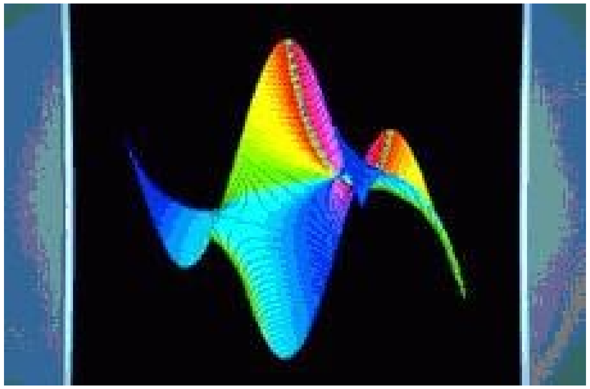
Binary Trees, Left and Right Paths, WKB Expansions, and Painlevé Transcendents

C. Knessl* and W. Szpankowski†

January 17, 2006



*Dept. Mathematics, Statistics & CS, University of Illinois at Chicago
†Department of Computer Science, Purdue University

Outline of the Talk

1. Formulation of Knuth's Problem
2. A Class of Functional Equations
3. Recurrence for Path Difference
4. WKB Method
5. Main Results
6. Painleve Transcendents
7. A New Class of Distributions?

Knuth's Problem

During the 10th Seminar on Analysis of Algorithms, MSRI, Berkeley, June 2004, Knuth posed the problem of analyzing the left and the right path length in a random binary trees.

In the standard model of binary trees generation, one selects uniformly a tree among all binary unlabeled ordered trees built on n nodes.

Let $\mathcal{T}_n$ denote a set of all binary trees built on n nodes. Then

$$|\mathcal{T}_n| = \binom{2n}{n}\frac{1}{n+1} = \text{Catalan number}.$$

A binary tree is chosen with probability $1/|\mathcal{T}_n|$.

Generating Functions

Let $N(p, q; n)$ be the number of binary trees with n nodes that have a total right path length p and a total left path length q.

The generating function of $N(p, q)$

$$G_n(w, v) = \sum_p \sum_q N(p, q; n) w^p v^q$$

satisfies the recurrence

$$G_{n+1}(w, v) = \sum_{i=0}^{n} w^i v^{n-i} G_i(w, v) G_{n-i}(w, v), \quad n \geq 0,$$

subject to the initial condition $G_0(w, v) = 1$.

Thus the triple transform

$$C(w, v, z) = \sum_{n=0}^{\infty} G_n(w, v) z^n = \sum_n \sum_p \sum_q N(p, q; n) z^n w^p v^q$$

satisfies

$$C(w, v, z) = 1 + zC(w, v, wz)C(w, v, vz).$$

This is Knuth's functional equation.

Catalan Numbers and Path Length

1. Set $w = v$ (and define $C((w,z) := C(w,w,z))$.

$$C(w,z) = 1 + zC^2(w,zw).$$

This describes the total path length.

Setting $w = 1$ we get

$$C(1,z) = 1 + zC^2(1,z)$$

which can be solved explicitly

$$C(1,z) = \left(1 - \sqrt{1-4z}\right)/(2z)$$

leading to the Catalan number $C_n = [z^n]C(1,z)$.

Solving for $C(w,z)$, we obtain the distribution of the total path length T_n.
Louchard and Takacs prove that

$$\Pr\{T_n/\sqrt{2n^3} \leq x\} \to W(x)$$

where $W(x)$ is the Airy distribution function defined by its moments.

Number of Trees with a Given Path Length

2. Setting $z = 1$ (and keeping $w = v$) in the previous equation we arrive at

$$C(w,1) = 1 + C^2(w,w)$$

which does not seem to be explicitly solvable.

Observe that $C(w,1) = \sum_t w^t \sum_n N(,t,n)$

$$[w^t]C(w,1) = \text{number of binary trees with path length=t}$$

During ANALCO'05 C. Knessl presented our paper where we proved that

$$[w^t]C(w,1) = |\mathcal{T}_t| = \frac{1}{(\log_2 t)\sqrt{\pi t}} 2^{\frac{2t}{\log_2 t}\left(1+c_1 \log^{-2/3} t+c_2 \log^{-1} t+O(\log^{-4/3} t)\right)}.$$

The leading term was also obtained by G. Seroussi.

Right Path Length, Area under Bernoulli Walk

3. Let us now set $v = 1$ in the triple transform equation. Then

$$C(w, 1, z) = 1 + zC(w, 1, wz)C(w, 1, z)$$

and $G_n(w, 1) = [z^n]C(w, 1, z)$ satisfies

$$G_{n+1}(w, 1) = \sum_{i=1}^{n} w^i G_i(w, 1) G_{n-i}(w, 1).$$

Observe that $G_n(w, 1)$ is the generating function of the right path length.

It was also studied by Takacs when he analyzed the area under a Bernoulli random walk.

Finally, it appears in the Kleitman-Winston conjecture.

Path Length Difference

Let

$$J = p - q$$

where, we recall,
p - the right path length,
q - the left path length.

Let $\mathcal{D}_n$ be a random variable representing the path difference. Then

$$P_{-}(J;n) = \mathrm{Prob}[\mathcal{D}_n = J] = \frac{1}{C_n} \sum_{i=0}^{\binom{n}{2}-|J|} N(i, i+|J|; n).$$

In terms of the generating function $G_n(w, v)$, we have

$$P_{-}(J;n) = \frac{1}{C_n}[w^J]G_n\left(w, \frac{1}{w}\right).$$

where, we recall

$$G_{n+1}(w,v) = \sum_{i=0}^{n} w^i v^{n-i} G_i(w,v) G_{n-i}(w,v).$$

The WKB Method

The WKB method was named after the physicists Wentzel, Kramers and Brillouin.

It assumes that a recurrence, functional equation or differential equation has an asymptotic solution, say $G(\xi; n)$, in the following form

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

where $\phi(\xi)$ and $A(\xi)$, $A^{(1)}(\xi)$, . . . are unknown functions.

These functions must be determined from the equation itself, often using another tool known as the asymptotic matching principle

Here is what Fedoryuk has to say about such approximations:
. . . It is necessary first of all to guess (and no other word will do) in what form to search for the asymptotic form. Of course, this stage – guessing the form of the asymptotic form – is not subject to any formalization. Analogy, experiments, numerical simulation, physical considerations, intuitions, random guesswork; these are the arsenal of means used by any research worker."

Scaling for J

It turns out that the most interesting behavior of J is around $O(n^{5/4})$ so we scale

$$J = \beta n^{5/4} = O(n^{5/4}).$$

and define for fixed β

$$P_-(J;n) \sim n^{-5/4} p_-(\beta)$$

To find the probability density function $p_-(\beta)$ we compute the double-sided Laplace transform

$$1 + \sqrt{\pi}\bar{H}(b) = \int_{-\infty}^{\infty} e^{\beta b} p_-(\beta) d\beta.$$

Then

$$\begin{aligned} p_-(\beta) &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{-\beta b}[1 + \sqrt{\pi}\bar{H}(b)]db \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\beta i x}[1 + \sqrt{\pi}\bar{H}(b)]dx \\ &= \frac{1}{\pi}\int_{0}^{\infty} \cos(\beta x) S(x) dx, \end{aligned}$$

where $S(x) = 1 + \sqrt{\pi}\bar{H}(ix)$.

Asymptotic Tails

We do not have an explicit formula for $p_-(\beta)$, but using the WKB method we find the following asymptotics as $\beta \to \infty$

$$p_-(\beta) \sim \sqrt{\frac{5}{6}}(5\beta)^{1/3} c_0 \exp\left(-\frac{3}{4}5^{1/3}\beta^{4/3}\right)\left[1 + O(\beta^{-4/3})\right]$$

with

$$c_0 \approx .5513288;$$

and

$$p_-(0) \approx .45727 \; ; \; p_-''(0) \approx -.71462.$$

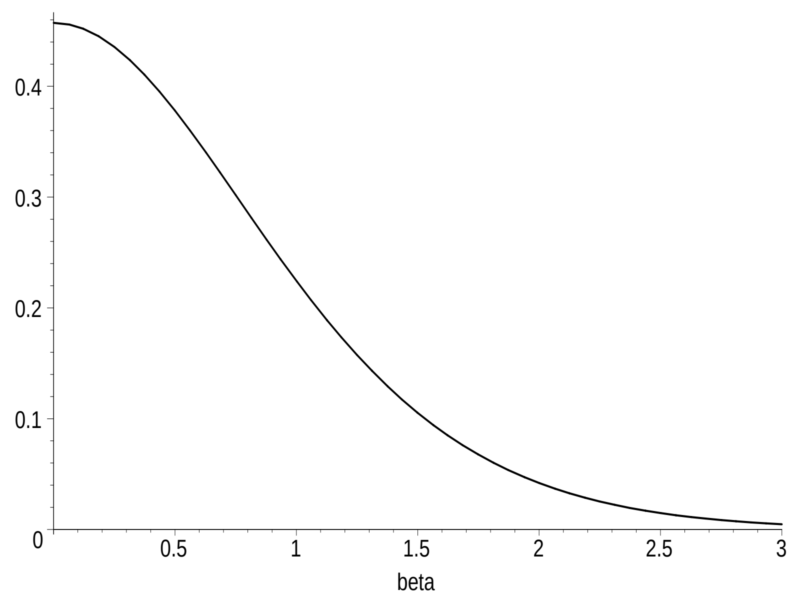


Figure 1: The density $p_-(\beta)$ for $\beta \in [0, 3]$.

Moments

Let $\bar{H}(b) = \sum_{m=0}^{\infty} b^{2m+2} \Delta_m$ and by setting

$$\Delta_m = \frac{1}{\Gamma\left(\frac{5}{2}m + 2\right)} \tilde{\Delta}_m$$

we find that $\tilde{\Delta}_m$ satisfies the nonlinear recurrence

$$\tilde{\Delta}_{m+1} = \frac{(5m+6)(5m+4)}{8} \tilde{\Delta}_m + \frac{1}{4} \sum_{\ell=0}^{m} \tilde{\Delta}_\ell \tilde{\Delta}_{m-\ell}, \quad m \geq 0$$

with $\Delta_0 = \tilde{\Delta}_0 = \frac{1}{4}$.

We thus observe that the even moments of the difference converge as follows

$$\frac{\mathbf{E}[\mathcal{D}_n^{2m+2}]}{n^{5(m+1)/2}} \to (2m+2)!\sqrt{\pi}\Delta_m.$$

These results agree with Janson (2006).

Using discrete WKB method we find

$$\Delta_m = C \frac{e^{m/2}}{\sqrt{m}} m^{-m/2} 10^{-m/2} \left[1 - \frac{4}{15m} + O(m^{-2})\right], \ m \to \infty.$$

Some Relationship

We now infer the behavior of $\bar{H}(b)$ for purely imaginary values of b. Define

$$\bar{H}(ix) = -y^{3/2}\Lambda(y) = -x^{6/5}\Lambda(x^{4/5}), \quad y = x^{4/5}$$

we find that $\Lambda(y)$ satisfies

$$0 = \int_0^y \Lambda(\xi)\Lambda(y-\xi)d\xi + 2y^2\Lambda(y) - 2\frac{\sqrt{y}}{\sqrt{\pi}} + \frac{4}{\sqrt{\pi}}\int_0^y \frac{\Delta'(\xi)}{\sqrt{y-\xi}}d\xi.$$

Painlav'e Transcendent

Let

$$U(\phi) = \int_0^\infty e^{-y\phi}\Lambda(y)dy$$

be the Laplace of $\Lambda(y)$. Then we have

$$0 = 2U''(\phi) + U^2(\phi) + 4\sqrt{\phi}U(\phi) - \phi^{-3/2}.$$

Setting

$$U(\phi) = -2\sqrt{\phi} + U_1(\phi)$$

we obtain

$$0 = U_1^2(\phi) + 2U_1''(\phi) - 4\phi,$$

with

$$U_1(\phi) = 2\sqrt{\phi} + \frac{1}{4}\phi^{-2}[1 + o(1)], \quad \text{for} \quad \phi \to \infty.$$

The second order nonlinear ODE as above is called the first Painlevè transcendent. This classic problem has been studied for over 100 years, and modern applications in nonlinear waves and random matrices have been found in recent years.

Some Properties

It is well known that each singularity of $U_1(\phi)$ is a double pole, and the Laurent expansion near any singularity at $\phi = -\nu$ has the form

$$U_1(\phi) = \frac{-12}{(\phi+\nu)^2} + O(1),\ \phi \to -\nu.$$

Let us denote by ν_* the singularity with the largest real part. Then

$$\Lambda(y) - \frac{1}{\sqrt{\pi}} y^{-3/2} \sim -12 y e^{-\nu_* y},\ y \to \infty.$$

This allows us to find asymptotic expression for the tail of $P_-(J)$ as presented above.

A New Class of Distributions?

We conjecture that the distribution of $\mathcal{D}_n$ can be only characterized by moments. In fact, we proved that

$$\frac{\mathbf{E}[\mathcal{D}_n^{2m+2}]}{n^{5(m+1)/2}} \to (2m+2)!\sqrt{\pi}\Delta_m.$$

where Δ_m satisfy the following nonlinear recurrence

$$\tilde{\Delta}_{m+1} = \frac{(5m+6)(5m+4)}{8}\tilde{\Delta}_m + \frac{1}{4}\sum_{\ell=0}^{m}\tilde{\Delta}_\ell\tilde{\Delta}_{m-\ell}, \quad m \geq 0.$$

There is a large class of important problems (e.g., quicksort, linear hashing, enumeration of trees in $\mathcal{T}_t$) in which limiting distribution cannot be explicitly found, but it can be expressed in terms of moments that satisfy a nonlinear recurrence like above.

Open Problem

Let Z be a (normalized) limiting distribution of a process for which we know that for some $a_m \to \infty$ we have

$$\frac{\mathbf{E}[Z^m]}{a_m} = c_m$$

such that in general c_m satisfies

$$c_{m+1} = \alpha_m + \beta_m c_m + \gamma_m \sum_{i=0}^{m} c_i c_{m-i}$$

with some initial conditions, and given α_m, β_m and γ_m.

Similar recurrences appear in the quicksort, linear hashing, path length in binary trees, area under Bernoulli walk, enumeration of trees with given path length, and many others.

Can we built a theory to characterize this class of distributions ?