|S|) = \mathbb{E}\{I_n(y,S)\}$ (noting that this does not depend upon y). Then*

$$\frac{\sup_y I_n(y,S)}{\inf_y I_n(y,S)} \to 1 \text{ in probability} \ .$$

Furthermore,

$$\frac{\sup_y I_n(y,S)}{\mu_n(|S|)} \to 1 \text{ in probability} \ ,$$

and similarly for $\inf_y I_n(y,S)$.

Remark. Theorem 2 remains valid if we replace I_n by N_n for both kinds of multidimensional tries. For the tries of the first kind, this is an immediate consequence of $N_n(y,S) = I_n(y,S) + n$. For the tries of the second kind, a bit more work is needed.

Before we prove this, let us introduce

$$N_{nk}(y,S) = \sum_{u \in T(y,S): \delta(u)=k} N(u) \ ,$$

the number of data strings among $X_1, \ldots, X_n$ that match (y,S) in their first k dits. Observe that $I_{nk}(y,S) \leq N_{nk}(y,S)$.

Lemma 4. *We have*

$$\mathbb{E}\{N_{nk}(y,S)\} = \frac{n}{2^{|S|k}}$$

and for $t > 0$,

$$\mathbb{P}\{N_{nk}(y,S) - \mathbb{E}\{N_{nk}(y,S)\} \geq t\mathbb{E}\{N_{nk}(y,S)\}\} \leq \exp\left(-\frac{tn}{3\,2^{|S|k}}\right) \ .$$

PROOF. Observe that $N_{nk}(y,S)$ is binomial with parameters n and $1/2^{|S|k}$. Thus,

$$\mathbb{E}\{N_{nk}(y,S)\} = \frac{n}{2^{|S|k}}$$

and by a tail inequality for the binomial distribution (see, e.g., Angluin and Valiant, 1979; or Hoeffding, 1963),

$$\mathbb{P}\{N_{nk}(y,S) - \mathbb{E}\{N_{nk}(y,S)\} \geq t\mathbb{E}\{N_{nk}(y,S)\}\} \leq \exp\left(-(t/3)\mathbb{E}\{N_{nk}(y,S)\}\right) \ ,$$

for $t > 0$. □

PROOF OF THEOREM 2. It suffices to prove the second part of Theorem 2. We prove the part for the supremum, as the infimum is treated in an analogous manner. We write μ_n instead of $\mu_n(|S|)$, and recall from Proposition 1 that $\mu_n = \Theta(n^{1-|S|/d})$. We first show the upper bound: for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{\sup_y \left(I_n(y,S) - \mu_n\right) > \epsilon\mu_n\right\} = 0 \ .$$

Note the following:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_y \left(I_n(y,S) - \mu_n\right) > \epsilon\mu_n\right\} \\
&\quad \leq \mathbb{P}\left\{\sum_{k=0}^{a} 2^{dk} > \epsilon\mu_n/3\right\} + \mathbb{P}\left\{\sum_{k=a}^{b} \sup_y \left(I_{nk}(y,S) - \mathbb{E}\{I_{nk}(y,S)\}\right) > \epsilon\mathbb{E}\{I_n(y,S)\}/3\right\} \\
&\qquad + \mathbb{P}\left\{\sup_y \sum_{k>b} I_{nk}(y,S) > \epsilon\mu_n/3\right\} \\
&\quad = I + II + III \ .
\end{aligned}$$

We pick the integers a and b as in the proof of Lemma 3:

$$a = \lceil 10\log_2 \log n\rceil \ , \ b = \left\lfloor \frac{(2-u)\log_2 n}{d+|S|}\right\rfloor$$

and $1 - |S|/d > u > 0$ is an appropriately small positive constant. Note that for n large enough, $(a+1)2^{da} < \epsilon\mu_n/3$, as the left-hand side grows as a polynomial of $\log n$ and the right hand side as a polynomial in n. We recall from the proof of Lemma 3 that for all n so large that $a \leq (1/d)\log_2(2(n-1)) \leq b$, $\mathbb{E}\{I_{nk}(y,S)\} \geq (\log n)^9$. Consider such large n. Next, by Lemma 2, we have

$$II \leq \sum_{k=a}^{b} 2^{db} \sup_y \mathbb{P}\left\{I_{nk}(y,S) - \mathbb{E}\{I_{nk}(y,S)\} > \frac{\epsilon}{3}\mathbb{E}\{I_{nk}(y,S)\}\right\}$$

because there are at most 2^{dk} different query strings y that can yield different values for $I_{nk}(y,S)$ when $k \leq b$. As in the proof of Lemma 3, this is further bounded as follows:

$$\begin{aligned}
II &\leq b\, 2^{db} \exp\left(-\frac{\epsilon^2\mu_n}{72+24\epsilon}\right) \\
&\leq (1+\log_2 n) n^{\frac{(2-u)d}{d+|S|}} \exp\left(-\frac{\epsilon^2(\log n)^9}{72+24\epsilon}\right) \\
&= O\left(\exp\left(-(\log n)^8\right)\right) \ .
\end{aligned}$$

We now consider III. Observe that

$$\sup_y \sum_{k \geq b} I_{nk}(y,S) \leq H_n \times \sup_y N_{nb}(y,S)$$

as each string X_i counted in $N_{nb}(y,S)$ can contribute to at most H_n nodes counted in $I_{nk}(y,S)$ with $k \geq b$.

By Lemma 4, with $\chi = (1-u/2)|S|/(d+|S|)$,

$$n^{\frac{d}{d+|S|}-\chi} \geq \mathbb{E}\{N_{nb}(y,S)\} = n^{\frac{d-|S|+u|S|+o(1)}{d+|S|}} \geq n^{\frac{d-|S|}{d+|S|}}$$

for all n large enough. For such large n, we have

$$\mathbb{P}\{N_{nb}(y,S) - \mathbb{E}\{N_{nb}(y,S)\} > t\mathbb{E}\{N_{nb}(y,S)\}\} \leq \exp\left(-tn^{\frac{d-|S|}{d+|S|}}/3\right) .$$

By the union bound,

$$\mathbb{P}\left\{\sup_y \frac{N_{nb}(y,S)}{\mathbb{E}\{N_{nb}(y,S)\}} > 1+t\right\} \leq 2^{db} \exp\left(-tn^{\frac{d-|S|}{d+|S|}}/3\right) \leq \exp\left(-tn^{\frac{d-|S|}{d+|S|}}/4\right)$$

for all n large enough. In particular, as the distribution of $N_n(y,S)$ does not depend upon y, we have for n large enough,

$$\begin{aligned}
III &= \mathbb{P}\left\{\sup_y \frac{\sum_{k>b} I_{nk}(y,S)}{\mu_n} > \frac{\epsilon}{3}\right\} \\
&\leq \mathbb{P}\left\{H_n \sup_y \frac{N_{nb}(y,S)}{\mu_n} > \frac{\epsilon}{3}\right\} \\
&\leq \mathbb{P}\{H_n > (3/d)\log_2 n\} + \mathbb{P}\left\{\sup_y \frac{N_{nb}(y,S)}{\mu_n} > \frac{\epsilon/3}{(3/d)\log_2 n}\right\} \\
&= \mathbb{P}\{H_n > (3/d)\log_2 n\} + \mathbb{P}\left\{\sup_y \frac{N_{nb}(y,S)}{\mathbb{E}\{N_{nb}(y,S)\}} > \frac{(\epsilon/3)\mu_n}{((3/d)\log_2 n)\sup_y \mathbb{E}\{N_{nb}(y,S)\}}\right\} \\
&\leq \frac{n^2}{2^{d((3/d)\log_2 n)}} + \mathbb{P}\left\{\sup_y \frac{N_{nb}(y,S)}{\mathbb{E}\{N_{nb}(y,S)\}} > \frac{(\epsilon/3)\beta n^{1-|S|/d}}{((3/d)\log_2 n)n^{\frac{d}{d+|S|}-\chi}}\right\} \\
&\qquad \text{(where } \beta > 0 \text{ is the lower bound of Proposition 1)} \\
&\leq \frac{1}{n} + \mathbb{P}\left\{\sup_y \frac{N_{nb}(y,S)}{\mathbb{E}\{N_{nb}(y,S)\}} > \frac{(\epsilon/3)\beta n^{-\frac{|S|^2}{d(d+|S|)}+\chi}}{((3/d)\log_2 n)}\right\} \\
&\leq \frac{1}{n} + \mathbb{P}\left\{\sup_y \frac{N_{nb}(y,S)}{\mathbb{E}\{N_{nb}(y,S)\}} > n^{\frac{(1-u)|S|-|S|^2/d}{d+|S|}}\right\} \\
&\qquad \text{(for } n \text{ large enough)} \\
&= O(1/n)
\end{aligned}$$

by the bound established above for the tail of the distribution of $\sup_y N_{nb}(y,S)$, after noting that $(1-u)|S| > |S|^2/d$ whenever u is chosen such that $u < 1 - |S|/d$. We conclude that $I + II + III \to 0$ for all $\epsilon > 0$.

For the lower bound we argue similarly. Since for any k, $\mathbb{E}\left\{I_{nk}(y,S)\right\} \le \mathbb{E}\left\{I_n(y,S)\right\}$,

$$\begin{aligned}
&\mathbb{P}\left\{\inf_y \left(I_n(y,S) - \mathbb{E}\left\{I_n(y,S)\right\}\right) < -\epsilon\mathbb{E}\left\{I_n(y,S)\right\}\right\} \\
&\le \mathbb{P}\left\{\exists y\, \exists a \le k \le b,\ I_{nk}(y,S) - \mathbb{E}\left\{I_{nk}(y,S)\right\} < -\epsilon\mathbb{E}\left\{I_{nk}(y,S)\right\}\right\} \\
&\le \sum_{k=a}^{b} 2^{db} \sup_y \mathbb{P}\left\{I_{nk}(y,S) - \mathbb{E}\left\{I_{nk}(y,S)\right\} < -\epsilon\mathbb{E}\left\{I_{nk}(y,S)\right\}\right\} \\
&\le \sum_{k=a}^{b} 2^{db} \exp\left(-\frac{\epsilon^2\mathbb{E}\left\{I_{nk}(y,S)\right\}}{22(1+\epsilon)}\right) \\
&\le n^{\frac{(2-u)d}{d+|S|}}(1+\log n)\exp\left(\frac{\epsilon^2(\log n)^9}{22(1+\epsilon)}\right) \\
&= O\left(\exp\left(-(\log n)^8\right)\right),
\end{aligned}$$

for every n large enough such that $a \le \frac{1}{d}\log_2(2(n-1)) \le b$. This completes the proof. □

The density model

For the density model, with density f on $[0,1]^d$ bounded by $F < \infty$, it is still true that $\mathbb{E}\{M_n\} = O(n^{1-|S|/d})$. The proof uses embedding. We generate $2Fn$ i.i.d. pairs (X_i, U_i), where X_i is uniform on $[0,1]^d$ and U_i is uniform $[0,1]$. The subset with index j such that $U_j F < f(X_j)$ forms a sample drawn from density f. Its size is binomial $(2Fn, 1/F)$, so that with probability at least $1 - 2\exp(-n/F)$ (Hoeffding's inequality, 1963), the sample size exceeds n. Therefore, the trie constructed with the uniform sample of size $2Fn$ contains as a subtree the one based on a sample of size n drawn from f. By applying Theorem 1, we thus have $\mathbb{E}\{M_n\} = O((2Fn)^{1-|S|/d})$.

References

D. Angluin and L. Valiant, "Fast probabilistic algorithms for Hamiltonian circuits and matchings," *Journal of Computer and System Sciences*, vol. 18, pp. 155–193, 1979.

J. L. Bentley and W. A. Burkhard, "Heuristics for partial-match retrieval in database design," *Information Processing Letters*, vol. 4(5), pp. 132–135, 1976.

S. Boucheron, G. Lugosi, and P. Massart, "A sharp concentration inequality with applications in random combinatorics and learning," *Random Structures and Algorithms*, vol. 16, pp. 277–292, 2000.

L. Devroye, "A probabilistic analysis of the height of tries and of the complexity of triesort," *Acta Informatica*, vol. 21, pp. 229–237, 1984.

L. Devroye, "A study of trie-like structures under the density model," *Annals of Applied Probability*, vol. 2, pp. 402–434, 1992.

P. Flajolet, "On the performance evaluation of extendible hashing and trie search," *Acta Informatica*, vol. 20, pp. 345–369, 1983.

P. Flajolet and C. Puech, “Partial match retrieval of multidimensional data,” *Journal of the ACM*, vol. 33, pp. 371–407, 1986.

P. Flajolet and R. Sedgewick, “Digital search trees revisited,” *Siam Journal on Computing*, vol. 15, pp. 748–767, 1986.

P. Flajolet and J. M. Steyaert, “A branching process arising in dynamic hashing, trie searching and polynomial factorization,” in: *Lecture Notes in Computer Science*, vol. 140, pp. 239–251, Springer-Verlag, New York, 1982.

E. H. Fredkin, “Trie memory,” *Communications of the ACM*, vol. 3, pp. 490–500, 1960.

W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

P. Jacquet and M. Régnier, “Trie partitioning process: limiting distributions,” in: *Lecture Notes in Computer Science*, vol. 214, pp. 196–210, 1986.

P. Kirschenhofer and H. Prodinger, “Some further results on digital trees,” in: *Lecture Notes in Computer Science*, vol. 226, pp. 177–185, Springer-Verlag, Berlin, 1986.

P. Kirschenhofer, H. Prodinger, and W. Szpankowski, “Multidimensional digital searching and some new parameters in tries,” *International Journal of Foundations of Computer Science*, vol. 4, pp. 69–84, 1993.

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

M. Ledoux, “On Talagrand’s deviation inequalities for product measures,” *ESAIM: Probability and Statistics*, vol. 1, pp. 63–87, 1996a.

M. Ledoux, “Isoperimetry and gaussian analysis,” in: *Lectures on Probability Theory and Statistics*, (edited by P. Bernard), pp. 165–294, Ecole d’Eté de Probabilités de St-Flour XXIV-1994, 1996b.

C. McDiarmid, “On the method of bounded differences,” in: *Surveys in Combinatorics*, (edited by J. Siemons), vol. 141, pp. 148–188, London Mathematical Society Lecture Note Series, Cambridge University Press, 1989.

C. McDiarmid, “Concentration,” in: *Probabilistic Methods for Algorithmic Discrete Mathematics*, (edited by M. Habib and C. McDiarmid and J. Ramirez-Alfonsin and B. Reed), pp. 195–248, Springer, New York, 1998.

H. Mendelson, “Analysis of extendible hashing,” *IEEE Transactions on Software Engineering*, vol. 8, pp. 611–619, 1982.

J. A. Orenstein, “Multidimensional tries used for associative searching,” Technical Report, School of Computer Science, McGill University, Montreal, 1982.

B. Pittel, “Asymptotical growth of a class of random trees,” *Annals of Probability*, vol. 13, pp. 414–427, 1985.

B. Pittel, “Path in a random digital tree: limiting distributions,” *Advances in Applied Probability*, vol. 18, pp. 139–155, 1986.

C. Puech and H. Yahia, “Quadtrees, octrees, hyperoctrees: a unified analytical approach to tree data structures used in graphics, geometric modeling and image processing,” in: *Proceedings of the Symposium on Computational Geometry*, pp. 272–280, ACM, New York, 1985.

M. Régnier, “On the average height of trees in digital searching and dynamic hashing,” *Information Processing Letters*, vol. 13, pp. 64–66, 1981.

W. Schachinger, “The variance of a partial match retrieval in a multidimensional symmetric trie,” *Random Structures and Algorithms*, vol. 7, pp. 81–95, 1995.

W. Schachinger, “Limiting distributions for the costs of partial match retrievals in multidimensional tries,” *Random Structures and Algorithms*, vol. 17, pp. 428–459, 2000.

W. Szpankowski, “Some results on V-ary asymmetric tries,” *Journal of Algorithms*, vol. 9, pp. 224–244, 1988.

W. Szpankowski, “Digital data structures and order statistics,” in: *Algorithms and Data Structures: Workshop WADS ’89 Ottawa*, vol. 382, pp. 206–217, Lecture Notes in Computer Science , Springer-Verlag, Berlin, 1989.

M. Talagrand, “New concentration inequalities in product spaces,” *Inventiones Mathematicae*, vol. 126, pp. 505–563, 1996a.

M. Talagrand, “A new look at independence,” *Annals of Probability*, vol. 24, pp. 1–34, 1996b.