# An Analysis of the Height of Tries with Random Weights on the Edges

N. Broutin L. Devroye*

September 10, 2007

**Abstract**

We analyze the weighted height of random tries built from independent strings of i.i.d. symbols on the finite alphabet $\{1, \ldots, d\}$. The edges receive random weights whose distribution depends upon the number of strings that visit that edge. Such a model covers the hybrid tries of de la Briandais (1959) and the TST of Bentley and Sedgewick (1997), where the search time for a string can be decomposed as a sum of processing times for each symbol in the string. Our weighted trie model also permits one to study maximal path imbalance. In all cases, the weighted height is shown be asymptotic to $c \log n$ in probability, where $c$ is determined by the behavior of the *core* of the trie (the part where all nodes have a full set of children) and the fringe of the trie (the part of the trie where nodes have only one child and form *spaghetti*-like trees). It can be found by maximizing a function that is related to the Cramér exponent of the distribution of the edge weights.

**Keywords:** data structure, trie, TST, random tree, height, branching random walk.

## 1 Introduction

Tries are tree-like data structures that have been introduced by de la Briandais (1959) and Fredkin (1960) in order to efficiently store and manipulate strings. They find a multitude of applications in computer science and telecommunications (see, e.g., Szpankowski, 2001; Flajolet, 2006). Consider $n$ strings, each consisting of an infinite sequence of symbols taken from a finite alphabet $\mathcal{A}$. We assume without loss of generality that $\mathcal{A} = \{1, 2, \ldots, d\}$. Each sequence defines an infinite path in an infinite $d$-ary position tree $T_\infty$. If the sequences are distinct, then the paths are distinct as well. We trim $T_\infty$ by cutting every branch below the shallowest node that belongs to a single path. The trie is the resulting finite tree, and the strings are stored in the leaves. In the usual array-based implementation of the data structure, the worst-case time to answer a search query corresponds to the height of the trie, i.e., the maximum number of edges on a path from the root. The heights of tries have been studied by many authors under various model of randomness for the sequences. For more information about general models, see Szpankowski (2001), Clément, Flajolet, and Vallée (2001), Flajolet (2006) and the references found there.

Here, we assume that the sequences are built using a *memoryless source*: each string is an infinite sequence of independent and identically distributed (i.i.d.) symbols distributed like $X$, where $\mathbf{P}\{X = i\} = p_i$, $1 \le i \le d$. In addition, we assume that the strings are

*Research of the authors was supported by NSERC Grant A3456 and a James McGill fellowship. Address: School of Computer Science, McGill University, Montreal H3A2K6 Canada. Email: `nicolas.broutin@m4x.org, luc@cs.mcgill.ca`.

independent. It is well-known that the height $H_n$ of a trie built from $n$ such sequences satisfies (Régnier, 1981; Devroye, 1984; Pittel, 1985; Szpankowski, 1991, 2001)

$$\frac{H_n}{\log n} \xrightarrow[n\to\infty]{} -\frac{2}{\log Q(2)} \quad \text{in probability,} \tag{1}$$

where

$$Q(b) = \sum_{i=1}^{d} p_i^b \tag{2}$$

is the probability that $b \geq 1$ independent characters are identical. This result holds when every leaf contains only one string. If the leaves can store up to $b$ strings, the tree is called a $b$-trie (see, e.g., Szpankowski, 2001) and its height $H_{n,b}$ is such that

$$\frac{H_{n,b}}{\log n} \xrightarrow[n\to\infty]{} -\frac{b+1}{\log Q(b+1)} \quad \text{in probability.}$$

The usual implementation of a trie uses an array for the branching structure of a node (Fredkin, 1960). Although this always ensures constant-time shunting of the words in the subtrees, the space required may become an issue for large alphabets: many pointers would be left unused. To avoid this, one can replace the array by variable size structures. There are essentially two solutions which have been considered. De la Briandais (1959) proposed to use linked-lists, and we shall call the implementation a *list-trie*. More recently, building on early ideas of Clampett (1964), Bentley and Sedgewick (1997) developed an elegant structure based on binary search trees going by the name of *bst-trie*, ternary search trie or TST for short.

These alternative implementations aim at a trade-off between storage space and speed, and the access time to children of a node is no longer constant. In particular, the height of the tree and the worst-case search time are different in general. List-tries and TST may be seen as *high-level* tries whose edges are weighted to reflect the internal *low-level* structure used to organize the children of a node. This point of view has been taken by Clément, Flajolet, and Vallée (1998, 2001) who thoroughly analyzed these *hybrid* implementations of tries. In particular, they analyzed the average size and average depth. The question of the worst-case search time in hybrid-tries was left open. This paper addresses the latter question by studying the *weighted height* of a general model of *weighted tries* that encompass hybrid tries.

The analysis requires the new ideas of Broutin and Devroye (2007a) who distinguish two different regions in the trie. We shall motivate the need for such a distinction and give more insight about the model using an example.

**Example: randomized list-tries**. Assume that the low level structure used to implement the set of subtrees at a node is a list. Assume for simplicity that the alphabet is $\{1, 2\}$ and that, for each node, an independent coin is flipped to decide which subtree will be first in the list. Then, one can easily see that the nodes do not all behave in the same way with respect to the costs: Towards the top of the tree, the nodes tend to have two children, and the cost of going to any of them is 1 or 2, each case occurring with probability 1/2. Towards the bottom of the tree, however, many nodes only have one child and the cost is then always 1.

Even this simple example shows that one should distinguish a region that is close to the root —the *core*— from the fringe of the tree —the *spaghettis*— (precise definitions will follow). We will see in the following (Lemma 1) that this simple binary distinction suffices to explain properties of tries such as the height and the profile. The distinction is not necessary to obtain parameters like the average cost of looking up a sequence because the number of nodes in the spaghettis is negligible compared to that in the core. This partly explains why the average weighted depth was know already (Clément et al., 1998, 2001). The situation is of course radically different if one is interested in the height for which every single node is relevant.

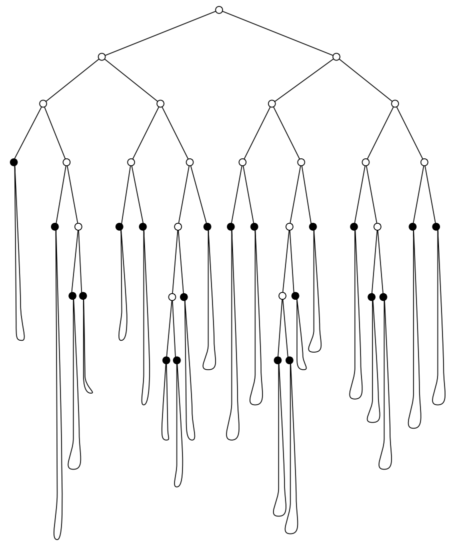

**Figure 1.** The structure of a trie: the bulk is the core, then some spaghetti-like trees hang down the core. Both the core and the spaghettis contribute significantly to the height of the trie. Observe also that the height may not be explained by a spaghetti born at one of the deepest nodes of the core. This latter fact will become clear later.

Our main result (Theorem 1) is a law of large numbers for the (weighted) height of a general model of random tries. Roughly speaking, we prove that the height of such tries on $n$ sequences is asymptotic to $c \log n$ in probability, where $c$ is characterized using large deviation techniques. The constant $c$ is the sum of the contributions of the core and the spaghettis.

Our method has several advantages. First, it yields the first order asymptotics of the height of hybrid-tries (see Clément et al., 1998, 2001, for more on this). But it also permits to shed new light on the family of digital trees. The profile of the core also happens to be the profile of digital search trees, a related model discussed more precisely later. In this sense, our methods unify the treatment of tries and digital search trees. This similarity goes further than the mere case of digital trees, and our methods rely on the branching processes treatment of trees of Broutin et al. (2007).

A detailed plan of the paper can be found at the end of Section 2, where we introduce the model more precisely and we sketch the key steps explaining our results. An early version of the results and the case of hybrid tries in particular in an extended abstract (see Broutin and Devroye, 2007b).

# 2 Random weighted tries

## 2.1 Constructing tries via an embedding

In this section, we propose an embedding to construct weighted tries. We will see in Section 7 that hybrid-tries, like list-tries and TST, can be seen as weighted tries built using this process.

Consider the distribution $\{p_1, \dots, p_d\}$ over the alphabet $\mathcal{A} = \{1, 2, \dots, d\}$. We assume without loss of generality that $1 > p_1 \geq p_2 \geq \cdots \geq p_d > 0$. We are given $n$ independent strings, each consisting of an infinite sequence of i.i.d. characters of $\mathcal{A}$ distributed as $X$, where $\mathbf{P}\{X = i\} = p_i$, $1 \le i \le d$. A weighted $b$-trie is built in two steps as follows: first it is given a *shape* (unweighted tree), the *weights* are then assigned to the edges using the shape.

THE SHAPE OF THE TRIE. Each string defines an infinite path in $T_\infty$. For a node $u \in T_\infty$, let $N_u^\star$ be the number of strings whose paths in $T_\infty$ intersect $u$. Then, for a natural number $b \ge 1$, the $b$-trie $T_{n,b}$ consists of the root together with the nodes whose parent $v$ has $N_v^\star > b$:

$$T_{n,b} = \{\text{all nodes } u \text{ whose parent } v \text{ has } N_v^\star > b\} \cup \{\text{root}\}.$$

We can then define the *cardinality* $N_u$ of a node $u \in T_\infty$ as the number of strings intersecting $u$ within $T_{n,b}$. Observe that we have $N_u = 0$ if $u \notin T_{n,b}$. The sequences are distinct with probability one, and the strings define distinct paths in $T_\infty$. Therefore, the trie $T_{n,b}$ is almost surely finite. The tree $T_{n,b}$ constitutes the *shape* of the weighted trie, and may be represented by the sequence $\{N_u : u \in T_\infty\}$. For the edge $e$ between $u$ and its $i$-th child we let $p_e = p_i$ and $E_e = -\log p_e$.

**Remark.** For a specified edge $e$, $E_e$ is deterministic. The values will later become random after some symmetrization process among the child edges of a node.

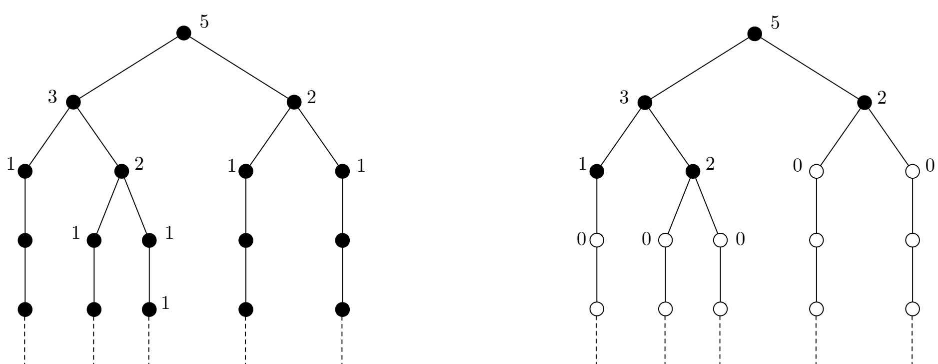


**Figure 2.** The way cardinalities are defined for a 2-trie. On the left we have represented the sequences and the values of $N_u^\star$, $u \in T_\infty$. On the right, the white nodes are those not in the 2-trie, and the labels are the values of the cardinalities $N_u$, $u \in T_\infty$.

As a position tree, $T_{n,b}$ potentially contains $2^d$ *types* of nodes, each type being characteristic of the subset of children that are present. The type of every node $u \in T_{n,b}$ is represented by a $d$-vector $\tau_u$: if $u_1, \dots, u_d$ are the $d$ children of $u$ in $T_\infty$, then we define

$$\tau_u = \left(\mathbf{1}[N_{u_i} \ge 1], \mathbf{1}[N_{u_2} \ge 1], \dots, \mathbf{1}[N_{u_d} \ge 1]\right),$$

where $\mathbf{1}[A] = 1$ if and only if the event $A$ holds. The weights of the edges to $u_1, \dots, u_d$ are assigned depending on the type $\tau_u$.

THE WEIGHTS. Consider a collection $\mathcal{Z}$ of random vectors $\{\mathcal{Z}^\tau, \tau \in \{0,1\}^d\}$, where $\mathcal{Z}^\tau = (Z_1^\tau, \dots, Z_d^\tau)$. For a fixed type $\tau \in \{0,1\}^d$, the components of $\mathcal{Z}^\tau$ may be dependent. Also, the members of the collection may be dependent. Each node of $T_\infty$ is given an independent copy of $\mathcal{Z}$. We always assume that there is a constant $a$ such that $|Z_i^\tau| \le a$ for all $\tau \in \{0,1\}^d$ and $i \in \{1, \dots, d\}$. Consider a node $u \in T_\infty$, and its sequence $\{\mathcal{Z}^\tau\}$. The weights of the child edges of $u$ are assigned using the vector $\mathcal{Z}^{\tau_u}$ only. In particular, the edge $e_i$ between $u$ and its $i$-th child in $T_\infty$ is given the weight

$$Z_{e_i} \;=\; Z_i^{\tau_u} \;=\; \sum_{\tau \in \{0,1\}^d} Z_i^\tau \cdot \mathbf{1}[\tau_u = \tau].$$

We use the notations $Z_i^\tau$ and $Z_e$ interchangeably. It should always be clear whether a subscript refers to an index or an edge. The weighted tree obtained in this way is called a *random trie with weight sequence distribution* $\{\mathcal{Z}^\tau, \tau \in \{0,1\}^d\}$ and source $\{p_1, \dots, p_d\}$.

**Example.** Let $d = 2$. The collection of weights $\mathcal{Z}$ such that $\mathcal{Z}^{(0,0)} = (0,0)$, $\mathcal{Z}^{(1,0)} = (X, 0)$, $\mathcal{Z}^{(0,1)} = (0, X)$ and $\mathcal{Z}^{(1,1)} = (X, X)$ where $X$ is a $[0,1]$-uniform random variable is acceptable.

**Example: Randomized list-tries.** Consider the example of binary list-tries of Section 1. The shape is the regular (unweighted) trie. The weight accounts for the costs to branch to subtrees. When the subtrees are ordered randomly in the list, we have

$$\mathcal{Z}^{(1,1)} = \begin{cases} (1,2) & \text{w.p.} \quad 1/2 \\ (2,1) & \text{w.p.} \quad 1/2, \end{cases}$$

and $\mathcal{Z}^{(0,1)} = \mathcal{Z}^{(1,0)} = \mathcal{Z}^{(0,0)} = (1,1)$. Observe that the weights associated to edges going to empty subtrees are irrelevant.

**Remark.** Our construction emphasizes an underlying structure consisting of independent random vectors associated with the nodes of $T_\infty$. However, in the *coupled* trie built from the embedding, the random variables $\{N_u : u \in T_\infty\}$ and $\{Z_e\}$ associated respectively with the nodes and the edges of $T_\infty$ are dependent: the values of $\{N_u : u \in T_\infty\}$ influence the types $\{\tau_u : u \in T_\infty\}$ which in turn influence the weights $\{Z_e\}$.

Let $\pi(u)$ be the set of edges on the path from $u$ to the root of $T_\infty$. The *weighted depth* of a node $u \in T_\infty$ is defined by $D_u = \sum_{e \in \pi(u)} Z_e$. We are interested in the weighted height of $T_{n,b}$ defined by

$$H_{n,b} = \max\{D_u : u \in T_{n,b}\}.$$

Surprisingly, if the weights are non-negative and bounded, $H_{n,b} \sim c_b \log n$ in probability, and $c_b$ depends only on

- the distribution $\{p_1, \dots, p_d\}$,
- the capacity $b$ of the leaves,
- the distribution of $\mathcal{Z}^{(1,\dots,1)}$, and
- the distributions of $\mathcal{Z}^\tau$, for all a permutations $\tau$ of $(1, 0, \dots, 0)$.

In particular, the first order asymptotics of $H_{n,b}$ stay the same if we modify the distribution of $\mathcal{Z}$ in such a way that the above parameters remain unchanged. In other words, the only nodes whose weights affect the first order term of the height are the ones having either $d$ children or a single child in $T_{n,b}$. This is easily understood by thinking of the structure of the shape of a trie.

## 2.2 Understanding the height using the structure of a trie

The reason why only the nodes with either one single child or $d$ children affect the first order asymptotics of the height is simply that the other types are negligible when looking at any path.

**Lemma 1.** *Let $T_{n,b}$ be a random trie. Let $m = m(n) \to \infty$ such that $m = o(\log n)$. There exists $\omega = \omega(n) \to \infty$, as $n \to \infty$ such that:*
*(a) with probability at least $1 - n^{-\omega}$, all the nodes $u$ with $N_u \ge \log^2 n$ have $\tau_u = (1, \dots, 1)$,*
*(b) the maximum number of nodes with $N_u \ge m(n)$ and $\tau_u \ne (1, \dots, 1)$ on a path from the root is $o(\log n)$ with probability at least $1 - n^{-\omega}$, and*

(c) *the maximum number of nodes with $N_u \le m(n)$ and degree at least two on a path from the root is at most $m(n) = o(\log n)$.*

*Proof.* Each time the degree is at least two, at least one string is put aside from the longest path. This can happen at most $m = o(\log n)$ times, and (c) follows. Therefore, we need only consider the portion of the paths visited by at least $m$ strings. The top of these paths, consisting of nodes $u$ with $N_u \ge \log^2 n$, is very likely to be free of any node with less than $d$ children: in this region, with probability $1 - o(1)$, *all* the nodes have $d$ children. For any node $u$, we have

$$\mathbf{P}\left\{\tau_u \neq (1,\dots,1) \mid N_u \ge \log^2 n\right\} \le d(1-p_d)^{\log^2 n}.$$

Moreover, the number of such nodes is polynomial in $n$. Indeed, writing $\mathscr{L}_k$ for the set of nodes at level $k$ in $T_\infty$, and setting for $k = \lceil \log_{1/p_1} n \rceil$,

$$\begin{aligned}\mathbf{P}\left\{\exists u \in \mathscr{L}_k : N_u \ge \log^2 n\right\} &\le d^{\log_{1/p_1} n+1} \cdot \mathbf{P}\left\{\mathrm{Bin}(n, p_1^k) \ge \log^2 n\right\}\\ &\le dn^{\log_{1/p_1} d} \cdot e^{-\frac{1}{2}\log^2 n},\end{aligned}$$

by the Chernoff bound (Chernoff, 1952). Therefore, by the union bound,

$$\begin{aligned}\mathbf{P}\left\{\exists u : N_u \ge \log^2 n, \tau_u \neq (1,\dots,1)\right\} &\le \mathbf{P}\left\{\exists u \in \mathscr{L}_j, j \le k : N_u \ge \log^2 n, \tau_u \neq (1,\dots,1)\right\}\\ &\quad + \mathbf{P}\left\{\exists u \in \mathscr{L}_k : N_u \ge \log^2 n\right\}\\ &\le 2dn^{\log_{1/p_1} d} \cdot (1-p_d)^{\log^2 n} + dn^{\log_{1/p_1} d} \cdot e^{-\frac{1}{2}\log^2 n}\\ &\le n^{-\omega_1}, \end{aligned} \tag{3}$$

for some $\omega_1 \to \infty$ as $n \to \infty$. This proves (a).

There is also a number of layers of nodes $u$ with $m(n) \le N_u < \log^2 n$. There are only $o(\log n)$ such layers in probability. To see this, let $\nu = \nu(n) \to \infty$ to be chosen later, and look at a node $v$, $\lceil \frac{1}{\nu} \log n \rceil$ levels below $u$ with $N_u \le \log^2 n$. Then,

$$\mathbf{P}\left\{N_v \ge m\right\} \le \mathbf{P}\left\{\mathrm{Bin}(\log^2 n, p_1^{\frac{1}{\nu}\log n}) \ge m\right\}. \tag{4}$$

The expected value of the binomial random variable above is

$$\ell = \log^2 n \cdot p_1^{\frac{1}{\nu}\log n} = \log^2 n \cdot n^{\frac{1}{\nu}\log p_1} \xrightarrow[n\to\infty]{} 0, \tag{5}$$

for $\nu = o(\log n / \log\log n)$. In particular, for $n$ large enough, $\ell \le m/2$. By the Chernoff bound for binomial random variables (see, e.g., Janson et al., 2000),

$$\mathbf{P}\left\{\mathrm{Bin}(\log^2 n, p_1^{\frac{a}{m}\log n}) \ge m\right\} \le \exp\left(-\ell\varphi\left(\frac{m}{2\ell}\right)\right), \tag{6}$$

where $\varphi(x) = (1+x)\log(1+x) - x$. Using (5), we see that, as $n \to \infty$,

$$\begin{aligned}\ell\varphi\left(\frac{m}{2\ell}\right) &= \left(\ell + \frac{m}{2}\right)\log\left(1 + \frac{m}{2\ell}\right) - \frac{m}{2}\\ &\sim \frac{m}{2}\cdot\log\left(\frac{m}{2\ell}\right)\\ &\sim \frac{m}{2}\log\left(\frac{m}{2}\right) - m\log\log n - \frac{m}{2\nu}\log p_1 \log n\\ &\sim \frac{m}{2\nu}\log\left(\frac{1}{p_1}\right)\log n,\end{aligned}$$

for $\nu = o(\log n/\log\log n)$. We now choose $\nu$ such that, in addition, $\nu = o(m)$ so that, by (4) and (6), $\mathbf{P}\left\{N_v \ge m\right\}$ decreases faster than any polynomial in $n$. The number of potential nodes $v$ is polynomial in $n$ since they lie $O(\log n)$ away from the root. It follows that the maximum number of levels between a node $u$ with $N_u \le \log^2 n$ and $v$ such that $N_v \le m$ is $O(\frac{\log n}{\nu}) = o(\log n)$ with probability at least $1 - n^{-\omega_2}$, for some $\omega_2 \to \infty$ as $n \to \infty$. With (3), this proves (b) with $\omega = \min\{\omega_1, \omega_2\}/2$. □

Lemma 1 justifies the distinction of two regions in a trie: a so-called *core*, that essentially consists of the nodes of out degree $d$, and *spaghetti*-like trees hanging from the core (see Broutin and Devroye, 2007a).

THE CORE OF A TRIE. What we call the core here should not be confused with the graph-theoretic core, which happens to be empty for trees (see, e.g., Janson et al., 2000). The core of the trie is defined to be the set of nodes $u \in T_\infty$ for which $N_u \geq m = m(n)$, for $m(n) \to \infty$ and $m(n) = o(\log n)$. The core is denoted by $\mathcal{C}$. By Lemma 1, on any path from the root, the number of nodes in the core which are not of type $\tau = (1, \dots, 1)$ is $o(\log n)$ in probability. As a consequence, when looking at a path of length $\Theta(\log n)$ in a weighted trie, the distribution of weights in the core should be closely approximated by $\mathcal{Z}^c = \mathcal{Z}^{(1,\dots,1)}$. The core can be described by its *logarithmic profile*

$$\phi(\alpha, t) = \lim_{n\to\infty} \frac{\log \mathbf{E} P_m(t \log n, \alpha \log n)}{\log n} \qquad \forall t, \alpha > 0, \tag{7}$$

where $P_m(k, h)$ denotes the number of nodes $u \in T_\infty$, $k$ levels away from the root with $N_u \geq m$ and $D_u \geq h$. In other words, assuming for now that the limit in (7) exists, we have $\mathbf{E} P_m(t \log n, \alpha \log n) = n^{\phi(\alpha,t)+o(1)}$, as $n \to \infty$. This will be proved, and the function $\phi(\cdot, \cdot)$ will be characterized in Section 4 (Theorem 3).

HANGING SPAGHETTIS. The *spaghettis* are the trees remaining when pulling out the core from the trie. They lie in the part of the trie where the nodes do not have $d$ children with high probability anymore: the types of the nodes may take all the values in $\{0, 1\}^d$. However, by Lemma 1, in any spaghetti, the number of nodes with at least two children is $o(\log n)$. Since the weights are bounded, these $o(\log n)$ terms contribute at most $o(\log n)$ to the height. Therefore, to first order, only the nodes of degree one affect the height. This explains why only $\mathcal{Z}^\tau$ for $\tau$ a permutation of $(1, 0, \dots, 0)$ matter in the weighted heights of spaghettis.

UNDERSTANDING THE HEIGHT. Both the core and the spaghettis contribute significantly to the height of a weighted trie. By figuring out what the core looks like, we can determine *when* the spaghettis take over. Roughly speaking, we then know if an edge's weight can be approximated by a component of $\mathcal{Z}^{(1,\dots,1)}$, characteristic of the core, or rather $\mathcal{Z}^\tau$, for $\tau$ a permutation of $(1, 0, \dots, 0)$, characteristic of the spaghettis. Each spaghetti is rooted at a node $u \in \partial\mathcal{C}$, the external node-boundary of the core $\mathcal{C}$ in $T_{n,b}$ (the nodes $u \in \partial\mathcal{C}$ are the children of some node $v$ in the core, but are not themselves in the core). Recall that $\mathscr{L}_k$ denotes the set of nodes at level $k$ in $T_\infty$. Then, if we write $W_u$ for the (weighted) height of the subtree rooted at $u$, we have

$$H_{n,b} = \max\{D_u + W_u : u \in \partial\mathcal{C}\} = \sup_{h,k}\{h + W_u : u \in \partial C \cap \mathscr{L}_k, D_u \geq h\},$$

where the nodes in $\partial C$ have been split into groups depending on their level $k$ and weighted depth $h$. Thus, we can rewrite

$$H_{n,b} = \sup_{h,k}\{h + \max\{W_u : u \in \partial\mathcal{C}, D_u \geq h, u \in \mathscr{L}_k\}\}.$$

We have separated the contributions of the core and the spaghettis,

$$h \qquad \text{and} \qquad \max\{W_u : u \in \partial\mathcal{C}, D_u \geq h, u \in \mathscr{L}_k\},$$

respectively. The height is simply the maximum value of the sum of these two terms, and we need to characterize them in order to pin down the asymptotic value of the height. The first term is just $h \sim \alpha \log n$, a parameter. The second one amounts to studying the weighted height of the *forest of conditionally independent random tries* rooted at the nodes $u \in \mathscr{L}_k$ with $D_u \geq h$. When $k \sim t \log n$ and $h \sim \alpha \log n$, we have

$$\lim_{n\to\infty} \frac{\max\{W_u : u \in \partial\mathcal{C}, D_u \geq h, u \in \mathscr{L}_k\}}{\log n} = \phi(\alpha, t) \cdot \gamma_b \quad \text{in probability,} \tag{8}$$

for some constant $\gamma_b$ characterizing the depths in spaghettis. This will be proved, and $\gamma_b$ will be characterized in Section 5 (Theorem 6) where we study forests of independent random tries. The two parameters $\phi(\cdot,\cdot)$ and $\gamma_b$ suffice to obtain the first order term of the asymptotic expansion of the height. Our main result is the following theorem.

**Theorem 1.** *Consider $T_{n,b}$, a weighted b-trie with non-negative weight sequence $\{\mathcal{Z}^\tau, \tau \in \{0,1\}^d\}$ built from n independent sequences with distribution $\{p_1, \dots, p_d\}$. Let $H_{n,b}$ be its weighted height. Assume that the weights $\mathcal{Z}^\tau$ are bounded. Let $\phi(\alpha, t)$ be the logarithmic weighted profile of the core of $T_{n,b}$ defined in (7), and $\gamma_b$ the constant defined in (8). Let*

$$c_b = \sup\left\{\alpha + \gamma_b \cdot \phi(\alpha, t) : \alpha, t > 0\right\}.$$

*Then $H_{n,b} = c_b \log n + o(\log n)$ in probability, as $n \to \infty$.*

**Remarks.** (a) For Theorem 1 to be useful in applications, we show that $\phi(\cdot,\cdot)$ and $\gamma_b$ are computable in Theorems 3 and 6, respectively.
(b) The definition of $c_b$ given makes it clear that, as long as the weights take positive values with positive probability, $c_b > 0$ is well and uniquely defined. We will see later that $c_b < \infty$.

The rest of the paper is organized as follows. The core and spaghettis are analyzed in detail in Sections 4 and 5, respectively. In Section 6, we prove Theorem 1. Finally, we present some applications in Section 7, with in particular, the heights of the trees of de la Briandais (1959) and of the TST of Bentley and Sedgewick (1997). The proofs are based on large deviations, in particular $\phi(\cdot,\cdot)$ and $\gamma_b$ are characterized in terms of large deviation rate functions.

# 3 Review of large deviations

In this section, we review large deviation theory. For a more complete treatment, see the textbooks of Deuschel and Stroock (1989), Dembo and Zeitouni (1998), or den Hollander (2000). We are interested in the case of *extended* random vectors $(Z, E)$, that is, for which $Z = -\infty$ may happen with positive probability. The reason will become clear when we analyze the behavior of spaghettis in Section 5. In the following, for a function $f$ taking values in $(-\infty, \infty]$, we define its domain $\mathcal{D}_f = \{x : f(x) < \infty\}$, with interior $\mathcal{D}_f^o$.

Let $\{X_i, 1 \leq i \leq n\}$ be a family of i.i.d. extended random vectors distributed like $(Z, E)$. Assume $Z \in [-\infty, \infty)$ and $E \in [0, \infty)$. Set $\kappa = \mathbf{P}\{Z > -\infty\}$. For $\alpha$ and $\rho$ real numbers, we are interested in the tail probability

$$\mathbf{P}\left\{\sum_{i=1}^n Z_i > \alpha n, \sum_{i=1}^n E_i < \rho n\right\}, \tag{9}$$

whose magnitude is dealt with by Cramér's theorem (Cramér, 1938). Define the cumulant generating function $\Lambda$ of the (extended) random vector $(Z, E)$ by

$$\Lambda(\lambda, \mu) = \log \mathbf{E}\left[\, e^{\lambda Z + \mu E} \,\middle|\, Z > -\infty \right] + \log \kappa \qquad \forall \lambda, \mu \in \mathbb{R}.$$

The tail probability in (9) is characterized using $\Lambda^\star(\cdot,\cdot)$, the Fenchel–Legendre or convex dual of $\Lambda$ (see Rockafellar, 1970): we define

$$\Lambda^\star(\alpha, \rho) = \sup_{\lambda,\mu}\{\lambda\alpha + \mu\rho - \Lambda(\lambda, \mu)\} \quad \forall \alpha, \rho \in \mathbb{R}.$$

**Theorem 2** (**Cramér**, see Dembo and Zeitouni 1998)**.** *Assume that $\{X_i, i \geq 1\}$ are i.i.d. random vectors, and that $(0,0) \in \mathcal{D}_\Lambda^o$. For $\alpha, \rho \in \mathbb{R}$, let $I(\alpha,\rho) \stackrel{\text{def}}{=} -\inf\{\Lambda^\star(x,y) : x > \alpha, y < \rho\}$. Then for any $\alpha, \rho \in \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left\{\sum_{i=1}^n Z_i > \alpha n, \sum_{i=1}^n E_i < \rho n\right\} = -I(\alpha,\rho).$$

*Moreover, the following explicit upper bound holds for all $n \geq 1$, and $\alpha, \rho \in \mathbb{R}$:*

$$\mathbf{P}\left\{\sum_{i=1}^n Z_i > \alpha n, \sum_{i=1}^n E_i < \rho n\right\} \leq e^{-nI(\alpha,\rho)}.$$

**Remark.** The explicit upper bound is analogous to the Chernoff bound (Chernoff, 1952) and holds because the quadrant $(\alpha,\infty) \times (0,\rho)$ is a convex set (see Exercise 2.2.38, p. 42, Dembo and Zeitouni, 1998).

Let $a \wedge b$ denote the minimum of two numbers $a$ and $b$ taking values in $\mathbb{R} \cap \{-\infty, \infty\}$. The rate function $I(\cdot,\cdot)$ has the following properties. For a proof, see Dembo and Zeitouni (1998).

**Lemma 2.** *(a) The function $I$ is convex,*
*(b) $I$ is lower-semicontinous: the level sets $\{(\alpha,\rho) : I(\alpha,\rho) \leq \ell\}$ are closed for all $\ell \in \mathbb{R}$,*
*(c) for any $x \in \mathbb{R}$, the function $x \wedge I(\cdot,\cdot)$ is continuous on $\mathbb{R}^2$.*

# 4 The core of a weighted trie

## 4.1 Asymptotic Behavior

Consider a weighted $b$-trie with weight distribution sequence $\{\mathcal{Z}^\tau : \tau \in \{0,1\}^d\}$ as defined as in Section 2. We consider $m = m(n) \to \infty$ with $m(n) = o(\log n)$. Let $\mathscr{L}_k$ be the set of nodes $k$ levels away from the root in $T_\infty$. Let $P_m(k,h)$ be the profile, i.e., the number of nodes $u \in \mathscr{L}_k$ with $D_u \geq h$ and $N_u \geq m$. Since $m \to \infty$, for $n$ large enough, we have $m \geq b$ and

$$P_m(k,h) = \sum_{u \in \mathscr{L}_k} \mathbf{1}[N_u \geq m, D_u \geq h].$$

The first step in characterizing the profile is to study its expected value. We then use some concentration arguments. The asymptotic properties of the expected profile are directly tied to large deviation theory (Dembo and Zeitouni, 1998). We have seen that the weights in the core should be closely approximated by $\mathcal{Z}^c = \mathcal{Z}^{(1,\dots,1)} = (Z_1^c, \dots, Z_d^c)$ by Lemma 1. Then, the random vector of interest here is $(Z, E) = (Z_K^c, -\log p_K)$, where $K$ is uniform in $\{1,\dots,d\}$ and $\mathcal{Z}^c = (Z_1^c, \dots, Z_d^c)$. Let $I(\cdot,\cdot)$ be the rate function associated with $(Z,E)$ appearing in Cramér's theorem (Theorem 2). For $a, b \in \mathbb{R} \cup \{-\infty,\infty\}$, let $a \vee b$ denote the maximum of $a$ and $b$.

**Theorem 3.** *Let $m = m(n) \to \infty$ with $m = o(\log n)$. For $\alpha \geq 0$, let $\phi(\alpha, 0) = 0$ and*

$$\forall t > 0 \qquad \phi(\alpha,t) = t \log d - t \cdot I\left(\frac{\alpha}{t}, \frac{1}{t}\right). \tag{10}$$

*If $\alpha, t \geq 0$ and $\phi(\alpha,t) > -\infty$, then $\mathbf{E}P_m(\lfloor t \log n \rfloor, \alpha \log n) = n^{\phi(\alpha,t)+o(1)}$, as $n \to \infty$. Moreover, for any $\epsilon > 0$, and any $a \in \mathbb{R}$,*

$$\exists n_o : \forall n \geq n_o, \forall \alpha, t > 0 \qquad \mathbf{E}P_m(\lfloor t \log n \rfloor, \alpha \log n) \leq n^{a \vee \phi(\alpha,t)+\epsilon}.$$

**Remarks.** (a) Observe that Theorem 3 justifies the definition of $\phi(\cdot,\cdot)$ in (7).
(b) The constraint that $m(n)$ is $o(\log n)$ is only used in the lower bound. However, the main reason why we choose $m = o(\log n)$ is for the spaghettis to contain each only $o(\log n)$ nodes of degree at least two.
(c) If $k \sim t \log n$ and $h \sim \alpha \log n$, we also have $\mathbf{E}P_m(k,h) = n^{\phi(\alpha,t)+o(1)}$. The proof is slightly more technical but does not shed any new light.

Unlike the profile of *unweighted* tries (Devroye, 2002, 2005; Park et al., 2006), that of *weighted* tries does not seem concentrated. In regular (unweighted) tries, the modification of one single sequence may affect $P_m(k,0)$ by at most one, and contrasts with the case of weighted tries where $P_m(k,h)$ may potentially change a lot. However, it is log-concentrated in the sense of the following theorem.

**Theorem 4.** *Let $m = m(n) \to \infty$ as $n \to \infty$ such that $m = o(\log n)$. Let $k \sim t \log n$ and $h \sim \alpha \log n$ for some positive constants $t$ and $\alpha$. Then, for all $\epsilon > 0$, as $n \to \infty$,*

$$\mathbf{P}\left\{P_m(k,h) \le n^{\phi(\alpha,t)-\epsilon}\right\} \xrightarrow[n\to\infty]{} 0.$$

*For all $a \in \mathbb{R}$, and all $\epsilon > 0$, there exists $n_o$ large enough such that*

$$\forall n \ge n_o \qquad \sup_{\alpha,t\ge 0} \mathbf{P}\left\{P_m(\lfloor t \log n\rfloor, \alpha \log n) \ge n^{a\vee\phi(\alpha,t)+\epsilon}\right\} \le n^{-\epsilon/2}.$$

**Remark.** In the upper bound, the use of $a \vee \phi(\alpha,t)$, $a \in \mathbb{R}$, is necessary since it is possible that $\phi(\alpha,t) = -\infty$. In such a case $n^{\phi(\alpha,t)+\epsilon} = 0$ for all $n \ge 2$, and of course,

$$\mathbf{P}\left\{P_m(\lfloor t \log n\rfloor, \alpha \log n) \ge n^{\phi(\alpha,t)+\epsilon}\right\} = 1.$$

**Lemma 3.** *Let $\phi(\cdot,\cdot)$ be the logarithmic profile as defined in (10). Then,*
(a) *the domain $\mathcal{D}_\phi \stackrel{\text{def}}{=} \{(\alpha,t) : \alpha, t > 0, \phi(\alpha,t) > -\infty\}$ of $\phi$ is bounded,*
(b) *$\phi(\cdot,\cdot)$ is concave, and*
(c) *$\phi(\cdot,\cdot)$ is continuous on $\mathcal{D}_\phi$, and for all $a \in \mathbb{R}$, $a \vee \phi(\cdot,\cdot)$ is continuous on $[0,\infty)^2$.*

*Proof.* We prove (a), the rest follows from Lemma 2. For all $x$ and $\rho$, we have

$$\Lambda^\star(x,\rho) = \sup_{\lambda,\mu}\{\lambda x + \mu\rho - \Lambda(\lambda,\mu)\} \ge \sup_{\lambda}\{\mu\rho - \Lambda(0,\mu)\}.$$

We find a lower bound on the cumulant generating function. For all $\mu < 0$,

$$\Lambda(0,\mu) = \log \mathbf{E}\left[e^{\mu E}\right] \ge -\mu \log p_1.$$

As a consequence, we see that for $\rho < -\log p_1$, and all $x$, taking $\mu \to -\infty$, $\Lambda^\star(x,\rho) = \infty$. Since $\phi(\alpha,t) = t\log d - \Lambda^\star(\alpha/t, 1/t)$, the result follows. $\square$

**Example: asymmetric randomized list-tries.** Consider asymmetric tries on $\{1,2\}$ with $p_1 = p > 1/2$ and $p_2 = q = 1-p$. A fair coin is flipped independently at each node to decide whether the character 1 or 2 would be first in the list. Therefore, the vector $\mathcal{Z}^c = (Z_1, Z_2)$ of search costs is such that $Z_1$ and $Z_2$ take values 1 or 2 with equal probability. In this example, the variables $E$ and $Z$ are independent and they are both linear functions of Bernoulli random variables (see Dembo and Zeitouni, 1998, Section 4.2 on transformations of large deviation functions). If we write $f(y) = y\log y + (1-y)\log(1-y) + \log 2$, then $\Lambda^\star(x,\rho) = \Lambda^\star_Z(x) + \Lambda^\star_E(\rho)$, where

$$\Lambda^\star_Z(\alpha) = f(\alpha - 1) \qquad \text{and} \qquad \Lambda^\star_E(\rho) = f\left(\frac{\rho + \log p}{\log p - \log q}\right).$$

The corresponding logarithmic profile $\phi(\cdot,\cdot)$ shown in Figure 3 is taken from this example.

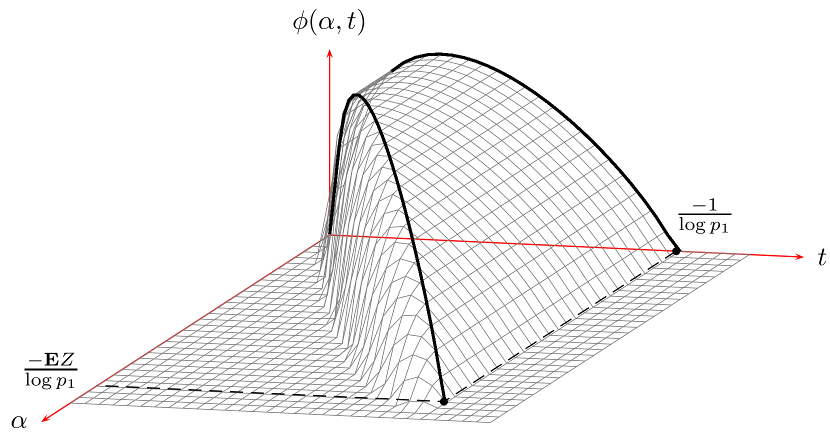


**Figure 3.** A typical logarithmic profile $0 \vee \phi(\alpha, t)$. The thick black lines represent $\phi(0, t)$ and $\phi(t\mathbf{E}Z, t)$. For $t_o$ constant, $\phi(\alpha, t_o)$ is constant for $\alpha \in [0, t_o\mathbf{E}Z]$.

## 4.2 The expected profile: Proof of Theorem 3

### 4.2.1 The upper bound

**Lemma 4.** *Let $m = m(n) \to \infty$. Let $\phi(\alpha, t)$ be given by (10). Then, for any $\epsilon > 0$ and any $a \in \mathbb{R}$, there exists $n_o$ large enough such that for all $n \geq n_o$,*

$$\forall \alpha, t > 0 \qquad \mathbf{E}P_m(\lfloor t \log n \rfloor, \alpha \log n) \leq n^{a \vee \phi(\alpha,t)+\epsilon}.$$

*Proof.* Let $\epsilon > 0$. When $t$ is small, there are not enough nodes in the $k$-th level $\mathscr{L}_k$ of $T_\infty$. For all $t < \epsilon / \log d$ and all $\alpha \geq 0$,

$$\mathbf{E}P_m(\lfloor t \log n \rfloor, \alpha \log n) \;\leq\; d^{t \log n} \;<\; d^{\epsilon \log_d n} \;=\; n^{\epsilon}.$$

Furthermore, if $\phi(\alpha, t) \geq 0$, $\epsilon \leq \phi(\alpha, t) + \epsilon \leq a \vee \phi(\alpha, t) + \epsilon$, and

$$\forall (\alpha, t) \in \mathcal{B} \qquad \mathbf{E}P_m(\lfloor t \log n \rfloor, \alpha \log n) \leq n^{a \vee \phi(\alpha,t)+\epsilon}, \tag{11}$$

where $\mathcal{B}$ is a small enough ball around the origin. It suffices now to consider the range $(\alpha, t) \notin \mathcal{B}$. In particular, in the rest of the proof, $t$ is bounded away from 0. Consider a *uniformly random* path $\{u_0, u_1, \dots, u_k, \dots\}$ in $T_\infty$: $u_0$ is the root, and for any integer $i \geq 0$, $u_{i+1}$ is a child of $u_i$ picked uniformly at random. Note that for $k \geq 0$, $u_k$ is a uniform node in $\mathscr{L}_k$, the set of nodes $k$ levels away from the root in $T_\infty$. Let

$$L_{u_k} = \prod_{e \in \pi(u_k)} p_e = \prod_{e \in \pi(u_k)} e^{-E_e}. \tag{12}$$

By definition of $P_m(k, h)$, we have

$$\mathbf{E}P_m(k, h) \;=\; d^k \cdot \mathbf{P}\left\{\mathrm{Bin}(n, L_{u_k}) \geq m, D_{u_k} \geq h\right\}.$$

The randomness coming from the binomial random variables is irrelevant for the order of precision we are after. Indeed, for any $\xi_1 \in [0, 1]$, we have

$$\mathbf{E}P_m(k, h) \;\leq\; d^k \cdot \mathbf{P}\left\{L_{u_k} \geq \xi_1, D_{u_k} \geq h\right\} + d^k \cdot \sup_{\xi \leq \xi_1} \mathbf{P}\left\{\mathrm{Bin}(n, \xi) \geq m\right\}. \tag{13}$$

In particular, if we set

$$\xi_1 = \frac{m d^{-k/m}}{e n^{1+1/\sqrt{m}}}, \tag{14}$$

the second term of (13) is easily bounded as follows

$$\sup_{\xi\le\xi_1} \mathbf{P}\left\{\mathrm{Bin}(n,\xi)\ge m\right\} \;\le\; \mathbf{P}\left\{\mathrm{Bin}(n,\xi_1)\ge m\right\} \;\le\; \binom{n}{m}{\xi_1}^m \;\le\; \left(\frac{en\xi_1}{m}\right)^m \;=\; \frac{d^{-k}}{n^{\sqrt{m}}}.$$

As a consequence, by definition of $D_{u_k}$ and (12),

$$\mathbf{E}P_m(k,h) \;\le\; d^k\cdot\mathbf{P}\left\{\sum_{e\in\pi(u_k)} Z_e\ge h,\ \sum_{e\in\pi(u_k)} E_e\le -\log\xi_1\right\}+\frac{1}{n^{\sqrt{m}}}. \tag{15}$$

There exists a constant $A>0$ such that, for all $t\le t_o$ and $\alpha\le\alpha_o$, we have, by (14),

$$-\log\xi_1 \;=\; \left(1+\frac{1}{\sqrt{m}}\right)\log n+1-\log m+\frac{\lfloor t\log n\rfloor}{m}\log d \;\le\; \left(1+\frac{A}{\sqrt{m}}\right)\log n,$$

for $n$ large enough. Hence, rewriting (15), we have

$$\mathbf{E}P_m(k,h) \;\;\le\;\; d^k\cdot\mathbf{P}\left\{\sum_{e\in\pi(u_k)} Z_e\ge h,\ \sum_{e\in\pi(u_k)} E_e\le\left(1+\frac{A}{m}\right)\log n\right\}+\frac{1}{n^{\sqrt{m}}}, \tag{16}$$

since $m=m(n)\to\infty$. By assumption, $\{E_e, e\in\pi(u_k)\}$ is a family of i.i.d. random variables. It is *not* the case for $\{Z_e, e\in\pi(u_k)\}$, and hence, not for $\{(Z_e,E_e), e\in\pi(u_k)\}$ either. However, by Lemma 1, the maximum number of nodes with less than $d$ children lying on a path down the root with $N\ge m(n)$ is $o(\log n)$ with probability $1-n^{-\omega}$, for some $\omega\to\infty$ as $n\to\infty$. Let $(Z_i^c,E_i)$, $i\ge1$, be i.i.d. random vectors distributed like $(Z^c,E)$. Then, for any $\delta>0$, and all $n$ large enough,

$$\mathbf{E}P_m(k,h)\le d^k\cdot\mathbf{P}\left\{\sum_{i=1}^k Z_i^c\ge\frac{\alpha}{t}\cdot k,\sum_{i=1}^k E_i\le\left(\frac{1}{t}+\delta\right)k\right\}+\frac{1}{n^{\sqrt{m}}}+\frac{d^k}{n^\omega}, \tag{17}$$

for $n$ large enough. Therefore, by Cramér's theorem (Theorem 2), we have, for any $\delta>0$, and $n$ large enough,

$$\mathbf{E}P_m(k,h)\le\exp\left(k\log d-kI\left(\frac{\alpha}{t},\frac{1}{t}+\delta\right)\right)+\frac{1}{n^{\sqrt{m}}}+\frac{d^k}{n^\omega}. \tag{18}$$

Recall that, by Lemma 3, $\mathcal{D}_\phi=\{(\alpha,t):\phi(\alpha,t)>-\infty\}$ is bounded. It follows that for $(\alpha,t)$ outside a slight compact blow-up $\mathcal{S}$ of $\mathcal{D}_\phi$, $I(\alpha/t,1/t+\delta)=\infty$. Then, $a\vee\phi(\alpha,t)=a$, for any $a\in\mathbb{R}$, and

$$\forall(\alpha,t)\notin\mathcal{S}\qquad \mathbf{E}P_m(\lfloor t\log n\rfloor,\alpha\log n)\;\le\; n^{a\vee\phi(\alpha,t)}+\frac{1}{n^{\sqrt{m}}}+\frac{d^k}{n^\omega}\;=\;n^{a\vee\phi(\alpha,t)+\epsilon}, \tag{19}$$

for $n$ large enough. It only remains to deal with the range $(\alpha,t)\in\mathcal{S}\setminus\mathcal{B}$. By (18), for any $x\in\mathbb{R}$,

$$\mathbf{E}P_m(k,h)\le\exp\left(k\log d-k\left[x\wedge I\left(\frac{\alpha}{t},\frac{1}{t}+\delta\right)\right]\right)+\frac{1}{n^{\sqrt{m}}}+\frac{d^k}{n^\omega}.$$

By Lemma 2, the function $x\wedge I(\cdot,\cdot)$ is continuous on $\mathbb{R}^2$, and uniformly continuous on $\mathcal{S}\setminus\mathcal{B}$. Thus, for any $\eta>0$, there exists $n$ large enough such that for all $(\alpha,t)\in\mathcal{S}\setminus\mathcal{B}$,

$$\begin{aligned}\mathbf{E}P_m(\lfloor t\log n\rfloor,\alpha\log n)\;&\le\;\exp\left(k\log d-k\left[x\wedge I\left(\frac{\alpha}{t},\frac{1}{t}\right)+\eta\right]\right)+\frac{1}{n^{\sqrt{m}}}+\frac{d^k}{n^\omega}\\&\le\;\exp\left(k\log d-k\left[x\wedge I\left(\frac{\alpha}{t},\frac{1}{t}\right)\right]+\frac{\epsilon}{2}\log n\right)+\frac{1}{n^{\sqrt{m}}}+\frac{d^k}{n^\omega},\end{aligned}$$

for $\eta$ small enough. Then,

$$\mathbf{E}P_m(\lfloor t\log n\rfloor, \alpha\log n) \le n^{tx\log d\vee\phi(\alpha,t)+\epsilon/2} + \frac{1}{n^{\sqrt{m}}} + \frac{d^k}{n^\omega}.$$

Choosing $x$ such that $tx\log d < a$ if $(\alpha,t)\in\mathcal{S}\setminus\mathcal{B}$, we obtain that, for $n$ large enough (independent of $\alpha$ and $t$),

$$\forall(\alpha,t)\in\mathcal{S}\setminus B \qquad \mathbf{E}P_m(\lfloor t\log n\rfloor, \alpha\log n) \le n^{a\vee\phi(\alpha,t)+\epsilon}. \tag{20}$$

Putting (11), (19) and (20) together proves the claim. This finishes the proof the lemma and the upper bound of Theorem 3. $\square$

#### 4.2.2 The lower bound

**Lemma 5.** *Let $m = m(n)\to\infty$ with $m = o(\log n)$. Let $\alpha, t\ge 0$ such that $\phi(\alpha,t) > -\infty$, then $\mathbf{E}P_m(\lfloor t\log n\rfloor, \alpha\log n) \ge n^{\phi(\alpha,t)+o(1)}$, as $n\to\infty$.*

*Proof.* If $t = 0$, $P_m(\lfloor t\log n\rfloor, \alpha\log n) = 1 = n^{o(1)} = n^{\phi(\alpha,t)+o(1)}$, by definition of $\phi$. We now assume that $t > 0$. We write $k = \lfloor t\log n\rfloor$ and $h = \alpha\log n$. As in the proof of the upper bound, let $u_k$ be a random node in $\mathscr{L}_k$, the set of nodes $k$ levels away from the root in $T_\infty$. We have

$$\mathbf{E}P_m(k,h) \quad = \quad d^k\cdot\mathbf{P}\left\{N_{u_k}\ge m, D_{u_k}\ge h\right\}.$$

By definition $N_{u_k}$ is distributed as $\mathrm{Bin}(n, L_{u_k})$, where $L_{u_k}$ is defined in (12). As a consequence, for any $\xi_2$,

$$\mathbf{E}P_m(k,h) \;\ge\; d^k\cdot\mathbf{P}\left\{L_{u_k}\ge\xi_2, D_{u_k}\ge h\right\}\cdot\inf_{\xi\ge\xi_2}\mathbf{P}\left\{\mathrm{Bin}(n,\xi)\ge n\right\},$$

Choosing $\xi_2 = m/n$, we see that

$$\begin{aligned}\inf_{\xi\ge\xi_2}\mathbf{P}\left\{\mathrm{Bin}(n,\xi)\ge m\right\} &= \mathbf{P}\left\{\mathrm{Bin}(n,\xi_2)\ge m\right\}\\ &\ge \mathbf{P}\left\{\mathrm{Bin}(n,\xi_2)\ge\mathbf{E}\mathrm{Bin}(n,\xi_2)\right\} \;=\; n^{o(1)},\end{aligned}$$

and it follows that

$$\begin{aligned}\mathbf{E}P_m(k,h) &\ge d^k\cdot\mathbf{P}\left\{L_{u_k}\ge\xi_2, D_{u_k}\ge h\right\}\cdot n^{o(1)}\\ &\ge d^k\cdot\mathbf{P}\left\{\sum_{e\in\pi(u_k)}E_e\le-\log\xi_2, \sum_{e\in\pi(u_k)}Z_e\ge h\right\}\cdot n^{o(1)}.\end{aligned}\tag{21}$$

The random vectors $(Z_e, E_e)$, $e\in\pi(u_k)$ are not i.i.d., since the path is likely to contain nodes of various types. We write:

$$\mathbf{E}P_m(k,h)\ge d^k\cdot\mathbf{P}\left\{\sum_{e\in\pi(u_k)}E_e\le-\log\xi_2, \sum_{e\in\pi(u_k)}Z_e\mathbf{1}[Z_e = Z_e^{(1,\dots,1)}]\ge h\right\}\cdot n^{o(1)}.$$

Let $(Z_i^c, E_i)$, $i\ge 1$, be i.i.d. vectors distributed like $(Z^{(1,\dots,1)}, E)$. By Lemma 1, there exists $\ell = o(\log n)$ such that the number of nodes along $\pi(u_k)$ with a type different from $(1,\dots,1)$ is at most $\ell$ with probability $1-o(1)$. Recall that $Z^c\ge -a$ for some $a\ge 0$. Thus,

$$\begin{aligned}\mathbf{E}P_m(k,h) &\ge d^k\cdot n^{o(1)}\cdot\mathbf{P}\left\{\sum_{i=1}^{k-\ell}Z_i^c\ge h+a\ell, \sum_{i=1}^{k}E_i\le-\log\xi_2\right\}\cdot(1-o(1))\\ &\ge d^k\cdot n^{o(1)}\cdot\mathbf{P}\left\{\sum_{i=1}^{k-\ell}Z_i^c\ge h+a\ell, \sum_{i=1}^{k-\ell}E_i\le-\log\xi_2+\ell\log p_d\right\},\end{aligned}$$

since $E \leq -\log p_d$. Recall that $-\log \xi_2 = (1+o(1)) \log n$, $k = \lfloor t \log n \rfloor$, $h = \alpha \log n$, and $\ell = o(\log n)$. Thus, for any $\delta > 0$, there exists $n$ large enough such that

$$\mathbf{E} P_m(k,h) \geq d^k \cdot n^{o(1)} \cdot \mathbf{P}\left\{ \sum_{i=1}^{k-\ell} Z_i^c \geq \left(\frac{\alpha}{t} + \delta\right)(k-\ell), \sum_{i=1}^{k-\ell} E_i \leq \left(\frac{1}{t} - \delta\right)(k-\ell) \right\}.$$

By Cramér's theorem (Theorem 2) and (21), this yields,

$$\mathbf{E} P_m(k,h) \;\geq\; d^k \cdot \exp\left(-kI\left(\frac{\alpha}{t} + \delta, \frac{1}{t} - \delta\right) + o(k)\right) \cdot n^{o(1)},$$

for any $\delta > 0$ and $n$ large enough. Now by assumption, $\phi(\alpha, t) > -\infty$ and hence $I(\alpha/t, 1/t) < \infty$. Since $\delta$ is arbitrary and $I(\cdot,\cdot)$ is continuous where it is finite, the claim is proven:

$$\mathbf{E} P_m(\lfloor t \log n \rfloor, \alpha \log n) \;\geq\; n^{\phi(\alpha,t)+o(1)},$$

where $\phi(\alpha, t)$ is given by (10). $\square$

### 4.3 Log-concentration of the profile: Proof of Theorem 4

The upper bound is straightforward if we combine Markov's inequality and the statement of Theorem 3. Let $a \in \mathbb{R}$. By Markov's inequality, for all $\alpha, t > 0$,

$$\mathbf{P}\left\{P_m(\lfloor t \log n \rfloor, \alpha \log n) \geq n^{a \vee \phi(t,\alpha)+\epsilon}\right\} \;\leq\; \frac{\mathbf{E} P_m(\lfloor t \log n \rfloor, \alpha \log n)}{n^{a \vee \phi(\alpha,t)+\epsilon}}.$$

By the uniform upper bound of Theorem 3, there exists $n_o$ large enough such that for all $n \geq n_o$ and for all $\alpha, t \geq 0$, we have $\mathbf{E} P_m(\lfloor t \log n \rfloor, \alpha \log n) \leq n^{a \vee \phi(\alpha,t)+\epsilon/2}$. It follows that

$$\sup_{\alpha,t\geq 0} \mathbf{P}\left\{P_m(\lfloor t \log n \rfloor, \alpha \log n) \geq n^{a \vee \phi(t,\alpha)+\epsilon}\right\} \;\leq\; n^{-\epsilon/2},$$

for such $n$, independently of $t$ or $\alpha$. We now focus on the lower bound. We first prove a weaker version that we will boost using standard techniques.

**Lemma 6.** *Let $\alpha, t > 0$ such that $\phi(\alpha, t) > 0$. Let $k \sim t \log n$ and $h \sim t \log n$. For any $\epsilon > 0$,*

$$\limsup_{n\to\infty} \mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} < 1.$$

*Proof.* From the previous section, we recall that all but $o(\log n)$ nodes have $d$ children on any path down the root in $T_{n,b}$. Then, all but the corresponding $o(\log n)$ random vectors are i.i.d.. We use a similar argument to relate the profile $P_m(k,h)$ to a Galton–Watson process. We construct our Galton–Watson tree using the variables $(Z^c, E) = (Z^{(1,\dots,1)}, E)$ of the embedding.

Let $\epsilon > 0$. By assumption, $\phi(\alpha, t) > 0$ and $I(\alpha/t, 1/t) < \log d$. Since the level sets of $I(\cdot,\cdot)$ are closed (Lemma 2), there exists an open ball $\mathcal{B}$ with center $(\alpha, t)$ such that $I(\alpha'/t', 1/t') < \log d$, for all $(\alpha', t') \in \mathcal{B}$. We enforce further the constraints: $\alpha' > \alpha$, $t' > t$ and $\alpha'/t' > \alpha/t$. For a node $u \in T_\infty$, let $L_u = \prod_{e\in\pi(u)} p_e$. Let $\ell$ be a natural number to be chosen later. The individuals of our process are some of the nodes of $\mathscr{L}_{s\ell}$, $s \geq 0$. A node $u$ is called *good* if either it is the root, or it lies $\ell$ levels below a good node $v$ and we have

$$D_u^c > D_v^c + \frac{\alpha'\ell}{t'} \qquad \text{and} \qquad L_u > L_v \cdot e^{-\ell/t'},$$

where $D_u^c = \sum_{e\in\pi(u)} Z_e^c$. The set of good nodes is a Galton–Watson process. For an integer $s \geq 0$, let $G_s$ be the number of good nodes in the $s$-th generation (at level $s\ell$ in $T_\infty$). Let $Y$ denote the progeny of an individual of the process. Then, the expected progeny is

$$\mathbf{E} Y \;=\; d^\ell \mathbf{P}\left\{ \sum_{e\in\pi(u_\ell)} Z_e^c > \frac{\alpha'\ell}{t'}, \sum_{e\in\pi(u_\ell)} E_e < \frac{\ell}{t'} \right\}.$$

Hence, by Cramér's theorem (Theorem 2),

$$\mathbf{E}Y \ \geq \ d^\ell \cdot \exp\left(-\ell I\left(\frac{\alpha'}{t'}, \frac{1}{t'}\right) + o(\ell)\right) \ = \ \exp\left(\ell \log d - \ell I\left(\frac{\alpha'}{t'}, \frac{1}{t'}\right) + o(\ell)\right).$$

By our choice of $\alpha'$ and $t'$, $I(\alpha'/t', 1/t') < \log d$. Then, for $\beta > 0$ small enough, there is $\ell$ large enough such that we have

$$\mathbf{E}Y \geq \exp\left(\ell \log d - \ell I\left(\frac{\alpha'}{t'}, \frac{1}{t'}\right) - \beta\ell\right) > 1. \tag{22}$$

Thus, the process $\{G_s, s \geq 0\}$ of good nodes is supercritical.

Let $A$ be the event that all the nodes with $N_u \geq \log^2 n$ are of type $(1, \ldots, 1)$. Let $B$ be the event that all the nodes with $nL_u \geq 2\log^2 n$ have $N_u \geq \log^2 n$. We have

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \leq \mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, B\right\} + \mathbf{P}\left\{\bar{A}\right\} + \mathbf{P}\left\{\bar{B}\right\},$$

where $\bar{A}$ and $\bar{B}$ are the complements of $A$ and $B$, respectively. If both $A$ and $B$ occur, then, the nodes with $nL_u \geq 2\log^2 n$ all have $d$ children. Writing $r = r(n) = \log^2 n$, for $n$ large enough, we have $m(n) \leq r(n)$, and

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, B\right\} \leq \mathbf{P}\left\{P_r(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, B\right\}.$$

By definition, on the event $A$, all the variables influencing $P_r(k,h)$ are distributed as $(Z^c, E)$. Also, for $k = \lfloor t \log n \rfloor$, for any good node $u$ at level $\ell\lfloor k/\ell \rfloor$ (in the $\lfloor k/\ell \rfloor$-th generation of the process), we have

$$nL_u \geq n \cdot \left(e^{-\ell/t'}\right)^{\lfloor k/\ell \rfloor} \geq n \cdot e^{-t \log n/t'} = n^{1-t/t'} \geq 2\log^2 n,$$

for $n$ large enough since $t < t'$. Hence, if $k = 0 \bmod \ell$, and $A \cap B$ occurs, $G_{\lfloor k/\ell \rfloor}$ is a lower bound on $P_r(k, \ell)$. If $k \neq 0 \bmod \ell$, the subtree of every good node lying at level $\lfloor k/\ell \rfloor$ contains at level $k$ a node with $N_u \geq d^{-\ell} \cdot n^{1-t/t'} \geq r$, for $n$ large enough. Moreover, the weights are nonnegative and all the nodes at level $k$ lying in the subtree of a good node at level $\ell \lfloor k/\ell \rfloor$ are such that

$$D_u \geq \lfloor k\ell \rfloor \cdot \frac{\alpha'}{t'} \geq (t \log n - 1) \cdot \frac{\alpha'}{t'} \geq \alpha \log n,$$

for $n$ large enough, by our choice of $\alpha'$ and $t'$. Thus, if $A \cap B$ occurs, for any $k \geq 0$, $G_{\lfloor k/\ell \rfloor}$ is a lower bound for $P_r(k,h)$. As a consequence,

$$\begin{aligned}
\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, B\right\} &\leq \mathbf{P}\left\{P_r(k,h) \leq n^{\phi(\alpha,t)-\epsilon}, A, B\right\} \\
&\leq \mathbf{P}\left\{G_{\lfloor k/\ell \rfloor} \leq n^{\phi(\alpha,t)-\epsilon}, A, B\right\} \\
&\leq \mathbf{P}\left\{G_{\lfloor k/\ell \rfloor} \leq n^{\phi(\alpha,t)-\epsilon}\right\}.
\end{aligned}$$

Now, by Lemma 1, for $n$ large enough, $\mathbf{P}\left\{\bar{A}\right\} \leq n^{-\omega}$, for some $\omega \to \infty$, as $n \to \infty$. Also, by the union bound and Chernoff's bound,

$$\mathbf{P}\left\{\bar{B}\right\} \leq d^k \cdot \mathbf{P}\left\{\mathrm{Bin}\left(n, \frac{2}{n}\log^2 n\right) \leq \log^2 n\right\} \leq d^k e^{-\frac{1}{8}\log^2 n} \leq e^{-\frac{1}{10}\log^2 n},$$

for $n$ large enough. It follows that

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \leq \mathbf{P}\left\{G_{\lfloor k/\ell \rfloor} \leq n^{\phi(\alpha,t)-\epsilon}\right\} + o(1), \tag{23}$$

as $n \to \infty$. Proving the claim reduces to showing that the first term in the right-hand side above is strictly less than one. For this purpose, we take advantage of asymptotic properties of supercritical Galton–Watson process.

By Doob's limit law (see, e.g., Athreya and Ney, 1972), there exists a random variable $W$ such that

$$\frac{G_s}{\mathbf{E}G_s} \xrightarrow[s\to\infty]{} W \quad \text{in distribution.}$$

The equation above gives us a handle on $G_{\lfloor k/\ell \rfloor}$ via the limit distribution $W$. Recall that

$$\mathbf{E}G_{\lfloor k/\ell \rfloor} \geq \exp\left(k \log d - kI\left(\frac{\alpha'}{t'}, \frac{1}{t'} + o(k)\right)\right).$$

Hence, by continuity of $I(\cdot,\cdot)$ at $(\alpha/t, 1/t)$, we can choose $\alpha'$ and $t'$ satisfying the previous constraints and close enough to $\alpha$ and $t$, respectively, that

$$\mathbf{E}G_{\lfloor k/\ell \rfloor} \geq n^{\phi(\alpha,t)-\epsilon/2+o(1)},$$

for $n$ large enough. It follows that

$$\mathbf{P}\left\{G_{\lfloor k/\ell \rfloor} \leq n^{\phi(\alpha,t)-\epsilon}\right\} \;=\; \mathbf{P}\left\{G_{\lfloor k/\ell \rfloor} \leq \mathbf{E}G_{\lfloor k/\ell \rfloor} \cdot n^{o(1)-\epsilon/2}\right\},$$

As a consequence, by (23),

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \;\leq\; \mathbf{P}\left\{\frac{G_{\lfloor k/\ell \rfloor}}{\mathbf{E}G_{\lfloor k/\ell \rfloor}} = o(1)\right\} + o(1) \xrightarrow[k\to\infty]{} \mathbf{P}\left\{W = 0\right\}.$$

The random variable $W$ is characterized by the Kesten–Stigum theorem (see, e.g., Athreya and Ney, 1972). Since the progeny $Y$ is bounded by $d^\ell$, we have $\mathbf{E}\left[Y \log(1+Y)\right] < \infty$, and hence, $\mathbf{P}\left\{W = 0\right\} = q$, the extinction probability of the Galton–Watson process. Recalling that the process is supercritical by (22), we have

$$\mathbf{P}\left\{P_m(k,h) \leq n^{\phi(\alpha,t)-\epsilon}\right\} \leq q + o(1) < 1,$$

for $n$ large enough. This completes the proof. □

We now proceed with the boosting argument. Let $\epsilon > 0$ be arbitrary. Observe first that, by Lemma 6, for some $q < 1$, and $n_o$ large enough,

$$\sup_{n \geq n_o} \mathbf{P}\left\{P_m(k,h) \leq n^{\phi(t,\alpha)-\epsilon/2}\right\} \leq q, \tag{24}$$

where $P_m(k,h)$ denotes the profile in a trie on $n$ sequences. Consider $\mathscr{L}_s$, the set of nodes $s$ levels away from the root, for $s = s(n) = \lfloor \log\log n \rfloor$. Each one of $N_u$, $u \in \mathscr{L}_s$ is distributed as a binomial $\mathrm{Bin}(n, \xi_u)$ with $\xi_u \geq p_d^s$. Let $J_s$ be the good event that for each $u \in \mathscr{L}_s$, $N_u \geq n_s$ with $n_s = \lceil n p_d^s/2 \rceil$. Using the union bound, and then Chernoff's bound for binomial random variables (Chernoff, 1952; Janson et al., 2000), we see that $J_s$ happens with high probability:

$$\mathbf{P}\left\{\bar{J}_s\right\} \;=\; \mathbf{P}\left\{\min_{u \in \mathscr{L}_s} N_u < n_s\right\} \;\leq\; d^s \cdot \mathbf{P}\left\{\mathrm{Bin}(n, p_d^s) < n_s\right\} \;\leq\; d^s \cdot e^{-n_s/8}. \tag{25}$$

Let $T_\infty(u)$ denote the subtree of $T_\infty$ rooted at a node $u$. Given the values of the first $s$ symbols of each string, the subtrees $T_\infty(u)$, $u \in \mathscr{L}_s$ are independent. Moreover, conditioning on $J_s$, each one of these trees behaves as a weighted trie with at least $n_s$ sequences. Let

$P^u_{n,m}(k,h)$ be the number of nodes $v \in \mathscr{L}_k \cap T_\infty(u)$ such that $N_v \geq m$ and $D_v \geq h$. Since the weights are bounded below by $-a$, say, we have

$$\begin{aligned}\mathbf{P}\left\{ P^u_{n,m}(k,h) \leq n^{\phi(t,\alpha)-\epsilon} \;\middle|\; J_s \right\} &\leq \sup_{u \in \mathscr{L}_s} \mathbf{P}\left\{ P_{N_u,m}(k-s, h-D_u) \leq n^{\phi(\alpha,t)-\epsilon} \right\} \\ &\leq \sup_{N \geq n_s} \mathbf{P}\left\{ P_{N,m}(k-s, h+as) \leq n_s^{\phi(\alpha,t)-\epsilon/2} \right\},\end{aligned}$$

for $n$ large enough. Hence, for $n$ large enough, since $k-s \sim t \log n$ and $h+as \sim \alpha \log n$,

$$\mathbf{P}\left\{ P^u_{n,m}(k,h) \leq n^{\phi(\alpha,t)-\epsilon} \;\middle|\; J_s \right\} \;\leq\; \sup_{N \geq n_s} \mathbf{P}\left\{ P_{N,m}(k-s, h+as) \leq n_s^{\phi(t,\alpha)-\epsilon/2} \right\} \;\leq\; q,$$

by (24). However, if $P^u_{n,m}(k,h)$ is large for any of the nodes $u \in \mathscr{L}_s$, then $P_m(k,h)$ is large as well:

$$\mathbf{P}\left\{ P_{n,m}(k,h) \leq n^{\phi(t,\alpha)-\epsilon} \;\middle|\; J_s \right\} \;\leq\; \mathbf{P}\left\{ \forall u \in \mathscr{L}_s, P^u_{n,m}(k,h) \leq n^{\phi(t,\alpha)-\epsilon} \;\middle|\; J_s \right\} \;\leq\; q^{d^s},$$

by independence. This finishes the proof of Theorem 4 since,

$$\mathbf{P}\left\{ P_{n,m}(k,h) \leq n^{\phi(\alpha,t)-\epsilon} \right\} \;\leq\; \mathbf{P}\left\{ P_{n,m}(k,h) \leq n^{\phi(\alpha,t)-\epsilon} \;\middle|\; J_s \right\} + \mathbf{P}\left\{ \bar{J}_s \right\} \;=\; o(1),$$

by (25) and our choice for $s$.

# 5 How long is a spaghetti?

## 5.1 Behavior and geometry

As we have seen in Section 2.2, the behavior of the spaghettis is captured by that of forests of independent tries. In preparation for the proof of Theorem 1, we aim at characterizing the maximum weighted height of a trie in such a forest.

Let $T^1, T^2, \ldots, T^n$ be $n$ independent $b$-tries. We assume that $T^i$ is a weighted $b$-trie on $m_i = m_i(n)$ sequences generated by the memoryless source with distribution $\{p_1, \ldots, p_d\}$. We also assume that for all $i$, $m/d \leq m_i \leq m$. The roots of $T^i$, $1 \leq i \leq n$, all lie at level zero. Then, we let $P^s(k,h)$ count the number of nodes $u$ at level $k$ with $D_u \geq h$ lying in any component $T^i$ of the forest. Since $T^i$ is a $b$-trie, we only count the nodes for which $N_u \geq b+1$. For now, we are only interested in $\mathbf{E}P^s(k,h)$, when $k \sim \rho \log n$ and $h \sim \gamma \log n$.

The point of view we adopt here is radically different from the one we used for the core: instead of counting the nodes using a *uniformly random node* among the $d^k$ ones in the $k$-th level, we consider here a *uniformly random sequence* and the corresponding node $v_k$ at level $k$. In other words, we write:

$$\mathbf{E}P^s(k,h) = n\mathbf{P}\left\{D_{v_k} \geq h, v_k \in T_{n,b}\right\},$$

whereas the core was studied using the formula $\mathbf{E}P(k,h) = d^k \cdot \mathbf{P}\left\{D_{u_k} \geq h, u_k \in T_{n,b}\right\},$ where $u_k$ is uniformly random among the $d^k$ nodes in the $k$-th level of $T_\infty$. This approach is very similar to the classical one and relies on the analysis of $(b+1)$-tuples of strings (see, e.g., Szpankowski, 2001).

Let us focus on one single $(b+1)$-tuple. Recall that

$$Q(b+1) = \sum_{i=1}^{d} p_i^{b+1}$$

is the probability that $(b+1)$ characters generated independently by the source $\{p_1, \ldots, p_d\}$ are identical. Assume that the $b+1$ sequences are identical up to the $k$-th position. So the paths corresponding to the $b+1$ strings agree at least until the $k$-th level in $T_\infty$. The values of the $(k+1)$-st characters influence the next step in the paths, and the weights on the corresponding edges. There are two possible situations: if the $b+1$ characters in position $k+1$ are not identical, the paths split and the $(b+1)$-tuple cannot influence the profile past level $k$. Otherwise, they are identical and the $b+1$ sequences have followed the same edge. We account for the tuple being split using a weight of $-\infty$, hence the need for extended random variables in Section 3. This case happens with probability $1 - Q(b+1)$. On the other hand if the paths did not split, they have followed the same edge $i$ with probability $p_i$, the weight is then that of the $i$-th edge. More precisely, let $\sigma(i)$ be the permutation of $(1, 0, \ldots, 0)$ with the 1 in the $i$-th position. Then, we define

$$Z^s = \begin{cases} -\infty & \text{w.p.} \quad 1 - Q(b+1) \\ Z_i^{\sigma(i)} & \text{w.p.} \quad p_i \cdot Q(b+1) \quad \forall i \in \{1, \ldots, d\}. \end{cases} \tag{26}$$

Let $\Lambda_s^\star$ denote the Fenchel-Legendre transform of the cumulant generating function $\Lambda_s$ associated with $Z^s$ (see definitions in Section 3). Let $I_s(\cdot)$ be the (one-dimensional) rate function associated with the variable $Z^s$ appearing in Cramér's theorem (Theorem 2), that is, $I_s(\gamma) = \inf\{\Lambda_s^\star(\gamma') : \gamma' > \gamma\}$ and

$$\mathbf{P}\left\{\sum_{i=1}^{k} Z_i^s \geq \gamma k\right\} = e^{-kI_s(\gamma)+o(k)},$$

as $k \to \infty$, where $Z_i^s$, $i \geq 1$, are i.i.d. copies of $Z^s$. For $\gamma \geq 0$, define $\psi(\gamma, 0) = 1$ and

$$\forall \rho > 0 \qquad \psi(\gamma, \rho) = 1 - \rho I_s\left(\frac{\gamma}{\rho}\right). \tag{27}$$

**Theorem 5.** *Let $T^i$, $1 \leq i \leq n$, be a forest of $n$ independent tries. Let $T^i$ store $m_i = m_i(n)$ sequences. Assume that $m/d \leq m_i \leq m$ for all $1 \leq i \leq n$, with $m = m(n) \to \infty$ and $m = o(\log n)$. Let $\rho, \gamma \geq 0$ such that $\psi(\gamma, \rho) > -\infty$, then,*

$$\mathbf{E}P^s(\lfloor \rho \log n \rfloor, \gamma \log n) = n^{\psi(\gamma,\rho)+o(1)}.$$

*Moreover, for any natural number $k$, any $\delta > 0$, and $n$ large enough, we have the explicit upper bound*

$$\mathbf{E}P^s(k, \gamma \log n) \leq m^{b+1} \cdot n \cdot \exp\left(-kI_s\left(\frac{(\gamma - \delta)\log n}{k}\right)\right). \tag{28}$$

A typical logarithmic profile of a forest of tries is shown in Figure 4. Observe in particular that the logarithmic profile decreases linearly along any fixed direction $\gamma/\rho$. In other words, the point $(0, 0, 1)$ casts a cone of projections on the horizontal plane. There is a *preferred* direction, corresponding to $(\gamma_b, \rho_b, 0)$, such that

$$\gamma_b = \sup_{\gamma,\rho>0}\{\gamma : \psi(\gamma, \rho) \geq 0\}.$$

This point is especially important since it characterizes the maximum weighted height of $T^1, \ldots, T^n$. Let $H^1, \ldots, H^n$ be the weighted heights of $T^1, \ldots, T^n$, respectively, and define

$$S_{n,b} = \max_{1 \leq i \leq n} H^i. \tag{29}$$

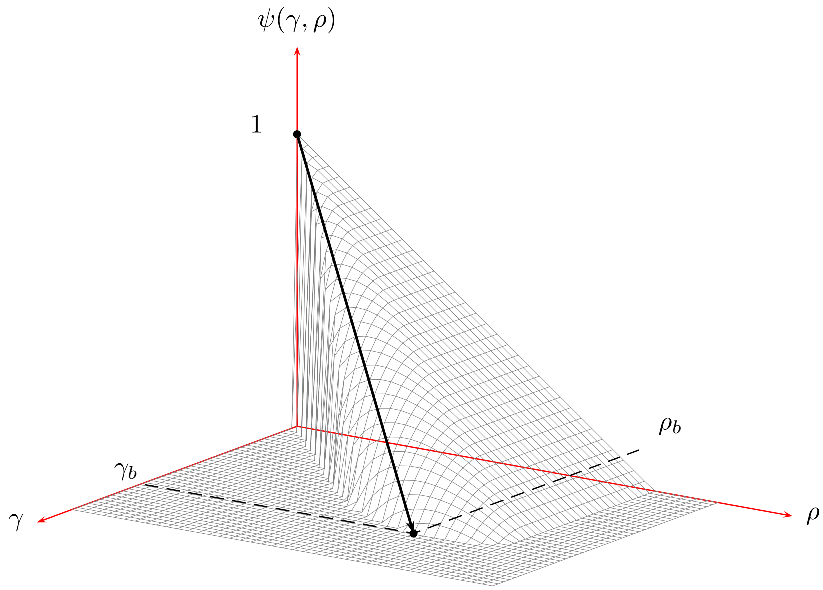


**Figure 4.** The profile generated by $n$ independent tries on roughly $m(n) = o(\log n)$ sequences each. The constant $\gamma_b$ characterizing the highest component (trie) of the forest is also shown.

**Theorem 6.** *Assume that $p_1 < 1$. Assume that $m(n) \to \infty$ and $m(n) = o(\log n)$. Let*

$$\gamma_b \stackrel{\text{def}}{=} \sup_{\gamma,\rho>0} \{\gamma : \psi(\gamma,\rho) \geq 0\}. \tag{30}$$

*Then, $S_{n,b} \sim \gamma_b \log n$ in probability, as $n \to \infty$. Furthermore, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for $n$ large enough,*

$$\mathbf{P}\left\{S_{n,b} \geq (\gamma_b + \epsilon)\log n\right\} \leq n^{-\delta}. \tag{31}$$

**Lemma 7.** *Let $Z^s$ be defined by (26). Let $\Lambda_s^\star$ and $I_s$ be the Fenchel-Legendre transform of the cumulant generating function $\Lambda_s$ and the rate function associated with $Z^s$, respectively. Let $\gamma_b = \sup_{\gamma,\rho}\{\gamma : \psi(\gamma,\rho) \geq 0\}$. We have*

$$\begin{aligned}
\gamma_b &= \sup\left\{\gamma : \exists\rho \quad \Lambda_s^\star(\rho) \leq \frac{\rho}{\gamma}\right\} \\
&= \sup_{\gamma,\rho>0} \left\{\gamma : \rho\Lambda_s^\star\left(\gamma/\rho\right) \leq 1\right\} \\
&= \inf\left\{\gamma : \forall\rho \quad \Lambda_s^\star(\rho) > \frac{\rho}{\gamma}\right\}.
\end{aligned}$$

*Proof.* We prove the first equality. Recall that $\psi(\gamma,\rho) = 1 - \rho I_s(\gamma/\rho)$. Assume that

$$\sup_{\gamma,\rho>0}\{\gamma : \psi(\gamma,\rho) \geq 0\} = \sup_{\gamma,\rho>0}\{\gamma : I_s(\rho) \leq \rho/\gamma\} = \gamma_o.$$

Let $\epsilon > 0$. Then, there exists $\gamma_1 > \gamma_o - \epsilon$ and $\rho_1$ such that $I_s(\rho_1) \leq \rho_1/\gamma_1$. By definition of $I_s(\cdot)$, there exists $\rho_1' > \rho_1 > 0$ such that $\Lambda_s^\star(\rho_1') \leq I(\rho_1) + \epsilon$. Therefore,

$$\Lambda_s^\star(\rho_1') \leq \frac{\rho_1'}{\gamma} + \epsilon \leq \frac{\rho_1'}{\gamma - \gamma_1\epsilon/\rho_1'},$$

since $\rho_1' > 0$. It follows that

$$\sup_{\gamma,\rho>0}\left\{\gamma : \Lambda_s^\star(\rho) \leq \frac{\rho}{\gamma}\right\} \geq \gamma_1 - \frac{\gamma_1\epsilon}{\rho_1'} \geq \gamma_o - \epsilon - \frac{\gamma_1\epsilon}{\rho_1'} \xrightarrow[\epsilon\downarrow 0]{} \gamma_o = \sup_{\gamma,\rho}\left\{\gamma : I_s(\rho) \leq \frac{\rho}{\gamma}\right\}.$$

Similar arguments prove the second inequality: assume that $\sup_{\gamma,\rho}\{\gamma : \Lambda_s^\star(\rho) \le \rho/\gamma\} = \gamma_2$. Then, there exist $\gamma_3 > \gamma_2 - \epsilon$ and $\rho_3$ such that $I_s(\rho_3) \le \rho_3/\gamma_3$. Moreover,

$$I_s(\rho_3 - \epsilon) \le \Lambda_s^\star(\rho_3 - \epsilon) \le \frac{\rho_3 - \epsilon}{\gamma_3} \cdot \frac{1}{1 - \epsilon/\rho_3}.$$

Therefore,

$$\sup_{\gamma,\rho} \left\{ \gamma : I_s(\rho) \le \frac{\rho}{\gamma} \right\} \ge \gamma_3 - \frac{\gamma_3 \epsilon}{\rho_3} \xrightarrow[\epsilon \downarrow 0]{} \gamma_2 = \sup_{\gamma,\rho} \left\{ \gamma : \Lambda_s^\star(\rho) \le \frac{\rho}{\gamma} \right\}.$$

The condition on the right-hand side of (30) implies that $\gamma_b$ is the largest $\gamma$ such that there exists $\rho$ satisfying $\Lambda_s^\star(\rho) \le \rho/\gamma$. In other words, if we plot $\rho \mapsto \Lambda_s^\star(\rho)$, then $1/\gamma_b$ is the slope of the most gentle line going through the origin and hitting the graph of $\Lambda_s^\star(\cdot)$, as shown in Figure 5. We have just proved the second equality. The third follows from the minimax principle. $\square$

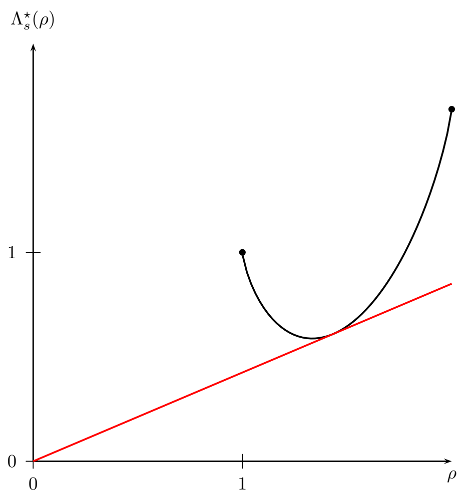


**Figure 5.** The constant $1/\gamma_b$ is the slope of the line going through the origin that is tangent to the curve $\{(\rho, \Lambda_s^\star(\rho))\}$.

Using this alternate definition of $\gamma_b$, we can characterize the value of $\gamma_b$.

**Lemma 8.** *Let $Z^s$ and $\psi(\cdot,\cdot)$ be defined by (26) and (27), respectively. Let $\gamma_b = \sup_{\gamma,\rho}\{\gamma : \psi(\gamma,\rho) \ge 0\}$. Assume that $Z^s$ is not almost surely null. If $Q(b) < 1$, then $\gamma_b \in (0,\infty)$. Otherwise, $\gamma_b = \infty$.*

*Proof.* We have $\inf_\rho \Lambda_s^\star(\rho) = -\log \mathbf{P}\{Z^s > -\infty\}$ (Dembo and Zeitouni, 1998; Broutin, 2007). Recall Lemma 7. If $Q(b) < 1$, then $\inf_\rho \Lambda_s^\star(\rho) = -\log Q(b) > 0$. Moreover the infimum is reached at $\rho = \mathbf{E}[\, Z^s \mid Z^s > -\infty\,] > 0$. The result follows (see Figure 5). On the other hand, if $Q(b) = 1$, then $\inf_\rho \Lambda_s^\star(\rho) = 0$ and $1/\gamma_b = 0$. $\square$

## 5.2 The profile of a forest of tries: Proof of Theorem 5

As we have sketched before, the proof of Theorem 5 relies on the analysis of $(b+1)$-tuples of sequences. Let $\gamma, \rho > 0$ such that $\psi(\gamma,\rho) > -\infty$. Let $k$ and $h$ be such that $h \sim \gamma \log n$, $k \sim \rho \log n$, as $n \to \infty$.

Consider a single sequence $X_1 X_2 \dots$. Let $v_k$ be the node of its associated path in $T_\infty$ lying at level $k$. Since the number of sequences lies between $nm/d$ and $n/m$, we have

$$\frac{nm}{d} \cdot \mathbf{P}\left\{ D_{v_k} \ge h, v_k \in \bigcup_{j=1}^n T^i \right\} \;\le\; \mathbf{E} P^s(k,h) \;\le\; nm\mathbf{P}\left\{ D_{v_k} \ge h, v_k \in \bigcup_{j=1}^n T^i \right\}. \quad (32)$$

Assume that $X_1X_2\dots$ is stored in $T^i$. Then, $v_k$ is a node of the forest if there are at least $b$ *other* sequences also stored in $T^i$ whose path intersect $v_k$. For a given set of $b$ *extra* sequences stored in $T^i$, let $F_j$ be the event that all the characters in $j$-th position are identical to $X_j$. There are $m_i^b$ choices for these $b$ sequences, and $m/d \le m_i \le m$. Therefore,

$$\mathbf{P}\left\{D_{v_k} \ge h, \bigcap_{j=1}^{k} F_j\right\} \;\le\; \mathbf{P}\left\{D_{v_k} \ge h, v_k \in \bigcup_{i=1}^{n} T^i\right\} \;\le\; m^b \mathbf{P}\left\{D_{v_k} \ge h, \bigcap_{j=1}^{k} F_j\right\}, \quad (33)$$

where the lower bound is obtained by considering a single set of $b$ sequences, and the upper bound follows by the union bound. Putting (32) and (33) together, we obtain, for $n$ large enough,

$$n \cdot \mathbf{P}\left\{D_{v_k} \ge h, \bigcap_{j=1}^{k} F_j\right\} \;\le\; \mathbf{E}P^s(k,h) \;\le\; nm^{b+1} \cdot \mathbf{P}\left\{D_{v_k} \ge h, \bigcap_{j=1}^{k} F_j\right\}. \quad (34)$$

However, $D_{v_k} = \sum_{e \in \pi(v_i)} Z_e$, and at most $m = o(\log n)$ nodes of any downward path in the forest have at least two children (in the forest). So for all but $o(\log n)$ levels, the node $v_j$ has type $\sigma(X_j)$, and the edge between $v_j$ and $v_{j+1}$ has weight

$$Z_{X_j}^{\sigma(X_j)}.$$

Since the weights are bounded, and for $h \sim \gamma \log n$, it follows that

$$\begin{aligned}
\mathbf{P}\left\{D_{v_k} \ge h, \bigcap_{1 \le j \le k} F_j\right\} &= \mathbf{P}\left\{\sum_{j=1}^{k} Z_{X_j}^{\sigma(X_j)} + o(\log n) \ge \gamma \log n, \bigcap_{1 \le j \le k} F_j\right\} \\
&= \mathbf{P}\left\{\sum_{j=1}^{k} \left(Z_{X_j}^{\sigma(X_j)} - \infty \mathbf{1}[\bar{F}_j]\right) + o(\log n) \ge \gamma \log n\right\}.
\end{aligned}$$

The summands in the probability above are precisely distributed as

$$Z^s = \begin{cases} -\infty & \text{w.p.} \quad 1 - Q(b+1) \\ Z_i^{\sigma(i)} & \text{w.p.} \quad p_i \cdot Q(b+1) \quad \forall i \in \{1, \dots, d\}. \end{cases}$$

defined in (26). It follows that

$$\mathbf{P}\left\{D_{v_k} \ge h, \bigcap_{1 \le j \le k} F_j\right\} = \mathbf{P}\left\{\sum_{j=1}^{k} Z_j^s + o(\log n) \ge \gamma \log n\right\}, \quad (35)$$

where $Z_j^s$, $1 \le j \le k$, are i.i.d. copies of $Z^s$. Let $\delta > 0$ be arbitrary. There is $n$ large enough such that $\gamma \log n + o(\log n) \ge (\gamma - \delta) \log n$. By Chernoff's bound,

$$\mathbf{E}P^s(k, \gamma \log n) \le nm^{b+1} \cdot \exp\left(-kI_s\left(\frac{(\gamma-\delta)\log n}{k}\right)\right),$$

which proves the explicit upper bound (28). Now, if $k \sim \rho \log n$ and $h \sim \gamma \log n$, there exists $n$ large enough that,

$$\frac{\gamma}{\rho} - \delta \;\le\; \frac{\gamma \log n + o(\log n)}{k} \;\le\; \frac{\gamma}{\rho} + \delta.$$

Using Cramér's Theorem in (35), and (34), we obtain

$$n \cdot e^{-kI_s(\gamma/\rho+\delta)+o(k)} \;\le\; \mathbf{E}P^s(k,h) \;\le\; nm^{b+1} \cdot e^{-kI_s(\gamma/\rho-\delta)},$$

as $n \to \infty$, where $I_s(x) = \inf\{\Lambda_s^\star(x') : x' > x\}$, and $\Lambda_s^\star$ is the rate function of $Z^s$. By definition of $\psi$ in (27), by continuity of $\Lambda_s^\star$ and $I_s$ at $\gamma/\rho \in \mathcal{D}^o_{\Lambda_s^\star}$ (see, e.g., Dembo and Zeitouni, 1998), and since $k \sim \rho \log n$, we have

$$\mathbf{E}P^s(k,h) \;=\; n^{1-\rho I_s(\gamma/\rho)+o(1)} \;=\; n^{\psi(\gamma,\rho)+o(1)},$$

as $n \to \infty$, since $m^{b+1} = n^{o(1)}$. This completes the proof of Theorem 5.

## 5.3 The longest spaghetti: Proof of Theorem 6

The upper bound. We use the first moment method (see, e.g., Alon et al., 2000). Let $\epsilon > 0$ be arbitrary. By the definition of $S_{n,b}$ in (29) and the union bound,

$$\begin{aligned}\mathbf{P}\{S_{n,b} \ge (\gamma_b+\epsilon)\log n\} \;&\le\; \sum_{k\ge 0} \mathbf{E}P^s(k,\gamma' \log n) \\ &\le\; nm^b \sum_{k\ge 0} \exp\left(-kI_s\left(\frac{(\gamma_b+\epsilon/2)}{\log n}\right)\right), \qquad (36)\end{aligned}$$

for $n$ large enough, by the explicit upper bound (28) of Theorem 5. We now split the sum in the right-hand side of (36) into two pieces, and then bound each one of them separately.

When $k$ is large, it is unlikely that we find a set of $b+1$ sequences in the same trie that all agree on the first $k$ characters. Recall that $\mathbf{P}\{Z^s > -\infty\} = Q(b+1)$, and hence $\inf_\rho I_s(\rho) = \inf_\rho \Lambda_s^\star(\rho) = -\log Q(b+1)$. Let $\delta > 0$ and define

$$K = K(n) = \frac{1+\delta}{-\log Q(b+1)} \cdot \log n.$$

Then, for $n$ large enough, we have

$$nm^b \sum_{k\ge K} \exp\left(-kI_s\left(\frac{(\gamma_b+\epsilon/2)\log n}{k}\right)\right) = O\left(nm^b e^{K\log Q(b+1)}\right) = O\left(n^{-\delta/2}\right). \qquad (37)$$

For the low values of $k$, we have to consider the weights. Observe first that, by definition of $\gamma_b$, there exists $\beta > 0$ such that

$$\inf_{k\ge K, n\ge 2} \left\{\frac{k}{\log n} \cdot I_s\left(\frac{\gamma+\epsilon/2}{k/\log n}\right)\right\} \;\ge\; \inf_{\rho>0} \left\{\rho \cdot I_s\left(\frac{\gamma+\epsilon/2}{\rho}\right)\right\} \;=\; 1+\beta.$$

Then, since $K = O(\log n)$,

$$nm^b \sum_{k\le K} \exp\left(-kI_s\left(\frac{(\gamma+\epsilon/2)\log n}{k}\right)\right) \;\le\; Km^b n^{-\beta} \;=\; O\left(n^{-\beta/2}\right), \qquad (38)$$

for $n$ large enough. Plugging both (37) and (38) in (36) proves that

$$\mathbf{P}\{S_{n,b} > (\gamma_b+\epsilon)\log n\} = O\left(n^{-\delta/2}\right) + O\left(n^{-\beta/2}\right),$$

which completes the proof of the upper bound (31).

The lower bound. Let $\epsilon > 0$. By assumption, $m(n) \to \infty$, and hence, there exists $n$ large enough that $m(n)/d \geq b+1$. We only consider one single tuple from the each of the $n$ tries $T^i$, $1 \leq i \leq n$. We then have $n$ independent realizations each one being at least

$$\max\left\{\sum_{j=1}^{k} Z_j^s : k \geq 0\right\} - o(\log n), \tag{39}$$

where $Z_j^s$, $j \leq 1$ are i.i.d. copies of $Z^s$ defined in (26). The largest of $n$ independent copies of (39) is a lower bound on $S_{n,b}$. Let $\xi_i$, $1 \leq i \leq n$ denote the sequence of indicators that the $i$-th realization is at least $(\gamma_b - \epsilon)\log n$. Let $M = \sum_{i=1}^n \xi_i$. We intend to prove that $M \geq 1$ with probability tending to one as $n \to \infty$. For this purpose, we use the second moment method. Since $\{\xi_i, 1 \leq i \leq n\}$ are independent, it suffices to prove that $\mathbf{E}M \to \infty$ (see, e.g., Alon et al., 2000, Corollary 4.3.4, p. 46). However, we have,

$$\begin{aligned}
\mathbf{E}M &= n \cdot \mathbf{P}\left\{\exists k : \sum_{j=1}^{k} Z_j^s - o(\log n) \geq (\gamma_b - \epsilon)\log n\right\} \\
&\geq n \cdot \mathbf{P}\left\{\sum_{j=1}^{k_o} Z_j^s \geq (\gamma_b - \epsilon/2)\log n\right\},
\end{aligned}$$

for any $k_o \geq 1$, and $n$ large enough. By the alternate definition of $\gamma_b$ provided by Lemma 7, there exists $\rho$ such that

$$\rho \cdot I_s\left(\frac{\gamma_b - \epsilon/2}{\rho}\right) < 1.$$

In particular, if we set $k_o = \lceil \rho \log n \rceil$, by Cramér's theorem,

$$\mathbf{E}M \geq n \cdot \mathbf{P}\left\{\sum_{j=1}^{k_o} Z_s^j \geq (\gamma_b - \epsilon/2)\log n\right\} = n \cdot \exp\left(-k_o I_s\left(\frac{\gamma_b - \epsilon/2}{\rho}\right) + o(k_o)\right) \xrightarrow[n\to\infty]{} \infty.$$

This completes the proof of Theorem 6.

# 6 The height of weighted tries

## 6.1 Projecting the profile

Recall the definitions of the core and spaghettis. Let $m = m(n) \to \infty$ with $m = o(\log n)$. The core $\mathcal{C}$ of a $b$-trie $T_{n,b}$ is the set of nodes $u \in T_{n,b}$ such that $N_u \geq m$. When removing the core $\mathcal{C}$ from $T_{n,b}$, we obtain a forest of trees, the *spaghettis* (see Figure 1). Each one of these trees is rooted at a node $u \in \partial\mathcal{C}$, the external node-boundary of $\mathcal{C}$ in $T_{n,b}$. In other words, the nodes $u \in \partial\mathcal{C}$ are the children of some node $v$ in the core, but are not themselves in the core. Recall that

$$\gamma_b = \sup_{\gamma,\rho>0}\{\gamma : \psi(\gamma,\rho) \geq 0\} \qquad \text{and} \qquad c_b = \sup_{\alpha,t>0}\{\alpha + \gamma_b\phi(\alpha,t)\},$$

where $\psi(\cdot,\cdot)$ and $\phi(\cdot,\cdot)$ denote the logarithmic profiles of a forest of tries and a single trie, respectively (see Sections 4 and 5).

The definition of $c_b$ can be interpreted as follows. Consider a point $(\alpha, t, \phi(\alpha,t))$. This point is mapped on the horizontal plane going through the origin via a *projection*. The direction of the projection is given by the vector $(1, 0, -1/\gamma_b)$. The direction along the $t$-axis is actually irrelevant, and any direction $(1, x, -1/\gamma_b)$ gives the same $\alpha$-coordinate for the image of the point $(\alpha, t, \phi(\alpha, t))$. The constant $c_b$ is the largest $\alpha$-coordinate of these projections.

The projection is not a mere interpretation of the *formula* for $c_b$. Indeed, Theorem 5 shows that a set of $P_m(k, h)$ tries on about $m(n)$ sequences each has a logarithmic profile that decays linearly in every direction. Observe also that the actual profile induces a *preferred* direction of projection $(1, -1/\rho_b, -1/\gamma_b)$, as shown in Figure 4. The projection of points $(\alpha, t, \phi(\alpha, t))$ using this preferred direction is depicted in Figure 6.

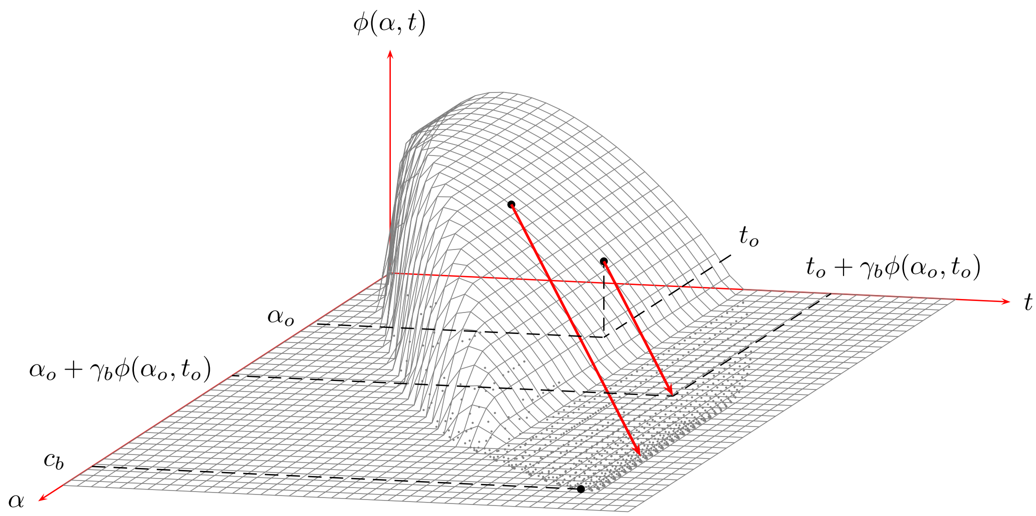


**Figure 6.** A geometric interpretation for the height: each point $(\alpha, t, \phi(\alpha, t))$ of the logarithmic profile of the core throws a line whose direction is given by $(1, -1/\rho_b, -1/\gamma_b)$. The line intersects the plane $\{\phi = 0\}$ at $(\alpha + \gamma_b\phi(\alpha, t)), t + \rho_b\phi(\alpha, t), 0)$. The constant $c_b$ is the largest coordinate of one of these point along the $\alpha$-axis.

## 6.2 Proof of Theorem 1

Put together, Lemmas 10 and 9 prove Theorem 1. We start with the lower bound which is easier.

**Lemma 9.** *Let $T_{n,b}$ be a b-trie as defined in Section 2. Let $H_{n,b}$ be its weighted height. Then, for any $\epsilon > 0$, $\mathbf{P}\{H_{n,b} \le (c_b - \epsilon)\log n\} \to 0$, as $n \to \infty$.*

*Proof.* Let $\epsilon > 0$. Recall that, by definition, $c_b = \sup\{\alpha + \gamma_b \cdot \phi(\alpha, t) : t, \alpha > 0\}$. Therefore, there exists $(\alpha_o, t_o)$ such that

$$\alpha_o + \gamma_b \cdot \phi(\alpha_o, t_o) \ge c_b - \epsilon/2. \tag{40}$$

Let $\alpha_o$ and $t_o$ now be fixed. Let $k = \lceil t_o \log n \rceil$ and $h = \alpha_o \log n$. Let $\mathcal{F}_k$ be the $\sigma$-algebra generated by the first $k$ characters of the $n$ strings. Consider the $N' = P_m(k, h)$ nodes $u$ at level $k$ for which $N_u \ge m$, $D_u \ge h$. Conditioning on $\mathcal{F}_k$, the $P_m(k, h)$ trees rooted at these nodes are independent. Following the setting of Section 5, $S_{N',b}$ denotes the weighted height of the tallest of these trees. We want to show that $h + S_{N',b}$ is a good enough lower bound on $H_{n,b}$. For this purpose, it suffices to lower bound $S_{N',b}$.

As we have sketched before, we are in the situation of a forest of independent random tries, and we intend to apply Theorem 6. Let $\delta > 0$ and $n' = n^{\phi(\alpha_o,t_o)-\delta}$. We have

$$\begin{aligned}\mathbf{P}\left\{ \frac{S_{N',b}}{\log n'} \le \gamma_b - \delta \,\middle|\, \mathcal{F}_k \right\} \quad \le \quad & \mathbf{P}\left\{ \frac{S_{N',b}}{\log n'} \le \gamma_b - \delta \,\middle|\, \mathcal{F}_k, P_m(k,h) \ge n^{\phi(\alpha_o,t_o)-\delta} \right\} \\ & + \mathbf{1}[P_m(k,h) \le n^{\phi(\alpha_o,t_o)-\delta}].\end{aligned}$$

Taking expected values, we obtain

$$\mathbf{P}\left\{ \frac{S_{N',b}}{\log n'} \;\le\; \gamma_b - \delta \right\} \le \mathbf{P}\left\{ \frac{S_{n',b}}{\log n'} \le \gamma_b - \delta \right\} + \mathbf{P}\left\{ P_m(k,h) \le n^{\phi(\alpha_o,t_o)-\delta} \right\}. \tag{41}$$

It only remains to bound both terms appearing in the right-hand side of (41). By Theorems 4 and 6, respectively, we have, for any $\delta > 0$,

$$\mathbf{P}\left\{ P_m(k,h) \le n^{\phi(t_o,\alpha_o)-\delta} \right\} \xrightarrow[n\to\infty]{} 0 \qquad \text{and} \qquad \mathbf{P}\left\{ \frac{S_{n',b}}{\log n'} \le \gamma_b - \delta \right\} \xrightarrow[n\to\infty]{} 0.$$

Therefore, with probability $1 - o(1)$,

$$\begin{aligned} S_{n',b} \;&\ge\; (\gamma_b - \delta) \cdot \log n' \\ &=\; (\gamma_b - \delta) \cdot (\phi(\alpha_o, t_o) - \delta) \cdot \log n \\ &>\; (\gamma_b \phi(\alpha_o, t_o) - \epsilon/2) \cdot \log n, \end{aligned}$$

for $\delta$ small enough. The weighted height $H_{n,b}$ of $T_{n,b}$ is at least $h + S_{N',b}$. It follows that, with probability $1 - o(1)$,

$$\frac{H_{n,b}}{\log n} \;\ge\; \alpha_o + \gamma_b \phi(\alpha_o, t_o) - \epsilon/2 \;\ge\; c_b - \epsilon,$$

by our choice of $\delta$ and (40). This completes the proof of the lower bound. □

**Lemma 10.** *Let $T_{n,b}$ be a b-trie as defined in Section 2. Let $H_{n,b}$ be its weighted height. Then, for any $\epsilon > 0$, $\mathbf{P}\{H_{n,b} \ge (c_b + \epsilon)\log n\} \to 0$, as $n \to \infty$.*

*Proof.* Let $\epsilon > 0$. Let $W_u$ denote the weighted height of the subtree of $T_{n,b}$ rooted at $u$. Recall that $\mathcal{C}$ denotes the set of nodes in the core. We have

$$\mathbf{P}\{H_{n,b} \ge (c_b + \epsilon)\log n\} \;\;\le\;\; \mathbf{P}\{\exists u \in \mathcal{C} : D_u + W_u \ge (c_b + \epsilon)\log n\}.$$

Let $\mathcal{C}_k = \mathcal{C} \cap \mathscr{L}_k$, where $\mathscr{L}_k$ is the set of nodes $k$ levels away from the root in $T_\infty$. Then,

$$\mathbf{P}\{H_{n,b} \ge (c_b + \epsilon)\log n\} \;\;\le\;\; \mathbf{P}\{\exists k, u \in \mathcal{C}_k : D_u + W_u \ge (c_b + \epsilon)\log n\}.$$

We can immediately restrict the range of $k$. Indeed, when $k$ is too large, it is unlikely that there is even one node $u$ in $\mathcal{C}_k$. By Lemma 3, $\{(\alpha, t) : \phi(\alpha, t) \ge 0\}$ is contained in a bounded set. Pick $t$ large enough that $\phi(0, t) \le -\epsilon < 0$. Let $K = K(n) = \lceil t \log n \rceil$. Then,

$$\begin{aligned} \mathbf{P}\{\exists k \ge K, u \in \mathcal{C}_k : D_u + W_u \ge (c_b + \epsilon)\log n\} \;&\le\; \mathbf{P}\{\exists u \in \mathcal{C}_K\} \\ &\le\; \mathbf{E} P_m(0, K) \\ &=\; n^{-\epsilon + o(1)}, \end{aligned}$$

by Theorem 3. Let $\mathcal{C}_k(h) = \{u \in \mathcal{C}_k : D_u \ge h\}$. By the union bound,

$$\begin{aligned} \mathbf{P}\{H_{n,b} \ge (c_b + \epsilon)\log n\} \;&\le\; \sum_{k \le K} \mathbf{P}\{\exists u \in \mathcal{C}_k : D_u + W_u \ge (c_b + \epsilon)\log n\} + o(1) \qquad (42) \\ &=\; \sum_{k \le K} \underbrace{\mathbf{P}\left\{ \exists h : \mathcal{C}_k(h) \ne \emptyset, h + \max_{u \in \mathcal{C}_k(h)} W_u \ge (c_b + \epsilon)\log n \right\}}_{R(k)} + o(1). \end{aligned}$$

The terms $R(k)$, $0 \le k \le K$, are influenced by two parameters: the number of trees in the forest, and the rate at which their weighted depths grow when their number is fixed. These parameters depend on the core and the spaghettis, respectively, as studied in Sections 4 and 5.

Bounding the growth of spaghettis. Let $k \le K$, and consider the corresponding term $R(k)$ in the sum above. Let $\mathcal{F}_k$ be the $\sigma$-algebra generated by the first $k$ symbols of the $n$ strings. Then,

$$R(k) = \mathbf{E}\left[\mathbf{P}\left\{\exists h : \mathcal{C}_k(h) \neq \emptyset, h + \max_{u \in \mathcal{C}_k(h)} W_u \ge (c_b + \epsilon)\log n \;\middle|\; \mathcal{F}_k\right\}\right].$$

However, given $\mathcal{F}_k$, the number of trees in the forest is fixed, and only the rate of growth of the spaghettis matters. More precisely, given $\mathcal{F}_k$, $\max\{W_u : u \in \mathcal{C}_k(h)\}$ is distributed like the longest of $P_m(k,h)$ independent weighted tries, each on at most $m(n)$ sequences. We bound the rate of growth of the spaghettis: by Theorem 6, for any $\beta > 0$ there exists $\delta > 0$ such that

$$\mathbf{P}\left\{\max_{u \in \mathcal{C}_k(h)} W_u \ge (\gamma_b + \beta)\log P_m(k,h) \;\middle|\; \mathcal{F}_k\right\} \le e^{-\delta \log P_m(k,h)}, \tag{43}$$

where $\gamma_b$ defined by (30). This bound is weak when $P_m(k,h)$ is small. In such a case, we shall rather use

$$\mathbf{P}\left\{\max_{u \in \mathcal{C}_k(h)} W_u \ge \frac{\epsilon}{2}\log n \;\middle|\; \mathcal{F}_k\right\} \le n^{-\delta\epsilon/(2\gamma_b + 2\beta)}. \tag{44}$$

Define the following good event:

$$A \stackrel{\text{def}}{=} \left\{\forall h : \max_{u \in \mathcal{C}_k(h)} W_u < \max\left\{(\gamma_b + \beta)\log P_m(k,h), \frac{\epsilon}{2}\log n\right\}\right\}.$$

Then, by (44) and (44), since the weights are bounded, we have, and for $\beta \le \gamma_b$, and all $k \le K$,

$$\mathbf{P}\left\{\bar{A}\right\} \;\le\; K\|Z\|_\infty \cdot n^{-\delta\epsilon/(4\gamma_b)} \;\le\; n^{-\delta\epsilon/(5\gamma_b)},$$

for $n$ large enough. Therefore, by definition of $R(k)$,

$$\begin{aligned} R(k) \;&\le\; \mathbf{P}\left\{\exists h : \mathcal{C}_k(h) \neq \emptyset, h + \max_{u \in \mathcal{C}_k(h)} W_u \ge (c_b + \epsilon)\log n \;\middle|\; A\right\} + \mathbf{P}\left\{\bar{A}\right\} \\ &\le\; \mathbf{P}\left\{\exists h : h + (\gamma_b + \beta)\log P_m(k,h) \ge \left(c_b + \frac{\epsilon}{2}\right)\log n\right\} + n^{-\delta\epsilon/(5\gamma_b)}. \end{aligned} \tag{45}$$

Bounding the number of spaghettis. We now bound the first term of (45), for which only the core matters. Let $\eta > 0$. The full range for $k$ and $h$ is obtained by setting

$$\begin{aligned} \lfloor t\log n \rfloor \;&\le\; k \;\le\; \lfloor (t + \eta/\|Z\|_\infty)\log n \rfloor \\ \alpha\log n \;&\le\; h \;\le\; (\alpha + \eta)\log n, \end{aligned}$$

and letting $t$ and $\alpha$ vary. For such $k$ and $h$,

$$P(k,h) \;\le\; P(\lfloor t\log n, (\alpha - \eta)\log n \rfloor) \cdot d^{\eta \log n/\|Z\|_\infty}. \tag{46}$$

Recall the definition of $c_b = \sup\{\alpha + \gamma_b\phi(\alpha,t)\} = \sup\{\alpha + \gamma_b\,[a \vee \phi(\alpha,t)]\}$, if $a < 0$. If $\phi(\alpha - \eta, t) > -\infty$, we write

$$\begin{aligned} &\mathbf{P}\left\{\frac{\log P_m(k,h)}{\log n} \ge \frac{c_b - \frac{h}{\log n} + \epsilon/2}{\gamma_b + \beta}\right\} \\ \le\;& \mathbf{P}\left\{\frac{\log P_m(k,h)}{\log n} \ge \frac{c_b - (\alpha - \eta) + \epsilon/2 - 2\eta}{\gamma_b + \beta}\right\} \\ \le\;& \mathbf{P}\left\{\frac{\log P_m(k,h)}{\log n} \ge \frac{\gamma_b\,[a \vee \phi(\alpha - \eta, t)] + \epsilon/2 - 2\eta}{\gamma_b + \beta}\right\}. \end{aligned}$$

Using (46), we can choose $\eta$ and $\beta$ small enough that the probability above is bounded by

$$\mathbf{P}\left\{P_m(\lfloor t\log n\rfloor, (\alpha-\eta)\log n) \geq n^{a\vee\phi(\alpha-\eta,t)+\epsilon/(4\gamma_b)}\right\} \quad \leq \quad n^{-\epsilon/(8\gamma_b)},$$

for $n$ large enough, by Theorem 4. This implies that

$$\sup_{h\leq K,h} \mathbf{P}\left\{\frac{\log P_m(k,h)}{\log n} \geq \frac{c_b - \frac{h}{\log n} + \epsilon/2}{\gamma_b+\beta}\right\} \;\leq\; n^{-\epsilon/(8\gamma_b)}.$$

As a consequence, recalling (42) and (45),

$$\begin{aligned}\mathbf{P}\left\{H_{n,b} \geq (c_b+\epsilon)\log n\right\} \quad &\leq \quad \sum_{k\leq K,h} n^{-\epsilon/(8\gamma_b)} + \sum_{k\leq K} n^{-\delta\epsilon/(5\gamma_b)} + o(1)\\ &\leq \quad O\left(n^{-\epsilon/(8\gamma_b)}\log^2 n\right) + O\left(n^{-\delta\epsilon/(4\gamma_b)}\log n\right) + o(1),\end{aligned}$$

since $P(k,h) = 0$ for all $k \leq K$ unless $h \leq K\|Z\|_\infty$. It follows that

$$\mathbf{P}\left\{H_{n,b} \geq (c_b+\epsilon)\log n\right\} \xrightarrow[n\to\infty]{} 0,$$

which completes the proof of the upper bound. □

# 7 Applications

## 7.1 Standard $b$-tries

We shall first consider simple well-known examples. We start with the case of standard unweighted trie. We show that the following theorem follows from Theorem 1.

**Theorem 7.** *Consider an unweighted $b$-trie $T_{n,b}$ on $n$ independent sequences of characters of $\{1,\ldots,d\}$ generated by a memoryless source with distribution $p_1 \geq \cdots \geq p_d > 0$. Let $H_{n,b}$ denote the height of $T_{n,b}$. Then,*

$$\frac{H_{n,b}}{\log n} \xrightarrow[n\to\infty]{} \frac{b+1}{-\log Q(b+1)}$$

*in probability, as $n \to \infty$.*

Theorem 7 is due to Szpankowski (1991). The case $b = 1$ was proved by Pittel (1985). See also Devroye et al. (1992). It has first been proved by considering the longest prefix of $(b+1)$-tuples of sequences, which is exactly what we do for the analysis of the spaghettis. It is interesting to note that for this case, one can obtain tight bounds on the height without distinguishing the core from the spaghettis. One of the reasons is that the weights are deterministic and identical for all the edges.

*Proof.* Here, we assume that $Z = 1$ almost surely. Then, $\phi(\alpha,t)$ is just the logarithmic profile studied by Park et al. (2006) in the binary case, or Broutin and Devroye (2007a). In this case, we have functions of one variable.

The core of the trie. For $1 \leq i \leq d$, we have $E = -\log p_i$ with probability $1/d$. We can compute the generating function of the cumulants: for any $\lambda,\mu\in\mathbb{R}$,

$$\Lambda(\lambda,\mu) = \log \mathbf{E}\left[e^{\lambda Z+\mu E}\right] = \lambda + \log\sum_{i=1}^d p_i^{-\mu} - \log d.$$

Then, the associated convex dual $\Lambda^\star$ is given by

$$\Lambda^\star(x,y) = \sup_{\lambda,\mu} \left\{ \lambda(x-1) + \mu y - \log \sum_{i=1}^d p_i^{-\mu} \right\} + \log d.$$

It follows that $\Lambda^\star(x,y)$ is infinite unless $x=1$. Writing $\mu = \mu(y)$ for the unique solution of

$$y = \frac{\partial \Lambda(\lambda,\mu)}{\partial \mu} = \frac{\sum_{i=1}^d \log p_i \cdot p_i^{-\mu}}{\sum_{i=1}^d p_i^{-\mu}}, \tag{47}$$

we have

$$\Lambda^\star(1,y) = \mu y - \log \sum_{i=1}^d p_i^{-\mu} + \log d.$$

For the height, the only relevant points of the profile are $\alpha \geq 1/\mathcal{E}$, where

$$\mathcal{E} = -\sum_{i=1}^d p_i \log p_i$$

is the entropy of the distribution $\{p_i, 1 \leq i \leq d\}$. In the range of interest, $\alpha \geq 1/\mathbf{E}E$, $I(1, 1/\alpha) = \Lambda^\star(1, 1/\alpha)$, and therefore

$$\phi(\alpha,\alpha) \;=\; \alpha \log d - \alpha \Lambda^\star\left(1, \frac{1}{\alpha}\right) \;=\; \mu(1/\alpha) + \alpha \log \sum_{i=1}^d p_i^{-\mu(1/\alpha)}. \tag{48}$$

For details see Broutin (2007) or Broutin and Devroye (2007a).

The behavior of spaghettis. In an unweighted trie, we have

$$Z^s = \begin{cases} 1 & \text{w.p.} \quad Q(b+1) \\ -\infty & \text{w.p.} \quad 1 - Q(b+1). \end{cases}$$

Therefore, for all $\lambda$,

$$\Lambda_s(\lambda) = \log \mathbf{E}\left[e^\lambda\right] + \log Q(b+1),$$

and hence $\Lambda_s^\star(x)$ is infinite unless $x=1$, in which case, we have $\Lambda_s^\star(1) = -\log Q(b+1)$. Then, by Lemma 7, we have

$$\gamma_b = \sup\left\{\gamma : \exists \rho \quad \Lambda_b^\star(\rho) \leq \frac{\rho}{\gamma}\right\} = \frac{1}{-\log Q(b+1)}.$$

The overall contribution. Now, by Theorem 1, the height $H_{n,b}$ of a random $b$-trie is asymptotic to $c_b \log n$ in probability, where

$$c_b = \sup_{\alpha>0} \left\{ \alpha + \frac{\phi(\alpha,\alpha)}{-\log Q(b+1)} \right\}.$$

This reduces to finding $\alpha_o$ such that

$$\left.\frac{\partial \phi(\alpha_o,\alpha_o)}{\partial \alpha}\right|_{\alpha=\alpha_o} = \log Q(b+1).$$

This occurs for $\alpha_o = Q(b+1)/\mathcal{E}(b+1)$, where

$$\mathcal{E}(b+1) = -\sum_{i=1}^d p_i^{b+1} \log p_i.$$

Indeed,

$$\frac{\partial \phi(\alpha,\alpha)}{\partial \alpha} = \log d - \Lambda^\star\left(1, \frac{1}{\alpha}\right) + \frac{1}{\alpha} \cdot \left.\frac{\partial \Lambda^\star(1,y)}{\partial y}\right|_{y=1/\alpha}.$$

Also, $\Lambda^\star(1,y) = \mu y - \Lambda(0,\mu)$, where $\mu = \mu(y)$ is defined in (47). For $\alpha = \alpha_o$, we have $\mu = \mu(1/\alpha_o) = -b-1$ and

$$\begin{aligned}
\left.\frac{\partial \phi(\alpha,\alpha)}{\partial \alpha}\right|_{\alpha_o} &= \log d - \Lambda^\star(1, 1/\alpha) + \frac{1}{\alpha} \cdot \left.\frac{\partial \Lambda^\star(1,y)}{\partial y}\right|_{y=1/\alpha_o} \\
&= \log d - \left(-(b+1)\frac{\mathcal{E}(b+1)}{Q(b+1)} - \log_d Q(b+1)\right) \\
&\quad + \frac{\mathcal{E}(b+1)}{Q(b+1)}\left(-(b+1) + \frac{\mathcal{E}(b+1)}{Q(b+1)} \left.\frac{\partial \mu(y)}{\partial y}\right|_{y=1/\alpha_o} - \frac{\mathcal{E}(b+1)}{Q(b+1)} \left.\frac{\partial \mu(y)}{\partial y}\right|_{y=1/\alpha_o}\right) \\
&= \log Q(b+1).
\end{aligned}$$

Now, observe that the line of support of $\phi(\alpha,\alpha)$ at $\alpha_o$ hits the vertical axis at

$$\phi(\alpha_o,\alpha_o) - \alpha_o \left.\frac{\partial \phi(\alpha,\alpha)}{\partial \alpha}\right|_{\alpha=\alpha_o} = -\left.\frac{\partial \Lambda^\star(1,y)}{\partial y}\right|_{y=1/\alpha_o} = b+1.$$

This implies that

$$c_b = \frac{b+1}{-\log Q(b+1)}.$$

This completes the proof of Theorem 7. For an illustration of this case, see Figure 7. $\square$

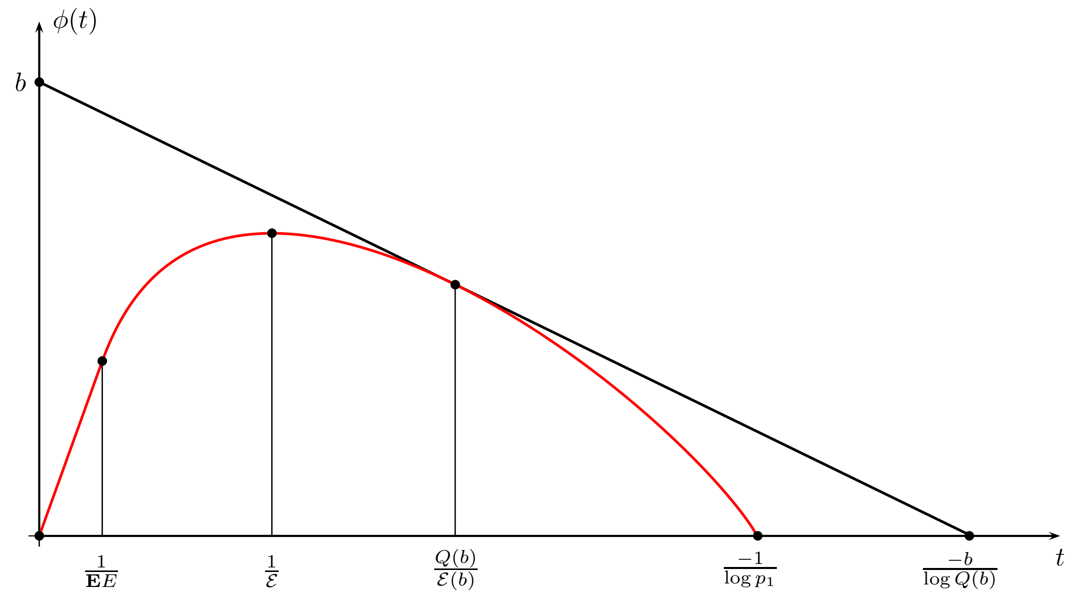


**Figure 7.** The diagram explaining the profile of the core (concave curve) and the behavior of the spaghettis (straight line) for the case of ordinary tries.

**Example: symmetric $b$-tries.** When $p_1 = p_2 = \cdots = p_d = 1/d$, the function $\Lambda^\star$ is degenerate in the sense that $\Lambda^\star$ is null at one point and infinite otherwise. Then, $\phi(\alpha, t)$ is triangular: $\phi(\alpha,\alpha)=\alpha\log d$, for $\alpha\le 1/\log d$ and $-\infty$ otherwise. In this case, $\log Q(b+1)=-b\log d$. It follows that

$$H_{n,b}\sim\left(\frac{1}{\log d}+\frac{1}{-\log Q(b+1)}\right)\log n=\left(1+\frac{1}{b}\right)\log_d n$$

in probability, as $n\to\infty$. In such a case, the contribution of the spaghettis is $1/b$ times that of the core. For instance, with ordinary tries, $b=1$ and the contribution of spaghettis is comparable to that of the core. Observe that this example is very special since all the spaghettis are born at the same level, which happens every time $p_1=\cdots=p_d=1/d$. This result was first obtained by Régnier (1981) in the case of a Poisson number of sequences. Flajolet and Steyaert (1982) and Flajolet (1983) obtained the limit distribution. See also Devroye (1984) and Pittel (1985).

| $b$ | 1 | 2 | 3 | 10 | 50 | 100 |
|---|---|---|---|---|---|---|
| $c_b(2)$ | 2.88539... | 2.16404... | 1.92359... | 1.58696... | 1.47154... | 1.45712... |
| $c_b(3)$ | 1.82047... | 1.36535... | 1.21365... | 1.00126... | 0.92844... | 0.91934... |
| $c_b(10)$ | 0.86858... | 0.65144... | 0.57905... | 0.47772... | 0.44298... | 0.43863... |

**Table 1.** Some numerical values for $c_b=c_b(d)$ the height of symmetric ordinary tries, as $b$ varies and $d\in\{2,3,10\}$.

## 7.2 Hybrid tries

Let $\mathcal{A}=\{1,\dots,d\}$ be the alphabet. Let $\{X^i, 1\le i\le n\}$ be the $n$ strings. In ordinary tries that use the array implementation, the order of the sequences is irrelevant. This is not the case any more in either the list-trie or the TST. In the following, we distinguish the *nodes* that constitute the high-level trie structure from the *slots* which make the low-level structure of a node, whether this latter is a linked-list or a binary search tree.

We now describe the way the internal structure of a node is constructed, in both list-tries and TSTs. Consider a node $u$. The subtree rooted at $u$ stores a subset of the sequences $X^i$, $1\le i\le n$. Let $\mathcal{N}_u\subset\{1,\dots,n\}$ be the set of their indices. So, in particular, the cardinality of $u$ is $N_u=|\mathcal{N}_u|$. The internal structure of the node is built using the sequences in increasing order of their index (see Figure 8). For a node $u$ at level $k$ in $T_\infty$, only the $k$-th characters of each sequence are used. Only the distinct characters matter. Let $\mathcal{A}_u\subset\mathcal{A}$ be the set of distinct characters appearing at the $k$-th position in the sequences $X^i$, $i\in\mathcal{N}_u$. The characters in $\mathcal{A}_u$ are ordered by first appearance. This induces a permutation $\sigma_u$ of $\mathcal{A}_u$. The internal structure of the node $u$ is built by successive insertions of the elements of $\sigma_u$ into an originally empty linked list, or binary search tree.

Both the list-tries and ternary search trees are built using the process we have just described. We shall now study each one of them more precisely.

## 7.3 List-tries

In the list-trie of de la Briandais (1959), the cost of branching to a character $a$ is just the index of $a$ in the permutation $\sigma_u$. For every node $u$, for which $\mathcal{A}_u=\mathcal{A}$, $\sigma_u$ is distributed as the sequence (in order) of first appearance of characters in an infinite string generated by the source. This fully describes the distribution of $Z$. That is, we have $Z_i$ is the index of $i$ in $\sigma$, and $Z=Z_K$, where $K$ is uniform in $\{1,\dots,d\}$. Observe that when $|\mathcal{A}_u|=1$, we have $Z=1$.

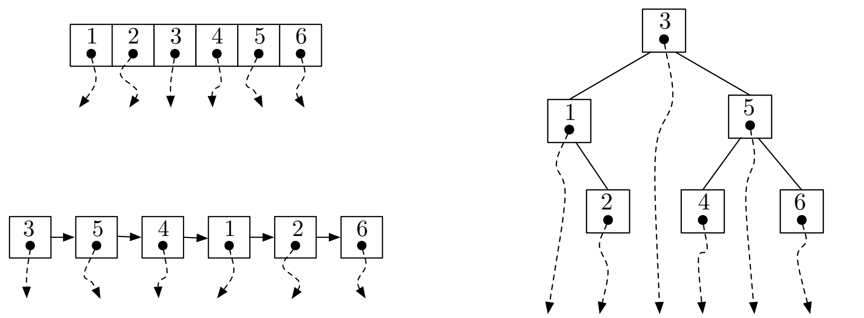


**Figure 8.** The different node structures used for the standard (top-left), list (bottom-left) and bst-trie (right) when the order of appearance of the characters is 3, 5, 4, 1, 2 and 6. The dashed arrows represent the pointers to further levels of the trie.

**Theorem 8.** *Let $H_{n,b}$ be the weighted height of a list-trie on $n$ sequences. Let $Z$ be as described above. Then, $H_{n,b} \sim c_b \log n$ in probability, as $n \to \infty$, where*

$$c_b = \sup_{\alpha,t>0} \left\{ \alpha + \frac{\phi(\alpha,t)}{-\log Q(b+1)} \right\},$$

*and $\phi(\cdot,\cdot)$ is the logarithmic profile of the trie weighted with $Z$.*

The theorem explains and characterizes the first term of the asymptotic expansion of the height for all distributions $\{p_1,\dots,p_d\}$ for $d<\infty$. For general distributions, it seems difficult to obtain a closed form for the height. We shall obtain more concrete values for a specific example.

**Example: symmetric list-tries.** In this case, for all $i$, we have $p_i = 1/d$ and $Z_i$ is uniform in $\{1,\dots,d\}$. Equivalently, $\mathcal{Z}^{(1,\dots,1)} = (1,2,\dots,d)$. Therefore, for any $\lambda,\mu \in \mathbb{R}$,

$$\Lambda(\lambda,\mu) \;=\; \log \mathbf{E}\left[e^{\lambda Z}\cdot d^{\mu}\right] + \mu \log d \;=\; \log\left(\sum_{i=1}^{d} e^{i\lambda}\right) + (\mu-1)\log d.$$

For $x\in[1,d]$, there exists $\lambda=\lambda(x)$ such that

$$x = \left.\frac{\partial \Lambda(\lambda,\mu)}{\partial \lambda}\right|_{(\lambda(x),1)} = \frac{\sum_{i=1}^{d} i e^{i\lambda}}{\sum_{i=1}^{d} e^{i\lambda}}. \tag{49}$$

Then, we have

$$\Lambda^\star(x,y) = \begin{cases} \lambda x - \log\left(\sum_{i=1}^{d} e^{i\lambda}\right) + \log d & \text{if } x\in[1,d], y=\log d \\ \infty & \text{otherwise.} \end{cases}$$

As for ordinary tries, in the range of interest,

$$\phi\left(\alpha, \frac{1}{\log d}\right) = 1 - \alpha\lambda(\alpha\log d) + \frac{\Lambda(\lambda(\alpha \log d),1)}{\log d}, \tag{50}$$

where $\lambda(\cdot)$ is defined in (49). In essence, $\phi(\alpha,t)$ is a function of $\alpha$ only. And we now write $\phi(\alpha)=\phi(\alpha,t)$ and $\Lambda(\lambda)=\Lambda(\lambda,1)$. By Theorem 8, looking for the constant $c_b$ boils down to finding $\alpha_o$ such that

$$\left.\frac{\mathrm{d}\phi(\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=\alpha_o} = \frac{1}{-\log Q(b+1)} = \frac{1}{b\log d},$$

and for this $\alpha_o$, we have

$$c_b = \alpha_o + \frac{\phi(\alpha_o)}{\log d}. \tag{51}$$

In other words, we have

$$\begin{aligned}
\left.\frac{\mathrm{d}\phi(\alpha)}{\mathrm{d}\alpha}\right|_{\alpha_o} &= -\lambda(\alpha_o \log d) - \alpha \left.\frac{\mathrm{d}\lambda(\alpha \log d)}{\mathrm{d}\alpha}\right|_{\alpha_o} + \frac{1}{\log d} \cdot \left.\frac{\mathrm{d}\Lambda(\lambda(\alpha \log d))}{\mathrm{d}\alpha}\right|_{\alpha_o} \\
&= -\lambda(\alpha_o \log d) - \alpha \left.\frac{\mathrm{d}\lambda(\alpha \log d)}{\mathrm{d}\alpha}\right|_{\alpha_o} + \frac{1}{\log d} \cdot \left.\frac{\mathrm{d}\Lambda(\lambda)}{\mathrm{d}\lambda}\right|_{\lambda(\alpha_o \log d)} \cdot \left.\frac{\mathrm{d}\lambda(\alpha \log d)}{\mathrm{d}\alpha}\right|_{\alpha_o} \\
&= -\lambda(\alpha_o \log d),
\end{aligned}$$

by (49), and hence $\lambda(\alpha_o \log d) = b \log d$. Hence, by (50) and (51),

$$c_b = \frac{1}{b \log d} + \frac{\Lambda(b \log d)}{b \log^2 d}.$$

Observe that this fully characterizes $c_b$ and holds for any symmetric weighted trie. For our case of symmetric list-tries, we obtain

$$c_b = c_b(d) = \frac{\log\left(\sum_{i=1}^{d} d^{bi}\right)}{b \log^2 d} \sim \frac{d}{\log d},$$

for large $d$. Some numerical values can be found in Table 2.

| $b$ | 1 | 2 | 3 | 10 | 50 | 100 |
|---|---|---|---|---|---|---|
| $c_b(2)$ | 3.28661... | 2.67491... | 2.52441... | 2.44289... | 2.44215... | 2.44206... |
| $c_b(3)$ | 3.12515... | 2.86870... | 2.83088... | 2.82022... | 2.81969... | 2.81963... |
| $c_b(10)$ | 4.92852... | 4.90959... | 4.90850... | 4.90723... | 4.90680... | 4.90675... |

**Table 2.** Some numerical values of $c_b = c_b(d)$ characterizing the height of symmetric list-tries.

**Remark.** Another equivalent description of the trees of de la Briandais is the following. We can consider list-tries in which children are added by a first-come-first-serve rule. Then, writing $(\sigma_1, \ldots, \sigma_d)$ for a uniformly random permutation of $\{1, \ldots, d\}$, we have $\mathcal{Z}^\tau = (1, \ldots, 1)$ a.s. for every permutation $\tau$ of $(1, 0, \ldots, 0)$, and $\mathcal{Z}^{(1,\ldots,1)} = (\sigma_1, \ldots, \sigma_d)$ with probability

$$p_{\xi_1} \cdot \frac{p_{\xi_2}}{1 - p_{\xi_1}} \cdot \frac{p_{\xi_3}}{1 - p_{\xi_1} - p_{\xi_2}} \cdots \frac{p_{\xi_{d-1}}}{1 - p_{\xi_1} - \cdots - p_{\xi_{d-2}}},$$

where $(\xi_1, \ldots, \xi_d)$ is the inverse permutation of $(\sigma_1, \ldots, \sigma_d)$, i.e., $\xi_i = j$ if $\sigma_j = i$.

### 7.4 Ternary search trees

In the ternary search trees introduced by Bentley and Sedgewick (1997), the implementation of a node uses a binary search tree. Hence, the cost of branching to a character $i \in \mathcal{A}$ at a node $u$ is the depth of $i$ in the binary search built from the (non-uniform) random permutation $\sigma_u$. When the node $u$ is of type $\tau_u = (1, \ldots, 1)$, the permutation $\sigma_u$ is distributed as the ordered list of first appearances of characters in an infinite string generated by the memoryless source with distribution $\{p_1, \ldots, p_d\}$.

Let $Z_i$ be distributed as the depth of $i$ in the binary search tree built from $\sigma$. Then, $\mathcal{Z}$ is distributed as $(Z_1, \ldots, Z_d)$ and $Z = Z_K$, where $K$ is uniform in $\{1, \ldots, d\}$. When $u$ is a non-branching node, i.e., $\tau_u$ is a permutation of $(1, 0, \ldots, 0)$, then the depth of the unique child is always one: $Z^s = 1$ almost surely. By Theorem 1, we obtain:

**Theorem 9.** *Let $H_{n,b}$ be the weighted height of a b-TST on n sequences. Let $\sigma$ be a permutation of $\{1,\ldots,d\}$ built by sampling from $\{1,\ldots,d\}$ according to $p_1,\ldots,p_d$. Let $Z$ be the depth of a random node in a binary search tree built from $\sigma$. Let*

$$c_b = \sup_{\alpha,t>0}\left\{\alpha + \frac{\phi(\alpha,t)}{-\log Q(b+1)}\right\},$$

*where $\phi(\alpha,t)$ is the logarithmic profile defined in (10). Then, $H_{n,b}\sim c_b\log n$ in probability, as $n\to\infty$.*

The random variable $Z$ is complicated to describe in other terms for general distributions $\{p_1,\ldots,p_d\}$. Some parameters like the average value and the variance of $Z_i$, $1\le i\le d$, have been studied by Clément et al. (1998, 2001) and Archibald and Clément (2006). For this case, describing $Z$ and $\phi(\alpha,t)$ in a way that would lead to $c_b$ seems more difficult than for list-tries.

**Example: Symmetric TST.** We assume here that $p_1=p_2=\cdots=p_d=1/d$. In this case, the permutation $\sigma$ is just a uniform random permutation. Hence, $Z_i$ is the depth of the key $i$ in a random binary search tree. Observe that unlike in the case of list-tries, $Z_i$, $1\le i\le d$, do *not* have the same distribution. This is easily seen, since, for instance as $d\to\infty$, $\mathbf{E}Z_1\sim\log d$ whereas $\mathbf{E}Z_{\lfloor d/2\rfloor}\sim 2\log d$. However, we are only interested in the distribution of $Z$, that is, the depth of a uniform random node. This distribution is known exactly for random binary search trees, and is due to Lynch (1965) and Brown and Shubert (1984):

$$\mathbf{P}\{Z=k\} = \frac{2^{k-1}}{d\cdot d!}\sum_{j=k}^{d}\begin{bmatrix} d\\ j\end{bmatrix}, \tag{52}$$

where $\begin{bmatrix} n\\ k\end{bmatrix}$ denotes the Stirling number of the first kind with parameter $n$ and $k$, that is the number of ways to divide $n$ objects into $k$ nonempty cycles (see Sedgewick and Flajolet, 1996; Mahmoud, 1992). Using (52), it is possible to compute the cumulant generating function $\Lambda$, and $\phi(\alpha,t)$. Numerical values for the constant $c_b=c_b(d)$ such that $H_n\sim c_b\log n$ in probability as $n\to\infty$ are given in Table 3. Observe that when $d=2$, TST are equivalent to list-tries. In general, using the computations we did in the case of symmetric list-tries,

$$\begin{aligned} c=c(d) &= \frac{1}{\log d}+\frac{1}{\log^2 d}\log\left(\sum_{i=1}^{d}\sum_{j=i}^{d}\frac{2^{i-1}}{d\cdot d!}\begin{bmatrix} d\\ j\end{bmatrix} d^i\right)\\ &= \frac{1}{\log d}+\frac{1}{\log^2 d}\log\left(\frac{(2d)\cdot(2d+1)\cdots(3d-1)-d!}{d!(2d-1)}\right)\sim\frac{d\log(27/4)}{\log^2 d}\end{aligned}$$

(see, e.g., Mahmoud, 1992, p. 79). Numerical values for the constant $c=c(d)$ are given in Table 3.

| $b$ | 1 | 2 | 3 | 10 | 50 | 100 |
|---|---|---|---|---|---|---|
| $c_b(2)$ | 3.28661... | 2.67491... | 2.52441... | 2.44289... | 2.44215... | 2.44206... |
| $c_b(3)$ | 2.90777... | 2.66010... | 2.65121... | 2.65118... | 2.65117... | 2.65116... |

**Table 3.** Some numerical values of $c_b=c_b(d)$ characterizing the height of symmetric ternary search trees.

### 7.5 Path imbalance

The question of path imbalance for random trees has been raised by Knuth in his keynote talk at the Conference on Analysis of Algorithms in 2004. The imbalance of a node in a binary tree is the difference between the number of left and right edges on its path from the root. Kuba and Panholzer (2007) have addressed this question and determined the limit distribution for the imbalance of a specified or randomly chosen node. Mahmoud (2007) and Christophi and Mahmoud (2007) have dealt with similar questions in tries. See also Janson (2006) who treated the related issue of left and right path lengths of random binary trees. We are interested in the *maximum* path imbalance in trees. The first order asymptotics have been described by Broutin and Devroye (2006) for binary search trees. Here, we use Theorem 1 to answer the question of the *maximum path imbalance in tries*. We assume that the tries are binary with symbol probabilities $p$ and $q = 1 - p$. In this case, the average imbalance is

$$\mathbf{E}\Delta_n = (p-q)\left(\frac{\log n}{\mathcal{E}} + O(1)\right),$$

and the rescaled imbalance of a random node has a Gaussian limit law :

$$\frac{\Delta_n - \frac{p-q}{\mathcal{E}}\log n}{\sqrt{\log n}} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{\log^2(pq)}{\mathcal{E}^3}\right)$$

(see Christophi and Mahmoud, 2007; Mahmoud, 2007). Observe in particular that when $p = q$, the typical range of $\Delta_n$ is $O(\sqrt{\log n})$. As we will see in the following, this contrasts with the behavior of the maximum imbalance in the trie which is $\Omega(\log n)$ as soon as $p > 0$. We let $B_n$ be the maximal path imbalance in a random $b$-trie. Then, as we will outline below, there is a constant $c_b$ depending upon $b$ and $p$ only such that

$$\frac{B_n}{\log n} \to c_b \text{ in probability}$$

as $n \to \infty$.

**Remark.** A related model involves tries whose edges are weighted with $+1$ or $-1$, independently for each edge using a fair coin flip. The values of the weights are then independent of the direction of the edges and we have

$$(Z,E) = \begin{cases} (+1, -\log p) & \text{w.p.} \quad 1/4 \\ (-1, -\log p) & \text{w.p.} \quad 1/4 \\ (+1, -\log q) & \text{w.p.} \quad 1/4 \\ (-1, -\log q) & \text{w.p.} \quad 1/4. \end{cases}$$

For this model, the average weight on the path to a random node has a centered Gaussian limit law, and the typical range is $O(\sqrt{\log n})$ for every $p \in (0,1)$.

THE CORE OF THE TRIE. In the case of imbalance, the vector of interest in the core is $\big((+1, -\log p), (-1, -\log q)\big)$ and hence a typical component (after symmetrization) is

$$(Z,E) = \begin{cases} (+1, -\log p) & \text{w.p.} \quad 1/2 \\ (-1, -\log q) & \text{w.p.} \quad 1/2. \end{cases}$$

Therefore,

$$\Lambda(\lambda,\mu) = \log\left(e^{\lambda - \mu \log p} + e^{-\lambda - \mu\log q}\right) - \log 2. \tag{53}$$

By definition, $\Lambda^\star(x,y) = \sup\{\lambda x + \mu y - \Lambda(\lambda,\mu)\}$ and $\Lambda(x,y) = \infty$ unless

$$\begin{cases} x &= \dfrac{\partial\Lambda(\lambda,\mu)}{\partial\lambda} &= \dfrac{e^{-\lambda-\mu\log p} - e^{-\lambda-\mu\log q}}{e^{-\lambda-\mu\log p} + e^{-\lambda-\mu\log q}} \\ y &= \dfrac{\partial\Lambda(\lambda,\mu)}{\partial\mu} &= \dfrac{-e^{-\lambda-\mu\log p}\log p - e^{-\lambda-\mu\log q}\log q}{e^{-\lambda-\mu\log p} + e^{-\lambda-\mu\log q}}. \end{cases}$$

Then we have

$$\left\{\begin{array}{ccccc} (1-x)e^{\lambda-\mu\log p} & + & (-1-\alpha)e^{-\lambda-\mu\log q} & = & 0 \\ (-\log q - y)e^{\lambda-\mu\log p} & + & (-\log q - y)e^{-\lambda-\mu\log q} & = & 0. \end{array}\right.$$

This leads to the equivalent system

$$\left\{\begin{array}{lcl} 2\lambda+\mu\log(q/p) & = & \log\left(\dfrac{1+x}{1-x}\right) \\ \dfrac{x+1}{x-1} & = & \dfrac{\log q+y}{\log p+y}. \end{array}\right. \tag{54}$$

For $x$ and $y$ satisfying (54),

$$\begin{aligned} \Lambda^\star(x,y) &= \lambda x+\mu y-\Lambda(\lambda,\mu) \\ &= \lambda x+\mu y+\log 2-\log\left(e^{\lambda-\mu\log p}+e^{-\lambda-\mu\log q}\right) \\ &= \lambda x+\mu y+\log 2-\log\left(e^{2\lambda+\mu\log(q/p)}+1\right)+\lambda+\mu\log q \\ &= \lambda(x+1)+\mu(y+\log q)+\log(1-x). \end{aligned}$$

Using the expression for $\lambda$ in terms of $\mu$, we can express $\Lambda^\star(x,y)$ independently of $\lambda$ or $\mu$. Finally, we obtain

$$\Lambda^\star(x,y)=\left\{\begin{array}{ll} \dfrac{x+1}{2}\log\left(\dfrac{x+1}{1-x}\right)+\log(1-x) & \text{if } \dfrac{1+x}{x-1}=\dfrac{\log q+y}{\log p+y} \\ \infty & \text{otherwise.} \end{array}\right.$$

The entire profile $\phi(\alpha,t)$ can be obtained from the values on the line where $\Lambda^\star$ takes finite values.

$$\phi(\alpha,t)=t\log 2-\frac{\alpha-t}{2}\log\left(\frac{\alpha+t}{t-\alpha}\right)-t\log\left(1-\frac{\alpha}{t}\right)\qquad\text{when}\quad \frac{\alpha+t}{\alpha-t}=\frac{t\log q+1}{t\log p+1}$$

or for such $\alpha$ and $t=t(\alpha)$ we have, in terms of $\alpha$ only,

$$\phi(\alpha,t)=\frac{\alpha\log p+1}{\log pq}\cdot\log\left(\frac{1-\alpha\log q}{1+\alpha\log p}\right)+\frac{2-\alpha\log(q/p)}{\log(pq)}\cdot\log\left(\frac{\alpha\log p+1}{\alpha\log(q/p)-2}\right). \tag{55}$$

The spaghettis. The spaghettis are characterized by the random variable $Z^s$ such that

$$Z^s=\left\{\begin{array}{lll} -\infty & \text{w.p.} & 1-Q(b+1) \\ +1 & \text{w.p.} & pQ(b+1) \\ -1 & \text{w.p.} & qQ(b+1). \end{array}\right.$$

To compute the constant $\gamma_b$, we find the expressions for $\Lambda_s(\cdot)$ and $\Lambda_s^\star(\cdot)$. We have

$$\forall\beta\in\mathbb{R}\qquad \Lambda_s(\beta)=\log Q(b+1)+\log\left(pe^{\beta}+qe^{-\beta}\right).$$

For every $\rho\in\mathbb{R}$, $\Lambda_s^\star(\gamma)<\infty$ if and only if there exists $\beta$ such that

$$\frac{\mathrm{d}\Lambda(\beta)}{\mathrm{d}\beta}=\frac{pe^{\beta}-qe^{-\beta}}{pe^{\beta}+qe^{-\beta}}=\rho,$$

and then $\Lambda_s^\star(\rho)=\rho\beta-\Lambda_s(\beta)$. This yields

$$\Lambda_s^\star(\rho)=\left\{\begin{array}{ll} \frac{\rho}{2}\log\left(\frac{q}{p}\cdot\frac{1+\rho}{1-\rho}\right)-\log Q(b+1)-\log\left(\sqrt{\frac{1+\rho}{1-\rho}}+\sqrt{\frac{1-\rho}{1+\rho}}\right)-\frac{\log(pq)}{2} & \text{if } \rho\in[-1,1] \\ +\infty & \text{otherwise.} \end{array}\right.$$

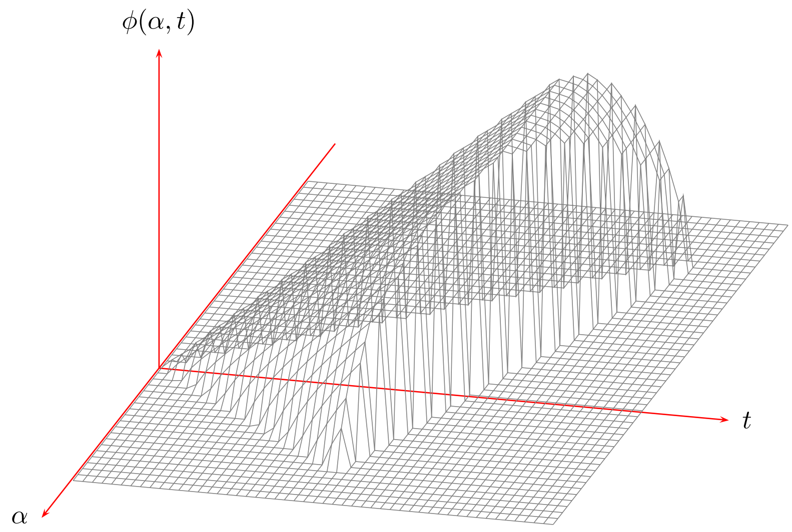


**Figure 9.** The weighted profile of the core for the question of path imbalance for $p = .3$. When looking at levels deep in the core, the nodes have mostly negative imbalance due to the fact that $p < 1/2$.

By Lemma 7 and the remark there, $\gamma_b$ is the given by the slope of a line with an end at the origin and which is tangent to the the curve $\{\Lambda_s^\star(\rho) : \rho \in [-1,1]\}$. In other words, $\gamma$ is given by

$$\frac{\mathrm{d}\Lambda_s^\star(\rho)}{\mathrm{d}\rho} = \frac{1}{\gamma} \qquad \text{and} \qquad \Lambda_s^\star(\rho) = \frac{\rho}{\gamma},$$

for some $\rho \in [-1,1]$. This leads to an explicit formula for $\gamma_b$. Indeed,

$$\frac{\mathrm{d}\Lambda_s^\star(\rho)}{\mathrm{d}\rho} = \frac{\rho}{2} \cdot \log\left(\frac{q}{p} \cdot \frac{1+\rho}{1-\rho}\right),$$

and hence

$$\frac{1}{\gamma_b} = \frac{\Lambda_s^\star(\rho_o)}{\rho_o} = \left.\frac{\mathrm{d}\Lambda^\star(\rho)}{\mathrm{d}\rho}\right|_{\rho=\rho_o} \qquad \text{where} \qquad \sqrt{\frac{1+\rho_o}{1-\rho_o}} + \sqrt{\frac{1-\rho_o}{1+\rho_o}} = \frac{1}{pqQ(b+1)}.$$

It follows that

$$\rho_o = \sqrt{1 - 4p^2q^2Q^2(b+1)} \qquad \text{and} \qquad \gamma_b = 2\Big/\rho_o \cdot \log\left(\frac{q}{p} \cdot \frac{1+\rho_o}{1-\rho_o}\right). \tag{56}$$

Numerical values are given in Table 4.

Overal contribution. We are in a case where $\Lambda^\star(\cdot,\cdot)$ is degenerate and takes finite values on a line only. So we have

$$c_b = \sup_{\alpha,t}\{\alpha + \gamma_b\phi(\alpha,t)\} = \sup\left\{\alpha + \gamma_b\phi(\alpha,t) : \frac{\alpha+t}{\alpha-t} = \frac{t\log q + 1}{t\log p + 1}\right\}.$$

Therefore, we have

$$c_b = \alpha_o + \gamma_b\alpha_o \qquad \text{where} \qquad \left.\frac{\mathrm{d}\phi(\alpha,t(\alpha))}{\mathrm{d}\alpha}\right|_{\alpha=\alpha_o} = -\frac{1}{\gamma_b},$$

where $\phi(\alpha,t(\alpha))$ and $\gamma_b$ are given by (55) and (56), respectively. It is this constant $c_b$ that is the weak limit of $B_n/\log n$. Numerical values can be found in Table 4 below.

| $b$ | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|
| $\gamma_b$ | 0.66760... | 0.46085... | 0.36055... | 0.29896... | 0.25664... | 0.15328... |
| $c_b$ | 1.16161... | 1.10845... | 1.09658... | 1.09313... | 1.09200... | 1.09136... |

**Table 4.** Numerical values for the constants $\gamma_b$ and $c_b$ in the example of the path imbalance when $p = 0.4$.

# 8 Concluding remarks

In addition to yielding the asymptotic properties of the height of hybrid tries, our method sheds some new light on the family of digital trees in general. In particular, the decomposition of the tree into a core and hanging spaghettis yields new connections between tries and the digital search tree. Coffman and Eve (1970) proposed digital search trees to improve the search costs in tries that are far from optimal. The main idea is to move the data from the external nodes to the internal nodes, which reduces the depth of the data, and hence the costs. This roughly speaking shaves off the spaghettis, as the weighted profile of digital search trees is identical to that of the core. Observe that the constant that would characterize the height of spaghettis in digital search trees would be $\gamma_b = 0$, since the height of a digital search tree on $m$ sequences is bounded by $m$ and digital search tree fits into the model of trees with bounded height of Broutin, Devroye, and McLeish (2007). So in our formalism, the height $H_{n,b}$ is asymptotic to $c_b \log n$ in probability with $c_b = \sup\{\alpha : \phi(\alpha, t) \geq 0\}$.

# 9 Acknowledgement

We thank Julien Clément for drawing our attention to the problem of the height of ternary search trees which lead to this document, and both referees for helping us improve the presentation.

# References

N. Alon, J. Spencer, and P. Erdős. *The Probabilitic Method.* Wiley, New York, NY, second edition, 2000.

M. Archibald and J. Clément. Average depth in binary search tree with repeated keys. In *Fourth Colloquium on Mathematics and Computer Science*, pages 309–320, 2006.

K. B. Athreya and P. E. Ney. *Branching Processes.* Springer, Berlin, 1972.

J. L. Bentley and R. Sedgewick. Fast algorithm for sorting and searching strings. In *Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 360–369, 1997.

N. Broutin. *Shedding New Light on Random Trees.* Phd thesis, McGill University, Montreal, 2007.

N. Broutin and L. Devroye. Large deviations for the weighted height of an extended class of trees. *Algorithmica*, 46:271–297, 2006.

N. Broutin and L. Devroye. The core of a trie. In *International Conference on Analysis of Algorithms*, 2007a. To appear.

N. Broutin and L. Devroye. The height of list tries and TST. In *International Conference on Analysis of Algorithms*, 2007b. (13 pages). To appear.

N. Broutin, L. Devroye, and E. McLeish. Weighted height of random trees. Manuscript (46 pages), 2007.

G.G. Brown and B.O. Shubert. On random binary trees. *Mathematics of Operations Research*, 9:43–65, 1984.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

C.A. Christophi and H. M. Mahmoud. One-sided variations on tries: path imbalance, climbing, and key sampling. In *Proceedings of the International Conference on Analysis of Algorithms (AofA)*, pages 301–310, 2007.

H. A. Clampett. Randomized binary searching with tree structures. *Communications of the ACM*, 7(3):163–165, 1964.

J. Clément, P. Flajolet, and B. Vallée. The analysis of hybrid trie structures. In *9th annual ACM-SIAM Symposium on Discrete Algorithms*, pages 531–539, Philadelphia, PA, 1998. SIAM Press.

J. Clément, P. Flajolet, and B. Vallée. Dynamical source in information theory: a general analysis of trie structures. *Algorithmica*, 29:307–369, 2001.

E. G. Coffman and J. Eve. File structures using hashing functions. *Communications of the ACM*, 13:427–436, 1970.

H. Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. In *Colloque Consacré à la Théorie des Probabilités*, volume 736, pages 5–23. Hermann, Paris, 1938.

R. de la Briandais. File searching using variable length keys. In *Proceedings of the Western Joint Computer Conference, Montvale, NJ, USA*. AFIPS Press, 1959.

A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications*. Springer Verlag, second edition, 1998.

F. den Hollander. *Large Deviations*. American Mathematical Society, Providence, RI, 2000.

J.-D. Deuschel and D.W. Stroock. *Large Deviations*. American Mathematical Society, Providence, RI, 1989.

L. Devroye. Laws of large numbers and tail inequalities for random tries and PATRICIA trees. *Journal of Computational and Applied Mathematics*, 142:27–37, 2002.

L. Devroye. Universal asymptotics for random tries and PATRICIA trees. *Algorithmica*, 42: 11–29, 2005.

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

L. Devroye. A note on the height of binary search trees. *Journal of the ACM*, 33:489–498, 1986.

L. Devroye. Branching processes and their application in the analysis of tree structures and tree algorithms. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Springer Series on Algorithms and Combinatorics*, pages 249–314, Berlin, 1998. Springer.

L. Devroye, W. Szpankowski, and B. Rais. A note on the height of suffix trees. *SIAM Journal on Computing*, 21:48–53, 1992.

P. Flajolet. The ubiquitous digital tree. In B. Durand and W. Thomas, editors, *STACS 2006, Annual Symposium on Theoretical Aspects of Computer Science*, volume 3884 of *Lecture Notes in Computer Science*, pages 1–22, Marseille, February 2006.

P. Flajolet. On the performance evaluation of extendible hashing and trie searching. *Acta Informatica*, 20:345–369, 1983.

P. Flajolet and J.M. Steyaert. A branching process arising in dynamic hashing, trie searching and polynomial factorization. In M. Nielsen and E.M. Schmidt, editors, *Automata, Languages and Programming: Proceedings of the 9th ICALP Conference*, volume 140 of *Lecture Notes in Computer Science*, pages 239–251. Springer, 1982.

E. Fredkin. Trie memory. *Communications of the ACM*, 3(9):490–499, 1960.

S. Janson. Left and right pathlengths in random binary trees. *Algorithmica*, 46:419–429, 2006.

S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley, New York, 2000.

M. Kuba and A. Panholzer. The left-right-imbalance of binary search trees. *Theoretical Computer Science*, 370:265–278, 2007. Manuscript.

W.C. Lynch. More combinatorial properties of certain trees. *Computing J.*, 7:299–302, 1965.

H. Mahmoud. *Evolution of Random Search Trees*. Wiley, New York, 1992.

H. M. Mahmoud. Imbalance in random digital trees. Submitted, 2007.

G. Park, H.K. Hwang, P. Nicodème, and W. Szpankowski. Profile of tries. Manuscript, 2006.

B. Pittel. Asymptotic growth of a class of random trees. *The Annals of Probability*, 13: 414–427, 1985.

M. Régnier. On the average height of trees in digital search and dynamic hashing. *Information Processing Letters*, 13:64–66, 1981.

R. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithm*. Addison-Wesley, Reading, MA, 1996.

W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York, 2001.

W. Szpankowski. On the height of digital trees and related problems. *Algorithmica*, 6: 256–277, 1991.