Basic hypergeometric series, digital search trees, and approximate counting

Helmut Prodinger
Technische Universitat Wien

October 16, 1995

[summary by Michèle Soria]

The transformation formula of Heine from the theory of basic hypergeometric functions allows very simple and pleasant derivations of explicit forms of the level polynomials of digital search trees [8], as well as of explicit forms of the probabilities in the "approximate counting" problem [7].

1. Basics about hypergeometric functions

This section contains basic notations and results about q-hypergeometric series (see e.g. [1, 3]).

q-Pochhammer symbol. Let us introduce the classical notations:

$$(a)_n = (1-a)(1-aq)\cdots(1-aq^{n-1}), \qquad (a)_0 = 1, \qquad (a)_\infty = \lim_{n\to\infty}(a)_n$$

and observe that

$$(a)_n = \frac{(a)_\infty}{(aq^n)_\infty} \tag{1}$$

Cauchy's Formula.

$$\sum_{n\geq 0}\frac{(a)_n t^n}{(q)_n} = \frac{(at)_\infty}{(t)_\infty}$$

Euler's identities. The special case $a = 0$ is generally attributed to Euler:

$$\sum_{n\geq 0}\frac{t^n}{(q)_n} = \frac{1}{(t)_\infty}$$

and the so called Euler formula is obtained by first substituting a/b by a and bt by b in Cauchy's formula, then setting $a = -1$ and $b = 0$:

$$(-t)_\infty = \sum_{n\geq 0}\frac{q^{\binom{n}{2}}t^n}{(q)_n}.$$

Heine's transformation. Cauchy's formula and equation (1) lead to Heine's formula:

$$\sum_{n\geq 0}\frac{(a)_n(b)_n}{(q)_n(c)_n}t^n = \frac{(at)_\infty(b)_\infty}{(c)_\infty(t)_\infty}\sum_{n\geq 0}\frac{(c/b)_n(t)_n}{(q)_n(at)_n}b^n$$

Setting $a=q$, $b=y$, $c=0$ and $t=z$ in Heine's transformation, one gets the simple formula

$$\sum_{n\geq 0}(y)_n z^n = (y)_\infty\sum_{n\geq 0}\frac{y^n}{(q)_n(1-zq^n)}$$

2. Level polynomials in digital search trees

A digital search tree is constructed like a binary search tree, but the decision to go down to the left or right is done accordingly to the binary representation of the key: if the first bit is 0, the item goes left and otherwise it goes right; then the second bit is used to go down further left or right, etc., until there is an empty node where the item can be stored. In order to study the average search cost, we are interested in $h_{n,k}$, the expected number of nodes on level k (by convention, the root is at level 0), in a tree built from n random data (i.e. in every decision, a bit 0 or 1 is equally likely).

The level polynomial $H_n(u)=\sum_{k\geq 0}h_{n,k}u^k$ satisfies (see e.g. [5]) $H_0(u)=0$, and for $n\geq 1$

$$H_n(u) = \sum_{k=1}^{n}\binom{n}{k}(-1)^{k-1}(u)_{k-1},$$

By probabilistic arguments, Louchard [6] gave an explicit formula for the coefficients of $H_n(u)$, that we shall derive here by means of hypergeometric functions. We introduce the bivariate generating function $H(u,x)=\sum_n H_n(u)x^n$ and obtain easily:

$$H(u,x) = \frac{x}{(1-x)^2}\sum_{k\geq 0}(u)_k\frac{x^k}{(x-1)^k}.$$

The use of Heine's formula gives

$$H(u,x) = \frac{x}{(1-x)^2}(u)_\infty\sum_{k\geq 0}\frac{u^k}{(q)_k\left(1-\frac{x}{x-1}q^k\right)}.$$

Then decomposing into partial fractions and applying Euler's formula leads to

$$H(u,x) = \frac{(u)_\infty}{1-x}\frac{1}{(u/q)_\infty} - (u)_\infty\sum_{k\geq 0}\frac{(u/q)^k}{(q)_k}\frac{1}{1-x(1-q^k)}.$$

From this expression we get

$$H_n(u)\equiv[x^n]H(u,x) = \frac{1}{1-u/q} - (u)_\infty\sum_{k\geq 0}\frac{(u/q)^k}{(q)_k}(1-q^k)^n,$$

The coefficient of u^l in $H_n(u)$ then transforms by Euler's formula in

$$h_{n,k}\equiv[u^l]H_n(u) = q^{-l} - \sum_{k=0}^{l}\frac{q^{-k}}{(q)_k}(1-q^k)^n(-1)^{l-k}\frac{q^{\binom{l-k}{2}}}{(q)_{l-k}}. \tag{2}$$

Other parameters of interest, such as partial sums ($[u^l]H_n(u)/(1-u)$) or leaf levels ($[u^l]1-(1-u/q)H_n(u)$) can be obtained immediately from (2).

3. Approximate counting via Euler transform

Approximate counting can be described by an automaton with states 1, 2,... Starting in state 1, we proceed step by step. In one step we may either advance from state i to state $i+1$ with probability q^i, or stay in state i with probability $1-q^i$. The interesting parameter is the state reached after n random steps. The original analysis of this problem was done by Flajolet [2] and consists of an enumerative part and an asymptotic part. We will show here how hypergeometric functions allow some shortcuts in the enumerative part. Let $p_{n,l}$ be the probability to be in state l after n random steps, and let $H_l(x) = \sum_{n\geq 0} p_{n,l} x^n$. Using a decomposition path from 1 to l into stages, it is not hard to see that

$$H_l(x) = \frac{x^{l-1} q^{\binom{l}{2}}}{\prod_{i=1}^{l}(1-x(1-q^i))} = \frac{\frac{1}{x}\left(\frac{x}{1-x}\right)^l q^{\binom{l}{2}}}{\left(\frac{xq}{x-1}\right)_l}.$$

We shall go to the expected value after n steps by means of the bivariate generating function $H(x,y) = \sum_{l\geq 0} H_l(x) y^l$. Setting $z = \frac{x}{x-1}$ and applying Heine's formula, we get

$$H(x,y) = \frac{1}{x} \frac{(q)_\infty (yz)_\infty}{(qz)_\infty} \sum_{n\geq 0} \frac{(z)_n q^n}{(q)_n (yz)_n}.$$

One more Heine transform, with $a=0$, $b=z$, $c=yz$ and $t=q$ leads to

$$H(x,y) = \frac{1}{x} \frac{(q)_\infty (yz)_\infty}{(qz)_\infty} \frac{(z)_\infty}{(q)_\infty (yz)_\infty} \sum_{n\geq 0} \frac{(y)_n (q)_n z^n}{(q)_n} = \frac{1}{x}(1-z) \sum_{n\geq 0} (y)_n z^n.$$

The expected value after n steps, $\sum_l l p_{n,l}$, is the coefficient of x^n in the partial derivative $H_y(x,y)$ taken at $y=1$. Since for $n \geq 1$

$$\frac{\partial}{\partial y}(y)_n \Big|_{y=1} = -(q)_{n-1},$$

we have

$$\sum_{l\geq 1} l H_l(x) = -\frac{1}{x}(1-z) \sum_{n\geq 1} (q)_{n-1} z^n.$$

And to get the quantity of interest we have to extract the coefficient of z^n in the last expression. This is done by using Euler's transform: if $f(x) = \sum_{n\geq 0} a_n x^n$ then

$$\frac{1}{1-x} f\left(\frac{x}{x-1}\right) = \sum_{n\geq 0} \sum_{k=0}^{n} \binom{n}{k} (-1)^k a_k x^n.$$

Thus

$$\sum_{l\geq 1} l p_{n,l} \equiv [x^n] \sum_{l\geq 1} l H_l(x) = 1 - \sum_{k=1}^{n} \binom{n}{k} (-1)^k q^k (q)_{k-1}.$$

This formula is equivalent to the one given in [4], where its asymptotic value is then obtained by Rice's Method.

Bibliography

[1] Andrews (George E.). – *The Theory of Partitions.* – Addison-Wesley, 1976, *Encyclopedia of Mathematics and its Applications*, vol. 2.

[2] Flajolet (P.). – Approximate counting: A detailed analysis. *BIT*, vol. 25, 1985, pp. 113–134.

[3] Gasper (George) and Rahman (Mizan). – *Basic Hypergeometric Series.* – Cambridge University Press, 1990, *Encyclopedia of Mathematics and its Applications*, vol. 35.

[4] Kirschenhofer (P.) and Prodinger (H.). – Approximate counting: An alternative approach. *RAIRO Informatique Théorique et Applications*, vol. 25, 1991, pp. 43–48.

[5] Knuth (Donald E.). – *The Art of Computer Programming.* – Addison-Wesley, 1973, vol. 3: Sorting and Searching.

[6] Louchard (G.). – Exact and asymptotic distributions in digital and binary search trees. *RAIRO Theoretical Informatics and Applications*, vol. 21, n° 4, 1987, pp. 479–495.

[7] Prodinger (H.). – Approximate counting via Euler transform. *Mathematica Slovaca*, vol. 44, n° 5, 1994, pp. 569–574.

[8] Prodinger (H.). – Digital search trees and basic hypergeometric functions. *EATCS Bulletin*, vol. 56, 1995, pp. 112–115.