Height Distributions in Digital Trees*

C. Knessl
Dept. Mathematics, Statistics & Computer Science
University of Illinois at Chicago
Chicago, Illinois 60607-7045
U.S.A.

W. Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

May 15, 2000

*Research supported by NSF Grants NCR-9415491 and C-CR 9804760

Outline

1. Digital Trees – Definitions
2. Recurrences for Height Distributions
3. Main Results for b-Tries
4. Main Results for PATRICIA Tries
5. Main Results for Digital Trees
6. Numerical Comparisons

Digital Trees – Definitions

We study heights in digital trees such as (generalized) tries and PATRICIA tries, built from n random strings generated by an **unbiased memoryless source** (i.e., all symbols are equally likely) over a binary alphabet.

Tries, b-Tries, PATRICIA Tries and Digital Search Trees

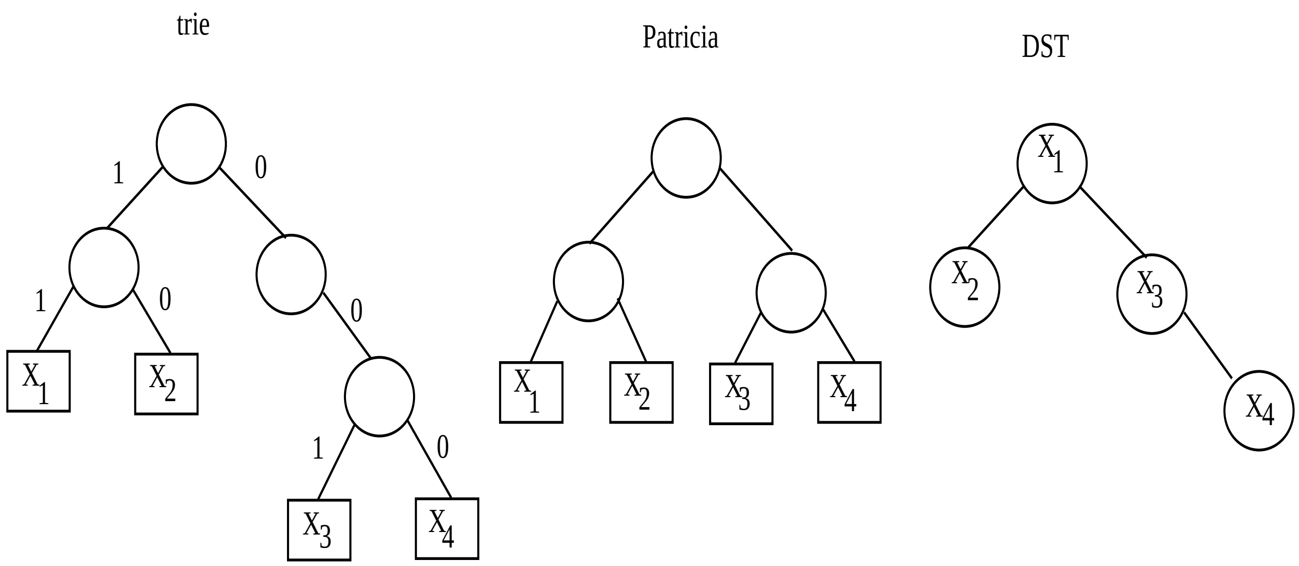


Figure 1: A trie, PATRICIA trie and a digital search tree (DST) built from the following four strings $X_1 = 11100\ldots$, $X_2 = 10111\ldots$, $X_3 = 00110\ldots$, and $X_4 = 00001\ldots$.

Recurrences for the b-Tries Height Distribution

Let $\mathcal{H}_n^T$ and $\mathcal{H}_n^P$ denote, respectively, the height of a b-trie and a PATRICIA trie. Their probability distributions are

$$\bar{h}_n^k = \Pr\{\mathcal{H}_n^T \leq k\} \qquad \text{and} \qquad h_n^k = \Pr\{\mathcal{H}_n^P \leq k\}.$$

Since

$$\mathcal{H}_n^T = \max\{\mathcal{H}_i^{LT}, \mathcal{H}_{n-i}^{RT}\} + 1,$$

where $\mathcal{H}_i^{LT}$ and $\mathcal{H}_{n-i}^{RT}$ denote, respectively, the left subtree and the right subtree of sizes i and $n-i$ (which happens with probability $2^{-n}\binom{n}{i}$). The probability distribution $\bar{h}_n^k$ satisfies the following recurrence

$$\bar{h}_n^{k+1} = 2^{-n} \sum_{i=0}^{n} \binom{n}{i} \bar{h}_i^k \bar{h}_{n-i}^k, \qquad k \geq 0$$

with the initial condition(s)

$$\bar{h}_n^0 = 1, \; n = 0, 1, 2, \ldots, b; \qquad \text{and} \qquad \bar{h}_n^0 = 0, \; n > b.$$

We note that for tries $\bar{h}_n^k = 0$ for $n > b2^k$ (corresponding to a balanced tree).

PATRICIA Height Distribution Recurrence

We have for all $n, k \geq 0$

$$h_n^{k+1} = 2^{-n+1} h_n^{k+1} + 2^{-n} \sum_{i=1}^{n-1} \binom{n}{i} h_i^k h_{n-i}^k, \quad k \geq 0$$

with the initial conditions

$$h_0^0 = h_1^0 = 1 \qquad \text{and} \qquad h_n^0 = 0, \quad n \geq 2.$$

Unlike b-tries, in a PATRICIA trie the left and the right subtrees cannot be empty (which occurs with probability 2^{-n+1}).

We note that for PATRICIA tries $h_n^k = 0$ for $n > 2^k$.

In addition, for PATRICIA we have the following boundary condition: $h_n^k = 1$ for $k \geq n$. It asserts that the height in a PATRICIA trie cannot be bigger than n (due to the elimination of all one-way branches).

DST Height Distribution Recurrence

Let $h_n^k = \Pr\{\mathcal{H}^{DST} \leq k\}$ be the height distribution in a DST. The following holds:

$$h_{n+1}^{k+1} = \sum_{i=0}^{n} \binom{n}{i} 2^{-n} h_i^k h_{h-i}^k, \quad k \geq 0$$

with the initial condition

$$h_0^0 = h_1^0 = 1; \quad h_n^0 = 0, n \geq 2.$$

The exponential generating function $H_k(z) = \sum_{n=0}^{\infty} \frac{h_n^k z^n}{n!}$ fulfills

$$H'_{k+1}(2z) = H_k^2(z), \qquad k \geq 0$$

but we do not know how to solve it.

Derivation of the Trie Height Distribution

We applied generating functions and the saddle point method to derive the height distribution.

Let

$$H^k(z) = \sum_{n \geq 0} \bar{h}_n^k \frac{z^n}{n!}.$$

The recurrence translates into $H^{k+1}(z) = [H^k(z/2)]^2$ and then

$$H^k(z) = \left(H^0(z2^{-k})\right)^{2^k}$$

with $H^0(z) = 1 + z + \cdots + z^b/b!$. By Cauchy's formula, we obtain

$$\bar{h}_n^k = \frac{n!}{2\pi i} \oint \left(1 + z2^{-k} + \frac{z^2}{2!}2^{-2k} + \cdots + \frac{z^b}{b!}2^{-bk}\right)^{2^k} z^{-n-1} dz$$

where the contour integral is a loop around the origin in the z-plane. This integral is evaluated by the saddle point method for various limits where $n, k \to \infty$.

Main Results – b-Tries

Theorem 1. *The distribution of the height of b-tries has the following asymptotic expansions for fixed b:*

(i) Right-Tail Region*:* $k \to \infty$, $n = O(1)$*:*

$$\Pr\{\mathcal{H}_n^T \le k\} = \bar{h}_n^k \sim 1 - \frac{n!}{(b+1)!(n-b-1)!} 2^{-kb}.$$

(iii) Left-Tail Region*:* $k, n \to \infty$ *with* $j = b2^k - n = O(1)$

$$\bar{h}_n^k \sim \sqrt{2\pi n} \frac{n^j}{j!} b^n \exp\left(-(n+j)\left(1 + b^{-1}\log b!\right)\right).$$

Central Regime

(ii) Central Regime: *$k, n \to \infty$ with $\xi = n2^{-k}$, $0 < \xi < b$:*

$$\bar{h}_n^k \sim A(\xi; b) e^{n\phi(\xi; b)},$$

where

$$\begin{aligned} \phi(\xi; b) &= -1 - \log \omega_0 + \frac{1}{\xi}\left(b \log(\omega_0 \xi) - \log b! \right. \\ &- \left. \log\left(1 - \frac{1}{\omega_0}\right)\right), \\ A(\xi; b) &= \frac{1}{\sqrt{1 + (\omega_0 - 1)(\xi - b)}}. \end{aligned}$$

In the above, $\omega_0 = \omega_0(\xi; b)$ is the solution to

$$1 - \frac{1}{\omega_0} = \frac{(\omega_0 \xi)^b}{b! \left(1 + \omega_0 \xi + \frac{\omega_0^2 \xi^2}{2!} + \cdots + \frac{\omega_0^b \xi^b}{b!}\right)}.$$

We note that when $b = 1$ or 2, we can obtain ω_0 explicitly:

$$\begin{aligned} \omega_0(\xi; 1) &= \frac{1}{1 - \xi}, \\ \omega_0(\xi; 2) &= \frac{2}{1 - \xi + \sqrt{1 + 2\xi - \xi^2}}. \end{aligned}$$

Limiting Distribution for the Trie Height

For arbitrary b, we have $\omega_0 \to \infty$ as $\xi \to b^-$ and $\omega_0 \to 1$ as $\xi \to 0^+$. More precisely,

$$\begin{aligned}\omega_0 &= 1 - \frac{\xi^b}{b!} + O(\xi^{b+1}), \qquad\qquad \xi \to 0\\ \omega_0 &= \frac{1}{b-\xi} + \frac{b-1}{b} + O(b-\xi), \quad \xi \to b.\end{aligned}$$

Observe that the height distribution is close to 1 in the **right-tail region**, and exponentially small in the **central region** and the **left-tail region**.
The most probability mass concentrates in the central region when $\xi \to 0$. Then the distribution has a phase-transition from zero to one (and the probability mass is $O(1)$ in this region).

For $\xi \to 0$ we obtain

$$\begin{aligned}\Pr\{\mathcal{H}_n^T \le k\} &\sim A(\xi)e^{n\phi(\xi)} \sim \exp\left(-\frac{n\xi^b}{(b+1)!}\right) \qquad \xi \to 0\\ &= \exp\left(-\frac{n^{1+b}2^{-kb}}{(b+1)!}\right).\end{aligned}$$

where the most mass is concentrated.

Asymptotic Distribution

Let us estimate the probability mass of $\mathcal{H}_n^T$ around

$$(1 + 1/b) \log_2 n + x$$

where x is a real fixed value. We observe that

$$\begin{aligned} \Pr\{\mathcal{H}_n^T &\leq (1 + 1/b) \log_2 n + x\} \\ &= \Pr\{\mathcal{H}_n^T \leq \lfloor (1 + 1/b) \log_2 n + x \rfloor\} \\ &\sim \exp\left(-\frac{1}{(1+b)!} 2^{-bx + b\langle (1+b)/b \cdot \log_2 n + x \rangle} \right) , \end{aligned}$$

where $\langle x \rangle$ is the fractional part of x, that is, $\langle x \rangle = x - \lfloor x \rfloor$.

Corollary 1. [Flajolet, Devroye, Pittel, Jacquet and Régnier] *While the limiting distribution of the height for b-tries does not exist, the following lower and upper envelopes can be established*

$$\begin{aligned} \liminf_{n\to\infty} \Pr\{\mathcal{H}_n^T &\leq (1 + 1/b) \log_2 n + x\} \\ &= \exp\left(-\frac{1}{(1+b)!} 2^{-b(x-1)} \right) , \\ \limsup_{n\to\infty} \Pr\{\mathcal{H}_n^T &\leq (1 + 1/b) \log_2 n + x\} \\ &= \exp\left(-\frac{1}{(1+b)!} 2^{-bx} \right) . \end{aligned}$$

Sketch of the Height Distribution

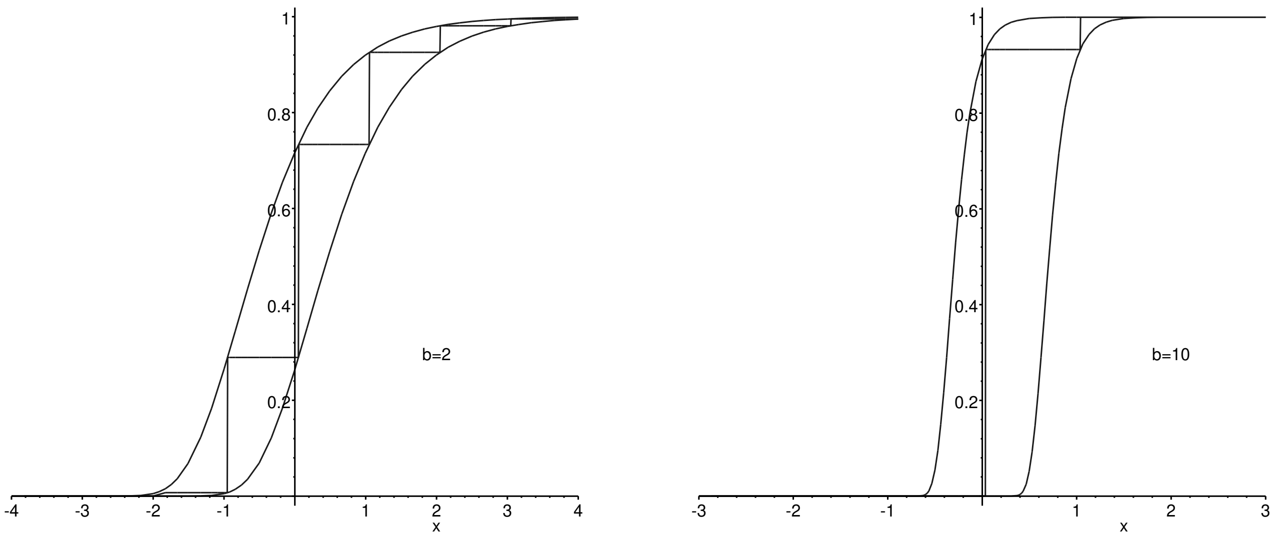


Figure 2: Asymptotic distributions for the height and their corresponding lower and upper bound for tries with $b = 2$ (left) and $b = 10$ (right).

Height for Tries by a Direct Method

We can recover the results of the height distribution for b-tries by using a direct method, that is, working directing with the recurrence. More precisely:

1. RIGHT-TAIL REGION: $k \to \infty$, $n = O(1)$:
   We use here **linearization** since the recurrence in this region reduces to a linear recurrence. Indeed, set $\bar{h}_n^k = 1 - G_n^k$, and then G_n^k satisfies

   $$G_n^{k+1} = 2\left(\frac{1}{2}\right)^n \sum_{i=0}^{n-b-1} \binom{n}{i} G_{n-i}^k - \left(\frac{1}{2}\right)^n \sum_{i=b+1}^{n-b-1} \binom{n}{i} G_i^k G_{n-i}^k.$$

   For $b+1 \leq n \leq 2b+1$, the non-linear term vanishes and we are left with a *linear* recurrence. We can solve it exactly by first solving the linear problem for $n \in [b+1, 2b+1]$. Then, using this solution to compute explicitly the nonlinear term in the above for $n \in [2b+2, 3b+2]$, etc.
2. LEFT-TAIL REGION: $k, n \to \infty$ with $j = b2^k - n$
   In this region we solve directly the recurrence for some sample values of j (say $j = b2^k, b2^k - 1, b2^k - 2$, etc) and guessing asymptotic expansion (which is later verified).

Direct Method – Continuation

3. CENTRAL REGIME: $k, n \to \infty$ with $\xi = n2^{-k}$, $0 < \xi < b$:
   Here we use **WKB method**. This method assumes that the asymptotic solution is a certain form, that is, defining $G(\xi; n) := G(n2^{-k}; n)$ we **postulate**

   $$G(\xi; n) \sim e^{n\phi(\xi)} \left[A(\xi) + \frac{1}{n} A^{(1)}(\xi) + \frac{1}{n^2} A^{(2)}(\xi) + \cdots \right]$$

   where $\phi(\xi)$ and $A^{(i}(\xi)$ are unknown. We compute asymptotically these functions by plugging the solution into the recurrence and finding them so that the recurrence is asymptotically satisfied.
4. We use finally **asymptotic matching** to relate the expansions to one another and to make sure that we covered all possible asymptotic regimes.

PATRICIA Height by the Direct Method

Using linearization in the right-tail region, direct computations in the left-tail region, WKB method in the central regime, and asymptotic matching to blend the above solutions, we prove the following result.

Theorem 2. *The distribution of PATRICIA tries has the following asymptotic expansions:*

(i) RIGHT-TAIL REGIME*:* $k, n \to \infty$ *with* $n - k = j = O(1)$*,* $j \geq 2$

$$\Pr\{\mathcal{H}_n^P \leq n-j\} = h_n^{n-j} \sim 1-\rho_0 K_j \cdot n! \cdot 2^{-n^2/2+(j-3/2)n},$$

where

$$\begin{aligned} K_j &= \frac{1}{j!} 2^{-j^2/2} 2^{3j/2} C_j, \\ C_j &= \frac{j!}{2\pi i} \oint \frac{z^{1-j} e^z}{2} \prod_{m=0}^{\infty} \left(\frac{1-\exp(-z 2^{-m-1})}{z 2^{-m-1}} \right) dz \end{aligned}$$

and $\rho_0 = \prod_{\ell=2}^{\infty} (1-2^{-\ell})^{-1} = 1.73137\ldots$

(ii) Central Regime: *$k, n \to \infty$ with $\xi = n2^{-k}$, $0 < \xi < 1$. Under the WKB hypothesis we have*

$$h_n^k \sim \sqrt{1 + 2\xi\Phi'(\xi) + \xi^2\Phi''(\xi)}e^{-n\Phi(\xi)}.$$

We know $\Phi(\xi)$ analytically only for $\xi \to 0$ and $\xi \to 1$. In particular, for $\xi \to 0$

$$\Phi(\xi) \sim \frac{1}{2}\rho_0 e^{\varphi(\log_2 \xi)}\xi^{3/2} \exp\left(-\frac{\log^2 \xi}{2\log 2}\right), \qquad \xi \to 0^+,$$

with

$$\begin{aligned}
\varphi(x) &= \frac{\log 2}{2}x(x+1) + \sum_{\ell=0}^{\infty} \log\left(\frac{1-\exp(-2^{x-\ell})}{2^{x-\ell}}\right) \\
&+ \sum_{\ell=1}^{\infty} \log(1-\exp(-2^{x+\ell})) \\
&= \Psi(x) - \frac{\log 2}{12} + \frac{1}{\log 2}\left(\frac{\gamma^2}{2} + \gamma(1) - \frac{\pi^2}{12}\right), \\
\Psi(x) &= \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \frac{1}{2\pi i\ell}\Gamma\left(1-\frac{2\pi i\ell}{\log 2}\right)\zeta\left(1-\frac{2\pi i\ell}{\log 2}\right)e^{2\pi i x\ell}.
\end{aligned}$$

In the above, $\Gamma(\cdot)$ *is the Gamma function,* $\zeta(\cdot)$ *is the Riemann zeta function,* $\gamma = -\Gamma'(1)$ *is the Euler constant, and* $\gamma(1)$ *is defined by the Laurent series* $\zeta(s) = 1/(s-1) + \gamma - \gamma(1)(s-1) + O((s-1)^2)$. *The function* $\Psi(x)$ *is periodic with a very small amplitude, i.e.,* $|\Psi(x)| < 10^{-5}$.

Moreover, for $\xi \to 1$ *the function* $\Phi(\xi)$ *becomes*

$$\Phi(\xi) \sim D_1 + (1-\xi)\log(1-\xi) - (1-\xi)(1+\log D_2), \qquad \xi \to 1^-$$

where $D_1 = 1 + \log(K_0^*)$ *and* $D_2 = K_1^* K_0^*/e$ *with*

$$\begin{aligned} K_0^* &= \prod_{\ell=1}^{\infty}\left(1 - 2^{-2^{\ell}+1}\right)^{2^{-\ell}} = .68321974\ldots, \\ K_1^* &= \prod_{\ell=1}^{\infty}\prod_{m=1}^{\infty}\left(1 - 2^{-2^{\ell+1}+2}\right)^{-1}\left[1 - 2^{-2^{\ell+m}+1}\right]^{2^{-m}} \\ &= 1.2596283\ldots \end{aligned}$$

(iii) Left-Tail Regime*:* $k, n \to \infty$ *with* $2^k - n = M = O(1)$

$$h_n^k \sim \frac{\sqrt{2\pi}}{M!} D_2^M n^{M+1/2} e^{-D_1 n}$$

where D_1 *and* D_2 *are defined above.*

PATRICIA Height – Discussions

1. Since $\Phi(\xi) > 0$, we see that in the regions (ii) and (iii), the PATRICIA height distribution is exponentially small in n, while in (i), $1 - h_n^k$ is super–exponentially small $(O(2^{-n^2/2}))$. Thus, (i) applies in the right tail of the distribution while (ii) and (iii) apply in the left tail. We wish to compute the range of k where h_n^k undergoes the transition from $h_n^k \approx 0$ to $h_n^k \approx 1$, as $n \to \infty$. This must be in the asymptotic matching region between (i) and (ii).

2. Using analytic depoissonization and asymptotic matching, we can show that C_j defined in (i) becomes as $j \to \infty$

$$C_j \sim \frac{j^{5/2}}{2} e^{\varphi(\alpha)} \exp\left(-\frac{1}{2}\frac{\log^2 j}{\log 2}\right),$$

where $\alpha = \langle \log_2 j \rangle$ is the fractional part. This will allow us to compute $\Phi(\xi)$ as $\xi \to 0$.

Discussions

3. The limiting form of (ii) as $\xi \to 0^+$ is (in the region the probability mass is $O(1)$)

$$h_n^k \sim \exp\left(-\frac{\rho_0}{2}e^{\varphi(\log_2 n)}\exp\left(-\frac{\log 2}{2}\left(\left(k+\frac{3}{2}-\log_2 n\right)^2-2\log_2 n-\frac{9}{4}\right)\right)\right)$$
$$= \exp\left(-\rho_0 e^{\varphi(\log_2 n)}2^{1/8}n\exp\left(-\frac{\log 2}{2}\left(k+1.5-\log_2 n\right)^2\right)\right)$$
$$= \exp\left(-\rho_0\cdot n\cdot\exp\left(-\frac{\log 2}{2}\left(k+1.5-\log_2 n\right)^2+\theta\right.\right.$$
$$\left.\left.+\Psi(\log_2 n)\right)\right)$$

where ρ_0 is defined in Theorem 2(i) and

$$\theta = \frac{1}{\log 2}\left(\frac{\gamma^2}{2}+\gamma(1)-\frac{\pi^2}{12}\right)+\frac{\log 2}{24} = -1.022401\ldots$$

while $|\Psi(\log_2 n)| \leq 10^{-5}$.

Discussions

4. We now set, for an integer ℓ,

$$\begin{aligned} k_\ell &= \left\lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \right\rfloor + \ell \\ &= \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} + \ell - \beta_n, \end{aligned}$$

where

$$\beta_n = \left\langle \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \right\rangle \in [0,1).$$

In terms of ℓ and β_n, the above becomes

$$\begin{aligned} \Pr\{\mathcal{H}_n^P &\le \lfloor \log_2 n + \sqrt{2\log_2 n} - 1.5 \rfloor + \ell\} \\ &\sim \exp\left(-\rho_0 e^{\theta + \Psi(\log_2 n)} 2^{-(\ell-\beta_n)^2/2 - (\ell-\beta_n)\sqrt{2\log_2 n}} \right). \end{aligned}$$

The form of the limiting distribution depends on the behavior of β_n!

Behavior of β_n

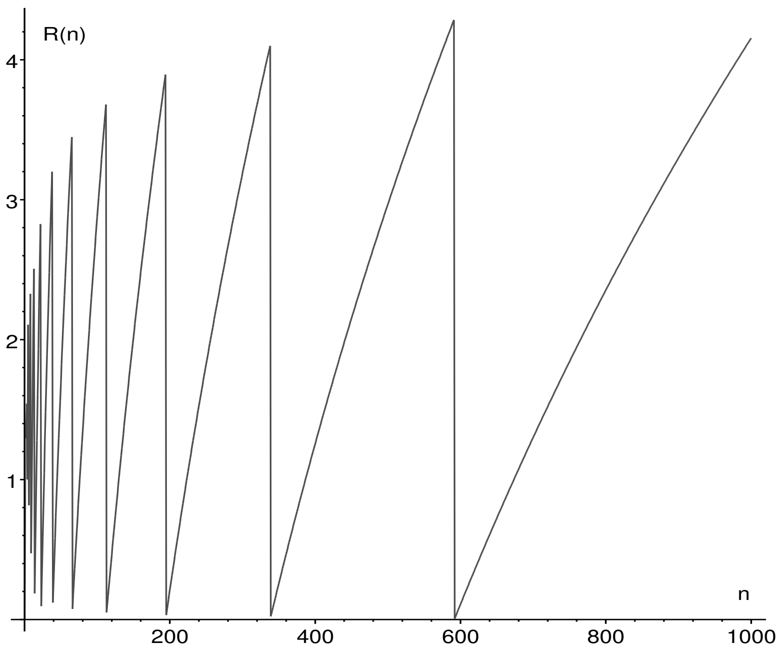


Figure 3: The function $R(n) = \langle \sqrt{2\log_2 n} \langle \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \rangle$ versus n.

Discussion

Copying the exponent from the last slide we must deal with

$$\exp\left(-\rho_0 e^{\theta+\Psi(\log_2 n)} 2^{-(\ell-\beta_n)^2/2-(\ell-\beta_n)\sqrt{2\log_2 n}}\right)$$

- If $0 < \beta_n < 1$, then the above is small for $\ell \le 0$, and it is close to one for $\ell \ge 1$. Thus the mass accumulates at one point $k = k_1$.
- But we can find a sequence n_i such that $n_i \to \infty$ as $i \to \infty$ but $\sqrt{2\log_2 n_i}\left\langle \log_2 n_i + \sqrt{2\log_2 n_i} - \frac{3}{2}\right\rangle$ remains bounded. Then, the expression above would be $O(1)$ for $\ell = 0$ (since $\beta_n\sqrt{2\log_2 n} = O(1)$). For $\ell = 1$, h_n^k would then be asymptotically close to 1. Thus, now the mass would accumulate at two points, namely, $k_0 = k_1 - 1$ and k_1.
- Finally, if $\beta_n = 1 - o(1)$ such that $(1 - \beta_n)\sqrt{2\log_2 n} = O(1)$, then the probability mass is concentrated on k_1 and $k_1 + 1$.

Conclusion

Corollary 2. *The asymptotic distribution of PATRICIA height is concentrated among the three points* $k_1 - 1$, k_1 *and* $k_1 + 1$ *where* $k_1 = \lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \rfloor + 1$, *that is,*

$$\Pr\{\mathcal{H}_n^P = k_1 - 1 \text{ or } k_1 \text{ or } k_1 + 1\} = 1 - o(1)$$

as $n \to \infty$. *More precisely: (i) there are subsequences* n_i *for which* $\Pr\{\mathcal{H}_{n_i}^P = k_1\} = 1 - o(1)$ *provided that*

$$R(n) = \sqrt{2\log_2 n} \left\langle \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \right\rangle \to \infty$$

as $i \to \infty$; *(ii) there are subsequences* n_i *for which* $\Pr\{\mathcal{H}_{n_i}^P = k_1 - 1 \text{ or } k_1\} = 1 - o(1)$ *provided that* $R(n_i) = O(1)$; *(iii) finally, there are subsequences* n_i *for which* $\Pr\{\mathcal{H}_{n_i}^P = k_1 \text{ or } k_1 + 1\} = 1 - o(1)$ *provided that* $\sqrt{2\log_2 n_i} - R(n_i) = O(1)$.

Corollary 3. *The average height of PATRICIA is*

$$\mathbf{E}[\mathcal{H}_n^P] = \log_2 n + \sqrt{2\log_2 n} + O(1).$$

Figures

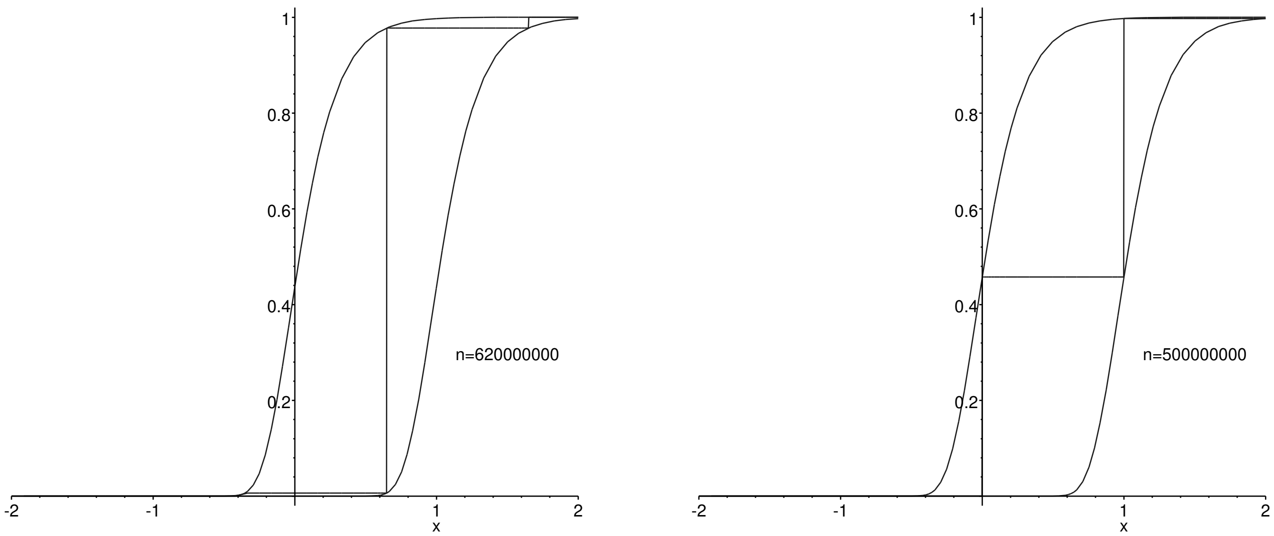


Figure 4: Asymptotic distributions for the height and their corresponding lower and upper bound for PATRICIA tries with $n = 6.2 \cdot 10^8$ (one-point distribution) and $n = 5 \cdot 10^8$ (two-points distribution).

DST Height

Using linearization in the right-tail region, direct computations in the left-tail region, WKB method in the central regime, and asymptotic matching to blend the above solutions, we prove the following result.

Theorem 3. *The distribution of the height for digital search trees built from n independent strings generated by an unbiased memoryless source (and also the length of the longest phrase in the Lempel-Ziv algorithm with a fixed number, n, of phrases) has the following asymptotic expansions:*

(i) Far Right-Tail Regime: $n, k \to \infty$, $n - k = j = O(1)$, $j \geq 2$

$$\Pr\{\mathcal{H}_n > k\} = 1 - h_n^k \sim 2^{-j^2/2} 2^{j/2} \frac{1}{(j-2)!} n^{j-2} 2^{-n^2/2} 2^{(j-1/2)n}.$$

Right-Tail Region

(ii) RIGHT-TAIL REGIME: $n, k \to \infty$, $0 < k/n < 1$ *or* $\alpha \equiv 1/(1-k/n) = n/(n-k) \in (1, \infty)$

$$1 - h_n^k \sim 2^{-k^2/2} 2^{-k/2} F_n^k,$$

where

$$\begin{aligned}
F_n^k &= \frac{n^n}{(n-k)^{n-k}k^k}\frac{\sqrt{\pi n}(n-k)^{3/2}}{\sqrt{2}\ k^{5/2}} I(\alpha),\\
I(\alpha) &= \frac{1}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{(\alpha-1)^m}\left[\prod_{\ell=1}^{m}(1-2^{-\ell})\right]^{-1}, \qquad \alpha > 2,\\
&= \frac{1}{2\pi i}\int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty}\frac{e^{z\log(\alpha-1)}}{\sin(\pi z)}\prod_{\ell=1}^{\infty}\exp\left(\frac{1-2^{\ell z}}{\ell(2^\ell-1)}\right)dz, \alpha > 1.
\end{aligned}$$

Central Region

(iii) Central Regime: $n, k \to \infty$, $\xi = n2^{-k}$, ξ *fixed and* $0 < \xi < 2$. *Under the WKB hypothesis we have*

$$\begin{aligned} h_n^k &\sim A(\xi)e^{-n\Phi(\xi)} \\ A(\xi) &= e^{-\xi\Phi'(\xi)-\Phi(\xi)}\sqrt{1+2\xi\Phi'(\xi)+\xi^2\Phi''(\xi)}. \end{aligned}$$

By asymptotic matching, we find that the function $\Phi(\xi)$ *satisfies, for* $\xi \to 2^-$,

$$\begin{aligned} \Phi(\xi) &\sim \frac{1}{2}\log\left(\frac{e^2}{2C_0}\right) + \frac{(2-\xi)}{2}\log(2-\xi) + \frac{(\xi-2)}{4}\log(2C_0), \\ C_0 &= \prod_{\ell=1}^{\infty}(1-2^{-\ell-1})^{-2^{-\ell}} = 1.20675\ldots, \end{aligned}$$

and for $\xi \to 0^+$

$$\begin{aligned} \Phi(\xi) &\sim \frac{(\log 2)^{3/2}}{2^{9/8}\log^2\xi}\xi^{1/2-1/\log 2}\exp\left(-\frac{[\log\xi - \log(-\log_2\xi)]^2}{2\log 2}\right) \\ &\times \quad Q_*\left(\log_2\xi - \log_2(-\log_2\xi) + \frac{1}{2}\right). \end{aligned}$$

Here $Q_*(z)$ is a periodic function of period one (see Figure)

$$Q_*(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-s^2/2} Q\left(z + \frac{is}{\sqrt{\log 2}}\right) ds$$

where $Q(z)$ has the Fourier series (see Figure)

$$\begin{aligned} Q(z) &= 2^{13/12} \prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-1} \exp\left[-\frac{\pi^2}{3\log 2}\right. \\ &- \left.\sum_{m=1}^{\infty} \frac{\cos(2m\pi z)}{m \sinh(2\pi^2 m/\log 2)} \exp\left(-\frac{2\pi^2 m}{\log 2}\right)\right]. \end{aligned}$$

An alternate representation is $Q(z) = \frac{e^{-\pi i z}}{\sin(\pi z)} e^{p(z)}$ where

$$e^{p(z)} = e^{-(\log 2) z^2/2} e^{\pi i (z + \frac{3}{2})} 2^{1/8} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2}\right] \vartheta_1\left(\frac{i}{2} z \log 2\right)$$

and $\vartheta_1(u)$ is the Jacobi theta function defined by

$$\begin{aligned} \vartheta_1(u) &= \vartheta_1(u|\tau) = 2 \sum_{m=1}^{\infty} (-1)^{m+1} q^{(m-1/2)^2} \sin((2m-1)u) \\ &= 2q^{1/4} \sin(u) \prod_{m=1}^{\infty} (1 - 2\cos(2u) q^{2m} + q^{4m})(1 - q^{2m}) \end{aligned}$$

where $q = 1/\sqrt{2} = e^{\pi i \tau}$ with $\tau = (i \log 2)/(2\pi)$.

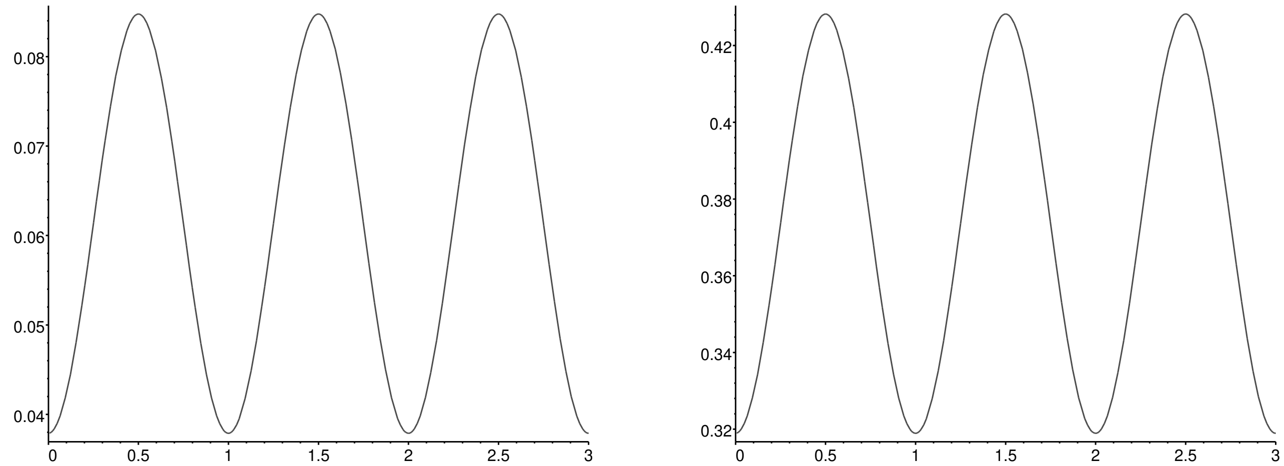
Left-Tail Regime

(iv) LEFT-TAIL REGIME: $n \to \infty$, $2^{k+1} - n = M = O(1)$, $M \geq 1$

$$h_n^k \sim \frac{2\sqrt{\pi}}{(M-1)!} \left(\frac{2C_0}{e^2}\right)^{2^k} 2^{(M-1/2)k}$$

where C_0 is defined above.

Graph of $Q(z)$ and $Q^*(z)$



(a) (b)

Figure 5: (a) A plot of $(Q(z) - 0.063716934676373540049220 0) \cdot 10^{24}$ over three periods; (b) A plot of $(Q_*(z) - 0.063716934676) \cdot 10^{12}$ over three periods.

These functions arise in the analysis of the following recurrence

$$\tilde{F}_j(n) = \sum_{m=0}^{j-2} (-1)^m \binom{n}{j-2-m} \prod_{N=1}^{m} (1-2^N)^{-1}.$$

DST Height - Discussion

1. Since $\Phi(\xi) > 0$ for $0 < \xi \le 2$ h_n^k is exponentially small for cases (iii) and (iv), while for cases (i) and (ii) $1 - h_n^k$ is exponentially small. For case (i) $1 - h_n^k$ is (roughly) $O(2^{-n^2/2})$, for case (ii) $1 - h_n^k = O(2^{-k^2/2})$, while for cases (iii) and (iv) $h_n^k = O(e^{-n\Phi(\xi)})$. For a fixed large n, most of the probability mass occurs in that range of k where h_n^k changes from $h_n^k \approx 0$ to $h_n^k \approx 1$. The mass must thus be concentrated in the asymptotic matching region between cases (ii) and (iii).

2. Let

$$k_\ell = \log_2 n + \sqrt{2\log_2 n} - \log_2(\sqrt{2\log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} - \beta(n) + \ell$$

where ℓ is an integer and

$$\beta(n) = \left\langle \log_2 n + \sqrt{2\log_2 n} - \log_2(\sqrt{2\log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \right\rangle .$$

Here $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x.

Using our previous estimates we obtain

$$h_n^k \approx \exp\left(\frac{-Q_*}{2\sqrt{\log 2}} e^{\beta-\ell-1/\log 2} 2^{\sqrt{2\log_2 n}(\beta-\ell)-(\beta-\ell)^2/2-\frac{3}{4}\log_2(2\log_2 n)} \right)$$

where $Q_* = Q_*(\log_2 n - \log_2(\sqrt{2\log_2 n}) + 1/2)$ and $\beta = \beta(n)$ is before.

The behavior of this distribution depends on the behavior of

$$\Delta(n) = \beta(n)\sqrt{2\log_2 n} - \frac{3}{2}\log_2(\sqrt{2\log_2 n})$$

as $n \to \infty$.

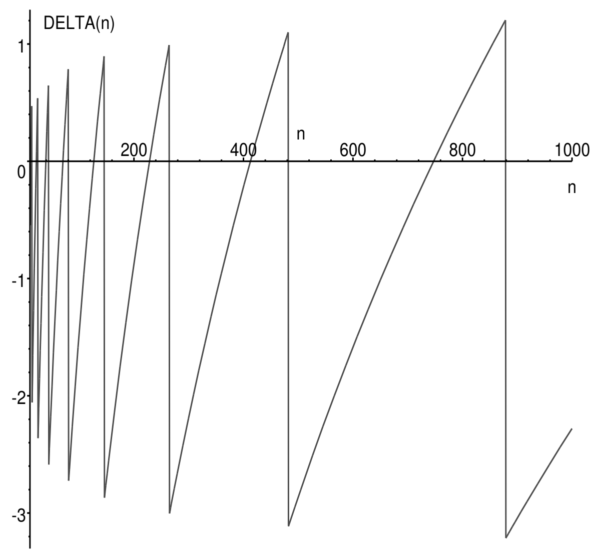


Figure 6: The function DELTA$(n) = \Delta(n) = \beta(n)\sqrt{2\log_2 n} - 1.5\log_2\sqrt{2\log_2(n)}$ versus n.

Concentration of the Asymptotic Distribution

(a) If $0 < \beta < 1$ and $\ell = 0$ then h_n^k is small as $n \to \infty$. If $0 < \beta < 1$ and $\ell = 1$, h_n^k is close to one. This shows that for $0 < \beta(n) < 1$ the mass becomes concentrated at a single point as $n \to \infty$.

(b) We now show that one can find special sequences $n(i)$ such that $n(i) \to \infty$ as $i \to \infty$ and along these sequences there is probability mass concentrated at two points, corresponding to $\ell = 0$ and $\ell = 1$.

We can easily find sequences $n(i)$ such that $\Delta(n(i))$ remains bounded, and in fact we can have $\Delta \to 0$ for some of these. If we consider $n(i)$ with $\Delta(n(i)) \to 0$, then as $n \to \infty$ for $\ell = 0$ h_n^k asymptotically becomes

$$\exp\left[-\frac{Q_*}{2\sqrt{\log 2}} \exp\left(-\frac{1}{\log 2}\right)\right].$$

This does not approach a limit as $n \to \infty$ due to the oscillatory behavior of $Q(z)$ and hence $Q_*(z)$.

Conclusion for DST

Corollary 4. *The asymptotic distribution of DST height is concentrated among the two points* $k_1 - 1$ *and* k_1*, that is,*

$$\Pr\{\mathcal{H}_n = k_1 - 1 \text{ or } k_1\} = 1 - o(1)$$

as $n \to \infty$*. More precisely: (i) there are subsequences* n_i *for which* $\Pr\{\mathcal{H}_{n_i} = k_1\} = 1 - o(1)$ *provided that*

$$\begin{aligned}\Delta(n_i) &= \sqrt{2\log_2 n_i}\left\langle \log_2 n_i + \sqrt{2\log_2 n_i} - \log_2(\sqrt{2\log_2 n_i}) + \frac{1}{\log 2} - \frac{1}{2}\right\rangle \\ &- \frac{3}{2}\log_2(\sqrt{2\log_2 n_i}) \to \infty\end{aligned}$$

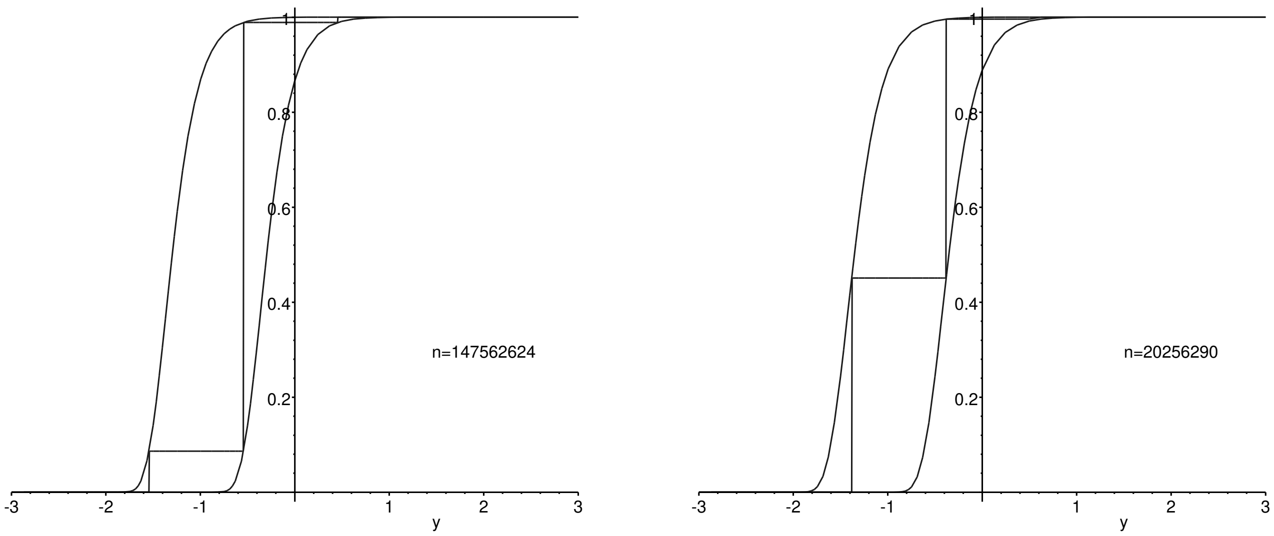
as $i \to \infty$*; (ii) there are subsequences* n_i *for which* $\Pr\{\mathcal{H}_{n_i} = k_1 - 1 \text{ or } k_1\} = 1 - o(1)$ *provided that* $\Delta(n_i) = O(1)$*.*

Corollary 5. *The average height is*

$$\mathbf{E}[\mathcal{H}_n] = \log_2 n + \sqrt{2\log_2 n} - \log_2(\sqrt{2\log_2 n}) + O(1)$$

as $n \to \infty$

DST Distribution - Figures



(a) (b)

Figure 7: Asymptotic distributions for the height of DST and their corresponding lower and upper bounds.

Numerical Results - PATRICIA Trie

Table 1: Right-Tail Comparison

n	$j = n - k$	$1 - h_n^k$ (Numerical)	$1 - h_n^k$ (Theorem 2(i))
10	2	1.782(10^{-7})	1.786(10^{-7})
	3	4.469(10^{-5})	4.571(10^{-5})
	4	3.040(10^{-3})	3.251(10^{-3})
	5 ($= n - k_1$)	7.071(10^{-2})	8.322(10^{-2})
20	2	2.684(10^{-39})	2.684(10^{-39})
	5	1.343(10^{-24})	1.343(10^{-24})
	10	3.463(10^{-8})	3.517(10^{-8})
	14 ($= n - k_1$)	0.2040	0.2676
30	2	5.176(10^{-99})	5.176(10^{-99})
	5	2.781(10^{-75})	2.781(10^{-75})
	10	8.200(10^{-44})	8.200(10^{-44})
	15	5.411(10^{-21})	5.417(10^{-21})
	20	2.050(10^{-6})	2.108(10^{-6})
	23 ($= n - k_1$)	0.1288	0.1616

Table 2: Left-Tail Comparison

n	k	M	h_n^k (Numerical)	h_n^k (Theorem 2(iii))
8	3	0	$5.062(10^{-2})$	$5.010(10^{-2})$
16	4	0	$5.032(10^{-4})$	$5.006(10^{-4})$
15	4	1	$4.734(10^{-3})$	$4.275(10^{-3})$
32	5	0	$3.544(10^{-8})$	$3.534(10^{-8})$
31	5	1	$6.667(10^{-7})$	$6.341(10^{-7})$
30	5	2	$6.197(10^{-6})$	$5.324(10^{-6})$
64	6	0	$1.248(10^{-16})$	$1.246(10^{-16})$
63	6	1	$4.694(10^{-15})$	$4.579(10^{-15})$
62	6	2	$8.780(10^{-14})$	$8.148(10^{-14})$
61	6	3	$1.089(10^{-12})$	$9.353(10^{-13})$

Table 3: Central Regime Comparison

n	k	ξ $(= n2^{-k})$	h_n^k (Numerical)	h_n^k $\xi \to 0$	h_n^k $\xi \to 0$
10	4	.63	.43975	.30024	.37067
	5	.31	.92929	.82862	.87673
	6	.16	.99696	.98542	.99145
20	5	.63	.12641	.09015	.11129
	6	.31	.79599	.68661	.72648
	7	.16	.98564	.97105	.97699
30	6	.47	.24414	.16249	.18627
	7	.23	.87115	.80820	.82855
	8	.12	.99283	.98760	.98948
50	7	.39	.23770	.15836	.17423
	8	.20	.87927	.83529	.84681
	9	.10	.99401	.99125	.99210
100	8	.39	.04845	.02508	.02759
	9	.20	.75483	.69772	.70733
	10	.10	.98623	.98259	.98342
150	9	.29	.15798	.10126	.10619
	10	.15	.87248	.84558	.84961
	11	.07	.99491	.99388	.99409
250	10	.24	.18706	.12842	.13209
	11	.12	.89950	.88226	.88426
	12	.06	.99668	.99618	.99627
500	11	.24	.03336	.01649	.01696
	12	.12	.80512	.77838	.78014
	13	.06	.99313	.99238	.99247