Discrete Mathematics and Theoretical Computer Science **6**, 2003, 91–100

Numerical Studies of the Asymptotic Height Distribution in Binary Search Trees

Charles Knessl[†]

Dept. of Mathematics, Statistics and Computer Science (M/C 249), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045, USA

received Sep 20, 2002, *revised Jun 14, 2003*, *accepted Jul 14, 2003*.

We study numerically a non-linear integral equation that arises in the study of binary search trees. If the tree is constructed from n elements, this integral equation describes the asymptotic (as $n \to \infty$) distribution of the height of the tree. The height is defined as the longest path in the tree. Our analysis supplements some asymptotic results we recently obtained (cf. Knessl and Szpankowski (2002)) for the tails of the distribution. The asymptotic height distribution is shown to be unimodal with highly asymmetric tails.

Keywords: binary search trees, height distribution

1 Introduction

A *binary search tree* is a fundamental data structure used in searching and sorting. It is defined as follows. There are n elements to be stored in the tree. A root node is created for the first element. Subsequent elements are directed to the left or right subtree according to whether they are less than or greater than the element in the root node. By this construction, the left and right subtrees are also binary search trees by themselves. Many classic sorting algorithms (such as QUICKSORT) can be conveniently represented by binary search trees (BST).

It is well known that the worst search time for this model is $O(n)$, but the average search time is only $O(\log n)$. We consider the average case performance and introduce the following probabilistic model. We take all $n!$ permutations of the n elements to be equally likely and analyze the height $\mathcal{H}_n$ of a BST constructed from n elements. The height is the longest path in the randomly built tree. Clearly $\mathcal{H}_n$ cannot exceed n and must exceed $\log_2 n$. In view of the probabilistic assumption $\mathcal{H}_n$ is a random variable and we set $L_n^k = \text{Prob}\{\mathcal{H}_n \le k\}$. The support lies in the range $k \le n < 2^k$.

There has been a lot of previous work on computing various aspects of this probability distribution, in the limit $n \to \infty$. Pittel (1984) showed that (almost surely) $\mathcal{H}_n \big/ \log n \to A_0$ as $n \to \infty$, with $A_0 \le A = 4.31107\ldots$. Devroye (1986) established that $E[\mathcal{H}_n] \sim A \log n$ as $n \to \infty$. This was refined to $E[\mathcal{H}_n] = A \log n + O(\log\log n)$ by Devroye and Reed (1995).

[†]This work was supported by NSF Grants DMS-99-71656 and DMS-02-02815, and NSA Grant MDA 904-03-1-0036.

In the past it had been conjectured that $E[\mathcal{H}_n] - A\log n \sim -\delta\frac{A}{A-1}\log\log n$ with $\delta = \frac{1}{2}$, but recent results of Reed (2000) (see also Knessl and Szpankowski (2002)) show that the correct value is $\delta = \frac{3}{2}$. There has also been some work on estimating the variance $var[\mathcal{H}_n]$. Experimental studies of Robson (1979) show that $E\big|\mathcal{H}_n - E[\mathcal{H}_n]\big|$ is bounded, suggesting that $var[\mathcal{H}_n] = O(1)$. This has been established rigorously by Drmota (1999, 2002) and by Reed (2000).

Very little seems to be known about the full distribution L_n^k. In Knessl and Szpankowski (2002) we used singular perturbation methods to analyze a recurrence relation satisfied by the distribution. Under some assumptions about the forms of various asymptotic expansions, we obtained expressions for L_n^k and $1-L_n^k$ for $n\to\infty$ and various ranges of k. In the range where most of the probability mass accumulates, we showed that L_n^k can be approximated by the solution of a non-linear integral equation. This was related to a functional equation studied by Drmota (1999). We established some asymptotic properties of the solution to this integral equation, but could not solve it exactly. Recently, Drmota (2003) established rigorously that the height distribution function satisfies this integral equation, in the limit $n\to\infty$. It was also shown that the equation has a unique solution that satisfies a certain auxiliary condition (cf. (4) and (7)).

In this note we supplement the results of Knessl and Szpankowski (2002) by numerically analyzing the integral equation. We thus obtain the shape of the asymptotic height distribution numerically. We state the problem more precisely in section 2, and the numerical results are discussed in section 3 and in the Figures and Tables therein.

2 Problem Statement

Let us denote by $\mathcal{H}_n$ the height of a binary search tree that stores n elements. Its probability distribution

$$L_n^k = \text{Prob}\{\mathcal{H}_n \le k\} \tag{1}$$

satisfies the non-linear recurrence

$$L_{n+1}^{k+1} = \frac{1}{n+1}\sum_{\ell=0}^{n} L_\ell^k L_{n-\ell}^k \tag{2}$$

subject to the initial condition $L_n^0 = \delta(n,0)$.

Setting

$$z = k - A\log n + \frac{3}{2}\frac{A}{A-1}\log\log n + c \tag{3}$$

and assuming that $L_n^k \sim f(z)$ we derived in Knessl and Szpankowski (2002) the following non-linear integral equation for $f(z)$:

$$\begin{aligned} f(z+1) &= \int_0^1 f(z-A\log x)f\big(z-A\log(1-x)\big)dx, \quad -\infty < z < \infty \\ f(-\infty) &= 0, \quad f(\infty) = 1. \end{aligned} \tag{4}$$

In (3) $A = 4.31107\ldots$ is the unique solution to $(x/2)^x = e^{x-1}$ in the range $x > 1$. We observe that if $f_0(z)$ is a solution to (4) then any translation (i.e., $f_0(z+C)$) is also a solution. Thus by retaining in the right side of (3) the arbitrary constant c, we can choose a convenient way to normalize the solution to (4) so as to make it unique. Recently, Drmota (2003) established rigorously that (4) has a unique solution, modulo the translation.

We let $f(z) = 1 - g(z)$ where clearly $g(z)$ will be small for $z \to \infty$. Then we can approximate, for z large, $g(z) \approx g_L(z)$ where g_L satisfies the linearized equation

$$g_L(z+1) = 2\int_0^1 g_L(z - A\log x)dx = 2\int_0^\infty g_L(z+At)e^{-t}dt. \tag{5}$$

Now (5) admits exponential solutions of the form $e^{-\nu z}$ provided that ν satisfies the characteristic equation

$$e^{-\nu} = \frac{2}{1+\nu A}. \tag{6}$$

We can easily show that $\nu = 1 - 1/A$ is a double root of (6) and that this is the only real solution for $\nu > 0$. There exist infinitely many complex solutions to (6) and these can be used to construct solutions to (4), e.g., by the method of successive iterations. However, the numerical and analytic studies in Knessl and Szpankowski (2002) show that these lead to solutions that are inappropriate (they typically oscillate and/or become negative). This again follows more rigorously from the work of Drmota (2003). Thus we write the general "acceptable" solution to (5) as

$$g_L(z) = \exp\left[-\left(1-\frac{1}{A}\right)z\right](\alpha z + \beta). \tag{7}$$

Knessl and Szpankowski (2002) also showed that if $\alpha = 0$ then the solution to the non-linear problem (4) becomes negative for $-z$ sufficiently large. Thus we have $\alpha > 0$ and in view of (3) normalize our solution by setting $\alpha = 1$.

Now we use (7) to construct a solution to the non-linear problem, with $f(z) = 1 - g(z)$, in the form

$$\begin{aligned} g(z) &= (z+\beta)e^{-az} + \sum_{m=2}^{\infty} e^{-maz}P_m(z) \\ &= \sum_{m=1}^{\infty} e^{-maz}P_m(z), \quad a = 1 - \frac{1}{A}, \end{aligned} \tag{8}$$

where $P_m(z)$ is a polynomial of degree m; we write

$$P_m(z) = \sum_{j=0}^{m} F(m,j)z^j. \tag{9}$$

Using (8) in (4) leads to

$$\begin{aligned} & e^{-ma}P_m(z+1) - 2\int_0^1 x^{maA}P_m(z - A\log x)dx \\ &= -\sum_{\ell=1}^{m-1}\int_0^1 x^{\ell aA}(1-x)^{(m-\ell)aA}P_\ell(z-A\log x)P_{m-\ell}(z-A\log(1-x))dx, \end{aligned} \tag{10}$$

for $m \geq 2$. Then by using (9) and comparing coefficients of z^M we are led to

$$\sum_{J=M}^{m}\binom{J}{M}F(m,J)\left\{e^{-m(1-1/A)} - \frac{2(J-M)!A^{J-M}}{[1+m(A-1)]^{J-M+1}}\right\} \tag{11}$$

$$= -[z^M] \sum_{\ell=1}^{m-1} \sum_{k_1=0}^{\ell} \sum_{k_2=0}^{m-\ell} z^{k_1+k_2} \sum_{i=k_1}^{\ell} \sum_{j=k_2}^{m-\ell} \binom{i}{k_1}\binom{j}{k_2}$$
$$\times D(\ell, m-\ell, i-k_1, j-k_2) F(\ell,i) F(m-\ell,j).$$

Here $[z^M]$ denotes the coefficient of z^M (we may replace $[z^M]z^{k_1+k_2}$ by $\delta(k_1+k_2, M)$) and

$$D(\alpha,\beta,\gamma,\delta) = A^{\gamma+\delta} \int_0^1 x^{\alpha(A-1)}(1-x)^{\beta(A-1)}(-\log x)^{\gamma}[-\log(1-x)]^{\delta} dx. \tag{12}$$

The above can be expressed in terms of derivatives of the Beta function. We also note that $F(1,1) = 1$ and $F(1,0) = \beta$.

Approximations to $g(z)$, and hence $f(z)$, may be obtained simply by truncating the sum in (8) at some large value $m = N$. However, this leads to problems for z negative and $-z$ sufficiently large, as discussed in section 3. We also note that given $P_1(z) = z+\beta$, each $F(m,j)$ is a polynomial in β in view of (11).

Finally we represent $f(z)$ in the contour integral form

$$f(z) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \exp\left(\frac{\eta}{A} e^{-z/A}\right) \mathcal{F}(\eta) d\eta, \tag{13}$$

which says that, after an appropriate variable change, $\mathcal{F}(\eta)$ is the (two-sided) Laplace transform of $f(z)$. Then from (4) it follows that

$$-\mathcal{F}'(\eta) = e^{-2/A}[\mathcal{F}(\eta e^{-1/A})]^2. \tag{14}$$

This is a functional-differential equation studied by Drmota (1999, 2002), who used the normalization condition $\mathcal{F}(0) = 1$, with which (14) has a unique analytic solution about $\eta = 0$, that is in fact an entire function (this is rigorously shown in Drmota (1999, 2002)). We note that our normalization (which took $\alpha = 1$ in (7)) is different from Drmota's; unfortunately it seems that neither can be used to infer the true value of c in (3). An important difference is that while (14) has a unique solution, our problem still has a one-parameter infinity of solutions, with β indexing the family. However we show numerically that only one value of β leads to a solution that can satisfy the condition $f(-\infty) = 0$ $(g(-\infty) = 1)$. The other solutions grow very rapidly as $z \to -\infty$ and apparently do not have Laplace transforms. Hence they are excluded from (14) by the form (13).

In Knessl and Szpankowski (2002) we also established that as $z \to -\infty$ $f(z)$ satisfies

$$f(z) \sim 2\sqrt{\frac{2\kappa}{\pi}} \frac{\sqrt{A\log 2}}{A\log 2 - 1} e^{-\omega z/2} \exp(-\kappa e^{-\omega z}) \tag{15}$$

where

$$\omega = \frac{\log 2}{A\log 2 - 1} = .3486294060\ldots$$

and κ is a constant, which is made unique by choosing $\alpha = 1$. Our derivation of (15) made certain assumptions about the asymptotic form; the numerical studies here provide more justification for this and also estimate the constant κ.

β	zeros of f_6	zeros of f_6'
-5	$-6.259, -5.004, -2.797$	$-5.975, -4.577, -2.060, 6.258$
-2	$-7.664, -6.697, -4.607$	$-7.466, -6.362, -4.021, 2.758$
-1.5	$-7.776, -7.548, -5.406$	$-7.689, -7.189, -4.888, 1.859$
-1.4	-5.681	$-7.656, -7.463, -5.218, 1.629$
-1.3	-5.995	$-5.615, 1.366$
-1	-6.713	$-6.410, .02408$
$-.97$	-6.765	$-6.464, -.2771$
$-.95$	-6.798	$-6.499, -.5591$
$-.92$	-6.847	$-6.549, -1.440$
$-.9112$	-6.861	$-6.563, -2.637$
$-.9111$	$-6.861, -2.808, -2.468$	$-6.563, -2.657$
$-.91$	$-6.863, -3.194, -2.072$	$-6.565, -2.846$
$-.9$	$-6.879, -3.898, -1.311$	$-6.581, -3.491$
$-.8$	$-7.027, -4.887, -.1593$	$-6.733, -4.475$
$-.7$	$-7.163, -5.231, .2318$	$-6.872, -4.823$
$-.5$	$-7.413, -5.673, .6764$	$-7.128, -5.276$
0	$-7.972, -6.448, 1.265$	$-7.700, -6.073$
1	$-8.985, -7.658, 1.878$	$-8.730, -7.302$
2	$-9.963, -8.737, 2.255$	$-9.718, -8.387$
5	$-12.92, -11.80, 2.940$	$-12.69, -11.45$

Tab. 1: The Zeros of f_6 and f_6' for various values of β.

3 Numerical Results

We define

$$f_N(z) = 1 - \sum_{m=1}^{N} e^{-amz} \left\{ \sum_{j=0}^{m} F(m,j) z^j \right\}, \quad N \geq 1 \tag{1}$$

with $g_N(z) = 1 - f_N(z)$. These correspond to approximate solutions to (4). The exact solution must also satisfy $f(-\infty) = 0$ and $f'(z) > 0$ for all z. Some of the problems arising in the convergence of $f_N(z)$ to $f(z)$ are illustrated by discussing f_N for a particular N, and we consider $N = 6$ in detail below.

A plot of $f_6(z) = f_6(z;\beta)$ for various β shows that typically both f_6 and f_6' have zeros and hence lead to unacceptable approximations to a probability distribution. Our goal is to define a criteria and choose an optimal value of β that somehow minimizes this "unacceptability". Then we shall increase N and obtain a sequence of optimal β that converges to the unique β for which we have $f(-\infty) = 0$ in (4).

In Table 1 we give the zeros of both f_6 and f_6' for various values of β. We note that for general N we have, as $z \to -\infty$, $f_N(z) \sim -e^{-aNz} F(N,N) z^N$. We can easily show that $F(m,m) > 0$ for all $m \geq 1$ and thus as $z \to -\infty$ $f_N(z) \to +\infty$ (resp. $-\infty$) for N odd (resp. N even). The data in Table 1 show that f_6 has exactly three zeros if $\beta \in (-\infty, \beta')$ or $\beta \in (\beta_*, \infty)$, and a single zero if $\beta \in (\beta', \beta_*)$. Here $\beta' \in (-1.5, -1.4)$ and $\beta_* \in (-.9112, -.9111)$. The derivative f_6' has four zeros if $\beta < \widehat{\beta}$ and two zeros if

N	β_{opt}	z_{min}	$\widetilde{z}_{\text{min}}$
6	$-.9111950$	-2.638	-2.991
7	$-.9117765$	-3.052	-3.398
8	$-.9119242$	-3.425	-3.763
9	$-.9119624$	-3.764	-4.097
10	$-.9119724$	-4.074	-4.403
11	$-.9119750693$	-4.360	-4.686
12	$-.9119757674$	-4.625	-4.950
13	$-.9119759527$	-4.873	-5.196
14	$-.9119760019$	-5.105	-5.427
15	$-.9119760150$	-5.324	-5.645

Tab. 2: β_{opt}, z_{min} and $\widetilde{z}_{\text{min}}$ for $N \leq 15$.

N	z_{max}	$h_{\text{max}} = h(z_{\text{max}})$
4	.3092	.1741
5	.2922	.1743
6	.2918	.1743
8	.2918	.1743
10	.2918	.1743

Tab. 3: Convergence near $z = z_{\text{max}}$.

$f_N(z)$	$z = -1$	$z = -2$	$z = -3$	$z = -4$
$N = 6$	.12954	.043850	.086020	4.4439
$N = 7$	.12953	.042793	.014072	.66115
$N = 8$	.12953	.042754	.0086837	.070850
$N = 9$	.12953	.042753	.0083977	.0062665
$N = 10$	.12953	.042753	.0083865	.0011049
$N = 11$	.12953	.042753	.0083862	.00079428
$N = 12$	.12953	.042753	.0083862	.00077984
$N = 13$	.12953	.042753	.0083862	.00077931
$N = 14$	.12953	.042753	.0083862	.00077929
$N = 15$	.12953	.042753	.0083862	.00077929

Tab. 4: Convergence of $f_N(z)$ for (fixed) negative values of z.

$\beta > \widehat{\beta}$, with $\widehat{\beta} \in (-1.4, -1.3)$. We define an optimal β as follows. For a given β we consider the minimum value of z such that $f_6(z)$ and $f_6'(z)$ are both positive for all z exceeding this value. More precisely we let $z_*^{(6)}(\beta) = \max\{z : f_6(z) = 0\}$ subject to the constraints that $f_6 > 0$ and $f_6' > 0$ for $z > z_*^{(6)}(\beta)$. Then β_{opt} is defined as the value of β that minimizes $z_*^{(6)}(\beta)$. Note that $z_*^{(6)}(\beta)$ may or may not exist for a particular β. When $N = 6$ Table 1 shows that it exists for all β exceeding about $-.9111$. We can define a more general $z_*^{(N)}(\beta)$ by setting it equal to the largest value of z where f_N' vanishes (if $z_*^{(N)}(\beta)$ fails to exist by the previous definition). In either case $f_N(z)$ can be an acceptable approximation to a probability distribution only for $z > z_*^{(N)}(\beta)$.

Our computational experience has shown that β_{opt} always corresponds to two roots of f_N coalescing into a double root (and thus a root of f_N'). When $N = 6$, $\beta_{\text{opt}} \approx -.9112$ and with this value f_6 has a double zero at $z_{\min} \approx -2.637$. Also, if we plot $f_6(z+1) - f_6(z)$, which would be an approximation to $\text{Prob}\{\mathcal{H}_n = k+1\}$, we find that it has a zero at $\widetilde{z}_{\min} \approx -2.9917$. Thus $f_6(z+1) - f_6(z)$ is an acceptable approximation to a probability density (or discrete distribution) only for $z > \widetilde{z}_{\min}$.

For arbitrary N we again compute β_{opt}, $z_{\min}$ and $\widetilde{z}_{\min}$. These are summarized in Table 2 for $N \leq 15$. The data suggest that β_{opt} converges rapidly to the value $-.9119760\ldots$. The sequences $z_{\min}^{(N)}$ and $\widetilde{z}_{\min}^{(N)}$ are converging to $-\infty$, but much more slowly, with the "gaps" $|z_{\min}^{(N+1)} - z_{\min}^{(N)}|$ decreasing with N. The sequence of functions $h_N(z) = f_N(z+1) - f_N(z)$ is converging to some unimodal positive function $h(z)$, with a maximum value of $h_{\max} = .1743\ldots$, which is attained at $z = z_{\max} = .2918\ldots$. The convergence near $z = z_{\max}$ is very rapid, as illustrated in Table 3. As z becomes negative, the convergence of $f_N(z)$ becomes much slower as $|z|$ increases. Also, there is a lot of cancelation in the sum in (1); for $N = 14$ and 15 we had to increase the precision to 20 digits in order to accurately do the calculation. In Table 4 we illustrate the convergence of $f_N(z)$ for (fixed) negative values of z. The value $N = 15$ is not sufficient to see convergence at $z = -5$. The minimum value of z that sees $f_{15}(z)$ settling to its limit is about $z = -4.3$. We find that $f(-4.1) = .00058157\ldots$, $f(-4.2) = .00042828\ldots$ and $f(-4.3) = .00031059\ldots$. We also see from Table 4 that once f_N settles to its limit value, it does so very quickly.

Next we test the asymptotic formula (15), which applies for $z \to -\infty$. The difficulties described above preclude us from computing $f(z)$ for large negative values. The constant κ in (15) could not be determined analytically. We can estimate it simply by numerically computing $f(z_0)$ for a certain negative z_0 (our results allow only $z_0 \geq -4.3$), comparing this to the right side of (15) (with $z = z_0$), and solving the resulting transcendental equation for κ. In Table 5 we estimate κ using various z. It would appear that

z	$f(z)$	κ
-1	.12953	2.0495
-2	.042753	2.0898
-3	.0083862	2.1100
-4	.00077929	2.1219
-4.1	.00058157	2.1236
-4.2	.00042828	2.1257
-4.3	.00031059	2.1287

Tab. 5: Estimating κ using various z.

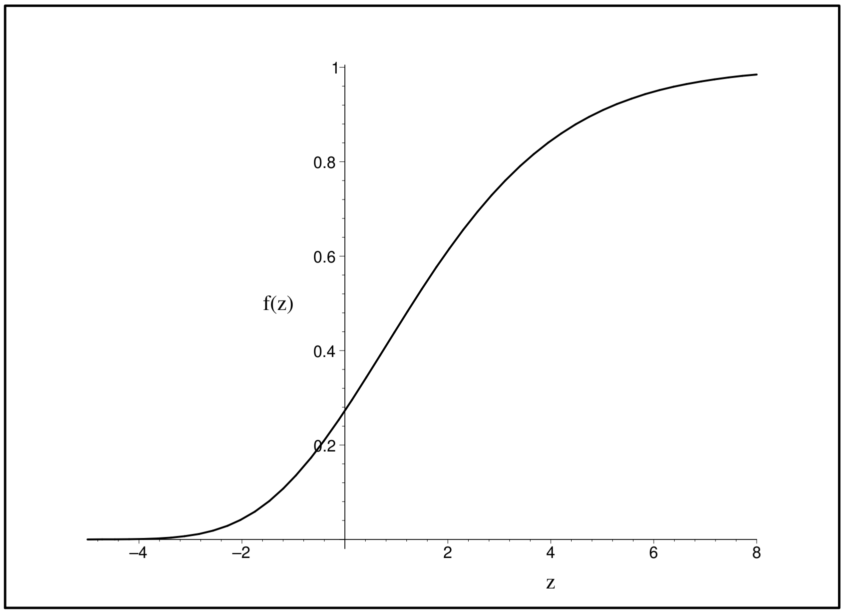

Fig. 1: The approximation to $f(z)$ using $N = 15$, for $z \in (-5, 8)$

$\kappa \approx 2.13$. It should be noted that the three values for $z < -4$ are more sensitive to error. Also, $f(z)$ should be very small due to the doubly-exponential convergence to zero as $z \to -\infty$. However, when $z = -4$, $-\omega z \approx 1.4$ which is not particularly large. A really accurate calculation of κ (and verification of (15)) would probably require that we calculate $f(z)$ accurately for values much more negative than -4.3.

In Figure 1 we plot $f_{15}(z)$ for $z \in [-5, 8]$ and in Figure 2 we plot $f_{15}(z+1) - f_{15}(z)$ for the same range. These are our final approximations to $f(z)$ and $f(z+1) - f(z)$. The second figure clearly illustrates the shape of the "density", showing its unimodal structure, the (roughly) exponential right tail and the very thin (roughly double exponential) left tail. These figures used $\beta = -.9119760150$. Note that we are approximating the discrete distribution (1) (or $L_n^{k+1} - L_n^k = \text{Prob}\{\mathcal{H}_n = k+1\}$) by the continuous function $h(z)$. For a given large n we can choose several values of k in (3) to make $z =$ "$O(1)$" and the corresponding values of L_n^k should lie close to the curve in Figure 1, for some appropriate c. The values of $L_n^{k+1} - L_n^k$ should then lie close to the curve in Figure 2 for this value of c. We have no analytic method for estimating the value of c. In Knessl and Szpankowski (2002) it was shown that if we had an asymptotic approximation to L_n^k valid on the scale $k, n \to \infty$ with $k/\log n$ fixed and $> A$, then we could use asymptotic matching to infer the value of c. However, we could not completely analyze this scale, which we refer to as the "near right tail" of the distribution. There $1 - L_n^k$ is algebraically small in n (for a fixed $k/\log n \in (A, \infty)$).

To summarize, we have presented an efficient numerical method for calculating the asymptotic height distribution in binary search trees. Our results yield the distribution's shape, but there is still the arbitrary translation arising from c in (3). Our results also suggest that the non-linear integral equation has, up to a translation, a unique solution that can represent a probability distribution. This was recently established

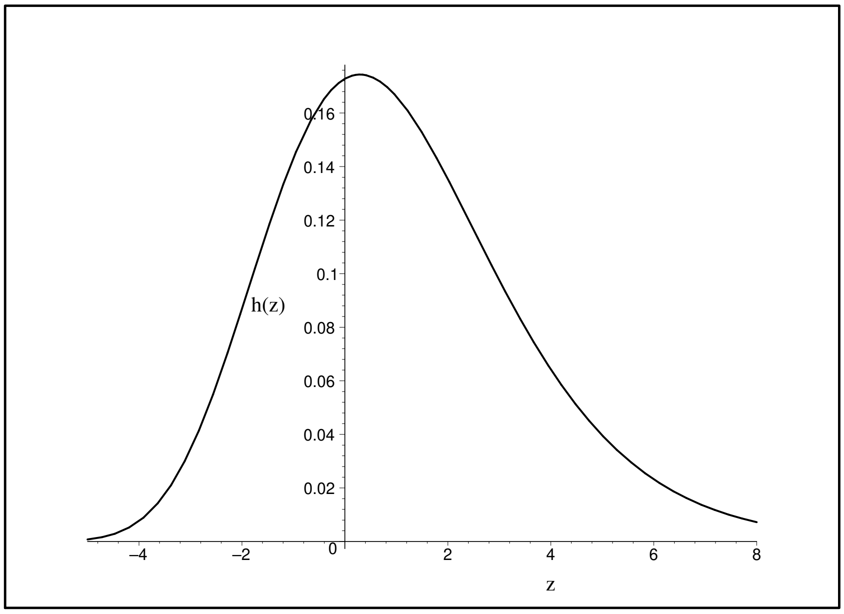

Fig. 2: The approximation to $h(z) = f(z+1) - f(z)$ using $N = 15$, for $z \in (-5, 8)$

rigorously by Drmota (2003). If $\alpha = 1$ in (7) this solution corresponds to $\beta = -.9119760\ldots$. There are still many issues regarding the convergence of L_n^k to $f(z)$ that need further work.

References

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

L. Devroye and B. Reed. On the variance of the height of random binary search trees. *SIAM J. Computing*, 24:1157–1162, 1995.

M. Drmota. An analytic approach to the height of binary search trees. *Algorithmica*, 29:89–119, 1999.

M. Drmota. The variance of the height of binary search trees. *Theoret. Comput. Sci.*, 270:913–919, 2002.

M. Drmota. An analytic approach to the height of binary search trees II. *Journal of the ACM*, 50:333–374, 2003.

C. Knessl and W. Szpankowski. The height of a binary search tree: the limiting distribution perspective. *Theoret. Comput. Sci.*, 289(1):649–703, 2002.

B. Pittel. On growing random binary trees. *J. Math. Anal. Appl.*, 103:461–480, 1984.

B. Reed. How tall is a tree. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 479–483, Portland, Oregon, May 21–23 2000.

J. Robson. The height of binary search trees. *Austral. Comput. J*, 11:151–153, 1979.