INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Analytic Variations on Bucket Selection and Sorting

Hosam Mahmoud, Philippe Flajolet, Philippe Jacquet, Mireille Régnier

N ° 3399

Avril 1998

THÈME 2

Rapport de recherche

ISSN 0249-6399

Analytic Variations on Bucket Selection and Sorting

Hosam Mahmoud, Philippe Flajolet, Philippe Jacquet, Mireille Régnier

Thème 2 — Génie logiciel
et calcul symbolique
Projet Algo

Rapport de recherche— Avril 1998 — 24 pages

Abstract: We provide complete average-case as well as probabilistic analysis of the cost of bucket selection and sorting algorithms. Two variations of bucketing (and flavors therein) are considered: distributive bucketing (large number of buckets) and radix bucketing (recursive with a small number of buckets, suitable for digital computation). For Distributive Selection a compound Poisson limit is established. For all other flavors of bucket selection and sorting, central limit theorems underlying the cost are derived by asymptotic techniques involving perturbation of Rice's integral and contour integration (saddle point methods). In the case of radix bucketing, periodic fluctuations appear in the moments of both the selection and sorting algorithms.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : (33) 01 39 63 55 11 – Télécopie : (33) 01 39 63 53 30

Variations analytiques sur la sélection et le tri par calcul d'adresse

Résumé : Cet article propose des analyses complètes en moyenne ainsi qu'en distribution des principaux algorithmes de sélection et de tri par calcul d'adresse. Les algorithmes considérés répartissent les données en sous-groupes disjoints, appelés des "seaux" (*"buckets"*). Deux variantes algorithmiques sont considérées: les algorithmes opérant par distribution (et utilisant un grand nombre de seaux) et les méthodes "lexicographiques" (qui sont récursives, reposent sur la décomposition en caractères des données, et utilisent un petit nombre de seaux). Pour la selection par distribution, une loi de Poisson composée décrit le coût de l'algorithme. Dans tous les autres cas, une loi centrale limite (gaussienne) décrit les fonctions de coûts. Les méthodes utilisées reposent sur l'asymptotique complexe (analyse de col, perturbation d'intégrales de Rice) et, dans le cas des algorithmes lexicographiques, elles mettent en évidence des phénomènes de fluctuation périodique.

ANALYTIC VARIATIONS ON BUCKET SELECTION AND SORTING

Hosam Mahmoud[1] Philippe Flajolet[2] Philippe Jacquet[2] Mireille Régnier[2]

April 14, 1998

ABSTRACT

We provide complete average-case as well as probabilistic analysis of the cost of bucket selection and sorting algorithms. Two variations of bucketing (and flavors therein) are considered: distributive bucketing (large number of buckets) and radix bucketing (recursive with a small number of buckets, suitable for digital computation). For Distributive Selection a compound Poisson limit is established. For all other flavors of bucket selection and sorting, central limit theorems underlying the cost are derived by asymptotic techniques involving perturbation of Rice's integral and contour integration (saddle point methods). In the case of radix bucketing, periodic fluctuations appear in the moments of both the selection and sorting algorithms.

1. Bucketing

Interest in distributive sorts peaked in the late 1970's as they offered efficient alternatives to standard comparison-based algorithms. The latter class of sorting algorithms provably has $\Omega(n \ln n)$ complexity both in the worst case and on average to sort n keys. Distributive Sort, invented by Dobosiewicz in 1978, provided for the first time a class of sorting algorithms with only $O(n)$ average cost under suitable uniformity assumptions on the data.

Bucketing is a term used for identifying numerical keys by intervals. A bucket sort algorithm accomplishes its task by "distributing" the keys into containers called buckets (hence the name Distributive Sort for some flavors of the algorithm); see Devroye (1986) for general background. Bucketing is usually achieved by simple arithmetic operations and is at the core of many fast algorithms like distributive sorting (Dobosiewicz (1978)), extendible hashing (Fagin, Nievergelt, Pippenger and Strong (1979)), selection by distributive partitioning (Alison and Noga (1981)), etc. Various associated data structures arise, such as tries (Knuth (1973)), hash trees and N-trees (Ehrlich (1981)). These algorithms and data structures in turn motivated several mathematical analyses like the analysis of distributive sorting (Devroye and Klincsek (1981)), the analysis of random hash trees and N-trees (Tamminen (1981)) and the analysis of tries (Flajolet (1983), Jacquet and Régnier (1986), Pittel (1986), Régnier and Jacquet (1989)).

In this paper we examine the behavior of various flavors of bucket selection and of bucket sorting. (Selection is the process of finding an order statistic, that is, finding the element of a given rank in an unsorted array.) The algorithmic side of the study is motivated by the need of general selection algorithms that work universally for all order statistics. Such a need arises for instance in inference (see Andrews, Bickel,

[1]Department of Statistics, The George Washington University, Washington, D.C. 20052, U.S.A.
[2]INRIA, Rocquencourt 78153-Le Chesnay, France.

Hampel, Huber, Rogers, and Tukey (1972)). Of course, there are many algorithms for *specific* order statistics like those for minima, maxima and medians that are custom-made for a specific situation while only few *universal* selection algorithms are known, of which the most notable is Quickselect (also known as FIND (Hoare (1961)). The algorithmic interest in bucket sorting is that for some flavors its average behavior is $O(n)$. The mathematical side of the study is motivated by the appearance of new types of recurrence equations and the reassurance gained from probabilistic analysis that these algorithms have good average behavior and a reasonable degree of concentration around average.

The difference between various flavors of bucketing algorithms is in the choice of the number of buckets and in what they do in the buckets. In one flavor, *Distributive Bucketing*, keys are distributed over a large number of buckets. In another flavor, *Radix Bucketing*, a recursive algorithm, a fixed number of buckets is used. (The terms fixed and large are with respect to n, the number of keys to be sorted, as $n \to \infty$.)

We shall denote the number of buckets by b, which may or may not depend on n. To unify the treatment across the various flavors, we assume that n keys are drawn from the Uniform—(0,1]. From a probabilistic standpoint, this assumption is equivalent to the general hypothesis that underlies hashing schemes and according to which "good" hash functions exist to distribute keys drawn from some domain (or probability distribution) uniformly over a hash table. The unit interval is divided into a number of equally long intervals $(0, 1/b], (1/b, 2/b], \ldots ((b-1)/b, 1]$, where b is a number to be specified later. Think of these intervals as indexed from left to right by $1, 2, \ldots, b$. In all bucketing flavors, a key K is thrown into the $\lceil bK \rceil$th bucket. This is one type of a *bucketing operation*.

For the rest of the paper we shall use the following standard notation. A binomial random variable on n trials and rate of success p per trial is denoted by $B(n,p)$; a Poisson random variable with parameter λ will be called $\mathcal{P}(\lambda)$; the normal random variable with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$. Convergence in distribution is denoted by $\xrightarrow{\mathcal{D}}$, whereas exact equality in distribution is denoted by $\stackrel{\mathcal{D}}{=}$. Throughout, we shall refer to the convergence in distribution to a normal law as a central limit result, where convergence in distribution is used in the usual probability sense of convergence at continuity points of the distribution function. Thus for example, a sequence of random variables satisfies $X_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$, when for each x:

$$\lim_{n\to\infty} \mathbf{P}\{X_n \le x\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

The operator $[z^n]$ extracts the nth coefficient of a generating function. The nth harmonic number $\sum_{j=1}^{n} 1/j$ is denoted by H_n.

The paper is organized as follows. The selection problem precedes sorting as it is perhaps the more interesting problem both mathematically and algorithmically. Section 2 is devoted to bucket selection, where a general functional equation valid for all flavors is derived. Two subsections of Section 2 specialize the problem to its various flavors: Subsection 2.1 is for the analysis of Distributive Selection (including its recursive flavor) where a compound Poisson limit is derived, and Section 2.2 is for the analysis of Radix Selection, where a central limit theorem is derived with periodic fluctuations in the moments of the cost. The method of proof for this case provides a first example of perturbation of Rice's integral.

Section 3 takes up bucket sorting in the same order—a general functional equation in the introduction of the section is followed by two subsections: 3.1 is for the analysis of Distributive Sort (including its recursive flavor) and Section 3.2 is for the analysis of Radix Sort. For all bucket sorts discussed, central limit theorems are found with periodic fluctuations in the moments of Radix Sort. Although central limit theorems were not known before in the context of bucket sorting, the methods of analysis are connected to some classical as well recent analyses in tries (Jacquet and Régnier (1988)) and hashing with linear probing (Flajolet, Poblete and Viola (1998)). Therefore our arguments in Section 3 will be sketchy. Section 4 concludes with a discussion.

2. Bucket Selection

We are given an unsorted array of n elements and a rank $m \in \{1, \ldots, n\}$ and we wish to identify the numeric key among the n data with rank m, i.e., the mth order statistic. Suppose N_j is the share of the jth bucket, $j = 1, \ldots, b$. Under our probability assumption, the shares $N_1, \ldots, N_b$ have a joint multinomial distribution on n trials and rate of success $1/b$ per trial for each bucket. In particular, the distribution of any individual bucket's share is $B(n, 1/b)$.

A bucket selection algorithm then continues its search for the desired order statistic by considering the keys in the bucket indexed i, for an index i satisfying

$$N_1 + \cdots + N_{j-1} < m \leq N_1 + \cdots + N_j; \tag{1}$$

interpret a sum as 0 when it is empty. It should be noted that *conditional* on the size of the bucket containing the desired order statistic, the rank of the order statistic is still uniformly distributed on the size of that bucket. That is, if the jth bucket, with size N_j, contains the desired order statistic, a subsequent search looks for a uniformly distributed rank over the set $\{1, \cdots, N_j\}$.

In choosing a large number b of buckets a natural value is $b = n$. The choice $b \gg n$ will not help much as it gives several empty buckets. The choice $b \ll n$ will lead to more collisions. The idea in $b = n$ is that on average a bucket contains one key. When the correct bucket containing the desired order statistic is identified, the process then terminates: the only key in the bucket must be the mth order statistic. Probabilistically, very few keys fall in the correct bucket and either a recursive action or switching to some other selection algorithm should lead to rather fast termination. The bucketing operation is performed on the initial list of keys, then followed by a stochastically small number of operations. In the recursive flavor the algorithm is applied within the buckets after rescaling interval lengths within each recursive call. In a direct recursive implementation, there are however rare instances where the algorithm takes a long time, or even possibly infinitely many operations to perform the desired task. (This situation arises only if infinite precision is available for keys; in practice finite precision is used and infinite recursion will never happen.) A bucket selection distributing the n keys over $b = n$ buckets at the first step will be called a *Distributive Selection* algorithm. Dobosiewicz's sorting algorithm (and the selection algorithm derived from it) avoids infinite recursion by adjusting (and readjusting within each recursive call) the boundaries of the given interval to the range of the data, thus ensuring that the minimum and maximum will fall in two different buckets; at least one key is removed at every recursive level guaranteeing finiteness.

A recursive flavor with a fixed number of buckets may be appealing because of its ease of implementation. *Radix Selection* uses a fixed number $b = B$ of buckets. The method may be especially attractive when keys are stored in digital form according to their expansion in some base (radix), with the binary expansion being the most typical at the level of machine data. When the digits or bits are readily available, the bucketing operation may be simplified to accessing the bits of a key or to hardwired machine arithmetic such as register shift instructions. The method is also directly applicable to non-numeric data, as in selecting from or sorting DNA strands.

In all cases we shall consider that m itself is chosen randomly according to a uniform distribution on its range. Thus m becomes a random variable $M_n =$ Uniform–$[1\,..\,n]$, independent of the shares of the buckets. There are two reasons for this choice: first it models naturally the way the algorithm is used on random inputs (a random array of data and a random rank to be selected); second it serves as an averaging process. This averaging process, introduced in Mahmoud, Modarres and Smythe (1995), works as a smoothing operator over all the fixed cases. The average cost of the randomized case is the average of the averages of the fixed cases, or a *grand average* stamped with the general character of the individual fixed cases. Similarly we can explore a grand variance. One can even speak of an "average distribution" which is the distribution of the randomized cost, and is an averaging of all distributions of the individual fixed cases and is indicative of their classes of probability distribitions.

At the outset of this analysis we write a distributional equation for all versions of the algorithm. Suppose that within the bucket containing the desired order statistic algorithm A, say, is used to complete the selection. Suppose such an algorithm performs Y_j operations of some kind to sort a random input of j keys. If A is some standard algorithm, the operation in question is typically a comparison of a pair of keys. Generally Y_j is a random variable having its own probability distribution. We have to take into account the possible difference between the bucketing operation, which typically involves arithmetics and ceils or floors, and the operations of A. If we take the unit cost to be that of a bucketing operation, a single operation of A may cost α. A typical value of α for comparison-based selection and sorting algorithms on a modern computer is 0.1–0.2 (a measure of the speed of a comparison of keys relative to a bucketing operation as the hashing-like operation already considered). In recursive flavors A will be the same algorithm used at the top level selecting or sorting by the same bucketing operations, i.e. $\alpha = 1$.

Let C_n be the number of bucketing operations involved in the selection of a rank

$M_n = \text{Uniform–}[1\,..\,n]$ and let I_j be the indicator of the event (1). Then, for $n \geq 2$,

$$C_n \stackrel{\mathcal{D}}{=} \alpha(I_1 Y_{N_1} + I_2 Y_{N_2} + \cdots + I_b Y_{N_b}) + n, \tag{2}$$

and $C_0 = 0$, and $C_1 = 0$. The quantity

$$Z_n := I_1 Y_{N_1} + I_2 Y_{N_2} + \cdots + I_b Y_{N_b} \tag{3}$$

is the only stochastic component of the cost and captures the essence of the extra cost after the first layer of bucketing. Introduce the probability generating functions $\phi_n(u) = \mathbf{E}[u^{Z_n}]$ and $\psi_n(u) = \mathbf{E}[u^{Y_n}]$.

Lemma 1 *For $n \geq 2$,*

$$\phi_n(u) = \frac{1}{nb^{n-1}} \sum_{j=1}^{\infty} j\psi_j(u)(n-1)^{n-j}\binom{n}{j}.$$

Proof. Coupled with a conditioning argument, Equation (3) gives us

$$\begin{aligned}
\phi_n(u) &= \sum_{\substack{i_1+\cdots+i_b=n \\ 0<m\leq n}} \mathcal{P}_n \mathbf{E}\big[u^{I_1 Y_{N_1}+I_2 Y_{N_2}+\cdots+I_b Y_{N_b}} \mid N_1 = i_1, \ldots, N_b = i_b, M_n = m\big] \\
&= \sum_{\substack{i_1+\cdots+i_b=n \\ 1\leq m\leq i_1}} \mathcal{P}_n \mathbf{E}[u^{Y_{i_1}}] + \sum_{\substack{i_1+\cdots+i_b=n \\ i_1<m\leq i_1+i_2}} \mathcal{P}_n \mathbf{E}[u^{Y_{i_2}}] \\
&\qquad + \cdots + \sum_{\substack{i_1+\cdots+i_b=n \\ i_1+i_2+\cdots+i_{b-1}<m\leq i_1+i_2+\cdots+i_b}} \mathcal{P}_n \mathbf{E}[u^{Y_{i_b}}],
\end{aligned}$$

where $\mathcal{P}_n$ is the probability $\mathbf{P}\{N_1 = i_1, \ldots, N_b = i_b, M_n = m\}$. By symmetry, the b sums are identical; we can use b copies of the first. We have assumed M_n and the shares of the buckets to be independent. Using this independence and the respective distributions, we arrive at

$$\begin{aligned}
\phi_n(u) &= b \sum_{i_1+\cdots+i_b=n} i_1 \mathbf{E}[u^{Y_{i_1}}] \binom{n}{i_1, \ldots, i_b} \frac{1}{b^n} \times \frac{1}{n} \\
&= \frac{1}{nb^{n-1}} \sum_{j=1}^{n} j\psi_j(u) \binom{n}{j} \sum_{i_2+\cdots+i_b=n-j} \frac{(n-j)!}{i_2! \ldots i_b!};
\end{aligned}$$

the sums above run over all non-negative integer solutions of their defining equation. $\square$

2.1. Distributive Selection

This version (with $b = n$) of bucket selection is easiest to analyze when the algorithm switches to a standard selection algorithm A within a bucket. Extracting the coefficient of u^k from the equality of Lemma 1 when specialized to the case $b = n$:

$$\mathbf{P}\{Z_n = k\} = \frac{1}{n^n} \sum_{j=1}^{\infty} j \mathbf{P}\{Y_j = k\}(n-1)^{n-j} \binom{n}{j}. \tag{4}$$

The term $n^{-n}(n-1)^{n-j}\binom{n}{j} = n^{-j}(1-1/n)^{n-j}\binom{n}{j}$ is the probability that $B(n, 1/n) = j$, which for any given j converges to $e^{-1}/(j-1)!$ by the standard approximation of the binomial distribution of $B(n, 1/n)$ to $\mathcal{P}(1)$. At any fixed k, passing to the limit (as $n \to \infty$) gives us

$$\lim_{n\to\infty} \mathbf{P}\{Z_n = k\} = \sum_{j=1}^{\infty} \mathbf{P}\{Y_j = k\} \frac{e^{-1}}{(j-1)!}. \tag{5}$$

The Poisson probabilities appearing on the right hand side indicate that the number of additional operations after the first level of bucketing is like the behavior of A on $\mathcal{P}(1)$ random number of keys. As the last limiting calculation is true for any fixed k, we can express the behavior of Z_n simply as convergence in distribution (see Chung (1974)).

Theorem 1 *In the selection of a randomly chosen rank from among n random keys by Distributive Selection, whose algorithm within a bucket makes Y_j operations for random selection in a file of size j, the extra cost after the first layer of bucketing satisfies a limiting compound Poisson law (in the sense of (5))*

$$Z_n \xrightarrow{\mathcal{D}} Y_{\mathcal{P}(1)+1}.$$

Within the buckets we may use a very simple algorithm, even an inefficient one, like a selection algorithm derived from the standard *Straight Selection Sort* (see Knuth (1973)). (We do not particularly care for a good algorithm in the buckets because with high probability each contains only a few number of keys.) Selection sort works by choosing successive minima, and a derived algorithm for finding the kth order statistic is one that stops right after finding the first k minima, using $(n-1) + (n-2) + \cdots + (n-k)$ comparisons. When A is this simplistic algorithm

$$\mathbf{E}[Z_n] = \frac{4}{3}.$$

On average, less than two additional comparisons will be needed in the bucket containing the order statistic and the average cost of the whole process is then

$$\mathbf{E}[C_n] = n + \frac{4}{3}\alpha + o(1).$$

Similarly,

$$\mathbf{Var}[C_n] \to \frac{29}{36}\alpha^2.$$

Note that the variance is $O(1)$ and does not grow with n, a highly desirable feature in algorithmic design.

For *Recursive Bucket Selection*, when A is the bucket selection itself that is used at the first stage, the analysis remains the same. However (4) becomes a rather interesting functional equation on distribution functions. In this flavor, the operations of A are themselves bucketing operations, costing one unit each ($\alpha = 1$) and

$$C_n - n \xrightarrow{\mathcal{D}} C_{\mathcal{P}(1)+1}.$$

No explicit characterization of the limit is known.

For the first two moments we can get good approximations as follows. Starting with (4),

$$\mathbf{E}[C_n] - n = \sum_{j=1}^{\infty} j\mathbf{E}[C_j]\,\mathbf{P}\{B(n, 1/n) = j\}. \tag{6}$$

No exact solution is known for this recurrence. An asymptotic guess will not be of much use either, because the *beginning terms* have more weight than those in the tail of the series. Nevertheless, accurate approximations can be made.

For a first approximation of averages, replace the binomial probabilities by the limiting Poisson probabilities as we did for the distribution of C_n itself:

$$\lim_{n\to\infty} \Big(\mathbf{E}[C_n] - n\Big) = \sum_{j=1}^{\infty} \mathbf{E}[C_j] \frac{e^{-1}}{(j-1)!}.$$

The series converges since $\mathbf{E}[C_j] \le 2j$ as can be proved by an easy induction on (6). We can therefore evaluate the numerical value of the series, producing as many digits as desired for any accuracy.

Corollary 1 *The average cost of Recursive Bucket Sort satisfies the asymptotic relation*

$$\mathbf{E}[C_n] = n + 3.011281835\ldots + o(1).$$

By a similar process (details suppressed) we can obtain the limiting variance—we find

$$\mathbf{Var}[C_n] \to 11.39004484\ldots.$$

In all the above versions of *Distributive Selection* (both recursive and non-recursive) we could have even considered second order (or higher) approximations by $\mathcal{P}(1)$ to $B(n, 1/n)$ and developed an asymptotic expansion for $\mathbf{E}[C_n]$, an approach first suggested by Gonnet (1984). For example, carrying the approximations of the *Recursive Distributive Selection* one step further, for any fixed j,

$$\mathbf{P}\Big\{B\Big(n, \frac{1}{n}\Big) = j\Big\} = \frac{e^{-1}}{j!}\Big(1 - \frac{1}{2n}\Big)\Big[1 - \frac{j}{2n}(j-3)\Big] + O\Big(\frac{1}{n^2}\Big).$$

So,

$$\mathbf{E}[C_n] = n + \sum_{j=0}^{\infty} \frac{\mathbf{E}[C_j]e^{-1}}{(j-1)!}\Bigg(1 - \Big(\frac{1}{2} + \frac{1}{2}j(j-3)\Big)\frac{1}{n} + O\Big(\frac{1}{n^2}\Big)\Bigg).$$

Again, all the series involved are convergent and can therefore be computed accurately up to any desired precision, yielding

$$\mathbf{E}[C_n] = n + 3.011281835\ldots - \frac{1.751820621\ldots}{n} + O\Big(\frac{1}{n^2}\Big).$$

The lower order terms are of the form $O(n^{-j})$, $j = 2, 3, \ldots$, and the coefficients of O can be determined, if so desired, by more elaborate calculation. By a completely analogous process we can obtain a series expansion for the variance.

2.2. Radix selection

As discussed in the introduction, *Radix Selection* is a recursive version of bucket selection with a fixed number of buckets. To analyze this version, we want to develop asymptotics for the expression of Lemma 1 when b is a fixed number, B say, as $n \to \infty$. It is usually the case that one develops generating functions for the sequence $\phi_n(u)$, however in our case, the recurrence suggests developing a generating function for $n\phi_n(u)$, as it is the combination that appears on both sides of rearranged version of the expression of Lemma 1. So, we introduce the bivariate exponential generating function

$$\Phi(u, z) = \sum_{n=0}^{\infty} n\phi_n(u)\frac{z^n}{n!}.$$

It then follows from Lemma 1, by multiplying its two sides by z^n and summing over $n \geq 2$ (the range of validity of the recurrence), that

$$\Phi(u, z) = B\Phi\Big(u, \frac{uz}{B}\Big)\exp\Big(\frac{uz(B-1)}{B}\Big) + z(1-u).$$

The function $z(1-u)$ appears to adjust for the boundary conditions.

We shall consider the case $B = 2$ (the most practical case for computer data in the form of bit strings). This case illustrates simply all the principles involved in the analysis. Extension to higher (fixed) B follows the same route and poses no particular difficulty; the result for general B is stated at the end of this section.

We start with the equation

$$\Phi(u, z) = 2\Phi\Big(u, \frac{uz}{2}\Big)e^{uz/2} + z(1-u), \tag{7}$$

When $u = 1$, the relation (7) simplifies and has the obvious solution ze^z, as it should, given the combinatorial origin of the problem. Introduce a suitable multiplier:

$$Q(u, z) = e^{uz/2}e^{u^2z/4}e^{u^3z/8}\cdots = \exp\left(\frac{uz}{2-u}\right),$$

so that $h(u, z) := \Phi(u, z)/Q(u, z)$ satisfies a simpler functional equation of the form $h(u, z) = 2h(u, uz/2) + a(u, z)$ that is solvable by indeterminate coefficients; then there result alternating (but finite) binomial sum expressions for $[z^n]\Phi(u, z)$, and these are natural candidates for a treatment by Rice's integral. The coefficients $h_k(u) = k!\,[z^k]h(u,z)$ satisfy

$$h_k(u) = 2^{1-k}u^k h_k(u) + a_k(u),$$

where $a_k(u) = k!\,[z^k]a(u,z)$ and $a(u,z) = z(1-u)/Q(u,z)$. Thus, one has

$$h_k(u) = (-1)^{k-1}\frac{k(1-u)}{1-2^{1-k}u^k}\left(\frac{u}{2-u}\right)^{k-1},$$

and

$$\phi_n(u) = \frac{1}{n}\sum_{k=1}^{n}\binom{n}{k}\left(\frac{u}{2-u}\right)^{n-k} h_k(u).$$

Inserting the expression for $h_k(u)$, we get

$$\phi_n(u) = \frac{1}{n}\left(\frac{u}{2-u}\right)^{n-1}\sum_{k=1}^{n}\binom{n}{k}(-1)^{k-1}k\frac{1-u}{1-2^{1-k}u^k}. \tag{8}$$

In particular, $\phi_0(u) = \phi_1(u) = 1$, and

$$\phi_2(u) = \frac{u^2}{2-u^2},\ \phi_3(u) = \frac{u^3(2+u^2)}{(2-u^2)(4-u^3)},\ \phi_4(u) = \frac{u^4(8+8u^2+4u^3+u^5)}{(2-u^2)(4-u^3)(8-u^4)}.$$

Observe that each $\phi_n(u)$ is a rational fraction with poles at $2^{1-1/k}\exp(2ij\pi/k)$ for $2 \le k \le n$, and with degree $\frac{1}{2}n(n+1)-1$. It is at any rate clear from (8) that each $\phi_n(u)$ is analytic in the disk $|u| \le 1.41$.

Expressions like (8) are natural candidates for a representation via integrals of the so-called Rice type. Define the beta function kernel

$$\beta_n(s) := \frac{(n-1)!}{(1-s)(2-s)\cdots(n-s)} = \frac{\Gamma(n)\Gamma(1-s)}{\Gamma(n+1-s)}.$$

Then, assuming that u is near 1, for instance $|u-1| \le \frac{1}{10}$, the standard theory (based on the residue theorem) provides

$$\phi_n(u) = -\left(\frac{u}{2-u}\right)^{n-1}\frac{1}{2i\pi}\int_D \frac{1-u}{1-2^{1-s}u^s}\beta_n(s)\,ds, \tag{9}$$

where D is a positively oriented contour that encircles the points $1,\ldots,n$ but no other singularity of the integrand. The next step consists in evaluating (9) by residues after enlarging the contour (see Flajolet and Sedgewick (1995) for an exposition of the general theory). The factor of $\beta_n(s)$ in the integrand has poles at

$$\chi_k(u) = \frac{\ln 2 + 2ik\pi}{\ln 2 - \ln u}.$$

Extending the integral in the usual way to a large contour then yields an exact representation

$$\phi_n(u) = \left(\frac{u}{2-u}\right)^{n-1}\frac{1-u}{\ln(2/u)}\sum_{k=-\infty}^{\infty}\beta_n\left(\chi_k(u)\right). \tag{10}$$

As $u \to 1$, the terms corresponding to $k \neq 0$ in the sum tend to definite limits while they are multiplied by a factor of $1-u$. Thus, the dominant contribution in (10) comes from terms corresponding to $k=0$. There, we have $\chi_0(1)=1$ while the factor $\Gamma(1-\chi_0(u))$ cancels out with $(1-u)$, so that we obtain, as expected, $\lim_{u\to 1}\phi_n(u)=1$.

The same limit process provides next the value of the average, $\mathbf{E}[C_n] := \phi_n'(1)$, namely,

$$\mathbf{E}[C_n] = 2n - \frac{3}{2} + \frac{H_{n-1}}{\ln 2} - \frac{1}{\ln 2}\sum_{k\in\mathbf{Z}\setminus\{0\}} \beta_n(\chi_k), \qquad \chi_k \equiv \chi_k(1) = 1 + \frac{2ik\pi}{\ln 2}.$$

By standard asymptotics of the gamma function

$$\begin{aligned}\mathbf{E}[C_n] &= 2n + \log_2 n - \frac{3}{2} + \frac{\gamma}{\ln 2} - \frac{1}{\ln 2}\sum_{k\in\mathbf{Z}\setminus\{0\}} \Gamma(1-\chi_k)e^{2ik\pi\log_2 n} + o(1) \\ &= 2n + \log_2 n + P(\log_2 n) + o(1),\end{aligned}$$

where P is the bounded periodic function given by the Fourier expansion above; here $\gamma = 0.5772156\ldots$ is Euler's constant. A similar but more complicated computation gives the variance of the distribution as

$$\mathbf{Var}[C_n] = \phi_n''(1) + \phi_n'(1) - \phi_n'(1)^2 = 2n + O(\ln^2 n).$$

We are now in a position to analyze the limit distribution. We want to prove that the normalized variable

$$C_n^\star = \frac{C_n - 2n}{\sqrt{2n}}$$

converges to the standard normal variate. For that purpose, it is enough to consider

$$\phi_n^\star(t) = e^{-t\sqrt{2/n}}\phi_n(e^{t/\sqrt{2n}}),$$

and establish that it converges pointwise to $e^{t^2/2}$ for any fixed t. This only requires a local analysis of $\phi_n(u)$ for u near 1, and is very much in line with our previous analyses. We thus reexamine Equation (10) with $u = e^{t/\sqrt{2n}}$. As before, only the term corresponding to $k=0$ dominates asymptotically when $u \to 1$, and we find

$$\phi_n^\star(t) = e^{-t\sqrt{2/n}}\left(\frac{e^{t/\sqrt{2n}}}{2-e^{t/\sqrt{2n}}}\right)^{n-1}\big(1+o(1)\big),$$

for *fixed* t as $n\to\infty$. But then, local expansions of the first factor yield

$$\phi_n^\star(t) = e^{t^2/2} + o(1).$$

We can then conclude on convergence to the standard normal distribution by continuity theorems for either Laplace transforms (taking t real) or Fourier transforms (taking t imaginary).

The proof generalizes immediately to the case of a general bucket parameter B fixed, where the bucketing operations amount to the extraction of digits from numbers in the B–ary number system. We summarize the results in the following.

Theorem 2 *Let C_n be the number of bucket operations (digit extractions) performed by Radix Select using B (fixed) buckets to find a randomly chose order statistic among n keys. If we set $\lambda_B = B/(B-1)$, then as $n \to \infty$,*

$$\mathbf{E}[C_n] = \lambda_B n + \log_B n + P(\log_B n) + o(1), \qquad \mathbf{Var}[C_n] = \frac{\lambda_B}{B-1} n + O(\ln^2 n),$$

where P is a smooth periodic function. The law of C_n is asymptotically normal:

$$\frac{C_n - \lambda_B n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{\lambda_B}{B-1}\Big).$$

We observe in passing that, from these analytic developments, the distribution of $X_n - n + 1$ "resembles" that of the sum of $n-1$ independent geometric variables with mean $1/B$. The analysis also suggests, via the Berry-Esseen inequalities, that the speed of convergence to the normal limit distribution is $O(n^{-1/2} \ln^2 n)$.

We know that normal limiting distributions are often associated with a perturbative analysis of singularities or critical points. For instance, Jacquet and Szpankowski (1991) derive a normal distribution from an alternating sum by a perturbative Mellin-like approach. The present case provides the first example of a perturbative analysis applied to Rice's integral representations.

3. Bucket Sorting

For whatever flavor of complete sorting by bucketing, the sorting process must continue in *all* buckets. Assume that the buckets' sorting algorithm uses Y_j operations to sort a file of size j. The distributional equation for C_n, the cost of sorting, is therefore like (2) but without indicators:

$$C_n \stackrel{\mathcal{D}}{=} \alpha(Y_{N_1} + Y_{N_2} + \cdots + Y_{N_b}) + n, \tag{11}$$

valid for $n \geq 2$; with $b = n$ in *Distributive Sort*, and $b = B$ (fixed) in *Radix Sort*. Again, the quantity

$$W_n := Y_{N_1} + Y_{N_2} + \cdots + Y_{N_b}$$

is the only stochastic component and we shall focus our study on it. Introduce the probability generating functions $\phi_n(u) = \mathbf{E}[u^{W_n}]$ and $\psi_n(u) = \mathbf{E}[u^{Y_n}]$. By a conditioning argument on the shares of the buckets, like that we used in selection, we have a decomposed representation as a product following from the conditional independence of the action in the buckets.

Lemma 2 *For $n \geq 2$,*

$$\phi_n(u) = \frac{n!}{b^n} \sum_{i_1 + \cdots + i_b = n} \frac{\psi_{i_1}(u)}{i_1!} \times \cdots \times \frac{\psi_{i_b}(u)}{i_b!};$$

the sum runs over all non-negative integer solutions of the equation $i_1 + \cdots + i_b = n$.

3.1. Distributive Sorting

For this form of bucket sorting we consider the functional equation of Lemma 2 when specialized to the case $b = n$. Specialized to the case $b = n$, the right hand side of this equation can be viewed as the nth coefficient in a generating function. We can express the specialized equation in the form

$$\phi_n(u) = \frac{n!}{n^n}[z^n]\Big(\sum_{j=0}^{\infty} \psi_j(u)\frac{z^j}{j!}\Big)^n.$$

This representation admits the following central limit result. The theorem follows from a rather general result in Flajolet, Poblete and Viola (1998) for hashing with linear probing which broadly states that, under suitable conditions, coefficients of bivariate generating functions raised to large powers follow a Gaussian law. We work through some of the details to obtain the first two moments as well in order to completely characterize the limiting normal distribution. We shall assume that within the buckets a "reasonable" sorting algorithm is used. All standard sorting algorithms have polynomial time worst-case behavior. Thus we assume for example that $Y_j \le j^\theta$, for some fixed $\theta > 0$.

Theorem 3 *Let C_n be the cost of Distributive Sort to sort n random keys. Suppose for some fixed $\theta > 0$, the algorithm applied in the buckets uses $Y_j \le j^\theta$ operations costing α units each (the unit being the cost of one bucketing operation). Then*

$$\frac{C_n - (1+\alpha\mu)n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \alpha^2\sigma^2),$$

where

$$\mu = e^{-1}\sum_{j=0}^{\infty}\frac{\mathbf{E}[Y_j]}{j!}, \qquad \sigma^2 = e^{-1}\sum_{j=0}^{\infty}\frac{\mathbf{E}[Y_j^2]}{j!} - \mu^2.$$

Proof. Consider $\phi_n(e^{it})$, and $\psi_n(e^{it})$, the characteristic functions of W_n and Y_n, respectively. Let us denote the bivariate generating function $\sum_{j=0}^{\infty}\psi_j(u)z^j/j!$ by $\Psi(u,z)$. Then by Cauchy's formula (see any standard book on complex analysis),

$$\phi_n(u) = \frac{n!}{2\pi i n^n}\oint_D \frac{\Psi^n(u,z)}{z^{n+1}}\,dz,$$

where D is any closed contour enclosing the origin. We choose D to be a particular contour consisting of the line segment connecting $c - iM$ to $c + iM$ (for some fixed $c > 0$ and some large M) and a closing (left) arc of the circle centered at the origin and passing through these two points. As $M \to \infty$, one can check that the integral on the arc approaches 0.

To evaluate the remaining line integral asymptotically by the saddle point method (see Wong (1988), for instance), write the last equation in the form:

$$\phi_n(u) = \frac{n!}{2\pi i n^n}\int_{c-i\infty}^{c+i\infty} e^{nh(u,z)}\,\frac{dz}{z}, \tag{12}$$

where by definition $h(u,z) = \ln\{\Psi(u,z)/z\}$. We shall eventually let $u \to 1$. The saddle point is the special value of z that solves the saddle point equation

$$z\Psi'(u,z) = \Psi(u,z); \tag{13}$$

The derivative is with respect to z. One can verify that near $u = 1$, $z = 1+O(1-u)$ is a saddle point of the integrand in (12). Therefore, we deform the line of integration to become one connecting $c - i\infty$ to $c + i\infty$ and going through $z = 1$ through an angle picking up the steepest descent of the function h.

Replacing $n!$ by its asymptotic Stirling's approximation, and using the saddle point method (as $u \to 1$)

$$\phi_n(u) \sim \frac{\Psi^n(u,1)}{e^n\sqrt{h''(u,1)}},$$

where the derivative in the denominator is with respect to z. Let us now set $u = e^{it}$, and expand $\Psi^n(e^{it}, z)$ around $t = 0$ (i.e. around $u = 1$) in powers of t with coefficients that are moments of the cost. Denoting the first two moments of Y_j respectively by μ_j and s_j, near $t = 0$ we obtain

$$\begin{aligned}\Psi^n(e^{it}, z) &= \exp\Big\{n \ln \sum_{j=0}^{\infty} \psi_j(e^{it})\frac{z^j}{j!}\Big\} \\ &= \exp\Big\{n \ln \Big[\sum_{j=0}^{\infty}\Big(\frac{z^j}{j!} + \frac{\mu_j z^j}{j!} it - \frac{s_j z^j}{2!\,j!}t^2\Big) + O(t^3)\Big]\Big\}.\end{aligned}$$

Letting $u \to 1$ implies $z \to 1$, and we have

$$\Psi^n(e^{it}, 1) = \exp\Big\{n \ln \Big[e\Big(1 + \mu i t - \sigma^2\frac{t^2}{2}\Big) + O(t^3)\Big]\Big\},$$

where

$$\mu := e^{-1}\sum_{j=0}^{\infty}\frac{\mu_j}{j!}, \qquad \sigma^2 := e^{-1}\sum_{j=0}^{\infty}\frac{s_j}{j!} - \mu^2.$$

Expanding the logarithm with the usual calculus equality $\ln(1+x) = x - x^2/2 + O(x^3)$, we get

$$\phi_n(e^{it}) = \mathbf{E}\big[e^{W_n it}\big] \sim \frac{e^{\mu n i t - \sigma^2 n t^2/2 + O(nt^3)}}{\sqrt{h''(e^{it},1)}}.$$

The assumption that the sorting algorithm within buckets is reasonable, with worst-case polynomial time, guarantees $\mu_j = O(j^\theta)$ and $s_j = O(j^{2\theta})$. Thus both series in μ and σ^2 converge and one can accurately compute their values up to any number of places.

Finally, set $t = v/\sqrt{n}$ for a fixed v, and let $n \to \infty$ (so indeed $u \to 1$). The reader can check that $h''(v/\sqrt{n}, 1) \to 1$. So,

$$\mathbf{E}\Big[\exp\Big\{\frac{C_n - (1+\alpha\mu)n}{\sqrt{n}} iv\Big\}\Big] \to e^{-\alpha^2\sigma^2 t^2/2};$$

the right hand side is the characteristic function of $\mathcal{N}(0, \alpha^2\sigma^2)$ and of course convergence of characteristic functions implies weak convergence. $\square$

For the standard *Straight Selection Sort* (see Knuth (1973)), $Y_j \equiv j(j-1)/2$. So, $\mu = 1/2$ and $\sigma^2 = 5/4$, and the particular central limit theorem becomes

$$\frac{C_n - (1+\alpha/2)n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{5}{4}\alpha^2\Big).$$

In the particular *Recursive Bucket Sort*, when A, the sorting algorithm within a bucket is bucket sort itself, the result still holds, although the saddle point argument becomes slightly more delicate. The subtlety here is that two technical points arise:

1. The function $\Phi(u,z) := \sum_{j=0}^{\infty} \phi_j(u) z^j/j!$ involves the *unknown* functions $\phi_j(u)$ and the saddle point equation (13) becomes:

$$z\Phi'(u,z) = \Phi(u,z).$$

   The expansion technique that led to determining the saddle point at $z = 1 + O(1-u)$ still holds verbatim without any changes.

2. The condition that $Y_j < j^\theta$, for $\theta > 0$, uniformly for all sample space points is no longer valid. As discussed in the introduction, there are rare cases where the Recursive Bucket Sort may take arbitrarily long time. However, all we really need in the proof is not a uniform bound on Y_j itself, but rather on its mean and variance. It is an easy induction to show that $\mu_j \leq 2j$ and $s_j \leq 6j$. Thus both series in μ and σ^2 converge and one can accurately compute their values up to any number of places from exact recurrences.

With $\alpha = 1$, the central limit theorem takes the following special form.

Corollary 2 *The cost of Recursive Bucket Sort satisfies*

$$\frac{C_n - (2.302023901\ldots)n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 6.456760413\ldots).$$

The cost here is the total number of bucketing operations performed throughout the entire sorting.

3.2. Radix Sort

For *Radix Sort*, $b = B$ fixed and the algorithm is recursive in the buckets ($\alpha = 1$). The bucketing operations here amount to the extraction of digits from numbers in the B–ary number system. Let $\xi_n(u)$ be the probability generating function of the cost C_n, and $\Xi(u,z) = \sum_{j=0}^{\infty} \xi_j(u) z^j/j!$. Equation (11) then becomes a distributional recurrence and Lemma 2 gives us the recursive functional equation

$$\Xi(u,z) = \Xi^B\Big(u, \frac{uz}{B}\Big) + z(1-u). \tag{14}$$

The analysis is adapted with minor change from Jacquet and Régnier (1988) with some methodological refinements borrowed from Jacquet and Szpankowski (1998).

Theorem 4 *Let C_n be the cost of Radix Sort (number of digit extractions) to sort n random keys. Then $(C_n - L_n)/\sqrt{V_n}$ tends in distribution and in moments to the standard normal variate $\mathcal{N}(0,1)$ with*

$$\begin{aligned} L_n &= \left(\ln n + \gamma + \frac{\ln B}{2} + P_1(\ln n)\right)\frac{n}{\ln B} + O(\ln n), \\ V_n &= n\big(C(B) + P_2(\ln n)\big) + O(\ln^2 n), \end{aligned}$$

where

$$C(B) = \frac{1}{\ln B}\Big(\frac{1}{4} + \ln 2 + 2\sum_{k=1}^{\infty}\big[\ln(1+B^{-k}) + (1+B^{-k})^{-2}\big]\Big),$$

and the functions $P_1(x)$, and $P_2(x)$ are periodic functions with small amplitude and period $\ln B$.

Remark: The average was found by Knuth and De Bruijn; the analysis for the case $B = 2$ is presented in Knuth (1973; pp. 131–134). Related variance analyses appear in Kirschenhofer, Prodinger, and Szpankowski (1989).

Proof of Theorem 4. The proof is divided into three parts: the mean value analysis, the variance analysis, and finally the limit distribution result.
Mean value analysis. Let $\mathbf{E}[C_n] = L_n$, and $L(z) = \sum_{n=0}^{\infty} L_n z^n e^{-z}/n!$ be the Poisson generating function of the sequence $\{L_n\}$. Taking derivatives of (14) with respect to u at $u = 1$ yields

$$L(z) = \frac{\partial}{\partial u}\big(\Xi(1,z)e^{-z}\big) = BL\Big(\frac{z}{B}\Big) + z - ze^{-z}, \tag{15}$$

with $L(0) = 0$. By straightforward iteration we get

$$L(z) = z\sum_{k=0}^{\infty}(1 - e^{-z/B^k}) \ .$$

The asymptotic equivalent of $L(z)$ is obtained via the Mellin transform $L^*(s) = \int_0^{\infty} L(x)x^{s-1}dx$ which has the closed form expression:

$$L^*(s) = -\frac{s\Gamma(s)}{1 - B^{s+1}},$$

defined for $\Re(s) \in (-2,-1)$ (For an extensive study of the Mellin transform see Flajolet, Gourdon, and Dumas (1995)). The singular expansion of $L^*(s)$ and a residue computation of the inverse Mellin transform $L(x) = \frac{1}{2i\pi}\int_{c-\infty}^{c+i\infty} L^*(s)x^{-s}\,ds$, for any $c \in (-2,-1)$, yield the asymptotic expansion,

$$L(x) = \Big[\ln(x) + \gamma + \frac{\ln B}{2} + P_1(\ln x)\Big]\frac{x}{\ln B} + O(x^{-M}),$$

for any arbitrary $M > 0$, and x varying in a cone included in the right half complex plane. The function $P_1(x)$ is identified by its Fourier transform:

$$P_1(x) = \sum_{k \in \mathbf{Z} \ \{0\}} \Gamma\Big(\frac{2ik\pi}{\ln B}\Big) e^{2ik\pi x/\ln B},$$

Now, we need to translate this asymptotic equivalent of $L(x)$ into an asymptotic result on L_n. To this end we make use of the depoissonization lemma introduced in Jacquet and Régnier (1988) and extensively developed in Jacquet and Szpankowski (1998).

In Jacquet and Szpankowski (1998) we find the following result. Assume the two following conditions:

(i) there exists a cone including the positive axis where the Poisson transform $L(z)$ has a polynomial growth of order $z^\beta h(z)$ where $h(z)$ is a slowly varying function.

(ii) outside the cone $|L(z)e^z|e^{-|z|}$ is a function that decays to zero faster than any polynomial.

Then $L_n = L(n) + O(n^{\beta-1}h(n))$. More generally there exists a double sequence b_{ij} such that

$$L_n = \sum_{i=0}^{i=k} \sum_{j=i}^{j=2k} b_{ij} n^i L^{(j)}(n) + O\big(n^{\beta-k} h(n)\big),$$

where $L^{(j)}(x)$ stands for the jth derivative of $L(x)$. The double sequence is identified by $\sum_{i=0}^\infty \sum_{j=i}^\infty b_{ij} x^i y^j = \exp(x\ln(1+y) - xy)$. In particular $L_n = L(n) - \frac{n}{2}L^{(2)}(n) + O(n^{\beta-2}h(n))$.

The condition (i) is satisfied for $\beta = 1$ and $L(x) = \ln x$. To check condition (ii) we take a cone $\mathcal{C}_\theta$ of angle $\pm\theta$ around the positive axis. We define the domain $\mathcal{D}_k$ as the domain outside the cone such that $|z| \in (B^k, B^{k+1})$. Notice that $z \in \mathcal{D}_k$ implies $z/B \in \mathcal{D}_{k-1}$. We define A_k to be the maximum value of $|L(z)e^z/z|e^{-|z|}$ on this domain, our objective is to prove that A_k tends to zero at a suitable rate. By using the functional equation we obtain the inequality:

$$\begin{aligned} A_{k+1} &\leq \max_{z \in \mathcal{D}_{k+1}} \Big\{ |B \frac{L(z/B)e^z}{z}|e^{-|z|} + |e^z - 1|e^{-|z|} \Big\} \\ &\leq A_k e^{(\cos\theta - 1)B^{k-1}} + e^{-B^{k-1}} + e^{(\cos\theta-1)B^{k-1}}, \end{aligned}$$

which proves that the sequence A_k converges to 0 at a rate at most $\exp((\cos\theta - 1)B^{k-1})$.

Variance analysis. We now turn to the variance and first introduce the Poisson generating function of the second factorial moment $S(z)$:

$$S(z) = \frac{\partial^2}{\partial u^2}\big(\Xi(1, z)e^{-z}\big).$$

After some algebra one finds

$$S(z) = BS\Big(\frac{z}{B}\Big) + 2BzL\Big(\frac{z}{B}\Big) + 2zL'\Big(\frac{z}{B}\Big) + (B^2 - B)L^2\Big(\frac{z}{B}\Big) + z^2. \tag{16}$$

Introduce the function $V(z)$ defined by $V(z) = S(z) + L(z) - L^2(z)$ which is the variance of the Poissonized cost. (It must be noticed that $V(z)$ is not the Poisson generating function of V_n.) The variance V_n is trivially obtained by the depoissonization of $V(z) + L^2(z)$ after subtracting L_n^2 and therefore depoissonization results prove to be of great help.

The function $V(z)$ satisfies the following functional equation:

$$V(z) = BV\Big(\frac{z}{B}\Big) + 2zL'\Big(\frac{z}{B}\Big) + 2ze^{-z}L(z) + z - ze^{-z} + z^2e^{-2z}.$$

We introduce the Mellin transform $V^*(s)$ of $V(x)$. It follows from (15) that the Mellin transform of $L'(x)$ is $(s-1)\Gamma(s)/(1-B^s)$. We get:

$$V^*(s) = \frac{1}{1-B^{s+1}}\left(\frac{2B^{s+1}s\Gamma(s+1)}{1-B^{s+1}} - \Gamma(s+1) + D(s)\right),$$

with

$$D(s) = \Gamma(s+2)\Big[2^{-2-s} + 2\sum_{k=0}^{\infty}(1-(1+B^{-k})^{-s-2}\Big].$$

The asymptotic expansion of $V(x)$ results from a residue analysis of the Mellin transform. The asymptotic expansion of V_n is delicate because of cancelations between periodic functions; see (Kirschenhofer, Prodinger, and Szpankowski (1989). Nevertheless, we can finesse the subtleties as was done in Jacquet and Régnier (1988): the relation

$$V_n = V(n) - n(L'(n))^2 + O\big(\ln^2 n\big),$$

holds as soon as $S(z) = V(z) + (L(z))^2$ satisfies condition (ii) of analytic depoissonization. This last check follows the same argument as for $L(z)$. First, it is clear that condition (ii) holds for $L'(z)$. Second, to prove condition (ii) for $S(z)$ we define A'_k to be the maximum value of $|S(z)e^z/z|e^{-|z|}$ in the domain $\mathcal{D}_k$. By using the functional equation (16) for $z \in \mathcal{D}_{k+1}$ we obtain the inequality:

$$\begin{aligned}
\Big|\frac{S(z)e^z}{z}\Big|e^{-|z|} \quad &\leq \quad A'_k e^{(\cos\theta-1)|z|} \\
&+ \quad 2\Big|\Big(BL\Big(\frac{z}{B}\Big) + L'\Big(\frac{z}{B}\Big)\Big)e^{z/B}\Big|e^{-|z|/B}\exp\Big\{(B-1)\Big(\cos\theta-1)\frac{|z|}{B}\Big) \\
&+ \quad (B^2-B)\Big(\Big|L\Big(\frac{z}{B}\Big)e^{\frac{z}{B}}\Big|e^{-|z|/B}\Big)^2\frac{1}{z}\exp\Big\{\Big(1-\frac{2}{B}\Big)(\cos\theta-1)|z|\Big\} \\
&+ \quad |z|e^{(\cos\theta-1)|z|}\Big\}.
\end{aligned}$$

This defines an inductive inequality between A'_{k+1} and A'_k, which implies that the sequence A'_k converges to 0 at a rate at most $\exp((\cos\theta-1)B^{k-1})$.

B	$C(B)$
2	$4.350088480\ldots$
3	$1.808375772\ldots$
4	$1.182662098\ldots$
5	$.9101381054\ldots$
6	$.7588387730\ldots$
7	$.6626599361\ldots$
8	$.5960035274\ldots$

Table 1: Coefficient of n in the asymptotic expansion of the variance of the cost of *Radix Sort*.

Differentiating (15) yields a functional equation for L' from which we get, after some algebra, the Mellin transform of $(L'(x))^2$:

$$\frac{2(s-1)\Gamma(s)}{(1-B^s)^2}+\frac{(2-2^{-s})\Gamma(s)}{1-B^s}+\frac{A(s)}{1-B^s},$$

where

$$\begin{aligned}A(s) \quad = \quad & \Gamma(s+1)(2^{-s}-2)-2^{-s-2}\Gamma(s+2)\\ & +2\Gamma(s+1)\sum_{k\geq 0}\left[1-(1+B^{-k})^{-s-1}\right]-2\Gamma(s)\sum_{k\geq 0}\left[1-(1+B^{-k})^{-s}\right]\\ & +2\Gamma(s+2)\sum_{k\geq 0}B^{-k}(1+B^{-k})^{-s-2}-2\Gamma(s+1)\sum_{k\geq 0}B^{-k}(1+B^{-k})^{-s-1}.\end{aligned}$$

Here $A(s)$ is analytic around $s=0$. Using the above we obtain the Mellin transform of $V(z)-z(L'(z))^2$:

$$\frac{\Gamma(s+1)[2^{-s-1}-1-2(s+1)]}{1-B^{s+1}}+\frac{D(s)-A(s+1)}{1-B^{s+1}}.$$

It follows that

$$V_n-n\big(L'(n)\big)^2\sim n\left(-\frac{2+\ln 2}{\ln B}+\frac{D(-1)-A(0)}{\ln B}+P_2(\ln n)\right).$$

Table 1 lists the average values $C(B)$ of V_n/n as a function of B, for the first few values of B.

Limit distribution. We now turn to the last part of the proof concerning the analysis of the limiting distribution. To this end we operate directly with the functional equation (14). We denote by $\kappa(u)$ the quantity $\ln B/(\ln B-\ln u)$. We note in passing that it satisfies:

$$B\Big(\frac{u}{B}\Big)^{\kappa(u)}=1.$$

We also note that $\kappa(1)=1$.

In preparation for depoissonization we again consider the cone $\mathcal{C}_\theta$. We will prove that there is a real neighborhood $U(1)$ of the variable u such that $\ln\Xi(u,z)$ exists and is $O(z^{\kappa(u)})$, when z tends to infinity inside the cone $\mathcal{C}_\theta$. We proceed in two steps.

First step: we prove that there exists $\alpha > 0$ such that for $z \in \mathcal{C}_\theta$: property $R(z)$ holds, with property $R(z)$ being $|\Xi(u,z)| \geq 2\exp(\alpha z^{\kappa(u)})$. We assume that $\nu < u/B < \rho$ for some $0 < \nu < \rho < 1$. We redefine the increasing domains $\mathcal{D}_k = \{z,\ z \in \mathcal{C}_\theta,\ A\nu \geq |z| \leq A\rho^{-k}\}$ for some $A > 0$. Since u is real we have the following property: $z \in \mathcal{D}_{k+1} - \mathcal{D}_k \Rightarrow uz/B \in \mathcal{D}_k$.

Let fix $\alpha < \cos\theta$. Let assume that u is in an interval such that $\kappa(u) > 1/2$ and $|u| < 2$. We fix A such that for all $x > A\nu$:

- $2e^{\alpha x} < e^{\cos\theta x}$;
- $3x < 2\exp(\alpha x^{1/2})$.

Since $|\Xi(1,z)| = |e^z| = \exp(\cos\theta|z|)$, by a compactness argument there exists a neighborhood $U(1)$ of 1 for u such that for all $(u,z) \in U(1) \times \mathcal{D}_0$: $|\Xi(u,z)| \geq 2\exp(\alpha|z|^{\kappa(u)}$. Therefore property $R(z)$ holds on $\mathcal{D}_0$.

By induction we prove that property $R(z)$ holds on all $\mathcal{D}_k$. Let us assume it holds on $\mathcal{D}_k$ and let us take $z \in \mathcal{D}_{k+1} - \mathcal{D}_k$. By applying equation (14) and using the fact that $uz/B \in \mathcal{D}_k$ we get

$$\begin{aligned} \big|\Xi(u,z)\big| &\geq \Big|\Xi^B\Big(u, \frac{zu}{B}\Big)\Big| - \big|(1-u)z\big| \\ &\geq 2^B\exp(\alpha|z|^{\kappa(u)}) - 3|z| \geq 2\exp(\alpha|z|^{\kappa(u)}), \end{aligned}$$

which proves by induction the property $R(z)$ over $\mathcal{D}_{k+1}$.

Remark. By similar arguments we can prove the counterpart of $R(z)$, $R'(z)$: $|\Xi(u,z)| < \frac{1}{2}\exp(\alpha'|z|^{\kappa(u)})$ for all $z \notin \mathcal{C}_\theta$ and for some $\alpha' < 1$, (notice that in this case $\alpha' > \cos\theta$).

Second step: we prove that for $(u,z) \in U(1) \times \mathcal{C}_\theta$, $\ln\Xi(u,z) = O(z^{\kappa(u)})$ when $z \to \infty$. We notice that the function $\ln\Xi(u,z)$ satisfies the functional equation

$$\ln\Xi(u,z) = B\ln\Xi\Big(u, \frac{uz}{B}\Big) - g(u,z), \tag{17}$$

with $g(z,u) = \ln(1-(1-u)z/\Xi(u,z))$ and according to the previous step $g(u,z) = O(\exp(-\alpha z^{\kappa(u)-\varepsilon}))$, with $\varepsilon > 0$ arbitrarily small, when $z \to \infty$.

Let A_k be the maximum value over $\mathcal{D}_k$ of function $|\ln\Xi(u,z)/z^{\kappa(u)}|$. To give an upper bound for A_{k+1} as a function of A_k, we take $z \in \mathcal{D}_{k+1} - \mathcal{D}_k$. Using equation (17) it follows that

$$\Big|\frac{\ln\Xi(u,z)}{z^{\kappa(u)}}\Big| \leq A_k + O\big(\exp(-\alpha z^{\kappa(u)-\varepsilon})\big),$$

and therefore $A_{k+1} \leq A_k + O(\exp(-(A\rho^{-k})^{\kappa(u)-\varepsilon}))$. Clearly $\limsup A_k < \infty$ and therefore the second step is proven.

We need an intermediate technical result in order to prove the convergence to normal distribution. We prove that, for all integers i and j,

$$\Xi_{u^iz^j}(u,z) := \frac{\partial^{i+j}}{\partial u^i \partial z^j}\ln\Xi(u,z) = O(z^{\kappa(u)+\varepsilon}).$$

Such intermediate results follow from two considerations: first we take the ith derivative of equation (14) in order to establish arbitrary polynomial growth of the factors $\Xi_{u^i z^j}(u,z)/\Xi(u,z)$. Second we take the jth derivative of the equation (17) and we use the polynomial order of $\Xi_{u^i z^j}$ in $\frac{\partial^{i+j}}{\partial u^i \partial z^j} g(u,z)$ in order to establish the expected result using similar arguments as with the A_k's.

Therefore we have the expression

$$\ln \Xi(e^t, z) = z + L(z)t + V(z)\frac{t^2}{2} + t^3 R(z,t).$$

Thanks to the intermediate technical results we have $R(z,t) = O(z^{\kappa(u)+\varepsilon})$ for t in a real neighborhood of 0. The normal limiting distribution of the Poisson transform $\Xi(u,z)e^{-z}$ comes from:

$$e^{-z}\Xi(e^{t/\sqrt{V(z)}}, z)\exp(-L(z)t/\sqrt{V(z)}\,) = \exp\Big(\frac{t^2}{2} + R(z,t)t^3 V^{-3/2}(z)\Big),$$

which tends to $e^{t^2/2}$.

In order to translate this result to $\phi_n(e^t)$ one needs to again refer to depoissonization. To this end we can use Theorems 8 and 9 of Jacquet and Szpankowski (1998) with $\beta = 1/2+\varepsilon$ which imply the convergence of $\phi_n(e^{t/\sqrt{V_n}})\exp(-L_n t/\sqrt{V_n})$ to the standard normal variate in distribution and in moments. We need before to check an equivalent of condition (ii), that is $|\Xi(e^{t/\sqrt{V_n}}, z)||e^{-n}|$ decreases faster than any polynomial in n when $z \notin \mathcal{C}_\theta$ and $|z| = n$. This property is easy to check since with property $R'(z)$:

$$\begin{aligned} \Big|\Xi(e^{t/\sqrt{V_n}}, z)\Big| &\le \frac{1}{2}\exp\Big(\alpha' n^{\kappa(e^{t/\sqrt{V_n}})}\Big) \\ &\le \frac{1}{2}\exp\Big((1 + O\Big(\frac{\ln n}{\sqrt{V_n}}\Big)\alpha' n\Big). \quad \square \end{aligned}$$

4. Discussion

We analyzed various flavors of bucket selection and sorting algorithms. A compound Poisson limit was found for *Distributive Selection*, and central limit theorems were found for the rest of selection and sorting methods. We considered distributive algorithms when the number of buckets is the same as the number of keys. Our analysis can be extended to other situations with a large number of buckets. For instance, only minor modification of the proofs of both *Distributive Selection* and *Distributive Sort* will give the analysis required for a distributive algorithm with $B = \lceil \beta n \rceil$, for a positive real constant β.

Small periodic fluctuations. typical in this type of analysis, appear in the results of recursive bucketing algorithms based on digital data. These fluctuations are of truly an ignorable magnitude, and should not affect the practioners' decision of whether to choose a bucketing algorithm or not.

The average-case result for *Binary Radix Sort has* long been known since Knuth (1973; pp. 131–134) published it for the first time. Our proof covers this old result.

The limiting distribution results are new in the context of radix sorting. Similar results hold for the external path length of tries (Jacquet and Régnier (1988)). In fact, the cost of *Radix Sort* is a special kind of path length in random B-ary tries on n keys—this cost is the total path length to all non-empty external nodes.

Acknowledgments. The authors are thankful to Marko Riedel for several helpful discussions. The first author wishes to thank INRIA for supporting his sabbatical leave. This work was supported in part by the Long Term Research Project *Alcom-IT* (# 20244) of the European Union.

References

1. Alison, D. and Noga, M. (1981). Selection by distributive partitioning. *Information Processing Letters*, **11**, 7–8.
2. Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W. and Tukey, J. (1972). *Robust Estimates of Location: Surveys and Advances.* Princeton University Press, Princeton, New Jersey.
3. Chung, K. (1974). *A Course in Probability Theory.*
4. Devroye, L. (1986). *Lecture Notes on Bucket Algorithms.* Birkhauser, Boston.
5. Devroye, L. and Klincsek, T. (1981). Average time behavior of distributive sorting algorithms. *Computing*, **26**, 1–7.
6. Dobosiewicz, W. (1978). Sorting by distributive partitioning. *Information Processing Letters*, **7**, 1–6.
7. Ehrlich, G. (1981). Searching and sorting real numbers. *J. Algorithms*, **2**, 1–14.
8. Fagin, R., Nievergelt, N., Pippenger, N. and Strong, H. (1979). Extendible Hashing—a fast access method for dynamic files. *ACM Transactions on Database Systems*, **4**, 315–344.
9. Flajolet, P. (1983). On the performance evaluation of extendible hashing. *Acta Informatica*, **20**, 345–369.
10. Flajolet, P., Gourdon, G., and Dumas, P. (1995) Mellin Transforms and Asymptotics: Harmonic Sums. *Theoretical Computer Science*, **144**, 3–58.
11. Flajolet, P., Poblete, P. and Viola, A. (1998). On the analysis of linear probing hashing. INRIA Technical Report 3265. To appear in *Algorithmica.*
12. Flajolet, P. and Sedgewick, R. (1995). Mellin transforms and asymptotics: Finite differences and Rice's integrals. *Theoretical Computer Science*, **144**, 101–124.
13. Gonnet, G. (1984). On direct addressing sort. *RAIRO: Technique et Science Informatiques*, **3**, 123–127.
14. Hoare, C. (1961). FIND (Algorithm 65). *Communications of ACM*, **4**, 321–322.
15. Jacquet, P. and Régnier, M. (1986). Trie partitioning process: Limiting Distributions. *Lecture Notes in Computer Science*, **214**, 196–210. Springer, New York.
16. Jacquet, P. and Régnier, M. (1988). Normal limiting distribution for the size and the external path length of tries, INRIA Research Report 827.
17. Jacquet, P. and Szpankowski W. (1991). Analysis of digital tries with markovian dependency. *IEEE Transactions on Information Theory*, **37**, 1470–1475.
18. Jacquet, P. and Szpankowski W. (1998). Analytical depoissonization and its applications, to be published in *Theoretical Computer Science*.

19. Kirschenhofer, P., Prodinger, H. and Szpankowski, W. (1989). On the variance of the external path length in a symmetric digital trie. *Discrete Applied Mathematics*, **25**, 129–143.

20. Knuth, D. (1973). *The Art of Computer Programming 3: Sorting and Searching.* Addison-Wesley, Reading, Massachusetts.

21. Mahmoud, H., Modarres, R. and Smythe, R. (1995). Analysis of Quickselect: An algorithm for order statistics. *RAIRO, Theoretical Informatics and Applications*, **29**, 255–276.

22. Pittel, B. (1986). Paths in a random digital tree: limiting distributions. *Advances in Applied Probability*, **18**, 139–155.

23. Régnier, M. and Jacquet, P. (1989). New results on the size of tries. *IEEE Transactions on Information Theory*, **35**, 203–205.

24. Tamminen, M. (1981). Analysis of N-trees. *RAIRO, Information Processing Letter*, **16**, 131–137.

25. Wong, R. (1989) *Asymptotic Approximations of Integrals.* Academic Press, Orlando, Florida.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105,
78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS
Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
(France)
http://www.inria.fr
ISSN 0249-6399