On approximate counting

Philippe Flajolet

INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.: 954 90 20

Rapports de Recherche

N° 153

ON APPROXIMATE COUNTING

Philippe FLAJOLET

Juillet 1982

ON APPROXIMATE COUNTING

Philippe FLAJOLET
INRIA
78150 Rocquencourt (France)

Abstract :

We analyze an algorithm of Todd, Martin, Langdon and Helman which can maintain an approximate counter up to m, using only about $\log_2 \log_2 m$ bits. The algorithm has a probabilistic nature ; we establish its soundness and prove that it has good convergence properties for a wide range of values. From our results follows that the algorithm will typically fall within a factor of 2 of the exact count. The algorithm is of potential use in applications related to data compression.

Résumé :

Nous présentons l'analyse d'un algorithme dû à Todd, Martin, Langdon et Helman qui permet de maintenir un compteur approximatif jusqu'à m en utilisant seulement $\log_2 \log_2 m$ bits environ. L'algorithme est de nature probabiliste ; nous en établissons la validité en montrant qu'il présente de bonnes propriétés de convergence pour un large domaine de valeurs. Il s'ensuit des méthodes développées dans cet article que l'algorithme fournit un résultat approché typiquement dans un rapport de 1 à 2 du résultat exact. L'algorithme possède des applications potentielles en compression de données.

INTRODUCTION

As shown by an easy information-theoretic argument, maintaining a counter whose values may range in the interval 1 to m essentially necessitates $\log_2 m$ bits. This lower bound is of course achieved by a standard binary counter. Todd et al [TO81] have proposed a probabilistic algorithm that maintains an approximate count using only about $\log_2 \log_2 m$ bits. This paper is devoted to a detailed analysis of their algorithm : we provide precise estimates on the probabilities of errors, from which the soundness of the algorithm can be assessed.

The analytic techniques used involue manipulation of generating functions related to a discrete time birth-process to which the algorithm is equivalent, the use of the Mellin integral transform and finally some simple identities that properly belong to the theory of integer partitions.

The algorithm itself was proposed for applications to data compression [TO81] when building an efficient variable length code to represent "non-random" binary data (see also [Ga78], [LaRi81]); there typically a large number of counters need to be maintained to gather statistics on the data to be compressed but high accuracy of each counter is not a critical factor in the design of almost-optimal codes.

1. - APPROXIMATE COUNTING

If the requirement of accuracy is dropped, a counter of value n can be replaced by another counter C containing $\lfloor \log_2 n \rfloor$ which only requires storing about $\log_2 \log_2 n$ bits. However since the fractional part of $\log_2 n$ is no longer preserved, there now arises the problem of deciding when to update the logarithmic counter in the course of successive incrementations. The idea of [To81] is to base this decision on probabilistic choices.

Approximate counting starts with counter C initialized to 1. After n increments, we expect C to contain a good approximation to $\lfloor \log_2 n \rfloor$; we should thus increase C by 1 after another n increments approximately. Since the exact value of n has not been kept and only C is known, the algorithm has to base its decision on the content of C alone. Approximate counting then treats the incrementation by the following procedure.

```
procedure increment (C : integer) ;
    Let DELTA (C) be a random variable which takes value 1 with probability 2^-C and value 0 with probability 1-2^-C ;
    C : = C + DELTA (C)
```

Our analysis will show that despite the number of random choices involved, the algorithm does not lose the count. More precisely we have :

Theorem 1 :

After n successive increments, the counter of approximate counting has average value

$$\overline{C}_n = \log_2 n + \frac{\gamma}{\ell n 2} - \lambda + \frac{1}{2} + \omega (\log_2 n) + 0 (\frac{1}{n^{0.98}})$$

where

and $\lambda = \sum_{n \geq 1} \frac{1}{2^n - 1} = 1.6067...$,

$\gamma = 0.577721...$

is the euler constant, and ω is a periodic function of mean value 0 and amplitude less than 10^{-5}.

The constant after $\log_2 n$ gives the asymptotic bias of the algorithm, and its numerical value is -0.27395... ; furthermore, calculations developped hereafter show that the actual bias for finite n varies very little with n : For n = 10, 100, 20000, the values of $\overline{C}_n - \log_2 n$ are respectively +0.0453, -0.2383, -0.2737.

Another interesting feature of the algorithm is the low dispersion of the results it produces. We can prove.

Theorem 2 :

After n successive increments, the standard deviation of the contents of the counter satisfies

$$\sigma_n = \sigma_\infty + \pi\ (\log_2 n) + \sigma(1)$$

where $\sigma_\infty = 0.8736..$ is a constant and π is a periodic function of mean value 0 and small amplitude.

In particular corresponding to n = 10, 100, 20000, we have $\sigma_n = 0.7776$, 0.8618, 0.8734. Thus typically the algorithm will estimate $\log_2 n$ with an error less than 1.

Finally, the methods developped here also permit evaluation of the probabilities of error. The distribution of values of approximate counting after n increments appear to be fairly narrowly centered around the average (better than merely predicted from the variance analysis using the Tchebycheff inequalities), and for instance in the case of n = 1024 (no that $\log_2 n = 10$) the following probabilities for the counter values, determined by proposition 1 below, are :

counter value	7	8	9	10	11	12	13
probability	0.0011	0.0602	0.3424	0.4218	0.1538	0.0195	0.0001

Thus the estimate on $\log_2 n$ will be off the exact result by more than 1 unit in only 8 % of the cases.

2. - BASIC PROBABILITIES

The possible evolutions of the algorithm can be seen as an evergrowing tree : we start from the counter set to 1 ; from this two situations can result : either the counter keeps its value 1 (this has probability $\frac{1}{2}$) or it is increased to 2 (with probability $\frac{1}{2}$) ; each of these possible stages has it self two possible outcomes. The corresponding tree of possibilities is given in figure 1, with edges labelled with the probabilities of corresponding transitions. From it, we see for instance that when n = 3, the probabilities of observing counter values 1,2,3,4 are respectively $\frac{8}{64}$, $\frac{38}{64}$, $\frac{17}{64}$, $\frac{1}{64}$.

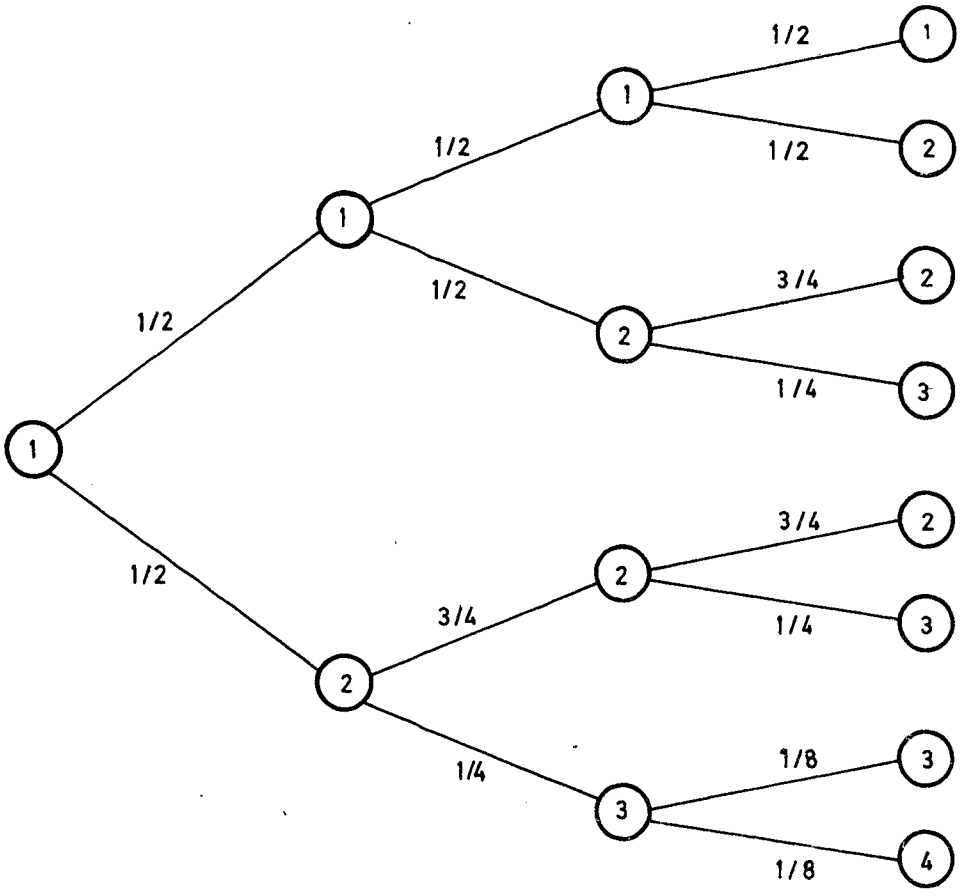
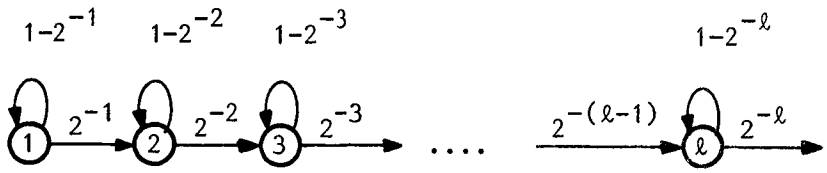


Figure 1 : The possible evolutions of approximate counting for n = 1,2,3

Another way of viewing the evolutions is by drawing a state diagram :



This is to be interpreted as follows : at state $\ell = 1,2,3....$, (i.e when the counter contains value ℓ), one increment causes the transition to state $\ell+1$ with probability $2^{-\ell}$, and the transition to state ℓ with probability $1-2^{-\ell}$. This is formally a discrete time pure birth process [K175].

Let $p_{n,\ell}$ be the probability that the counter contains value ℓ after n applications of the stochastic increment procedure. To compute $p_{n,\ell}$, observe that the probability of reaching state ℓ through n_1 transitions from state 1 to state 1, n_2 transitions from state 2 to state 2... n_ℓ transitions from state ℓ to state ℓ is :

$$(1-2^{-1})^{n_1}\, 2^{-1}\, (1-2^{-2})^{n_2}\, 2^{-2} \ldots\ldots\, (1-2^{-(\ell-1)})^{n_{\ell-1}}\, 2^{-(\ell-1)}\, (1-2^{-\ell})^{n_\ell},$$

with the condition that :

$$n_1 + n_2 + \ldots + n_\ell + \ell - 1 = n.$$

Summing over all possible intermediary transitions, we thus find

$$P_{n,\ell} = 2^{-\binom{\ell}{2}} \sum_{\substack{n_1+n_2+\ldots+n_\ell \\ +\ell-1=n}} (1-2^{-1})^{n_1}\, (1-2^{-2})^{n_2} \ldots (1-2^{-\ell})^{n_\ell} \qquad (1)$$

where as usual $\binom{\ell}{2} = \frac{\ell(\ell-1)}{2}$.

If we introduce the corresponding generating functions (for each ℓ) :

$$H_\ell(x) = \sum_{n \geq 0} p_{n,\ell}\, x^n, \qquad (2)$$

we observe that (1) expresses the $p_{n,\ell}$ as the coefficients of a Cauchy product of simpler functions, so that :

$$H_\ell(x) = \frac{2^{-\binom{\ell}{2}}\, x^{\ell-1}}{(1-\alpha_1 x)\,(1-\alpha_2 x) \ldots (1-\alpha_\ell x)} \quad \text{with } \alpha_j = 1-2^{-j} \qquad (3)$$

We obtain an expression different from (1), and indeed simpler to estimate numerically, by decomposing H_ℓ into partial fractions. Since we expect the asymptotically dominant contributions in the $p_{n,\ell}$ to come from the dominant poles, we set :

$$H_\ell(x) = \sum_{j=0}^{\ell-1} \frac{C_j}{1-x\alpha_{\ell-j}}, \qquad (4)$$

and start evaluating $C_0, C_1, \ldots$

We have

$$C_j = \lim_{x \to \alpha_{\ell-j}} H_\ell(x)\ (1-\alpha_{\ell-j}x),$$

so that :

$$C_0 = \frac{2^{-\binom{\ell}{2}}\ (1-2^{-\ell})^{-(\ell-1)}}{(1-\frac{\alpha_1}{\alpha_\ell})\ (1-\frac{\alpha_2}{\alpha_\ell})\ \ldots\ (1-\frac{\alpha_{\ell-1}}{\alpha_\ell})}$$

which after simplification using $(\alpha_j - \alpha_\ell) = 2^{-j}\ (1-2^{-\ell+j})$ gives :

$$C_0 = \frac{1}{(1-\frac{1}{2})\ (1-\frac{1}{4})\ \ldots\ (1-\frac{1}{2^{\ell-1}})}$$

similarly, we find

$$C_1 = \frac{-1}{(1-\frac{1}{2})\ (1-\frac{1}{4})\ \ldots\ (1-\frac{1}{2^{\ell-2}})} \cdot \frac{1}{(1-\frac{1}{2})},$$

and in general

$$C_j = \frac{(-1)^j\ 2^{-\binom{j}{2}}}{Q_j\ Q_{\ell-1-j}} \tag{5}$$

where for all k :

$$Q_k = \prod_{i=1}^{k} (1-\frac{1}{2^i}). \tag{6}$$

Now from (4) immediately follows an expression for the coefficients $p_{n,\ell}$ of $H_\ell(x)$:

$$p_{n,\ell} = \sum_{j=0}^{\ell-1} C_j\ \alpha_{\ell-j}^n,$$

whence with (5), (6) :

Proposition 1 :

The probability $p_{n,\ell}$ of having counter value ℓ after n increments is

$$p_{n,\ell} = \sum_{j=0}^{\ell-1} \frac{(-1)^j \, 2^{-\binom{j}{2}}}{\prod_{i=1}^{j} (1 - \frac{1}{2^i}) \prod_{i=1}^{\ell-1-j} (1 - \frac{1}{2^i})} (1-2^{-(\ell-j)})^n . \qquad (7)$$

This expression permits an easy numerical calculation of the probabilities involved in approximate counting.

3. - CONTINUING WITH APPROXIMATIONS

The expression of Proposition 1 is not as bad as it looks. Firstly the product

$$Q = \prod_{i=1}^{\infty} (1 - \frac{1}{2^i}) \qquad (8)$$

is convergent and simple comparisons with the geometric series show that

$$|Q - Q_k| = O\left(\frac{1}{2^k}\right)$$

with Q_k defined in (6). In particular the Q_k are always in the interval defined by Q = 0.28878... and 1, and the denominators in (7) are bounded.

Secondly the very fast decrease of coefficients $2^{-\binom{j}{2}}$ shows that numerically the significant contribution comes from small values of the index j, and asymptotically only values of j less than $O(\sqrt{\log_2 n})$ need to be considered.

Lastly the exponential approximation

$$(1-a)^n \approx e^{-an}$$

is usually justified in this class of problems (see e-g [Kn 73, p 131]).

We first prove that for ℓ small enough compared to the probabilities $p_{n,\ell}$ are small.

Proposition 2 :

For $\ell < \log_2 n - 2 \log_2 \log n$, the probabilities $p_{n,\ell}$ satisfy

$$p_{n,\ell} = O\left(\log n \; e^{-\log^2 n}\right).$$

uniformly in n and ℓ.

Proof :

Since we have

$$\left(1 - \frac{1}{2^\ell}\right)^n > \left(1 - \frac{4}{2^\ell}\right)^n > \left(1 - \frac{8}{2^\ell}\right)^n \cdots$$

and

$$Q_k > Q \quad \text{for all } k,$$

the $p_{n,\ell}$ can be bounded by

$$p_{n,\ell} < \frac{1}{Q^2}\, \ell \left(1 - \frac{1}{2^\ell}\right)^n$$

$$< \frac{1}{Q^2}\, \ell \exp\left(n \log\left(1 - \frac{1}{2^\ell}\right)\right). \qquad (9)$$

Now observing that for $u \in]0;1[$:

$$\exp(n \log(1-u)) = \exp\left(-nu - \frac{nu^2}{2} - \ldots.\right)$$

$$< e^{-nu},$$

we obtain from (9) :

$$p_{n,\ell} < \frac{1}{Q^2}\, \ell\, e^{-\frac{n}{2^\ell}}$$

$$= O\,(\log n\ e^{-\log^2 n})$$

which is thus exponentially small □

Now when ℓ is large enough, we can prove that the $p_{n,\ell}$ approach a limiting distribution in the following sense :

Proposition 3 :

Let ϕ be the function defined by

$$\phi(x) = \frac{1}{Q} \sum_{j=0}^{\infty} \frac{(-1)^j\, 2^{-\binom{j}{2}}\, e^{-x2^j}}{\prod_{i=1}^{j} (1 - \frac{1}{2^i})}$$

Then for $\ell > \log_2 n - 2 \log_2 \log n$, we have

$$p_{n,\ell} = \phi\,(\frac{n}{2^\ell}) + O\,(\frac{1}{n^{0.99}})$$

where the O (.) term is uniform in n and ℓ.

Proof :

The proof proceeds by stages using the previously mentioned approximations.

(i) Truncation of the sum : let $r = r(n) = 2\,(\log_2 n)^{1/2}$. We set

$$p'_{n,\ell} = \sum_{j=0}^{r} \frac{(-1)^j\, 2^{-\binom{j}{2}}}{Q_j\, Q_{\ell-1-j}}\ (1-2^{-(\ell-j)})^n. \qquad (10)$$

Obviously

$$|p_{n,\ell} - p'_{n,\ell}| \leq \frac{1}{Q^2} \sum_{j>r} 2^{-\binom{j}{2}} \tag{11}$$

$$= O\left(\frac{1}{n^2}\right)$$

(ii) Simplification of the denominators : define

$$p''_{n,\ell} = \frac{1}{Q} \sum_{j=0}^{r} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j} \left(1-2^{-(\ell-j)}\right)^n ; \tag{12}$$

using the fact that

$$|Q - Q_{\ell-1-j}| = O\left(\frac{1}{2^{\ell-1-j}}\right),$$

since the sum of $p''_{n,\ell}$ comprises $(r + 1)$ terms :

$$|p'_{n,\ell} - p''_{n,\ell}| = O\left(\frac{r(n)}{2^{\ell-1-r(n)}}\right)$$

$$= O\left(\frac{r(n)\ 2^{r(n)}\ \log^2 n}{n}\right)$$

$$= O\left(\frac{1}{n^{0.99}}\right). \tag{13}$$

(iii) Using the exponential approximation : given the conditions on ℓ and j, $u = 2^{-(\ell-j)}$ is always small, so that :

$$(1-u)^n = e^{n\log(1-u)}$$

$$= e^{-nu}\ e^{O(nu^2)}$$

$$= e^{-nu}\ (1 + O(nu^2))$$

since $nu^2 < 1$ for n large enough. Thus setting :

$$p'''_{n,\ell} = \frac{1}{Q} \sum_{j=0}^{r} \frac{(-1)^j 2^{-\binom{j}{2}}}{0_j} e^{-\frac{n}{2^\ell} 2^j} \tag{14}$$

we have

$$|p''_{n,\ell} - p'''_{n,\ell}| = 0\ (r(n) n\ 2^{2r(n)}\ 2^{-2\ell})$$

$$= 0\ (\frac{1}{n^{0.99}}). \tag{15}$$

(iv) Completing the sum : $p'''_{n,\ell}$ is a partial sum of $\phi\ (\frac{n}{2^\ell})$;

using again the majorization of (11), we find :

$$|p'''_{n,\ell} - \phi\ (\frac{n}{2^\ell})| = 0\ (\frac{1}{n^2}). \tag{16}$$

Thus putting together equations (10) to (16) proves Proposition 3. □

We last need information on the tail of the distribution, corresponding to values of ℓ larger than $\log_2 n$.

Proposition 4 :

For $\ell = 2 \log_2 n + \delta$ with $\delta \geq 0$, we have

$$p_{n,\ell} = 0\ (\frac{2^{-\delta}}{n^{0.99}})$$

uniformly in n and δ.

Proof (sketch) :

The proof mimics the previous one ; let us choose this time

$$r = \log_2 n + \delta$$

as the splitting value for the index in the sum giving $p_{n,\ell}$. In part (i), we now have :

$$\left| \sum_{j=r+1}^{\ell-1} \frac{(-1)^j \, 2^{-\binom{j}{2}}}{Q_j \, Q_{\ell-1-j}} \left(1-2^{-(\ell-j)}\right)^n \right| < \frac{1}{Q^2} \sum_{j>r} 2^{-\binom{j}{2}} = O\left(2^{-\delta} \, 2^{-\log^2 n}\right). \qquad (17)$$

Parts (ii) and (iii) now lead to error bounds of the form

$$O\left(\frac{2^{-\delta}}{n^{0.99}}\right)$$

since

$$2^{-(\ell-j)} = \frac{1}{n}.$$

And finally we can again complete the sum as in (iv) introducing error terms of the form (17).

We have thus proved :

$$p_{n,\ell} = \phi\left(\frac{n}{2^\ell}\right) + O\left(\frac{2^{-\delta}}{n^{0.99}}\right). \qquad (18)$$

Since $\phi(x)$ is clearly differentiable at $x = 0$, we have

$$\phi\left(\frac{n}{2^\ell}\right) = O\left(\frac{n}{2^\ell}\right) = O\left(\frac{2^{-\delta}}{n}\right). \qquad (19)$$

Thus combining (18) and (19) completes the proof of the proposition. □

In the sequel we shall need properties of function ϕ. Some of them appear to be related to classical identities in the theory of partitions. Our starting point is the following identity [Co 70]

$$\prod_{m=0}^{\infty} (1 + ut^m) = \sum_{k=1}^{\infty} \frac{u^k \, t^{\binom{k}{2}}}{(1-t)\,(1-t^2)\ldots(1-t^k)}. \qquad (20)$$

The coefficient of $[u^k t^n]$ in the left hand side member counts the number of partitions of integer n into distinct summands and, with a simple transformation on partitions, the right-hand side can be similarly interpreted (see also [Co70] for an algebraic proof). Instantiating (20) with u=-1 and $t = \frac{1}{2}$, shows that

$$\sum_{k=0}^{\infty} \frac{(-1)^k \, 2^{-\binom{k}{2}}}{Q_k} = 0 \tag{21}$$

and thus :

$$\phi(0) = 0,$$

as could be expected. We shall also need the following identities :

$$\sum_{k=1}^{\infty} \frac{(-1)^k \, k2^{-\binom{k}{2}}}{(1-\frac{1}{2})\ldots(1-\frac{1}{2^k})} = -(1-\tfrac{1}{2})(1-\tfrac{1}{4})(1-\tfrac{1}{8})\ldots, \tag{22}$$

$$\sum_{k=2}^{\infty} \frac{(-1)^k \, k(k-1)2^{-\binom{k}{2}}}{(1-\frac{1}{2})\ldots(1-\frac{1}{2^k})} = 2\,[(1-\tfrac{1}{2})(1-\tfrac{1}{4})(1-\tfrac{1}{8})\ldots] \sum_{n=1}^{\infty} \frac{1}{2^n-1}, \tag{23}$$

which are easily obtained by successive differentiation of (21) with respect to u, setting then u = -1 and $t = \frac{1}{2}$.

4. - DETERMINATION OF ASYMPTOTIC EXPANSIONS

The above developments suggest approximating $\overline{C}_n$ with the value F(n) where function F is defined for all $x \geq 0$ by :

$$F(x) = \sum_{\ell \geq 1} \ell \, \phi\left(\frac{x}{2^\ell}\right), \tag{24}$$

For large x, F can be estimated using mellin transform techniques.

We first prove

Lemma 1 :

The expected value $\overline{C}_n$ satisfies :

$$\overline{C}_n = F(n) + 0\left(\frac{1}{n^{0.98}}\right).$$

Proof :

Let us define the 3 intervals :

$$I_1 = [1 \; ; \; \log_2 n - 2 \log_2 \log n[$$

$$I_2 = [\log_2 n - 2 \log_2 \log n \; .;. \; 2 \log_2 n[$$

$$I_3 = [2 \log_2 n \; .;. \; \infty \; [,$$

and for $j = 1,2,3$:

$$C^{(j)} = \sum_{\ell \in I_j} \ell_{pn,\ell} \qquad F^{(j)} = \sum_{\ell \in I_j} \ell \; \phi \left(\frac{n}{2^\ell}\right).$$

By Proposition 2, we have :

$$|C^{(1)} - F^{(1)}| = 0\left(\log^2 n \; e^{-\log^2 n}\right) ;$$

similarly, by Proposition 3

$$|C^{(2)} - F^{(2)}| = 0\left(\frac{\log n}{n^{0.99}}\right),$$

and by Proposition 4 :

$$|C^{(3)} - F^{(3)}| = 0\left(\sum_{\delta>0} \frac{2^{-\delta}}{n^{0.99}}\right) = 0\left(\frac{1}{n^{0.99}}\right).$$

The three last equalities imply Lemma 1. □

We are thus left with estimating the behaviour of F(x) as given by (24). To that purpose, we use the Mellin integral transform which for a real function f is defined by (see [Do37]) :

$$f^*(s) = \mathcal{M}[f(x) ; s] = \int_0^\infty f(x)\, x^{s-1}\, dx. \tag{25}$$

This transform is useful for studying harmonic sums like (24) : from the obvious functional property

$$\mathcal{M}[f(ax) ; s] = a^{-s} f^*(s) \qquad a > 0 \tag{26}$$

follows formally that the Mellin transform of F is

$$F^*(s) = \Big(\sum_{\ell \geq 1} \ell\, 2^{+\ell s}\Big).\ \phi^*(s). \tag{27}$$

The Mellin transform of ϕ is itself computed using (26) repeatedly : from the definition of ϕ (again formally) we expect

$$\phi^*(s) = \frac{1}{Q} \sum_{j \geq 0} \frac{(-1)^j\, 2^{-\binom{j}{2}}\, 2^{-js}}{Q_j}\, \Gamma(s), \tag{28}$$

since, as is classically known [WW07]:

$$\mathcal{M}[e^{-x} ; s] = \int_0^\infty e^{-x}\, x^{s-1}\, dx = \Gamma(s). \tag{29}$$

Thus formally, we have :

$$F^*(s) = \frac{2^s \Gamma(s)}{(2^s-1)^2}\, \xi(s) \tag{30}$$

where

$$\xi(s) = \frac{1}{Q} \sum_{j=0}^{\infty} \frac{(-1)^j\, 2^{-\binom{j}{2}}}{(1-\frac{1}{2})\ldots(1-\frac{1}{2^j})}\, 2^{-js} \tag{31}$$

Analytically the integral in (29) is defined for $Re(s) > 0$.
For $s : -1 < R_e(s) < 0$, we have

$$\int_0^\infty (e^{-x}-1)\, x^{s-1} dx = \Gamma(s).$$

Using (21), we also have

$$\phi(x) = \frac{1}{Q} \sum_{j\geq 0} \frac{(-1)^j\, 2^{-\binom{j}{2}}\, (e^{-x2^j}-1)}{Q_j}$$

whence the integral defining the Mellin transform of ϕ is defined for $-1 < Re(s) < 0$. Actually (28) holds for any s $Re(s) > -1$ and ϕ^* has a removeable singularity at s=0. It is finally easy to see that (27) holds provided the sum there is convergent, which requires $Re(s) < 0$. Thus equations (30), (31) are justified for s in the stripe $-1 < Re(s) < 0$; there the integral of the form (25) expressing $F^*(s)$ is absolutely convergent.

The singularities of $F^*(s)$ are related to the terms in the asymptotic expansion of $F(x)$ when $x \to \infty$ [Do 55]. To see that, we use the inversion theorem for Mellin transforms which gives

$$F(x) = \frac{1}{2i\pi} \int_{d-i\infty}^{d+i\infty} F^*(s) x^{-s}\, ds \qquad (32)$$

where d can be taken arbitrarily inside the domain of absolute convergence of the integral giving $F^*(s)$. Here, we may take any d in the interval $]-1 ; 0[$.

By Cauchy's residue theorem, assuming the contour of integration can be moved to the right with $F^*(s)$ meromorphic :

$$F(x) = \frac{1}{2i\pi} \int_{e+i\infty}^{\ell-i\infty} F^*(s) x^{-s}\, ds - \sum_s \text{Res}\, (F^*(s) x^{-s}) \qquad (33)$$

where the summation is extended to all poles of $F^*(s)$ in the stripe $d < R_e(s) < e$.

The first integral should be $O(x^{-e})$ representing smaller and smaller terms (for large x) as e increases. A simple computation shows that if $F^*(s)$ has a pole of order k at $s_o = \sigma_o + it_o$, then

$$\operatorname*{Res}_{s=s_o} (F^*(s)x^{-s}) = x^{-s_o} P_{k-1} (\log x)$$

where P_{k-1} is a polynomial of degree k-1.

Since

$$x^{-s_o} = x^{-\sigma_o} e^{-it_o \log x}$$

we thus see that successive poles of F^* starting from the left yield successive terms in the asymptotic expansion of F(x) for $x \to \infty$.

We shall therefore first identify singularities of $F^*(s)$ for $\mathrm{Re}(s) \geq 0$ and then return to a formal justification of (33).

(i) $F^*(s)$ has a double pole at s = 0 as the following expansions show :

$$\Gamma(s) = \frac{1}{s}\Gamma(s+1) = \frac{1}{s}(1-\gamma s + O(s^2)) \tag{34}$$

$$\frac{2^s}{(2^s-1)^2} = \frac{1}{s^2(\log 2)^2}(1 + O(s^2)) \tag{35}$$

$$\xi(s) = \xi(0) + s\xi'(0) + \frac{s^2}{2}\xi''(0) + O(s^3) \tag{36}$$

We already know from (18) that

$$\xi(0) = 0.$$

Using (19), we can transform $\xi'(0)$:

$$\xi'(0) = -\frac{1}{Q} \sum_{j \geq 0} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j} \cdot j \log 2$$

$$= \log 2.$$

Similarly with (20) :

$$\xi''(0) = \sum_{j\geq 0} \frac{(-1)^j \, 2^{-\binom{j}{2}}}{Q_j} \, j^2 (\log 2)^2$$

$$= \frac{(\log 2)^2}{Q} \left[2Q \sum_{n\geq 1} \frac{1}{2^n - 1} - Q \right]$$

$$= (\log 2)^2 \left[2 \sum_{n\geq 1} \frac{1}{2^n - 1} - 1 \right].$$

Thus for ξ around 0 :

$$\xi(s) = + s \log 2 \, (1 + s \log 2 \, (\lambda - \tfrac{1}{2}) + 0 \, (s^2)) \tag{37}$$

with

$$\lambda = \sum_{n\geq 1} \frac{1}{2^n - 1}.$$

We find after some calculations :

$$\operatorname{Res}_{s=0} (x^{-s} F^*(s)) = - \log_2 x - \frac{\gamma}{\log 2} + \lambda - \frac{1}{2} \tag{38}$$

(ii) $F^*(s)$ also has a simple pole at $\chi_k = \frac{2ik\pi}{\log 2}$ for all $k \in \mathbb{Z} \backslash 0$. Due to the periodicity of $\xi(s)$ and 2^s, we can use some of the previous expansions ; in particular around χ_k :

$$\xi(s) = (s - \chi_k) \, \xi'(0) + 0 \, (s-\chi_k)^2 = \log 2 \, (s-\chi_k) + 0 \, (s-\chi_k)^2$$

$$\frac{2^s}{(2^s-1)^2} = \frac{1}{(s-\chi_k)^2 \, (\log 2)^2} \, (1 + 0 \, (s-\chi_k))$$

$$\Gamma(s) = \Gamma(\chi_k) + 0 \, (s-\chi_k).$$

Thus :

$$\operatorname{Res}_{s=\chi_k} \left(x^{-s} F^*(s)\right) = \frac{\Gamma(\chi_k)}{\log 2} e^{-\chi_k \log x} . \tag{39}$$

To conclude with the proof of the theorem, we only need to establish (33). To that purpose we use the rectangular contours

$$R(M,e) = R_1 + R_2 + R_3 + R_4$$

where

$$R_1 = \{d+it \;/\; t \in [-M \;;\; M]\}$$

$$R_2 = \{u+iM \;/\; u \in [d \;;\; e]\}$$

$$R_3 = \{e+it \;/\; t \in [-M \;;\; M]\}$$

$$R_4 = \{u-iM \;/\; u \in [d \;;\; e]\}$$

with R oriented clockwise. For any positive e and iM not equal to one of the χ_k, we have by cauchy's theorem applied to contour R and to integrand $F^*(0)\, x^{-s}$:

$$\frac{1}{2i\pi}\left[\int_{d-iM}^{d+iM} + \int_{d+iM}^{e+iM} + \int_{e+iM}^{e-iM} + \int_{e-iM}^{d-iM}\right] = -\sum \operatorname{Res}\left(F^*(s)x^{-s}\right)$$

where the sum is extended to all poles s with

$$-M < \operatorname{Im}(s) < +M , \quad d < \operatorname{Re}(s) < e$$

If we let M tend to infinity - keeping e fixed - in such a way that $M = \frac{(2k+1)\pi}{\log 2}$ for some integer k, we observe that, along the contour, $\xi(s)$ and $\frac{2^s}{(2^s-1)^2}$ stay uniformly bounded. The very fast decrease of $\Gamma(s)$ when Im(s) tends to infinity [WW07] entails that the second and fourth integrals then tend to 0.

The first term converges to F(x) by the inversion formula (32). As to the third one, it is bounded in modulus by :

$$\frac{x^{-e}}{2\pi} \int_{-\infty}^{+\infty} |F(e+it)|\, dt < A(e)\, x^{-e}$$

for all $x > 0$. On the right hand side, the sum is a partial sum of a Fourier series of $\log_2 x$, which is also convergent.

We have therefore established that

$$F(x) = -\log_2 x - \frac{\gamma}{\log 2} + \lambda - \frac{1}{2} + \frac{1}{\log 2} \sum_{k \in \mathbb{Z} \setminus 0} \Gamma(\chi_k)\, e^{-2ik\pi \log_2 \chi} + O(x^{-e}) \quad (40)$$

For any positive e. Combining (40) with Lemma 1, and taking e=1 establishes Theorem 1. In passing, we have proved :

Corollary :

The periodic function that expresses the fluctuations of $\overline{C}_n$ is

$$\omega(u) = \frac{1}{\log 2} \sum_{k \in \mathbb{Z} \setminus 0} \Gamma(\chi_k)\, e^{-2ik\pi u} \quad (41)$$

Such periodicities are not of unfrequent occurrence in the analysis of algorithms : a function similar to ω turns up in the analysis of radix exchange sort, as shown in [Kn73, p131]where an integration contour similar to ours is used.

Let us last briefly mention how to prove Theorem 2 relative to the variance. After n increments, the variance of the counter content is :

$$V_n = \sum \ell^2 p_{n,\ell} - \overline{C}_n^2. \quad (42)$$

To handle the first sum, we first approximate it by G(n) where

$$G(x) = \sum_{\ell} \ell^2 \phi\left(\frac{x}{2^\ell}\right), \quad (43)$$

introducing only vanishing error terms. The Mellin transform of (43) is

$$G^*(s) = \frac{2^s(2^s+1)}{(2^s-1)^3}\,\Gamma(s)\,\xi(s) \qquad (44)$$

which now has a triple pole at $s = 0$, and double poles at $s = \frac{2ik\pi}{\ell n 2}$.

Thus :

$G(x) = O(\log^2 x)$ as $x \to \infty$.

One can actually determine the terms in the asymptotic expansion of G up to $o(1)$ error terms. The main terms in G(n) cancel with those of $\overline{C}_n^2$ and we are left with the result of Theorem 2.

5. - CONCLUSIONS

The previous analysis has shown precisely that the performances of approximate counting remain remarkably stable with n (the number of increments).

Actually, [To80] report an overall performance of a system using approximate counting which is only a few percent off a reference system using exact counts. We can also remark that the accuracy of approximate counting could be improved for instance by keeping several counters and averaging their contents : this can be done efficiently representing actually one counter in binary and encoding the other ones by their difference to that reference counter ; when averaging, one should take out the asymptotic bias of -0.27395 to obtain an asymptotically almost unbiased estimation. Several conceivable variants of the algorithm can be analyzed using the previously developped methods.

Acknowledgements :

The author would like to express his gratitude to N. MARTIN for introducing him to the subject and for several enriching discussions on approximate counting.

BIBLIOGRAPHY

[Co70] L. COMTET
"L'Analyse Combinatoire",
2 vol., P.U.F., Paris (1970).

[Do55] G. DOETSCH
"Handbuch der Laplace. Transformation",
Birkhauser Verlag, Basel (1955).

[Ga78] R.G. GALLAGER
"Variations of a Theme by Huffmann",
IEEE Trans. IT, 24 (1978) pp 669-674.

[La81] G. LANGDON - J. RISSANEN
"Compression of black white Images with Binary Arithmetic Coding",
IEEE Trans. on Communications (1981).

[To81] S. TODD - N. MARTIN - G. LANGDON - D. HELMAN
"Dynamic Statistics Collection for Compression Coding",
Unpublished manuscript, 12 p. (1981).

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique