Analysis of Hashing Algorithms and a New Mathematical Transform [*] [†]

by

Alfredo Viola

Waterloo, Ontario, Canada, 1995

[*] This report is based on the author's PhD thesis. Many results are joint work with J. Ian Munro and Patricio V. Poblete.

[†] Supported in part by the Natural Science and Engineering Research Council of Canada under grant number A-8237, the Information Technology Research Centre of Ontario and FONDECYT(Chile) under grants 1940271 and 1950622.

Abstract

The main contribution of this report is the introduction of a new mathematical tool that we call the Diagonal Poisson Transform, and its application to the analysis of some linear probing hashing schemes. We also present what appears to be the first exact analysis of a linear probing hashing scheme with buckets of size b.

First, we present the Diagonal Poisson Transform. We show its main properties and apply it to solve recurrences, find inverse relations and obtain several generalizations of Abel's summation formula.

We follow with the analyisis of LCFS hashing with linear probing. It is known that the Robin Hood linear probing algorithm minimizes the variance of the cost of successful searches for all linear probing algorithms. We prove that the variance of the LCFS scheme is within lower order terms of this optimum.

Finally we present the first exact analysis of linear probing hashing with buckets of size b. From the generating function for the Robin Hood heuristic, we obtain exact expressions for the cost of successful searches when the table is full. Then, with the help of Singularity Analysis, we find the asymptotic expansion of this cost up to $O((bm)^{-1})$, where m is the number of buckets. We also give upper and lower bounds when the table is not full. We conclude with a new approach to study certain recurrences that involves truncated exponentials. A new family of numbers that satisfies a recurrence resembling that of the Bernoulli numbers is introduced. These numbers may prove helpful in studying recurrences involving truncated generating functions.

Acknowledgements

This thesis owes its existence largely to the strong support of my supervisors Professor Ian Munro and Professor Patricio Poblete who introduced me to the area of Analysis of Algorithms. Among other things, Ian was very generous with my financial support and his unmatched intuition gave rise to several fruitful conversations. With his insight, Patricio encouraged me in my search for conceptual solutions to my research problems. Their example will be an inspiration for my future research. I also wish to thank the other members of my thesis committee: Professor Prabhakar Ragde, Professor Anna Lubiw, Professor Bruce Richmond and Professor Kevin Compton for their helpful feedback. I thank Bruce especially for the generous gift of his time to speak with me on topics related to asymptotic analysis.

I am very thankful to Professor Gaston Gonnet for his advice in several important aspects of my studies. Gaston was my supervisor for my Master's degree and supported the first year of my Ph.D. studies. As co-director of the Symbolic Computation Group at the University of Waterloo, he initiated the Maple project, and I would like to extend my gratitude to all the developers of this powerful system. Not only did Maple assist us in making conjectures about the results we wished to prove, but it was also used to check most of the solutions presented in this thesis. Many thanks go to Professor Philippe Flajolet who pointed out several references related to analytic methods for average-case analysis of algorithms and to singularity analysis that played an essential role in the asymptotic results presented in Chapter 5.

Professor Frank Tompa was my advisor during the first year of my program. I am grateful to him for his support at a time when I had to make some important decisions. It was also a pleasure for me to work with Professor Ming Li on topics not related to this thesis.

I would like to acknowledge the support I received from the members of the faculty and staff of the Computer Science Department who were kind and efficient in dealing with my requests. A special thanks goes to Wendy Rush who was always willing to help me with administrative problems.

My life in Canada was made enjoyable by all the friends that I have had the opportunity to meet while I was here. My warmest appreciation goes to Glenn Paulley and Leslie Cornwell for all their support and friendship. I will particularly remember all those times we met to play Bridge. With my good friends Andrej Brodnik and David Clark, we had the opportunity to discuss each other's theses. Their comments and suggestions were greatly appreciated. Moreover Andy, David, Glenn and I shared one of my most enjoyable activities in my life at University: for almost two years, we devoted one hour each week to play Bridge. I also want to express my gratitude to Darrell Raymond for his unconditional support when help was needed. Mariano Consens and I worked together for four years administering the Uruguayan mailing list, a duty that I really enjoyed fulfilling. I want to express my gratitude to Mariano for his support and advice, especially during my final year of studies.

I want to mention in a very special way Daniel Panario and Lucia Moura. Together we shared some of the most beautiful times in our stay in Canada, We also shared difficult moments and important decisions, and their personal advice always brought new light to me. Furthermore, I had the pleasure to work with Daniel on several problems not related to the results presented here. Daniel read early drafts of this thesis and his observations were warmingly welcomed.

I am also thankful to Jorge Sotuyo, Julio Villafuerte and Marcela Diaz, Tiziana Digiorgio and Giovanni Cascante, Claudia Iturriaga-Velazquez and Alex Lopez-Ortiz, Catalina Alvarez, Igor Benko and Jasna Jurjovec, Ricardo Baeza-Yates and Susana Contreras, Tim Snider, Rolf Fagerberg, Tom Papadakis, Rene Mayorga, the Brazilian community in Waterloo, and my host family Ted, Carlene, Lawrence, Matthew and Mickey Goddard, for all the pleasant memories I left behind.

Finally, I want to thank my family for their unconditional love during these years. My warmest feelings go to my wife Graciela and to my daughter Manuelita for providing meaning in my life. Almost one year ago, Graciela and Manuelita returned to Uruguay, and two months later I visited them for one short week. Half an hour before I left to return to Canada, Manuelita grabbed my hand and began a conversation that marked one of the special moments in my life. In many ways, this conversation guided me in this my final year of research. My memory of it was so strong that one day I was inspired to write a short story in Spanish about it. I feel that Manuelita deserves a privileged place in this thesis and so, at the beginning of each chapter I quote some fragments of this story, starting with its first sentence immediately prior to Chapter 1 and ending with its last prior to Chapter 6.

To my daughter Manuelita and the moon, the sources of my inspiration and love.

Contents

Chapter 1

Introduction

To me, the moon always meant mystery, magic, and mystique; but above all romanticism, love, life, hope, and happiness.

1.1 Introduction

The idea of hashing seems to have been originated by H. P. Luhn, in an internal IBM memorandum in January 1953 [46]. The first major paper published in the area is the classic article by Peterson [69]. In this work, Peterson defines open addressing in general, and gives empirical statistics about linear probing hashing. He also notes the degradation in performance when records are deleted. Moreover, he acknowledges that the open addressing idea was devised in 1954 by A.L. Samuel, G.M. Amdahl, and E. Boehme. A good early survey of the area is the paper by W. Buchholz [12]. Nevertheless, as noted by Knuth [46], the word "hashing" to identify this technique appeared for the first time in the literature in the survey of Morris [63], although it had been in common usage for several years. In that paper he introduced the idea of random probing (with secondary clustering).

Linear probing is the simplest collision resolution for open addressing. It works reasonably well for tables that are not too full, but as the load factor increases, its performance deteriorates rapidly. The longer a contiguous sequence of key grows, the more likely collisions with this sequence will occur when new keys are inserted. Furthermore, one insertion may coalesce two long clusters. This phenomenon is called primary clustering.

The main application of linear probing is to retrieve information in secondary storage devices when the load factor is not too high, as first proposed by Peterson [69]. It was also proposed by Larson as a method to handle overflow records in linear hashing schemes [56, 57]. One reason for the use of linear probing is that it preserves locality of reference between successive probes, thus avoiding long seeks [55].

The first published analysis of linear probing for buckets of size 1, was done by Konheim and Weiss [52]. However, this algorithm was first analyzed by Knuth in 1962 [46, 82], who stated that this analysis had a strong influence in the structure of his series "The Art of Computer Programming". A different approach to the analysis of this hashing scheme, based on the application of ballot theorems, was presented by Mendelson and Yechiali [61]. Pflug and Kessler [70] study the case in which the keys are nonuniformly distributed. They do an asymptotic analysis for the case in which the size of the table tends to infinity while the load factor is constant. Pittel [71], also presents an asymptotic analysis of the probable largest cost of a successful search. Finally, Aldous [3] studies the case when the access probabilities of the keys are not uniform.

Operating primarily in the context of double hashing, several authors [10, 4, 36] observed that a collision could be resolved in favor of *any* of the keys involved, and used this additional degree of freedom to decrease the expected search time in the table. We obtain the standard schemes by letting the incoming key probe its next location. Celis *et al.* [16, 17] were the first to observe that collisions could be resolved having *variance reduction* as a goal. They defined the Robin Hood heuristic, in which each collision occurring on each insertion is resolved in favor of the key that is farthest away from its home location. Later, Poblete and Munro [73] defined the last-come-first-served heuristic, where collisions are resolved in favor of the incoming key, and others are moved ahead one position in their probe sequences. In both cases, the reduction of the variance can be used to speed up searches by replacing the standard search algorithm by a "mean-centered" one that first searches in the vicinity of where we would expect the element to have "drifted" to, rather than its initial probe location.

Very little work has been done with respect to the analysis of open addressing hashing schemes with buckets of size b. Larson [55] presents an asymptotic analysis for uniform hashing while Ramakrishna [77] studies random probing but he only gives numerical solutions. For linear probing, Blake and Konheim [9] present an asymptotic analysis, and Mendelson [60] derive exact expressions but only solves them numerically. Knuth [46] presents an approximate analysis (based on the Poisson approximation of the binomial distribution) generalizing the model presented by Schay and Spruth [81]. He completes the ideas introduced by M. Tainiter [85].

1.1.1 General References

There are several good and classical references for different areas related with the research presented in this report.

Two good sources of information for hashing techniques are [46] by D. Knuth and [35] by Gonnet and Baeza-Yates. These books, together with [47] and [48], also describe a wide class of data structures and algorithms related to sorting, searching, selection, arithmetic, random numbers generators and text databases. They also present theoretical results on the complexity of these algorithms.

A good survey about analytic methods for average-case analysis with applications to analyzing sorting algorithms, algorithms on trees, hashing and dynamic algorithms can be found in [87] by Vitter and Flajolet.

Other sources for advanced mathematical methods in the analysis of algorithms are [39, 40, 27, 34].

[33] is a good synthetic presentation of the use of complex analysis to estimate the asymptotic growth of coefficients of generating functions. A source for other methods of asymptotic analysis is the classical book by de Bruijn [20]. This is a very useful problem solving oriented book. More recently, and as an excellent source of information, we have the survey by Odlyzko [65]. For background related with complex analysis one may consult [2, 42].

Finally, we should mention some references related to automatic average-case analysis of algorithms. Flajolet *et al.* [26] present a theoretical framework for a powerful system developed for just such computations [25]. This system, called $\Lambda\Upsilon\Omega$, is oriented to the analysis of an important class of algorithms that operate over decomposable data structures. There is a considerable amount of research devoted to improving the capabilities of this software.

1.2 Organization and Guide for the Reader

The main topic of this report is the introduction of a new mathematical tool that we call the Diagonal Poisson Transform, and its application to the analysis of some linear probing hashing schemes. We also present what we believe to be the first exact analysis of a linear probing hashing scheme with buckets of size b.

In Chapter 2, we describe the basic notation and the mathematical machinery that we are going to use. These tools include probability generating functions, basic binomial coefficient identities, the Bernoulli numbers, the Euler-Maclaurin summation formula, a family of functions called the Q-functions, and multisection of summations. The Stirling numbers of the second kind play an important rôle in our analyses and so, we present their main properties as well as the derivation of new identities related to them. We also present the main ideas of *Singularity Analysis* [31], a technique that is used to find asymptotic expansions of the coefficients of generating functions directly from their singularities. The Cayley tree function is also introduced together with some generalizations of it. These functions are essential in the analysis of linear probing hashing with buckets presented in Chapter 5.

In Chapter 3, we present two standard models that are extensively used in the analysis of hashing algorithms: the *Poisson* model and the *exact filling* model. Actually, these models are deeply related by the Poisson Transform [37]. We present this transform, and prove several important properties of it. However, to perform our analyses we require a new mathematical transform, called the Diagonal Poisson Transform. We show the main properties of the transform and apply it to solve recurrences, find inverse relations and obtain several generalizations of Abel's summation formula.

We follow with the analysis of LCFS hashing with linear probing done in Chapter 4. It was shown in [14] that the Robin Hood linear probing algorithm minimizes the variance of the cost of successful searches for all linear probing algorithms. We prove that the variance of the LCFS scheme is within lower order terms of this optimum. This result also appears in [75]. Chapter 4 concludes with an alternative analysis of the standard linear probing algorithm.

In Chapter 5, we present the first exact analysis of linear probing hashing with buckets. From the generating function for the Robin Hood heuristic, we obtain exact expressions for the cost of successful searches when the table is full. Then, with the help of Singularity Analysis, we find the asymptotic expansion of this cost up to $O((bm)^{-1})$. We also give upper and lower bounds when the table is not full. The technical results of this report conclude with a new approach to study certain recurrences that involve truncated exponentials. A new family of numbers that satisfies a recurrence resembling that of the Bernoulli numbers is introduced. These numbers may prove helpful in studying recurrences involving truncated generating functions.

Finally, we conclude in Chapter 6 with a summary of our results and some suggestions for possible future research.

Chapter 2

Mathematical Background

The happiest moments of my life, as well as the most difficult ones, have been witnessed by her mothering look.

In this chapter we present the mathematical machinery that will be used in our analyses. In Sections 2.1, 2.3 and 2.4 we describe the basic properties we need for the derivation of our results. In Section 2.5 we introduce a family of functions that play a central rôle in our analyses. Finally, in Section 2.6, we describe the Stirling numbers of the second kind, and we prove some important lemmata that will be used in Chapter 4.

2.1 Mathematical Notation

We use the now standard notation for asymptotic analysis, introduced by Bachmann in 1894 [5]. Given two functions $f, g : N \to R$, we say that $f(n) = O(g(n))$ if there exists a constant $C > 0$ and $n_0 \in N$ such that

$$\mid f(n) \mid \leq C \mid g(n) \mid \qquad \text{for all } n \geq n_0. \tag{2.1}$$

We also use the "little oh" notation introduced by Landau [54], saying that $f(n) = o(g(n))$ if for each constant $C > 0$, there exists $n_C > 0$ such that

$$\mid f(n) \mid\leq C \mid g(n) \mid \qquad \text{for all } n \geq n_C. \tag{2.2}$$

We assume the reader is familiar with the O notation and the manipulation of such terms. A good introduction to this topic can be found in [39].

Given a function $F(x_1, \ldots, x_n, z)$ we use the following operators:

$$\mathbf{U}_z F(x_1, \ldots, x_n, z) = F(x_1, \ldots, x_n, 1) \qquad \text{(unit)}, \tag{2.3}$$

and

$$\mathbf{D}_z^k F(x_1, \ldots, x_n, z) = \frac{\partial^k F(x_1, \ldots, x_n, z)}{\partial z^k} \qquad \text{(differentiation)} \tag{2.4}$$

The Bernoulli numbers are denoted by B_k. They are defined by the implicit recurrence relation

$$\sum_{j=0}^{m} \binom{m+1}{j} B_j = [m = 0] \qquad m \geq 0 \tag{2.5}$$

(following the notation presented in [39] we use $[S]$ to represent 1 if S is true, and 0 otherwise). These numbers are named after Jakob Bernoulli who discovered the sum [8]:

$$\sum_{r=0}^{k-1} r^i = \frac{1}{i+1} \sum_{j=0}^{i} \binom{i+1}{j} B_j k^{i+1-j}. \tag{2.6}$$

We obtain an asymptotic in k for fixed i by considering only the term for $j = 0$ in (2.6)

$$\sum_{r=0}^{k-1} r^i = O\left(\frac{k^{i+1}}{i+1}\right). \tag{2.7}$$

These numbers also appear in the Euler-Maclaurin summation formula [22, 59],

$$\begin{aligned}\sum_{a\le k<b} f(k) &= \int_a^b f(x)dx - \frac{1}{2}f(x)\mid_a^b + \sum_{k=1}^{r} \frac{B_{2k}}{(2k)!}\mathbf{D}_x^{2k-1}f(x)\mid_a^b & (2.8)\\ &+ O((2\pi)^{-2r})\int_a^b \mid \mathbf{D}_x^{2r}f(x)\mid dx. & (2.9)\end{aligned}$$

Other properties of the Bernoulli numbers can be found in [39].

The harmonic numbers are denoted by H_m and are defined as

$$H_m = \sum_{k=1}^{m}\frac{1}{k} = \log(m) + \gamma + O\left(\frac{1}{m}\right), \tag{2.10}$$

where $\gamma = .5772156649\ldots$ is Euler's constant.

equally likely to occur, the probability of empty location

2.2 Exponential Generating Functions

Given a sequence f_n, we define its exponential generating function (egf) as $F(z) = \sum_{n\ge 0} f_n \frac{z^n}{n!}$. In our analyses we use an important convolution formula for egf's. If $F(z)$ and $G(z)$ are the egf's for the sequences f_n and g_n, then $H(z) = F(z)G(z)$ is the egf for the sequence

$$h_n = \sum_k \binom{n}{k} f_k g_{n-k} \tag{2.11}$$

In Section 5.7 we work with truncated exponential generating functions. We define

$$[A(z)]_n \equiv \sum_{k=0}^{n} a_k \frac{z^k}{k!} \tag{2.12}$$

(we use $\equiv$, to define functions).

2.3 Probability Generating Functions

If X is an integer-valued random variable, denote $p_i = \text{Prob}[X = i], i = 1 \ldots n$. The generating function for the probability distribution p_i is defined by

$$P_{m,n}(z) = \sum_{i \geq 0} p_i z^i. \tag{2.13}$$

We use the following well known properties of generating functions [39]:

$$E[X] = \mathbf{U}_z \mathbf{D}_z P_{m,n}(z), \tag{2.14}$$

$$V[X] = \mathbf{U}_z \mathbf{D}_z^2 P_{m,n}(z) + E[X] - E[X]^2, \tag{2.15}$$

where $E[X]$ and $V[X]$ are the expected value and the variance of X respectively.

If $f(z) = \sum_{n \geq 0} f_n z^n$, then $[z^n]f(z) \equiv f_n$.

2.4 Binomial Coefficients

The *binomial coefficients* are defined by

$$\binom{r}{k} = \begin{cases} \frac{r^{\underline{k}}}{k!} & \text{integer k} \geq 0\text{, real r} \\ 0 & \text{integer k} < 0 \end{cases} \tag{2.16}$$

where $r^{\underline{k}}$ is the k^{th} *falling factorial power* of r, defined as

$$r^{\underline{k}} = r(r-1)\ldots(r-k+1) \quad \text{real } r, \text{ integer } k \geq \mathit{0} \tag{2.17}$$

We list here some useful properties of the binomial coefficients [39]. Let n, k, m be integers and r real. Then,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (n \geq k \geq 0) \tag{2.18}$$

$$\binom{n}{k} = 0 \qquad (k < 0) \tag{2.19}$$

$$\binom{n}{k} = \binom{n}{n-k} \qquad (n \geq 0) \tag{2.20}$$

$$\binom{r}{k} = \frac{r}{k}\binom{r-1}{k-1} \qquad (k > 0) \tag{2.21}$$

$$\binom{r}{k} = \binom{r-1}{k} + \binom{r-1}{k-1} \tag{2.22}$$

$$\binom{r}{k} = (-1)^k \binom{k-r-1}{k} \tag{2.23}$$

$$\binom{r}{m}\binom{m}{k} = \binom{r}{k}\binom{r-k}{m-k} \tag{2.24}$$

$$\sum_k \binom{r}{k} x^k y^{r-k} = (x+y)^r \tag{2.25}$$

$$\sum_{k \le n} \binom{r+k}{k} = \binom{r+k+1}{n} \tag{2.26}$$

$$\sum_{k \le n} (-1)^k \binom{r}{k} = (-1)^n \binom{r-1}{n} \tag{2.27}$$

$$\sum_{0 \le k \le n} \binom{k}{m} = \binom{n+1}{m+1} \qquad (m, n \ge 0) \tag{2.28}$$

$$\sum_{n \ge 0} \binom{n+m}{n} z^m = \frac{1}{(1-z)^{m+1}} \tag{2.29}$$

$$\tag{2.30}$$

We use the notation (i, j) for the "symmetric binomial coefficients" introduced by Comtet [19], defined as

$$(i, j) = \binom{i+j}{j} = \binom{i+j}{i} \tag{2.31}$$

2.5 The Q functions

The Q functions are a family of sums of the form

$$Q_r(m, n) = \sum_{i \ge 0} (i, r) \frac{n^{\underline{i}}}{m^i}. \tag{2.32}$$

In [49] a more general class of Q functions is presented, several properties are proved, and a Q-Algebra is defined. These generalized Q functions play a central rôle in the analysis of hashing with linear probing [46], representation of equivalence relations [51], interleaved memory [50], counting of labelled trees [62], optimal caching [49] and random mappings [47, 11].

Some useful properties of the Q functions are [14]:

$$Q_r(m, n) = Q_{r-1}(m, n) + \frac{n}{m} Q_r(m, n-1) \tag{2.33}$$

(This comes from the fact that $(i,r) = (i-1,r) + (i,r-1)$).

$$Q_{-1}(m,n) = 1 \tag{2.34}$$

$$Q_r(m,n) = \frac{m}{r}(Q_{r-1}(m,n+1) - Q_{r-1}(m,n)) \tag{2.35}$$

(This is a consequence of $\Delta n^{\underline{i}} \equiv (n+1)^{\underline{i}} - n^{\underline{i}} = in^{\underline{i-1}}$).

$$Q_r(m,m-1) = \frac{m}{r}Q_{r-2}(m,m) \tag{2.36}$$

(This is a consequence of (2.33) and (2.35). In particular, given (2.34), it implies that $Q_1(m,m-1) = m$).

$$Q_0(m,m-1) = \frac{\sqrt{2\pi}}{2}\sqrt{m} - \frac{1}{3} + \frac{\sqrt{2\pi}}{24}m^{-1/2} - \frac{4}{135m} + O(m^{-3/2}) \tag{2.37}$$

(The proof of this expansion can be found in [48]).

For fixed $\alpha, 0 \leq \alpha < 1$, we have the expansions:

$$Q_r(m,\alpha m) = \frac{1}{(1-\alpha)^{r+1}} - \frac{(r+1)(r+2)\alpha}{2(1-\alpha)^{r+3}}m^{-1} + O(m^{-2}) \tag{2.38}$$

$$Q_r(m,\alpha m - 1) = \frac{1}{(1-\alpha)^{r+1}} - \frac{(r+1)(r\alpha+2)}{2(1-\alpha)^{r+3}}m^{-1} + O(m^{-2}). \tag{2.39}$$

An asymptotic series for $Q_0(m,m-1)$ was first derived by Ramanujan [78, 79]. The function $Q_0(m,m-1)$ is also known as the Ramanujan's Q function. A detailed analysis of it is found in [28].

2.6 Stirling Numbers of the Second Kind

The Stirling numbers of the second kind count all the possible ways of partitioning a set of n elements into k nonempty subsets without distinguishing between the subsets. Following the notation of [39], we denote these numbers by $\left\{ {n \atop k} \right\}$. They are named after James Stirling (1692-1770). These are some of their properties for m, n, k non negative integers [39]:

$$\left\{ {n \atop 0} \right\} = [n = 0] \tag{2.40}$$

$$\left\{ {n \atop k} \right\} = \left\{ {n-1 \atop k-1} \right\} + k\left\{ {n-1 \atop k} \right\} \tag{2.41}$$

$$\left\{ {n \atop k} \right\} = 0 \quad \text{if} \quad k > n \tag{2.42}$$

$$\left\{ {n \atop n} \right\} = 1 \tag{2.43}$$

$$\left\{ {n+1 \atop n} \right\} = \binom{n+1}{2} \tag{2.44}$$

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} k^m = (-1)^n n! \left\{ {m \atop n} \right\} \quad m \geq 0 \tag{2.45}$$

$$\sum_{k=0}^{n} \left\{ {k \atop m} \right\} \binom{n}{k} = \left\{ {n+1 \atop m+1} \right\} \tag{2.46}$$

$$\sum_{k=0}^{m} k \left\{ {k+n \atop k} \right\} = \left\{ {m+n+1 \atop m} \right\} \tag{2.47}$$

We also need to prove the following lemma:

Lemma 2.1

$$\left\{ {n+2 \atop n} \right\} = 3\binom{n+3}{4} - 2\binom{n+2}{3}. \tag{2.48}$$

Proof:

Using properties (2.47) and (2.44) we find

$$\begin{aligned}
\left\{ {n+2 \atop n} \right\} &= \sum_{k=0}^{n} k \left\{ {k+1 \atop k} \right\} = \sum_{k=0}^{n} k \binom{k+1}{2} && (2.49) \\
&= 3\sum_{k=0}^{n} \frac{(k+2-2)}{3} \binom{k+1}{2} && (2.50) \\
&= 3\sum_{k=0}^{n} \binom{k+2}{3} - 2\sum_{k=0}^{n} \binom{k+1}{2} && (2.51) \\
&= 3\binom{n+3}{4} - 2\binom{n+2}{3}. && (2.52)
\end{aligned}$$

QED

As a consequence, we have the following sums that will prove useful in Chapter 4.

$$\sum_{n \geq 0} \left\{ {n+1 \atop n+1} \right\} x^n = \frac{1}{1-x} \tag{2.53}$$

$$\sum_{n \geq 0} \left\{ {n+2 \atop n+1} \right\} x^n = \frac{1}{(1-x)^3} \tag{2.54}$$

$$\sum_{n\geq 0}\left\{{n+3 \atop n+1}\right\}x^n = \frac{3}{(1-x)^5} - \frac{2}{(1-x)^4} \tag{2.55}$$

More generally, using (2.41), we can prove that $u_p \equiv \sum_{n\geq 0}\left\{{n+1+p \atop n+1}\right\}x^n$ satisfies

$$u_0 = \frac{1}{1-x} \tag{2.56}$$

$$u_p = \frac{1}{1-x}\mathbf{D}_x(xu_{p-1}) \qquad p > 0. \tag{2.57}$$

Lemma 2.2

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k}(k+1)^{n+p} = (-1)^n n!\left\{{n+p+1 \atop n+1}\right\} \qquad p \geq 0. \tag{2.58}$$

Proof:

If we use equations (2.45) and (2.46) then

$$\begin{aligned}
&\sum_{k=0}^{n}(-1)^k\binom{n}{k}(k+1)^{n+p} = \\
&\sum_{k=0}^{n}(-1)^k\binom{n}{k}\sum_{j=0}^{n+p}\binom{n+p}{j}k^j = \sum_{j=0}^{n+p}\binom{n+p}{j}\sum_{k=0}^{n}(-1)^k\binom{n}{k}k^j \\
= \;& (-1)^n n!\sum_{j=0}^{n+p}\left\{{j \atop n}\right\}\binom{n+p}{j} = (-1)^n n!\left\{{n+p+1 \atop n+1}\right\}.
\end{aligned} \tag{2.59}$$

$\mathcal{QED}$

Lemma 2.3

$$\sum_{k\geq 0}e^{-(k+1)x}\frac{(k+1)^{k+p}}{k!}x^k = \sum_{n\geq 0}\left\{{n+p+1 \atop n+1}\right\}x^n \qquad p \geq 0. \tag{2.60}$$

Proof:

We use the Taylor expansion of the exponential and Lemma 2.2. Hence

$$\begin{aligned}
\sum_{k\geq 0}e^{-(k+1)x}\frac{(k+1)^{k+p}}{k!}x^k &= \sum_{k\geq 0}\frac{(k+1)^{k+p}}{k!}x^k\sum_{j\geq 0}(-1)^j\frac{(k+1)^j}{j!}x^j \qquad (2.61)\\
\{\text{letting } n = j+k\} \qquad &= \sum_{n\geq 0}\frac{(-1)^n}{n!}x^n\sum_{k=0}^{n}(-1)^k\binom{n}{k}(k+1)^{n+p}
\end{aligned}$$

$$= \sum_{n\geq 0} \left\{ {n+p+1 \atop n+1} \right\} x^n. \tag{2.62}$$

$\mathcal{QED}$

We will also require an analogous formula when $p = -1$. In this case Lemma 2.2 does not hold for $n = 0$, because $n + p = -1 < 0$, and so (2.46) is not valid. However, the following lemma holds:

Lemma 2.4

$$\sum_{k\geq 0} e^{-(k+c)x} \frac{(k+c)^{k-1}}{k!} x^k = \frac{1}{c}. \tag{2.63}$$

Proof:

This proof is similar to the one of Lemma 2.3, but we must take care when $n = 0$.

$$\begin{aligned}
\sum_{k\geq 0} e^{-(k+c)x} \frac{(k+c)^{k-1}}{k!} x^k &= \sum_{k\geq 0} \frac{(k+c)^{k-1}}{k!} x^k \sum_{j\geq 0} (-1)^j \frac{(k+c)^j}{j!} x^j && (2.64)\\
\{\text{letting } n = j+k\} &= \sum_{n\geq 0} \frac{(-1)^n}{n!} x^n \sum_{k=0}^{n} (-1)^k \binom{n}{k} (k+c)^{n-1} \\
&= \frac{1}{c} + \sum_{n\geq 1} \frac{(-1)^n}{n!} x^n \sum_{k=0}^{n} (-1)^k \binom{n}{k} (k+c)^{n-1} \\
&= \frac{1}{c} + \sum_{n\geq 1} \left\{ {n \atop n+1} \right\} x^n = \frac{1}{c}, && (2.65)
\end{aligned}$$

where the last equality holds by (2.42). $\mathcal{QED}$

Lemma 2.5

$$\sum_{k\geq 0} e^{-kx} \frac{k^{k+p}}{k!} x^k = \sum_{n\geq 0} \left\{ {n+p \atop n} \right\} x^n \quad p \geq 0. \tag{2.66}$$

Proof:

The Taylor expansion of the exponential and (2.45) give

$$\begin{aligned}
\sum_{k\geq 0} e^{-kx} \frac{k^{k+p}}{k!} x^k &= \sum_{k\geq 0} \frac{k^{k+p}}{k!} x^k \sum_{j\geq 0} (-1)^j \frac{k^j}{j!} x^j && (2.67)\\
\{\text{letting } n = j+k\} &= \sum_{n\geq 0} \frac{(-1)^n}{n!} x^n \sum_{k=0}^{n} (-1)^k \binom{n}{k} k^{n+p}
\end{aligned}$$

$$= \sum_{n\geq 0} \left\{ {n+p \atop n} \right\} x^n. \tag{2.68}$$

$\mathcal{QED}$

When $p=-1$, the following lemma holds.

Lemma 2.6

$$\sum_{k\geq 1} e^{-kx} \frac{k^{k-1}}{k!} x^k = x. \tag{2.69}$$

Proof:

Again, the Taylor expansion of the exponential and (2.45) give

$$\begin{aligned} \sum_{k\geq 1} e^{-kx} \frac{k^{k-1}}{k!} x^k &= \sum_{k\geq 1} \frac{k^{k-1}}{k!} x^k \sum_{j\geq 0} (-1)^j \frac{k^j}{j!} x^j & (2.70) \\ \{\text{letting } n = j+k\} \qquad &= \sum_{n\geq 1} \frac{(-1)^n}{n!} x^n \sum_{k=1}^{n} (-1)^k \binom{n}{k} k^{n-1} \\ &= x + \sum_{n\geq 2} \left\{ {n-1 \atop n} \right\} x^n = x. & (2.71) \end{aligned}$$

$\mathcal{QED}$

Knuth, in [49], presents other useful properties of these numbers.

$$\sum_{k\geq 0} k \left\{ {k+r-1 \atop k} \right\} \frac{n^{\underline{k}}}{n^k} = n^r, \tag{2.72}$$

and for fixed m

$$\left\{ {k+m \atop k} \right\} = \frac{k^m}{2^m m!} + O\left(k^{2m-1}\right). \tag{2.73}$$

2.7 Asymptotic Analysis

Some of the problems we present in this report give rise to very complicated asymptotic analyses. Fortunately, there exist fairly synthetic and powerful methods that permit us to extract the asymptotic form of the coefficients of some complicated generating functions directly from their singularities.

These methods originated in the work of Darboux in the last century [66]. We will use the *Singularity Analysis* approach by Flajolet and Odlyzko [31, 64, 29].

Their main idea, is to show that it is sufficient to determine local asymptotic expansions near a singularity, and such expansions can be "transferred" to coefficients. A detailed presentation of this method can be found in [31] and [33]. This technique applies to algebraic-logarithmic functions whose singular expansions involve fractional powers and logarithms. One of the important features of the method, is that it requires only local asymptotic properties of the function to be analyzed. Therefore, it is very suitable for functions that are only indirectly accessible through functional equations, as for example the Cayley generating function.

One of their results that we will use is

Theorem 2.1 (Singularity Analysis) *Let $f(z)$ be a function analytic in a domain*

$$\mathcal{D} = \{z : \ \mid z \mid \leq s_1, \mid Arg(z-s) \mid > \frac{\pi}{2} - \eta\}, \tag{2.74}$$

where s, $s_1 > s$, and η are three positive real numbers. Assume that, with $\sigma(u) = u^{\alpha} log^{\beta}(u)$ and $\alpha \notin \{0, -1, -2, \ldots\}$, we have

$$f(z) \sim \sigma\left(\frac{1}{1-z/s}\right) \quad as \quad z \to s \in \mathcal{D}. \tag{2.75}$$

Then, the Taylor coefficients of f(z) satisfy

$$[z^n]f(z) \sim s^{-n}\frac{\sigma(n)}{n\Gamma(\alpha)}. \tag{2.76}$$

So, for example [30], if we use Theorem 2.1 we have

$$[z^n]\frac{1}{\sqrt{1-4z}}\sqrt{\frac{1}{4z}\log\frac{1}{1-4z}} \sim \frac{4^n}{\sqrt{\pi n}}\sqrt{\log n} \tag{2.77}$$

2.8 Lagrange Inversion Formula

This inversion formula is very useful for solving certain kinds of functional equations, and in some cases gives explicit solutions. There is an immense literature on this problem, and here we only present the main theorem. Lagrange first presented this formula in 1770 [21] and also mentions it in [53]. These references were taken from [19]. We present here the formulation given in [34]

Theorem 2.2 *Let $\phi(u) = \sum_{j=0}^{\infty} \phi_j u^j$ be a formal power series with $\phi_0 \neq 0$, and let $Y(z)$ be the unique formal power series solution of the equation $Y = z\phi(Y)$. The coefficients of Y, Y^k, and $\psi(Y)$ (for an arbitrary series ψ) are given by*

$$[z^n]Y(z) \quad = \frac{1}{n}\left[u^{n-1}\right](\phi(u))^n \tag{2.78}$$

$$[z^n]\,Y^k(z) \quad = \frac{k}{n}\left[u^{n-k}\right](\phi(u))^n \tag{2.79}$$

$$[z^n]\,\psi(Y(z)) = \frac{1}{n}\left[u^{n-1}\right](\phi(u))^n\,\mathbf{D}_u\psi(u). \tag{2.80}$$

2.9 Generalizations of the Cayley Tree Function

In Chapter 5 we require several generalizations of the function $f(z)$, defined implicitly by $f(z) = ze^{f(z)}$. This function appears in problems related with the counting of rooted labelled trees [34, 38, 88]. A standard application of the Lagrange Inversion Formula [34, 38, 88], shows that we can write $f(z)$ as

$$f(z) = \sum_{k\geq 1}\frac{k^{k-1}}{k!}z^k \tag{2.81}$$

Following the notation presented in [18], we define

$$f_p(z) = \sum_{k\geq 1}\frac{k^{k+p}}{k!}z^k \quad \text{and} \quad g_{q,y}(z) = \sum_{k\geq 0}\frac{(y+k)^{k+q}}{k!}z^k \tag{2.82}$$

When $p = 0$, then it is convenient to begin the summation for $f_p(z)$ at $k = 0$ rather than $k = 1$, so that the constant coefficient is 1. Therefore, the Cayley function $f(z)$ is $f_{-1}(z)$. The two most important identities we need are [18]

$$z\mathbf{D}_z f(z) \quad = \quad \frac{f(z)}{1-f(z)} = \frac{1}{1-f(z)} - 1 \tag{2.83}$$

and

$$g_{y,0}(z) \quad = \quad \left(\frac{f(z)}{z}\right)^y \frac{1}{1-f(z)} \tag{2.84}$$

If we notice that $z\mathbf{D}_z f_p(z) = f_{p+1}(z)$, then by iteration of (2.83), we can write the functions $f_p(z)$, as combinations of powers of $1/(1-f(z))$.

With the help of the Implicit Function Theorem [23], and the functional equation that defines $f(z)$, it is shown in [33, 18] that

Lemma 2.7 *The function $f(z)$ has a dominant singularity at $z_0 = 1/e$, and its singular expansion at z_0 is*

$$f(z) = 1 - 2^{1/2}\sqrt{1-ez} + \frac{2}{3}(1-ez) + O((1-ez)^{3/2}) \tag{2.85}$$

Following the notation given in [18], we write $\delta = 2^{1/2}\sqrt{1-ez}$.

Therefore, by Theorem 2.1, using (2.83) and (2.84), we are able to find asymptotic expansions for the family of generating functions $f_p(z)$ and $q_{q,y}(z)$.

If we use the Stirling formula and the binomial theorem, we find that [18]

$$\left[\frac{z^n}{n!}\right]\delta^{-s} \sim \frac{\sqrt{\pi} n^{n+\frac{s-1}{2}}}{\Gamma\left(\frac{s}{2}\right) 2^{\frac{s-1}{2}}}\left(1+\frac{3s^2-6s+2}{24n}+O\left(\frac{1}{n^2}\right)\right) \tag{2.86}$$

Equation (2.86) is valid for all values of s, provided we define $1/\Gamma(-k) = 0$, for k a positive natural number.

2.10 Multisection of Series

Let $A(z) = \sum_{k\geq 0} a_k z^k$. Sometimes, we do not want the generating function of a_k, but rather the generating function of a_{bk+t}, for some fixed $b > 0$ and $0 \leq t < b$. Therefore, we want $\mathcal{A}_{b,t}(z) = \sum_{k\geq 0} a_{bk+t} z^{bk+t}$.

Let $r = e^{\frac{2\pi \mathbf{i}}{b}}$, where $\mathbf{i} = \sqrt{-1}$. That is, r, is a primitive b-th root of unity. Then, we can write [19, 80]

$$\mathcal{A}_{b,r}(z) = \frac{1}{b}\sum_{j=0}^{b-1} r^{-tj} A\left(r^j z\right) \tag{2.87}$$

or, equivalently

$$\sum_{k\geq 0} a_{bk+t} z^{bk+t} = \frac{1}{b}\sum_{j=0}^{b-1} e^{-\frac{2\pi \mathbf{i}}{b}tj} A\left(e^{\frac{2\pi \mathbf{i}}{b}j} z\right) \tag{2.88}$$

Therefore, if we know local asymptotic expansions for $A(z)$ near its dominant singularities, then, by (2.88), we can use singularity analysis to find the asymptotics of a_{bk+d}, when k goes to infinity.

We use this multisection approach to some generalizations of the Cayley generating function in Chapter 5.

Chapter 3

The Diagonal Poisson Transform

I have had several night walks with Manuelita, and often our celestial mother was illuminating us with her sweet light.

3.1 The Poisson Transform

There are two standard models that are extensively used in the analysis of hashing algorithms: the *exact filling* model and the *Poisson filling* model.

Under the exact filling model, we have a fixed number of keys, n, that are distributed among m locations, and all m^n possible arrangements are equally likely to occur.

Under the Poisson model, we assume that each location receives a number of keys that is Poisson distributed with parameter x, and is *independent* of the number of keys going elsewhere. This implies that the total number of keys, N, is itself a Poisson distributed random variable with parameter mx.

$$\text{Prob}\,[N = n] = \frac{e^{-mx}(mx)^n}{n!} \qquad n \;=\; 0, 1, \ldots \tag{3.1}$$

This model was first considered in hashing analysis by Fagin *et al.* [24] in 1979.

It is generally agreed that the Poisson model is simpler to analyze than the exact filling model. The main difference is the fact that in the Poisson model, the number of keys in each location is independent of the number of keys in other places. This is not the case in the exact filling model. Gonnet and Munro in [37], observed that these models are deeply related. They showed that the results from one model can be *transformed* into the other, and that this transformation can be inverted.

Consider a hash table of size m with n elements. Let P be a property (e.g. cost of a successful search) of a random element of the table, and $f(m, n)$ be the result of applying a linear operator f (e.g. an expected value) to the probability generating function of P that was found using the exact filling model. Then $\tilde{f}_m(x)$, the result of computing the same linear operator f to the probability generating function of P computed using a model with m random independent Poisson distributed objects each with parameter x, is

$$\begin{aligned} \tilde{f}_m(x) &= \sum_{n \geq 0} f(m, n) Pr\{N = n\} \\ &= e^{-mx} \sum_{n=0}^{\infty} f(m, n) \frac{(mx)^n}{n!} \end{aligned} \tag{3.2}$$

We may use (3.2) to define $\mathcal{P}_m[f(m, n); x]$, the *Poisson transform* (also called *Poisson generating function* [32, 43]) of $f(m, n)$, as

$$\mathcal{P}_m[f(m, n); x] = \tilde{f}_m(x) = e^{-mx} \sum_{n=0}^{\infty} f(m, n) \frac{(mx)^n}{n!} \tag{3.3}$$

If $\mathcal{P}_m[f(m, n); x]$ has a MacLaurin expansion in powers of x, then we can retrieve the original sequence $f(m, n)$ by the following inversion theorem [37]:

Theorem 3.1 *If $\mathcal{P}_m[f(m,n);x] = \sum_{i\geq 0} a_i x^i$ is the Poisson transform of f(m,n) then* $f(m,n) = \sum_{i=0}^{\infty} a_i \frac{n^{\underline{i}}}{m^i}$.

This theorem is easily proved by multiplying each side of (3.3) by e^{mx} (or its power series), and equating the powers of x on both sides.

So we can study a hashing problem under the more convenient model, and then transfer the results to the other by using the Poisson transform or its inverse.

The results obtained under the Poisson filling model can also be interpreted as an approximation of those one would obtain under the exact filling model, if $n = mx$. This approximation can be formalized by means of an asymptotic expansion. Poblete, in [72], presents an approximation theorem and gives an explicit form for all the terms of the expansion.

Theorem 3.2 *For $x = n/m$,*

$$f(m,n) = \tilde{f}_m(x) + \sum_{j\geq 1} \left(\frac{1}{n}\right)^j \sum_{i\geq 2} c_{i,j} x^i \tilde{f}_m^{(i)}(x). \tag{3.4}$$

Here

$$c_{i,j} = \frac{1}{i!} \sum_{k\geq 0} (-1)^{i-k+j} \binom{j}{k} \left[{k \atop k-j} \right] \tag{3.5}$$

and $\tilde{f}_m^{(i)}(x) = D^i \tilde{f}_m(x)$

where $\left[{k \atop k-j} \right]$ denotes the Stirling numbers of the first kind.

For most situations, this approximation is satisfactory. However, it cannot be used when we have a full, or almost full table (x is very close to 1).

Some of the transforms presented in [37] are

$$\mathcal{P}_m[f(m,n);x] = \tilde{f}_m(x) = e^{-mx} \sum_{n=0}^{\infty} f(m,n) \frac{(mx)^n}{n!} \tag{3.6}$$

$$\mathcal{P}_m[\alpha f(m,n) + \beta g(m,n);x] = \alpha \mathcal{P}_m[f(m,n);x] + \beta \mathcal{P}_m[g(m,n);x] \tag{3.7}$$

α, β constants

$$\mathcal{P}_m[1;x] = 1 \tag{3.8}$$

$$\mathcal{P}_m\left[\frac{n^{\underline{k}}}{m^k};x\right] = x^k \tag{3.9}$$

$$\mathcal{P}_m[Q_r(m,n);x] = \frac{1}{(1-x)^{r+1}} \tag{3.10}$$

$$\mathcal{P}_m[m(f(m,n+1) - f(m,n));x] = \mathbf{D}_x \mathcal{P}_m[f(m,n);x] \tag{3.11}$$

$$\mathcal{P}_m\left[\frac{1}{m}\sum_{k=0}^{n-1} f(m,k);x\right] = \int_0^x \mathcal{P}_m[f(m,n);t]dt \tag{3.12}$$

We require several new transformations.

Theorem 3.3 *The following properties of the Poisson Transform hold:*

$$e^{-x}\mathcal{P}_m[f(m,n);x] = \mathcal{P}_{m+1}\left[\left(\frac{m}{m+1}\right)^n f(m,n);x\right] \tag{3.13}$$

$$e^{x}\mathcal{P}_m[f(m,n);x] = \mathcal{P}_{m-1}\left[\left(\frac{m}{m-1}\right)^n f(m,n);x\right] \tag{3.14}$$

$$\mathcal{P}_m\left[\frac{f(m,n+1)}{n+1};x\right] = \frac{1}{mx}\left(\mathcal{P}_m[f(m,n);x] - f(m,0)e^{-mx}\right) \tag{3.15}$$

$$\mathcal{P}_m\left[\frac{1}{n+1}\sum_{k=0}^{n} f(m,k);x\right] = \frac{1}{x}\int_0^x \mathcal{P}_m[f(m,n);t]dt \tag{3.16}$$

$$\mathcal{P}_m\left[n^{\underline{k}} f(m,n-k);x\right] = (mx)^k\mathcal{P}_m[f(m,n);x] \tag{3.17}$$

$$\mathcal{P}_m\left[\binom{n}{k} f(m,n-k);x\right] = \frac{(mx)^k}{k!}\mathcal{P}_m[f(m,n);x] \tag{3.18}$$

$$\mathcal{P}_m\left[c^n f(m,n);x\right] = e^{(c-1)mx}\mathcal{P}_m[f(m,n);cx] \tag{3.19}$$

$$\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k} f(m,n-k);x\right] = e^{mx}\mathcal{P}_m[f(m,n);x] \tag{3.20}$$

$$\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k} f(m,k)g(m,n-k);x\right] = e^{mx}\mathcal{P}_m[f(m,n);x]\mathcal{P}_m[g(m,n);x] \tag{3.21}$$

$$\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k} p^k f(m,k)q^{n-k}g(m,n-k);x\right] = \underset{p+q=1}{\mathcal{P}_m[f(m,n);px]\mathcal{P}_m[g(m,n);qx]} \tag{3.22}$$

$$\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k} p^k f(pm,k)q^{n-k}g(qm,n-k);x\right] = \underset{p+q=1}{\mathcal{P}_{pm}[f(pm,n);x]\mathcal{P}_q[g(qm,n);x]} \tag{3.23}$$

Proof: These proofs are based on the definition of the Poisson Transform.
<u>(3.13):</u>

$$\begin{aligned} e^{-x}\mathcal{P}_m[f(m,n);x] &= e^{-(m+1)x}\sum_{n=0}^{\infty}\left(\frac{m}{m+1}\right)^n f(m,n)\frac{(m+1)^n}{n!}x^n \\ &= \mathcal{P}_{m+1}\left[\left(\frac{m}{m+1}\right)^n f(m,n);x\right] \end{aligned}$$

(3.14):

$$\begin{aligned}
e^{x}\mathcal{P}_m[f(m,n);x] &= e^{-(m-1)x}\sum_{n=0}^{\infty}\left(\frac{m}{m-1}\right)^{n} f(m,n)\frac{(m-1)^n}{n!}x^n \\
&= \mathcal{P}_{m-1}\left[\left(\frac{m}{m-1}\right)^{n} f(m,n);x\right]
\end{aligned}$$

(3.15):

$$\begin{aligned}
\mathcal{P}_m\left[\frac{f(m,n+1)}{n+1};x\right] &= e^{-mx}\sum_{n=0}^{\infty}\frac{f(m,n+1)}{n+1}\,\frac{(mx)^n}{n!} \\
&= \frac{e^{-mx}}{mx}\sum_{n=1}^{\infty} f(m,n)\frac{m^n}{n!}x^n \\
&= \frac{1}{mx}\left(\mathcal{P}_m[f(m,n);x] - f(m,0)e^{-mx}\right)
\end{aligned}$$

(3.16):

It follows directly from (3.12) and (3.15).

(3.17):

$$\begin{aligned}
\mathcal{P}_m\left[n^{\underline{k}}f(m,n-k);x\right] &= e^{-mx}\sum_{n=k}^{\infty} f(m,n-k)\frac{(mx)^n}{(n-k)!} \\
&= (mx)^k e^{-mx}\sum_{n=0}^{\infty} f(m,n)\frac{(mx)^n}{n!} \\
&= (mx)^k\mathcal{P}_m[f(m,n);x]
\end{aligned}$$

(3.18):

Divide both sides of (3.17) by $k!$.

(3.19):

$$\begin{aligned}
\mathcal{P}_m\left[c^n f(m,n);x\right] &= e^{-mx}\sum_{n=0}^{\infty} f(m,n)\frac{(cmx)^n}{n!} \\
&= e^{(c-1)mx}e^{-m(cx)}\sum_{n=0}^{\infty} f(m,n)\frac{m^n}{n!}(cx)^n \\
&= e^{(c-1)mx}\mathcal{P}_m[f(m,n);cx]
\end{aligned}$$

<u>(3.20):</u>

$$\begin{aligned}
\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k}f(m,n-k);x\right] &= e^{-mx}\sum_{n=0}^{\infty}\sum_{k=0}^{n}\binom{n}{k}f(m,n-k)\frac{(mx)^n}{n!}\\
&= \sum_{k=0}^{\infty}\frac{(mx)^k}{k!}\left(e^{-mx}\sum_{n=k}^{\infty}f(m,n-k)\frac{(mx)^{n-k}}{(n-k)!}\right)\\
&= e^{mx}\mathcal{P}_m[f(m,n);x]
\end{aligned}$$

<u>(3.21):</u>

$$\begin{aligned}
&\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k}f(m,k)g(m,n-k);x\right]\\
&\quad= e^{-mx}\sum_{n=0}^{\infty}\sum_{k=0}^{n}\binom{n}{k}f(m,k)g(m,n-k)\frac{(mx)^n}{n!}\\
&\quad= e^{mx}\left(e^{-mx}\sum_{k=0}^{\infty}f(m,k)\frac{(mx)^k}{k!}\right)\left(e^{mx}\sum_{n=k}^{\infty}g(m,n-k)\frac{(mx)^{n-k}}{(n-k)!}\right)\\
&\quad= e^{mx}\mathcal{P}_m[f(m,n);x]\mathcal{P}_m[g(m,n);x]
\end{aligned}$$

<u>(3.22):</u>

$$\begin{aligned}
&\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k}p^kf(m,k)q^{n-k}g(m,n-k);x\right]\\
&\quad= e^{-m(p+q)x}\sum_{n=0}^{\infty}\sum_{k=0}^{n}\binom{n}{k}p^kf(m,k)q^{n-k}g(m,n-k)\frac{(mx)^n}{n!}\\
&\quad= \left(e^{-mpx}\sum_{k=0}^{\infty}f(m,k)\frac{(mpx)^k}{k!}\right)\left(e^{-mqx}\sum_{n=k}^{\infty}g(m,n-k)\frac{(mqx)^{n-k}}{(n-k)!}\right)\\
&\quad= \mathcal{P}_m[f(m,n);px]\mathcal{P}_m[g(m,n);qx]
\end{aligned}$$

<u>(3.23):</u>

$$\begin{aligned}
&\mathcal{P}_m\left[\sum_{k=0}^{n}\binom{n}{k}p^kf(pm,k)q^{n-k}g(qm,n-k);x\right] =\\
&\quad= e^{-m(p+q)x}\sum_{n=0}^{\infty}\sum_{k=0}^{n}\binom{n}{k}p^kf(pm,k)q^{n-k}g(qm,n-k)\frac{(mx)^n}{n!}\\
&\quad= \left(e^{-mpx}\sum_{k=0}^{\infty}f(pm,k)\frac{(mpx)^k}{k!}\right)\left(e^{-mqx}\sum_{n=k}^{\infty}g(qm,n-k)\frac{(mqx)^{n-k}}{(n-k)!}\right)
\end{aligned}$$

$$= \mathcal{P}_{pm}[f(pm,n);x]\mathcal{P}_{qm}[g(qm,n);x]$$

$\mathcal{QED}$

3.2 The Diagonal Poisson Transform

In Chapter 4, we present a new methodology to study some linear probing hashing algorithms. The main tool in this analysis is the introduction of a new transform which we call the *Diagonal Poisson Transform.* This transform, first introduced by Poblete *et al.* [75], is used in section 3.5 to solve (4.30), the main recurrence of this analysis.

3.2.1 Motivation for the New Transform

Let P be a property (e.g.: cost of a successful search) of a random (but fixed) element $\bullet$ into a table of size m with $n+1$ elements, as is shown in Figure 3.1. Since the table is circular, without loss of generality we may assume that the last location is empty and $\bullet$ is among precisely $i+1$ consecutive occupied locations preceding the last one. Let $f_{m,n}$ be the result of applying a linear operator f (e.g.: an expected value) to the probability generating function of P that was found using the exact filling model.

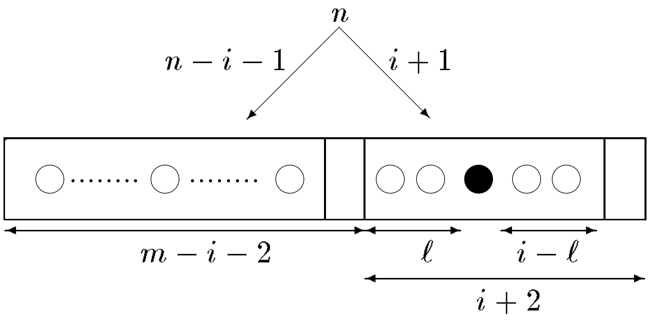

Figure 3.1:

Since f is linear, we can express $f_{m,n}$ as the sum of the following conditional probabilities

$$f_{m,n} = \sum_{i\geq 0} P_{m,n}(B_i) f_{i+2,i} \tag{3.24}$$

where $P_{m,n}(B_i)$= Prob[$\bullet \in$ cluster of size $i+1$].

There are $(m-i-2)^{n-i-1}(m-n-2)$ ways of inserting $n-i-1$ elements in a table of size $m-i-2$ while leaving the last location of the table empty. Furthermore, there are $(i+2)^i$ ways of inserting $i+1$ elements into a table of size $i+2$, so that the last position of the table is empty. Moreover, there are $i+1$ candidates for $\bullet$ and $m^n(n+1)$ ways of inserting the elements in the table. Therefore,

$$f_{m,n} = \sum_{i \geq 0} \binom{n+1}{i+1} \frac{(m-i+2)^{n-i-1}(m-n-2)(i+2)^i(i+1)}{m^n(n+1)} f_{i+2,i} \tag{3.25}$$

If we apply the Poisson Transform to both sides of (3.25) then

$$\begin{aligned}
&\mathcal{P}_m[f_{m,n};x] = \\
= \quad & e^{-mx} \sum_{n=0}^{\infty} \frac{(mx)^n}{n!} \sum_{i \geq 0} \binom{n+1}{i+1} \frac{(m-i-2)^{n-i-1}(m-n-2)(i+2)^i(i+1)}{m^n(n+1)} f_{i+2,i} \\
= \quad & e^{-mx} \sum_{i=0}^{\infty} \frac{(i+2)^i x^i}{i!} f_{i+2,i} \sum_{n \geq i} \frac{x^{n-i}}{(n-i)!} (m-i-2)^{n-i-1}(m-n-2) \\
= \quad & e^{-mx} \sum_{i=0}^{\infty} \frac{(i+2)^i x^i}{i!} f_{i+2,i}(1-x)e^{-(m-i-2)x} \\
= \quad & (1-x) \sum_{i=0}^{\infty} e^{(i+2)x} \frac{(i+2)^i x^i}{i!} f_{i+2,i} \qquad (3.26)
\end{aligned}$$

So, if we define

$$\mathcal{D}_c[f(n);x] = (1-x) \sum_{n \geq 0} e^{-(n+c)x} \frac{((n+c)x)^n}{n!} f(n) \tag{3.27}$$

as a new transform, then $\mathcal{P}_m[f_{m,n};x] = \mathcal{D}_2[f(n+2,n);x]$.

3.2.2 Properties of the Diagonal Poisson Transform

We define $\grave{f}_c(x)$, the Diagonal Poisson Transform of $f(n)$, as

$$\grave{f}_c(x) \equiv \mathcal{D}_c[f(n);x] = (1-x) \sum_{n \geq 0} e^{-(n+c)x} \frac{((n+c)x)^n}{n!} f(n). \tag{3.28}$$

The name diagonal Poisson transform comes from the similarity with the Poisson transform. If we consider an infinite matrix where the rows represent the values of m and the columns represent the values of n, we may easily see the relationship. The Poisson transform has m fixed, while n varies from 0 to infinity; hence, it follows a row of this matrix. The diagonal Poisson transform, has the property that $m - n = c$, where c is a constant. Therefore, it follows a principal diagonal of the matrix. The grave accent in the notation $\grave{f}_c(x)$ was introduced to illustrate this property.

Some useful properties of this transform are:

Theorem 3.4

$$\mathcal{D}_c[\alpha f(n) + \beta g(n); x] = \alpha\ \mathcal{D}_c[f(n); x] + \beta\ \mathcal{D}_c[g(n); x] \quad \alpha, \beta \text{ constants} \tag{3.29}$$

$$\mathcal{D}_c[1; x] = 1 \tag{3.30}$$

$$\mathcal{D}_c\left[\frac{n^{\underline{k}}}{(n+c)^k}; x\right] = x^k \tag{3.31}$$

$$\mathcal{D}_c[Q_r(n+c, n); x] = \frac{1}{(1-x)^{r+1}} \tag{3.32}$$

$$\mathcal{D}_c[(n+1)f(n); x] = \left(1 - c + \frac{c}{1-x}\right)\mathcal{D}_c[f(n); x] + x\mathbf{D}_x\left(\frac{\mathcal{D}_c[f(n); x]}{1-x}\right) \tag{3.33}$$

$$\mathcal{D}_c\left[\frac{f(n)}{n+1}; x\right] = \frac{e^{-(c-1)x}(1-x)}{x}\int_0^x e^{(c-1)t}\mathcal{D}_c[f(n); t]dt \tag{3.34}$$

$$\mathbf{D}_x\left(x^c\frac{\mathcal{D}_c[f(n); x]}{1-x}\right) = x^{c-1}\mathcal{D}_c[(n+c)f(n); x] \tag{3.35}$$

Proof:

For the proofs we just use the definition of the Diagonal Poisson Transform.

(3.29):

$$\begin{aligned}
&\mathcal{D}_c[\alpha f(n) + \beta g(n); x] \\
= \ &(1-x)\sum_{n\geq 0} e^{-(n+c)x}\frac{((n+c)x)^n}{n!}(\alpha f(n) + \beta g(n)) \\
= \ &\alpha\ (1-x)\sum_{n\geq 0} e^{-(n+c)x}\frac{((n+c)x)^n}{n!}f(n) + \beta\ (1-x)\sum_{n\geq 0} e^{-(n+c)x}\frac{((n+c)x)^n}{n!}g(n) \\
= \ &\alpha\ \mathcal{D}_c[f(n); x] + \beta\ \mathcal{D}_c[g(n); x].
\end{aligned}$$

(3.30):

$$\begin{aligned}
\mathcal{D}_c[1; x] &= (1-x)\sum_{n\geq 0} e^{-(n+c)x}\frac{((n+c)x)^n}{n!} \\
&= (1-x)\sum_{n\geq 0}\sum_{k\geq 0}(-1)^k\frac{((n+c)x)^k}{k!}\ \frac{((n+c)x)^n}{n!} \\
\{\text{letting } j = n+k\} \qquad &= (1-x)\sum_{j\geq 0}\frac{(-x)^j}{j!}\sum_{n\geq 0}(-1)^n\binom{j}{n}(n+c)^j.
\end{aligned}$$

For the inner sum, we use (2.45) for $m = j$ and $n = j$, and then

$$\begin{aligned}(1-x)\sum_{j\geq 0}\frac{(-x)^j}{j!}\sum_{n\geq 0}(-1)^n\binom{j}{n}(n+c)^j &= (1-x)\sum_{j\geq 0}\frac{(-x)^j}{j!}(-1)^j j!\left\{{j \atop j}\right\}\\ &= (1-x)\sum_{j\geq 0} x = 1.\end{aligned}$$

(3.31):

$$\begin{aligned}\mathcal{D}_c\left[\frac{n^{\underline{k}}}{(n+c)^k};x\right] &= x^k(1-x)\sum_{n\geq k} e^{-(n+c)x}\frac{((n+c)x)^{n-k}}{(n-k)!}\\ &= x^k(1-x)\sum_{n\geq 0} e^{-(n+k+c)x}\frac{((n+k+c)x)^n}{n!}\\ &= x^k\mathcal{D}_{k+c}[1;x] = x^k,\end{aligned}$$

where the last equality holds by (3.30).
(3.32): By (3.10) and Theorem 3.6 (Transfer Theorem).
(3.33):

$$\begin{aligned}&\left(1-c+\frac{c}{1-x}\right)\mathcal{D}_c[f(n);x] + x\mathbf{D}_x\left(\frac{\mathcal{D}_c[f(n);x]}{1-x}\right)\\ =\ &\left(1-c+\frac{c}{1-x}\right)\mathcal{D}_c[f(n);x] + \sum_{n\geq 0} e^{-(n+c)x}\frac{((n+c)x)^n}{n!}f(n)(n-(n+c)x)\\ =\ &\left(1-c+\frac{c}{1-x}\right)\mathcal{D}_c[f(n);x] + \mathcal{D}_c[(n+c)f(n);x] - \frac{c}{1-x}\mathcal{D}_c[f(n);x]\\ =\ &(1-x)\sum_{n\geq 0} e^{-(n+c)x}\frac{((n+c)x)^n}{n!}f(n)(1-c+n+c)\\ =\ &\mathcal{D}_c[(n+1)f(n);x].\end{aligned}$$

(3.34): This is the inverse relation of (3.33).
(3.35):

$$\begin{aligned}&\mathbf{D}_x\left(x^c\frac{\mathcal{D}_c[f(n);x]}{1-x}\right)\\ =\ &\sum_{n\geq 0} e^{-(n+c)x}\frac{(n+c)^n x^{n+c-1}}{n!}f(n)((n+c)-(n+c)x)\\ =\ &x^{c-1}\mathcal{D}_c[(n+c)f(n);x].\end{aligned}$$

$\mathcal{QED}$

We are now able to prove the Inversion Theorem.

Theorem 3.5 (Inversion Theorem) *If $\mathcal{D}_c[f(n);x] = \sum_{k\geq 0} a_k x^k$ is the diagonal Poisson transform of $f(n)$ then $f(n) = \sum_{k\geq 0} a_k \frac{n^{\underline{k}}}{(n+c)^k}$.*

Proof: By (3.29) and (3.31) we know

$$\mathcal{D}_c\left[\sum_{k\geq 0} a_k \frac{n^{\underline{k}}}{(n+c)^k};x\right] = \sum_{k\geq 0} a_k \mathcal{D}_c\left[\frac{n^{\underline{k}}}{(n+c)^k};x\right] = \sum_{k\geq 0} a_k x^k = \mathcal{D}_c[f(n);x]. \quad (3.36)$$

$\mathcal{QED}$

A useful corollary of the Inversion Theorem is the following inversion formula

Corollary 3.1

$$\frac{(-1)^n}{n!}(n+c)\sum_{k\geq 0}(-1)^k \binom{n}{k}(k+c)^{n-1} b_k = a_n \Leftrightarrow b_n = \sum_{k\geq 0} a_k \frac{n^{\underline{k}}}{(n+c)^k}. \quad (3.37)$$

This inversion formula can be easily checked by finding the Diagonal Poisson Transform of b_n, and considering the coefficients of x^n in the Taylor expansion of this transform.

A very natural question is to characterize the set of functions $f(m,n)$ such that their Poisson Transform coincide with the Diagonal Poisson Transform of $f(n+c,c)$. The functions presented in (3.30)-(3.32), satisfy this condition. The next theorem completely characterizes this set of functions. Therefore we will be able to transfer known properties from one transform to the other.

Theorem 3.6 (Transfer Theorem) *Let $\tilde{a}_m(x) = \mathcal{P}_m[f(m,n);x]$ and $\grave{b}_c(x) = \mathcal{D}_c[f(n+c,n);x]$. Then $\tilde{a}_m(x) = \grave{b}_c(x)$ if and only if $\tilde{a}_m(x)$ does not depend on m.*

Proof: The necessity condition is trivial: if $\tilde{a}_m(x)$ depends on m, then it cannot be equal to $\grave{b}_c(x)$, because the latter does not depend on m.

Now suppose $\tilde{a}_m(x) = \tilde{a}(x)$ and let $\tilde{a}(x) = \sum_{k\geq 0} a_k x^k$ and $\grave{b}_c(x) = \sum_{k\geq 0} b_k x^k$. Then by Theorem 3.1 and the Inversion Theorem,

$$f(m,n) = \sum_{i\geq 0} a_i \frac{n^{\underline{i}}}{m^i} \quad (3.38)$$

and

$$f(n+c,n) = \sum_{i\geq 0} b_i \frac{n^{\underline{i}}}{(n+c)^i}. \quad (3.39)$$

Then, if we substitute $m = n + c$ in (3.38),

$$f(n+c,n) = \sum_{i\geq 0} a_i \frac{n^{\underline{i}}}{(n+c)^i}. \qquad (3.40)$$

Therefore, (3.39) and (3.40) are two expansions for $f(n+c,n)$. Both expansions are rational functions in n with the same denominator. Hence, the numerators should be equal. As both numerators are polynomials in n, their coefficients should be equal. Then, $a_i = b_i$ for $i \geq 0$. As a consequence, $\tilde{a}(x) = \grave{b}_c(x)$. $\mathcal{QED}$

Finally, we would like to find an explicit characterization of the functions that satisfy the Transfer Theorem. This characterization comes as a very nice consequence of Theorem 3.1, the Inversion Theorem, and the Transfer Theorem.

Corollary 3.2 *A function $f(m,n)$ satisfies the conditions of the Transfer Theorem if and only if $f(m,n) = \sum_{k\geq 0} a_k \frac{n^{\underline{k}}}{m^k}$, where the a_k do not depend on m.*

For the case $n = m$, these functions are exactly those studied by Knuth in [49], where he defines a Q-Algebra to study them.

Let $\tilde{a}(x) = \mathcal{P}_m[f(m,n);x]$ and $\grave{b}(x) = \mathcal{D}_c[f(n+c,n);x]$, and then suppose $\tilde{a}(x) = \grave{b}(x)$. If we consider the Taylor expansion of $e^{mx}\tilde{a}(x)$ and $e^{mx}\grave{b}(x)$, then the coefficients of x^n from both expansions should be equal. As a consequence we have the following equation

$$\sum_{k=0}^{n} \frac{m^k}{k!} f(m,k) = \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} (k+c)^k (m-c-k)^{n-k} f(k+c,k) \qquad (3.41)$$

Hence, the functions that satisfy Corollary 3.2 are the solutions of (3.41).

3.3 Generalizations of Abel's formula

In chapter 4 we require some generalizations of Abel's formula

$$\sum_{k\geq 0} \binom{n}{k} (k+c_1)^{k-1} (n-k+c_2)^{n-k} = \frac{(n+c_1+c_2)^n}{c_1} \quad (c_1 \neq 0). \qquad (3.42)$$

We study them with the help of the Diagonal Poisson Transform. After finding the transform of the sum, we use the inversion properties of the Diagonal Poisson Transform to find the final result. Some of these sums have been studied in [80]. They also appear in other fields such as coding theory, pattern matching, data compression, random mappings and multiprocessing systems [45, 86, 44, 58, 11, 18]. Asymptotics for some special cases of these sums have also been studied recently [18, 84].

We now study the first sum:

Lemma 3.1

$$\begin{aligned}
&\mathcal{D}_{c_1+c_2}\left[\frac{1}{(n+c_1+c_2)^n}\sum_{k\geq 0}\binom{n}{k}(k+c_1)^{k+p}(n-k+c_2)^{n-k+q};x\right]\\
=\ &\frac{1}{1-x}\mathcal{D}_{c_1}[(n+c_1)^p;x]\mathcal{D}_{c_2}[(n+c_2)^q;x]. \qquad (3.43)
\end{aligned}$$

Proof: If we use the definition of the Diagonal Poisson Transform, then

$$\begin{aligned}
&\mathcal{D}_{c_1+c_2}\left[\frac{1}{(n+c_1+c_2)^n}\sum_{k\geq 0}\binom{n}{k}(k+c_1)^{k+p}(n-k+c_2)^{n-k+q};x\right]\\
=\ &(1-x)\sum_{n\geq 0}e^{-(n+c_1+c_2)x}\frac{(n+c_1+c_2)^n x^n}{n!}\sum_{k\geq 0}\binom{n}{k}\frac{(k+c_1)^{k+p}(n-k+c_2)^{n-k+q}}{(n+c_1+c_2)^n}\\
=\ &(1-x)\sum_{k\geq 0}e^{-(k+c_1)x}\frac{(k+c_1)^{k+p}x^k}{k!}\sum_{n-k\geq 0}e^{-(n-k+c_2)x}\frac{(n-k+c_2)^{n-k+q}x^{n-k}}{(n-k)!}\\
=\ &(1-x)\sum_{k\geq 0}e^{-(k+c_1)x}\frac{(k+c_1)^{k+p}x^k}{k!}\sum_{n\geq 0}e^{-(n+c_2)x}\frac{(n+c_2)^{n+q}x^n}{n!}\\
=\ &\frac{1}{1-x}\mathcal{D}_{c_1}[(n+c_1)^p;x]\mathcal{D}_{c_2}[(n+c_2)^q;x]. \qquad (3.44)
\end{aligned}$$

$\mathcal{QED}$

If $c_1 = c_2 = 1$ and we use Lemma 2.3, we obtain the following:

Corollary 3.3

$$\begin{aligned}
&\mathcal{D}_2\left[\frac{1}{(n+2)^n}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k+p}(n-k+1)^{n-k+q};x\right]\\
=\ &(1-x)\sum_{n\geq 0}\left\{{n+p+1 \atop n+1}\right\}x^n\sum_{n\geq 0}\left\{{n+q+1 \atop n+1}\right\}x^n \quad (p,q\geq 0). \qquad (3.45)
\end{aligned}$$

When $p = -1$, we use Lemma 2.4, and arrive at:

Corollary 3.4

$$\begin{aligned}
&\mathcal{D}_2\left[\frac{1}{(n+2)^n}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k+q};x\right]\\
=\ &(1-x)\sum_{n\geq 0}\left\{{n+q+1 \atop n+1}\right\}x^n \quad (q\geq 0). \qquad (3.46)
\end{aligned}$$

Moreover, we find Abel's identity by using Lemma 2.4 and Lemma 3.1, for $p = -1$ and $q = 0$.

Corollary 3.5

$$\mathcal{D}_{c_1+c_2}\left[\frac{1}{(n+c_1+c_2)^n}\sum_{k\geq 0}\binom{n}{k}(k+c_1)^{k-1}(n-k+c_2)^{n-k};x\right]=\frac{1}{c_1}\quad c_1\neq 0$$

Another interesting case is obtained when $p = 0$, $q = 0$, $c_1 = 0$, and $c_2 = 0$. Then

$$\mathcal{D}_0\left[\frac{1}{n^n}\sum_{k\geq 0}\binom{n}{k}k^k(n-k)^{n-k};x\right]=\frac{1}{1-x}. \tag{3.47}$$

So after using (3.32) for $c = 0$, we derive the following identity proven by Cauchy [15]

$$\frac{1}{n^n}\sum_{k\geq 0}\binom{n}{k}k^k(n-k)^{n-k}=Q_0(n,n). \tag{3.48}$$

The second sum we have to study is

Lemma 3.2

$$\begin{aligned}&\mathcal{D}_{c_1+c_2}\left[\sum_{k\geq 0}\binom{n}{k}\frac{(k+c_1)^{k+p}(n-k+c_2)^{n-k-q}(n-k)^{\underline{q}}f(n-k-q)}{(n+c_1+c_2)^n};x\right]\\ =\;&\frac{x^q}{1-x}\mathcal{D}_{c_1}[(n+c_1)^p;x]\mathcal{D}_{c_2+q}[f(n);x]\end{aligned} \tag{3.49}$$

Proof: If we use the definition of the Diagonal Poisson Transform and the equality $n! = n^{\underline{q}}(n-q)!$, then

$$\begin{aligned}&\mathcal{D}_{c_1+c_2}\left[\sum_{k\geq 0}\binom{n}{k}\frac{(k+c_1)^{k+p}(n-k+c_2)^{n-k-q}(n-k)^{\underline{q}}f(n-k-q)}{(n+c_1+c_2)^n};x\right]\\ =\;&(1-x)\sum_{n\geq 0}e^{-(n+c_1+c_2)x}\frac{(n+c_1+c_2)^n x^n}{n!}\\ &\qquad\sum_{k\geq 0}\binom{n}{k}\frac{(k+c_1)^{k+p}(n-k+c_2)^{n-k-q}(n-k)^{\underline{q}}f(n-k-q)}{(n+c_1+c_2)^n}\\ =\;&(1-x)\sum_{k\geq 0}e^{-(k+c_1)x}\frac{(k+c_1)^{k+p}x^k}{k!}\\ &\qquad\sum_{n-k\geq 0}e^{-(n-k+c_2)x}\frac{(n-k+c_2)^{n-k-q}(n-k)^{\underline{q}}f(n-k-q)x^{(n-k)}}{(n-k)!}\end{aligned}$$

$$
\begin{aligned}
&= (1-x)\sum_{k\geq 0} e^{-(k+c_1)x}\frac{(k+c_1)^{k+p}x^k}{k!}\sum_{n\geq 0} e^{-(n+c_2)x}\frac{(n+c_2)^{n-q}n^{\underline{q}}f(n-q)x^n}{n!} \\
&= (1-x)\sum_{k\geq 0} e^{-(k+c_1)x}\frac{(k+c_1)^{k+p}x^k}{k!}\sum_{n\geq 0} e^{-(n+c_2+q)x}\frac{(n+c_2+q)^n f(n)x^{n+q}}{n!} \\
&= \frac{x^q}{1-x}\mathcal{D}_{c_1}[(n+c_1)^p;x]\mathcal{D}_{c_2+q}[f(n);x] \qquad (3.50)
\end{aligned}
$$

$\mathcal{QED}$

If c_1=1, then we can use Lemma 2.3, and obtain the following important result.

Corollary 3.6

$$
\begin{aligned}
&\mathcal{D}_{c_2+1}\left[\sum_{k\geq 0}\binom{n}{k}\frac{(k+1)^{k+p}(n-k+c)^{n-k-q}(n-k)^{\underline{q}}f(n-k-q)}{(n+c_2+1)^n};x\right] \\
&= x^q\sum_{n\geq 0}\left\{{n+p+1 \atop n+1}\right\}x^n\mathcal{D}_{c_2+q}[f(n);x] \qquad (3.51)
\end{aligned}
$$

3.4 Inverse Relations

Inverse relations are very important in the study of combinatorial identities. Probably the most remarkable one is the Lagrange inversion formula [48, 19, 38, 34, 88]. This tool is used to solve some functional equations, and in several cases it can give explicit formulae for the solutions. Another famous relation is the Möbius inversion formula, of wide application in number theory [41]. Riordan in [80] presents a very large library of inverse relations that are very general and varied. In this section we show how we can derive some classic and new inverse relations with the use of the Poisson and Diagonal Poisson transforms.

3.4.1 Binomial Transform

If we denote

$$a(m,n) = \sum_k \binom{n}{k}(-1)^k b(m,k) \qquad (3.52)$$

and use (3.20) and then (3.19) for $c=-1$, we have

$$\mathcal{P}_m\left[a(m,n);x\right] = e^{mx}\mathcal{P}_m[(-1)^n b(m,n);x] = e^{-mx}\mathcal{P}_m[b(m,n);-x]. \qquad (3.53)$$

Moreover, if we substitute x by $-x$ in (3.53), we also have the symmetric equality

$$\mathcal{P}_m\left[b(m,n);x\right] = e^{-mx}\mathcal{P}_m[a(m,n);-x] \tag{3.54}$$

So we have easily derived the inversion formulae

$$a(m,n) = \sum_k (-1)^k \binom{n}{k} b(m,k) \tag{3.55}$$

$$\text{and} \qquad b(m,n) = \sum_k (-1)^k \binom{n}{k} a(m,k). \tag{3.56}$$

In [46], Knuth used this relation to define a *transform* that maps sequences of real numbers onto sequences of real numbers. This is called the *Binomial Transform* of $a(m,n)$. Poblete *et al.* [74] developed the theory of this transform, and show how it can be used to analyze the performance of skip lists, a probabilistic data structure introduced by W. Pugh [76, 68]. Several of the properties presented there can be proven using the Poisson Transform.

3.4.2 Abel Inverse Relations

In [80], Riordan presents several Abel inverse relations that are associated with Abel's generalization of the binomial theorem. We can derive some of these relations using the Diagonal Poisson Transform. Furthermore, we present a new class of Abel inverse relations. First we need to prove the following lemma

Lemma 3.3

$$\textit{Let } A(n) = \sum_{k\geq 0}\binom{n}{k}\frac{(k+c_1)^k B(k)(n-k)^{\underline{q}}(n-k+c_2)^{n-k-q}g(n-k-q)}{(n+c_1+c_2)^n} \tag{3.57}$$

$$\textit{then } \mathcal{D}_{c_1+c_2}\left[A(n);x\right] = \frac{x^q}{1-x}\mathcal{D}_{c_1}[B(n);x]\mathcal{D}_{c_2+q}[g(n);x] \tag{3.58}$$

Proof: This proof is very similar to that of Lemma 3.2.

$$\begin{aligned}
&\mathcal{D}_{c_1+c_2}\left[\sum_{k\geq 0}\binom{n}{k}\frac{(k+c_1)^k B(k)(n-k+c_2)^{n-k-q}(n-k)^{\underline{q}}g(n-k-q)}{(n+c_1+c_2)^n};x\right] \\
= \quad & (1-x)\sum_{n\geq 0} e^{-(n+c_1+c_2)x}\frac{(n+c_1+c_2)^n x^n}{n!} \\
&\qquad \sum_{k\geq 0}\binom{n}{k}\frac{(k+c_1)^k B(k)(n-k+c_2)^{n-k-q}(n-k)^{\underline{q}}g(n-k-q)}{(n+c_1+c_2)^n}
\end{aligned}$$

$$\begin{aligned}
&= (1-x)\sum_{k\geq 0} e^{-(k+c_1)x}\frac{(k+c_1)^k x^k}{k!}B(k) \\
&\qquad \sum_{n-k\geq 0} e^{-(n-k+c_2)x}\frac{(n-k+c_2)^{n-k-q}(n-k)^{\underline{q}}g(n-k-q)x^{(n-k)}}{(n-k)!} \\
&= (1-x)\sum_{k\geq 0} e^{-(k+c_1)x}\frac{(k+c_1)^k x^k}{k!}B(k)\sum_{n\geq 0} e^{-(n+c_2)x}\frac{(n+c_2)^{n-q}n^{\underline{q}}g(n-q)x^n}{n!} \\
&= (1-x)\sum_{k\geq 0} e^{-(k+c_1)x}\frac{(k+c_1)^k x^k}{k!}B(k)\sum_{n\geq 0} e^{-(n+c_2+q)x}\frac{(n+c_2+q)^n g(n)x^{n+q}}{n!} \\
&= \frac{x^q}{1-x}\mathcal{D}_{c_1}[B(n);x]\mathcal{D}_{c_2+q}[g(n);x] \qquad (3.59)
\end{aligned}$$

$\mathcal{QED}$

Now suppose we know $\mathcal{D}_{c_2+q}[g(n);x]$. Then, we write the Diagonal Poisson Transform of $B(n)$, as a function of that of $A(n)$, with an identity that resembles (3.58). Let us define $G(n)$ as a function that satisfies

$$\mathcal{D}_{-c_2-q}[G(n);x] = \frac{(1-x)^2}{\mathcal{D}_{c_2+q}[g(n);x]}. \qquad (3.60)$$

So by (3.58) and (3.60) we obtain

$$\mathcal{D}_{c_1}[B(n);x] \;=\; \frac{x^{-q}}{1-x}\mathcal{D}_{c_1+c_2}\,[A(n);x]\,\mathcal{D}_{-c_2-q}[G(n);x]. \qquad (3.61)$$

Then, by Lemma 3.3 we find

$$B(n) = \sum_{k\geq 0}\binom{n}{k}(k+c_1+c_2)^k A(k)\frac{(n-k)^{\underline{-q}}(n-k+c_2)^{n-k+q}G(n-k+q)}{(n+c_1)^n}. \qquad (3.62)$$

The inverse relation is obtained by defining

$$\begin{aligned}
a_n &= (n+c_1+c_2)^n A(n) \\
b_n &= (n+c_1)^n B(n) \\
c_2 &= z
\end{aligned}$$

and substituting these values in (3.62). Therefore, we arrive at

$$a_n = \sum_{k\geq 0}\binom{n}{k}(n-k)^{\underline{q}}\,(n-k+z)^{n-k-q}g(n-k-q)b_k \qquad (3.63)$$

$$\text{and} \qquad b_n = \sum_{k\geq 0}\binom{n}{k}(n-k)^{\underline{-q}}\,(n-k-z)^{n-k+q}G(n-k+q)a_k. \tag{3.64}$$

We obtain several useful special cases for various choices of $g(n)$.

3.4.3 A New Abel Inverse Relation

Consider $g(n) = Q_{r-2}(n+z+q, n)$. Then, by (3.32),

$$\mathcal{D}_{z+q}[g(n); x] \;=\; (1-x)^{-r-1}, \tag{3.65}$$

and therefore

$$\mathcal{D}_{-z-q}[G(n); x] \;=\; \frac{(1-x)^2}{\mathcal{D}_{z+q}[g(n); x]} = (1-x)^{r+1} \tag{3.66}$$

Then, by (3.32), we obtain $G(n) = Q_{-r-2}(n-z-q, n)$. So (3.63) and (3.64) give us the following inversion formulae

$$a_n \;=\; \sum_{k\geq 0}\binom{n}{k}(n-k)^{\underline{q}}\,(n-k+z)^{n-k-q}Q_{r-2}(n-k+z, n-k-q)b_k \tag{3.67}$$

$$b_n \;=\; \sum_{k\geq 0}\binom{n}{k}(n-k)^{\underline{-q}}\,(n-k-z)^{n-k+q}Q_{-r-2}(n-k-z, n-k+q)a_k. \tag{3.68}$$

The most interesting feature of this pair of inverse relations is its symmetry in z, q, and r. Since $Q_{-1}(m,n) = 1$, then for $q = 0$ and $r = 1$ (3.67) and (3.68) simplify to

$$a_n = \sum_{k\geq 0}\binom{n}{k}(n-k+z)^{n-k}b_k \tag{3.69}$$

$$\text{and} \qquad b_n = \sum_{k\geq 0}\binom{n}{k}(z^2-n+k)(n-k-z)^{n-k-2}a_k. \tag{3.70}$$

(3.69) and (3.70) are studied in [80].

We can find more inverse relations by replacing $g(n)$ in (3.58) with other functions whose Diagonal Poisson Transforms are known, and using (3.61).

3.5 Solving Recurrences with the Diagonal Poisson Transform

In the analysis presented in Chapter 4 we require a solution to the recurrence:

$$H_i = B_i + \sum_{k \geq 0} \binom{i}{k} (k+1)^{k+p} (i+d) H_{i-k-1}. \tag{3.71}$$

Writing $h_i = \frac{H_i}{(i+c)^i(i+1)}$ and $b_i = \frac{B_i}{(i+c)^i(i+1)}$ we are to solve

$$\begin{aligned} h_i &= b_i + \sum_{k \geq 0} \binom{i}{k} (k+1)^{k+p} \frac{i+d}{(i+c)^i(i+1)} (i-k+c-1)^{i-k-1}(i-k)h_{i-k-1} \\ &= b_i + \frac{i+d}{i+1} i! \sum_{0 \leq k < i} \frac{(k+1)^{k+p}}{k!} \frac{(i-k+c-1)^{i-k-1}}{(i-k-1)!} \frac{h_{i-k-1}}{(i+c)^i} \\ &= b_i + \left(1 + \frac{d-1}{i+1}\right) a_i, \end{aligned} \tag{3.72}$$

where a_i denotes the factor that multiplies $\frac{i+d}{i+1}$. Applying the diagonal Poisson transform to both sides of (3.72) we get

$$\grave{h}_c(x) = \grave{b}_c(x) + \mathcal{D}_c[a_i; x] + (d-1)\ \mathcal{D}_c\left[\frac{a_i}{i+1}; x\right], \tag{3.73}$$

where (3.73) holds by the linearity property of the transform.

Now, we only have to find the values of $\mathcal{D}_c[a_i; x]$ and $\mathcal{D}_c[\frac{a_i}{i+1}; x]$. For the first transform, we can use Corollary 3.6, for $c_2 = c-1, q = 1$ and $f(n) = h_n$. Then, we have

$$\mathcal{D}_c[a_i; x] \quad = \quad \left(x \sum_{n \geq 0} \left\{ {n+p+1 \atop n+1} \right\} x^n \right) \grave{h}_c(x) = s_p(x)\grave{h}_c(x), \tag{3.74}$$

where $s_p(x)$ denotes the sum involving the Stirling coefficients.

For the second transform, we use (3.34) and (3.74) and obtain

$$\mathcal{D}_c\left[\frac{a_i}{i+1}; x\right] = \frac{e^{-(c-1)x}(1-x)}{x} \int_0^x e^{(c-1)t} s_p(t)\grave{h}_c(t)dt. \tag{3.75}$$

Finally, we arrive at the following integral equation:

$$\grave{h}_c(x) = \grave{b}_c(x) + s_p(x)\grave{h}_c(x) + \frac{(d-1)e^{-(c-1)x}(1-x)}{x} \int_0^x e^{(c-1)t} s_p(t)\grave{h}_c(t)dt. \tag{3.76}$$

After solving the integral equation and using (3.33), we obtain the following solution

$$\grave{h}_c(x) = \frac{(1-x)e^{(d-c)x}}{x^d(1-s_p(x))} e^{(d-1)A(x)} \int_0^x x^{d-1} e^{(c-d)t} e^{-(d-1)A(t)} \mathcal{D}_c[(i+1)b_i; t] dt, \qquad (3.77)$$

where $A(x) = \int_{t=1}^{x} (1-t)/(t(1-s_p(t))dt$.

We use (3.77) to solve (4.30), the main recurrence studied in Chapter 4.

Chapter 4

Analysis of LCFS Hashing with Linear Probing

On January 8, 1995, my wife Graciela returned to Uruguay, and with her went Manuelita.

4.1 Motivation

The simplest collision resolution scheme for open addressing hash tables is *linear probing*, which uses the cyclic probe sequence

$$h(K), h(K)+1, \ldots m-1, 0, 1, \ldots, h(K)-1 \tag{4.1}$$

assuming the table slots are numbered from 0 to $m-1$. Linear probing works reasonably well for tables that are not too full, but as the load factor increases, its performance deteriorates rapidly.

If A_n denotes the number of probes in a successful search in a hash table of n elements (assuming all elements in the table are equally likely to be searched), and if we assume that the hash function h takes all the values in $0 \ldots m-1$ with equal probabilities, then we know from [46, 35]:

$$\mathbf{E}[A_n] = \frac{1}{2}(1 + Q_0(m, n-1)) \tag{4.2}$$

$$\mathbf{V}[A_n] = \frac{1}{3}Q_2(m, n-1) - \frac{1}{4}Q_0(m, n-1)^2 - \frac{1}{12} \tag{4.3}$$

where the functions $Q_i(m,n)$ are a generalization of Ramanujan's Q-function studied in Section 2.5. For a table with $n = \alpha m$ elements, and fixed $\alpha < 1$ and $n, m \to \infty$, these quantities depend (essentially) only on α:

$$\mathbf{E}[A_{\alpha m}] = \frac{1}{2}\left(1 + \frac{1}{1-\alpha}\right) - \frac{1}{2(1-\alpha)^3 m} + O\left(\frac{1}{m^2}\right) \tag{4.4}$$

$$\mathbf{V}[A_{\alpha m}] = \frac{1}{3(1-\alpha)^3} - \frac{1}{4(1-\alpha)^2} - \frac{1}{12} - \frac{1+3\alpha}{2(1-\alpha^5)m} + O\left(\frac{1}{m^2}\right) \tag{4.5}$$

For a full table, these approximations are useless, but the properties of the Q functions can be used to obtain the following expressions:

$$\mathbf{E}[A_m] = \frac{\sqrt{2\pi m}}{4} + \frac{1}{3} + \frac{1}{48}\sqrt{\frac{2\pi}{m}} + O\left(\frac{1}{m}\right) \tag{4.6}$$

$$\mathbf{V}[A_m] = \frac{\sqrt{2\pi m^3}}{12} + \left(\frac{1}{9} - \frac{\pi}{8}\right)m + \frac{13\sqrt{2\pi m}}{144} - \frac{47}{405} - \frac{\pi}{48} + O\left(\frac{1}{\sqrt{m}}\right) \tag{4.7}$$

It is clear from these expressions that not only is the expected search time high, but also the variances are quite large, and therefore the expected value is not a very reliable predictor for the actual running time of a successful search.

It was shown in [14] that the Robin Hood linear probing algorithm minimizes the variance for all linear probing algorithms. This variance, for a full table, is $\Theta(m)$, instead of the $\Theta(m^{3/2})$ of the standard algorithm. They derived the following expressions for the variance of the successful search time:

$$\begin{aligned}
\mathbf{V}[A_n] &= \frac{1}{2}Q_1(m,n-1) - \frac{1}{4}Q_0(m,n-1)^2 - \frac{1}{6}Q_0(m,n-1) + \frac{1}{6}\frac{n-1}{m} - \frac{1}{12} \\
\mathbf{V}[A_{\alpha m}] &= \frac{1}{4(1-\alpha)^2} - \frac{1}{6(1-\alpha)} - \frac{1}{12} + \frac{\alpha}{6} - \frac{1}{6m} - \frac{1+2\alpha}{3(1-\alpha^4)m} + O\left(\frac{1}{m^2}\right) \\
\mathbf{V}[A_m] &= \frac{4-\pi}{8}m + \frac{1}{9} - \frac{\pi}{48} + \frac{1}{135}\sqrt{\frac{2\pi}{m}} + O\left(\frac{1}{m^2}\right) \qquad (4.8)
\end{aligned}$$

In this chapter we study the effect of the LCFS (last-come-first-served) heuristic on the linear probing scheme. Surprisingly, the variance of this scheme is much less than that of the standard first come first served approach and within lower order terms of the minimal (Robin Hood) method. Some of the results presented here also appear in [75].

4.2 Analysis of Last-Come-First-Served Linear Probing Hashing

Consider a hash table of size m, with $n+1$ elements inserted using the last-come-first-served linear probing algorithm. We will consider a randomly chosen element as a "tagged" one, and denote it by $\bullet$. Define $P_{m,n}(z)$ as the probability generating function for the cost of searching for this tagged element. We first derive a recurrence for $P_{m,n}(z)$.

We define an almost full hash table of size m as a hash table of size m with $m-1$ elements inserted in such a way that the last location is empty.

Following the analysis of the standard linear probing algorithm given in [46], we use the function $\hat{f}(m,n)$ to denote the number of ways to create a table of size m, with n elements inserted so that the last location is empty. If all the possible m^n arrangements are equally likely to occur, the probability of empty location being the last is $(1-n/m)$. It follows that

$$\hat{f}(m,n) = m^{n-1}(m-n). \qquad (4.9)$$

Without loss of generality, we may assume that after inserting the first n elements, the hash table is as shown in Figure 4.1, and that as a result of the insertion of the $(n+1)^{st}$ element, the last location of the table is filled. We may see the table as a concatenation of two tables of sizes $m-i-2$ and $i+2$ with $n-i-1$ and $i+1$ elements respectively. We may also assume that $\bullet$ belongs to the last cluster of the hash table. Consider now the insertion of the last element. With probability $1/(n+1)$, this element is $\bullet$, and so its cost is 1 (generating function z). With probability $n/(n+1)$ the new element is not $\bullet$. If we assume this insertion does not force $\bullet$ to move, then we have the recurrence:

$$P_{m,n}(z) = \frac{z}{n+1} + \frac{n}{n+1}P_{m,n-1}(z) \qquad (4.10)$$

We must, of course, include a correction term to account for this shortcoming. As we can see in Figure 4.1, the last insertion increments the cost of searching for • when it maps into any of the first $\ell+1$ positions of the last cluster.

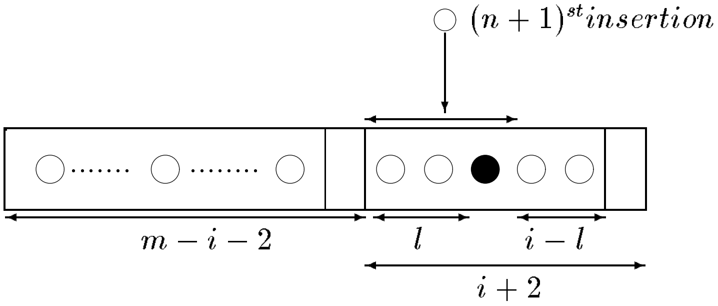

Figure 4.1:

In order to study the correction term, we introduce two auxiliary functions. Given a table of size $\ell+r+2$, we define $F_{\ell,r}(z)$ as the generating function for the number of ways of inserting $\ell+r+1$ elements in the table, where one element is tagged (•) with z keeping track of its cost, such that the rightmost location is empty, and such that there are ℓ elements to the left of • and r elements to its right. Figure 4.2 helps to understand this definition. It is easy to see that if we insert a new element in any of the first $\ell+1$ locations of the table, the cost of • increases by one. By the definition of $F_{\ell,r}(z)$ we know that

$$\mathbf{U}_z F_{\ell,r}(z) = \hat{f}(\ell+r+2, \ell+r+1) = (\ell+r+2)^{\ell+r}. \tag{4.11}$$

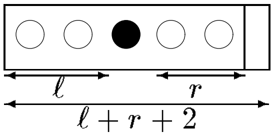

Figure 4.2:

We define $C_i(z)$ as the generating function for the number of ways of inserting $i+1$ elements into a table of size $i+2$, where one element is tagged (•), and such that the rightmost location is empty. z keeps track of the cost of •. Since • may be any of the $i+1$ elements inserted we have

$$C_i(z) = \sum_{\substack{\ell+r=i \\ \ell,r\geq 0}} F_{\ell,r}(z). \tag{4.12}$$

Equations (4.12) and (4.11) imply that

$$\mathbf{U}_z C_i(z) = (i+2)^i (i+1). \tag{4.13}$$

The function $C_i(z)/\mathbf{U}_z C_i(z)$ is the probability generating function for the cost of a successful search for • in an almost full table of size $i+2$. Therefore, by (4.2) we have

$$\frac{\mathbf{U}_z \mathbf{D}_z C_i(z)}{\mathbf{U}_z C_i(z)} = \frac{1}{2}(1 + Q_0(i+2, i)), \tag{4.14}$$

because the expected successful search time for a linear probing scheme is independent of the discipline used to resolve collisions [69, 46].

We now have the tools to find the correction term $T_{m,n}(z)$.

There are $\sum_{\ell+r=i}(\ell+1)F_{\ell,r}(z)$ possibilities that the insertion in an almost full table of size $i+2$ increments the cost of searching for •. Moreover, there are $\hat{f}(m-i-2, n-i-1)$ ways of inserting $n-i-1$ elements in a table of size $m-i-2$, in such a way that the last location in the table is empty. Furthermore, there are $\binom{n}{i+1}$ ways to divide the n inserted elements in two sets of sizes $n-i-1$ and $i+1$. Since this is valid for $0 \le i \le n-1$, we have the following correction term:

$$T_{m,n}(z) = \frac{z-1}{m^n(n+1)} \sum_{i=0}^{n-1} \binom{n}{i+1} \hat{f}(m-i-2, n-i-1) \sum_{\ell+r=i} (\ell+1) F_{\ell,r}(z). \tag{4.15}$$

The increment in cost is 1, therefore we have to use the factor $(z-1)$. Since we are counting number of ways, and want probability generating functions, we have to divide by a normalization factor $m^n(n+1)$: there are m^n ways of inserting n elements in a table of size m, and there are $n+1$ possibilities for the choice of the tagged element. Therefore, if we consider (4.10) and (4.15) together, we have the following recurrence for $P_{m,n}(z)$:

$$P_{m,n}(z) \quad = \quad \frac{z}{n+1} + \frac{n}{n+1} P_{m,n-1}(z) + T_{m,n}(z) \tag{4.16}$$

with $P_{m,0} = z$, as it is the probability generating function for the cost of searching for • when it is the only element in the table. If we define $R_{m,n}(z) \equiv (n+1)P_{m,n}(z)$, then recurrence (4.16) is transformed into the linear recurrence

$$R_{m,n}(z) \quad = \quad R_{m,n-1}(z) + z + (n+1)T_{m,n}(z). \tag{4.17}$$

This leads us to the solution

$$R_{m,n}(z) = (n+1)z + \sum_{k=1}^{n} (k+1) T_{m,k}(z) \tag{4.18}$$

and so,

$$P_{m,n} = z + \frac{1}{n+1}\sum_{k=1}^{n}(k+1)T_{m,k}(z) \tag{4.19}$$

To further simplify (4.19), we need the following lemma:

Lemma 4.1

$$\begin{aligned} S(m,n,i) &\equiv \sum_{k=i+1}^{n}\binom{k}{i+1}\frac{(m-i-2)^{k-i-2}(m-k-1)}{m^k} \\ &= \binom{n+1}{i+2}\frac{(m-i-2)^{n-i-1}}{m^n} \end{aligned} \tag{4.20}$$

Proof:

$$\begin{aligned} S(m,n,i) &= \sum_{k=i+1}^{n}\binom{k}{i+1}\frac{(m-i-2)^{k-i-2}(m-i-2+i+1-k)}{m^k} \\ &= \sum_{k=i+1}^{n}\binom{k}{i+1}\frac{(m-i-2)^{k-i-1}}{m^k} \\ &\quad -\sum_{k=i+1}^{n}\frac{k}{k-i-1}\binom{k-1}{k-i-2}\frac{(m-i-2)^{k-i-2}(k-i-1)}{m^k} \\ &= \sum_{k=i+1}^{n}\binom{k}{i+1}\frac{(m-i-2)^{k-i-1}}{m^k} \\ &\quad -\sum_{k=i+1}^{n-1}(k+1)\frac{k}{i+1}\frac{(m-i-2)^{k-i-1}}{m^{k+1}} \\ &= \sum_{k=i+1}^{n}\binom{k}{i+1}\frac{(m-i-2)^{k-i-1}}{m^k}\frac{(m-k-1)}{m} \\ &\quad +\binom{n+1}{i+2}\frac{(m-i-2)^{n-i-1}}{m^n}\frac{(i+2)}{m} \\ &= \left(1-\frac{i+2}{m}\right)S(m,n,i)+\binom{n+1}{i+2}\frac{(m-i-2)^{n-i-1}}{m^n}\frac{(i+2)}{m}. \end{aligned}$$

So, we have an equation in $S(m,n,i)$, and the lemma follows immediately. $\mathcal{QED}$

Then, if we define $G_i(z) \equiv \sum_{\ell+r=i}(\ell+1)F_{\ell,r}(z)$, using Lemma 4.1 and equations (4.9), (4.15) and (4.19), we find

$$\begin{aligned} P_{m,n}(z) &= z+\frac{z-1}{n+1}\sum_{k=1}^{n}\frac{1}{m^k}\sum_{i=0}^{k-1}\binom{k}{i+1}(m-i-2)^{k-i-2}(m-k-1)G_i(z) \\ &= z+\frac{z-1}{n+1}\sum_{i=0}^{n-1}G_i(z)\sum_{k=i+1}^{n}\binom{k}{i+1}\frac{(m-i-2)^{k-i-2}(m-k-1)}{m^k} \\ &= z+\frac{z-1}{m^n(n+1)}\sum_{0\le i\le n-1}\binom{n+1}{i+2}(m-i-2)^{n-i-1}G_i(z) \qquad (4.21)\end{aligned}$$

Following the ideas presented in [37] we will find the Poisson transform $\tilde{P}_m(x,z)$ of $P_{m,n}(z)$. So, we first obtain an accurate analysis under a Poisson-filling model, and then after using the inversion theorem of the Poisson transform we convert $\tilde{P}_m(x,z)$ back to $P_{m,n}(z)$. If we use the definition of the Poisson transform we obtain

$$\begin{aligned} \tilde{P}_m(x,z) &= z+(z-1)e^{-mx}\sum_{n\ge 0}\frac{(mx)^n}{m^n(n+1)!}\sum_{i=0}^{n-1}\binom{n+1}{i+2}(m-i-2)^{n-i-1}G_i(z) \\ &= z+(z-1)e^{-mx}\sum_{i\ge 0}\frac{x^{i+1}}{(i+2)!}G_i(z)\sum_{n-i-1\ge 0}\frac{((m-i-2)x)^{n-i-1}}{(n-i-1)!} \\ &= z+(z-1)\sum_{i\ge 0}e^{-(i+2)x}\frac{x^{i+1}}{(i+2)!}G_i(z). \qquad (4.22)\end{aligned}$$

Now, we have to find a recurrence for $G_i(z)$, and try to solve it. Note that $G_i(z)$ is defined in almost full tables of size $i+2$. If we use (4.11) and the definition of $G_i(z)$ we may easily check that for $z=1$

$$\mathbf{U}_z G_i(z) = \frac{(i+1)(i+2)^{i+1}}{2}. \qquad (4.23)$$

4.2.1 A Recurrence for $G_i(z)$

We first present a recurrence for $F_{\ell,r}(z)$, which is required to derive the recurrence we need. We have a table of size $\ell+r+2$, with $\ell+r$ elements inserted, and want to see what happens when we add the $(\ell+r+1)^{st}$ element. There are four cases as described in Figure 4.3. When the tagged element is moved one position, the label z of the arrow shows that we need z as a factor in the recurrence.

Case a) is the insertion of the tagged element. In this case case the generating function is z times the number of ways of generating a table of size $\ell+r+2$ with $\ell+r$ elements, in such a way that the last cluster is of size k. For a fixed k, this factor is $\binom{\ell+r}{k}(k+$

$1)^{k-1}(\ell+r-k+1)^{\ell+r-k-1}$. Since k ranges from 0 to r, the contribution is

$$F_{\ell,r}(z) \leftarrow z \sum_{0\leq k\leq r} \binom{\ell+r}{k} (k+1)^{k-1}(\ell+r-k+1)^{\ell+r-k-1}. \tag{4.24}$$

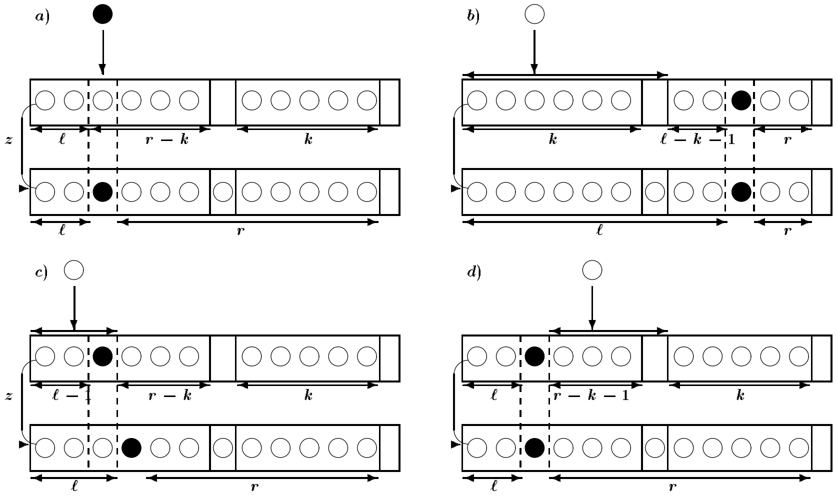

Figure 4.3:

For the last three cases, we assume that the inserted element is not the tagged one. Case *b*) is the insertion of an element in the cluster that precedes the one that has •. The cost of searching for • does not increase. We have $k+1$ different positions where the new element may hash. The number of ways of generating the upper table is the product of the number of ways of generating the first cluster and the number of ways of generating the second one. For a fixed k, the number of ways of generating the second cluster is $F_{\ell-k-1,r}(z)$, while we have $\binom{\ell+r}{k}(k+1)^{k-1}$ ways of generating the first one. Since k ranges from 0 to $\ell-1$,

$$F_{\ell,r}(z) \leftarrow \sum_{0\leq k\leq \ell-1} \binom{\ell+r}{k} (k+1)^{k-1}F_{\ell-k-1,r}(z)(k+1). \tag{4.25}$$

Case *c*) is the insertion of an element to the left of the tagged element. Now, the cost of searching for it increases by 1, and therefore we multiply by z. We have ℓ positions where the element may hash. Following a similar analysis as for the previous cases we have

$$F_{\ell,r}(z) \leftarrow \ell z \sum_{0\leq k\leq r} \binom{\ell+r}{k} (k+1)^{k-1}F_{\ell-1,r-k}(z). \tag{4.26}$$

Case *d*) is the insertion of an element to the right of •. Again, in this case the cost of searching for • does not increase. We have $r - k$ positions where the element may hash. Therefore,

$$F_{\ell,r}(z) \leftarrow \sum_{0 \le k \le r-1} \binom{\ell+r}{k} (k+1)^{k-1} F_{\ell,r-k-1}(z)(r-k). \tag{4.27}$$

Putting the contributions of (4.24),(4.25),(4.26) and (4.27) together, and noting that in cases *b*), *c*) and *d*) we may omit the limits in the sum if we assume that $F_{\ell,r}(z) = 0$ for $l < 0$ and $r < 0$, we have the recurrence:

$$\begin{aligned} F_{\ell,r}(z) &= z \sum_{0 \le k \le r} \binom{\ell+r}{k} (k+1)^{k-1} (\ell + r - k + 1)^{\ell+r-k-1} \\ &\quad + \sum_{0 \le k \le r} \binom{\ell+r}{k} (k+1)^{k-1} \left(F_{\ell-k-1,r}(z)(k+1) + \ell z F_{\ell-1,r-k}(z) \right. \\ &\quad \left. + (r-k) F_{\ell,r-k-1}(z) \right). \end{aligned} \tag{4.28}$$

If we sum both sides of (4.28) for $\ell + r = i$, we have

$$\begin{aligned} G_i(z) &= \sum_{\ell+r=i} (\ell+1) F_{\ell,r}(z) \\ &= \sum_{k \ge 0} \binom{i}{k} (k+1)^{k-1} \left(z(i-k+1)^{i-k-1} \sum_{k \le r \le i} (i-r+1) \right. \\ &\quad + \sum_{\ell+r=i} (\ell+1)(k+1) F_{\ell-k-1,r}(z) + \sum_{\ell+r=i} z\ell(\ell+1) F_{\ell-1,r-k}(z) \\ &\quad \left. + \sum_{\ell+r=i} (\ell+1)(i-\ell-k) F_{\ell,r-k-1}(z) \right) \\ &= \sum_{k \ge 0} \binom{i}{k} (k+1)^{k-1} \left(z \frac{(i-k+1)^{i-k}(i-k+2)}{2} \right. \\ &\quad + \sum_{(\ell-k-1)+r=i-k-1} ((\ell-k+1)+k+2)(k+1) F_{\ell-k-1,r}(z) \\ &\quad + \sum_{(\ell-1)+(r-k)=i-k-1} z((\ell-1)+1)((\ell-1)+2) F_{\ell-1,r-k}(z) \\ &\quad \left. + \sum_{\ell+(r-k-1)=i-k-1} (\ell+1)(i-\ell-k) F_{\ell,r-k-1}(z) \right) \end{aligned}$$

$$
\begin{aligned}
= {} & \frac{z}{2}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k}(i-k+2) \\
& +\sum_{k}\binom{i}{k}(k+1)^{k-1}\sum_{\ell+r=i-k-1}F_{\ell,r}(z)\,((\ell+k+2)(k+1) \\
& \qquad + z(\ell+1)(\ell+2)+(\ell+1)(i-\ell-k))\,.
\end{aligned}
$$

So, if we use the definition of $G_i(z)$ and $C_i(z)$, we arrive at the following recurrence for $G_i(z)$:

$$
\begin{aligned}
G_i(z) = {} & \frac{z}{2}\sum_{k}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k}(i-k+2) \\
& +\sum_{k}\binom{i}{k}(k+1)^{k-1}\Big((i+3)G_{i-k-1}(z)+(k+1)^2C_{i-k-1}(z)\Big) \\
& +(z-1)\sum_{k}\binom{i}{k}(k+1)^{k-1}\sum_{\ell+r=i-k-1}(\ell+1)(\ell+2)F_{\ell,r}(z). \qquad (4.29)
\end{aligned}
$$

Later we will require the value of $\mathbf{U}_z\mathbf{D}_zG_i(z)$. So, we need to prove the following:

Lemma 4.2

$$
\begin{aligned}
\mathbf{U}_z\mathbf{D}_zG_i(z) = {} & \frac{(i+2)^{i+2}}{2}+\frac{(i+2)^i}{6}-\frac{7(i+2)^{i+1}}{12}Q_0(i+2,i) \\
& +(i+3)\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}\mathbf{U}_z\mathbf{D}_zG_{i-k-1}(z). \qquad (4.30)
\end{aligned}
$$

Proof: If in (4.29) we take derivatives with respect to z and evaluate at $z=1$, we have

$$
\begin{aligned}
\mathbf{U}_z\mathbf{D}_zG_i(z) = {} & \sum_{k}\binom{i}{k}(k+1)^{k-1}\frac{(i-k+1)^{i-k}(i-k+2)}{2} \\
& +(i+3)\sum_{k}\binom{i}{k}(k+1)^{k-1}\mathbf{U}_z\mathbf{D}_zG_{i-k-1}(z) \\
& +\sum_{k}\binom{i}{k}(k+1)^{k+1}\mathbf{U}_z\mathbf{D}_zC_{i-k-1}(z) \\
& +\sum_{k}\binom{i}{k}(k+1)^{k-1}\sum_{\ell+r=i-k-1}(\ell+1)(\ell+2)\mathbf{U}_zF_{\ell,r}(z).
\end{aligned}
$$

If we use (4.14) and (4.23), then

$$\begin{aligned}\mathbf{U}_z\mathbf{D}_zG_i(z) = {} & (i+3)\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}\mathbf{U}_z\mathbf{D}_zG_{i-k-1}(z)\\ & +\frac{1}{2}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+1}(i-k+1)^{i-k-1}(i-k)Q_0(i-k+1,i-k-1)\\ & +\frac{1}{2}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+1}(i-k+1)^{i-k}\\ & +\frac{1}{6}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k}\\ & +\frac{1}{3}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k+2}. \qquad (4.31)\end{aligned}$$

If we divide by $(i+2)^i$, the second sum of the right hand side of (4.31) has the form

$$s(i) \;=\; \frac{1}{(i+2)^i}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+1}(i-k+1)^{i-k-1}(i-k)h_{i-k-1}. \qquad (4.32)$$

So, we have a sum that is the same as that studied in Corollary 3.6, for $p=1$, $q=1$, $c_1=c_2=1$, and $f(n)=Q_0(n+2,n)$. If we use (3.32) and (2.54) then, the Diagonal Poisson Transform of $s(i)$ is

$$\mathcal{D}_2[s(i);x] \;=\; \frac{x}{(1-x)^3}\frac{1}{1-x}=\frac{1}{(1-x)^4}-\frac{1}{(1-x)^3}. \qquad (4.33)$$

Dividing by $(i+2)^i$, the next three addends of (4.31) have the form

$$s(i)=\frac{1}{(i+2)^i}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+p}(i-k+1)^{i-k+q}. \qquad (4.34)$$

So, we can use Corollary 3.3 for the following values of $(p,q)=(1,0)$, and Corollary 3.4 for $q=0$ and $q=2$. Defining

$$\begin{aligned}r(i) \;\equiv\; & \frac{1}{2}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+1}(i-k+1)^{i-k-1}(i-k)Q_0(i-k+1,i-k-1)\\ & +\frac{1}{2}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+1}(i-k+1)^{i-k}\end{aligned}$$

$$\begin{aligned}
&+\frac{1}{6}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k}\\
&+\frac{1}{3}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k+2},
\end{aligned}$$

we have by (2.53), (2.54) and (2.55),

$$\begin{aligned}
\mathcal{D}_2\left[\frac{r(i)}{(i+2)^i};x\right] &= \frac{1}{2}\left(\frac{1}{(1-x)^4}-\frac{1}{(1-x)^3}\right)\\
&\quad +\frac{1}{2}(1-x)\frac{1}{(1-x)^3}\frac{1}{(1-x)}\\
&\quad +\frac{1}{6}(1-x)\frac{1}{(1-x)}\\
&\quad +\frac{1}{3}(1-x)\frac{1}{(1-x)}\left(\frac{3}{(1-x)^5}-\frac{2}{(1-x)^4}\right)\\
&= \frac{1}{6}-\frac{2}{3(1-x)^3}+\frac{3}{2(1-x)^4}. \qquad (4.35)
\end{aligned}$$

Using (3.30) and (3.32) to find the inverse of the transform (4.35), and (2.33), (2.35) to simplify the expressions we obtain, we find

$$\begin{aligned}
r(i) &= \frac{1}{6}-\frac{2}{3}Q_2(n+2,n)+\frac{3}{2}Q_3(n+2,n)\\
&= \frac{(i+2)^{i+2}}{2}+\frac{(i+2)^i}{6}-\frac{7(i+2)^{i+1}}{12}Q_0(i+2,i). \qquad (4.36)
\end{aligned}$$

Substituting this value for $r(i)$ back into (4.36), we obtain

$$\begin{aligned}
\mathbf{U}_z\mathbf{D}_zG_i(z) &= \frac{(i+2)^{i+2}}{2}+\frac{(i+2)^i}{6}-\frac{7(i+2)^{i+1}}{12}Q_0(i+2,i)\\
&\quad +(i+3)\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}\mathbf{U}_z\mathbf{D}_zG_{i-k-1}(z). \qquad (4.37)
\end{aligned}$$

$\mathcal{QED}$

It is interesting to note that setting z to 1 in (4.29) and applying (4.13), we have

$$\begin{aligned}
\mathbf{U}_zG_i(z) &= \sum_k\binom{i}{k}(k+1)^{k-1}\frac{(i-k+1)^{i-k}(i-k+2)}{2}\\
&\quad +\sum_k\binom{i}{k}(k+1)^{k-1}\left((i+3)\mathbf{U}_zG_{i-k-1}(z)+(k+1)^2\mathbf{U}_zC_{i-k-1}(z)\right)
\end{aligned}$$

$$\begin{aligned}
&= (i+3)\sum_{k}\binom{i}{k}(k+1)^{k-1}\mathbf{U}_z G_{i-k-1}(z) \\
&\quad +\frac{1}{2}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k} \\
&\quad +\frac{1}{2}\sum_{k\geq 0}\binom{i}{k}(k+1)^{k-1}(i-k+1)^{i-k+1} \\
&\quad -\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+1}(i-k+1)^{i-k} \\
&\quad +\sum_{k\geq 0}\binom{i}{k}(k+1)^{k+1}(i-k+1)^{i-k+1}. \qquad (4.38)
\end{aligned}$$

We can use Corollary 3.3 to find the values of the sums that do not involve $\mathbf{U}_z G_{i-k-1}(z)$. This gives us a recurrence for $\mathbf{U}_z G_i(z)$, to which we apply formula (3.77) for $c=2$, $d=3$ and $p=-1$. This reverifies the special case (4.38) previously given as (4.23).

4.3 Verification of Known Results

In this section we rewrite (4.22) as a function of $\mathcal{D}_2[g_i(z);x]$ and then verify that $\mathbf{E}[A_{n+1}] = \frac{1}{2}(1+Q_0(m,n))$.

Define $\grave{g}_2(x,z)$ as $\mathcal{D}_2[g_i(z);x]$, where $g_i(z) = \frac{G_i(z)}{(i+2)^i(i+1)}$, then

$$\begin{aligned}
\frac{\partial(x\tilde{P}_m(x,z))}{\partial x} &= \tilde{P}_m(x,z) + x\frac{\partial \tilde{P}_m(x,z)}{\partial x} \\
&= z + (z-1)\sum_{i\geq 0} e^{-(i+2)x}\frac{(i+1)(i+2)^i x^{i+1}}{(i+2)!}g_i(z) \\
&+ x(z-1)\sum_{i\geq 0} e^{-(i+2)x} g_i(z)\left[-\frac{(i+1)(i+2)^{i+1}x^{i+1}}{(i+2)!} + \frac{(i+1)^2(i+2)^i x^i}{(i+2)!}\right] \\
&= z + (z-1)x\sum_{i\geq 0} e^{-(i+2)x}\frac{(i+1)(i+2)^i x^i}{(i+2)!}g_i(z)(1-(i+2)x+(i+1)) \\
&= z + (z-1)x(1-x)\sum_{i\geq 0} e^{-(i+2)x}\frac{(i+2)^i x^i}{i!}g_i(z) \\
&= z + (z-1)x\grave{g}_2(x,z). \qquad (4.39)
\end{aligned}$$

Therefore we derive

$$\tilde{P}_m(x,z) = \frac{1}{x}\int_0^x \left(z + t(z-1)\grave{g}_2(t,z)\right)dt = z + \frac{z-1}{x}\int_0^x t\grave{g}_2(t,z)dt. \tag{4.40}$$

Taking derivatives with respect to z we obtain

$$\mathbf{U}_z\mathbf{D}_z\tilde{P}_m(x,z) = 1 + \frac{1}{x}\int_0^x t\mathbf{U}_{\mathbf{z}}\grave{g}_2(t,z)dt \tag{4.41}$$

$$\mathbf{U}_z\mathbf{D}_z^2\tilde{P}_m(x,z) = \frac{2}{x}\int_0^x t\mathbf{U}_z\mathbf{D}_z\grave{g}_2(t,z)dt. \tag{4.42}$$

From (4.23), $\mathbf{U}_{\mathbf{z}}g_i(z) = (i+2)/2$, therefore $\mathbf{U}_{\mathbf{z}}\grave{g}_2(x,z) = \mathcal{D}_2\left[\frac{i+2}{2};x\right]$. By (3.35) we know $\mathcal{D}_2\left[\frac{(i+2)}{2};t\right] = \frac{1}{2}\left(\frac{1}{(1-t)^2} + \frac{1}{(1-t)}\right)$. Therefore, if we substitute into (4.41) and integrate, we find that $\mathbf{U}_z\mathbf{D}_z\tilde{P}_m(x,z) = \frac{1}{2}\left(1 + \frac{1}{1-x}\right)$.

Since $1/(1-x)$ is the Poisson transform of $Q_0(m,n)$, we have given an alternative proof of (4.2) to that of [46].

4.4 Solving the recurrence for $\mathbf{U}_z\mathbf{D}_z g_i(z)$

In (4.40), we wrote $\tilde{P}_m(x,z)$ as a function of $\mathcal{D}_2[g_i(z);x]$, and in (4.42) we found the value of $\mathbf{U}_z\mathbf{D}_z^2\tilde{P}_m(x,z)$ as a function of $\mathbf{U}_z\mathbf{D}_z\grave{g}_2(x,z)$. However, we still do not know the value of $\mathbf{U}_z\mathbf{D}_z\grave{g}_2(x,z)$.

Equation (4.30) is the special case of (3.71) with $c=2$, $d=3$ and $p=-1$. Since $p=-1$, $s_p(x)=x$. Therefore, the general solution simplifies to

$$\grave{h}_2(x) = \frac{e^x}{x}\int_0^x e^{-t}\mathcal{D}_2[(i+1)b_i;t]dt. \tag{4.43}$$

In (4.42) $\grave{h}_2(x) = \mathbf{U}_z\mathbf{D}_z\grave{g}_2(x,z)$. Applying (4.43) to (4.42):

$$\begin{aligned}
\mathbf{U}_z\mathbf{D}_z^2\tilde{P}_m(x,z) &= \frac{2}{x}\int_0^x e^u\left(\int_0^u e^{-t}\mathcal{D}_2[(i+1)b_i;t]dt\right)du \\
&= \frac{2}{x}\int_0^x e^{-t}\mathcal{D}_2[(i+1)b_i;t]\left(\int_t^x e^u du\right)dt \\
&= \frac{2}{x}\int_0^x \left(e^{x-t}-1\right)\mathcal{D}_2[(i+1)b_i;t]dt.
\end{aligned} \tag{4.44}$$

In (4.30) we have $(i+1)b_i = \frac{(i+2)^2}{2} + \frac{1}{6} - \frac{7(i+2)}{12}Q_0(i+2,i)$. If we use (3.35),(3.30) and (3.32) for $c=2$, we arrive at the final result:

$$\mathbf{U}_z\mathbf{D}_z^2\tilde{P}_m(x,z) = \frac{2}{x}\int_0^x \left(e^{x-t}-1\right)\left(\frac{3}{2(1-t)^4} - \frac{2}{3(1-t)^3} + \frac{1}{6}\right)dt$$

$$= \frac{1}{3(1-x)} + \frac{1}{2(1-x)^2} - \frac{1}{3} - \frac{1}{3x}(e^x - 1) - \frac{e^{x-1}}{6x}\left(Ei(1) - Ei(1-x)\right) \quad (4.45)$$

where $Ei(1) - Ei(1-x) = \int_{1-x}^{1} \frac{e^t}{t} dt$. The function $Ei(x)$ is the exponential integral function [1]. Next we apply the inversion formulae presented in [37] to find $\mathbf{U}_z\mathbf{D}_z^2 P_{m,n}(z)$.

4.4.1 Finding $\mathbf{U}_z\mathbf{D}_z^2 P_{m,n}(z)$

Since the Poisson transform is linear, we need only find the inverse of each summand of (4.45). We find easily the inverse of the first three, by (3.7), (3.8) and (3.10). With more work, we find the inverse of the other two addends. With a change of variable $t = 1 - v$ we have $\frac{e^{x-1}}{x} \int_{1-x}^{1} \frac{e^t}{t} dt = \frac{e^x}{x} \int_0^x \frac{e^{-v}}{1-v} dv$. To find the inverse transform of the function $e^{-x}/(1-x)$, we may use (3.13). Then, applying formulae (3.16) and (3.14), we arrive at the relation

$$\begin{aligned} &\tilde{P}_m\left[\frac{1}{6}\left(\frac{m+1}{m}\right)^n \frac{1}{n+1}\sum_{k=0}^{n}\left(\frac{m}{m+1}\right)^k Q_0(m,k); x\right] \\ = &\frac{e^{x-1}}{6x}\left(Ei(1) - Ei(1-x)\right) \end{aligned} \quad (4.46)$$

Using a similar analysis, we find the remaining inverse transform

$$\tilde{P}_m\left[\frac{m+1}{3(n+1)}\left(\frac{m+1}{m}\right)^n - \frac{m}{3(n+1)}; x\right] = \frac{1}{3x}(e^x - 1), \quad (4.47)$$

and have proven

Lemma 4.3

$$\begin{aligned} \mathbf{U}_z\mathbf{D}_z^2 P_{m,n}(z) &= \frac{1}{3}Q_0(m,n) + \frac{1}{2}Q_1(m,n) - \frac{1}{3} - \frac{m+1}{3(n+1)}\left(\frac{m+1}{m}\right)^n \\ &+ \frac{m}{3(n+1)} - \frac{1}{6}\left(\frac{m+1}{m}\right)^n \frac{1}{n+1}\sum_{k=0}^{n}\left(\frac{m}{m+1}\right)^k Q_0(m,k). \end{aligned} \quad (4.48)$$

4.5 Analysis of the Variance

As a consequence of Lemma 4.3 and using (2.15) we have the following theorem.

Theorem 4.1

$$\begin{aligned} \mathbf{V}[A_{n+1}] &= \frac{1}{2}Q_1(m,n) - \frac{1}{4}{Q_0}^2(m,n) + \frac{1}{3}Q_0(m,n) - \frac{m+1}{3(n+1)}\left(\frac{m+1}{m}\right)^n \\ &- \frac{1}{12} + \frac{m}{3(n+1)} - \frac{1}{6}\left(\frac{m+1}{m}\right)^n \frac{1}{n+1}\sum_{k=0}^{n}\left(\frac{m}{m+1}\right)^k Q_0(m,k). \end{aligned} \quad (4.49)$$

If we use the approximation theorem, Theorem 3.2, we have the following result for a table with $n = \alpha m$ elements, for fixed $0 \leq \alpha < 1$ and $n, m \to \infty$.

Theorem 4.2

$$\begin{aligned}\mathbf{V}[A_{\alpha m}] &= \frac{1}{4(1-\alpha)^2} + \frac{1}{3(1-\alpha)} - \frac{1}{3\alpha}(e^{\alpha} - 1) \\ &\quad - \frac{e^{\alpha-1}}{6\alpha}\left(Ei(1) - Ei(1-\alpha)\right) - \frac{1}{12} + O\left(\frac{1}{m}\right) \qquad (4.50)\end{aligned}$$

Now, we want to study the asymptotic behavior of the variance for a full table ($n = m - 1$). We know by (2.37) the asymptotic behavior of $Q_0(m, m-1)$, and we have $Q_1(m, m-1) = m$. Then the only difficulty is with the asymptotic expansion of the last summand of $\mathbf{V}[A_m]$. This is done in two steps. First, in Lemma 4.4, we find the asymptotic expansion of $\frac{1}{m}\sum_{k=0}^{m-1} Q_0(m,k)$ up to $o(1/\sqrt{m})$. Then we generalize the ideas presented in this lemma to find the expansion for our original sum.

Lemma 4.4

$$\frac{1}{m}\sum_{k=1}^{m} Q_0(m,k) = \sum_{k=1}^{m} \frac{m^{\underline{k}}}{km^k} = \frac{H_m}{2} + \frac{\ln 2}{2} + \frac{1}{3}\sqrt{\frac{\pi}{2m}} + o\left(\frac{1}{\sqrt{m}}\right). \qquad (4.51)$$

Proof: In [7], Bender gives the first term of the approximation, but we would like some lower order terms. First, note that $\frac{m^{\underline{k}}}{km^k}$ is a monotone decreasing function of k. So,

$$\begin{aligned}\sum_{k=m^{7/12}}^{m} \frac{m^{\underline{k}}}{km^k} &= \sum_{k=m^{7/12}}^{m} \frac{m!}{k(m-k)!m^k} \\ &< m\frac{m!}{m^{7/12}(m-m^{7/12})!m^{m^{7/12}}} = O(m^{\frac{5}{12}}e^{-\frac{m^{1/6}}{2}}), \qquad (4.52)\end{aligned}$$

that is exponentially small. Therefore, we only have to consider the sum of the first $m^{7/12}$ terms.

The sum may be rewritten as

$$\begin{aligned}\sum_{k=1}^{m^{7/12}} \frac{m^{\underline{k}}}{km^k} &= \sum_{k=1}^{m^{7/12}} \frac{1}{k}\prod_{j=0}^{k-1}\left(1 - \frac{j}{m}\right) = \sum_{k=1}^{m^{7/12}} \frac{1}{k} e^{\left(\sum_{j=0}^{k-1} \ln\left(1-\frac{j}{m}\right)\right)} \\ &= \sum_{k=1}^{m^{7/12}} \frac{1}{k} e^{-\left(\sum_{j=0}^{k-1}\sum_{i=i}^{\infty} \frac{1}{i}\left(\frac{j}{m}\right)^i\right)} = \sum_{k=1}^{m^{7/12}} \frac{1}{k} e^{-\left(\sum_{i=1}^{\infty} \frac{1}{im^i}\sum_{j=0}^{k-1} j^i\right)}\end{aligned}$$

$$= \sum_{k=1}^{m^{7/12}} \frac{1}{k} \prod_{i=1}^{\infty} e^{-\frac{1}{im^i}\sum_{j=0}^{k-1} j^i} . \qquad (4.53)$$

If we use formulae (2.6) and (2.7) and the asymptotic expansion of e^x, we have

$$\begin{aligned}
\sum_{k=1}^{m^{7/12}} \frac{1}{k} \prod_{i=1}^{\infty} e^{-\frac{1}{im^i}\sum_{j=0}^{k-1} j^i} &= \sum_{k=1}^{m^{7/12}} \frac{1}{k} e^{-k^2/2m} e^{k/2m} e^{-k^3/6m^2} \left(1 + O\left(\frac{k^2}{m^2} + \frac{k^4}{m^3} + \frac{k^6}{m^4}\right)\right) \\
&= \sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m}}{k} \left(1 + \frac{k}{2m}\right)\left(1 - \frac{k^3}{6m^2}\right) \\
&= \sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m}}{k} \left(1 + \frac{k}{2m}\right)\left(1 - \frac{k^3}{6m^2}\right) O\left(\frac{k^2}{m^2} + \frac{k^4}{m^3} + \frac{k^6}{m^4}\right) \\
&= \sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m}}{k} + \sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m}}{2m} - \sum_{k=1}^{m^{7/12}} \frac{k^2 e^{-k^2/2m}}{6m^2} \\
&\quad + \sum_{k=1}^{m^{7/12}} e^{-k^2/2m} O\left(\frac{k}{m^2} + \frac{k^3}{m^3} + \frac{k^5}{m^4}\right). \qquad (4.54)
\end{aligned}$$

The Euler-Maclaurin summation formula can be used to find good estimates for (4.54). This formula is

$$\begin{aligned}
\sum_{a \le k < b} f(k) &= \int_a^b f(x)dx - \frac{1}{2} f(x) \mid_a^b + \sum_{k=1}^{r} \frac{B_{2k}}{(2k)!} f^{(2k-1)}(x) \mid_a^b \\
&\quad + O((2\pi)^{-2r}) \int_a^b \mid f^{2r}(x) \mid dx. \qquad (4.55)
\end{aligned}$$

We may see that the contribution of the last sum in (4.54) is $O(1/m)$, and therefore we need only examine the first three sums.

The first sum can be rewritten as

$$\sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m}}{k} = \sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m} - 1}{k} + \sum_{k=1}^{m^{7/12}} \frac{1}{k}. \qquad (4.56)$$

The first sum can be approximated by an integral, and the second sum gives us the harmonic numbers. Using (2.10), we apply the Euler-Maclaurin formula to the first sum, giving

$$\begin{aligned}\sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m}-1}{k} + \sum_{k=1}^{m^{7/12}} \frac{1}{k} &= \left(\frac{-1}{12}\ln m - \frac{\gamma}{2} + \frac{\ln 2)}{2} + O\left(\frac{1}{m^{7/12}}\right)\right) \\ &\quad + \left(\frac{7}{12}\ln m) + \gamma + O\left(\frac{1}{m^{7/12}}\right)\right) \\ &= \frac{1}{2}(\ln m) + \gamma + \ln 2)) + O\left(\frac{1}{m^{7/12}}\right) \\ &= \frac{H_n}{2} + \frac{\ln 2}{2} + o\left(\frac{1}{\sqrt{m}}\right). \end{aligned} \tag{4.57}$$

We apply the Euler-Maclaurin formula to the other two sums and find:

$$\sum_{k=1}^{m^{7/12}} \frac{e^{-k^2/2m}}{2m} = \frac{1}{2}\sqrt{\frac{\pi}{2m}} + O\left(\frac{1}{n}\right) \tag{4.58}$$

and

$$-\sum_{k=1}^{m^{7/12}} \frac{k^2 e^{-k^2/2m}}{6m^2} = -\frac{1}{6}\sqrt{\frac{\pi}{2m}} + O\left(\frac{1}{n}\right). \tag{4.59}$$

The lemma follows from (4.57), (4.58) and (4.59). $\mathcal{QED}$

Lemma 4.5

$$\sum_{k=0}^{m-1} \left(\frac{m}{m+1}\right)^k Q_0(m,k) = \frac{m}{e}\left(\frac{H_m}{2} + \frac{\ln 2}{2} + Ei(1) - \gamma - \frac{2}{3}\sqrt{\frac{\pi}{2m}}\right) + o\left(\frac{1}{\sqrt{m}}\right).$$

Proof: The key ideas are similar to those used to prove Lemma 4.4. We use the following well known generating function

$$\frac{1}{(1+z)^k} = \sum_{n\geq 0}(-1)^n \binom{n+k-1}{n} z^n. \tag{4.60}$$

The definition of $Q_0(m,k)$ can be used to rewrite the sum

$$\begin{aligned}\sum_{k=0}^{m-1}\left(\frac{m}{m+1}\right)^k \frac{Q_0(m,k)}{m} &= \frac{1}{m}\sum_{k=0}^{m-1} \frac{1}{(1+1/m)^k} \sum_{i=0}^{k} \frac{k^{\underline{i}}}{m^i} \\ &= \frac{1}{m}\sum_{i=0}^{m-1} \frac{i!}{m^i} \sum_{k=i}^{m-1} \binom{k}{i} \frac{1}{(1+1/m)^k}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{i=0}^{m-1}\frac{i!}{m^i}\sum_{k=i}^{m-1}\binom{k}{i}\sum_{r\geq 0}\binom{r+k-1}{r}\frac{(-1)^r}{m^r} \\
&= \frac{1}{m}\sum_{r\geq 0}\frac{(-1)^r}{m^r}\sum_{i=0}^{m-1}\frac{i!}{m^i}\sum_{k=i}^{m-1}\binom{r+k-1}{r}\binom{k}{i}. \qquad (4.61)
\end{aligned}
$$

Now, we find the value of the innermost sum. We have

$$
\begin{aligned}
\sum_{k=i}^{m-1}\binom{k+r-1}{r}k^{\underline{i}} &= \frac{(i+r-1)!}{r!}\sum_{k=i}^{m-1}k\binom{k+r-1}{i+r-1} \\
&= \frac{(i+r-1)!}{r!}\sum_{k=i}^{m-1}(k+r-r)\binom{k+r-1}{i+r-1} \\
&= \frac{(i+r-1)!}{r!}\left((i+r)\sum_{k=i}^{m-1}\binom{k+r}{i+r}-r\sum_{k=i}^{m-1}\binom{k+r-1}{i+r-1}\right) \\
&= \frac{(i+r-1)!}{r!}\left((i+r)\binom{m+r}{i+r+1}-r\binom{m+r-1}{i+r}\right) \\
&= a_r(i,m)-a_{r-1}(i,m), \qquad (4.62)
\end{aligned}
$$

where

$$
a_r(i,m) = i!(i,r)\binom{m+r}{i+r+1} = (m,r)\frac{m^{\underline{i+1}}}{i+r+1}. \qquad (4.63)
$$

Defining

$$
\begin{aligned}
b_r(m) &\equiv (m,r)m\sum_{i=1}^{m}\frac{m^{\underline{i}}}{(i+r)m^i} \qquad (4.64) \\
b_{-1}(m) &\equiv 0, \qquad (4.65)
\end{aligned}
$$

and using (4.62), we may rewrite (4.61) as

$$
\begin{aligned}
&\frac{1}{m}\sum_{r\geq 0}\frac{(-1)^r}{m^r}\sum_{i=0}^{m-1}\frac{i!}{m^i}\sum_{k=i}^{m-1}\binom{r+k-1}{r}\binom{k}{i} = \frac{1}{m}\sum_{r\geq 0}\frac{(-1)^r}{m^r}\left(b_r(m)-b_{r-1}(m)\right) \\
&= \frac{1}{m}\left(1+\frac{1}{m}\right)\sum_{r\geq 0}\frac{(-1)^r}{m^r}b_r(m) = \left(1+\frac{1}{m}\right)\sum_{r\geq 0}\frac{(-1)^r}{m^r}(m,r)\sum_{i=1}^{m}\frac{m^{\underline{i}}}{(i+r)m^i} \\
&= \left(1+\frac{1}{m}\right)\sum_{r\geq 0}\frac{(-1)^r(m+r)^{\underline{r}}}{m^r r!}\sum_{i=1}^{m}\frac{m^{\underline{i}}}{(i+r)m^i}. \qquad (4.66)
\end{aligned}
$$

Equation (4.66) is simplified by discarding terms known to be $o(1/\sqrt{m})$. First we know that $1/\ln(m)! = o(1/m)$, and therefore we can discard all the terms for $r > \ln m$. Then, for $r \leq \ln m$, we know that $(m+r)^{\underline{r}} = m^r + O(r^2 m^{r-1})$. and so $(m+r)^{\underline{r}}/m^r = 1 + O(r^2/m)$. Now, if we use Lemma 4.4, as $r \geq 0$, the innermost sum of (4.66) is $O(\ln m)$. Therefore we have

$$\begin{aligned}
&\frac{m+1}{m} \sum_{r\geq 0} \frac{(-1)^r (m+r)^{\underline{r}}}{m^r r!} \sum_{i=1}^{m} \frac{m^{\underline{i}}}{(i+r)m^i} \\
&= \sum_{r=0}^{\ln m} \frac{(-1)^r (m+r)^{\underline{r}}}{m^r r!} \sum_{i=1}^{m} \frac{m^{\underline{i}}}{(i+r)m^i} + o\left(\frac{1}{m}\right) && (4.67) \\
&= \sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \sum_{i=1}^{m} \frac{m^{\underline{i}}}{(i+r)m^i} + O\left(\frac{\ln m}{m}\right) && (4.68) \\
&= \sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \sum_{i=1}^{m^{7/12}} \frac{m^{\underline{i}}}{(i+r)m^i} + O\left(\frac{\ln m}{m}\right). && (4.69)
\end{aligned}$$

We continue with a line of reasoning similar to the proof of Lemma 4.4. We may check that if $r = O(\ln m)$, then all the expansions given by the Euler-Maclaurin formula are exactly the same for all the terms up to $O(1/\sqrt{m})$. This is the main reason to bound the sum up to $\ln m$ terms. Hence, we have the following derivation, where the equalities are up to $o(1/\sqrt{m})$ (we omit this term, so the text is more readable)

$$\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \sum_{i=1}^{m^{7/12}} \frac{m^{\underline{i}}}{(i+r)m^i} = \sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \sum_{k=1}^{m^{7/12}} \frac{1}{k+r} e^{-k^2/2m} e^{k/2m} e^{-k^3/6m^2} =$$

$$\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \sum_{k=1}^{m^{7/12}} \frac{1}{k+r} e^{-\frac{(k+r)^2}{2m}} e^{\frac{(2r+1)(k+r)}{2m}} e^{-\frac{(k+r)^3}{6m^2}} =$$

$$\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \sum_{k=1}^{m^{7/12}} \frac{1}{k+r} e^{-\frac{(k+r)^2}{2m}} \left(1 + \frac{(2r+1)(k+r)}{2m}\right) \left(1 - \frac{(k+r)^3}{6m^2}\right) =$$

$$\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \left(\sum_{k=1}^{m^{7/12}} \frac{e^{-\frac{(k+r)^2}{2m}}}{(k+r)} + (2r+1) \sum_{k=1}^{m^{7/12}} \frac{e^{-\frac{(k+r)^2}{2m}}}{2m} - \sum_{k=1}^{m^{7/12}} \frac{(k+r)^2 e^{-\frac{(k+r)^2}{2m}}}{6m^2} \right) =$$

$$\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \left(\sum_{k=r+1}^{m^{7/12}} \frac{e^{-k^2/2m}}{k} + (2r+1) \sum_{k=r+1}^{m^{7/12}} \frac{e^{-k^2/2m}}{2m} - \sum_{k=r+1}^{m^{7/12}} \frac{k^2 e^{-k^2/2m}}{6m^2} \right) =$$

$$\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \left(\sum_{k=r+1}^{m^{7/12+r}} \frac{e^{-k^2/2m} - 1}{k} + \sum_{k=r+1}^{m^{7/12+r}} \frac{1}{k} + (2r+1) \sum_{k=r+1}^{m^{7/12+r}} \frac{e^{-k^2/2m}}{2m} \right.$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad\qquad\qquad - \sum_{k=r+1}^{m^{7/12+r}} \frac{k^2 e^{-k^2/2m}}{6m^2} \Bigg) = \\
&\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \left(\left(-\frac{\ln m}{12} - \frac{\gamma}{2} + \frac{\ln 2}{2} \right) + \left(\frac{7}{12} \ln m + \gamma - H_r \right) \right. \\
&\qquad\qquad\qquad \left. + \left(\frac{2r+1}{4} \sqrt{\frac{2\pi}{m}} \right) - \left(\frac{1}{12} \sqrt{\frac{2\pi}{m}} \right) \right) = \\
&\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \left(\left(\frac{H_m}{2} + \frac{\ln 2}{2} + \frac{1}{3} \sqrt{\frac{\pi}{2m}} \right) + \left(r \sqrt{\frac{\pi}{2m}} - H_r \right) \right) = \\
&\sum_{r=0}^{\ln m} \frac{(-1)^r}{r!} \left(\frac{H_m}{2} + \frac{\ln 2}{2} - \frac{2}{3} \sqrt{\frac{\pi}{2m}} \right) - \sum_{r=1}^{\ln m} \frac{(-1)^r H_r}{r!} = \\
&\frac{1}{e} \left(\frac{H_m}{2} + \frac{\ln 2}{2} - \frac{2}{3} \sqrt{\frac{\pi}{2m}} \right) - \sum_{r=1}^{\infty} \frac{(-1)^r H_r}{r!} + O\left(\frac{1}{\ln(m)!} \right) = \\
&\frac{1}{e} \left(\frac{H_m}{2} + \frac{\ln 2}{2} - \frac{2}{3} \sqrt{\frac{\pi}{2m}} - \gamma + E_i(1) \right). \qquad (4.70)
\end{aligned}
$$

The last equation requires some explanation. If we define $H(z) = \sum_{k \geq 1} H_k z^k / k!$, then we must find $H(-1)$.It is easy to check that $z \frac{\partial H(z)}{\partial z} = zH(z) + e^z - 1$. Solving the differential equation, we evaluate the result in $z = -1$, and have $H(-1) = (\gamma - E_i(1))/e$. $\mathcal{QED}$
From (2.37), Theorem 4.1 and Lemma 4.5 we have

Theorem 4.3

$$
\begin{aligned}
\mathbf{V}[A_m] \;=\;& \frac{4-\pi}{8} m + \frac{\sqrt{2\pi m}}{4} - \frac{1}{12} H_m + \left(\frac{1}{9} - \frac{\ln 2}{12} - \frac{\pi}{48} - \frac{Ei(1)}{6} - \frac{e}{3} + \frac{\gamma}{6} \right) \\
&+ \frac{181}{2160} \sqrt{\frac{2\pi}{m}} + o\left(\frac{1}{\sqrt{m}} \right), \qquad (4.71)
\end{aligned}
$$

Comparing with (4.8), we have shown that for a full table, the last-come-first-served heuristic on a linear probing hash table achieves the optimal variance for the distribution of successful searches, up to lower order terms.

4.6 Analysis of the Standard Linear Probing Hashing Algorithm

In a footnote ([46, p.529]), D.E. Knuth acknowledges that the standard linear probing hashing was the first nontrivial algorithm he had ever analyzed satisfactorily. He did this analysis in 1962. However, the first published analysis of this algorithm was done by Konheim and Weiss in 1966 [52]. In this section, we present a different analysis of this algorithm, based on similar ideas as those used to analyze the LCFS linear probing algorithm.

We define $P_{m,n}(z)$ as the probability generating function for the cost for searching $\bullet$ in a table of size m with $n+1$ elements inserted. As observed in section 3.2.1, we have $\mathcal{P}_m[P_{m,n}(z); x] = \mathcal{D}_2[P_{n+2,n}(z); x]$. Therefore, we only have to study $P_{n+2,n}(z)$.

There are two cases as indicated in Figure 4.4.

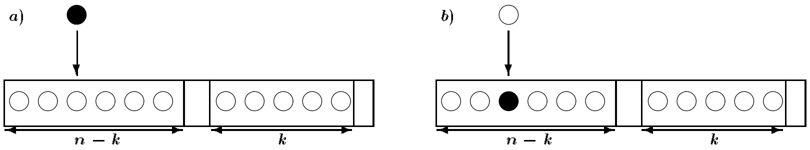

Figure 4.4:

In case a), we insert $\bullet$. There are $(k+1)^{k-1}$ ways of creating a table of size $k+1$, with k elements inserted in such a way that the last location is empty. Similarly, there are $(n-k+1)^{n-k-1}$ ways of creating a table of size $n-k+1$, with $n-k$ elements inserted in such a way that the last location is empty. Since $\bullet$ can hash into any of the first $n-k$ locations of the cluster, the cost for inserting $\bullet$ will be $\sum_{j=0}^{n-k} z^{j+1}$. Since we are working with probability generating functions, we have to divide by the normalization factor $(n+2)^n(n+1)$, as there are $(n+2)^n$ ways of inserting n elements in a table of size $n+2$ and there are $n+1$ different possibilities for choosing $\bullet$. Therefore, for case a) we have

$$P_{n+2,n}(z) \sim \sum_{k\geq 0} \binom{n}{k} \frac{(k+1)^{k-1}(n-k+1)^{n-k-1}}{(n+2)^n(n+1)} \sum_{0\leq j\leq n-k} z^{j+1}. \qquad (4.72)$$

In case b), the element inserted is not $\bullet$, therefore, the cost for searching it, does not increase. There are $(n+2)$ places where the new element can hash. There are $(k+1)^{k-1}$ ways of creating a table of size $k+1$, with k elements inserted in such a way that the last location is empty. There are $(n-k+1)^{n-k-1}(n-k)P_{k-1}(z)$ ways to create a table of size $n-k+1$ with $n-k$ elements inserted, one of them $\bullet$, with z tracking the cost of retrieving $\bullet$, in such a way that the last location of the table is empty. Then, for case b), we have

$$P_{n+2,n}(z) \quad \sim \quad (n+2)\sum_{k\geq 0} \binom{n}{k} \frac{(k+1)^{k-1}(n-k+1)^{n-k-1}}{(n+2)^n(n+1)} (n-k)P_{n-k-1}. \qquad (4.73)$$

Adding (4.72) and (4.73), we find

$$\begin{aligned}P_{n+2,n}(z) &= \frac{1}{(n+2)^n(n+1)}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k-1}\sum_{0\leq j\leq n-k} z^{j+1}\\ &+\frac{(n+2)}{(n+2)^n(n+1)}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k-1}(n-k)P_{n-k-1}(z).\end{aligned}$$

Moreover, $P_{n+2,n}(z)$ verifies recurrence (3.71) with parameters $d=2$, $c=2$ $p=-1$ and $B_n(z)=\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k-1}\sum_{j=0}^{n-k} z^{j+1}$. By (3.77) we have $\mathcal{D}_2[P_{n+2,n}(z);x]$ $=\frac{1}{x}\int_0^x \mathcal{D}_2[(n+1)B_n(z);t]dt$.

Since we need $\mathbf{U}_z\mathbf{D}_z\mathcal{D}_2[P_{n+2,n}(z);x]$ and $\mathbf{U}_z\mathbf{D}_z^2\mathcal{D}_2[P_{n+2,n}(z);x]$, then we have to find the values of $\mathbf{U}_z\mathbf{D}_z\mathcal{D}_2[(n+1)B_n(z);x]$ and $\mathbf{U}_z\mathbf{D}_z^2\mathcal{D}_2[(n+1)B_n(z);x]$. If we differentiate $(n+1)B_n(z)$ and evaluate at $z=1$ we have

$$\begin{aligned}\mathbf{U}_z\mathbf{D}_z(n+1)B_n(z) &= \frac{1}{(n+2)^n}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k-1}\sum_{j=0}^{n-k}(j+1)\\ &= \frac{1}{2(n+2)^n}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k+1}\\ &\quad+\frac{1}{2(n+2)^n}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k}\\ &= \frac{1}{2}Q_1(n+2,n)+\frac{1}{2}. \qquad (4.74)\end{aligned}$$

Using (3.30) and (3.32) we have

$$\mathcal{D}_2\left[\frac{1}{2}Q_1(n+2,n)+\frac{1}{2};x\right]=\frac{1}{2(1-x)^2}+\frac{1}{2}, \qquad (4.75)$$

and then

$$\begin{aligned}\mathbf{U}_z\mathbf{D}_z\mathcal{P}_m[P_{m,n}(z);x] &= \mathbf{U}_z\mathbf{D}_z\mathcal{D}_2[P_{n+2,n}(z);x] \qquad (4.76)\\ &= \frac{1}{2x}\int_0^x\left(\frac{1}{(1-t)^2}+1\right)dt \qquad (4.77)\\ &= \frac{1}{2}\left(\frac{1}{1-x}+1\right). \qquad (4.78)\end{aligned}$$

So, by (3.8) and (3.10), we find

$$\mathbf{E}[A_{n+1}]=\frac{1}{2}(1+Q_0(m,n)) \qquad (4.79)$$

as expected.

With respect to the second moment, we find

$$\begin{aligned}\mathbf{U}_z\mathbf{D}_z^2\mathcal{D}_2[(n+1)B_n(z);x] &= \frac{1}{3(n+2)^n}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k+2} \\ &\quad -\frac{1}{3(n+2)^n}\sum_{k\geq 0}\binom{n}{k}(k+1)^{k-1}(n-k+1)^{n-k} \\ &= \frac{(n+2)^2-2(n+2)Q_0(n+2,n)+Q_2(n+2,n)-1}{3}.\end{aligned}$$

Then, by (3.30) and (3.32) we arrive at

$$\begin{aligned}&\mathcal{D}_2\left[\frac{1}{3}\left((n+2)^2-2(n+2)Q_0(n+2,n)+Q_2(n+2,n)\right)-\frac{1}{3};x\right] \\ &\qquad = \frac{1}{3}\left(\frac{3}{(1-x)^4}-\frac{2}{(1-x)^3}-1\right), \qquad (4.80)\end{aligned}$$

and therefore,

$$\begin{aligned}\mathbf{U}_z\mathbf{D}_z^2\mathcal{P}_m[P_{m,n}(z);x] &= \mathbf{U}_z\mathbf{D}_z^2\mathcal{D}_2[P_{n+2,n}(z);x] && (4.81)\\ &= \frac{1}{3x}\int_0^x\left(\frac{3}{(1-t)^4}-\frac{2}{(1-t)^3}-1\right)dt && (4.82)\\ &= \frac{1}{3}\left(\frac{1}{(1-x)^3}-1\right). && (4.83)\end{aligned}$$

Finally, by (3.8) and (3.10), we have

$$\mathbf{U}_z\mathbf{D}_z^2P_{m,n}(z)=\frac{1}{3}(Q_2(m,n)-1), \qquad (4.84)$$

and as a consequence, we obtain

$$\begin{aligned}\mathbf{V}[A_n] &= \frac{1}{3}\left(Q_2(m,n)-1\right)+\frac{1}{2}\left(Q_0(m,n)+1\right)-\left(\frac{1}{2}\left(Q_0(m,n)+1\right)\right)^2 \\ &= \frac{1}{3}Q_2(m,n)-\frac{Q_0^2(m,n)}{4}-\frac{1}{12}. && (4.85)\end{aligned}$$

as we know from [46, 35].

Chapter 5

Linear Probing Hashing with Buckets

While I was kissing Manuelita, she said: "When daddy is with me, he will kiss me. However, while he is in Canada, I will kiss the moon and he will also kiss her".

5.1 Introduction

The problem of storing information in a computer memory or a peripheral device has been widely studied. Several data structures have been proposed that work well on secondary storage devices such as magnetic disks. Two of the most popular techniques are B-trees (and its variations) introduced by Bayer and McCreight [6], and hashing with buckets (Peterson in [69] presented the first major paper in this area). Two good sources of information for this problem are [46] and [35]. More recently, O'Neil [67] presents some applications to data bases.

Several methods for handling overflow records in hash tables have been proposed. Many of these methods are based on *open addressing* [69]. The key of each record uniquely determines a probe sequence that is followed for storing or retrieving the record. The most basic algorithm for conflict resolution under open addressing is linear probing.

In this chapter we present an exact analysis for the average cost of a successful search in a linear probing hash table with buckets of size b. In [9], Blake and Konheim studied the asymptotic behavior of the algorithm as the number of records and buckets tend together to infinity so that their ratio is constant. Mendelson [60], derived exact formulae for the problem, but only solved them numerically.

We present an analysis of Robin Hood linear probing hashing [16, 17] with buckets of size b. This algorithm is introduced in section 5.3. It is well known [69], that in a hash table accessed by linear probing, the average number of probes for a successful search is independent of the collision resolution strategy used, and this is true for any set of keys. Therefore our analysis gives an exact solution for the algorithm studied in [60], and solves the open problem presented by D. Knuth in question 6.4.56 in [46].

This chapter is divided as follows. Section 5.2 contains preliminary definitions and theorems. In section 5.3 we introduce the Robin Hood heuristic, and in sections 5.4, 5.5 and 5.6 the main results are proved. Finally, in section 5.7 we present a different point of view to study some aspects of the problem.

5.2 Some Preliminaries

We define $Q_{m,n,d}$ as the number of ways of inserting n records in a table with m buckets of size b, so that a given (say the last) bucket of the table contains more than d empty slots. The subscript b will be omitted, as it is a fixed parameter. There cannot be more empty slots than the size of the bucket so $Q_{m,n,b} = 0$. For each of the m^n possible arrangements, the last bucket has 0 or more empty slots, and so $Q_{m,n,-1} = m^n$. Observe that $Q_{m,n,0}$ gives the number of ways of inserting n records into a table with m buckets, so that the last bucket is not full. For notational convenience, we define $Q_{0,n,d} = [n = 0]$. In [60], Mendelson proves

Theorem 5.1 *For $0 \leq d \leq b-1$, and $m > 0$,*

$$Q_{m,n,d} = \begin{cases} \sum_{j=0}^{n} \binom{n}{j} Q_{m-1,j,d} & (0 \leq n < mb \text{ -}d); \\ 0 & (n \geq mb\text{-}d). \end{cases}$$

It does not seem possible to find a closed formula for $Q_{m,n,d}$. However, as we shall see, for the average cost of a successful search we only require $\sum_{d=0}^{b-1} Q_{m,n,d}$. The following theorem, tells us that this sum is surprisingly simple.

Theorem 5.2

$$\sum_{d=0}^{b-1} Q_{m,n,d} = bm^n - nm^{n-1} \qquad (0 \leq n \leq bm). \tag{5.1}$$

Proof:

Let $P_{m,n,j} = \frac{Q_{m,n,j-1} - Q_{m,n,j}}{m^n}$. $P_{m,n,j}$ is the probability of inserting n records in a table with m buckets of size b so that the last bucket of the table contains exactly j empty slots. Then, as $Q_{m,n,b} = 0$,

$$Q_{m,n,d} = m^n \sum_{j=d+1}^{b} P_{m,n,j} \tag{5.2}$$

As a consequence, we find the following identity

$$\begin{aligned} \sum_{d=0}^{b-1} Q_{m,n,d} &= m^n \sum_{d=0}^{b-1} \sum_{j=d+1}^{b} P_{m,n,j} && (5.3) \\ &= m^n \sum_{j=1}^{b} P_{m,n,j} \sum_{d=0}^{j-1} 1 && (5.4) \\ &= m^n \sum_{j=1}^{b} j P_{m,n,j}. && (5.5) \end{aligned}$$

The last sum gives the expected number of empty slots in a given bucket. There is an average of $\frac{n}{m}$ records in each bucket of capacity b. Therefore the expected number of empty slots in a given bucket is $b - \frac{n}{m}$, and the theorem is proved. $\mathcal{QED}$

We will need the exponential generating function of $\sum_{d=0}^{b-1} Q_{m,j,d}$ for $0 \leq j \leq bm$. This is easily obtained using Theorem 5.2 as

$$\sum_{d=0}^{b-1} Q_{m,d}(x) = \sum_{j=0}^{bm} \sum_{d=0}^{b-1} Q_{m,j,d} \frac{x^j}{j!}$$

$$= \sum_{j=0}^{bm} (bm^j - jm^{j-1}) \frac{x^j}{j!}. \tag{5.6}$$

5.3 Robin Hood Linear Probing

When a new record moves to an occupied location in an open addressing hash table, the usual solution is to let the incoming key try again in some other bucket. Thus, the standard collision resolution strategy can be called "First-Come-First-Served". Operating in the context of double hashing, Celis *et al.* [16, 17] defined the Robin Hood heuristic, under which each collision occurring on each insertion is resolved in favor of the record that is farthest away from its home bucket. We will focus on the same heuristic but in the context of linear probing (as did Carlsson *et al.* in [14] for buckets of capacity one).

Figure 5.1 shows the result of inserting records with the keys 36, 77, 24, 79, 56, 69, 49, 18, 38, 97, 78, 10, 58, 29, 30 and 16 in a table with ten buckets of size two, and with hash function $h(x) = x \bmod 10$, and resolving collisions by linear probing using the Robin Hood heuristic.

a	78	49	79	30	24		16	56	97	38
	29	69	10				36	77	18	58
	0	1	2	3	4	5	6	7	8	9

Figure 5.1:

When there is a collision in bucket i and this bucket is full, then the record that has probed the least number of buckets, probes bucket $(i + 1) \bmod m$. In the case of a tie, we (arbitrarily) move the record whose key has largest value.

a	49	79			24		36	77	18	58
	69	10					56	97	38	78
	0	1	2	3	4	5	6	7	8	9

Figure 5.2:

Figure 5.2 shows the partially filled table after inserting 58. When we want to insert 29, bucket 9 is full. Both keys in bucket 9 are in their second probe position, and 29 is in its first, so it has to try bucket 0. At bucket 0, all three keys are in their second probe position. Then we arbitrarily choose 69, the key with largest value, to probe bucket 1. At bucket 1, both 69 and 79 are in their third probe bucket, while 10 is in its second. So, 10 has to move to bucket 2, where it is inserted. Figure 5.3 shows the table after inserting 29.

a	29	69	10		24		36	77	18	58
	49	79					56	97	38	78
	0	1	2	3	4	5	6	7	8	9

Figure 5.3:

The following properties are easily verified:

- At least one record is in its home bucket.
- The keys are stored in nondecreasing order by hash value, starting at some location k and wrapping around. In our example, $k = 6$ (the second slot of the third bucket).
- If a fixed rule is used to break ties among the candidates to probe their next probe bucket (eg: by sorting these keys in increasing order), then the resulting table is independent of the order in which the records were inserted [16].

5.4 Linear Probing Sort

To analyze Robin Hood linear probing with buckets, we first have to discuss some ideas presented in [14] and [37].

For $b = 1$, when the hash function is order preserving (that is, if $x < y$ then $h(x) < h(y)$), a variation of the Robin Hood linear probing algorithm can be used to sort [37], by successively inserting the n records in an initially empty table. In this case, instead of letting the excess records from the rightmost bucket of the table wrap around to bucket zero, we can use an *overflow area* consisting of buckets m, $m + 1$, etc. The number of buckets needed for this overflow area is an important performance measure for this sorting algorithm.

In this section we study the average number of records that overflow when the buckets have capacity b. Then, in section 5.5 we show how this analysis is related to the study of the cost of successful searches in the Robin Hood linear probing algorithm.

Let $W_{m,n}(w)$ be the generating function for the number of records that go to the overflow area when n keys are inserted in a table with m buckets, each with capacity b. Since b is a given parameter, this subscript is omitted. Let us also define $W_{m,n,k} = [w^k]W_{m,n}(w)$.

The records inserted in the table can be divided in two sets, as shown in Figure 5.4. The hash table can be seen as a concatenation of two tables of size $m - 1$, and 1 respectively.

If $n - k \geq b$, then $n - k - b$ records go to the overflow area as a consequence of being inserted in the last bucket of the table. To this number we have to add the records that go to the overflow area when k records are inserted in the table of size $m - 1$. Then, for this case, the probability generating function for the number of records that overflow is $W_{m-1,k}(w)w^{n-k-b}$.

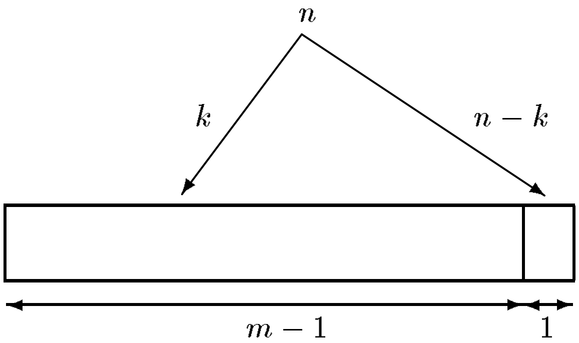

Figure 5.4:

Therefore, as a first approximation

$$W_{m,n}(w) \sim \sum_{0\le k\le n} \binom{n}{k} \left(\frac{m-1}{m}\right)^{k} \left(\frac{1}{m}\right)^{n-k} W_{m-1,k}(w)w^{n-k-b} \tag{5.7}$$

since there are $\binom{n}{k}$ ways of choosing the $n-k$ records that hash to the last bucket, and the probability that any record hashes to a given bucket is $1/m$.

However, we have to make a correction because, when $n-k<b$, there is no overflow caused by the records inserted in the last bucket of the table. In such a case, the following correction term is needed

$$\sum_{0\le i<b-(n-k)} W_{m-1,k,i}\left(1-w^{i+n-k-b}\right). \tag{5.8}$$

Then, by (5.7) and (5.8), we have the following recurrence for the probability generating function of the size of overflow

$$\begin{aligned} W_{m,n}(w) &= \sum_{0\le k\le n} \binom{n}{k} \left(\frac{m-1}{m}\right)^{k} \left(\frac{1}{m}\right)^{n-k} \\ &\left(W_{m-1,k}(w)w^{n-k-b} + \sum_{0\le i<b-(n-k)} W_{m-1,k,i}\left(1-w^{i+n-k-b}\right)\right). \end{aligned} \tag{5.9}$$

As a consequence of this correction term, the values of $W_{m,n,i}$ for $0\le i<b$ have to be studied separately. So, the first bucket of the overflow area is analyzed with a different approach.

5.4.1 First Bucket of the Overflow Area

Let $D_{m,n,r} \equiv Q_{m,n,b-r-1} - Q_{m,n,b-r}$, be the number of ways of inserting n records so that the last bucket has exactly r records, for $0 \leq r \leq b$. Also define $B_{m,n,r} \equiv m^n W_{m,n,r}$. We want to find $B_{m,n,r}$ for $0 \leq r < b$.

Theorem 5.3

$$B_{m,n,r} = D_{m+1,n,r} - \sum_{j=1}^{r} \binom{n}{j} B_{m,n-j,r-j}. \tag{5.10}$$

Proof: $B_{m,n,r}$ can first be approximated by $D_{m+1,n,r}$. However, we do not want any record to hash to bucket m. This situation should be considered when $1 \leq r \leq b$.

For a fixed j with $1 \leq j \leq r$, $B_{m,n-j,r-j}$ counts the number of ways of inserting $n-j$ records in a table of size m, such that $r-j$ records go to overflow. Since there should be r records in the overflow area, then j records have to hash to bucket m. There are $\binom{n}{j}$ different ways of choosing these j records. So, for a fixed j, the number of forbidden configurations is $\binom{n}{j}B_{m,n-j,r-j}$. Then, the lemma is proven by letting j vary from 1 to r.

$\mathcal{QED}$

As a solution of (5.10), we have

Theorem 5.4

$$B_{m,n,r} = \sum_{j=0}^{r} (-1)^j \binom{n}{j} D_{m+1,n-j,r-j}. \tag{5.11}$$

Proof: By Theorem 5.3, we have

$$D_{m+1,n,r} = \sum_{j=0}^{r} \binom{n}{j} B_{m,n-j,r-j}, \tag{5.12}$$

and since $B_{m,n-j,r-j}$ and $D_{m+1,n-j,r-j}$ both vanish when $j > r$ (as $0 \leq r-j \leq b$), then

$$D_{m+1,n,r} = \sum_{j=0}^{n} \binom{n}{j} B_{m,n-j,r-j}. \tag{5.13}$$

For a fixed r, let $\overline{B}_{m,n-j} \equiv B_{m,n-j,r-j}$ and $\overline{D}_{m+1,n-j} \equiv D_{m+1,n-j,r-j}$. Also define $\overline{B}_m(z) \equiv \sum_{n\geq 0} \overline{B}_{m,n}\frac{z^n}{n!}$ and $\overline{D}_{m+1}(z) \equiv \sum_{n\geq 0} \overline{D}_{m+1,n}\frac{z^n}{n!}$. Then, by (5.13)

$$\overline{D}_{m+1,n} = \sum_{j=0}^{n} \binom{n}{j} \overline{B}_{m,n-j}. \tag{5.14}$$

This identity is directly translated into an equation in their respective exponential generating functions as

$$\overline{D}_{m+1}(z) = e^{z}\overline{B}_{m}(z). \tag{5.15}$$

If (5.15) is solved for $\overline{B}_m(z)$, and then we consider the coefficient of $\frac{z^n}{n!}$ on both sides, the following inverse relation is obtained

$$\overline{B}_{m,n} = \sum_{j=0}^{n}(-1)^j\binom{n}{j}\overline{D}_{m+1,n-j}, \tag{5.16}$$

and so,

$$\begin{aligned} B_{m,n,r} &= \sum_{j=0}^{n}(-1)^j\binom{n}{j}D_{m+1,n-j,r-j} && (5.17)\\ &= \sum_{j=0}^{r}(-1)^j\binom{n}{j}D_{m+1,n-j,r-j}. && (5.18)\end{aligned}$$

$\mathcal{QED}$

Corollary 5.1

$$W_{m,n}(w) = \sum_{0\le k\le n}\binom{n}{k}\left(\frac{m-1}{m}\right)^{n-k}\left(\frac{1}{m}\right)^{k}$$
$$\left(W_{m-1,n-k}(w)w^{k-b} + \sum_{0\le i<b-k}\left(1-w^{i+k-b}\right)\sum_{j=0}^{i}(-1)^j\binom{n-k}{j}\frac{D_{m,n-k-j,i-j}}{(m-1)^{n-k}}\right). \tag{5.19}$$

5.4.2 Distribution of the Size of the Overflow Area

In this section we use the Poisson Transform to find $E[W_{m,n}]$. Let us define

$$T_m(x,w) \equiv e^{-mx}\sum_{n\ge 0}W_{m,n}(w)\frac{(mx)^n}{n!} = \mathcal{P}_m[W_{m,n}(w);x] \tag{5.20}$$

$$\text{and}\qquad R_m(x,w) \equiv e^{mx}T_m(x,w) = \sum_{n\ge 0}R_{m,n}(w)x^n. \tag{5.21}$$

First we will find $a_i, i \ge 0$ that satisfy

$$\mathbf{U}_w\mathbf{D}_wT_m(x,w) = \mathcal{P}_m[E[W_{m,n}];x] = \sum_{i\ge 0}a_ix^i, \tag{5.22}$$

and then, by Theorem 3.1,

$$E[W_{m,n}] = \sum_{i \ge 0} a_i \frac{n^{\underline{i}}}{m^i} \tag{5.23}$$

By Corollary 5.1, and the definition of $R_{m,n}(w)$,

$$\begin{aligned} R_{m,n}(w) = {} & \frac{1}{w^b} \sum_{0 \le k \le n} R_{m-1,n-k}(w) \frac{w^k}{k!} \\ & + \frac{1}{n!} \sum_{0 \le k \le n} \binom{n}{k} \sum_{0 \le i < b-k} \left(1 - w^{i+k-b}\right) \sum_{j=0}^{i} (-1)^j \binom{n-k}{j} D_{m,n-k-j,i-j}. \end{aligned} \tag{5.24}$$

Let us first concentrate on the last sum of (5.24). The following lemma will be useful for this purpose.

Lemma 5.1

$$\sum_{k=0}^{\ell} (-1)^k \binom{n}{k} \binom{n-k}{\ell-k} = [\ell = 0]. \tag{5.25}$$

Proof:

By (2.24), we have

$$\sum_{k=0}^{\ell} (-1)^k \binom{n}{k} \binom{n-k}{\ell-k} \quad = \quad \binom{n}{\ell} \sum_{k=0}^{\ell} (-1)^k \binom{\ell}{k} = [\ell = 0]. \tag{5.26}$$

$\mathcal{QED}$

If $s = i + k$, then

$$\begin{aligned} & \frac{1}{n!} \sum_{0 \le k \le n} \binom{n}{k} \sum_{0 \le i < b-k} \left(1 - w^{i+k-b}\right) \sum_{j=0}^{i} (-1)^j \binom{n-k}{j} D_{m,n-k-j,i-j} && (5.27) \\ = {} & \frac{1}{n!} \sum_{0 \le k \le n} \binom{n}{k} \sum_{0 \le s < b} \left(1 - w^{s-b}\right) \sum_{j=0}^{s-k} (-1)^j \binom{n-k}{j} D_{m,n-k-j,s-k-j} && (5.28) \\ = {} & \frac{1}{n!} \sum_{0 \le s < b} \left(1 - w^{s-b}\right) \sum_{0 \le k \le n} \binom{n}{k} \sum_{j=0}^{s-k} (-1)^j \binom{n-k}{j} D_{m,n-k-j,s-k-j}. && (5.29) \end{aligned}$$

Actually, the upper bound of the sum indexed by k may be s instead of n. If $n < s$, when $n < k \le s$, $\binom{n}{k} = 0$ because $n \ge 0$. Moreover, if $n > s$, when $s < k \le n$, the sum indexed by j is 0, because $s-k<0$. If we use Lemma 5.1 and define $\ell = k+j$, then

$$\begin{aligned}
& \frac{1}{n!}\sum_{0\le s<b}\left(1-w^{s-b}\right)\sum_{0\le k\le n}\binom{n}{k}\sum_{j=0}^{s-k}(-1)^j\binom{n-k}{j}D_{m,n-k-j,s-k-j}\\
= \; & \frac{1}{n!}\sum_{0\le s<b}\left(1-w^{s-b}\right)\sum_{0\le k\le s}\binom{n}{k}\sum_{j=0}^{s-k}(-1)^j\binom{n-k}{j}D_{m,n-k-j,s-k-j}\\
= \; & \frac{1}{n!}\sum_{0\le s<b}\left(1-w^{s-b}\right)\sum_{0\le k\le s}\binom{n}{k}\sum_{\ell=k}^{s}(-1)^{\ell-k}\binom{n-k}{\ell-k}D_{m,n-\ell,s-\ell}\\
= \; & \frac{1}{n!}\sum_{0\le s<b}\left(1-w^{s-b}\right)\sum_{0\le \ell\le s}(-1)^{\ell}D_{m,n-\ell,s-\ell}\sum_{k=0}^{\ell}(-1)^k\binom{n}{k}\binom{n-k}{\ell-k}\\
= \; & \frac{1}{n!}\sum_{0\le s<b}\left(1-w^{s-b}\right)D_{m,n,s}. \qquad (5.30)
\end{aligned}$$

So, by (5.24) and (5.30) we can write

$$\begin{aligned}
R_{m,n}(w) &= \frac{1}{w^b}\sum_{0\le k\le n}R_{m-1,n-k}(w)\frac{w^k}{k!}+\frac{1}{n!}\sum_{0\le s<b}\left(1-w^{s-b}\right)D_{m,n,s}\\
&= \frac{1}{w^b}\sum_{0\le k\le n}R_{m-1,n-k}(w)\frac{w^k}{k!}+\frac{1}{n!}\sum_{0\le s<b}\left(1-w^{s-b}\right)\left(Q_{m,n,b-s-1}-Q_{m,n,b-s}\right)\\
&= \frac{1}{w^b}\sum_{0\le k\le n}R_{m-1,n-k}(w)\frac{w^k}{k!}+\frac{1}{n!}\sum_{0<s\le b}\left(1-w^{-s}\right)\left(Q_{m,n,s-1}-Q_{m,n,s}\right)\\
&= \frac{1}{w^b}\sum_{0\le k\le n}R_{m-1,n-k}(w)\frac{w^k}{k!}+A_{m,n}(w), \qquad (5.31)
\end{aligned}$$

where $A_{m,n}(w)$ denotes the sum indexed by s. If

$$A_m(x,w)=\sum_{n\ge 0}A_{m,n}(w)x^n \qquad (5.32)$$

then,

$$\begin{aligned}
R_m(x,w) &= \frac{1}{w^b}\sum_{n\ge 0}\left(\sum_{k=0}^{n}R_{m-1,n-k}(w)\frac{w^k}{k!}\right)x^n+A_m(x,w)\\
&= \frac{1}{w^b}\sum_{k\ge 0}\frac{(wx)^k}{k!}\sum_{n\ge k}R_{m-1,n-k}(w)x^{n-k}+A_m(x,w)\\
&= \frac{e^{wx}}{w^b}R_{m-1}(x,w)+A_m(x,w). \qquad (5.33)
\end{aligned}$$

Since (5.33) is a linear recurrence with $R_0(x,w)=1$, we find

$$R_m(x,w) = \frac{e^{mxw}}{w^{bm}} + \sum_{k=1}^{m} \frac{e^{(m-k)xw}}{w^{b(m-k)}} A_k(x,w). \tag{5.34}$$

Finally, by the definition of $T_m(x,w)$,

$$\begin{aligned} \mathcal{P}_m[W_{m,n};x] &= e^{-mx} R_m(x,w) \\ &= \frac{e^{mx(w-1)}}{w^{bm}} + \sum_{k=1}^{m} e^{-kx} \frac{e^{(m-k)x(w-1)}}{w^{b(m-k)}} A_k(x,w). \end{aligned} \tag{5.35}$$

Let us study now $A_k(x,w)$. From its definition,

$$\begin{aligned} A_k(x,w) &= \sum_{n\geq 0} A_{k,n}(w) x^n \\ &= \sum_{n\geq 0} \frac{x^n}{n!} \sum_{0<s\leq b} \left(1-w^{-s}\right)\left(Q_{k,n,s-1} - Q_{k,n,s}\right) \\ &= \sum_{0<s\leq b} \left(1-w^{-s}\right) \sum_{n\geq 0} \frac{x^n}{n!} \left(Q_{k,n,s-1} - Q_{k,n,s}\right) \\ &= \sum_{0<s\leq b} \left(1-w^{-s}\right)\left(Q_{k,s-1}(x) - Q_{k,s}(x)\right). \end{aligned} \tag{5.36}$$

As a consequence,

$$\mathbf{U}_w A_k(x,w) = 0, \tag{5.37}$$

and by (5.6),

$$\begin{aligned} \mathbf{U}_w \mathbf{D}_w A_k(x,w) &= \sum_{0<s\leq b} s\left(Q_{k,s-1}(x) - Q_{k,s}(x)\right) \\ &= \sum_{0<s\leq b} Q_{k,s}(x) \\ &= \sum_{j=0}^{bk} (bk^j - jk^{j-1}) \frac{x^j}{j!}. \end{aligned} \tag{5.38}$$

Finally, since

$$\begin{aligned} \mathbf{U}_w \mathbf{D}_w \left(\frac{e^{(m-k)(w-1)x}}{w^{b(m-k)}} \right) &= \mathbf{U}_w \left(\frac{e^{(m-k)(w-1)x}(m-k)(wx-b)}{w^{b(m-k)+1}} \right), \\ &= (m-k)(x-b) \end{aligned} \tag{5.39}$$

then by (5.22), (5.35), (5.37), (5.38) and (5.39),

$$\mathcal{P}_m[E[W_{m,n}];x] \;=\; m(x-b)+\sum_{k=1}^{m} e^{-kx}\sum_{j=0}^{bk}(bk^j - jk^{j-1})\frac{x^j}{j!}. \tag{5.40}$$

This sum can be further simplified. If $n = i + j$, then

$$\begin{aligned}
&\sum_{k=1}^{m} e^{-kx}\sum_{j=0}^{bk}(bk^j - jk^{j-1})\frac{x^j}{j!}\\
=\;&\sum_{k=1}^{m}\sum_{i\geq 0}(-1)^i\frac{(kx)^i}{i!}\sum_{j=0}^{bk}(bk^j - jk^{j-1})\frac{x^j}{j!}\\
=\;&\sum_{k=1}^{m}\sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{j=0}^{min(n,bk)}(-1)^j\binom{n}{j}k^{n-j}(bk^j - jk^{j-1})\\
=\;&\sum_{k=1}^{m}\sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{j=0}^{bk}(-1)^j\binom{n}{j}k^{n-j}(bk^j - jk^{j-1}) &(5.41)\\
=\;&\sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\sum_{j=0}^{bk}(-1)^j\binom{n}{j}(bk-j). &(5.42)
\end{aligned}$$

Step (5.41) needs some justification when $n < bk$, as it may cause problems when $n < j \leq bk$. In this range, $\binom{n}{j} = 0$, and so $min(n, bk)$ can be substituted by bk as the upper bound of the sum indexed by j.

To continue the simplification, we require an identity that is a special case of (2.27):

$$\sum_{j=0}^{bk}(-1)^j\binom{n}{j} = (-1)^{bk}\binom{n-1}{bk}. \tag{5.43}$$

Therefore, from (5.42),

$$\begin{aligned}
&\sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\sum_{j=0}^{bk}(-1)^j\binom{n}{j}(bk-j)\\
=\;&\sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\left(bk\sum_{j=0}^{bk}(-1)^j\binom{n}{j} - \sum_{j=0}^{bk}(-1)^j j\binom{n}{j}\right)\\
=\;&\sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\left(bk\sum_{j=0}^{bk}(-1)^j\binom{n}{j} - n\sum_{j=1}^{bk}(-1)^j\binom{n-1}{j-1}\right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\left(bk\sum_{j=0}^{bk}(-1)^j\binom{n}{j}+n\sum_{j=0}^{bk-1}(-1)^j\binom{n-1}{j}\right)\\
&= \sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\left(bk(-1)^{bk}\binom{n-1}{bk}+n(-1)^{bk-1}\binom{n-2}{bk-1}\right)\\
&= \sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\left((-1)^{bk}(n-1)\binom{n-2}{bk-1}-n(-1)^{bk}\binom{n-2}{bk-1}\right)\\
&= \sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}(-1)^{bk+1}k^{n-1}\binom{n-2}{bk-1}\\
&= \sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}(-1)^{bk+1}k^{n-1}\left((-1)^{bk-1}\binom{bk-n}{bk-1}\right)\\
&= \sum_{n\geq 0}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\binom{bk-n}{bk-1}\\
&= bm-mx+\sum_{n\geq 2}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\binom{bk-n}{bk-1}. \qquad (5.44)
\end{aligned}$$

Finally, from (5.40) and (5.44),

$$\mathcal{P}_m[E[W_{m,n}];x] \;=\; \sum_{n\geq 2}(-1)^n\frac{x^n}{n!}\sum_{k=1}^{m}k^{n-1}\binom{bk-n}{bk-1} \qquad (5.45)$$

Moreover, by (5.23),

$$\begin{aligned}
E[W_{m,n}] &= \sum_{i\geq 2}\frac{n^{\underline{i}}}{m^i}\frac{(-1)^i}{i!}\sum_{k=1}^{m}k^{i-1}\binom{bk-i}{bk-1}\\
&= \sum_{i\geq 2}\binom{n}{i}\frac{(-1)^i}{m^i}\sum_{k=1}^{m}k^{i-1}\binom{bk-i}{bk-1}. \qquad (5.46)
\end{aligned}$$

It is important to note that for $b=1$, (2.45) can be used with $m=i$ and $n=i-1$ to calculate the inner sum. Then,

$$\begin{aligned}
E[W_{m,n}] &= \sum_{i\geq 2}\frac{n^{\underline{i}}}{m^i}\frac{(-1)^{i+1}}{i!}\sum_{k=1}^{m}(-1)^k k^{i-1}\binom{i-2}{k-1}\\
&= \sum_{i\geq 2}\frac{n^{\underline{i}}}{m^i}\frac{(-1)^{i+1}}{i!(i-1)}\sum_{k=1}^{m}(-1)^k k^{i}\binom{i-1}{k}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\geq 2} \frac{n^{\underline{i}}}{m^i} \frac{(-1)^{i+1}}{i!(i-1)} (-1)^{i-1} (i-1)! \left\{ {i \atop i-1} \right\} \\
&= \sum_{i\geq 2} \frac{n^{\underline{i}}}{m^i} \frac{1}{i!(i-1)} (i-1)! \frac{i(i-1)}{2} \\
&= \frac{1}{2} \sum_{i\geq 2} \frac{n^{\underline{i}}}{m^i} \\
&= \frac{1}{2} \left(Q_0(m,n) - 1 - \frac{n}{m} \right), \qquad (5.47)
\end{aligned}
$$

as was derived in [14] and [37].

5.5 Analysis of Robin Hood Linear Probing

In this section we find the average cost of a successful search for a random record in a hash table with m buckets of size b that contains $n+1$ records. Without loss of generality, we search for a record that hashes to bucket 0. Moreover, since the order of the insertion is not important, we assume that this record was the last one inserted.

If we look at the table after the first n records have been inserted, all the records that hash to bucket 0 (if any) will be occupying contiguous buckets, near the beginning of the table. The buckets preceding them will be occupied by records that wrapped around from the right end of the table, as can be seen in Figure 5.1. The key observation here is that those records are exactly the ones that would have gone to the overflow area. Furthermore, it is easy to see that the number of records in this overflow area does not change when the records that hash to bucket 0 are removed.

Let $S_{m,n}(y)$ be the probability generating function for the cost of a successful search for a random record that hashes to 0 in a table with m buckets of capacity b that contains $n+1$ records. As before, the subscript b will be omitted, as it is a given constant.

The cost of retrieving a record that hashes to 0 can be divided in two parts.

- The number of records (k) that wrap around the table. In other words, the size of the overflow area.

- The number of records $(i+1)$ that hash to bucket 0.

So the cost of (separately) retrieving all records that hash to bucket 0 is represented by the generating function

$$
y \sum_{r=0}^{i} y^{\left\lfloor \frac{k+r}{b} \right\rfloor}. \qquad (5.48)
$$

The y outside the sum, denotes that the cost is at least 1 (the first bucket). The exponent of y in the sum represents the fact that to retrieve the $(r+1)^{st}$ record that hashes to

0, the k records that go to overflow plus the first r records that hash to 0, have to be probed. Since the buckets have size b, we have to divide this cost by b. Hence $1+\lfloor\frac{k+r}{b}\rfloor$ is the number of buckets probed to retrieve the $(r+1)^{st}$ record that hashes to bucket 0. Therefore, the cost of retrieving a random record that hashes to 0, given that k records overflow from the end of the table and $i+1$ records hash to 0, has the generating function

$$\frac{y}{i+1}\sum_{r=0}^{i} y^{\lfloor\frac{k+r}{b}\rfloor}. \tag{5.49}$$

If the table contains $n+1$ records and $i+1$ of them hash to bucket 0, then only the remaining $n-i$ records that hash to buckets 1 through $m-1$ influence the size of the overflow area. Remember from section 5.4 that $W_{m-1,n-i,k}$ is the probability that k records overflow when we insert $n-i$ records in a table of size $m-1$ (as bucket 0 is not considered). Then,

$$\sum_{k\geq 0} W_{m-1,n-i,k}\frac{y}{i+1}\sum_{r=0}^{i} y^{\lfloor\frac{k+r}{b}\rfloor} \tag{5.50}$$

represents the cost of retrieving a random record that hashes to 0, given that $i+1$ of them hash to this bucket. We need now to average over all i. There are $\binom{n}{i}$ different possibilities to choose the i records that hash to 0 (besides the last one inserted), and the probability of a record hashing to 0 is $\frac{1}{m}$. Finally, we find the generating function

$$\begin{aligned} S_{m,n}(y) &= \sum_{i=0}^{n}\binom{n}{i}\left(\frac{1}{m}\right)^{i}\left(\frac{m-1}{m}\right)^{n-i}\sum_{k\geq 0} W_{m-1,n-i,k}\frac{y}{i+1}\sum_{r=0}^{i} y^{\lfloor\frac{k+r}{b}\rfloor} \\ &= \frac{y}{(n+1)m^n}\sum_{i=0}^{n}\binom{n+1}{i+1}(m-1)^{n-i}\sum_{k\geq 0} W_{m-1,n-i,k}\sum_{r=0}^{i} y^{\lfloor\frac{k+r}{b}\rfloor}. \end{aligned} \tag{5.51}$$

5.5.1 Average Cost of a Successful Search

The expected number of buckets inspected on a successful search is $E[S_{m,n}]=$ $\mathbf{U}_y\mathbf{D}_yS_{m,n}(y)$. By (5.51),

$$E[S_{m,n}] = \sum_{i=0}^{n}\binom{n+1}{i+1}\frac{(m-1)^{n-i}}{(n+1)m^n}\sum_{k\geq 0} W_{m-1,n-i,k}\sum_{r=0}^{i}\left(\left\lfloor\frac{k+r}{b}\right\rfloor+1\right) \tag{5.52}$$

As a first approximation, we can use the relation $x-1<\lfloor x\rfloor\leq x$, and therefore

$$\sum_{i=0}^{n}\binom{n+1}{i+1}\frac{(m-1)^{n-i}}{(n+1)m^n}\sum_{k\geq 0} W_{m-1,n-i,k}\sum_{r=0}^{i}\frac{k+r}{b} \tag{5.53}$$

$$\begin{aligned}
&< E[S_{m,n}] \\
&\leq \sum_{i=0}^{n} \binom{n+1}{i+1} \frac{(m-1)^{n-i}}{(n+1)m^n} \sum_{k\geq 0} W_{m-1,n-i,k} \sum_{r=0}^{i} \left(\frac{k+r}{b} + 1 \right). \qquad (5.54)
\end{aligned}$$

Since $W_{m,n}(w)$ is a probability generating function, $\mathbf{U}_w W_{m,n}(w) = 1$. Therefore, the difference between (5.54) and (5.53) is bounded by

$$\begin{aligned}
&\sum_{i=0}^{n} \binom{n+1}{i+1} \frac{(m-1)^{n-i}}{(n+1)m^n} \sum_{r=0}^{i} \sum_{k\geq 0} W_{m-1,n-i,k} \\
&= \sum_{i=0}^{n} \binom{n+1}{i+1} \frac{(m-1)^{n-i}}{(n+1)m^n} (i+1) \\
&= \frac{1}{m^n} \sum_{i=0}^{n} \binom{n}{i} (m-1)^{n-i} = 1. \qquad (5.55)
\end{aligned}$$

To analyze the lower bound (5.53), we first study the inner sum

$$\begin{aligned}
&\sum_{k\geq 0} W_{m-1,n-i,k} \sum_{r=0}^{i} \frac{k+r}{b} \\
&= \sum_{k\geq 0} W_{m-1,n-i,k} \left((i+1)\frac{k}{b} + \frac{i(i+1)}{2b} \right) \\
&= \frac{i+1}{b} \sum_{k\geq 0} k W_{m-1,n-i,k} + \frac{i(i+1)}{2b} \sum_{k\geq 0} W_{m-1,n-i,k} \\
&= \frac{i+1}{b} E[W_{m-1,n-i}] + \frac{i(i+1)}{2b} \qquad (5.56)
\end{aligned}$$

and so,

$$\begin{aligned}
&\sum_{i=0}^{n} \binom{n+1}{i+1} \frac{(m-1)^{n-i}}{(n+1)m^n} \sum_{k\geq 0} W_{m-1,n-i,k} \sum_{r=0}^{i} \frac{k+r}{b} \\
&= \sum_{i=0}^{n} \binom{n+1}{i+1} \frac{(m-1)^{n-i}}{(n+1)m^n} \left(\frac{i+1}{b} E[W_{m-1,n-i}] + \frac{i(i+1)}{2b} \right) \\
&= \frac{1}{b} \sum_{i=0}^{n} \binom{n}{i} \frac{(m-1)^{n-i}}{m^n} E[W_{m-1,n-i}] + \frac{n}{2b} \sum_{i=1}^{n} \binom{n-1}{i-1} \frac{(m-1)^{n-i}}{m^n} \\
&= \frac{1}{b} \sum_{i=0}^{n} \binom{n}{i} \frac{(m-1)^{n-i}}{m^n} E[W_{m-1,n-i}] + \frac{nm^{n-1}}{2bm^n}. \qquad (5.57)
\end{aligned}$$

In order to study the first sum in (5.57), we use (5.46), and so

$$\begin{aligned}
&\frac{1}{bm^n}\sum_{i=0}^{n}\binom{n}{i}(m-1)^{n-i}E[W_{m-1,n-i}] \\
=\;& \frac{1}{bm^n}\sum_{i=0}^{n}\binom{n}{i}(m-1)^{n-i}\sum_{j\geq 2}\binom{n-i}{j}\frac{(-1)^j}{(m-1)^j}\sum_{k=1}^{m}k^{j-1}\binom{bk-j}{bk-1} \\
=\;& \frac{1}{bm^n}\sum_{j\geq 2}\binom{n}{j}(-1)^j(m-1)^{n-j}\sum_{k=1}^{m}k^{j-1}\binom{bk-j}{bk-1}\sum_{i=0}^{n-j}\binom{n-j}{i}\frac{1}{(m-1)^i} \\
=\;& \frac{1}{bm^n}\sum_{j\geq 2}\binom{n}{j}(-1)^j(m-1)^{n-j}\sum_{k=1}^{m}k^{j-1}\binom{bk-j}{bk-1}\left(1+\frac{1}{m-1}\right)^{n-j} \\
=\;& \frac{1}{b}\sum_{j\geq 2}\binom{n}{j}\frac{(-1)^j}{m^j}\sum_{k=1}^{m}k^{j-1}\binom{bk-j}{bk-1} \\
=\;& \frac{1}{b}E[W_{m,n}]. \qquad (5.58)
\end{aligned}$$

Then, by (5.53), (5.54), (5.57) and (5.58) we have the following bounds

$$\frac{E[W_{m,n}]}{b}+\frac{n}{2bm} < E[S_{m,n}] \leq \frac{E[W_{m,n}]}{b}+\frac{n}{2bm}+1. \qquad (5.59)$$

Nevertheless, we can give an exact expression for a full table ($n = bm - 1$). Every real number x can be written as $x = \lfloor x \rfloor + \{x\}$, where $\{x\}$ denotes the fractional part of x [39]. The bounds given in (5.59) are based on the approximation of $\lfloor \frac{k+r}{b} \rfloor$ made in (5.53) and (5.54). This term appears after taking derivatives in (5.51) with respect to y. We could have replaced the exponent of y in (5.51) by

$$1+\left\lfloor\frac{k+r}{b}\right\rfloor = 1+\frac{k+r}{b}-\left\{\frac{k+r}{b}\right\}. \qquad (5.60)$$

When we take derivatives, the upper bound (5.54) is obtained from the first two addends of the right hand side of (5.60).

When the table is full, we can give an interpretation for the coefficient of $y^{\left\{\frac{k+r}{b}\right\}}$ in (5.51). The cost of searching for a random record in the table can be divided in two parts. The first is the number of buckets we have to probe. We add one to the cost, every time a new bucket is probed. The second part is the location of the record inside the bucket. In our model we do not consider this cost, and this is the discrepancy we have from $\frac{k+r}{b}$ (total cost of the two parts) and $\lfloor \frac{k+r}{b} \rfloor$ (cost of the first part). Since the table is full, the record to be searched has the same probability ($1/b$) of being in any position inside its bucket. Therefore, for the special case of a full table, the probability generating function for the second part is

$$G_{m,bm-1}(y) = \sum_{j=0}^{b-1} \frac{y^{\frac{j}{b}}}{b} \tag{5.61}$$

and therefore,

$$\mathbf{U}_y \mathbf{D}_y G_{m,bm-1}(y) = \sum_{j=0}^{b-1} \frac{j}{b^2} = \frac{b-1}{2b}. \tag{5.62}$$

So, we have proven

Lemma 5.2

$$\frac{1}{bm^{bm}} \sum_{i=0}^{bm-1} \binom{bm}{i+1} (m-1)^{bm-1-i} \sum_{k \geq 0} W_{m-1,bm-1-i,k} \sum_{r=0}^{i} \left\{ \frac{k+r}{b} \right\} \quad = \quad \frac{b-1}{2b}.$$

The most notable feature of Lemma 5.2, is that this sum is independent of m. Now, we can use it to prove

Theorem 5.5

$$E[S_{m,bm-1}] = \frac{E[W_{m,bm-1}]}{b} + \frac{m-1}{2bm} + 1. \tag{5.63}$$

Proof: We have to subtract (5.62) from the upper bound given in (5.59) for $n = bm - 1$. Then,

$$\begin{aligned} E[S_{m,bm-1}] &= \frac{E[W_{m,bm-1}]}{b} + \frac{bm-1}{2bm} + 1 - \frac{b-1}{2b} & (5.64) \\ &= \frac{E[W_{m,bm-1}]}{b} + \frac{m-1}{2bm} + 1. & (5.65) \end{aligned}$$

$\mathcal{QED}$

It is important to note that when $b = 1$, Theorem 5.5 tells us that

$$E[S_{m,m-1}] = \frac{1}{2}\left(1 + Q_0(m, m-1)\right). \tag{5.66}$$

as we already know by (4.2).

As a corollary, we can improve the bounds given in (5.59).

Corollary 5.2

$$\frac{E[W_{m,n}]}{b} + \frac{n}{2bm} + 1 - \frac{b-1}{2b} \leq E[S_{m,n}] \leq \frac{E[W_{m,n}]}{b} + \frac{n}{2bm} + 1. \tag{5.67}$$

5.6 Asymptotic Analysis

By Theorem 5.5, only the asymptotic behavior of $E[W_{m,bm-1}]$ has to be studied. For this purpose, we use the method of singularity analysis [31].

Our approach is as follows. We will first find an exponential generating function for $E[W_{m,bm-1}]$. As we shall see, this generating function is related with some variations of the Cayley generating function, introduced in chapter 2. Then we use multisection of series to express this generating function as a combination of known series. Finally, we use singularity analysis to find the desired asymptotics.

5.6.1 The Exponential Generating Function

First we require the following technical lemma.

Lemma 5.3 *Let* $I(v_c) = \int_v^{v_c} dv_{c-1} \int_v^{v_{c-1}} dv_{c-2} \ldots \int_v^{v_2} dv_1$. *Then,* $I(v_c) = \frac{(v_c-v)^{c-1}}{(c-1)!}$.

Proof: The proof is by induction on c.

If $c = 2$, then $\int_v^{v_2} dv_1 = (v_2 - v)$.

For the induction step, we have $I(v_c) = \int_v^{v_c} I(v_{c-1}) dv_{c-1}$. Then,

$$I(v_c) = \int_v^{v_c} \frac{(v_{c-1}-v)^{c-2}}{(c-2)!} dv_{c-1} \tag{5.68}$$

$$= \frac{(v_c-v)^{c-1}}{(c-1)!} \tag{5.69}$$

$\mathcal{QED}$

By (5.46), and using (2.23), we can express $E[W_{m,bm-1}]$ as follows

$$E[W_{m,mb-1}] = \sum_{i\geq 2} \binom{mb-1}{i} \frac{(-1)^i}{m^i} \sum_{k=1}^{m} (-1)^{bk-1} k^{i-1} \binom{i-2}{bk-1} \tag{5.70}$$

$$= -b \sum_{i\geq 2} \binom{mb-1}{i} \frac{(-1)^i}{(bm)^i} \sum_{k=1}^{m} (-1)^{bk} (bk)^{i-1} \binom{i-2}{bk-1}. \tag{5.71}$$

More generally, we will find the exponential generating function of

$$B_{a,c,d,n} \equiv \sum_{i\geq c} \binom{n}{i} (n+a)^{n-i} (-1)^i \sum_{k=1}^{m} (-1)^{bk} (bk)^{i-c+d} \binom{i-c}{bk-1}. \tag{5.72}$$

As usual, we omit the subscript b. If we denote

$$A_{i,d} \equiv (-1)^i \sum_{k=1}^{m} (-1)^{bk} (bk)^{i+d} \binom{i}{bk-1}, \tag{5.73}$$

then the outer sum in (5.72) can be rewritten as

$$B_{a,c,d,n} = \sum_{i \geq c} \binom{n}{i} (n+a)^{n-i} A_{i-c,d} \tag{5.74}$$

and so

$$E[W_{m,bm-1}] = \frac{-b}{(bm)^{bm-1}} B_{1,2,1,bm-1}. \tag{5.75}$$

The first goal is to find an exponential generating function for $B_{a,c,d,n}$.

$$\begin{aligned}
B_{a,c,d}(z) &= \sum_{n \geq c} \left(\sum_{i \geq c} \binom{n}{i} (n+a)^{n-i} A_{i-c,d} \right) \frac{z^n}{n!} \\
&= \sum_{i \geq c} A_{i-c,d} \frac{z^i}{i!} \sum_{n \geq i} (n+a)^{n-i} \frac{z^{n-i}}{(n-i)!} \\
&= \sum_{i \geq c} A_{i-c,d} \frac{z^i}{i!} \sum_{n \geq 0} (n+i+a)^n \frac{z^n}{n!}.
\end{aligned} \tag{5.76}$$

If $f(z)$ is the Cayley generating function defined in chapter 2, and we use (2.84), with $y = i + a$, then the inner sum of (5.76) can be simplified as follows

$$\begin{aligned}
B_{a,c,d}(z) &= \sum_{i \geq c} A_{i-c,d} \frac{z^i}{i!} \left(\frac{f(z)}{z} \right)^{a+i} \frac{1}{1 - f(z)} \\
&= \left(\frac{f(z)}{z} \right)^a \frac{1}{1 - f(z)} \sum_{i \geq c} A_{i-c,d} \frac{f(z)^i}{i!} \\
&= \left(\frac{f(z)}{z} \right)^a \frac{1}{1 - f(z)} \sum_{i \geq 0} A_{i,d} \frac{f(z)^{i+c}}{(i+c)!}.
\end{aligned} \tag{5.77}$$

Then, if we denote the exponential generating function of $A_{i,d}$ by $A_d(z)$, and use

Lemma 5.3, (5.77) tells us that for $d > 0$,

$$\begin{aligned}
B_{a,c,d}(z) &= \left(\frac{f(z)}{z} \right)^a \frac{1}{1 - f(z)} \int_0^{f(z)} dv_{c-1} \int_0^{v_{c-1}} \ldots \int_0^{v1} A_d(v) dv \\
&= \left(\frac{f(z)}{z} \right)^a \frac{1}{1 - f(z)} \int_0^{f(z)} A_d(v) dv \int_v^{f(z)} I(v_{c-1}) dv_{c-1} \\
&= \left(\frac{f(z)}{z} \right)^a \frac{1}{1 - f(z)} \int_0^{f(z)} \frac{(f(z) - v)^{c-1}}{(c-1)!} A_d(v) dv
\end{aligned}$$

$$= \left(\frac{f(z)}{z}\right)^a \frac{1}{1-f(z)} \int_0^z \frac{(f(z)-f(u))^{c-1}}{(c-1)!} A_d(f(u)) \mathbf{D}_u f(u) du. \tag{5.78}$$

Therefore, by (5.78), we have to find $A_d(z)$. By the definition of $A_{i,d}$,

$$\begin{aligned} A_d(z) &= \sum_{i\geq 0}\left((-1)^i \sum_{k\geq 1}(-1)^{bk}(bk)^{i+d}\binom{i}{bk-1}\right)\frac{z^i}{i!} \\ &= \sum_{k\geq 1}\frac{z^{bk}}{(bk-1)!}(bk)^{bk+d}\sum_{i\geq bk-1}\frac{(-z)^{i-bk}}{(i-bk+1)!}(bk)^{i-bk} \\ &= \sum_{k\geq 1}\frac{z^{bk}}{(bk-1)!}(bk)^{bk+d}\sum_{i\geq 0}\frac{(-z)^{i-1}}{i!}(bk)^{i-1} \\ &= -\frac{1}{z}\sum_{k\geq 1}\frac{z^{bk}}{(bk)!}(bk)^{bk+d}\sum_{i\geq 0}\frac{(-bkz)^i}{i!} \\ &= -\frac{1}{z}\sum_{k\geq 1}\frac{z^{bk}}{(bk)!}(bk)^{bk+d}e^{-bkz} \\ &= -\frac{1}{z}\sum_{k\geq 1}\frac{(bk)^{bk+d}}{(bk)!}\left(ze^{-z}\right)^{bk}. \end{aligned} \tag{5.79}$$

However, by (5.78), we do not need $A_d(z)$, but rather $A_d(f(z))$. Since we have

$f(z)e^{-f(z)} = z$,

$$\begin{aligned} A_d(f(z)) &= -\frac{1}{f(z)}\sum_{k\geq 1}\frac{(bk)^{bk+d}}{(bk)!}\left(f(z)e^{-f(z)}\right)^{bk} \\ &= -\frac{1}{f(z)}\sum_{k\geq 1}\frac{(bk)^{bk+d}}{(bk)!}z^{bk}. \end{aligned} \tag{5.80}$$

We have a case of multisection of series, as presented in chapter 2. By (2.82), we are dealing with a b-section of $f_d(z)$. So, by (2.88) for $t=0$,

$$A_d(f(z)) = -\frac{1}{bf(z)}\sum_{j=0}^{b-1} f_d\left(e^{\frac{2\pi i}{b}j}z\right). \tag{5.81}$$

So, (5.78) can be rewritten as

$$B_{a,c,d}(z) = -\frac{1}{b(c-1)!}\sum_{j=0}^{b-1}\left(\frac{f(z)}{z}\right)^a \frac{1}{1-f(z)}$$

$$\int_0^z (f(z)-f(u))^{c-1} f_d\left(e^{\frac{2\pi\mathbf{i}}{b}j}z\right)\frac{\mathbf{D}_u f(u)}{f(u)}du. \tag{5.82}$$

Although several interesting special cases can be derived from (5.82), we will only deal with the special case $a=1$, $c=2$ and $d=1$.

Since $f_1(z)=z\mathbf{D}_z[z\mathbf{D}_z f(z)]$, (2.83) can be applied twice, and so,

$$\begin{aligned} f_1(z) &= z\mathbf{D}_z\left[\frac{f(z)}{1-f(z)}\right]=z\mathbf{D}_z\left[\frac{1}{1-f(z)}-1\right] \\ &= \frac{z\mathbf{D}_z f(z)}{(1-f(z))^2}=\frac{f(z)}{(1-f(z))^3}. \end{aligned} \tag{5.83}$$

Therefore, (5.81) can be rewritten as

$$A(f(z))=-\frac{1}{bf(z)}\sum_{j=0}^{b-1}\frac{f\left(e^{\frac{2\pi\mathbf{i}}{b}j}z\right)}{\left(1-f\left(e^{\frac{2\pi\mathbf{i}}{b}j}z\right)\right)^3} \tag{5.84}$$

Finally, by putting (5.82) and (5.84) together we obtain

$$\begin{aligned} B_{1,2,1}(z) &= -\frac{1}{b}\left(\frac{f(z)}{z}\right)\frac{1}{1-f(z)}\sum_{j=0}^{b-1}\int_0^z(1-f(u))\frac{f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)}{\left(1-f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)\right)^3}\frac{\mathbf{D}_u f(u)}{f(u)}du \\ &+ \frac{1}{b}\left(\frac{f(z)}{z}\right)\sum_{j=0}^{b-1}\int_0^z\frac{f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)}{\left(1-f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)\right)^3}\frac{\mathbf{D}_u f(u)}{f(u)}du. \end{aligned} \tag{5.85}$$

Moreover, the first integral in (5.85) can be simplified by using (2.83).

$$\begin{aligned} &\int_0^z(1-f(u))\frac{f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)}{\left(1-f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)\right)^3}\frac{\mathbf{D}_u f(u)}{f(u)}du \\ &= \int_0^z\frac{f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)}{\left(1-f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)\right)^3}\frac{du}{u} \\ &= \int_0^{e^{\frac{2\pi\mathbf{i}}{b}j}z}\frac{f(u)}{(1-f(u))^3}\frac{du}{u} \\ &= \int_0^{e^{\frac{2\pi\mathbf{i}}{b}j}z}\frac{1}{(1-f(u))^2}\frac{\mathbf{D}_u f(u)}{du} \end{aligned}$$

$$= \frac{1}{\left(1 - f\left(e^{\frac{2\pi \mathbf{i}}{b}j}z\right)\right)} - 1. \tag{5.86}$$

Furthermore, when $j = 0$, the second integral in (5.85) can also be simplified.

$$\int_0^z \frac{\mathbf{D}_u f(u)}{(1 - f(u))^3} du = \frac{1}{2\left(1 - f\left(z\right)\right)^2} - \frac{1}{2}. \tag{5.87}$$

Finally, if we substitute (5.86) and (5.87) into (5.85), and use (2.83) then

$$\begin{aligned} B(z) &= -\frac{1}{2b}\left(\frac{f(z)}{z}\right)\frac{1}{(1 - f(z))^2} \\ &+ \frac{1}{b}\left(\frac{f(z)}{z}\right)\frac{1}{(1 - f(z))} - \frac{1}{2b}\left(\frac{f(z)}{z}\right) \\ &- \frac{1}{b}\left(\frac{f(z)}{z}\right)\frac{1}{1 - f(z)}\sum_{j=1}^{b-1}\left(\frac{1}{\left(1 - f\left(e^{\frac{2\pi \mathbf{i}}{b}j}z\right)\right)} - 1\right) \\ &+ \frac{1}{b}\left(\frac{f(z)}{z}\right)\sum_{j=1}^{b-1}\int_0^z \frac{f\left(e^{\frac{2\pi \mathbf{i}}{b}j}u\right)}{\left(1 - f\left(e^{\frac{2\pi \mathbf{i}}{b}j}u\right)\right)^3}\frac{du}{u(1 - f(u))}. \end{aligned} \tag{5.88}$$

5.6.2 Singularity Analysis

For simplicity, we will do singularity analysis on $-bzB(z)$. Let $r = e^{\frac{2\pi \mathbf{i}}{b}}$ be a b-th root of unity and let $z_j = r^{-j}/e$. Sometimes, depending on the context, z_j will be also denoted by u_j. Then if $\delta_j(z) = 2^{1/2}\sqrt{1 - z/z_j}$, by Lemma 2.7 [18, 33], $f(r^j z)$, admits the singular expansion at $z = z_j$

$$1 - \delta_j(z) + \frac{1}{3}\delta_j^2(z) + O(\delta_j(z)^3). \tag{5.89}$$

Since $f(z)$ is analytic at $z = z_j, j \neq 0$, then by (2.83)

$$f(z) = f(z_j) - \frac{f(z_j)}{2(1 - f(z_j))}\delta_j(z)^2 + O(\delta_j(z)^4). \tag{5.90}$$

First, let concentrate on the integral that appears in (5.88). For each j, the integrand has 2 singularities, one at u_j and the other u_0.

Around $u = u_j$, by (5.83) and (5.89)

$$f_1(r^j u) = \frac{f(r^j u)}{(1 - f(r^j u))^3} = \delta_j(u)^{-3} + O(\delta_j(u)^{-1}). \tag{5.91}$$

Moreover, $\frac{1}{u(1-f(u))}$ is analytic at u_j, because $j > 0$, and $f(u)$ has its only singularity at u_0. Then,

$$\frac{1}{u(1-f(u))} = \frac{1}{u_j(1-f(u_j))} + O(\delta_j(u)^2). \tag{5.92}$$

Therefore,

$$\frac{f_1(r^j u)}{u(1-f(u))} \quad = \quad \frac{\delta_j(u)^{-3}}{u_j(1-f(u_j))} + O(\delta_j(u)^{-1}). \tag{5.93}$$

We also know

$$\int_0^z \frac{\delta_j(u)^{-3}}{u_j} du = \int_0^z \frac{2^{-3/2}(1-u/u_j)^{-3/2}}{u_j} du = \delta_j(z)^{-1} - \sqrt{2} \tag{5.94}$$

and

$$\int_0^z \frac{\delta_j(u)^{-1}}{u_j} du = \int_0^z \frac{2^{-1/2}(1-u/u_j)^{-1/2}}{u_j} du = -\delta_j(z) + \sqrt{2}.. \tag{5.95}$$

Then, around $z = z_j$ we have

$$\int_0^z \frac{f\left(e^{\frac{2\pi \mathbf{i}}{b}j}u\right)}{\left(1-f\left(e^{\frac{2\pi \mathbf{i}}{b}j}u\right)\right)^3} \frac{du}{u(1-f(u))} = \frac{\delta_j(z)^{-1}}{(1-f(z_j))} + O(\delta_j(z)). \tag{5.96}$$

Similarly, around $u = u_0$, we find

$$\frac{1}{1-f(u)} = \delta_0(u)^{-1} - \frac{2}{3} + O(\delta_0(z)) \tag{5.97}$$

and

$$f_1(r^j u) = \frac{f(u_{-j})}{(1-f(u_{-j}))^3} + O(\delta_0(u)^2). \tag{5.98}$$

So by (5.95) we can conclude that around $z = z_0$,

$$\int_0^z \frac{f\left(e^{\frac{2\pi \mathbf{i}}{b}j}u\right)}{\left(1-f\left(e^{\frac{2\pi \mathbf{i}}{b}j}u\right)\right)^3} \frac{du}{u(1-f(u))}. \quad \sim \quad O(\delta_0(z)) \tag{5.99}$$

So from (5.90), (5.96) and (5.99), we find

$$\begin{aligned} & f(z)\sum_{j=1}^{b-1}\int_0^z \frac{f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)}{\left(1-f\left(e^{\frac{2\pi\mathbf{i}}{b}j}u\right)\right)^3}\frac{du}{u(1-f(u))} \\ \sim\ & \sum_{j=1}^{b-1}\left(\frac{f(z_j)\delta_j(z)^{-1}}{(1-f(z_j))}+O(\delta_j(z))\right)+O(\delta_0(z)). \end{aligned} \tag{5.100}$$

The other addends of (5.88), can be studied by using (5.89) and (5.90). So,

$$\begin{aligned} -bzB(z)\ \sim\ & \frac{\delta_0(z)^{-2}}{2}-\frac{\delta_0(z)^{-1}}{6}+O(\delta_0(z)) \\ -\ & \delta_0(z)^{-1}+O(\delta_0(z)) \\ +\ & \sum_{j=1}^{b-1}\left(\frac{f(z_j)\delta_j(z)^{-1}}{(1-f(z_j))}+O(\delta_j(z))\right)+\sum_{j=1}^{b-1}\frac{f(z_{-j})\delta_0(z)^{-1}}{(1-f(z_{-j}))}+O(\delta_0(z)) \\ -\ & \sum_{j=1}^{b-1}\left(\frac{f(z_j)\delta_j(z)^{-1}}{(1-f(z_j))}+O(\delta_j(z))\right)+O(\delta_0(z)) \\ =\ & \frac{\delta_0(z)^{-2}}{2}-\frac{7}{6}\delta_0(z)^{-1}+\sum_{j=1}^{b-1}\frac{f(z_{-j})\delta_0(z)^{-1}}{(1-f(z_{-j}))}+O(\delta_0(z)). \end{aligned} \tag{5.101}$$

Once the asymptotic expansion (5.101) is obtained, we can find the asymptotic expansion of B_n. In fact, by (5.75) we require the asymptotic expansion of $-bB_n/(n+1)^n$.

First, by the binomial theorem and Stirling's formula, we find [18]

$$\left[\frac{z^n}{n!}\right]\delta_0(z)^{-s}\sim\frac{\sqrt{\pi}n^{n+\frac{s-1}{2}}}{\Gamma\left(\frac{s}{2}\right)2^{\frac{s-1}{2}}}\left(1+\frac{3s^2-6s+2}{24n}+O\left(\frac{1}{n^2}\right)\right) \tag{5.102}$$

Because z is a factor of the left hand side of (5.101), we require the asymptotic behavior of $\frac{1}{n+1}\left[\frac{z^{n+1}}{(n+1)!}\right]\delta_0(z)^{-s}$. Since $n+1=mb$, by (5.101) and (5.102) we arrive at

Theorem 5.6

$$E[W_{mb,mb-1}]=\frac{1}{2}\sqrt{\frac{bm\pi}{2}}-\frac{7}{6}+\sum_{j=1}^{b-1}\frac{f(e^{\frac{2\pi\mathbf{i}}{b}j-1})}{(1-f(e^{\frac{2\pi\mathbf{i}}{b}j-1}))}+\frac{1}{24}\sqrt{\frac{\pi}{2bm}}+O\left(\frac{1}{bm}\right). \tag{5.103}$$

Then, by Theorem 5.5, we obtain our main theorem.

Theorem 5.7

$$\begin{aligned} bE[S_{m,bm-1}] &= \frac{\sqrt{2\pi}}{4}(bm)^{1/2} + \frac{1}{3} + \sum_{j=1}^{b-1} \frac{1}{(1 - f(e^{\frac{2\pi\mathbf{i}}{b}j-1}))} \\ &+ \frac{\sqrt{2\pi}}{48}(bm)^{-1/2} + O\left((bm)^{-1}\right). \end{aligned} \qquad (5.104)$$

As a particular case, when $b = 1$, we find

$$E[S_{m,m-1}] = \frac{\sqrt{2\pi}}{4} m^{1/2} + \frac{1}{3} + \frac{\sqrt{2\pi}}{48} m^{-1/2} + O\left(m^{-1}\right) \qquad (5.105)$$

as we already know [46].

5.7 A New Approach to the Study of $Q_{m,n,d}$

In this section we present a different approach to the study of the numbers $Q_{m,n,d}$, by introducing exponential generating functions. In the process, we define a new family of numbers that satisfy a recurrence resembling that of the Bernoulli numbers. We feel that this approach may be helpful in solving problems involving recurrences with truncated generating functions. So even though no new results related with hashing probing with buckets are obtained, we feel that this approach deserves a special study in its own right.

By (2.11) Theorem 5.1 gives the following recurrence relation

$$\begin{aligned} Q_{0,d}(z) &= 1 \\ Q_{m,d}(z) &= [e^z Q_{m-1,d}(z)]_{bm-d-1} \qquad (m \geq 1) \end{aligned} \qquad (5.106)$$

where $Q_{m,d}(z) = \sum_{n\geq 0} Q_{m,n,d}\frac{z^n}{n!}$. The main problem is that we are dealing with a recurrence that involves truncated generating functions.

Our strategy is to find an exponential generating function $T_d(z)$ such that

$$Q_{m,d}(z) = [T_d(z)e^{mz}]_{bm-d-1} \qquad (5.107)$$

where $T_d(z) = \sum_{k\geq 0} T_{k,d}\frac{z^k}{k!}$, for some coefficients $T_{k,d}$ to be determined, and independent of m. Again, b is an implicit parameter.

The intuition behind this idea is as follows. From (5.106), we obtain $Q_{m,d}(z)$ by multiplying the truncated generating function $Q_{m-1,d}(z)$ by the series e^z and then taking only the first $bm - d - 1$ terms of it. Moreover, $Q_{0,d}(z)$ is the first term of e^z. It is clear that without any truncations $Q_{m,d}(z)$ would be e^{mz}. However we have to consider a correcting factor originated by these truncations and this is the reason for defining this generating function $T_d(z)$. Then (5.107) gives a nonrecursive definition of $Q_{m,d}(z)$ that involves the truncated product of two series. The interesting aspect of this approach is that $T_d(z)$ does not depend on m. Furthermore, the only dependency on m is captured in the well known series that converges to e^{mz}. This section is devoted to the study of some properties of the numbers $T_{k,d}$.

By (2.11) and assuming (5.107),

$$Q_{m,n,d} = \sum_{k\geq 0} \binom{n}{k} T_{k,d} m^{n-k}. \qquad (0 \leq n < mb - d). \tag{5.108}$$

Actually, as we will see below, we need

$$Q_{m,d}(z) \;=\; [T_d(z)e^{mz}]_{b(m+1)-d-1} \tag{5.109}$$

Equation (5.109) is not an immediate consequence of Theorem 5.1 because the recursive definition of $Q_{m,n,d}$ is valid only up to $n = bm - d - 1$. So we have to prove

Lemma 5.4

$$Q_{m,n,d} = \sum_{k\geq 0} \binom{n}{k} T_{k,d} m^{n-k} \qquad (bm - d \leq n < (m+1)b - d). \tag{5.110}$$

By Theorem 5.1 and (2.11) we can reformulate (5.110) as

$$\sum_{k\geq 0} \binom{n}{k} T_{k,d} m^{n-k} = 0 \qquad (bm - d \leq n < (m+1)b - d). \tag{5.111}$$

The reason for Lemma 5.4 is as follows. By (5.106) and (5.107) we have

$$\begin{aligned}
Q_{m,d}(z) &= [e^z Q_{m-1,d}(z)]_{bm-d-1} \\
&= \left[e^z \left[T_d(z)e^{(m-1)z}\right]_{(m-1)b-d-1}\right]_{bm-d-1} \\
&= \left[\sum_{n\geq 0} \frac{z^n}{n!} \sum_{n=0}^{(m-1)b-d-1} \left(\sum_{k\geq 0} \binom{n}{k} T_{k,d}(m-1)^{n-k}\right) \frac{z^n}{n!}\right]_{bm-d-1} \\
&= \left[\sum_{n\geq 0} \frac{z^n}{n!} \sum_{n=0}^{bm-d-1} \left(\sum_{k\geq 0} \binom{n}{k} T_{k,d}(m-1)^{n-k}\right) \frac{z^n}{n!}\right]_{bm-d-1} \qquad (5.112)\\
&= \sum_{n=0}^{bm-d-1} \left(\sum_{j\geq 0} \binom{n}{j} \sum_{k\geq 0} \binom{j}{k} T_{k,d}(m-1)^{j-k}\right) \frac{z^n}{n!} \\
&= \sum_{n=0}^{bm-d-1} \left(\sum_{k\geq 0} \binom{n}{k} T_{k,d} \sum_{j=0}^{n-k} \binom{n-k}{j} (m-1)^{n-k-j}\right) \frac{z^n}{n!}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=0}^{bm-d-1} \left(\sum_{k\geq 0} \binom{n}{k} T_{k,d} m^{n-k} \right) \frac{z^n}{n!} \\
&= [T_d(z)e^{mz}]_{bm-d-1} .
\end{aligned}
$$

Note that (5.111) (and therefore Lemma 5.4) is required at step (5.112) above. Lemma 5.4 will follow as a consequence of Theorem 5.8.

The numbers $T_{k,d}$ satisfy some nice properties. The following can indeed be used as definition.

Theorem 5.8

$$\sum_j \binom{k}{j} \left(\left\lfloor \frac{k+d}{b} \right\rfloor \right)^{k-j} T_{j,d} = [k=0]. \tag{5.113}$$

To prove this theorem we require

Lemma 5.5

$$T_{k,d} = 0 \qquad 1 \leq k \leq b-d-1. \tag{5.114}$$

Proof:

If $m=1$, by Theorem 5.1

$$Q_{1,n,d} = \sum_{k\geq 0} \binom{n}{k} Q_{0,k,d} = \sum_{k\geq 0} \binom{n}{k} [k=0] = 1 \quad 0 \leq n \leq b-d-1 \tag{5.115}$$

and so by (5.108)

$$Q_{1,n,d} = \sum_{k\geq 0} \binom{n}{k} T_{k,d} \quad 0 \leq n \leq b-d-1 \tag{5.116}$$

If $n=0$, by (5.116), $T_{0,d}=1$.

We prove the lemma by induction on n. Note that as (5.116) is valid only up to $n=b-d-1$, so is this induction proof.

For $n=1$

$$Q_{1,1,d} \;=\; \binom{1}{0} T_{0,d} + \binom{1}{1} T_{1,d} = \binom{1}{0} 1 + \binom{1}{1} T_{1,d} = 1 \tag{5.117}$$

and so $T_{1,d}=0$.

Now, if we assume this lemma holds for up to $n = k - 1$, then for $n = k$,

$$Q_{1,k,d} = \sum_{j \geq 0} \binom{k}{j} T_{j,d} = \binom{k}{0} 1 + \binom{k}{k} T_{k,d} = 1 \tag{5.118}$$

and so $T_{k,d} = 0$. $\mathcal{QED}$

Since $\lfloor \frac{k+d}{b} \rfloor = 0$, for $0 \leq k \leq b - d - 1$ as a consequence we obtain

Corollary 5.3

$$\sum_j \binom{k}{j} \left(\left\lfloor \frac{k+d}{b} \right\rfloor \right)^{k-j} T_{j,d} = [k = 0] \quad (0 \leq k \leq b - d - 1). \tag{5.119}$$

Proof of Theorem 5.8:

When $0 \leq k \leq b - d - 1$ the theorem holds by Corollary 5.3.

Let $s = mb - d$ and $0 \leq r \leq b - 1$, for $m \geq 1$. By Theorem 5.1 we have

$$Q_{m+1,s+r,d} = \sum_{k=0}^{s+r} \binom{s+r}{k} Q_{m,k,d} = \sum_{k=0}^{s-1} \binom{s+r}{k} Q_{m,k,d} \tag{5.120}$$

as $Q_{m,k,d} = 0$ if $k \geq s$. Then by (5.108) we obtain

$$\sum_{k=0}^{s+r} \binom{s+r}{k} T_{k,d}(m+1)^{s+r-k} = \sum_{k=0}^{s-1} \binom{s+r}{k} \sum_{j=0}^{k} \binom{k}{j} T_{j,d} m^{k-j} \tag{5.121}$$

If we manipulate the right hand side of (5.121), and use (2.24), then

$$\begin{aligned}
\sum_{k=0}^{s-1} \binom{s+r}{k} \sum_{j=0}^{k} \binom{k}{j} T_{j,d} m^{k-j} &= \sum_{j=0}^{s-1} \binom{s+r}{j} T_{j,d} \sum_{k=j}^{s-1} \binom{s+r-j}{k-j} m^{k-j} \\
&= \sum_{j=0}^{s-1} \binom{s+r}{j} T_{j,d} \sum_{k=0}^{s-1-j} \binom{s+r-j}{k} m^{k} \\
&= \sum_{j=0}^{s-1} \binom{s+r}{j} T_{j,d} (m+1)^{s+r-j} \\
&- \sum_{j=0}^{s-1} \binom{s+r}{j} T_{j,d} \sum_{k=s-j}^{s+r-j} \binom{s+r-j}{k} m^{k}.
\end{aligned} \tag{5.122}$$

So considering together (5.121) and (5.122),

$$\sum_{k=s}^{s+r}\binom{s+r}{k}T_{k,d}(m+1)^{s+r-k} = -\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}\sum_{k=s-j}^{s+r-j}\binom{s+r-j}{k}m^k. \tag{5.123}$$

By changing the variable k to $k+s-j$ on the right hand side of (5.123) and then using (2.24) we find

$$\begin{aligned}\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}\sum_{k=s-j}^{s+r-j}\binom{s+r-j}{k}m^k &= \sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}\sum_{k=0}^{r}\binom{s+r-j}{s+k-j}m^{k+s-j}\\ &= \sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}\sum_{k=0}^{r}\binom{s+r-j}{r-k}m^{k+s-j}\\ &= \sum_{k=0}^{r}\binom{s+r}{r-k}\sum_{j=0}^{s-1}\binom{s+k}{j}T_{j,d}m^{k+s-j}\\ &= \sum_{k=0}^{r}\binom{s+r}{s+k}\sum_{j=0}^{s-1}\binom{s+k}{j}T_{j,d}m^{k+s-j}.\end{aligned}$$

After substituting the variable k by $k+s$ on the left hand side of (5.123), we obtain the identity

$$\sum_{k=0}^{r}\binom{s+r}{s+k}T_{s+k,d}(m+1)^{r-k} = -\sum_{k=0}^{r}\binom{s+r}{s+k}\sum_{j=0}^{s-1}\binom{s+k}{j}T_{j,d}m^{k+s-j}. \tag{5.124}$$

Now we prove the theorem by induction on r. Note that (5.124) is valid only if $r \le b-1$.

If $r=0$ in (5.124), then

$$T_{s,d} = -\sum_{j=0}^{s-1}\binom{s}{j}T_{j,d}m^{s-j} \tag{5.125}$$

and so

$$\sum_{j=0}^{s}\binom{s}{j}T_{j,d}m^{s-j} = 0. \tag{5.126}$$

By induction hypothesis, suppose that for $0 \le i \le r-1$, then

$$\sum_{j=0}^{s+i}\binom{s+i}{j}T_{j,d}m^{i+s-j} = 0 \tag{5.127}$$

and therefore

$$\sum_{j=0}^{s-1}\binom{s+i}{j}T_{j,d}m^{i+s-j} = -\sum_{j=s}^{s+i}\binom{s+i}{j}T_{j,d}m^{i+s-j}. \tag{5.128}$$

So for $i=r$, we can derive for the left hand side of (5.124)

$$\begin{aligned}
& -\sum_{k=0}^{r}\binom{s+r}{s+k}\sum_{j=0}^{s-1}\binom{s+k}{j}T_{j,d}m^{k+s-j} \\
= \; & -\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} + \sum_{k=0}^{r-1}\binom{s+r}{s+k}\sum_{j=s}^{s+k}\binom{s+k}{j}T_{j,d}m^{k+s-j} \\
= \; & -\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} + \sum_{k=0}^{r-1}\binom{s+r}{s+k}\sum_{j=0}^{k}\binom{s+k}{s+j}T_{s+j,d}m^{k-j} \\
= \; & -\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} + \sum_{j=0}^{r-1}\binom{s+r}{s+j}T_{s+j,d}\sum_{k=j}^{r-1}\binom{r-j}{k-j}m^{k-j} \\
= \; & -\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} + \sum_{j=0}^{r-1}\binom{s+r}{s+j}T_{s+j,d}\sum_{k=0}^{r-j-1}\binom{r-j}{k}m^{k} \\
= \; & -\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} + \sum_{j=0}^{r-1}\binom{s+r}{s+j}T_{s+j,d}\left((m+1)^{r-j} - m^{r-j}\right) \\
= \; & -\sum_{j=0}^{s-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} - \sum_{j=s}^{s+r-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} \\
& \qquad + \sum_{j=0}^{r-1}\binom{s+r}{s+j}T_{s+j,d}(m+1)^{r-j} \\
= \; & -\sum_{j=0}^{s+r-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} + \sum_{j=0}^{r-1}\binom{s+r}{s+j}T_{s+j,d}(m+1)^{r-j}.
\end{aligned} \tag{5.129}$$

Finally consider (5.124) and (5.129) together. Then

$$T_{s+r,d} = -\sum_{j=0}^{s+r-1}\binom{s+r}{j}T_{j,d}m^{r+s-j} \tag{5.130}$$

and so

$$\sum_{j=0}^{s+r}\binom{s+r}{j}T_{j,d}m^{r+s-j} = 0. \tag{5.131}$$

Since $k = s + r = mb - d + r$, then as $0 \le r \le b - 1$,

$$\left\lfloor \frac{n+d}{b} \right\rfloor = \left\lfloor \frac{bm+r}{b} \right\rfloor = m. \tag{5.132}$$

Therefore, after putting (5.131) and (5.132) together, we have proved the theorem for $mb - d \le k \le (m+1)b - d - 1$. Since this proof is valid for each $m \ge 1$, the theorem follows. $\mathcal{QED}$

As an important consequence of Theorem 5.8 we obtain the proof of Lemma 5.4. **Proof of Lemma 5.4:** By Theorem 5.8, for $0 \le r \le b - 1$, we have

$$\sum_{j=0}^{bm-d+r} \binom{bm-d+r}{j} T_{j,d} m^{bm-d+r-j} = 0. \tag{5.133}$$

The theorem follows easily, because by Theorem 5.1, $Q_{m,mb-d+r} = 0$, for $r \ge 0$. $\mathcal{QED}$

From Theorem 5.8 we can derive a recurrence to generate the numbers $T_{k,d}$ as follows

$$\begin{aligned} T_{0,d} &= 1 \\ T_{k,d} &= -\sum_{j=0}^{k-1} \binom{k}{j} \left(\left\lfloor \frac{k+d}{b} \right\rfloor \right)^{k-j} T_{j,d} \qquad (k > 0) \end{aligned} \tag{5.134}$$

A very curious property of these numbers is

Theorem 5.9

$$\sum_{d=0}^{b-1} T_{k,d} = \begin{cases} b & (\text{k} = 0) \\ -1 & (\text{k} = 1) \\ 0 & (\text{k} > 1). \end{cases} \tag{5.135}$$

Proof: By (5.108) and Theorem 5.2,

$$\begin{aligned} \sum_{d=0}^{b-1} Q_{m,n,d} &= \sum_{d=0}^{b-1} \sum_{k \ge 0} \binom{n}{k} T_{k,d} m^{n-k} & (5.136) \\ &= \sum_{k \ge 0} \binom{n}{k} m^{n-k} \sum_{d=0}^{b-1} T_{k,d} & (5.137) \\ &= bm^n - nm^{n-1}. & (5.138) \end{aligned}$$

Since this is an identity of two polynomials on m, the theorem follows immediately. $\mathcal{QED}$

There is also an inverse relation as follows.

Theorem 5.10

$$T_{n,d} = \sum_{k\geq 0} \binom{n}{k} (-1)^{n-k} Q_{m,k,d} m^{n-k} \qquad (n \leq (m+1)b - d - 1). \tag{5.139}$$

Proof: By (5.107) and Lemma 5.4,

$$Q_{m,d}(z) \;=\; \left[T_d(z)e^{mz}\right]_{(m+1)b-d-1} \tag{5.140}$$

and therefore we find the inverse relation

$$T_d(z) \;=\; \left[Q_{m,d}(z)e^{-mz}\right]_{(m+1)b-d-1}. \tag{5.141}$$

After taking the coefficient of $\frac{z^n}{n!}$ on both sides of (5.141), we obtain the result claimed.
$\mathcal{QED}$

It is interesting to note that this inverse relation is independent of the value of m, as long as $n \leq (m+1)b - d - 1$.

5.7.1 The Exponential Generating Function for $T_{k,0}$

In this section we find an implicit formula for $T_0(z)$. By (5.113),

$$\sum_{k\geq 0}\left(\sum_j \binom{k}{j}\left(\left\lfloor\frac{k}{b}\right\rfloor\right)^{k-j} T_{j,0}\right)\frac{z^k}{k!} = 1 \tag{5.142}$$

It is convenient to define $k = bs + \ell$ with $0 \leq \ell \leq b-1$. Let us study the left hand side of (5.142).

$$\begin{aligned}
&\sum_{k\geq 0}\left(\sum_j \binom{k}{j}\left(\left\lfloor\frac{k}{b}\right\rfloor\right)^{k-j} T_{j,0}\right)\frac{z^k}{k!}\\
&= \sum_{\ell=0}^{b-1}\sum_{s\geq 0}\sum_j \binom{bs+\ell}{j} s^{bs+\ell-j} T_{j,0}\frac{z^{bs+\ell}}{(bs+\ell)!}\\
&= \sum_{\ell=0}^{b-1}\sum_{j\geq 0} T_{j,0}\frac{z^j}{j!}\sum_{s\geq\lceil\frac{j-\ell}{b}\rceil} (bs)^{bs+\ell-j}\frac{(z/b)^{bs+\ell-j}}{(bs+\ell-j)!}
\end{aligned} \tag{5.143}$$

The inner sum is a b-section of

$$S(z) = \sum_{k\geq j-\ell} k^{k+\ell-j}\frac{(z/b)^{k+\ell-j}}{(k+\ell-j)!} \tag{5.144}$$

Therefore, if r is a b-th root of unity,

$$\begin{aligned}
&\sum_{\ell=0}^{b-1}\sum_{j\geq 0} T_{j,0}\frac{z^j}{j!}\sum_{s\geq\lceil\frac{j-\ell}{b}\rceil}(bs)^{bs+\ell-j}\frac{(z/b)^{bs+\ell-j}}{(bs+\ell-j)!}\\
=\;&\sum_{\ell=0}^{b-1}\sum_{j\geq 0} T_{j,0}\frac{z^j}{j!}\frac{1}{b}\sum_{n=0}^{b-1} r^{-n(\ell-j)}S(r^n z)\\
=\;&\frac{1}{b}\sum_{n=0}^{b-1}\sum_{j\geq 0} T_{j,0}\frac{z^j}{j!}\sum_{\ell=0}^{b-1} r^{-n(\ell-j)}\sum_{k\geq j-\ell} k^{k+\ell-j}\frac{(r^n z/b)^{k+\ell-j}}{(k+\ell-j)!}\\
=\;&\frac{1}{b}\sum_{n=0}^{b-1}\sum_{j\geq 0} T_{j,0}\frac{z^j}{j!}\sum_{\ell=0}^{b-1} r^{-n(\ell-j)}\sum_{k\geq 0}(k+j-\ell)^k\frac{(r^n z/b)^k}{k!} \qquad (5.145)
\end{aligned}$$

We now use (2.84) for the inner sum, and so

$$\begin{aligned}
&\frac{1}{b}\sum_{n=0}^{b-1}\sum_{j\geq 0} T_{j,0}\frac{z^j}{j!}\sum_{\ell=0}^{b-1} r^{-n(\ell-j)}\sum_{k\geq 0}(k+j-\ell)^k\frac{(r^n z/b)^k}{k!}\\
=\;&\frac{1}{b}\sum_{n=0}^{b-1}\sum_{j\geq 0} T_{j,0}\frac{z^j}{j!}\sum_{\ell=0}^{b-1} r^{-n(\ell-j)}\left(\frac{f(r^n z/b)}{r^n z/b}\right)^{j-\ell}\frac{1}{1-f(r^n z/b)}\\
=\;&\frac{1}{b}\sum_{n=0}^{b-1}\frac{1}{1-f(r^n z/b)}\sum_{j\geq 0} T_{j,0}\frac{(bf(r^n z/b))^j}{j!}\sum_{\ell=0}^{b-1}\left(\frac{z/b}{f(r^n z/b)}\right)^{\ell}\\
=\;&\frac{1}{b}\sum_{n=0}^{b-1}\frac{1}{1-f(r^n z/b)}\frac{\left(\frac{z/b}{f(r^n z/b)}\right)^b-1}{\frac{z/b}{f(r^n z/b)}-1}\sum_{j\geq 0} T_{j,0}\frac{(bf(r^n z/b))^j}{j!} \qquad (5.146)
\end{aligned}$$

Since $f(z)=ze^{f(z)}$, then $(z/b)/f(r^n z/b)=r^{-n}e^{-f(r^n z/b)}$, and as $r^{-nb}=1$, we have proved

Theorem 5.11

$$\frac{1}{b}\sum_{n=0}^{b-1}\frac{T_0(bf(r^n z/b)}{1-f(r^n z/b)}\;\frac{e^{-bf(r^n z/b)}-1}{r^{-n}e^{-f(r^n z/b)}-1}=1. \qquad (5.147)$$

When $b=1$, then (5.147) simplifies to

$$T_0(f(z))=1-f(z). \qquad (5.148)$$

and therefore $T_0=1$, $T_1=-1$, and $T_k=0, k\geq 2$, as we already know.

It would be very interesting to study (5.147) for other values of b.

Chapter 6

Conclusions and Future Work

Every night of the full moon, when I look to the sky, I know that far away a four year old girl is in deep communication with me, and is asking the moon to reunite her with her father very soon.

6.1 Conclusions

In this report we introduce a new mathematical transform that we call the Diagonal Poisson Transform. This transform, which resembles the Poisson Transform, is the main tool in the analysis presented in Chapter 4. In Chapter 3 we use it to study in a unified way various general classes of "Abel-like" recurrences, sums, and inverse relations.

In Chapter 4 we study the effect of the LCFS heuristic on the linear probing hashing scheme. We prove that, up to lower order terms, this heuristic achieves the optimal variance for the distribution of successful searches.

Finally, in Chapter 5, we present the first exact analysis of a problem related with an open addressing hashing scheme and multi-record buckets. We study the average cost for a successful search of a random element in a linear probing hash table with buckets of size b. We obtain the generating function for the Robin Hood heuristic, and then, for a full table, find an asymptotic expansion up to $O((bm)^{-1})$. In Section 5.7 we introduce a new family of numbers that verify a recurrence that resembles that of the Bernoulli numbers. These numbers may be used to give an alternative derivation of the analysis made in Chapter 5 and may prove very helpful in studying recurrences involving truncated generating functions.

Most of the formulae we have derived in this report have been checked with the assist of the Maple system [13].

6.2 Future Work

Several problems arise from the results presented in this report.

It would be very interesting to find new areas that can be studied with the help of the Diagonal Poisson Transform. This tool seems to be particularly useful when "Abel-like" problems arise. Furthermore, we would like to find problems in which new classes of recurrences, sums or inverse relations can be studied using it. Other problems of mathematical interest involve finding new properties of this transform, as well as to define an algebra (similar to the Q-Algebra defined by Knuth [49]) of the functions that satisfy the Transfer Lemma.

For the analysis of hashing with buckets, we would like to find an exact expression for the variance, as well as an asymptotic expansion when the table is full. It would also be interesting to study the variance for other heuristics such as the standard FCFS or the LCFS approach.

Another area of research is to study other open addressing schemes such as uniform or random probing. For uniform probing, Larson [55] presents an asymptotic analysis, in which $m, n \to \infty$ while the ratio m/n is constant. Later, for random probing, Ramakrishna [77] gives explicit expressions for the cost of successful searches. However, he only solves them numerically. New ideas have to be introduced to analyze these algorithms. The methodology used in Chapter 5 to do the asymptotic analysis could be used in the analysis of these schemes.

It would be very interesting to better understand the numbers $T_{k,d}$ defined in Section 5.7. A development of a theory for them may help in studying other recurrences that involve truncated generating functions. These numbers seem not to appear in *The Encyclopedia of Integer Sequences* [83], although some special cases were handled by the Superseeker. We would like to find other problems in which these numbers appear.

Bibliography

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

[2] L.V. Ahlfors. *Complex Analysis.* McGraw-Hill, 1966.

[3] D.J. Aldous. Hashing with linear probing, under non-uniform probabilities. Technical Report TR-88, University of California, Berkeley. Dept. of Statistics, February 1987.

[4] O. Amble and D.E. Knuth. Ordered hash tables. *Computer Journal*, 17(2):135–142, 1974.

[5] P. Bachmann. *Die analytische Zahlentheorie.* Teubner, Leipzig, 1894.

[6] R. Bayer and E.M. McCreight. Organization and maintenance of large ordered indexes. *Acta Informatica*, 1(3):173–189, 1972.

[7] E.A. Bender. Asymptotic methods in enumeration. *SIAM Review*, 16(4):485–515, 1974.

[8] J. Bernoulli. *Ars Conjectandi, opus posthumum.* Basel, 1713. Reprinted in *Die Werke von Jakob Bernoulli*, volume 3, 107-286.

[9] I.F. Blake and A.G. Konheim. Big buckets are (are not) better! *J. ACM*, 24(4):591–606, October 1977.

[10] R.P. Brent. Reducing the retrieval time of scatter storage techniques. *C. ACM*, 16(2):105–109, 1973.

[11] A. Broder. Two counting problems solved via string encodings. In A. Apostolico and Z. Galil, editors, *Combinatorial Algorithms on Words*, volume 12 of NATO Advance Science Institute Series. Series F: Computer and System Sciences, pages 229–240. Springer Verlag, 1985.

[12] W. Buchholz. File organization and addressing. *IBM Systems Journal*, 2:86–111, 1963.

[13] B.W.Char, K.O.Geddes, G.H.Gonnet, B.L.Leong, M.B.Monagan, and S.M.Watt. *MAPLE V Reference Manual.* Springer-Verlag, 1991.

[14] S. Carlsson, J.I. Munro, and P.V. Poblete. On linear probing hashing. Unpublished Manuscript.

[15] A. Cauchy. Exercises de mathématiques. pages 62–73. 1826.

[16] P. Celis. *Robin Hood Hashing.* PhD thesis, Computer Science Department, University of Waterloo, April 1986. Technical Report CS-86-14.

[17] P. Celis, P.-Å. Larson, and J.I. Munro. Robin hood hashing. In *26th IEEE Sympusium on the Foundations of Computer Science*, pages 281–288, 1985.

[18] K. J. Compton and C. Ravishankar. Expected deadlock time in a multiprocessing system. *JACM*, 42(3):562–583, 1995.

[19] L. Comtet. *Advanced Combinatorics.* Reidel, Dordrecht, 1974.

[20] N. G. de Bruijn. *Asymptotic Methods in Analysis.* North Holland, third edition, 1958. Reprinted by Dover, 1981.

[21] J.-L. Lagrange (de la Grange). Nouvelle méthode pour résoudre les équations littérales par le moyen des séries. *Mém. Acad. Roy. Sci. Belles-Lettres de Berlin*, 24, 1770.

[22] L. Euler. Methodus generalis summandi progressiones. *Commentarii academiæ scientiarum Petropolitanæ*, 6:68–97, 1732. Reprinted in his *Opera Omnia*, series 1, volume 14, 42-72.

[23] M. A. Evgrafov. *Analytic Functions.* Dover Publications, Inc., New York, 1978.

[24] R. Fagin, J. Nievergelt, N. Pippenger, and H. R. Strong. Extendible hashing - a fast access method for dynamic files. *ACM Transactions on Database Systems*, 4(3):315–344, 1979.

[25] P. Flajolet, , B. Salvy, and P. Zimmermann. Lambda-upsilon-omega. the 1989 cookbook. Research Report 1073, INRIA, Aug 1989.

[26] P. Flajolet, , B. Salvy, and P. Zimmermann. Automatic average-case analysis of algorithms. *Theoretical Computer Science*, 79:37–109, 1991.

[27] P. Flajolet. Mathematical methods in the analysis of algorithms and data structures. In E. Börger, editor, *Trends in Theoretical Computer Science*, pages 225–304. Computer Science Press, Rockville, MD, 1988.

[28] P. Flajolet, P. Grabner, P. Kirschenhofer, and H. Prodinger. On Ramanujan's Q–function. Research Report 1760, INRIA, Oct 1992.

[29] P. Flajolet and A. M. Odlyzko. The average height of binary trees and other simple trees. *Journal of Computer and System Sciences*, 25:171–213, 1982.

[30] P. Flajolet and A. M. Odlyzko. Random mapping statistics. In J.-J. Quisquater and J. Vandewalle, editors, *Advances in Cryptology*, volume 434 of *Lecture Notes in Computer Science*, pages 329–354. Springer Verlag, 1990. Proceedings of EUROCRYPT'89, Houtalen, Belgium, April 1989.

[31] P. Flajolet and A. M. Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3(2):216–240, 1990.

[32] P. Flajolet, M Régnier, and R. Sedgewick. Some uses of the mellin integral transform in the analysis of algorithm. In A. Apostolico and Z. Galil, editors, *Combinatorial Algorithms on Words*, volume 12 of NATO Advance Science Institute Series. Series F: Computer and System Sciences, pages 241–254. Springer Verlag, 1985. (invited lecture).

[33] P. Flajolet and R. Sedgewick. The average case analysis of algorithms: Complex asymptotics and generating functions. Research Report 2026, INRIA, Sept 1993.

[34] P. Flajolet and R. Sedgewick. The average case analysis of algorithms: Counting and generating functions. Research Report 1888, INRIA, Apr 1993.

[35] G.H. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures.* Addison-Wesley, 1991. Second Edition.

[36] G.H. Gonnet and J.I. Munro. Efficient ordering of hash tables. *SIAM Journal on Computing*, 8(3):463–478, 1979.

[37] G.H. Gonnet and J.I. Munro. The analysis of linear probing sort by the use of a new mathematical transform. *Journal of Algorithms*, 5:451–470, 1984.

[38] I. P. Goulden and D. M. Jackson. *Combinatorial Enumeration.* John Wiley, New York, 1983.

[39] R.L. Graham, D.E. Knuth, and O.Patashnik. *Concrete Mathematics.* Addison-Wesley Publishing Company, 1989.

[40] D.H. Greene and D.E. Knuth. *Mathematics for the Analysis of Algorithms.* Birkhäuser, Boston, 1990. Third Edition.

[41] G.H. Hardy and E.M Wright. *An Introduction to the Theory of Numbers.* Oxford University Press, 1979.

[42] P. Henrici. *Applied and computational complex analysis.* J. Wiley, New York, 1974. Three volumes.

[43] P. Jacquet and M Régnier. Trie partitioning process: Limiting distributions. In A. Apostolico and Z. Galil, editors, *Proceedings of the 11th Colloquim on Trees in Algebra and Programming (CAAP)*, volume 214 of Lecture Notes in Computer Science, pages 196–210. Springer Verlag, March 1986.

[44] P. Jacquet and W. Szpankowski. Asymptotic behaviour of the lempel-ziv parsing scheme and digital search trees. *Theoretical Computer Science*, 144, 1995.

[45] T. Kløve. Bounds for the worst case probability of undetected error. *IEEE Information Theory*, 41:298–300, 1995.

[46] D.E. Knuth. *The Art of Computer Programming*, volume 3. Addison-Wesley Publishing Company, 1973.

[47] D.E. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley Publishing Company, 1973.

[48] D.E. Knuth. *The Art of Computer Programming*, volume 1. Addison-Wesley Publishing Company, 1973.

[49] D.E. Knuth. Analysis of optimum caching. *Journal of Algorithms*, 6:181–199, 1985.

[50] D.E. Knuth and G. S. Rao. Activity in an interleaved memory. *IEEE Transactions on Computers*, C-24:943–944, 1975.

[51] D.E. Knuth and A. Schönhage. The expected linearity of a simple equivalence algorithm. *Theoretical Computer Science*, 6:281–315, 1978.

[52] A.G. Konheim and B. Weiss. An occupancy discipline and applications. *SIAM Journal on Applied Mathematics*, 6(14):1266–1274, 1966.

[53] J.-L. Lagrange and A.-M. Legendre. Rapport sur deux mémoires d'analyse du professeur bürmann. *Mémmoires de l'Institut National des Sciences ...*, 2 (an VII):13–17, 1799.

[54] E. Landau. *Handbuch der Lehre von der Verteilung der Primzahlen.* Two volumes. Teubner, Leipzig, 1909.

[55] P.-Å. Larson. Analysis of uniform hashing. *JACM*, 30(4):805–819, 1983.

[56] P.-Å. Larson. Linear hashing with overflow-handling by linear probing. *ACM Transaction on Database Systems*, 10(1):75–89, 1985.

[57] P.-Å. Larson. Linear hashing with separators - a dynamic hashing scheme achieving one-acess retrieval. *ACM Transaction on Database Systems*, 13(3):366–388, 1988.

[58] G. Louchard and W. Szpankowski. Average profile and limiting distribution for a phrase size in the lempel-ziv parsing algorithm. *IEEE Information Theory*, 41, 1995.

[59] C. MacLaurin. *Collected Letters.* edited by Stella Mills. Shiva Publishing, Nantwich, Cheshire, 1982.

[60] H. Mendelson. Analysis of linear probing with buckets. *Information Systems*, 8(3):207–216, 1983.

[61] H. Mendelson and U. Yechiali. A new approach to the analysis of linear probing schemes. *J. ACM*, 27:474–483, 1980.

[62] J. W. Moon. Counting labelled trees. *Canadian Mathematical Monographs*, 1970.

[63] R. Morris. Scatter storage techniques. *CAMC*, 11(1):38–44, 1968.

[64] A. M. Odlyzko. Periodic oscillations of coefficients of power series that satisfy functional equations. *Advances in Mathematics*, 44:180–205, 1982.

[65] A. M. Odlyzko. Asymptotic enumeration methods. In R. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics.* 1995.

[66] F. W. J. Olver. *Asymptotics and Special Functions.* Academic Press, 1974.

[67] P. O'Neil. *Data Base. Principles, Programming and Performance.* Morgan Kaufmann Publishers, Inc., 1994.

[68] T. Papadakis. *Skip Lists and Probabilistic Analysis of Algorithms.* PhD thesis, Computer Science Department, University of Waterloo, May 1993. Technical Report CS-93-28.

[69] W. W. Peterson. Addressing for random-access storage. *IBM Journal of Research and Development*, 1(2):130–146, 1957.

[70] G.Ch. Pflug and H.W. Kessler. Linear probing with a nonuniform address distribution. *JACM*, 34(2):397–410, 1987.

[71] B. Pittel. Linear probing: The probable largest search time grows logarithmically with the number of records. *Journal of Algorithms*, 8:236–249, 1987.

[72] P.V. Poblete. Approximating functions by their poisson transform. *Information Processing Letters*, 23:127–130, 1986.

[73] P.V. Poblete and J.I. Munro. Last-come-first-served hashing. *Journal of Algorithms*, 10:228–248, 1989.

[74] P.V. Poblete, J.I. Munro, and T. Papadakis. The binomial transform and its application to the analysis of skip lists. In *3rd European Symposium on Algorithms*, 1995.

[75] P.V. Poblete, A. Viola, and J.I. Munro. The analysis of a hashing secheme by a new transform. In *2nd European Symposium on Algorithms*, 1994.

[76] W. Pugh. Skip lists: A probabilistic alternative to balanced trees. *Comm. ACM*, 33(6):668–676, 1990.

[77] M.V. Ramakrishna. Analysis of random probing hashing. *Information Processing Letters*, 31:83–90, 1989.

[78] S. Ramanujan. Question 294. *Journal of the Indian Mathematical Society*, 3:128, 1911.

[79] S. Ramanujan. On question 294. *Journal of the Indian Mathematical Society*, 4:151–152, 1912.

[80] J. Riordan. *Combinatorial Identities.* Wiley, New York, 1968.

[81] G. Schay and W. G. Spruth. Analysis of a file addressing method. *CACM*, 5(8):459–462, 1962.

[82] R. Sedgewick. Mathematical analysis of combinatorial algorithms. In G. Louchard and G. Latouche, editors, *Probability Theory and Computer Science*, pages 123–205. Academic Press, Inc., 1983.

[83] N.J.A. Sloane and S. Plouffe. *The Encyclopedia of Integer Sequences.* Academic Press, 1995.

[84] W. Szpankowski. On asymptotics of certain sums arising in coding theory. 1995. Unpublished Manuscript.

[85] M. Tainiter. Addressing for random-access storage with multiple bucket capacities. *JACM*, 10:307–315, 1963.

[86] J. H. van Lint. *Introduction to Coding Theory.* Springer-Verlag, New York, 1982.

[87] J.S. Vitter and P. Flajolet. Average-case analysis of algorithms and data structures. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume A, pages 431–524. Elsevier, Amsterdam, 1990.

[88] H. S. Wilf. *Generatingfunctionology.* Academic Press, 1994.