Generating functions in the analysis of m-versions of approximate counting, binary search trees and other structures

Helmut Prodinger

Stellenbosch

June 10, 2013

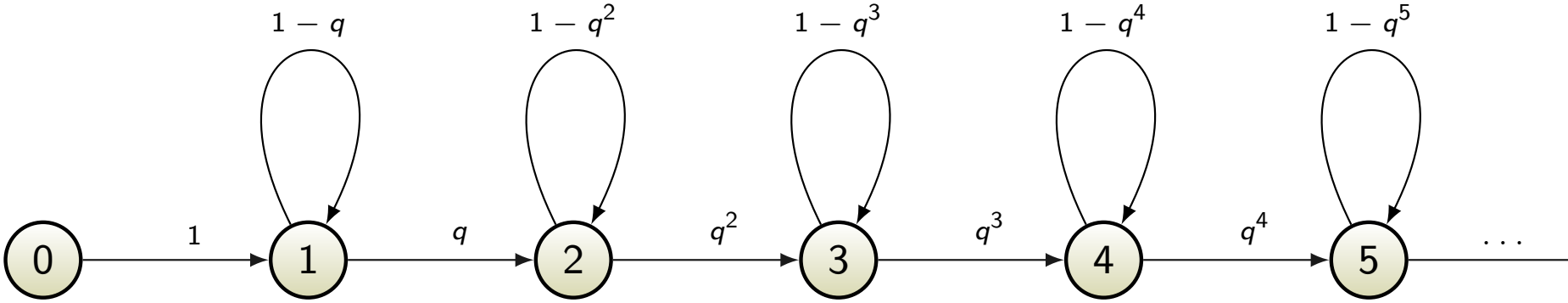

Figure: State diagram of the evolution of the counter in *approximate counting*. Normally, the (deterministic) initial step from state 0 to state 1 is not shown.

In approximate counting, one is interested in the state one is in after n random steps. This state number is interpreted as the value of a certain counter. This model was first analyzed by Flajolet and remains popular to this day. Flajolet derived explicit expressions to reach state k after n random steps, and computed expectation and variance of this parameter.

Theorem

The average and the variance of Cichon's approximate counter scheme admit the following asymptotic expansions as $N \to \infty$:

$$E_N \sim m\Big[\log_2 N - \log_2 m + \frac{1}{2} - \alpha + \frac{\gamma}{L} + \delta(\log_2 N)\Big],$$

$$V_N \sim m\Big[1 - \alpha - \beta + \frac{2\tau}{L} + \delta_V(\log_2 N)\Big].$$

$$Q_n := (1 - \frac{1}{2}) \dots (1 - \frac{1}{2^n})$$

,

$$Q(x) := \prod_{k \geq 1} (1 - \frac{x}{2^k})$$

$$Q_n = Q(1)/Q(2^{-n})$$

$$P_{N,I} = \sum_{n_1+\cdots+n_m=N} \frac{\binom{N}{n_1,\ldots,n_m}}{m^N} \sum_{I_1+\cdots+I_m=I} p_{n_1,I_1} \cdots p_{n_m,I_m}.$$

$$G_n(u) := \sum_{I\geq 0} p_{n,I} u^I.$$

$$G_n'(1) = 1 - \sum_{k=1}^{n} \binom{n}{k} (-1)^k 2^{-k} Q_{k-1},$$

$$G_n''(1) = \sum_{k=1}^{n} \binom{n}{k} (-1)^k 2^{1-k} Q_{k-1}(T_{k-1} - 1),$$

with

$$T_k = \sum_{j=1}^{k} \frac{1}{2^k - 1}.$$

Theorem

The first and second factorial moments have the following explicit expressions:

$$E_N = m - m\sum_{k=1}^{N}\binom{N}{k}(-1)^k(2m)^{-k}Q_{k-1},$$

$$\begin{aligned} E_N^{(2)} &= m(m-1) - 2m^2\sum_{j=1}^{N}\binom{N}{j}(-1)^j(2m)^{-j}Q_{j-1} \\ &\quad + m(m-1)\sum_{k=1}^{N}\binom{N}{k}(-1)^k(2m)^{-k}\sum_{j=1}^{k-1}\binom{k}{j}Q_{j-1}Q_{k-1-j} \\ &\quad + 2m\sum_{k=1}^{N}\binom{N}{k}(-1)^k(2m)^{-k}Q_{k-1}T_{k-1}. \end{aligned}$$

$$\psi(N) := \sum_{j=1}^{N-1} \binom{N}{j} Q_{j-1} Q_{N-j-1}$$

$$= 2Q_\infty^2 \sum_{s,h\geq 0} \frac{2^{-sN}}{Q_s Q_{s+h}} \left((1+2^{-h})^N - 1 - 2^{-hN} \right)$$

$$- (2^N - 2) Q_\infty^2 \sum_{s\geq 0} \frac{1}{Q_s^2} 2^{-Ns}$$

not good enough!

$$\psi(z) = 2\sum_{n\geq 0} a_{n+1} Q_{z+n-1} \sum_{\lambda\geq 1} \binom{z}{\lambda} \frac{1}{2^{\lambda+n}-1}$$
$$- 2\sum_{n\geq 0} a_{n+1} Q_{z+n-1} \frac{1}{1-2^{-z-n}} + 2^z \sum_{n\geq 0} a_{n+1} Q_{z+n-1}.$$

with

$$a_{n+1} = (-1)^n 2^{-\binom{n+1}{2}} / Q_n,$$

Approach by Fuchs (inspired by Hwang)

$$\mathcal{P} := \frac{\log 2}{L} - \alpha - \beta + \frac{2}{L}\tau \quad \text{with} \quad \tau := \sum_{k\geq 1} \frac{(-1)^{k-1}}{k(2^k-1)}$$

$$\alpha = \sum_{j\geq 1} \frac{1}{2^j-1} \quad \text{and} \quad \beta = \sum_{j\geq 1} \frac{1}{(2^j-1)^2}.$$

$$\mathcal{F} := \frac{(q)_\infty}{L} \sum_{j,l,h\geq 0} \frac{(-1)^j q^{\binom{j+1}{2}+l+h}}{(q)_j (q)_l (q)_h} \frac{\log(q^{h+j}+q^{l+j})}{q^{h+j}+q^{l+j}-1}.$$

Insertion cost in digital search trees (m-version)

$$H'_N(1) = \sum_{k=2}^{N} \binom{N}{k} (-1)^k Q_{k-2} m^{1-k},$$

$$H''_N(1) = 2 \sum_{k=2}^{N} \binom{N}{k} (-1)^{k-1} Q_{k-2} T_{k-2} m^{1-k}$$

Binary search trees
unsuccessful search

$$g_n(z) = \frac{1}{n!} 2z(2z+1)\dots(2z+n-2),$$

probability that k comparisons are needed, is

$$P_{n,k} = \frac{1}{n!}[z^k]2z(2z+1)\dots(2z+n-2).$$

From this

$$E_n = g_n'(1) = \sum_{j=0}^{n-2} \frac{2}{j+2} = 2\sum_{j=2}^{n} \frac{2}{j} = 2(H_n - 1)$$

and

$$E_n^{(2)} = 8 \sum_{2\le i<j\le n} \frac{1}{ij} = 8 \sum_{1\le i<j\le n} \frac{1}{ij} - 8 \sum_{1<j\le n} \frac{1}{j} = 4H_n^2 - 4H_n^{(2)} - 8H_n + 8.$$

$$\mathcal{G}_n(z) = m^{1-N} \sum_{n=1}^{N} \binom{N-1}{n-1} (m-1)^{N-n} g_n(z).$$

For the moments, we need three lemmas:

Lemma

$$S_n = \sum_{k=1}^{n} \binom{n}{k} x^k \frac{1}{k} = \sum_{k=1}^{n} \frac{(1+x)^k}{k} - H_n.$$

$$T_n = \sum_{k=1}^{n} \binom{n}{k} x^k H_k = H_n(1+x)^n - (1+x)^n \sum_{k=1}^{n} \frac{1}{k(1+x)^k}.$$

$$\begin{aligned} U_n &= \sum_{k=1}^{n} \binom{n}{k} x^k \sum_{1 \le i < j \le k} \frac{1}{ij} \\ &= (1+x)^n \frac{H_n^2 - H_n^{(2)}}{2} - (1+x)^n \sum_{k=1}^{n} \frac{1}{k(1+x)^k} H_{n-k} \\ &\quad + (1+x)^n \sum_{k=1}^{n} \frac{1}{k(1+x)^k} H_k - (1+x)^n \sum_{k=1}^{n} \frac{1}{k^2(1+x)^k}. \end{aligned}$$

$$\mathcal{E}_N \sim 2H_{N-1} - 2\log m + \frac{2m}{N} - 2;$$

$$\mathcal{E}_N^{(2)} + 4\mathcal{E}_N \sim 4(H_{N-1}^2 - H_{N-1}^{(2)}) - 8\sum_{k=1}^{N-1} \frac{1}{k}\Big(1 - \frac{1}{m}\Big)^k H_{N-1-k}$$
$$+ 8C_1(m) - 8C_2(m)$$
$$+ \frac{8m}{N}H_N - \frac{8m}{N}\log m - \frac{8m}{N}\sum_{k=0}^{N-1} \frac{1}{N-k}\Big(1 - \frac{1}{m}\Big)^k.$$

with

$$C_1(m) = \sum_{k\geq 1} \frac{1}{k}\Big(1 - \frac{1}{m}\Big)^k H_k \quad \text{and} \quad C_2(m) = \sum_{k\geq 1} \frac{1}{k^2}\Big(1 - \frac{1}{m}\Big)^k.$$

Successful search in binary search trees. The model is that the comparisons to find all possible nodes are *added*, and this count is than divided by the total number of nodes. This parameter has the following probability generating function:

$$R_n(z) = \frac{z(2z)(2z+1)\dots(2z+n-1)}{nn!(2z-1)} - \frac{z}{n(2z-1)}.$$

It translates into the m-model as follows:

$$\mathcal{R}_N(z) = m^{-N}\frac{1}{N}\sum_{n=1}^{N}\binom{N}{n}(m-1)^{N-n}nR_n(z);$$

classical:

$$nR_n'(1) = nE_n = 2(n+1)H_n - 3n,$$
$$nR_n''(1) = nE_n^{(2)} = 4(n+1)(H_n^2 - H_n^{(2)}) - 12nH_n - 4H_n + 16n.$$

$$\mathcal{E}_N = 2H_{N-1} - 2\sum_{k=1}^{N-1} \frac{1}{k}\left(1-\frac{1}{m}\right)^k + \frac{2m}{N} - \frac{2m}{N}\left(1-\frac{1}{m}\right)^N - 3$$
$$+ \frac{2m}{N}H_N - \frac{2m}{N}\sum_{k=1}^{N} \frac{1}{k}\left(1-\frac{1}{m}\right)^k.$$

Further,

$$\begin{aligned}
\mathcal{E}_N^{(2)} &= 4(H_{N-1}^2 - H_{N-1}^{(2)}) + \frac{4m}{N}(H_N^2 - H_N^{(2)}) \\
&\quad - 8\sum_{k=1}^{N-1}\frac{1}{k}\Big(1-\frac{1}{m}\Big)^k H_{N-1-k} + 8\sum_{k=1}^{N-1}\frac{1}{k}\Big(1-\frac{1}{m}\Big)^k H_k - 8\sum_{k=1}^{N-1}\frac{1}{k^2}\Big(1- \\
&\quad + \frac{4m}{N}H_N - \frac{4m}{N}\sum_{k=1}^{N}\frac{1}{k}\Big(1-\frac{1}{m}\Big)^k \\
&\quad - \frac{8m}{N}\sum_{k=1}^{N-1}\frac{1}{N-k}\Big(1-\frac{1}{m}\Big)^k - \frac{8m}{N^2} + \frac{8m}{N}H_N\Big(1-\frac{1}{m}\Big)^N \\
&\quad - \frac{8m}{N}\sum_{k=1}^{N}\frac{1}{k}\Big(1-\frac{1}{m}\Big)^k H_{N-k} + \frac{8m}{N}\sum_{k=1}^{N}\frac{1}{k}\Big(1-\frac{1}{m}\Big)^k H_k - \frac{8m}{N}\sum_{k=1}^{N}\frac{1}{k^2} \\
&\quad - 12H_{N-1} + 12\sum_{k=1}^{N-1}\frac{1}{k}\Big(1-\frac{1}{m}\Big)^k - \frac{12m}{N} + \frac{12m}{N}\Big(1-\frac{1}{m}\Big)^N + 16.
\end{aligned}$$

Theorem

The expectation and variance of the number of comparisons in a successful search related to m binary search trees of altogether N nodes, are given by

$$\mathcal{E}_N = 2\log\frac{N}{m} - 3 + 2\gamma + O\Big(\frac{1}{N}\Big),$$

$$\mathcal{V}_N = 2\log\frac{N}{m} - 4\log^2 m + 4 - \frac{2}{3}\pi^2$$

$$+ 2\gamma + 8C_1(m) - 8C_2(m) + O\Big(\frac{1}{N}\Big).$$

(internal) path length: sum of the distances of all the nodes to the root. In the m-version, it is simply the sum of the path lengths in the m individual trees. It is known that the probability generating functions satisfy

$$g_n(z) = \frac{z^{n-1}}{n} \sum_{k=1}^{n} g_{k-1}(z) g_{n-k}(z), \quad g_0(z) = 1,$$

whence

$$\mathcal{G}_N(z) = m^{-N} \sum_{n=0}^{N} \binom{N}{n} (m-1)^{N-n} g_n(z).$$

It is known that

$$g_n'(1) = 2(n+1)H_n - 4n,$$

$$g_n''(1) = 4(n+1)^2(H_n^2 - H_n^{(2)}) - 4(n+1)(4n+1)H_n + n(23n+17).$$

Theorem

The expectation and variance of the internal path length of m binary search trees of altogether N nodes, are given by

$$\mathcal{E}_N = \frac{N}{m}\left[2\log\frac{N}{m} + 2\gamma - \frac{4}{m}\right] + O(\log N),$$

$$\mathcal{V}_N = \frac{N^2}{m^2}\left[7 - \frac{2}{3}\pi^2 + 8C_1(m) - 8C_2(m) - 4\log^2 m\right] + O(N\log N).$$

Depth and path length of m-Plane Oriented Recursive Trees

The expectation of the depth of a random node in a PORT of size n is given by

$$\mathbb{E}(D_n) = \left(1 - \frac{1}{2n}\right)\widehat{H}_n - \frac{1}{2};$$

the variance is

$$\mathbb{V}(D_n) = \left(1 - \frac{1}{2n}\right)\left[\widehat{H}_n - \widehat{H}_n^{(2)}\right] - \mathbb{E}(D_n)^2,$$

with

$$\widehat{H}_n = \sum_{k=1}^{n} \frac{1}{2k-1} \quad \text{and} \quad \widehat{H}_n^{(2)} = \sum_{k=1}^{n} \frac{1}{(2k-1)^2}.$$

These quantities can be expressed in terms of traditional harmonic numbers,

$$H_n = \sum_{k=1}^{n} \frac{1}{k} \quad \text{and} \quad H_n^{(2)} = \sum_{k=1}^{n} \frac{1}{k^2},$$

but it is useful to have a special notation here.

The expectation of the path length of a PORT of size n is given by

$$\mathbb{E}(P_n) = \left(n - \frac{1}{2}\right)\widehat{H}_n - \frac{n}{2};$$

the variance is

$$\mathbb{V}(P_n) = n^2\left(\frac{3}{2} - \widehat{H}_n^{(2)}\right) + n\left(\widehat{H}_n^{(2)} - \widehat{H}_n - \frac{3}{4}\right) + \frac{1}{2}\widehat{H}_n - \frac{1}{4}\widehat{H}_n^{(2)}.$$

$f(z)$	$a_n := [z^n]f(z)/C_n$	$\sum_{n\geq 1} a_n z^n$
$\sqrt{1-4z}$	-2	$\dfrac{-2z}{1-z}$
$\dfrac{1}{\sqrt{1-4z}}$	$2(2n-1)$	$\dfrac{2z(1+z)}{(1-z)^2}$
$\dfrac{1}{(1-4z)^{3/2}}$	$2(2n+1)(2n-1)$	$\dfrac{2z(3+6z-z^2)}{(1-z)^3}$
$\dfrac{1}{(1-4z)^{1/2}}\log\dfrac{1}{1-4z}$	$4(2n-1)\widehat{H}_n$	$\dfrac{4z^{3/2}}{(1-z)^2}\log\dfrac{1}{1-z}$ $+\dfrac{8z^{3/2}}{(1-z)^2}\log(1+\sqrt{z}\,)+\dfrac{4z}{(1-z)^2}$
$\dfrac{1}{(1-4z)^{1/2}}\log^2\dfrac{1}{1-4z}$	$8(2n-1)(\widehat{H}_n^2-\widehat{H}_n^{(2)})$	$\dfrac{4z^{3/2}}{(1-z)^2}\log^2\dfrac{1}{1-z}$ $+\dfrac{8(1+2\log(2))z^{3/2}}{(1-z)^2}\log\dfrac{1}{1-z}$ $+\dfrac{16z^{3/2}}{(1-z)^2}\mathrm{Li}_2\Big(\dfrac{1-\sqrt{z}}{2}\Big)$ $-\dfrac{8z^{3/2}}{(1-z)^2}\log^2(1+\sqrt{z}\,)$ $+\dfrac{16(1+\log(2))z^{3/2}}{(1-z)^2}\log(1+\sqrt{z}\,)$ $+\dfrac{8\log^2(2)z^{3/2}}{(1-z)^2}-\dfrac{4\pi^2 z^{3/2}}{3(1-z)^2}$
$\dfrac{1}{(1-4z)^{3/2}}\log\dfrac{1}{1-4z}$	$4(2n+1)(2n-1)$ $\times(\widehat{H}_{n+1}-1)$	$\dfrac{16z^{3/2}}{(1-z)^3}\log\dfrac{1}{1-z}$ $+\dfrac{32z^{3/2}\log(1+\sqrt{z}\,)}{(1-z)^3}$ $-\dfrac{12z^2}{(1-z)^3}+\dfrac{4z}{(1-z)^3}$

$\frac{1}{(1-4z)^{1/2}} \log^2 \frac{1}{1-4z}$	$8(2n-1)(\widehat{H}_n^2 - \widehat{H}_n^{(2)})$	$\frac{4z^{3/2}}{(1-z)^2} \log^2 \frac{1}{1-z} + \frac{8(1+2\log(2))z^{3/2}}{(1-z)^2} \log \frac{1}{1-z} + \frac{16z^{3/2}}{(1-z)^2} \mathrm{Li}_2\Big(\frac{1-\sqrt{z}}{2}\Big) - \frac{8z^{3/2}}{(1-z)^2} \log^2(1+\sqrt{z}\,) + \frac{16(1+\log(2))z^{3/2}}{(1-z)^2} \log(1+\sqrt{z}\,) + \frac{8\log^2(2)z^{3/2}}{(1-z)^2} - \frac{4\pi^2 z^{3/2}}{3(1-z)^2}$
$\frac{1}{(1-4z)^{3/2}} \log \frac{1}{1-4z}$	$4(2n+1)(2n-1)$ $\times(\widehat{H}_{n+1} - 1)$	$\frac{16z^{3/2}}{(1-z)^3} \log \frac{1}{1-z} + \frac{32z^{3/2}\log(1+\sqrt{z}\,)}{(1-z)^3} - \frac{12z^2}{(1-z)^3} + \frac{4z}{(1-z)^3}$
$\frac{1}{(1-4z)^{3/2}} \log^2 \frac{1}{1-4z}$	$8(2n-1)(2n+1)$ $\times[(\widehat{H}_{n+1} - 1)^2$ $-(\widehat{H}_{n+1}^{(2)} - 1)]$	$\frac{16z^{3/2}}{(1-z)^3} \log^2 \frac{1}{1-z} - \frac{16(1-4\log(2))z^{3/2}}{(1-z)^3} \log \frac{1}{1-z} + \frac{64z^{3/2}}{(1-z)^3} \mathrm{Li}_2\Big(\frac{1-\sqrt{z}}{2}\Big) - \frac{32z^{3/2}}{(1-z)^3} \log^2(1+\sqrt{z}\,) - \frac{32(1-2\log(2))z^{3/2}}{(1-z)^3} \log(1+\sqrt{z}\,) - \frac{16\pi^2 z^{3/2}}{3(1-z)^3} + \frac{32\log^2(2)z^{3/2}}{(1-z)^3} + \frac{48z^2}{(1-z)^3}$

$$F(z) := \sum_{n \geq 1} L_n'(1) z^n$$

is known. Then

$$\mathcal{L}_N'(1) = [z^N] \frac{m^2}{m-(m-1)z} F\Big(\frac{z}{m-(m-1)z}\Big),$$

This is often called an Euler transform
A catalogue was created, and singularity analysis was used.

Theorem

Expectation and variance of the depth of a random node in m-PORTs of size N admit the following asymptotic expansions:

$$\mathbb{E}(\mathcal{D}_N) = \frac{1}{2}\log\frac{N}{m} + \frac{1}{2}\gamma + \log 2 - \frac{1}{2} + O\Big(\frac{\log N}{N}\Big),$$

$$\mathbb{V}(\mathcal{D}_N) = \frac{1}{2}\log\frac{N}{m} + \frac{\gamma}{2} - \frac{\pi^2}{8} - \frac{1}{4} + \log 2 + O\Big(\frac{\log N}{N}\Big).$$

Theorem

Expectation and variance of the path length in m-PORTs of size N admit the following asymptotic expansions:

$$\mathbb{E}(\mathcal{P}_N) = \frac{N}{2m}\left(\log\frac{N}{m} + \frac{1}{2}\gamma + \log 2 - 1\right) + O(\log N),$$

$$\mathbb{V}(\mathcal{P}_N) = \frac{N^2}{m^2}\left(\frac{3}{2} - \frac{\pi^2}{8}\right) + O(N \log N).$$

WORDS

$$E_n = \sum_{k=1}^{n} \binom{n}{k} (-1)^k \omega(k).$$

Therefore

$$\mathcal{E}_N = \sum_{k=1}^{N} \binom{N}{k} (-1)^k \omega(k) m^{1-k}.$$

$$E_n^{(2)} = \sum_{k=1}^{n} \binom{n}{k} (-1)^k \theta(k).$$

$$\begin{aligned} \mathcal{E}_N^{(2)} &= m \sum_{k=1}^{N} \binom{N}{k} (-1)^k \theta(k) m^{-k} \\ &\quad + (m-1)m \sum_{k=1}^{N} \binom{N}{k} (-1)^k m^{-k} \sum_{j=1}^{k-1} \binom{k}{j} \omega(j)\omega(k-j). \end{aligned}$$

Maximum
Left-to-right maximum
(cumulative)
were investigated in this way.