The Analysis of Linear Probing Sort by the Use of a New Mathematical Transform

Gaston H. Gonnet
J. Ian Munro

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1

Research Report CS-82-06

February 7, 1982

The Analysis of Linear Probing Sort by the Use of a New Mathematical Transform

Gaston H. Gonnet
J. Ian Munro
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1

Abstract

We present a variant of the distribution sort approach which makes use of extra storage to sort a list of n elements in an average of about $(2+\sqrt{2})n = 3.412...n$ probes into a table. An accurate analysis of this technique is made by introducing a transform from a Poisson approximation to the exact (finite) distribution. This analysis also leads to the solution of an interesting parking problem.

Keywords

sorting; analysis of algorithms; asymptotic analysis; hashing; linear probing; mathematical transform; Poisson distribution; discrete mathematics;

1. Introduction.

Our interest, in this paper, is in the detailed analysis of the behaviour of algorithms and in the development of techniques for performing such analyses. In particular we present a new interpolation-based sorting algorithm and an accurate analysis of its expected time and space requirements. In performing these analyses we first obtain an accurate analysis under a Poisson-filling model and then by use of a new transform are able to convert (expected values) to the exact model. We feel that the introduction of this transform lemma is one of the main contributions of the paper.

2. A Sorting Algorithm.

Consider the following approach to sorting a sequence of n numbers which are assumed to be uniformly distributed over a bounded range (or are easily transformed into such a distribution [Gonnet,5]). As each element is read, it is interpolated into one of the first m positions in an array. (Note that m will be taken to be greater than n unlike most other interpolation sort methods [Knuth,6, MacLaren,8].) If a conflict arises, then the smaller element takes the location in question and the larger element moves forward to the next location, and the process repeats until we find an empty location. (This may, ultimately, cause elements to overflow beyond position m.) After insertion of all elements, a single pass through the array compresses the file to the first n locations.

In the more formal outline below, we let A denote a "large enough" array and interpolate(x,m) be the function which linearly interpolates the data into integers in the range $1,...,m$.

```
for i := 1 to n do
   begin
      Read the next input, x;
      j := interpolate(x, m);
      while A[j] not empty do
         begin
            if x < A[j] then interchange (x,A[j]);
            j:= j + 1
         end
      A[j] := x;
   end
i := 1; j := 1;
while j ≤ n do ,
   begin
      if A[i] not empty then begin A[j] := A[i]; j := j + 1
                             end
      i := i + 1
   end.
```

This method is very much like the process of inserting elements into a hash table in which conflicts are resolved by linear probing. There are, however, two differences, one minor and the other more significant.

(i) We move elements already in the table in a manner similar to the ordered hashing technique of Amble and Knuth [2] (this is of little consequence),

and

(ii) there may be an overflow beyond location m. In the case of hashing this wraps around to the beginning of the table.

There are two standard models used to solve this type of problem: the *Poisson* model and the *exact filling* model. In the Poisson model we assume that each location receives a number of keys that is Poisson distributed with parameter α and is *independent* of the number of keys going elsewhere. (This implies that the total number of keys is itself a random variable whose expected value is $m\alpha$.) In the exact filling model we have n keys (balls) to be distributed among m locations (boxes) and all m^n possible arrangements are equally likely to occur.

Both models have been used extensively in the analysis of hashing algorithms. It is generally agreed that the Poisson-filling model is simpler to analyze than the exact-filling model. The main difference lies on the fact that in the Poisson model, the number of elements in each group (or box) is independent of the number of elements in other places. This is not so in the exact-filling model.

Although the Poisson model is satisfactory for most situations, there is at least one case where it fails badly. This case is exemplified by a full or almost full hashing table. In such situations the formulas are simply not applicable or give a tremendous error.

Several algorithms have been analyzed using both models. In most cases it was demonstrated that the results of the two models coincided asymptotically. In this paper we show that the results from one model can be transformed into the other and furthermore we derive an asymptotic series that represents the relation.

We perform the analysis of linear probing sort under both models. There are two issues of concern in the analysis of the algorithm:

(i) the size of the overflow, that is, the size of array beyond location m which should be taken to be relatively sure of success

and

(ii) the number of comparisons, or inspections of data locations, which are required

Consider first the overflow.

3. Analysis of the Overflow

Konheim and Weiss [7] have considered the special case of determining the probability that there is no overflow. They pose the problem in terms of drivers randomly deciding to park as they move along a street containing m parking places. If the driver comes to the end of the street, his parking effort is deemed unsuccessful. They show the probability of n drivers all finding places is

$$(1+\frac{1}{m})^n\left[1-\frac{n}{m+1}\right] \tag{1}$$

and in particular as $m,n\to\infty$ but $n/m=\alpha$, the probability becomes

$$(1-\alpha)e^{\alpha}. \tag{2}$$

In this context, we are interested in the average and in the distribution of the number of drivers who fail to park.

We begin our study by considering $p_{i,j}$, the probability that j keys overflow from the first i table locations. (Note then, for example, $p_{0,j}=0$.) A recursive relation $p_{i,j}$ in terms of $p_{i-1,k}$, can be developed.

$$p_{i,0} = p_{i-1,0}\,r_0 + p_{i-1,1}\,r_0 + p_{i-1,0}\,r_1$$
$$p_{i,1} = p_{i-1,0}\,r_2 + p_{i-1,1}\,r_1 + p_{i-1,2}\,r_0$$
$$\vdots$$
$$p_{i,j} = \sum_{k=0}^{j+1}{}' p_{i-1,k}r_{j-k+1} \qquad (j>0) \tag{3}$$

where r_i denotes the probability of having i keys map to one location. Under a Poisson filling model with parameter α, $r_i = e^{-\alpha}\alpha^i/i!$

The analysis below is based on this model. We will then translate expressions for expected values to the exact model.

Let $P_i(z) = \sum_j p_{ij}z^j$ be the generating function of the p_{ij}, then the above relation can be rewritten as

$$P_i(z) = \frac{P_{i-1}(z)\,R(z) + (z-1)\,p_{i-1,0}\,r_0}{z} \tag{4}$$
$$P_0(z) = 1,$$

where $R(z) = \sum_i r_i z^i = e^{\alpha(z-1)}$. The overflow, W_m, is described by the generating function $P_m(z)$ and so our primary interest is in the expected value, $E[W_m] = P_m'(1)$ and in the variance $\sigma^2(W_m) = P_m''(1) + P_m'(1)\,(1-P_m'(1))$. We can establish that if the sequence $P_i(z)$ converges,

$$P_\infty(z) = P(z) = \frac{(1-z)p_{\infty,0}r_0}{R(z)-z} = \frac{(1-z)(1-\alpha)}{R(z)-z} \tag{5}$$

and so, with a couple of applications of l'Hospital's Rule, it will follow that

$$E[W_\infty] = \frac{\alpha^2}{2(1-\alpha)} \tag{6}$$

and

$$\sigma^2(W_\infty) = \frac{6\alpha^2 - 2\alpha^3 - \alpha^4}{12(1-\alpha)^2} \tag{7}$$

We emphasize that at this point we still must demonstrate

(i) that $P_i(z)$ does converge to $P(z)$, and in doing so obtain approximations to relevant moments of finite $P_i(z)$ under the Poisson model

(ii) that the moments obtained under the Poisson model can be translated into the exact model.

We continue with step (i).

As it happens, $P_i(z)$ does converge to $P(z)$. To establish this asymptotic behaviour, however, we must resort to more detailed derivations, considering the case of finite m.

First note

Lemma 3.1: No $P_i(z)$ has a pole at $z=0$.

Proof: by induction, $P_0(z)$ does not have a pole at $z=0$ and by inspection of equation (4) we see that if $P_{i-1}(z)$ does not have a pole at $z=0$, $P_i(z)$ will not have one either. ▪

Next we can use the original recurrence for $p_{i,j}$ to derive a closed form for $P_i(z)$ in terms of $p_{j,0}$ $(j<i)$

Lemma 3.2:
$$P_i(z) = \gamma^i + \beta\sum_{j=0}^{i-1} p_{i-j-1,0}\gamma^j \tag{8}$$

where $\gamma = R(z)/z$ and $\beta = e^{-\alpha}(z-1)/z$.

Proof: Follows as a direct application of the solution for recurrence equations used for eq. (4). ▪

Equating the terms independent of z in eq. (8) we derive the following recursion for $p_{i,0}$:

$$p_{i,0} = e^{-i\alpha}\frac{(i\alpha)^i}{i!} + e^{-\alpha}\sum_{j=0}^{i-1} p_{i-j-1,0}\, e^{-j\alpha}\,\frac{(j\alpha)^j}{j!}\left(1-\frac{j\alpha}{j+1}\right). \tag{9}$$

Using these lemmata we are able to express $p_{i,0}$ in terms of a truncated power series expansion for e^x.

$$e_i(x) = \sum_{j=0}^{i} x^j/j! \tag{10}$$

Lemma 3.3:
$$p_{i,0} = e^{-i\alpha}(e_{i+1}((i+1)\alpha) - \alpha e_i((i+1)\alpha)) \tag{11}$$

Proof: To prove this lemma we use induction on i. The result is trivial for $i=0$. If the result is true for $j<i$, then we can substitute (11) in (9) and after simplifying $e^{-i\alpha}$ it remains only to show:

$$e_{i+1}((i+1)\alpha)-\alpha e_i((i+1)\alpha) = \frac{(i\alpha)^i}{i!}+\sum_{j=0}^{i-1}[e_{i-j}((i-j)\alpha)-\alpha e_{i-j-1}((i-j)\alpha)]\,\frac{(j\alpha)^j}{j!}\left(1-\frac{j\alpha}{j+1}\right) . \tag{12}$$

The above equation is far from trivial. We present a full proof of it using an embedding technique very well suited to treat the function $e_i(x)$.

First note that the left side is a polynomial in α of degree i (the α^{i+1} terms cancel) as are those of the right. This fact is of considerable help since we can add terms of order α^{i+1} or higher on the left and right without altering the embedded polynomial of degree i.

In doing so we complete the functions $e_i(x)$ to e^x using the equality

$$e_i(x) = e^x + O(x^{i+1}) \tag{13}$$

noting that only terms in α^{i+1} or higher are added. The claimed equality is now

$$e^{(i+1)\alpha} - \alpha e^{(i+1)\alpha} = \frac{(i\alpha)^i}{i!}+\sum_{j=0}^{i-1}(e^{(i-j)\alpha} - \alpha e^{(i-j)\alpha})\,\frac{(j\alpha)^j}{j!}\left(1-\frac{j\alpha}{j+1}\right) + O(\alpha^{i+1}), \tag{14}$$

including $-\dfrac{(i\alpha)^{i+1}}{(i+1)!}$ on the right hand side and regrouping we obtain

$$(1-\alpha)e^{(i+1)\alpha} = e^{i\alpha}(1-\alpha)\sum_{j=0}^{i} e^{-j\alpha}\frac{(j\alpha)^j}{j!}\left(1-\frac{j\alpha}{j+1}\right)+O(\alpha^{i+1}) \tag{15}$$

and finally completing the sum with more terms of order α^{i+1} or higher

$$e^{\alpha} = \sum_{j=0}^{\infty}\left[\frac{j\alpha}{e^{\alpha}}\right]^j\left(1-\frac{j\alpha}{j+1}\right)\frac{1}{j!}. \tag{16}$$

This is an Abel expansion of e^{α} which can be derived also from the Lagrange inversion formula [Riordan, 10 S.3.2]. This type of summation can be, rather systematically, treated with the aid of the transcendental function $w(x)$ defined by $w(x)e^{w(x)}=x$. It is well known that

$$\sum_{j=1}^{\infty}\frac{j^{j-1}x^j}{j!} = -w(-x) \tag{17}$$

from which it is easy to verify that

$$\sum_{j=1}^{\infty}\frac{j^j x^{j-1}}{j!} = w'(-x) = \frac{1}{e^{w(-x)}-x} \tag{18}$$

and

$$\sum_{j=1}^{\infty}\frac{j^{j+1}x^{j+1}}{(j+1)!} = \frac{e^{w(-x)}}{1+w(-x)} - 1 \tag{19}$$

Consequently, regrouping terms in (16) we need only show

$$e^{\alpha} = 1 + \sum_{j=1}^{\infty}\left[\frac{\alpha}{e^{\alpha}}\right]^j\frac{j^j}{j!} - \sum_{j=1}^{\infty}e^{\alpha}\left[\frac{\alpha}{e^{\alpha}}\right]^{j+1}\frac{j^{j+1}}{(j+1)!} \tag{20}$$

which, noting that $w(-\alpha e^{-\alpha}) = -\alpha$ by definition, reduces to the identity

$$e^{\alpha} = 1 + \frac{\alpha}{e^{\alpha}} \frac{1}{e^{-\alpha}-\alpha e^{-\alpha}} - e^{\alpha}\left(\frac{e^{-\alpha}}{1-\alpha} - 1\right)$$

■

From Lemma 3.3 and the definition of $e_i(x)$ we can derive not only the bounds of Konheim and Weiss, but also that the convergence in terms of i of $p_{i,0}$ to $(1-\alpha)e^{\alpha}$ is exponential:

Theorem 3.1: for $\alpha<1$

$$p_{i,0} = (1-\alpha)e^{\alpha} + O(i^{-1/2}e^{(1-\alpha+\ln\alpha)i}) \tag{21}$$

(note that for $\alpha<1$, $1-\alpha+\ln\alpha<0$) and for $\alpha=1$

$$p_{i,0} = O(i^{-1/2}) \tag{22}$$

Proof: For $\alpha<1$ we can see, from the definition of $e_i(x)$ that

$$\begin{aligned} &|e^{-i\alpha}(e_{i+1}((i+1)\alpha)-e^{-(i+1)\alpha})| \qquad (23)\\ &\quad \le e^{-i\alpha}\frac{((i+1)\alpha)^{i+2}}{(i+2)!}\frac{1}{1-\frac{(i+1)\alpha}{i+3}} \\ &\quad = O(e^{(1-\alpha+\ln\alpha)i}i^{-1/2}) \qquad iff\ \alpha<1\ . \end{aligned}$$

Consequently, applying the above twice on equation (11) we find that for fixed α eq. (21) holds true. (Note that when α approaches 1, $1-\alpha+\ln\alpha \approx -\frac{(1-\alpha)^2}{2}$).

For $\alpha=1$, using the definition of $p_{i,0}$ we conclude that

$$\begin{aligned} p_{i,0} &= e^{-i}(e_{i+1}(i+1)-e_i(i+1)) \\ &= e^{-i}\frac{(i+1)^{i+1}}{(i+1)!} \approx \frac{e}{\sqrt{2\pi i}} = O(i^{-1/2}). \end{aligned}$$ ■

Another line of attack is to apply Lemma 3.3 (note the α^{i+1} term is 0) and express $p_{i,0}$ as a power series in α. This leads to a technically useful lemma.

Lemma 3.4: $$p_{i,0}=(1-\alpha)e^{\alpha} + O(\alpha^{i+2}) \tag{24}$$

Lemma 3.2 together with the convergence of $p_{i,0}$ demonstrated above leads to:

Lemma 3.5: $\lim_{i\to\infty} P_i(z) = \frac{(1-\alpha)(z-1)}{z-R(z)} = P(z)$ in the interval $[1, 1/\alpha]$.

[In fact the lemma holds for a much larger range, however, the proof is easier in the restricted range and no more is necessary for our main results.]

Taking derivatives with respect to z and evaluating at $z=1$ we use the last two lemmata to demonstrate

$$P_i'(1) = R'(1) + P_{i-1}'(1) + p_{i-1,0}r_0 - 1 \tag{25}$$

$$= \sum_{j=0}^{i-1} \alpha - 1 + p_{j,0} e^{-\alpha}.$$

Using eq. (24) we find that for $j < i$

$$P_i'(1) = P_j'(1) + \sum_{k=j}^{i-1} O(\alpha^{k+2}). \tag{26}$$

Since we know that $\lim_{i \to \infty} P_i'(1) = \frac{\alpha^2}{2(1-\alpha)}$ we conclude that

$$P_i'(1) = P_\infty'(1) - \sum_{j \geq i} O(\alpha^{j+2}) = \frac{\alpha^2}{2(1-\alpha)} + O(\alpha^{i+2}) \tag{27}$$

Theorem 3.2: The expected overflow of table of size m with load factor α is $\frac{\alpha^2}{2(1-\alpha)} + O(\alpha^{m+2})$.

For example computing directly from equations (25) and (11) we find:

$$P_1'(1) = \frac{\alpha^2}{2} - \frac{\alpha^3}{6} + \frac{\alpha^4}{24} \cdots$$

$$P_2'(1) = \frac{\alpha^2}{2} + \frac{\alpha^3}{2} - \frac{5\alpha^4}{8} + \frac{47\alpha^5}{120} \cdots$$

$$P_3'(1) = \frac{\alpha^2}{2} + \frac{\alpha^3}{2} + \frac{\alpha^4}{2} - \frac{49\alpha^5}{30} + \frac{1331\alpha^6}{720} \cdots$$

As previously noted, this result is valid under a model in which each location receives an independent random number of keys with Poisson distribution. To translate this result to the model in which n keys are distributed randomly among the m locations, we will use the following development.

4. The Mathematical Transform

Let $f(m,n)$ be an expected value computed using a model of n objects randomly distributed among m locations. Let $g(m,\alpha)$ be the equivalent expected value computed using a model with m random independent Poisson distributed objects each with paramenter α. Then

$$g(m,\alpha) = \sum_{n=0}^{\infty} f(m,n) Pr\{X_1 + X_2 + \cdots + X_m = n\} \tag{28}$$

where X_i is a random variable with Poisson distribution and parameter α. The above equality is easy to verify once we realize that the distribution of the X's, conditioned to their sum being n, coincides with the random distribution of n objects in m places. In some sense this is a non-interesting conversion since in general $f(m,n)$ is more difficult to compute than $g(m,\alpha)$. In what follows, we will be able to invert this equality, i.e., obtain $f(m,n)$ from $g(m,\alpha)$.

It is known that the sum of independent Poisson distributed variables is also Poisson, and hence

$$g(m,\alpha) = \sum_{n=0}^{\infty} f(m,n) \frac{(m\alpha)^n e^{-m\alpha}}{n!} \tag{29}$$

Note that this is valid for any value of α; therefore it is an identity in α. In our case we know the function $g(m,\alpha)$, and want to find its inverse transform according to this identity. To do this, we simply write it as

$$e^{m\alpha} g(m,\alpha) = \sum_{n=0}^{\infty} f(m,n) \frac{(m\alpha)^n}{n!}. \tag{30}$$

Note that by definition, or by the use of equation (29),

$$\lim_{\alpha \to 0} g(m,\alpha) = f(m,0)$$

is finite and furthermore, any order of derivative of $g(m,\alpha)$ with respect to α has a finite limit when $\alpha \to 0$. Consequently $g(m,\alpha)$ has a unique Maclaurin expansion in powers of α. From 30, equating powers of α and using $n^{\underline{i}}$ to denote the descending factorial $n^{\underline{i}} = n!/(n-i)!$, we obtain:

Lemma 3.6: If $g(m,\alpha) = \sum a_i \alpha^i$ is an expected value in a model consisting of m independent, (Poisson-distributed with parameter α) number of objects (the a_i may be functions of m) then

$$f(m,n) = \sum a_i \frac{n^{\underline{i}}}{m^n} \tag{31}$$

is the corresponding expected value in a model consisting of n objects randomly distributed among m locations.

For example, in the case of hashing, the Poisson-filling model assumes that to each location there is a Poisson-distributed number of keys initially probing to it. In the exact-filling model, the first probes of the n keys are randomly distributed among m locations such that each of the m^n distributions is equally likely to occur.

A table of useful transforms follows

Poisson Model	Exact Model
$g(m,\alpha)$	$f(m,n)$
$\sum_{n\geq 0} f(m,n)\ \frac{(m\alpha)^n e^{-m\alpha}}{n!}$	$f(m,n)$
$g'_\alpha(m,\alpha)$	$m[f(m,n+1) - f(m,n)]$
$\int_0^\alpha g(m,t)\ dt$	$\frac{1}{m}\sum_{k=0}^{n-1} f(m,k)$
α^r	$\frac{n^x}{m^r}$
$\frac{1}{(1-\alpha)^r}$	$Q_r(m,n) = \sum_{k\geq 0}\binom{r+k}{k}\frac{n^k}{m^k}$
$e^{\alpha a}$	$\left(\frac{m+a}{m}\right)^n$
$\alpha^r e^{\alpha a}$	$\frac{n^x}{m^r}\left(\frac{m+a}{m}\right)^{n-r}$
$\int_0^a \frac{1-e^{-mt}}{t}dt$	H_n
$\frac{e^{-a}\alpha^i}{i!}$	$\binom{n}{i}(m-1)^{n-i}m^{-n}$

Intuitively, $f(m,n)$ is well approximated by $g(m,\alpha)$ with $\alpha = n/m$. We will now formalize this approximation by finding an asymptotic approximation of $f(m,n)$ in terms of $g(m,n/m)$. To do this we will approximate $\frac{n^{\underline{i}}}{m^i}$ by α^i with $\alpha=n/m$.

Consequently

$$\begin{aligned} f(m,n) &= \sum_i a_i \frac{n^{\underline{i}}}{m^i} = \sum_i a_i \alpha^i + \sum_i a_i \left(\frac{n^{\underline{i}}}{m^i} - \alpha^i\right) \\ &= g(m,\alpha) + \sum_i a_i \left(\frac{n^{\underline{i}}}{m^i \alpha^i} - 1\right)\alpha^i \\ &= g(m,\alpha) + \sum_i a_i \left(\frac{n^{\underline{i}}}{n^i} - 1\right)\alpha^i \end{aligned} \tag{32}$$

Using the Stirling asymptotic approximation of factorial to compute $n^{\underline{i}}$ and doing some computation we derive

$$f(m,n) = g(m,\alpha) + \sum a_i \left(-\frac{i(i-1)}{2n} - \frac{i(i-1)(i-2)(3i-1)}{24n^2} + O(n^{-3})\right)\alpha^i \tag{33}$$

Now if $g(m,\alpha)$ can be differentiated with respect to α a sufficient number of times we know that

$$g_\alpha^{II}(m,\alpha) = \sum a_i i(i-1)\alpha^{i-2},$$
$$g_\alpha^{III}(m,\alpha) = \sum a_i i^2\alpha^{i-3},$$

and so on.

Substituting in the above we find the first 3 terms of the asymptotic expansion:

Theorem 3.3 (Approximation theorem)

$$f(m,n) = g(m,\alpha) - \frac{\alpha g_\alpha^{II}(m,\alpha)}{2m} + \frac{\alpha}{24m^2}[3\alpha g_\alpha^{IV}(m,\alpha)+8g_\alpha^{III}(m,\alpha)] - \frac{\alpha}{48m^3}[\alpha^2 g_\alpha^{VI}(m,\alpha) + 8\alpha g_\alpha^{V}(m,\alpha) + 12 g_\alpha^{IV}(m,\alpha)]+\ldots. \tag{34}$$

for $\alpha = n/m$.

Depending on the asymptotic behaviour in m of the derivatives of $g(m,\alpha)$ with respect to α, the above may be a proper asymptotic expansion in terms of m. (In all our real examples this is the case). For example in the analysis of liner probing hashing, we may want approximations of $g(m,\alpha) = \frac{1}{1-\alpha}$ which is independent of m and consequently we obtain proper asymptotic series. The function $g(m,\alpha) = e^{\alpha m}$ has derivatives in α that are of increasing order in m, consequently Theorem 3.3 does not produce an asymptotic series in m; in this case the transform is $f(m,n) = 2^n$

We are now able to complete the proof of the main results of linear probing sort. Let $W_{m,n}$ denote the overflow of a table with m locations and n entries. Note that $n^{\underline{i}} = 0$ for $i>n$, consequently for $n\le m+1$, Theorem 3.2 gives an exact transform.

Theorem 3.4: if $n\le m+1$

$$E[W_{m,n}] = \frac{1}{2}\sum_{i=2}^{n}\frac{n^{\underline{i}}}{m^i} \tag{35}$$

and

$$E[W_{m,m}] = \sqrt{m\pi/8} + 2/3 + O(m^{-1/2}) \tag{36}$$

Proof: The auxiliary functions $Q_r(m,n)$ prove useful. In particular we can show

$$E[W_{m,n}] = \frac{Q_0(m,n)-1-\alpha}{2} = \frac{n(n-1)}{2m^2}Q_0(m,n-2)$$
$$= \frac{n^2}{2m(m-n)}+O(m^{-1}).$$

Similar values on the variance of $W_{m,n}$ can also be achieved.

Theorem 3.5:

$$\sigma^2(W_{m,n}) = \frac{n^2(6m^2-2mn-n^2)}{12m^2(m-n)^2} + O(m^{-1}) \quad \text{for } n<m \tag{37}$$

and

$$\sigma^2(W_{m,m}) = \frac{(4-\pi)m}{8} + \frac{1}{9} - \frac{\pi}{48} + O(m^{-1/2})$$

Sketch of Proof: The proof is completely analogous to those of the preceeding two theorems as the second moment of $W_{m,n}$ is computed. The variance is then determined as

$$\sigma^2(W_{m,n}) = \frac{5m^2+4mn-n(n-2)}{12m^2} + \frac{Q_1(m,n)}{2} + \frac{3n-4m}{6m}Q_0(m,n) - \frac{Q_0(m,n)^2}{4},$$

leading to the result. ▪

If we are to use the interpolation sorting algorithm, the key issue regarding the overflow is not its mean or standard deviation, but in fact the probability of it exceeding a given bound. Such probability is given by the following theorem:

Theorem 3.6:

$$Pr\{W_m > k\} \le (1-\alpha)\sum_{j\ge 0} \frac{e^{-j\alpha}(j\alpha)^{j+k+1}}{(j+k+1)!} \tag{38}$$

Proof:

$$\begin{aligned} Pr\{W_m > k\} &= \sum_{j>k} p_{m,j} \le \sum_{j>k} p_{\infty,j} \\ &\le \sum_{j>k} \frac{1}{2\pi i}\int \frac{P(z)}{z^{j+1}}dz = \frac{1}{2\pi i}\int \frac{P(z)dz}{z^{k+1}(z-1)} \\ &\le \frac{1}{2\pi i}\int \frac{(1-\alpha)dz}{z^{k+1}(z-R(z))} = \frac{1-\alpha}{2\pi i}\int \sum_{j\ge 0}\left[\frac{R(z)}{z}\right]^j \frac{dz}{z^{k+2}} \\ &\le (1-\alpha)\sum_{j\ge 0} \frac{e^{-j\alpha}(j\alpha)^{j+k+1}}{(j+k+1)!} \end{aligned}$$

▪

A careful bounding of the summation in Theorem 3.6 by the saddle point method [de Bruijn,3] (in this case the main contribution is proportional to the maximum) and the application of the transform leads to

Theorem 3.7:

$$\ln(Pr\{W_{m,n} > k\}) \approx -2k\left(1-\frac{n}{m}\right) \tag{39}$$

Exact computations in Theorem 3.6 lead to statements such as

"For a table 80% full the probability of exceeding 36 locations is less than 10^{-7}."

5. On the Number of Comparisons

The algorithm clearly falls into two parts, the insertion phase and the compression phase. The latter requires $m + W_{m,n}$ probes into the table, and so the former is the topic of this section. To determine this number of probes we start with the analysis of linear probing hashing.

It is well known that the number of comparisons to search or build a linear probing hash table is independent of the order in which the keys are inserted (Knuth [6], Peterson [9]). In the insertion phase of our algorithm, the interchanges are equivalent to changing the order of insertion. Consequently the

total number of comparisons does not depend on the order of the input nor on the outcome of these interchanges.

The number of comparisons would be exactly given by the number of comparisons in linear probing hashing except for the fact we do not wrap-around the table. It can be seen that in no situation does our algorithm requires more comparisons than an equivalent hashing table. While the converse is not true, the gain is not significant either. If we simply ignore the $W_{m,n}$ keys that overflow, the number of comparisons is in this phase $C(m,n)$, is bounded by

$$H(m,n-W_{m,n}) \le C(m,n) \le H(m,n)$$

where $H(m,n)$ denotes the number of comparisons (or probes) to insert n elements into a linear probing hash table of size m. The exact value of $H(m,n)$ is well known [6,9], and so we immediately have a rather good approximation to $C(m,n)$, the total number of comparisons required by linear probing sort.

We can, however, apply the technique of the preceding section to produce a much better analysis.

The total number of comparisons can be expressed as the sum of three components: the initial comparisons; additional comparisons inside the table and additional comparisons in the overflow area. For the Poisson model this is simply

$$E[\textit{total comparisons}] = C(m,\alpha) \tag{40}$$
$$= m\alpha + \sum_{i=1}^{m} P_i'(1) + E[\frac{W_{m,\alpha}(W_{m,\alpha}-1)}{2}].$$

From the analysis in the previous section we conclude that

Lemma 4.1:

$$E[\frac{W_{m,\alpha}(W_{m,\alpha}-1)}{2}] = P_m''(1) \tag{41}$$
$$= \frac{1}{4(1-\alpha)^2} - \frac{5}{6(1-\alpha)} + \frac{\alpha^2+4\alpha+7}{12} + O(\alpha^{m+1})$$

The main problem remains finding a closed formula for the summation. As before we are mainly interested in the expression of the sum up to terms of $O(\alpha^{m+2})$.

Using eqs. (25) and (11) we can rewrite the sum as

$$\sum_{i=1}^{m} P_i'(1) = \tag{42}$$
$$\sum_{i=1}^{m}\sum_{j=0}^{i-1}\left\{(\alpha-1)+e^{-j\alpha}(e_{j+1}((j+1)\alpha)-\alpha e_j((j+1)\alpha))e^{-\alpha}\right\}$$
$$=\sum_{i=1}^{m}\sum_{j=0}^{i-1} e^{-(j+1)\alpha}\left[e_{j+1}((j+1)\alpha)-e^{(j+1)\alpha}-\alpha(e_j((j+1)\alpha)-e^{(j+1)\alpha})\right]$$
$$=\sum_{j=0}^{m-1}(m-j)e^{-(j+1)\alpha}\left[e_{j+1}((j+1)\alpha)-e^{(j+1)\alpha}-\alpha(e_j((j+1)\alpha)-e^{(j+1)\alpha})\right]$$

Now let us compute the coefficient of the term in α^{m+1} in the above summation, it follows after checking the limits of summations that

$$[\alpha^{m+1}]\{\sum_{i=1}^{m} P_i'(1)\} = \tag{43}$$

$$= \sum_{j=0}^{m-1} (m-j) \sum_{k=0}^{m-j-1} \frac{(-j-1)^k}{k!} \left[-\frac{(j+1)^{m+1-k}}{(m+1-k)!} + \frac{(j+1)^{m-k}}{(m-k)!} \right]$$

$$= \sum_{j=0}^{m-1} \sum_{k=0}^{m-j-1} \frac{(m-j)(j+1)^m}{m!} \left[-\binom{m+1}{k} \frac{j+1}{m+1} + \binom{m}{k} \right] (-1)^k$$

To solve the above summation we make use of the formulae

$$\sum_{j=0}^{m-k-1} \binom{m}{k} (-1)^k \sum_{j=0}^{m-k-1} (j+1)^m = \frac{(m+1)!}{2} \tag{44}$$

$$\sum_{k=0}^{m-1} \binom{m}{k} (-1)^k \sum_{j=0}^{m-k-1} (j+1)^{m+1} = \frac{(3m+1)(m+2)!}{24}$$

$$\sum_{k=0}^{m-1} \binom{m+1}{k} (-1)^k \sum_{k=0}^{m-k-1} (j+1)^{m+1} = \frac{m(m+1)!}{2}$$

$$\sum_{k=0}^{m-1} \binom{m+1}{k} (-1)^k \sum_{k=0}^{m-k-1} (j+1)^{m+2} = \frac{(3m+1)m(m+2)!}{24}$$

These formulae are readily verified by substituting $(j+1)^m$ for a sum of descending factorials of $j+1$ using the Stirling numbers of the second kind (Abramowitz [1]). The inner summation becomes also a descending factorial and all summations except the first vanish.

Substituting in the above, after expressing $m-j = m+1-(j+1)$ we obtain

$$[\alpha^{m+1}]\{\sum_{i=1}^{m} P_i'(1)\} = \frac{1}{m!} \left[-\frac{m(m+1)!}{2} + \frac{(3m+1)m(m+2)!}{24(m+1)} \right. \tag{45}$$

$$\left. + \frac{(m+1)(m+1)!}{2} - \frac{(3m+1)(m+2)!}{24} \right]$$

$$= \frac{-3m^2+5m+10}{24}$$

In general, the term in α^{j+1} will be

$$[\alpha^{j+1}]\{\sum_{i=1}^{m} P_i'(1)\} = \frac{-3j^2+5j+10}{24} + \frac{(m-j)}{2}$$

since, due to eq (27) the coefficient of α^{j+1} in $P_i'(1)$ is $1/2$ for $i>j$. Consequently

$$\sum_{i=1}^{m} P_i'(1) = \sum_{i=1}^{m} \left\{ \frac{m-i}{2} + \frac{-3i^2+5i+10}{24} \right\} \alpha^{i+1} + O(\alpha^{m+2}) \tag{46}$$

which after normal summation methods yields

Lemma 4.2

$$\sum_{i=1}^{m} P_i'(1) = \frac{m\alpha^2}{2(1-\alpha)} - \frac{\alpha^2(16\alpha-10\alpha^2)}{24(1-\alpha)^3} + O(\alpha^{m+2}) \tag{47}$$

Finally, putting all the results together we find

is much simpler to average values for the whole table.)

The total number of additional comparisons is then $\frac{m\alpha^2}{2(1-\alpha)}$ and its transform is $\frac{m}{2}[-1-\frac{n}{m}+Q_0(m,n)]$. The average number of accesses per key is the latter divided by n plus 1 (first access):

$$C_n = \frac{1}{2} + \frac{m}{2n}(Q_0(m,n)-1) = \frac{1}{2} + \frac{Q_0(m,n-1)}{2}$$

which coincides with the previously reported results. Furthermore, we can apply the approximation theorem and obtain new asymptotic approximations for C_n. For example, using the first three terms we obtain

$$C_n = \frac{1}{2}[1+\frac{m}{m-n}] - \frac{1}{2m(1-\alpha)^3} + \frac{\alpha+2}{2m^2(1-\alpha)^5} + O(m^{-3})$$

where $\alpha = n/m$.

Another example can be found in Knuth's problem 6.4.56 [M48]: find the exact average number of buckets accessed in linear probing hashing using buckets of size b. This is solved by computing the transform of an equation given in the same text [Knuth,6.4, eq (61)].

In direct chaining hashing it is relatively standard to use a Poisson-filling model. In this case several measures are known in both models. Further measures are investigated and converted in [Gonnet,4].

7. References

[1] Abramowitz, M. and Stegun, I.A.: *Handbook of Mathematical Functions*, Dover, New York, 1972.

[2] Amble, O. and Knuth, D.E.: Ordered Hash Tables. The Computer J. 17(2):135-142, (May 1974).

[3] de Bruijn, N.G.: *Asymptotic Methods in Analysis*, North Holland. Amsterdam 1970.

[4] Gonnet, G.H.: Expected Length of the Longest Probe Sequence in Hash Code Searching; JACM, 28(2):289-304 (Apr. 1981).

[5] Gonnet, G.H., Rogers, L.D. and George, A.: An Algorithmic and Complexity Analysis of Interpolation Search. Acta Informatica, 13(1):39-46, (Jan 1980).

[6] Knuth, D.E.: *The Art of Computer Programming: Sorting and Searching*, Vol III, Addison-Wesley, Don Mills, Ont. 1973.

[7] Konheim, A.G. and Weiss, B.: An Occupancy Discipline and Applications. SIAM J. Applied Math. 14:1266-1274, (1966).

[8] MacLaren, M.D.: Internal Sorting by Radix Plus Sifting, J.ACM, 13(3):404-411, (July 1966).

[9] Peterson, W.W.: Addressing for Random-Access Storage. IBM Journal of Research and Development, 1(4):130-146, (Apr 1957).

[10] Riordan, J.: *Combinatorial Identities*, John Wiley, New York, 1968.