Progress in Computer Science
No. 6

Birkhäuser
Boston · Basel · Stuttgart

Luc Devroye

Lecture Notes on Bucket Algorithms

1986

Birkhäuser
Boston · Basel · Stuttgart

Author:

Luc Devroye
School of Computer Science
McGill University
Montreal H3A 2K6
Canada

Library of Congress Cataloging in Publication Data

Devroye, Luc.
Lecture notes on bucket algorithms.

(Progress in computer science ; no. 6)
Bibliography: p.
Includes index.
1. Data structures (Computer science) 2. Algorithms.
I. Title. II. Title: Bucket algorithms. III. Series.
QA76.9.D35D48 1985 005.7'3 85-26850
ISBN 0-8176-3328-6

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Devroye, Luc:
Lecture notes on bucket algorithms / Luc Devroye. –
Boston ; Basel ; Stuttgart : Birkhäuser, 1985.

(Progress in computer science ; No. 6)
ISBN 3-7643-3328-6 (Stuttgart . . .)
ISBN 0-8176-3328-6 (Boston)
NE: GT

Printed in Germany
ISBN 0-8176-3328-6
ISBN 3-7643-3328-6

TABLE OF CONTENTS

PREFACE

Hashing algorithms scramble data and create pseudo-uniform data distributions. Bucket algorithms operate on raw untransformed data which are partitioned into groups according to membership in equi-sized d-dimensional hyperrectangles, called cells or buckets. The bucket data structure is rather sensitive to the distribution of the data. In these lecture notes, we attempt to explain the connection between the expected time of various bucket algorithms and the distribution of the data. The results are illustrated on standard searching, sorting and selection problems, as well as on a variety of problems in computational geometry and operations research.

The notes grew partially from a graduate course on probability theory in computer science. I wish to thank Elizabeth Van Gulick for her help with the manuscript, and David Avis, Hanna Ayukawa, Vasek Chvatal, Beatrice Devroye, Hossam El Gindy, Duncan McCallum, Magda McCallum, Godfried Toussaint and Sue Whitesides for making the School of Computer Science at McGill University such an enjoyable place. The work was supported by NSERC Grant A3456 and by FCAC Grant EQ-1679.

INTRODUCTION

It is not a secret that methods based upon the truncation of data have good expected time performance. For example, for nice distributions of the data, searching is often better done via a hashing data structure instead of via a search tree. The speed one observes in practice is due to the fact that the truncation operation is a constant time operation.

Hashing data structures have not received a lot of attention in the 1970's because they cannot be fit into the comparison-based computational model. For example, there is no generally accepted lower bound theory for algorithms that can truncate real numbers in constant time. The few analyses that are available (see Knuth (1973), Gonnet (1981,1984) and the references found there) relate to the following model: the data points are uniformly distributed over either [0,1] or $\{1,...,M\}$. The uniform model is of course motivated by the fact that it is often possible to find a good hash function $h(.)$, i.e. a function of the data points which distributes the data evenly over its range. In the vast majority of the cases, $h(.)$ is not a monotone function of its argument when the argument is an integer or a real number. Non monotone functions have the undesirable side-effect that the data are not sorted. Although this is not important for searching, it is when the data need to be listed in sorted order rather frequently. If the data form a **data base**, i.e. each data point can be considered as a point in R^d with $d > 1$, then range queries can be conveniently handled if the data are hashed via monotone functions. There is an ever increasing number of applications in **computational geometry** (see the general survey articles by Toussaint (1980,1982) where applications in pattern recognition are highlighted ; and the survey article on bucket methods by Asano, Edahiro, Imai, Iri and Murota (1985)) and **computer graphics**, in which the data points should preserve their relative positions because of the numerous geometrical operations that have to be carried out on them. Points that are near one another should stay near. In **geographic data processing**, the cellular organization is particularly helpful in storing large amounts of data such as satellite data (see the survey article by Nagy and Wagle, 1979). Many tests in **statistics** are based upon the partition of the space in equal intervals, and the counts of the numbers of points in these intervals. Among these, we cite the popular chi-square test, and the empty cell test. See for example Kolchin, Sevast'yanov and Chistyakov (1978) and Johnson and Kotz (1977) for applications in statistics. In **economic surveys** and **management science**, the histogram is a favorite tool for visualizing complex data. The histogram is also a superb tool for statisticians in exploratory data analysis. In all these examples, the order in the data must be preserved.

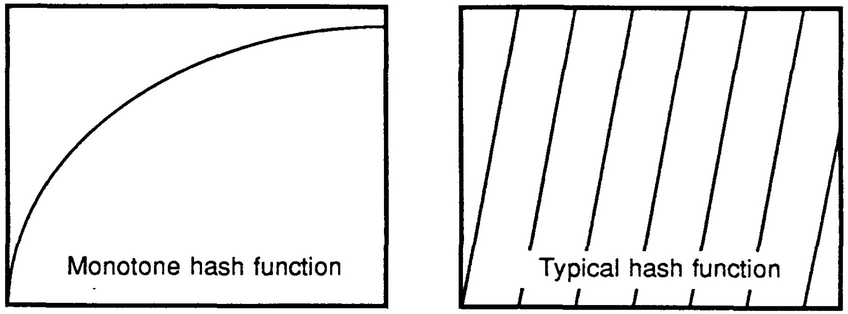


Figure 0.1.

If we use monotone or order-preserving hash functions, or no hash functions at all, the uniform distribution model becomes suspect. At best, we should assume that the data points are random vectors (or random variables) that are independent and identically distributed. The randomness is important because we are not interested here in worst-case performance. Expected time can only be analyzed if some randomness is assumed on the part of the data. The independence assumption can be defended in some situations, e.g. in the context of data bases for populations. Unfortunately in some geometrical applications, particularly image processing, the independence assumption is just not good enough. Notice that if pixels in a screen were selected independently and according to a given distribution, then the composed picture would be a "pure noise" picture. In a sense, the more information we have in a picture, the more dependence we see between the pixels. Finally, if we accept the independence assumption, we might as well accept the identical distribution assumption, except if there is some nonstationary (time-dependent) element in the data collection process.

We will only deal with d-dimensional real numbers and with distributions that have densities. The complexities of various algorithms are measured in terms of fundamental operations. Typically, truncation or hashing is one such operation. We will of course assume that real numbers can be truncated and / or hashed in time independent of the size or the precision of the number - recall that a similar assumption about comparing two real numbers is needed in the well-known comparison-based complexity theory. Densities are convenient because they free us from having to consider discretization problems: if a distribution is atomic (i.e., it puts its mass on a countable set), and enough data points are drawn from this distribution, the number of colliding values increases steadily. In fact, if n independent identically distributed random vectors are considered with any atomic distribution, then $N/n \to 0$ almost surely as $n \to \infty$ where N is the number of different values. Meaningful asymptotics are only possible if either the atomic distribution varies with n, or the distribution is non-atomic. There is another key argument in favor of the use of densities: they provide a compact description of the distribution, and are easily visualized or plotted.

When independent random vectors with a common density are partitioned by means of a d-dimensional grid, the number of grid locations (or buckets) with at least 2 points has a distribution which depends upon the density in question. The density affects the frequency of collisions of data points in buckets. For example, if the density is very peaked, the buckets near the peak are more likely to contain a large number of points. We want to investigate how this crowding affects the performance of algorithms of bucket or grid algorithms.

Throughout this set of notes, we will consider a d-dimensional array of equi-sized rectangles (which we will call a grid), and within each rectangle, points are kept in a chain (or linked list). The number of rectangles will be denoted by m, and the data size by n. We will not consider infinite grids such as $\{[i,i+1) \mid i \text{ integer}\}$ because infinite arrays cannot be stored. However, because data may grow not only in size but also in value as $n \to \infty$, we will consider at times grid sizes m that are data value dependent. In any case, m is usually a function of n.

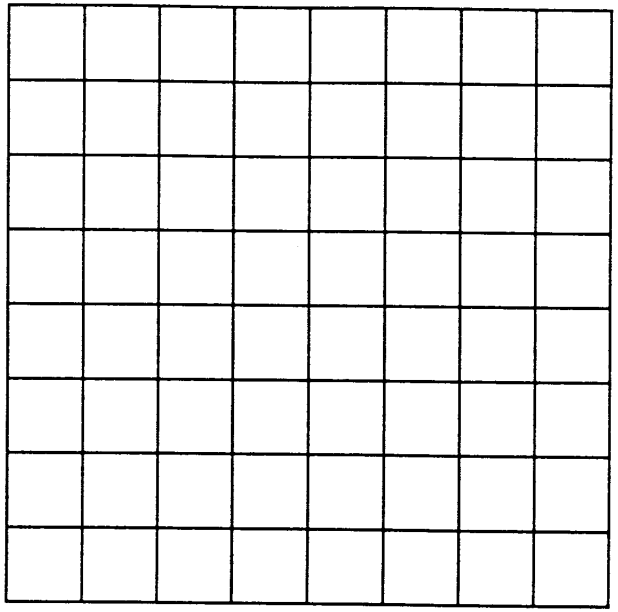


Figure 0.2.
2d Grid

The purpose of this collection of notes is to give a variety of probability theoretical techniques for analyzing various random variables related to the bucket structure described above. Such random variables include for example, the average search time, the time needed for sorting, the worst-case search time and other nonlinear functions of the cardinalities $N_1, \ldots, N_m$ of the buckets. The probability theoretical techniques have several features: they are general (for example, the Lebesgue density theorem is needed in crucial places in order to avoid having to impose any smoothness conditions on the densities), and whenever possible, appropriate probability inequalities are invoked (for example, heavy use is made of Jensen's inequality (see e.g. Chow and Teicher (1978)) and Chernoff's exponential bounding technique (Chernoff (1952))). Since $N_1, \ldots, N_m$ is multinomially distributed for a data-independent grid, and the N_i's are thus not independent, it is sometimes useful to use an embedding method that relates the multinomial vector to a vector of independent Poisson random variables. This method is commonly called Poissonization. Even in our Poissonization, we choose to rely on inequalities because only inequalities will

help us in the assessment of the expected time performance for particular values of n.

The point is that we do not wish to give an exhaustive description of known results in the field, or to present a list of exotic applications. We start very slowly on standard problems such as one-dimensional sorting and searching, and will move on to multidimensional applications towards the end of the notes. These applications are in the areas of computational geometry, operations research (e.g. the traveling salesman problem) and pattern recognition (e.g. the all-nearest neighbor problem).

In chapter 1, we have the simplest of all possible settings: the random variables $X_1, \ldots, X_n$ have a common density f on [0,1], and [0,1] is divided into m equal intervals $A_i = \left[\frac{i-1}{m}, \frac{i}{m}\right)$, $1 \le i \le m$. We are concerned with the simplest possible measures of performance in searching and sorting such as the average successful search time (called D_S) and the number of element comparisons for sorting (called C). If $m = n$, and f is uniform on [0,1], then each interval receives on the average one data point. It is well-known that $E(D_S) = O(1)$ and $E(C) = O(n)$ in that case. It is also known that the density f affects the distribution of quantities such as D_S and C. We will see that $E(D_S) \sim 1 + \frac{1}{2}\int f^2$ and $E(C) \sim \frac{n}{2}\int f^2$ as $n \to \infty$. The factor $\int f^2$, which is a measure of the peakedness of the density f, affects the performance in a dramatic way. For example, when $\int f^2 = \infty$, we have $E(C)/n \to \infty$ and $E(D_S) \to \infty$ as $n \to \infty$. In other words, bucket sorting takes linear expected time if and only if $\int f^2 < \infty$.

While most users will be quite satisfied with information about $E(C)$, some may doubt whether the expected value is a good measure of the state of affairs. After all, $E(C)$ is an estimate of the time taken per sort if averaged over a large number of sorts. The actual value of C for one individual sort could be far away from its mean. Fortunately, this is not the case. We will see that $C/n \to \frac{1}{2}\int f^2$ in probability as $n \to \infty$: thus, if $\int f^2 < \infty$, $C/E(C) \to 1$ in probability. For large n, even if we time only one sort, it is unlikely that $C/E(C)$ is far away from 1. Of course, similar results are valid for D_S and the other quantities.

We can take our analysis a bit further and ask what the variation is on random variables such as C. In other words, how small is $C - E(C)$ or $D_S - E(D_S)$? This too is done in chapter 1. The answer for C is the following:

$$Var(C) \sim n\left[\int f^3 - \left(\int f^2\right)^2 + \frac{1}{2}\int f^2\right].$$

In other words, $C - E(C)$ is of the order of $\sqrt{n}$ whereas $E(C)$ itself is of the order of n. Variances are used by statisticians to obtain an upper bound for

$$P(C - E(C) \geq \epsilon)$$

via the Chebyshev-Cantelli inequality:

$$P(C - E(C) \geq \epsilon) \leq \frac{Var(C)}{\epsilon^2 + Var(C)} .$$

Sometimes, this inequality is very loose. When ϵ is large compared to $\sqrt{n}$, there are much better (exponential) inequalities which provide us with a lot of confidence and security. After all, if C is extremely unlikely to be much larger than $E(C)$, then the usual worst-case analysis becomes almost meaningless.

We close chapter 1 with an attempt at reducing the dependence upon f. The idea is to apply the bucket method again within each bucket. This will be called double bucketing. The rather surprising result is that double bucketing works. For example, when $\int f^2 < \infty$, we have

$$\frac{E(C)}{n} \sim \frac{1}{2} \int_0^1 e^{-f} \leq \frac{1}{2} .$$

The detailed analysis of chapter 1 is well worth the effort. The development given there can be mimicked in more complicated contexts. It would of course be unwise to do so in these notes. Rather, from chapter 2 on, we will look at various problems, and focus our attention on expected values only. From chapter 2 onwards, the chapters are independent of each other, so that interested readers can immediately skip to the subject of their choice.

In chapter 2, the data $X_1, \ldots, X_n$ determine the buckets: the interval $[\min X_i , \max X_i]$ is partitioned into n equal intervals. This introduces additional dependence between the bucket cardinalities. The new factor working against us is the size of the tail of the distribution. Infinite tails force $\min X_i$ and $\max X_i$ to diverge, and if the rate of divergence is uncontrolled, we could actually have a situation in which the sizes of the intervals increase with n in some probabilistic sense. The study of $E(D_S)$, $E(C)$ and other quantities requires auxiliary results from the theory of order statistics. Under some conditions on f, including $\int f^2 < \infty$, we will for example see that

$$\frac{E(C)}{n} \sim E(\max_{1 \leq i \leq n} X_i - \min_{1 \leq i \leq n} X_i) . \int f^2 ,$$

i.e. the asymptotic coefficient of n is the expected range of the data (this measures the heaviness of the tail of f) times $\int f^2$, the measure of peakedness. Unless f vanishes outside a compact set, it is impossible to have $E(C) = O(n)$.

In chapter 3, we look at multidimensional problems in general. The applications are so different that a good treatment is only possible if we analyze

$$\sum_{i=1}^{m} g(N_i)$$

where $g(.)$ is a "work function", typically a convex positive function. The main result of the chapter is that for $m = n$, the expected value of this sum is $O(n)$ if and only if f has compact support, and

$$\int g(f) < \infty$$

provided that $g(.)$ is a "nice" function. Some applications in computational geometry and operations research are treated in separate sections of the chapter.

In some problems, we need to have assurances that the expected worst-case is not bad. For example, in the simple one-dimensional bucket data structure, the worst-case search time for a given element is equal to the maximal cardinality. Thus, we need to know how large

$$\max(N_i)$$

is. This quantity is analyzed in chapter 4. If f is bounded on a compact set of R^d, and $m = n$ then its expected value is asymptotic to $\frac{\log n}{\log \log n}$. If f is not bounded, then its expected value could increase faster with n. This result is for example applied to Shamos' two dimensional convex hull algorithm.

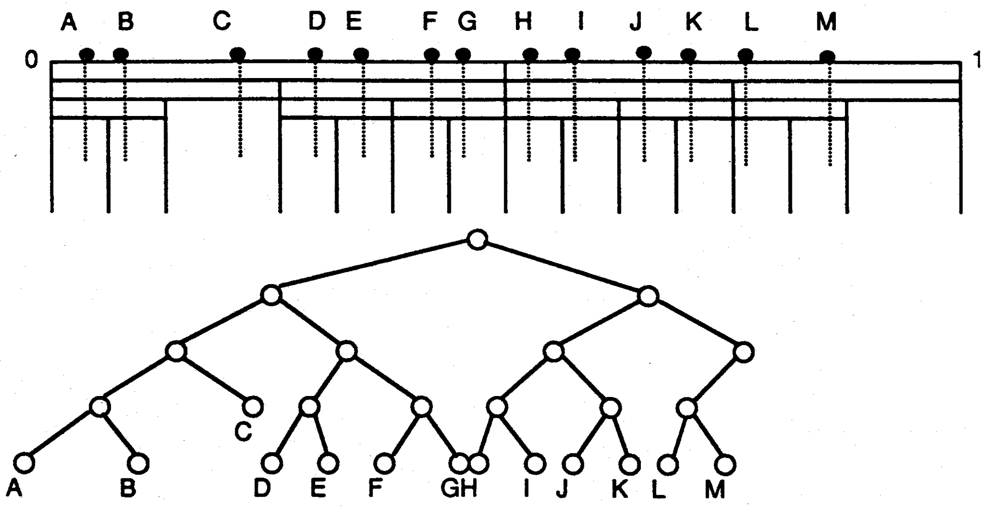


Figure 0.3.
Binary trie for points distributed on [0,1].

It is sometimes important to have bucket structures which are allowed to grow and shrink dynamically, i.e. structures that can handle the operations insert and delete efficiently. The essential ingredient in such a structure is an auxiliary array of bucket cardinalities. One can choose to split individual buckets once a certain threshold value is reached. This leads to a tree structure. If a bucket can hold at most one element, then one obtains in fact a binary trie (Knuth, 1973). Another strategy consists of splitting all buckets in two equi-sized buckets simultaneously as soon as the global cardinality reaches a certain level. In this manner, the number of buckets is guaranteed to be a power of two, and by manipulating the threshold, one can assure that the ratio of points to buckets is a number between 1 and 2 for example. This has the additional advantage that individual bucket counts are not necessary. Also, no pointers for a tree structure are needed, since data points are kept in linked lists within buckets. This dyadic dynamic structure is at the basis of the extendible hash structure described and analyzed in Fagin, Nievergelt, Pippenger and Strong (1979), Tamminen (1981) and Flajolet (1983). Tamminen (1985) compares extendible hashing with ordinary bucketing and various types of tries. See Tamminen (1985) and Samet (1984) for multidimensional tries. To keep these notes simple, we will not analyze any tree structures, nor will we specifically deal with dynamic bucket structures.

A last remark about the grid size m. Usually, we will choose m such that $m = m(n) \sim cn$ for some constant $c > 0$. (The ratio m/n will be called c_n,

so that $c_n \to c$ as $n \to \infty$.) We do so because we are mainly interested in searching and sorting. Roughly speaking, we can expect to sort the data in time $O(n)$ and to search for an element in time $O(1)$. If $m = o(n)$, the average number of points per interval grows unbounded, and we cannot hope to sort the data in time $O(n)$. On the other hand, if $m/n \to \infty$, the overhead due to housekeeping (e.g., traveling from bucket to bucket), which is proportional to m, and the storage requirements are both superlinear in n. Thus, there are few situations that warrant a sublinear or superlinear choice for m.

While we do generally speaking have some control over m, the grid size, we do not have the power to determine d, the dimension. Raw bucket algorithms perform particularly poorly for large values of d. For example, if each axis is cut into two intervals, then the grid size is 2^d. There are problems in which 2^d is much larger than n, the sample size. Thus, storage limitations will keep us from creating fine mazes in large dimensions. On the other hand, if rough grids are employed, the distribution of points is probably more uneven, and the expected time performance deteriorates.

Chapter 1

ANALYSIS OF BUCKET SORTING AND SEARCHING

1.1. EXPECTED VALUES.

In this chapter, f is a density on [0,1], which is divided into m intervals

$$A_i = \left[\frac{i-1}{m}, \frac{i}{m} \right), 1 \le i \le m \ .$$

The quantities of interest to us here are those that matter in sorting and searching. If sorting is done by performing a selection sort within each bucket and concatenating the buckets, then the total number of element comparisons is

$$C = \sum_{i=1}^{m} \frac{N_i(N_i-1)}{2} = \frac{1}{2}(T-n)$$

where, by definition,

$$T = \sum_{i=1}^{m} N_i^2 \ .$$

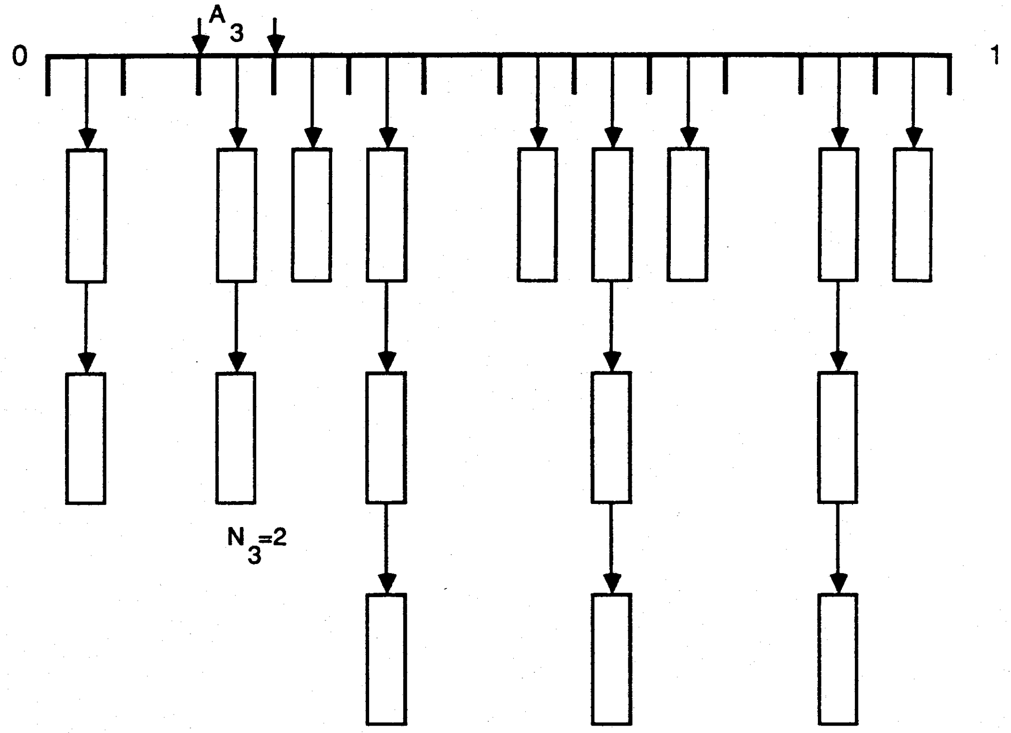


Figure 1.1.
Bucket structure with n=17 points, m=12 buckets.

The other work takes time proportional to m, and is not random. Selection sort was only chosen here for its simplicity. It is clear that for quadratic comparison-based sorting methods, we will eventually have to study T.

To search for an element present in the data, assuming that all elements are equally likely; to be queried, takes on the average

$$D_S = \frac{1}{n}\left[\sum_{i=1}^{m}\frac{1}{2}N_i(N_i+1)\right] = \frac{1}{n}\left[\frac{T}{2}+\frac{n}{2}\right]=\frac{T}{2n}+\frac{1}{2}$$

comparisons. This will be referred to as the ASST (Average Successful Search Time). Note that D_S is a function of the N_i's and is thus a random variable.

To search for an element not present in the data (i.e., an unsuccessful search), we assume that the element queried is X_{n+1}, independent of the data and distributed as X_1. The expected number of comparisons conditional on the data is

$$D_U = \sum_{i=1}^{m} N_i \int_{A_i} f = \sum_{i=1}^{m} N_i p_i$$

where only comparisons with non-empty cells in the data structure are counted. D_U will be called the AUST (Average Unsuccessful Search Time), and p_i is the integral of f over A_i.

The properties of this simple bucket structure for sorting and searching have been studied by Maclaren (1966), Dobosiewicz (1978) and Akl and Meijer (1982). In this chapter, we will unravel the dependence upon f. To get a rough idea of the dependence, we will start with the expected values of the quantities defined above.

Theorem 1.1.

Let f be an arbitrary density on [0,1]. Then, even if $\int f^2 = \infty$,

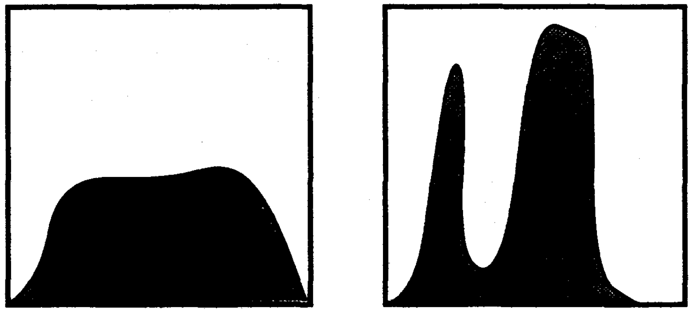
$$E(T)/n \sim 1 + \frac{1}{c}\int f^2 ;$$

$$E(C)/n \sim \frac{1}{2c}\int f^2 ;$$

$$E(D_S) \sim 1 + \frac{1}{2c}\int f^2 ;$$

$$E(D_U) \sim \frac{1}{c}\int f^2 .$$

Furthermore, $E(T) = o(n^2)$, $E(C) = o(n^2)$, $E(D_U) = o(n)$ and $E(D_S) = o(n)$.



Density with low value for square integral

Density with high value for square integral

Figure 1.2.

Theorem 1.1 sets the stage for this paper. We see for example that $E(T) = O(n)$ if and only if $\int f^2 < \infty$. Thus, for hashing with chaining, $\int f^2$ measures to some extent the influence of f on the data structure: it is an indicator of the peakedness of f. In the best case ($\int f^2 < \infty$), we have linear expected time behavior for sorting, and constant expected time behavior for searching. This fact was first pointed out in Devroye and Klincsek (1981). Under stricter conditions on f (f bounded, etc.), the given expected time behavior was established in a series of papers; see e.g. Dobosiewicz (1977), Weide (1978), Meijer and Akl (1980) and Akl and Meijer (1982). Theorem 1.1 gives a characterization of the densities with $\int f^2 = \infty$ in terms of quantities that are important in computer science. It also provides us with the form of the "best" density. Because $\int f^2 \geq (\int f)^2 = 1$ (Jensen's inequality), and $\int f^2 = 1$ for the uniform density on [0,1], we see that all the expected values in Theorem 1.1 are minimal for the uniform density.

Theorem 1.1 does not give the rate of increase of $E(T)$ as a function of n when $\int f^2 = \infty$. However, even though $T = \sum_{i=1}^{m} N_i^2$ can reach its maximal value n^2 (just set $N_1 = n$, $N_2 = \cdots = N_m = 0$), we have $E(T) = o(n^2)$ for **all** densities f. Thus, hashing with chaining when used for even the most peaked density, must dramatically improve the expected time for sorting and searching when n is large.

Proof of Theorem 1.1.

The proof is based upon a fundamental Lemma that will be useful in several places:

Lemma 1.1.

(i) $\max_i p_i = o(1)$ as $m \to \infty$.

(ii) For all $r \geq 1$, $n^r \sum_{i=1}^{m} p_i^r \leq \left(\frac{n}{m}\right)^r m \int f^r$.

(iii) For all $r \geq 1$,

$$\sum_{i=1}^{m} (np_i)^r \sim \left(\frac{1}{c}\right)^{r-1} n \int f^r \text{, and } \sum_{i=1}^{m} p_i^r = o\left(\sum_{i=1}^{m} p_i^{r-1}\right).$$

Proof of Lemma 1.1.

(i) follows from the absolute continuity of f, i.e. for each $\epsilon > 0$ we can find a $\delta > 0$ such that for all sets A with $\int_A dx < \delta$, we have $\int_A f < \epsilon$.

(ii) follows from Jensen's inequality:

$$\sum_{i=1}^{m} (np_i)^r = \sum_{i=1}^{m} \left(\frac{n}{m}\right)^r \left(m \int_{A_i} f\right)^r \leq \left(\frac{n}{m}\right)^r \sum_{i=1}^{m} m \int_{A_i} f^r = \left(\frac{n}{m}\right)^r m \int f^r .$$

(iii) follows from (ii) and a small additional argument: the upper bound in (ii) $\sim (\frac{1}{c})^{r-1} n \int f^r$. Furthermore, by Fatou's Lemma and the Lebesgue density theorem (see Lemma 5.10 for one version of this theorem), we have

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{i=1}^{m} (np_i)^r = \liminf_{n \to \infty} \frac{1}{n} \left(\frac{n}{m}\right)^r \sum_{i=1}^{m} \left(m \int_{A_i} f\right)^r$$

$$= \liminf_{n \to \infty} \frac{1}{n} \left(\frac{n}{m}\right)^r m \int f_n^r \quad \text{(where } f_n(x) = mp_i \text{ for } x \in A_i\text{)}$$

$$\geq \liminf_{n \to \infty} \left(\frac{n}{m}\right)^{r-1} \int \liminf_{n \to \infty} f_n^r$$

$$= \left(\frac{1}{c}\right)^{r-1} \int f^r \quad \text{(because } f_n \to f \text{ for almost all } x\text{).}$$

Note that f_n is the histogram approximation of f.

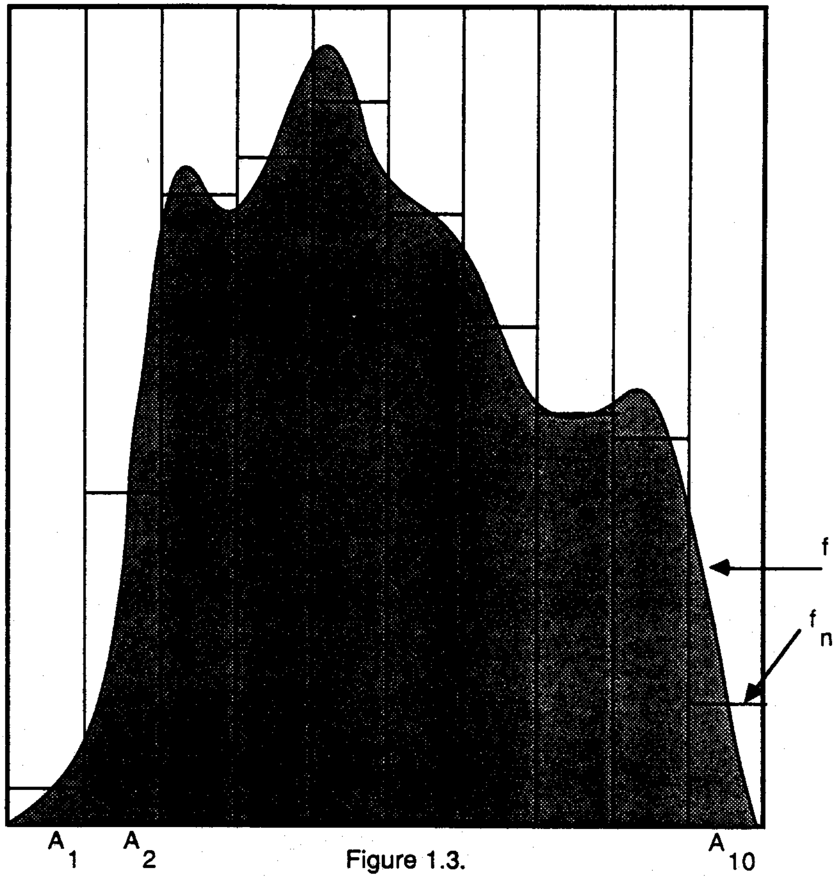


Figure 1.3.
Density f and its histogram approximation

The second half of (iii) follows from (i) and the inequality $\sum_{i=1}^{m} p_i^{\,r} \le \max p_i \cdot \sum_{i=1}^{m} p_i^{\,r-1}$.

The proof of Theorem 1.1 is simple. Observe that each N_i is a binomial (n, p_i) random variable and thus $E(N_i^{\,2}) = (np_i)^2 + np_i(1-p_i)$. Thus,

$$E(T) = E\left(\sum_{i=1}^{m} N_i^{\,2}\right) = \sum_{i=1}^{m} (n^2 p_i^{\,2} + np_i(1-p_i)) = (n^2 - n)\sum_{i=1}^{m} p_i^{\,2} + n$$

$$\sim \sum_{i=1}^{m} (np_i)^2 + n \sim \frac{n}{c}\int f^{\,2} + n$$

by Lemma 1.1 (iii). Also, by Lemma 1.1 (iii), $\sum_{i=1}^{m} p_i^{\,2} = o(1)$, so that $E(T) = o(n^2)$. All the other statements in the Theorem follow from the relations:

$$C = \frac{1}{2}\sum_{i=1}^{m} (N_i^{\,2} - N_i) = \frac{1}{2}(T - n)\,.$$

$$D_S = \frac{1}{n}\sum_{i=1}^{m} \frac{1}{2}(N_i^{\,2} + N_i) = \frac{1}{2} + \frac{T}{2\,n}\,,$$

and

$$D_U = \sum_{i=1}^{m} p_i N_i \;\; (E(D_U) = n\sum_{i=1}^{m} p_i^{\,2})\,.$$

1.2. WEAK CONVERGENCE.

In the previous section, we obtained an asymptotic expression for $E(T)$. One should not exaggerate the importance of such a quantity unless it is known that $T - E(T)$ is usually "small". For example, if we could show that $T/E(T) \to 1$ in probability, then we would be satisfied with our criterion $E(T)$. In addition, since $T/E(T)$ is closed to 1 for large n, the value of T obtained in one particular case (i.e., run; simulation) is probably representative of nearly all the values that will be obtained in the future for the same n. The main result here is

Theorem 1.2.

Let $\int f^2 < \infty$. Then :

$$T/n \to 1+\frac{1}{c}\int f^2 \text{ in probability ;}$$

$$C/n \to \frac{1}{2c} \int f^2 \text{ in probability ;}$$

$$D_S \to 1 + \frac{1}{2c} \int f^2 \text{ in probability ;}$$

and

$$D_U \to \frac{1}{c}\int f^2 \text{ in probability .}$$

The proof of the Theorem uses Poissonization to handle the fact that $N_1, \ldots, N_m$ are dependent random variables. For some properties of the Poisson distribution used here, we refer to section 5.1. We proceed now by extracting a key Lemma:

Lemma 1.2.

Let $\int f^2 < \infty$. Let N_i be Poisson (np_i) random variables $1 \leq i \leq m$. Then

$$\lim_{K\to\infty} \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{m} E(Y_i) = 0$$

where Y_i is either

(i) $E(N_i^2 I_{N_i^2 \geq K})$; or

(ii) $E(N_i^2)P(N_i^2 \geq K)$; or

(iii) $E(N_i^2)\, I_{np_i \geq K}$,

and I is the indicator function.

Proof of Lemma 1.2.

It is useful to recall a simple association inequality: if ϕ,ψ are nondecreasing nonnegative functions of their arguments, and X is an arbitrary real-valued random variable, then $E(\phi(X)\psi(X)) \geq E(\phi(X))E(\psi(X))$ (see e.g. Lehmann (1966), Esary, Proschan and Walkup (1967), and Gurland (1968)). For example, applied here,

$$E(N_i^2)P(N_i^2 \geq K) \leq E(N_i^2 I_{N_i^2 \geq K}).$$

Thus, we need not consider (ii). We will deal with (iii) first.

$$\frac{1}{n} \sum_{i=1}^{m} E(N_i^2)\, I_{np_i \geq K} = \frac{1}{n} \sum_{i=1}^{m} (n^2 p_i^2 + np_i)\, I_{np_i \geq K}$$

$$= \frac{1}{n} \sum_{i=1}^{m} ((\frac{n}{m})^2 (m\int_{A_i} f\,)^2 + n\int_{A_i} f\,)\, I_{np_i \geq K}$$

(where f_n is the function of section 1.1)

$$\leq \frac{1}{n} \sum_{i=1}^{m} (\frac{n^2}{m})\int_{A_i} f^2 + (n\int_{A_i} f\,)\, I_{p_i \geq K/n} \text{ (Jensen's inequality)}$$

$$= \int_0^1 (\frac{n}{m} f^2 + f\,) I_{f_n \geq Km/n} .$$

Now, $n/m \to 1/c$. Also, $I_{f_n \geq Km/n} \leq I_{f \geq Kc/2}$ for almost all x for which $f(x) > 0$, and all n large enough (this uses the fact that $f_n \to f$ for almost all x; see section 5.3.) Since $\int f^2 < \infty$, we thus have by the Lebesgue dominated convergence theorem,

$$\limsup_{n\to\infty} \int_0^1 (\frac{n}{m} f^2 + f) I_{f_n \geq Km/n} \leq \int_0^1 (\frac{1}{c} f^2 + f) I_{f \geq Kc/2}$$

and this can be made arbitrarily small by choosing K large enough.

Consider now (i). Let $L > 0$ be an arbitrary constant, depending upon K only.

$$\frac{1}{n} \sum_{i=1}^{m} E(N_i^2 I_{N_i^2 \geq K}) \leq \frac{1}{n} \sum_{i=1}^{m} E(N_i^2 I_{N^2 p_i^2 + np_i \geq L})$$

$$+ \frac{1}{n} \sum_{i=1}^{m} E(N_i^2 I_{N_i^2 \geq K, n^2 p_i^2 + np_i < L}) .$$

A simple application of (iii) shows that the first term on the right-hand-side has a limit supermum that is $o(1)$ as $L \to \infty$. Thus, we should choose L in such a way that $L \to \infty$ as $K \to \infty$. The second term on the right-hand-side is

$$\frac{1}{n} \sum_{\substack{i=1 \\ n^2 p_i^2 + np_i < L}}^{m} \sum_{j \geq \sqrt{K}} j^2 (np_i)^j e^{np_i} / j!$$

$$\leq \frac{1}{n} (\sum_{n^2 p_i^2 < L,\, i \leq m} 1). \left(\sum_{j \geq \sqrt{K}} j^2 (\sqrt{L})^j e^{-\sqrt{L}} / j! \right|$$

$\leq (c + o(1)) E(Y^2 I_{Y \geq \sqrt{K}})$ (where Y is Poisson ($\sqrt{L}$) distributed)

$\leq (c + o(1)) E(Y^3/\sqrt{K})$ (by Chebyshev's inequality)

$$= (c + o(1)) (L^{3/2} + 3L + \sqrt{L}) / \sqrt{K} .$$

This tends to 0 as $K \to \infty$ when we choose $L = K^{1/4}$. The proof of Lemma 1.2 is complete.

Proof of Theorem 1.2.

The results for C and D_S follow from the result for T. One possible Poissonization argument goes as follows: let $n' = n - n^{3/4}$, $n'' = n + n^{3/4}$. Let N', N'' be Poisson (n') and Poisson (n'') respectively. Let N_i' be a number of X_j' s, $1 \leq j \leq N''$ belonging to A_i. It is clear that $N_1', \ldots, N_m'$ are independent Poisson random variables with parameters $n'' p_i$, $1 \leq i \leq m$. Finally, let $T' = \sum_{i=1}^{m} N_i'^2$, $T'' = \sum_{i=1}^{m} N_i''^2$. For arbitrary $\epsilon > 0$ we have

$$\left[\frac{T}{n} > (1+\epsilon)(1+\frac{1}{c}\int f^2) \right] \subseteq [N'' < n] \cup \left[\frac{T''}{n} > (1+\epsilon)(1+\frac{1}{c}\int f^2) \right].$$

Using Theorem 5.5, we have $P(N'' < n) = P(\frac{N - n''}{n''} < -n^{3/4}/n'')$ $\leq 2\exp(-n^{3/2}/(2n''(1+\frac{n^{3/4}}{n''})))$. Thus, for all n large enough,

$$P(\frac{T}{n} > (1+\epsilon)(1+\frac{1}{c}\int f^2)) \leq o(1) + P(\frac{T''}{n''} > (1+\frac{\epsilon}{2})(1+\frac{1}{c}\int f^2)).$$

Similarly,

$$P(\frac{T}{n} < (1-\epsilon)(1+\frac{1}{c}\int f^2)) \leq P(N' > n) + P(\frac{T'}{n} < (1-\epsilon)(1+\frac{1}{c}\int f^2))$$

$$\leq 2\exp(-n^{3/2}/(2+o(1))n') \leq P(\frac{T'}{n'} < (1-\frac{\epsilon}{2})(1+\frac{1}{c}\int f^2)),$$

all n large enough. Now, all the probabilities involving T' and T'' are $o(1)$ if both T'/n' and T''/n'' tend to $1+\frac{1}{c}\int f^2$ in probability. Thus, the statements about T, C and D_S are valid if we can show the statement about T where $T = \sum_{i=1}^{m} N_i^2$ and $N_1, \ldots, N_m$ are independent Poisson random variables with parameters np_i, $1 \leq i \leq m$.

First, we note that by Lemma 1.1,

$$E(T) = \sum_{i=1}^{m} (n^2 p_i^2 + np_i) \sim n(1+\frac{1}{c}\int f^2).$$

To show that $(T-E(T))/n \to 0$ in probability (which is all that is left), we could verify the conditions of the weak law of large numbers for triangular arrays of non-identically distributed random variables (see e.g., Loeve (1963, p. 317)). Instead, we will proceed in a more direct fashion. We will show the stronger result that $E(|T-E(T)|)/n \to 0$. We have

$$|T-E(T)| \le |\sum_{i=1}^{m} (N_i^2 - E(N_i^2)) I_{|N_i^2 - E(N_i^2)| \ge K}|$$

$$+ |\sum_{i=1}^{m} (N_i^2 - E(N_i^2)) I_{|N_i^2 - E(N_i^2)| \le K}| = |I| + II ,$$

$$E(|I|) \le \sum_{i=1}^{m} [E(N_i^2 I_{N_i^2 \ge K/2}) + E(N_i^2) I_{E(N_i^2) \ge K/2} + E(N_i^2) P(N_i^2 \ge K/2$$

$$+ E(N_i^2) I_{E(N_i^2) \ge K/2}],$$

$$II = |\sum_{i=1}^{m} Y_i| \le |\sum_{i=1}^{m} (Y_i - E(Y_i))| + |\sum_{i=1}^{m} E(Y_i)| = III + IV ,$$

where $Y_i = (N_i^2 - E(N_i^2)) I_{|N_i^2 - E(N_i^2)| \le K}$,

$$IV = |E(I)| ,$$

and

$$E(IV) \le E(|I|) .$$

Now, first choose K large enough so that $\limsup_{n \to \infty} E(|I|)/n < \epsilon$, where ϵ is an arbitrary positive integer. (This can be done in view of Lemma 1.2.) Now, we need only show that for every K, $E(III)/n \to 0$. But this is an immediate consequence of the fact that the $Y_i - E(Y_i)$ terms are independent zero mean bounded random variables (see e.g. section 16 of Loeve (1963)).

This completes the first part of the proof of Theorem 1.2. The argument for D_U is left as an exercise: first, argue again by Poissonization that it suffices to consider independent N_i's that are Poisson (np_i) distributed. Then note that we need only show that $\sum_{i=1}^{m} p_i (N_i - np_i) \to 0$ in probability.

1.3. VARIANCE.

The results obtained so far are more qualitative than quantitative: we know now for example that $E(T)$ grows as $n(1+\frac{1}{c}\int f^2)$ and that $|T-E(T)|/n$ tends to 0 in probability and in the mean. Yet, we have **not** established just how close T is to $E(T)$. Thus, we should take our analysis a step further and get a more refined result. For example, we could ask how large $Var(T)$ is. Because of the relations between C, D_S and T, we need only consider $Var(T)$ as $Var(C) = Var(T)/4$ and $Var(D_S) = Var(T)/(4n^2)$. $Var(D_U)$ is treated separately.

Theorem 1.3.

A. For all f, we have

$$\frac{1}{n} Var(T) \to \frac{4}{c^2}(\int f^3 - (\int f^2)^2) + \frac{2}{c}\int f^2 \ge \frac{2}{c}\int f^2$$

where the right-hand-side remains valid even if $\int f^2$ or $\int f^3$ are infinite. (To avoid $\infty - \infty$, consider only the lowest bound in such situations.)

B. For all f

$$n\ Var(D_U) \to c^{-2} (\int f^3 - (\int f^2)^2).$$

Here, the right-hand-side should formally be considered as ∞ when either $\int f^2 = \infty$ or $\int f^3 = \infty$.

We note that for all f, $(\int f^2)^2 \le \int f^3$ (Jensen's inequality), and that equality is reached for the uniform density on [0,1]. Thus, once again, the uniform density mimimizes the "cost", now measured in terms of variances. In fact, for the uniform density, we have $Var(D_U) = 0$, all n, and $Var(T) = 2n - 4 - \frac{4}{n} + \frac{6}{n^2}$ when $c = 1$, $m = n$.

For the proof of Theorem 1.3, the reader should consult section 5.1 first. We note here that the Poissonization trick of section 1.2 is no longer of any use because the variance introduced by it, say, $Var(T' - T)$ for $n' = n$ (see notation of the proof of Theorem 1.1), grows as n, and is thus asymptotically nonnegligible.

Proof of Theorem 1.3.

Consider T first. We will repeatedly use Lemma 5.1 because $N_1, \ldots, N_m$ are multinomial $(n, p_1, \ldots, p_m)$. Thus, omitting the fact that we are constantly summing for i and j from 1 to m we have

$$E^2(T) = \sum E^2(N_i) + \sum_{i \neq j} E(N_i^2) E(N_j^2)$$

$$= \sum [n^2(n-1)^2 p_i^4 + 2n^2(n-1)p_i^3 + n^2 p_i^2]$$

$$+ \sum_{i \neq j} [n^2(n-1)^2 p_i^2 p_j^2 + n^2(n-1)(p_i p_j^2 + p_i^2 p_j) + n^2 p_i p_j]$$

where we used the fact that $E^2(N_i) = n(n-1)p_i^2 + np_i$. Using various expressions from Lemma 5.1, we have

$$E(T^2) = \sum E(N_i^4) + \sum_{i \neq j} E(N_i^2 N_j^2)$$

$$= \sum [np_i + 7n(n-1)p_i^2 + 6n(n-1)(n-2)p_i^3 + n(n-1)(n-2)(n-3)p_i^4]$$

$$+ \sum_{i \neq j} [n(n-1)(n-2)(n-3)p_i^2 p_j^2 + n(n-1)(n-2)(p_i p_j^2 + p_i^2 p_j) + n(n-1)p_i p_j$$

Because $Var(T) = E(T^2) - E^2(T)$, we have

$$Var(T) = \sum [(-4n^3 + 10n^2 - 6)p_i^4 + (4n^3 - 16n^2 + 12)p_i^3 + (6n^2 - 7n)p_i^2 + np_i$$

$$+ \sum_{i \neq j} [(-4n^3 + 10n^2 - 6)p_i^2 p_j^2 + (-2n^2 + 2n)(p_i^2 p_j + p_i p_j^2) + (-n)p_i p_j].$$

Using the facts that $\sum p_i = 1$, $\sum_{i \neq j} p_i p_j = \sum p_i(1 - p_i) = 1 - \sum p_i^2$, $\sum_{i \neq j} p_i^2 p_j = \sum p_i^2(1 - p_i) = \sum p_i^2 - p_i^3$, and $\sum_{i \neq j} p_i^2 p_j^2 = \sum p_i^2 (\sum p_j^2 - p_i^2) = (\sum p_i^2)^2 - \sum p_i$, we see that

$$Var(T) = (-4n^3 + 10n^2 - 6)(\sum p_i^2)^2$$

$$+ (4n^3 - 12n^2 - 4n + 12) \sum p_i^3 + (2n^2 - 2n) \sum p_i^2 .$$

By Lemma 1.1, we have for all constants $r \geq 1$, $\sum p_i^r \sim (nc)^{-(r-1)} \int f^r$. Thus, if $\int f^2 < \infty$,

$$Var(T) \sim -4n^3(nc)^{-2}(\int f^2)^2 + 4n^3(nc)^{-2} \int f^3 + 2n^2(nc)^{-1} \int f^2 ,$$

which gives us our expression. The right-hand-side of this expression is nonsense if both $\int f^2$ and $\int f^3$ are ∞. In that case, note that $(\sum p_i^2)^2 \leq \sum p_i^3$ (by Jensen's inequality), and that thus, because $\sum p_i^3 = o(\sum p_i^2)$,

$$Var(T) \geq (-2n^2 - 4n + 6) \sum p_i^3 + (2n^2 - 2n) \sum p_i^2 \sim 2n^2 \sum p_i^2$$

so that $Var(T)/n \to \infty$. This concludes the proof of the first half of Theorem 1.3.

We have $E(D_U) = \sum np_i^2 \sim \frac{1}{c} \int f^2$, and

$$E(D_U^2) = E(\sum p_i^2 N_i^2) + E(\sum_{i \neq j} p_i p_j N_i N_j)$$

$$= \sum p_i^2 (np_i + n(n-1)p_i^2) + \sum_{i \neq j} p_i^2 p_j^2 n(n-1)$$

$$= n \sum p_i^3 + n(n-1)(\sum p_i^2)^2 .$$

Thus,

$$Var(D_U) = n \sum p_i^3 - n(\sum p_i^2)^2 .$$

If $\int f^3 < \infty$, then $Var(D_U) \sim \frac{1}{nc^2}(\int f^3 - (\int f^2)^2)$. If $\int f^3 = \infty$ but $\int f^2 < \infty$, this is still true. If both integrals are infinite, we need an additional argument. For example, let J be the collection of indices for which $p_i > a/m$, where $a > 0$ is a constant. We have, by the inequality $(u+v)^2 \leq 2u^2+2v^2$,

$$Var(D_U) \geq n\sum_J p_i^3 - 2n(\sum_J p_i^2)^2 + n\sum_{J^c} p_i^3 - 2n(\sum_{J^c} p_i^2)^2$$

where J^c is the complement of J. By Jensen's inequality,

$$\sum_J p_i^3 \sum_J p_i \geq (\sum_J p_i^2)^2 ,$$

and similarly for J^c. Thus, we have

$$Var(D_U) \geq n(\sum_J p_i^2)^2((\sum_J p_i)^{-1} - 2) + n(\sum_{J^c} p_i^2)^2((\sum_{J^c} p_i)^{-1} - 2).$$

It is a simple exercise to show that $m\sum_{J^c} p_i^2 \to \int_{f \leq a} f^2$, $\sum_{J^c} p_i \to \int_{f \leq a} f$, $\sum_J p_i \to \int_{f > a} f$, $m\sum_J p_i^2 \to \infty$. For any choice of a with $\int_{f > a} f \in (0,1)$, we have thus $n \; Var(D_U) \to \infty$.

1.4. LARGE DEVIATION INEQUALITIES.

Often, one would like to know the likelihood of the event $[C > x]$ (or of $[D_S > x]$ or $[D_U > x]$), and in the absence of an exact answer, good upper bounds for the corresponding probabilities $P(C > x)$, $P(D_S > x)$ and $P(D_U > x)$ will do too. For the sake of simplicity, we will derive such upper bounds for $P(D_U > x)$. The analysis for C and D_S is considerably more complicated.

First, we observe that there is little hope to get a small bound unless x exceeds $E(D_U) \sim \frac{1}{c_n}\int f^2$. Thus, we will ask for upper bounds for the probability

$$P(D_U > \frac{1}{c_n}\int f^2 (1+\epsilon)) , \epsilon > 0.$$

From Markov's inequality and Theorem 1.1, we have

$$P(D_U > \frac{1}{c_n}\int f^2 (1+\epsilon)) \leq \frac{E(D_U)}{\frac{1}{c_n}\int f^2 (1+\epsilon)} \leq \frac{1}{1+\epsilon} ,$$

valid for all f . Unfortunately, this bound requires large values of ϵ to be useful. By restricting ourselves to smaller classes of densities, we can obtain smaller upper bounds.

For example, by the Chebyshev-Cantelli inequality and $E(D_U) \leq c_n^{-1}\int f^2$, we have

$$P(D_U \geq (1+\epsilon)c_n^{-1}\int f^2) \leq P(D_U - E(D_U) \geq \epsilon c_n^{-1}\int f^2)$$

$$\leq Var(D_U)/(Var(D_U) + \epsilon^2(\int f^2)^2 c_n^{-2})$$

$$\leq (\int f^3/(\int f^2)^2 - 1)/(n\epsilon^2)$$

if $\int f^2 < \infty$. The upper bound is obviously useless when $\int f^3 = \infty$. When $\int f^3 < \infty$, it decreases with n for every $\epsilon > 0$. Unfortunately, the decrease is only as $1/n$. Better rates can be obtained at the expense of stricter conditions on f . For example, we can hope to obtain bounds that decrease as $(n\epsilon^2)^{-r}$ for arbitrary $r > 1$ provided that $\int f^p < \infty$ for an appropriately big constant p .

The conditions $\int f^p < \infty$, $p > 1$, are conditions restricting the size of the infinite peaks of f . The strongest possible peak condition is "$f \leq C$ for some constant C" . In that case, we can obtain an exponential inequality:

Theorem 1.4.

Assume that $\sup f \leq C < \infty$. For all $\epsilon > 0$, we have

$$P(D_U \geq (1+\epsilon)c_n^{-1}\int f^2) \leq \exp(-A(\epsilon)n)$$

where

$$A(\epsilon) = \sup_{r>0} r\epsilon\int f^2 - \frac{r^2}{2}\int f^3 e^{rC} > 0.$$

In particular, if $\epsilon = \epsilon_n$ varies with n in such a way that $\epsilon_n \downarrow 0$, then

$$A(\epsilon_n) \sim \frac{1}{2}\epsilon_n^2 (\int f^2)^2/\int f^3 ;$$

and if $\epsilon_n \uparrow \infty$, then

$$A(\epsilon_n) \sim \frac{1}{2C}\int f^2 \epsilon_n \log \epsilon_n .$$

Proof of Theorem 1.4.

The proof is based upon Chernoff's bounding technique and a simple expression for the moment generating function of the multinomial distribution (see Lemma 5.2). Let $t > 0$ be an arbitrary number. Then

$$P(D_U = \sum_{i=1}^{m} N_i p_i > (1+\epsilon)\frac{1}{c_n}\int f^2)$$

$$\leq E(\exp(-t(1+\epsilon)\frac{1}{c_n}\int f^2 + t \sum_{i=1}^{m} N_i p_i))$$

$$= \exp(-t\frac{1}{c_n}\int f^2(1+\epsilon)) (\sum_{i=1}^{m} p_i \exp(tp_i))^n .$$

Let us recall the definition of the function f_n from Lemma 1.1. Using the fact that $e^u - 1 \leq u + \frac{u^2}{2}e^u$ for $u > 0$, we have the following chain of equalities and inequalities (where the first expression is equal to the last expression of the chain given above):

$$\exp(-tc_n^{-1}(1+\epsilon)\int f^2). (\int f_n \exp(\frac{t}{m} f_n)dx)^n$$

$$= \exp(-tc_n^{-1}(1+\epsilon)\int f^2). (1+ \int f_n (\exp(\frac{t}{m} f_n)-1)dx)^n$$

$$\leq \exp(-tc_n^{-1}(1+\epsilon)\int f^2). (1+\frac{t}{m}\int f_n^2 + \frac{t^2}{2m^2}\int f_n^3 \exp(\frac{t}{m} f_n))^n$$

$$\leq \exp(-tc_n^{-1}(1+\epsilon)\int f^2 + tc_n^{-1}\int f_n^2 + n\frac{t^2}{2m^2}\int f_n^3 \exp(\frac{t}{m}C))$$

$$\leq \exp(-tc_n^{-1}\epsilon\int f^2 + n\frac{t^2}{2m^2}\int f^3 \exp(\frac{t}{m}C)).$$

Here we also used the inequality $(1+u) \leq \exp(u)$, and the fact that $\int f_n^s \leq \int f^s$ for all $s \geq 1$ (Lemma 1.1). The first half of the Theorem follows from the choice $t = rm$. Now, as $\epsilon \downarrow 0$, we see that the supremum is reached for $r = r(\epsilon) > 0$, and that $A(\epsilon)$ is asymptotic to the value $\sup_{r>0} r\epsilon\int f^2 - \frac{1}{2}r^2\int f^3$. The latter supremum, for each $\epsilon > 0$, is reached for $r = \epsilon\int f^2/\int f^3$. Resubstitution gives the desired solution, $A(\epsilon) \sim \frac{1}{2}\epsilon^2(\int f^2)^2/\int f^3$.

When $\epsilon \uparrow \infty$, it is easy to see that the supremem in the expression for $A(\epsilon)$ is reached for $r(\epsilon) \uparrow \infty$. By standard functional iterations, applied to the equation $r(\epsilon) = \frac{1}{C}\log(\epsilon\int f^2/(r(\epsilon)\int f^3))$, we see that $A(\epsilon) \sim$ the value of the expression to be optimized, at $r = \frac{1}{C}\log(\epsilon\int f^2/(\int f^3 \frac{1}{C}\log\epsilon))$, which gives us our solution.

Remark.

The inequality of Theorem 1.4 for $\epsilon_n \downarrow 0$, $n\,\epsilon_n{}^2 \uparrow \infty$, is called a **moderate deviation inequality.** It provides us with good information about the tail of the distribution of D_U for values of the order of magnitude of the mean of D_U plus a few standard deviations of D_U. On the other hand, when ϵ_n is constant or tends to ∞, we have **large deviation inequalities.** As a rule, these should give good information about the extreme tail of the distribution, where the central limit theorem is hardly at work. For example, it appears from the form of the inequality that the extreme tail of D_U drops off at the rate of the tail of the Poisson distribution.

1.5. DOUBLE BUCKETING.

The results that we have obtained until now qualify the statement that T is close to $n\,(1+\frac{1}{c}\int f^{\,2})$ when $\int f^{\,2} < \infty$. The presence of $\int f^{\,2}$ in this expression is disappointing. Perhaps we could hope to reduce the direct influence of f on the quantities that are of interest to us by hashing the n intervals a second time: each interval A_i is subdivided into N_i equal subintervals. This method will be referred to as the "double bucketing" method. The idea of double bucketing is obviously not novel (see for example Maclaren, 1966). In fact, we could keep on dividing intervals until all data points are in separate intervals. The structure thus obtained is called an N-tree (Ehrlich (1982), Tamminen (1982)). Some analysis for restricted classes of densities is given in these papers. Recursive bucketing when applied to sorting is analyzed in Doboslewicz (1978) and Van Dam, Frenk and Rinnooy Kan (1983).

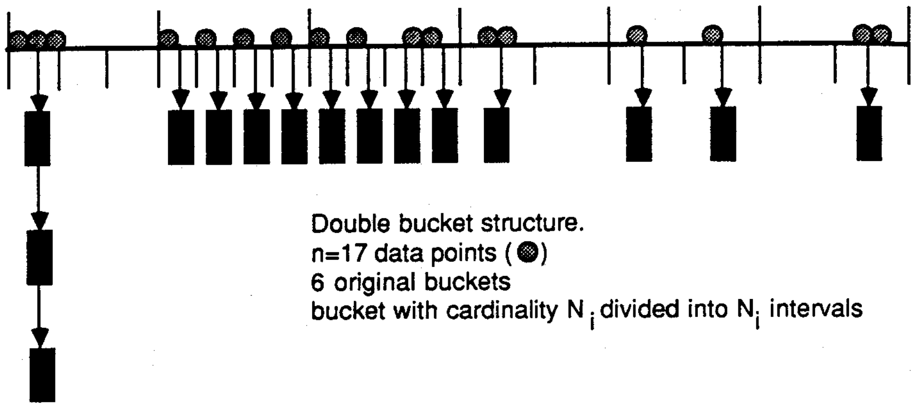
What we will try to show here is that most of the benefits of recursive bucketing are obtained after two passes, i.e. with double bucketing. The structure that we will analyze is obtained as follows:

Step 1.

Let $A_i = [\frac{i-1}{n}, \frac{i}{n})$, $1 \le i \le n$. For each A_i, keep a lined list of $X_j{}'s$ falling in it. Let N_i be the cardinality of A_i.

Step 2.

For $i = 1$ to n do : if $N_i \ge 1$, divide A_i into N_i equal intervals A_{ij}, and keep for each A_{ij} linked lists of the data points in it. Let N_{ij} be the cardinality of A_{ij}.



Double bucket structure.
n=17 data points (●)
6 original buckets
bucket with cardinality N_i divided into N_i intervals

Figure 1.4.

The quantities that we will consider here are

$$T = \sum_{i=1}^{n} \sum_{j=1}^{N_i} N_{ij}{}^2 ,$$

$$C = \sum_{i=1}^{n} \left(\frac{1}{2} \sum_{j=1}^{N_i} (N_{ij}{}^2 - N_{ij})\right) = \frac{1}{2}(T-n),$$

$$D_S = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \frac{1}{2}(N_{ij}{}^2 + N_{ij}) = \frac{1}{2n}(T+n),$$

and

$$D_U = \sum_{i=1}^{n} \sum_{j=1}^{N_i} p_{ij} N_{ij}$$

where all the summations $\sum_{j=1}^{N_i}$ for $N_i = 0$ must be omitted, and $p_{ij} = \int_{A_{ij}} f$ when A_{ij} is defined. We note that the first division is into n

intervals. The generalization towards a division into m intervals is straighforward.

Theorem 1.5.

If $\int f^2 < \infty$, then the double bucketing structure gives

$$E(T)/n \sim 1 + \int_0^1 e^{-f} \; ; E(C)/n \sim \frac{1}{2} \int_0^1 e^{-f} \; ; E(D_S) \sim \frac{1}{2}(2+\int_0^1 e^{-f})$$

and

$$E(D_U) \rightarrow 1 ..$$

If we compare these asymptotic expressions with those for ordinary bucketing when $m=n$, i.e. $E(T)/n \sim 1+\int f^2$, we see that double bucketing is strictly better for all f. This follows from Jensen's inequality and the fact that $e^{-u} \leq 1-u+\frac{1}{2}u^2$:

$$\int f^2 > \frac{1}{2}\int f^2 \geq \int_0^1 e^{-f} \geq \exp(-\int_0^1 f) = \frac{1}{e} .$$

For all f with $\int f^2 < \infty$, we have

$$\lim_{n \to \infty} \frac{E(T)}{n} \in [1+\frac{1}{e}, 2) .$$

Thus, the limit of $E(T)/n$ is uniformly bounded over all such f. In other words, double bucketing has the effect of eliminating all peaks in densities with $\int f^2 < \infty$. Let us also note in passing that the lower bound for $E(T)/n$ is reached for the uniform density on [0,1], and that the upper bound can be approached by considering densities that are uniform on [0,1], and that the upper bound can be approached by considering densities that are uniform on $[0,\frac{1}{K}]$ $(\int_0^1 e^{-f} = 1-\frac{1}{K}+\frac{1}{K}e^{-K})$ and letting $K \rightarrow \infty$. The fact that the properties of the double bucketing structure are basically independent of the density f was observed independently by Tamminen (1985). The same is a fortiori true for N-trees (Ehrlich (1981), Van Dam, Frenk and Rinnooy Kan (1983), Tamminen (1983)).

Proof of Theorem 1.5.

In the proof, all summations $\sum_{j=1}^{N_i}$ for which $N_i = 0$ should be omitted, to avoid trivialities. We start with a lower bound for $E(T)$.

$$E(T) = \sum_{i=1}^{n} E(I_{N_i \geq 1} \sum_{j=1}^{N_i} [(N_i^2-N_i)(p_{ij}/p_i)^2 + N_i p_{ij}/p_i])$$

$$= \sum_{i=1}^{n} E(N_i) + \sum_{i=1}^{n} E((N_i^2-N_i) \sum_{j=1}^{N_i} (p_{ij}/p_i)^2)$$

$$\geq n + \sum_{i=1}^{n} E((N_i^2-N_i) \sum_{j=1}^{N_i} (\frac{1}{N_i})^2)$$

$$= n + \sum_{i=1}^{n} E((N_i-1)_+) \text{ (where } u_+=\max(u,0))$$

$$= n + \sum_{i=1}^{n} E(N_i-1) + \sum_{i=1}^{n} P(N_i=0)$$

$$= n + \sum_{i=1}^{n} P(N_i=0)$$

$$= n + \sum_{i=1}^{n} (1-p_i)^n \text{ (where } p_i = \int_{A_i} f \text{)}$$

$$\geq n + \sum_{i=1}^{n} \exp(-np_i/(1-p_i)) \text{ (because } 1-u \geq \exp(-u/(1-u)), 0 \leq u < 1)$$

$$= n + n \int_0^1 \exp(-f_n/(1-f_n/n)) \text{ (where } f_n(x) = np_i , x \in A_i)$$

$$\sim n + n \int_0^1 e^{-f}$$

by the Lebesgue dominated convergence theorem and Lemma 5.10.

We now derive an upper bound for $E(T)$. For any integer K, we have

$$E(T) = n + \sum_{i=1}^{n} E(V_i{}') + \sum_{i=1}^{n} E(V_i{}'')$$

where

$$V_i{}' = (N_i{}^2 - N_i) \sum_{j=1}^{N_i} (p_{ij}/p_i)^2 I_{N_i \le K}$$

and

$$V_i{}'' = (N_i{}^2 - N_i) \sum_{j=1}^{N_i} (p_{ij}/p_i)^2 I_{N_i > K} .$$

The statements about $E(T)$, $E(C)$ and $E(D_S)$ in Theorem 1.5 are proved if we can show that

$$\lim_{K \to \infty} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E(V_i{}') = \int_0^1 e^{-f} \ ;$$

$$\lim_{K \to \infty} \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E(V_i{}'') = 0.$$

We will use the function $g_n(x) = E(V_i{}')$, $x \in A_i$. Clearly,

$$g_n(x) \le K \ E(N_i) = Knp_i = Kf_n(x) \, , \, x \in A_i \, ,$$

$$\int f_n = 1, \text{ all } n \ ; \ f_n \to f \ \text{ almost all } x.$$

Thus, by an extended version of the Lebesgue dominated convergence theorem (see e.g. Royden (1968, p. 89)), we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E(V_i{}') = \lim_{n \to \infty} \int_0^1 g_n = \int_0^1 \lim_{n \to \infty} g_n$$

provided that the limit of g_n exists almost everywhere. Consider now a sequence of couples (i,j) such that $x \in A_{ij} \subseteq A_i$ for all n. We have by Lemma 5.11, $nN_i p_{ij} \to f(x)$ for almost all x, uniformly in N_i, $1 \le N_i \le K$. From this, we conclude that

$$g_n(x) \sim E((N_i - 1)_+ I_{N_i \le K}) \, , \text{ almost all } x.$$

Consider only those x's for which $f(x) > 0$, and Lemma 5.11 applies. Clearly, N_i tends in distribution to Z where Z is a Poisson $(f(x))$ random variable (this follows from $np_i \to f(x)$ (Chow and Teicher (1978, p. 36-37))). Since $(N_i - 1)_+ I_{N_i \le K}$ forms a sequence of bounded random variables, we also have convergence of the moments, and thus,

$$g_n(x) \sim E((Z-1)_+ I_{Z \le K}) = f(x) - 1 + e^{-f(x)} - E((Z-1)_+ I_{Z > K})$$

for all such x, i.e. for almost all $x(f)$. Thus,

$$\lim_{K \to \infty} \lim_{n \to \infty} \frac{1}{n} E(V_i{}') = \lim_{K \to \infty} \int_0^1 (f(x) - 1 + e^{-f(x)} - E((Z-1)_+ I_{Z > K})) \, dx$$

$$= \int_0^1 e^{-f} \ .$$

Here we needed the fact that $\lim_{K \to \infty} \int_0^1 E((Z-1)_+ I_{Z > K}) \, dx = 0$, which is a simple consequence of the Lebesgue dominated convergence theorem (note that $\int_0^1 E(Z) dx = 1$). Also,

$$\frac{1}{n} \sum_{i=1}^{n} E(V_i{}'') \le \sum_{i=1}^{n} E(N_i{}^2 I_{N_i > K}) .$$

Define the function $h_n(x) = E(N_i{}^2 I_{N_i > K})$, $x \in A_i$, and the function $h(x) = E(Z^2 I_{Z > K})$ where Z is Poisson $(f(x))$ distributed. We know that $h_n(x) \le E(N_i{}^2) \le np_i + (np_i)^2 = f_n(x) + f_n{}^2(x) \to f(x) + f^2(x)$, almost all x; and that $\int f_n + f_n{}^2 \to \int f + f^2$. Thus, by an extension of the Lebesgue dominated convergence theorem, we have

$$\frac{1}{n} \sum_{i=1}^{n} E(V_i{}'') \leq \int_0^1 h_n \rightarrow \int_0^1 \lim_{n \to \infty} h_n$$

provided that the almost everywhere limit of h_n exists. For almost all x, N_i tends in distribution to Z. Thus, for such x,

$$|h_n - h| \leq \sum_{j=1}^{\infty} j^2 |P(N_i = j) - P(Z = j)| \to 0$$

(see e.g. Simons and Johnson, 1971). But $\int_0^1 h \to 0$ as $K \to \infty$ since $\int_0^1 E(Z^2) = \int_0^1 f + f^2 < \infty$, and $E(Z^2 I_{Z > K}) \to 0$ for almost all x. This concludes the proof of

$$\limsup_{K \to \infty} \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E(V_i{}'') = 0.$$

We will only sketch the proof for

$$E(D_U) = E(\sum_{i=1}^{n} \sum_{j=1}^{N_i} p_{ij} N_{ij}) = E(\sum_{i=1}^{n} p_i N_i \sum_{j=1}^{N_i} (p_{ij}/p_i)^2) .$$

First, it is easily seen that

$$E(D_U) \geq E(\sum_{i=1}^{n} p_i N_i / N_i) = \sum_{i=1}^{n} p_i = 1.$$

Also, if we follow the treatment to obtain an upper bound for $E(T)$, we come across terms $V_i{}'$ and $V_i{}''$ in which $(N_i{}^2 - N_i)$ is now replaced by $p_i N_i$. Mimicking the Poisson approximation arguments for $E(T)$, we obtain $\limsup_{n \to \infty} E(D_U) \leq 1$ when $\int f^2 < \infty$. This concludes the proof of Theorem 1.5.

Chapter 2

DENSITIES WITH UNBOUNDED SUPPORT

2.1. MAIN RESULTS.

In chapter 1, we have analyzed in some detail what happens when f is known to have support contained in [0,1]. In first approximation, the main term in the asymptotic expressions for $E(T)/n$ and $E(D_U)$ contain the factor $\int f^2$, which is scale-dependent. If we were to divide the interval $[M_n, M_n{}^*] =$ [min X_i, max X_i] into m equal-sized sub-intervals, these expected values would obviously not be scale-dependent because the distribution of $N_1, \ldots, N_m$ is scale invariant.

We could repeat all of chapter 1 for this more general setting if tedium was no deterrent. There is a new ingredient however when f has infinite tails because M_n and / or $M_n{}^*$ diverges in those cases. The results in this chapter rely heavily on some results from the theory of order statistics. The technicalities are deferred to section 2.2. The following notation will be introduced:

$$M_n = \min_{1 \leq i \leq n} X_i ,$$

$$M_n{}^* = \max_{1 \leq i \leq n} X_i ,$$

$$R_n = \text{range}(X_1, \ldots, X_n) = M_n{}^* - M_n ,$$

$$x_i = M_n + \frac{i-1}{m}(M_n{}^* - M_n) , 1 \leq i \leq m+1 ,$$

$$p_i = \int_{x_i}^{x_{i+1}} f , 1 \leq i \leq m ,$$

$$p = \int_{M_n}^{M_n{}^*} f ,.$$

s = ess sup X_1 – ess inf X_1 = width of support of f.

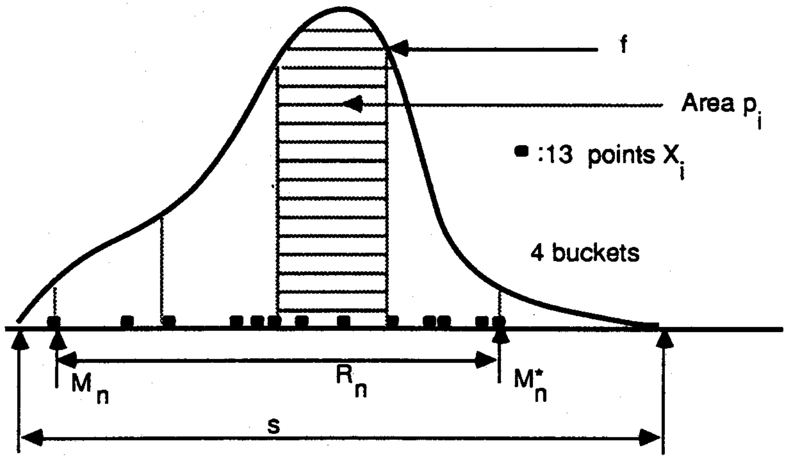


Figure 2.1.

Theorem 2.1.

Let f be a density on R^1 with $\int f^2 < \infty$. Then

$$\text{(i)}\quad \liminf_{\delta\downarrow 0}\ \liminf_{n\to\infty} \frac{E(T)}{(n(1+E(\min(R_n c_n^{-1}\int f^2, \delta n)))} \geq 1$$

and

$$\text{(ii)}\quad \limsup_{n\to\infty} \frac{E(T)}{n(1+E(\min(R_n c_n^{-1}\int f^2, n)))} \leq 1 .$$

In particular, if $s < \infty$, we have

$$\liminf_{n\to\infty} \frac{E(T)}{n} = \limsup_{n\to\infty} \frac{E(T)}{n} = 1 + s\,\frac{1}{c}\int f^2.$$

Theorem 2.1 shows that there is a close relation between $E(T)$ and the range R_n. For densities with no tails, we have a generalization of Theorem 1.1. It is noteworthy that $1+\frac{s}{c}\int f^2$, the limit value of $E(T)/n$, is scale invariant. When $s = \infty$, it is not clear at all how $E(\min(R_n c_n^{-1}\int f^2, n))$ varies with n. For example, is this quantity close to $E(R_n)c_n^{-1}\int f^2$ (which is easier to handle)? Thus, to apply Theorem 2.1 in concrete examples, some results are needed for R_n. Some of these are stated in Lemma 2.1.

We will work with the following quantities: $X = X_1$ has density f and distribution function $F(x) = P(X \leq x) = 1-G(x)$; the integrals

$$\bar{F}(x) = \int_{-\infty}^{x} F(t)dt \ ;\ \bar{G}(x) = \int_{x}^{\infty} G(t)dt$$

will also be useful. We recall that

$$E(|X|) = \bar{G}(0)+\bar{F}(0) = \int_0^\infty G(t)dt + \int_{-\infty}^{0} F(t)dt \ .$$

Lemma 2.1

Let $\delta > 0$ be arbitrary. Then:

(i) $E(\min(R_n, \delta n))\uparrow$.

(ii) $\limsup_{n\to\infty} E(\min(R_n, \delta n)) < \infty$ if and only if $s < \infty$.

(iii) $\limsup_{n\to\infty} E(R_n) < \infty$ if and only if $s < \infty$.

(iv) $E(R_n) = \infty$ for all $n \geq 2$ if and only if $E(R_n) = \infty$ for some $n > 2$ if and only if $E(|X|) = \infty$.

(v) $E(|X|) < \infty$ implies $E(R_n) = o(n)$.

(vi) $E(|X|) < \infty$,

$$\lim_{a \downarrow 0} \liminf_{x \to \infty} \frac{\breve{G}(ax)}{\breve{G}(x)} = \infty \quad (\frac{0}{0} = \infty)$$

and

$$\lim_{a \downarrow 0} \liminf_{x \to \infty} \frac{\breve{F}(-ax)}{\breve{F}(-x)} = \infty$$

imply

$$E(\min(R_n, \delta n)) \sim E(R_n) \text{ for all } \delta > 0.$$

(vii) Are equivalent:

$$\limsup_{n \to \infty} E(\min(R_n, \delta n))/n > 0 \text{ for all } \delta > 0 \text{ ;}$$

$$\limsup_{n \to \infty} E(\min(R_n, \delta n))/n > 0 \text{ for some } \delta > 0 \text{ ;}$$

$$\limsup_{x \to \infty} |x| P(|X| > x) > 0.$$

(viii) Are equivalent:

$$\liminf_{n \to \infty} E(\min(R_n, \delta n))/n > 0 \text{ for all } \delta > 0 \text{ ;}$$

$$\liminf_{n \to \infty} E(\min(R_n, \delta n))/n > 0 \text{ for some } \delta > 0 \text{ ;}$$

$$\liminf_{x \to \infty} |x| P(|X| > x) > 0.$$

Lemma 2.1 in conjunction with Theorem 2.1 gives us quite a bit of information about $E(T)$. For example, we have

Theorem 2.2.

If $\int f^2 < \infty$, then are equivalent:

$$\liminf_{n \to \infty} E(T)/n < \infty \text{ ;}$$

$$\limsup_{n \to \infty} E(T)/n < \infty \text{ ;}$$

$$s < \infty .$$

(And if $s < \infty$, this lim inf is equal to this lim sup. Its value $1+\frac{s}{c}\int f^2$.)

Theorem 2.2 follows from Lemma 2.1 (i), (ii) and Theorem 2.1. In Devroye and Klincsek (1980), one finds a slightly stronger result: $E(T) = 0(n)$ if and only if $s < \infty$ and $\int f^2 < \infty$. In the next chapter, this will be generalized to R^d, so we don't have to bother with an R^1 version of it here.

We also have

Theorem 2.3.

If $\int f^2 < \infty$, then condition (vi) of Lemma 2.1 implies that

$$E(T) \sim n(1+\frac{1}{c} E(R_n) \int f^2) .$$

Theorems 2.2 and 2.3 cover all the small-tailed distributions with little oscillation in the tails. In Akl and Meijer (1982) the upper bound part of Theorem 2.3 was obtained for bounded densities. The actual limiting expression of $E(T)$ shows the interaction between the effect of the peaks ($\int f^2$) and the effect of the tails ($E(R_n)$). Note that $E(R_n)\int f^2$ is a scale-invariant and translation-invariant quantity: it is solely determined by the shape of the density. It is perhaps interesting to see when condition (vi) of Lemma 2.1 is valid.

Example 2.1. (Relatively stable distributions.)

A relatively stable distribution is one for which

$$\text{(i)} \lim_{x \to \infty} \frac{G(ax)}{G(x)} = \infty \text{ , all } a \in (0,1) \text{ ; } (\frac{0}{0} = \infty) \text{ ;}$$

and

$$\text{(ii)} \lim_{x \to \infty} \frac{F(-ax)}{F(-x)} = \infty \text{ , all } a \in (0,1) \text{ .}$$

If we use the notation $M_n^* = \max(X_1^+, \ldots, X_n^+)$ where $u^+=\max(u,0)$ then it should be noted that if $P(X > 0) > 0$, (i) is equivalent to

(iii) $M_n^* \to 1$ in probability for some sequence a_n

(Gnedenko, 1943). In that case we can take $a_n = \inf(x: G(x) \le \frac{1}{n})$ where $G(x)=P(X \ge x)$, or in short, $a_n = G^{-1}(\frac{1}{n})$ (Dehaan, 1975, pp. 117). We note that (i) is equivalent to $\tilde{G}(0) < \infty$, $\tilde{G}(x)/(xG(x)) \to 0$ as $x \to \infty$; or to $\tilde{G}(0) < \infty$, $\int_x^\infty tdF(t)/(xG(x)) \to 1$ as $x \to \infty$.

For relatively stable distributions, we have $E(R_n) \sim F^{-1}(\frac{1}{n})+G^{-1}(\frac{1}{n})$ (Pickands, 1968). It is very easy to check that condition (vi) follows from the relative stability of the distribution of X. When

$$\text{(iv)} \lim_{x \to \infty} \frac{xf(x)}{G(x)} = \infty ,$$

we know that (iii) holds (Geffroy, 1958; Dehaan, 1975, Theorem 2.9.2). condition (iv) comes close to being best possible because if f is nonincreasing and positive for all x, then (iii) implies (iv) (Dehaan, 1975, Theorem 2.9.2).

Example 2.2. (Normal distribution.)

For the normal distribution with density $(2\pi)^{-1/2}\exp(-x^2/2)$, we have relative stability and square integrability. In particular, $E(R_n) \sim 2G^{-1}(\frac{1}{n}) \sim 2\sqrt{2\log n}$ (see e.g. Galambos, 1978, pp. 65), and thus

$$E(T) \sim n\ (1+2\sqrt{2\log n}\ \int f^2) = n\ (1+\sqrt{\frac{2}{\pi}\log n}\ ) \sim \sqrt{\frac{2}{\pi}}\, n\sqrt{\log n}\ .$$

Example 2.3. (Exponential distribution.)

For density $f(x) = e^{-x}$, $x > 0$, we have relative stability and square integrability. Thus, because $E(R_n) \sim \log n$,

$$E(T) \sim n(1+\log n\ \int f^2) \sim \frac{1}{2}\, n\ \log n\ .$$

Example 2.4. (Regularly varying distribution functions.)

Condition (vi) of Lemma 2.1 is satisfied for all distributions for which

(i) $\tilde{G}(x) = 0$ for all x large enough; or $\tilde{G}$ is regularly varying with coefficient $\rho < 0$ (i.e., $\tilde{G}(ax)/\tilde{G}(x) \to a^\rho$ for all $a > 0$ as $x \to \infty$).

(ii) $\tilde{F}(x) = 0$ for all x large enough; or $\tilde{F}$ is regularly varying with coefficient $\rho < 0$ (i.e., $\tilde{F}(ax)/\tilde{F}(x) \to a^\rho$ for all $a > 0$ as $x \to \infty$).

In (i) and (ii) we can replace the functions $\tilde{G}$ and $\tilde{F}$ by G and F if we wish provided that we add the condition that the coefficient of regular variation be $\rho < -1$. The latter fact follows from the observation that as $x \to \infty$, $\tilde{G}(x) \sim xG(x)/(-\rho-1)$ (Dehaan, 1975, Theorem 1.2.1).

Example 2.5. (Upper bounds for $E(R_n)$.)

One of the by-products of Theorem 2.1 is that

$$\limsup_{n \to \infty} \frac{E(T)}{n\ E(R_n)\,\frac{1}{c}\int f^2} \le 1.$$

Thus, good upper bounds for $E(R_n)$ give us good upper bounds for $E(T)/n$. For example, we have

$$E(R_n) \le E(\max_i X_i^+ - \min_i X_i^-)$$

$$\le E^{1/r}(\max_i X_i^{+r}) + E^{1/r}((-\min_i X_i^-)^r)\ \text{,all } r \ge 1\ ,$$

$$\le 2\, n^{1/r}\ E^{\frac{1}{r}}(|X|^r).$$

Thus, depending upon the heaviness of the tail of X, we obtain upper bounds for $E(T)$ that increase as $n^{1+1/r}$. We can do better when the moment generating function of X is finite in a neighborhood of the origin, i.e.

$$E(e^{t|X|}) < \infty\ \text{, for all } t \text{ in some interval } [0,\epsilon).$$

Since $u^r \le (\frac{r}{e\;t})\,e^{tu}$, $u \ge 0$,we have

$$E(R_n) \le \frac{2r}{e\;t} n^{1/r}\; E^{1/r}(e^{t|X|})$$

$$= 2\,\frac{\log n}{t}\exp(\frac{\log(E(e^{t|X|}))}{\log n})$$,for such t , and all $n \ge e$,

where we took $r = \log n$. For the t ' s in the interval $[0,\epsilon)$, we have a: $n \to \infty$,

$$E(R_n) \le (2+o(1))\,\frac{\log n}{t}.$$

Thus, the best result is obtained by setting t equal to ϵ. In particular, 1 $E(e^{t|X|}) < \infty$ for all $t > 0$ (such as for the normal density), then

$$E(R_n) = o(\log n),$$

and thus

$$E(T) = o(n\;\log n).$$

Theorem 2.2 treats densities with compact support, while Theorem 2.3 cov ers quite a few densities with finite moment. We will now skip over some dens! ties in a gray area: some have a finite first moment but do not satisfy (vi) c Lemma 2.1, and some have infinite first moment $E(|X|)$, but have relativel small tails. The worst densities are described in Theorem 2.4:

Theorem 2.4.

Let $\int f^2 < \infty$. Then

(i) $\limsup_{n\to\infty} E(T)/n^2 > 0$ if and only if $\limsup_{x\to\infty} |x|P(|X| > x) > 0$;

(ii) $\liminf_{n\to\infty} E(T)/n^2 > 0$ if and only if $\liminf_{x\to\infty} |x|P(|X| > x) > 0$;

(iii) $E(T) = o(n^2)$ if and only if $\limsup_{x\to\infty} |x|P(|X| > x) = 0$;

(Note that $T \le n^2$ for all densities, and that statement (i) implies $E(|X|) = \infty$.)

Thus, the Cauchy density $f(x)=\frac{1}{\pi}(1+x^2)^{-1}$, which satisfies (ii), must have $E(T) \ge cn^2$ for some positive constant c. If we compare Theorem 2.4 with the results of chapter 1, we notice that heavy tails are much more of a nuisance than infinite peaks: indeed, regardless of which f is chosen on [0,1], we have $E(T) = o(n^2)$; but even moderately tailed densities can lead to a lower bound for $E(T)$ of the form cn^2. Let us also point out that there are densities with $E(|X|) = \infty$ for all n, but $E(\min(R_n, \delta n)) = o(n)$ for all $\delta > 0$: just take $F(x) = 1-1/((1+x)\log(x+e))$, $x > 0$.

We conclude this section by noting that

$$E(D_S) \sim E(T)/(2n) + 1/2,$$

$$E(C) \sim \frac{E(T)-n}{2} \sim \frac{1}{2}n^2 \sum_{i=1}^{m} (p_i/p)^2 ,$$

and

$$E(D_U) \sim E(n\sum_{i=1}^{m} p_i{}^2/p).$$

Nearly all that was said about $E(T)$ remains easily extendible to $E(C)$, $E(D_S)$ and $E(D_U)$. For example, if $s < \infty$,

$$E(D_S) \sim 1 + \frac{s}{2c}\int f^2 ,$$

$$E(D_U) \sim \frac{s}{c}\int f^2$$

and

$$E(C) \sim n\ \frac{s}{2c}\int f^2 .$$

If $s = \infty$, we have $E(C) \sim \frac{1}{E(D_S}) \sim E(T)/(2n)$ and $E(D_U) \sim E(T)/n$.

We finally note that the quantity $s\int f^2$ is scale invariant and that for all densities it is at least equal to 1, in view of

$$1 = (\int_{\text{support of } f} f)^2 \le \int f^2 \int_{\text{support of } f} dx = s\int f^2 .$$

2.2. PROOFS.

Proof of Lemma 2.1.

Fact (i) is trivial. For fact (ii), we note that if $s = \infty$, we have $R_n \to \infty$ almost surely, and thus, $\liminf_{n\to\infty} E(\min(R_n,\delta n)) \ge E(\liminf_{n\to\infty} \min(R_n,\delta n)) = \infty$. Also, in all cases, $s \ge R_n$, and we are done. Fact (iii) is proved as (ii).

For item (iv), we note that $E(R_n) \le 2nE(|X|)$, that $E(R_n) \uparrow$ and that $E(R_2) = E(|X_1 - X_2|) \ge \inf_x E(|X-x|) = \infty$ when $E(|X|) = \infty$.

To show (v), it suffices to prove that $E(\max(\,|X_1|, \ldots, |X_n|\,)) = o(n)$. Let $|X_1|$ have distributed function F on $[0,\infty)$. Then for all $\epsilon > 0$,

$$E(\max(|X_1|, \ldots, |X_n|)) = \int_0^\infty 1-(1-F(x))^n\, dx$$

$$\le n\epsilon + \int_{n\epsilon}^\infty (1-(1-F(x))^n)dx \le n\epsilon + n\int_{n\epsilon}^\infty F(x)dx = n\epsilon + o(n),$$

and we are done.

We will now prove (vi). Since $\min(R_n,\delta n) \le R_n$, we need only show that $\liminf_{n\to\infty} E(\min(R_n,\delta n))/E(R_n) \ge 1$ for all $\delta > 0$. Let us define $x^+ = \max(x,0)$, $x^- = \min(x,0)$, $R^+ = \max(X_1^+, \ldots, X_n^+)$, $R^- = \min(X_1^-, \ldots, X_n^-)$. We will show that $E(R_n - \min(R_n,\delta n))/E(R_n) \to 0$ for all $\delta > 0$ and all nondegenerate distribution with $s = \infty$ (for otherwise, the statement is trivially true). Clearly, it suffices to show that for all $\delta > 0$, $E(R^+ - \min(R^+,\delta n))/E(R_n) \to 0$. If X^+ has finite support, we see that this follows from (ii). Thus, we need only consider the case that X^+ has infinite support. Now, $E(R_n) \ge E((R^+ - X)I_{R^+>0}) \ge E(R^+ I_{R^+>0}) - E(|X|) = E(R^+) - E(|X|) = \int_0^\infty 1-(1-G(t))^n\, dt - E(|X|) \sim \int_0^\infty 1-(1-G(t))^n\, dt$. Also, $E(R^+ - \min(R^+,\delta n)) = \int_{\delta n}^\infty 1-(1-G(t))^n\, dt$. We have reduced the problem to that of showing that for all $\delta > 0$,

$$\int_{\delta n}^\infty 1-(1-G(t))^n\ dt \Big/ \int_0^\infty 1-(1-G(t))^n\ dt \to 0.$$

We will need the following inequality:

$$\frac{1}{2}\min(nu,1) \le 1-(1-u)^n \le \min(nu,1)\text{ , all } n \ge 1,\ u \in [0,1].$$

This follows from $1-nu \le (1-u)^n \le e^{-nu}$; $e^{-t} \le \frac{1}{2}$ for $t \ge 1$; and $e^{-t} \le 1-\frac{t}{2}$ for $t \in [0,1]$. Thus, if $a_n = \inf(x : G(x) \le \frac{1}{n})$ and n is so large that $a_n > 0$, we have

$$\frac{1}{2} \le \int_0^\infty 1-(1-G(t))^n\, dt \ / \ (a_n + n\int_{a_n}^\infty G(t)dt\ ) \le 1.$$

Thus, we need only show that

$$n\int_{\delta n}^\infty G(t)\ dt \ / \ (a_n + n\int_{a_n}^\infty G(t)dt) \to 0\text{ , all } \delta > 0.$$

By our assumption in (vi), we have $\int_{a_n}^{\infty} G(t)dt \; / \int_{\delta n}^{\infty} G(t)dt \to \infty$ when $a_n/n \not\to 0$ (and this in turn follows of course from the fact that $\int_0^{\infty} G(t)dt < \infty$ implies $tG(t) \to 0$ as $t \to \infty$). This concludes the proof of (vi).

We will now prove (vii) and (viii) for R^+ and $\limsup_{x\to\infty}$ (or $\liminf_{x\to\infty}$) $xG(x) > 0$. The extension of the result to R_n is left as an exercise. For $\in \in (0,\delta)$ we have the following chains of inequalities:

$$\frac{1}{n}E(\min(R^+,\delta n)) = \frac{1}{n}\int_0^{\delta n} 1-(1-G(t))^n\,dt = \frac{1}{n}(\int_0^{\epsilon n} + \int_{\epsilon n}^{\delta n})$$

$$\leq \frac{1}{n}(\epsilon n + n\int_{\epsilon n}^{\delta n} G(t)dt) \leq \epsilon + \delta n G(\epsilon n) = \epsilon + \frac{\delta}{\epsilon}\epsilon n\; G(\epsilon n);$$

and

$$\frac{1}{n}E(\min(R^+,\delta n)) \geq \frac{1}{n}\int_0^{\delta n} 1-e^{-nG(t)}\,dt \geq (1-e^{-nG(\delta n)}).$$

This proves that $\limsup_{x\to\infty} xG(x) > 0$ is equivalent to $\limsup_{n\to\infty} E(\min(R^+,\delta n))/n > 0$ for all $\delta > 0$ or for some $\delta > 0$; and that similar statements are true for the limit infimum. This concludes the proof of Lemma 2.1.

We are left with the proof of Theorem 2.1. This will be taken care of in small steps. From the observation that conditional on M_n, M_n^*, the N_i's are binomially distributed with parameters $n-2$, p_i/p, we deduce the following:

Lemma 2.2.

(i) $T \leq n^2$.

(ii) $E(T|M_n,M_n^*) \leq n(1+\frac{R_n}{c_n p^2}\int_{M_n}^{M_n^*} f^2)$.

(iii) $E(T|M_n,M_n^*) \geq (n-2)^2 \sum_{i=1}^{m} p_i^2$.

Proof of Lemma 2.2.

Part (i) is obviously true. Parts (ii) and (iii) follow from

$$E(T|M_n,M_n^*) = \sum_{i=1}^{m} [(\frac{(n-2)p_i}{p})^2 + \frac{(n-2)p_i}{p}(1-\frac{p_i}{p})]$$

$$= n-2 + [(n-2)^2-(n-2)]\sum_{i=1}^{m}(p_i/p)^2$$

and the fact that

$$\sum_{i=1}^{m} p_i^2 = \sum_{i=1}^{m} (\int_{x_i}^{x_{i+1}} f \;/(R_n/m))^2 (R_n/m)^2 \leq (R_n/m)\int_{M_n}^{M_n^*} f^2 .$$

Proof of Theorem 2.1 (i)

We start from Lemma 2.2 (iii). Let $\delta > 0$ be a sufficiently small number. Then

$$\sum_{i=1}^{m} p_i^2 = \sum_{i=1}^{m} (x_{i+1}-x_i)(\int_{x_i}^{x_{i+1}} f \;/(x_{i+1}-x_i))^2$$

$$\geq \sum_{i=1}^{m} \frac{1}{m} R_n \int_{x_i}^{x_{i+1}} f^{\,2}(R_n/m, x)\, dx$$

$$\text{(where } f\,(a, x) = \inf_{\substack{z \leq x \leq y \\ y-z=a}} \int_z^y f \,/\,|y-z|)$$

$$= \frac{1}{m} R_n \int_{M_n}^{M_n^{\,*}} f^{\,2}(R_n/m, x).$$

Find values $A(\delta)$ and $A^*(\delta)$ such that $\int_{-\infty}^{A(\delta)} f^{\,2} = \frac{\delta}{3} f^{\,2} \int_{A^*(\delta)}^{\infty} f^{\,2} = \frac{\delta}{3} f^{\,2}$, and a value $B(\delta)$ such that

$$\int f^{\,2}(a, x) > (1-\frac{\delta}{3}) \int f^{\,2}, \text{ all } 0 < a \leq B(\delta).$$

Thus, if A is the event $[M_n < A(\delta),\ M_n^{\,*} > A^*(\delta)]$ and B is the event $[R_n/m \leq B(\delta)]$, we have on $A \cap B$, for $a = R_n/m$,

$$\int_{M_n}^{M_n^{\,*}} f^{\,2}(a, x) \geq \int f^{\,2}(a, x) - \int_{M_n^{\,*}}^{\infty} f^{\,2}(a, x) - \int_{\infty}^{M_n} f^{\,2}(a, x)$$

$$\geq (1-\frac{\delta}{3}) \int f^{\,2} - 2\frac{\delta}{3} \int f^{\,2} = (1-\delta) \int f^{\,2}$$

Thus,

$$\sum_{i=1}^{m} p_i^{\,2} \geq I_{A \cap B}\ (1-\delta)\ \frac{1}{m} R_n \int f^{\,2}.$$

We also have

$$\sum_{i=1}^{m} p_i^{\,2} \geq I_{A \cap B^c}\ C(\delta)$$

where

$$C(\delta) = \sup_{\substack{A(\delta) \leq z < y \leq A^*(\delta) \\ y-z=B(\delta)}} (\int_z^y f\,)^2.$$

Note that as $\delta \downarrow 0$, we have $B(\delta) \to 0$ and thus $C(\delta) \to 0$. Combining these bounds gives

$$\sum_{i=1}^{m} p_i^{\,2} \geq I_A\ \min((1-\delta)\frac{1}{m} R_n \int f^{\,2}, C(\delta)) = I_A\ Z(R_n)$$

where $Z(R_n)$ is an increasing function of R_n. By Gurland's inequalities (Gurland, 1968) we have $E(I_A\ Z(R_n)) \geq P(A)\, E(Z(R_n))$. We also know that $P(A) \to 1$ for all $\delta \in (0,1)$. Thus, with a little extra manipulation we obtain the following bound:

$$E(T)/n \geq (1+o(1))(1+E(\min(\frac{1}{c_n} R_n (1-\delta) \int f^{\,2}, nC(\delta))))$$

$$\geq (1+ o(1))(1+(1-\delta)E(\min(\frac{1}{c_n} R_n \int f^{\,2}, nC(\delta)))).$$

This concludes the proof of Theorem 2.1 (i).

Proof of Theorem 2.1 (ii).

From Lemma 2.2, we have

$$E(T\,|M_n, M_n^{\,*})/n \leq \min(n,\ 1+(R_n/(c_n p^2)) \int f^{\,2})$$

$$\leq 1 + \min(n, (R_n/(c_n p^2)) \int f^{\,2}).$$

Let us take expectations on both sides of this inequality. For arbitrary $\epsilon > 0$ we have

$$E(T)/n \leq 1+E(\min(n, (R_n/(c_n p^2)) \int f^{\,2}$$

$$\leq E(\min(n, (R_n(1+\epsilon)/c_n) \int f^{\,2})) + 1 + n\ P(\frac{1}{p^2} > 1+\epsilon)$$

$$\leq E(\min(n, (R_n/c_n) \int f^{\,2})) + \frac{\epsilon}{c_n} \int f^{\,2} + n\ P(p < 1/\sqrt{1+\epsilon}) + 1.$$

The proof is complete if we can show that the last probability is $o(1)$ for every $\epsilon > 0$. Let U_1, U_2 be independent uniform [0,1] random variables, and note that p is distributed as $U_1^{1/n} U_2^{1/(n-1)}$. Thus,

$$P(p < 1/\sqrt{1+\epsilon}) \leq P(U_1^{1/n} < (1+\epsilon)^{-1/4}) + P(U_2^{1/(n-1)} < (1+\epsilon)^{-1/4})$$
$$\leq 2(1+\epsilon)^{-(n-1)/4} ,$$

and we are done.

2.3. A SUPERLINEAR NUMBER OF BUCKETS.

For many infinite-tailed distributions, we know precisely how $E(T)$ varies asymptotically. For example, for densities covered by Theorem 2.3,

$$E(T) \sim n(1 + \frac{1}{c} E(R_n) \int f^2)$$

when $m \sim cn$. We also have in those cases, by the proof of Theorem 2.1 (ii),

$$E(T) \leq n(1 + 2\frac{1}{(1+\epsilon)^{(n-1)/4}} + \frac{\epsilon}{c_n} \int f^2 + \frac{E(R_n)}{c_n} \int f^2) ,$$

for arbitrary $\epsilon > 0$. Here $c_n = m/n$. When we sort, there is an additional cost of the form Am for some constant $A > 0$ due to the time needed to initialize and concatenate the buckets. If $E(R_n) \to \infty$, it is easy to see that in the upper bound,

$$E(T) \leq n \frac{E(R_n)}{c_n} \int f^2 (1+o(1))$$

provided that $E(R_n)/c_n \to \infty$. If we balance the two contributions to the cost of searching with respect to m, then we will find that it is best to let m increase at a faster-than-linear pace. For example, consider the minimization of the cost function

$$Am + \frac{n E(R_n)}{(\frac{m}{n})} \int f^2 .$$

The minimum is attained at

$$m = n \sqrt{\frac{E(R_n)}{A} \int f^2} ,$$

and the minimal value of the cost function is

$$2n \sqrt{AE(R_n) \int f^2} .$$

If we had picked $m \sim cn$, then the main contribution to the sorting time would have come from the selection sort, and it would have increased as a constant times $n E(R_n)$. The balancing act reduces this to about $n \sqrt{E(R_n)}$, albeit at some cost: the space requirements increase at a superlinear rate too. Futhermore, for the balancing to be useful, one has to have a priori information about $E(R_n)$.

Let us consider a few examples. For the normal distribution, we would optimally need

$$m \sim n \sqrt{\frac{1}{A} \sqrt{\frac{2}{\pi}} \log n} \quad ,$$

and obtain

$$Am \sim E(T) \sim n \sqrt{A \sqrt{\frac{2}{\pi}} \log n} .$$

For the exponential distribution, we have

$$m \sim n \sqrt{\frac{1}{2A} \log n} \quad ,$$
$$Am \sim E(T) \sim n \sqrt{\frac{A}{2} \log n} .$$

Similarly, for all distributions with finite $\int |x|^r f(x)dx$, $\int f^2(x)dx$, we can choose m such that

$$Am \sim E(T) \leq C\, n^{1+\frac{1}{2r}}$$

for some constant C.

The idea of a superlinear number of buckets to reduce the expected time can also be used advantageously when $\int f^2 = \infty$ and f has compact support. When preprocessing is allowed, as in the case of searching, and space requirements are no obstacle, we could choose m so large that $E(D_S)$ and $E(D_U)$ are both $O(1)$. To illustrate this point, we use the bound for $E(T)$ used in the proof of Theorem 2.1 (ii), and the fact that

$$D_S = \frac{T}{2n} + \frac{1}{2} .$$

Thus, when $\int f^2 < \infty$, $E(R_n) \to \infty$, we can choose

$$m \sim nE(R_n) \int f^2 ,$$

and conclude that

$$\limsup_{n \to \infty} \frac{E(T)}{n} \leq 2 ,$$

$$\limsup_{n \to \infty} E(D_S) \leq \frac{3}{2} .$$

We stress again that the idea of a superlinear number of buckets seems more useful in problems in which a lot of preprocessing is allowed, such as in ordinary searching and in data base query problems.

Chapter 3

MULTIDIMENSIONAL BUCKETING.

3.1. MAIN THEOREM.

Several algorithms in computer science operate on points in R^d by first storing the points in equal-sized cells, and then traveling from cell to cell, to obtain some solution. Often these algorithms have good expected time behavior when the points are sufficiently smoothly distributed over R^d. This will be illustrated here by exhibiting necessary and sufficient conditions on the distribution of the points for linear expected time behavior.

Our model is as follows: $X_1, \ldots, X_n$ are independent random vectors from R^d with common density f. We let C_n be the smallest closed rectangle covering $X_1, \ldots, X_n$. Each side of C_n is divided into $n' = \lfloor n^{1/d} \rfloor$ equal-length intervals of the type $[a, b)$; the rightmost intervals are of the type $[a, b]$. Let A be the collection of all rectangles **(cells)** generated by taking d-fold products of intervals. Clearly, A has m cells where

$$n \geq m \geq (n^{1/d} - 1)^d \geq n(1 - dn^{-1/d}).$$

The cells will be called $A_1, \ldots, A_m$, and N_i will denote the number of X_j's in cell A_i. Thus, to determine all the cell memberships takes time proportional to n. Within each cell, the data are stored in a linked list for the time being.

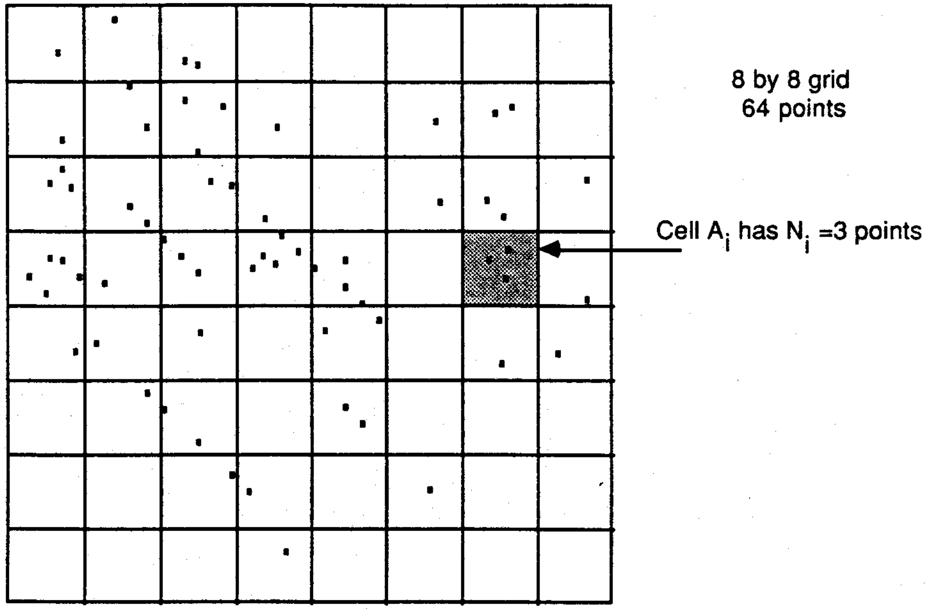


Figure 3.1.

The cell structure has been used with some success in computational geometry (see for example, Shamos (1978), Weide (1978), Bentley, Weide and Yao (1980), and Asano, Edahiro, Imai, Iri and Murota (1985)). Often it suffices to travel to each cell once and to do some work in the i-th cell that takes time $g(N_i)$ for some function g (or at least, is bounded from above by $ag(N_i)$ and from below by $bg(N_i)$ for some appropriate constants a,b: this slightly more general formulation will not be pursued here for the sake of simplicity).

For example, one heuristic for the traveling salesman problem would be as follows: sort the points within each cell according to their y-coordinate, join these points, then join all the cells that have the same x-coordinate, and finally join all the long strips at the ends to obtain a traveling salesman path (see e.g. Christofides (1976) or Papadimitriou and Steiglitz (1976)). It is clear that the work here is $O(n)+\sum_{i=1}^{m} g(N_i)$ for $g(u)=u^2$ or $g(u)=u\ \log(u+1)$ depending upon the type of sorting algorithm that is used. The same serpentine path construction is of use in minimum-weight perfect planar matching heuristics (see e.g. Iri, Murota, and Matsui 1981, 1983).

If we need to find the two closest points among $X_1, \ldots, X_n$ in $[0,1]^d$, it clearly suffices to consider all pairwise distances $d(X_i, X_j)$ for X_i and X_j at most a_d (a constant depending upon d only) cells apart, provided that the grid is constructed by cutting each side of $[0,1]^d$ into $n' = \lfloor n^{1/d} \rfloor$ equal pieces. Using the inequality $(u_1+u_2+...+u_k)^2 \leq 2^{k-1}(u_1{}^2+...+u_k{}^2)$, it is not hard to see that the total work here is bounded from above by $O(n)$ plus a constant times $\sum_{i=1}^{m} N_i{}^2$.

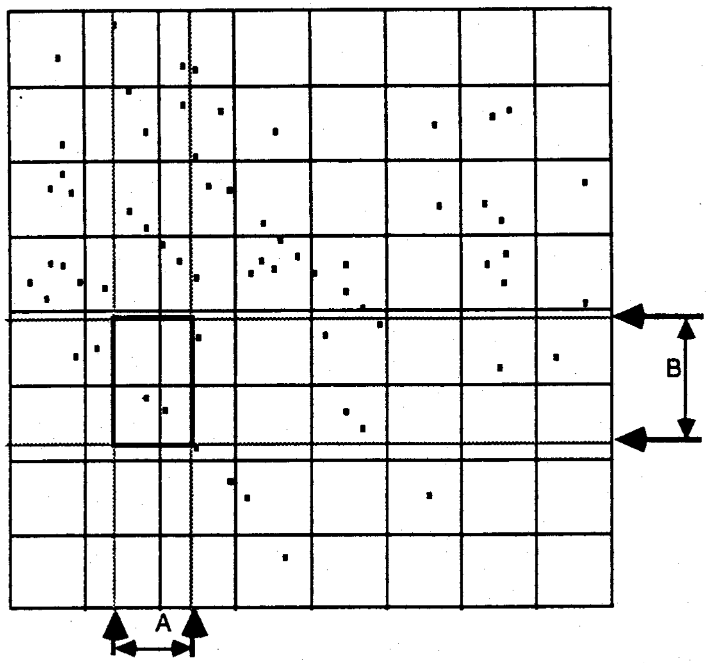


Figure 3.2.
Range search problem: report all points in the intersection of A and B. Grid to be used in solution is also shown.

For multidimensional sorting and searching, we refer to section 3.2. In section 3.2, a few brief remarks about the point-location and point enclosure problems will be included. The point enclosure problem can be considered as a special case of range searching, i.e. the problem of retrieving all points satisfying certain characteristics. If for example we want to retrieve all points for which the coorc nates are between certain threshold values, then we can speak of an orthogon range query. In the survey articles of Bentley and Friedman (1979) and 'Asan Edahiro, Imai, Iri and Murota (1985), some comparisons between cell structur and other structures for the range search problem are made. The range searc problem has one additional parameter, namely the number of points retrieve Query time is usually measured in terms of the number of retrieved points plus function of n. If most queries are large, then it makes sense to consider larg cells. In other words, the cell size should not only depend upon n and f, bu also on the expected size of the query rectangle (see e.g. Bentley, Stanat and Wi liams, 1977). In addition, new distributions must be introduced for the locatic and size of the query rectangle, thus complicating matters even further. Fc these reasons, the range search problem will not be dealt with any further in th collection of notes. The traveling salesman problem is briefly dealt with in sec tion 3.3, and in section 3.4, we will look at some closest point problems in compu tational geometry. The latter problems differ in that the time taken by the algc rithm is no longer a simple sum of an univariate function of cell cardinalities, bu a sum of a multivariate function of cell cardinalities (usually of the cardinality c a central cell and the cardinalities of some neighboring cells). In the entir chapter, we will deal with a work function g. Initially, the time of an algorithr is given by

$$T = \sum_{i=1}^{m} g(N_i)$$

for some function g satisfying:

(i) g is nonnegative and $g(u)/u \uparrow \infty$ as $u \uparrow \infty$.

(ii) $g(u)/u^k \downarrow 0$ as $u \to \infty$ for some finite constant k;

(iii) g is convex on $[0,\infty)$; $g(0) = 0$.

Remark.

We would like to point out that (i) and (ii) imply that g is continuous anc that $g(0)=0$. Examples of functions $g(.)$ satisfying the listed conditions arı $g(u) = u^r$, some $r \geq 1$, and $g(u) = u \log(u+1)$.

Theorem 3.1.

Let f be an arbitrary density on R^d. Then are equivalent:

(i) $\liminf_{n \to \infty} E(T)/n < \infty$;

(ii) $\limsup_{n \to \infty} E(T)/n < \infty$;

(iii) $\int g(f(x))\, dx < \infty$.

Proof of Theorem 3.1.

The proof is in three parts:

A. f compact support, $\int g(f) = \infty \Rightarrow \liminf_{n \to \infty} E(T)/n = \infty$.

B. f compact support, $\int g(f) < \infty \Rightarrow \liminf_{n \to \infty} E(T)/n < \infty$.

C. f does not have compact support, $\liminf_{n \to \infty} E(T)/n = \infty$.

In the proof, we will use the symbols $p_i = \int_{A_i} f$, $C = \bigcup_{i=1}^{m} A_i$, $p = \int_C f$. The following fact will be needed a few times: given C,

$$Y_i < N_i < W_i + 2d \ , 1 \leq i \leq m \ , n > 2d \ ,$$

where Y_i is a binomial $(n-2d, p_i)$ random variable, W_i is a binomial $(n, p_i/p)$ random variable, and "$<$" denotes "is stochastically smaller than", i.e.

$$P(Y_i \geq x) \leq P(N_i \geq x) \leq P(W_i + 2d \geq x) \text{, all } x.$$

Proof of A.

Let C_0 be the smallest closed rectangle covering the support of f , and lei $f_n(x)$ be the function defined by the relations: $f_n(x) = 0,\ x \notin C$ $f_n(x) = (n-2d)p_i,\ x \in A_i$. We have

$$E(T) = \sum_{i=1}^{m} E(g(N_i)) = \sum_{i=1}^{m} E(E(g(N_i)|C))$$

$$\geq \sum_{i=1}^{m} E(E(g(Y_i)|C))$$

$$\geq \sum_{i=1}^{m} E(\frac{1}{2}g((n-2d)p_i - \sqrt{(n-2d)p_i}))$$

(by Lemma 5.4, if we agree to let $g(u)=0$ for $u \leq 0$)

$$= E(\int \frac{m}{2\lambda(C)}\ g(f_n - \sqrt{f_n})).\ (\lambda \text{ denotes Lebesgue measure})$$

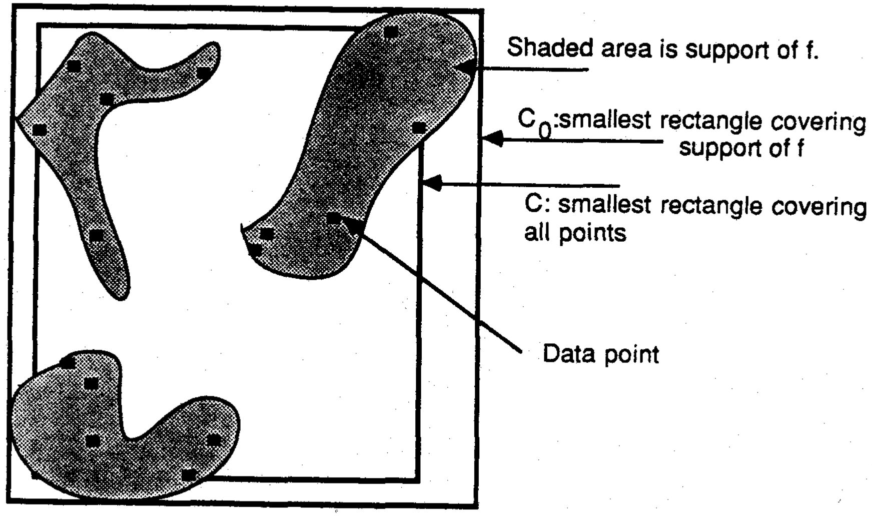

Shaded area is support of f.
C0:smallest rectangle covering support of f
C: smallest rectangle covering all points
Data point

Figure 3.3.

Thus, by Fatou's lemma,

$$\liminf_{n\to\infty} E(T)/n \geq E(\int \liminf_{n\to\infty} (\frac{1}{2\lambda(C)} g(f_n - \sqrt{f_n})))$$

where the inner limit infimum is with respect to a.e. convergence. Now, for almost all $\omega \in \Omega$ (where $(\Omega, \mathbf{F}, \mathbf{P})$ is our probability space with probability element ω), we have $C \to C_0$ and thus $\lambda(C) \to \lambda(C_0)$. But then, by Lemma 5.11, for almost all $(x,\omega) \in R^d \times \Omega$, we have $f_n(x) \to f(x)$. Thus, the Fatou lower bound given above is

$$\int (2\lambda(C_0))^{-1} g(f - \sqrt{f}\ )$$

$$\geq \int_{f \geq 4} (2\lambda(C_0))^{-1} g(f/2) \geq \int_{f \geq 4} (2\lambda(C_0))^{-1} g(f)/2^k = \infty$$

when $\int g(f) = \infty$ (for $\int_{f \leq 4} g(f) \leq g(4)\lambda(C_0) < \infty$).

Proof of B.

$$E(T) \leq \sum_{i=1}^{m} E(E(g(W_i + 2d|C)) \leq \sum_{i=1}^{m} E(E(g(2W_i)|C) + g(4d))$$

$$\leq mg(4d) + 2^k \sum_{i=1}^{m} E(E(g(W_i)|C))$$

$$\leq mg(4d) + 2^k \sum_{i=1}^{m} (aE(g(np_i/p)) + ag(1))$$

where a is the constant of Lemma 5.4 (and depends upon k only). Thus, to show that $E(T) = O(n)$, we need only show that $\sum_{i=1}^{m} E(g(np_i/p)) = O(n)$. Now,

$$E(g(np_i/p)) \leq E(g(2np_i)) + g(n)P(p < 1/2).$$

The last term is uniformly bounded in n as we will now prove. First, we have $g(n) \leq n^k g(1)$. We will show that $P(p < 1/2) \leq 2d \exp(-n/(4d))$ for all n. Because the function $u^k e^{-u}$, $u > 0$, is uniformly bounded, we see that $\sup_n g(n) P(p < 1/2) < \infty$. Indeed,

$$[p < 1/2] \subseteq \bigcup_{j=1}^{d} [p_j' < 1 - 1/(2d)]$$

where p_j' is the integral of f over all x's whose j-th component lies between the minimal and maximal j-th components of all the X_i's. But by the probability integral transform, when $U_1, \ldots, U_n$ are independent uniform [0,1] random variables,

$$P(p_j' < 1-1/(2d)) \leq 2P(\min(U_1, \ldots, U_n > 1/(4d)) = 2(1-1/(4d))^n$$

$$\leq 2 \exp(-n/(4d)) .$$

Finally, by Jensen's inequality,

$$\sum_{i=1}^{m} E(g(2np_i)) = \sum_{i=1}^{m} E(g(2n\lambda(A_i) \int_{A_i} f / \lambda(A_i)))$$

$$\leq \sum_{i=1}^{m} E(\int_{A_i} g(2n\lambda(A_i) f) / \lambda(A_i))$$

$$\leq E(\frac{m}{\lambda(C)} \int_C g(f) \max(2n\lambda(C)/m, (2n\lambda(C)/m)^k))$$

$$\leq m \int g(f) \max(2\frac{n}{m}, (2\frac{n}{m})^k \lambda(C_0)^{k-1})$$

and B follows since $m \sim n$.

Proof of C.

By a bound derived in the proof of A and by the second inequality of Lemma 5.4, we need only show that

$$\frac{1}{m} \sum_{i=1}^{m} E(g(\lfloor(n-2d)p_i\rfloor) = \infty .$$

when f does not have compact support. By our assumptions on g, $(n-2d)$ can be replaced by n. We may assume without loss of generality that the first component of X_1 has unbounded support. Let $(a_1, b_1), \ldots, (a_d, b_d)$ be ϵ and $1-\epsilon$ quantiles of all the marginal distributions where $\epsilon \in (0,1/2)$ is chosen such that $B = \times_{j=1}^{d} (a_j, b_j)$ satisfies $\int_B f = \frac{1}{2}$. Let Q be the collection of A_j's intersecting with B, and let q be the cardinality of Q. Set $p_j' = \int_{A_j \cap B} f$, and let Z be the indicator of the event $B \subseteq C$. Clearly,

$$\frac{1}{m} \sum_{i=1}^{m} E(g\lfloor np_i \rfloor) \geq \frac{1}{m} E(\sum_{A_j \in Q} g\lfloor np_i' \rfloor)$$

$$\geq E(\frac{q}{m} \frac{Z}{q} \sum_{A_j \in Q} g\lfloor np_i' \rfloor) \geq E(\frac{q}{m} Z g(\frac{1}{q} \sum_{A_j \in Q} \lfloor np_i' \rfloor))$$

$$\geq E(\frac{q}{m} Z \; g(\frac{n}{2q} - 1)) \geq E(Z \; (\frac{1}{2} - \frac{q}{n}) \; g(\frac{n}{2q} - 1)/(\frac{n}{2q} - 1)).$$

where we used Jensen's inequality. Since $g(u)/u \uparrow \infty$, we need only show that for any constant M, however large,

$$\liminf_{n \to \infty} P(Z=1, n/2q - 1 \geq M) > 0.$$

Now, let U, V be the minimum and the maximum of the first components of $X_1, \ldots, X_n$. When $Z = 1$, we have

$$q \leq m^{\frac{d-1}{d}} \left(\frac{b_1 - a_1}{\frac{V-U}{m^{1/d}}} + 2 \right),$$

and thus

$$P(Z=1, n \geq 2q(M+1))$$

$$\geq P(Z=1, (b_1-a_1)m/(V-U)+2m^{\frac{d-1}{d}} \leq n/(2(M+1)))$$

$$\geq 1-P(Z=0) - P((b_1-a_1)m/(V-U) \geq \frac{n}{4(M+1)})$$

$$-P(2m^{\frac{d-1}{d}} \geq \frac{n}{4(M+1)}).$$

The second term of the last expression is $o(1)$ for obvious reasons. The third term is $o(1)$ since $m \sim n$ and $V-U \rightarrow \infty$ in probability as $n \rightarrow \infty$. The last term is $o(1)$ since $m \sim n$. This concludes the proof of C.

3.2. SORTING AND SEARCHING.

When $d=1$, and elements within each bucket A_i are sorted by an n^2 sorting algorithm (such as selection sort, or insertion sort), Theorem 3.1 applies with $g(u)=u^2$. The data can be sorted in expected time $O(n)$ if and only if f has compact support and

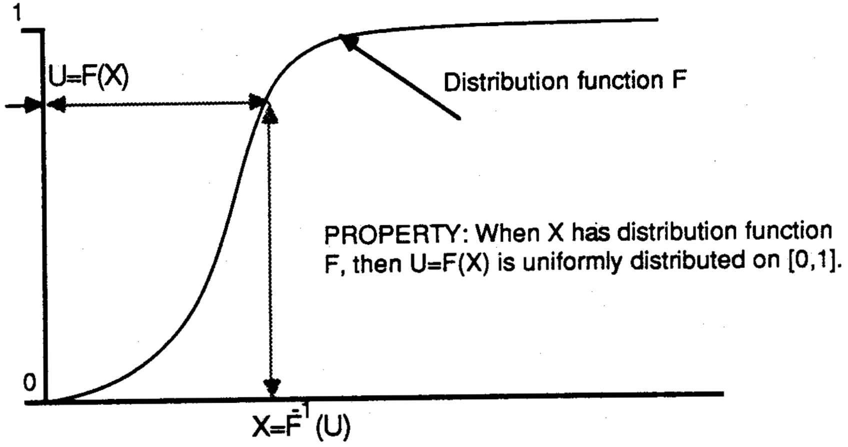
$$\int f^2 < \infty.$$

If however we employ an expected time $n \log n$ sorting algorithm based upon comparisons only (such as heapsort, quicksort or tree insertion sort), the data can be sorted in expected time $O(n)$ if and only if f has compact support and

$$\int f \log^+ f < \infty .$$

The latter condition is only violated for all but the most peaked densities. These results generalize those of Devroye and Klincsek (1981). We should mention here that if we first transform arbitrary data by a mapping $h : R^1 \rightarrow [0,1]$ that is continuous and monotone, construct buckets on [0,1], and then carry out a subsequent sort within each bucket as described above, then often $E(T) = O(n)$: in other words, with little extra effort, we gain a lot in expected time. The ideal

transformation h uniformizes, i.e. we should try to use $F(x)$ where F is the distribution function of the data. In general, we can take h in such a way that it is equal to $F(\frac{x-\mu}{\sigma})$ where F is a fixed distribution function, μ is a sample estimate of location (mean, median, etc.) and σ is a sample estimate of scale (standard deviation, etc.). This should in many cases give satisfactory results. It is probably advantageous to take robust estimates of location and scale, i.e. estimates that are based upon the sample quantiles. Meijer and Akl (1980) and Weide (1978) give variations of a similar idea. For example, in the former reference, F is piecewise linear with cut-points at the extrema and a few sample quantiles. One should of course investigate if the theoretical results remain valid for transformations F that are data-dependent.



PROPERTY: When X has distribution function F, then U=F(X) is uniformly distributed on [0,1].

Figure 3.4.

The conditions on f mentioned above are satisfied for all bounded densities f. It is nice exercise to verify that if a transformation

$$h(x) = x/(1+|x|)$$

is used and $f(x) \leq a \exp(-b|x|^c)$ for some $a, b, c > 0$, then the density of the transformed density remains bounded. Thus, for the large class of densities with exponentially dominated tail, we can sort the transformed data in average time $O(n)$ by any of the bucket-based methods discussed above.

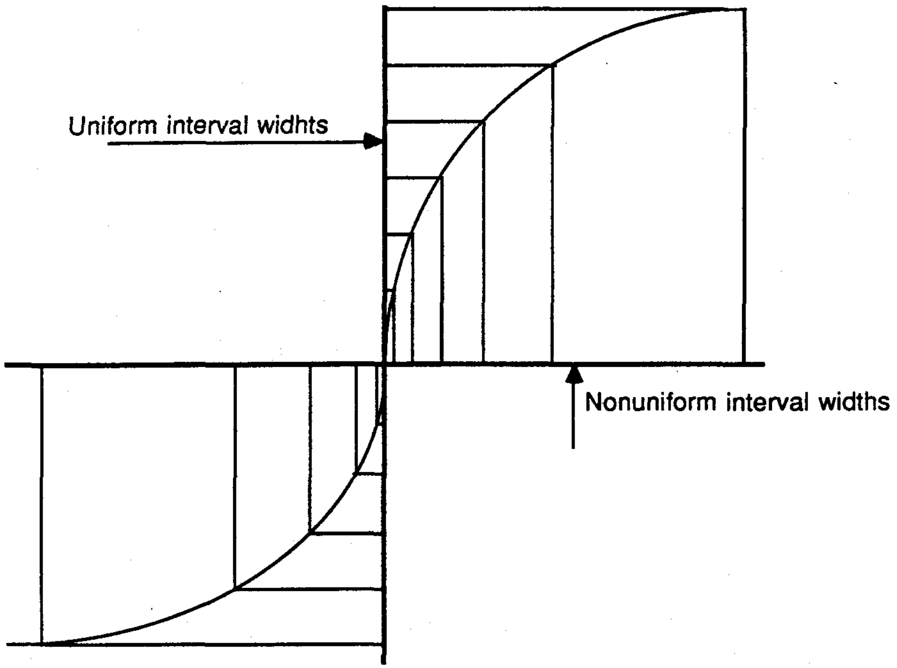


Figure 3.5.
A nonlinear transformation useful for distribution with unbounded support.

For the expected number of comparisons in a successful or unsuccessful search of linked list based buckets, we obtain without effort from Theorem 3.1 the value $O(1)$ (even when $d \neq 1$) when f has compact support and $\int f^2 < \infty$. These conditions are necessary too. If within each bucket the $X_i' s$ are ordered according to their first component, and are stored in a binary search tree or a balanced binary tree such as a 2-3 tree, condition $\int f^2 < \infty$ can be replaced by $\int f \log^+ f < \infty$. Just apply the Theorem with $g(u)=u \log(u+1)$, and note that $\int f \log(f+1) < \infty$ is equivalent to $\int f \log^+ f < \infty$ because $\log^+ u \le \log(1+u) \le \log^+ u + \log 2$. For a more detailed analysis, the quantity $T = \sum_{i=1}^{m} N_i^2$ of chapter 1 must be replaced now by $T = \sum_{i=1}^{m} N_i \log(N_i+1)$. Most of chapters 1 and 2 can be repeated for this new quantity. We leave it as

an exercise to show that

$$E(T) \le a + bn + n\, E(\log(1+\min(n, R_n)))$$

$$\le a + bn + n \log(1 + E(R_n))$$

for some constants $a, b > 0$ when $\int f \log(f+1) < \infty$. Hence, if f is any density with a finite moment generating function in a small neighborhood of the origin, we obtain $E(T) = O(n \log\log n)$. Examples of such densities are the exponential and normal densities. This extends an interesting observation reported in Akl and Meijer (1982).

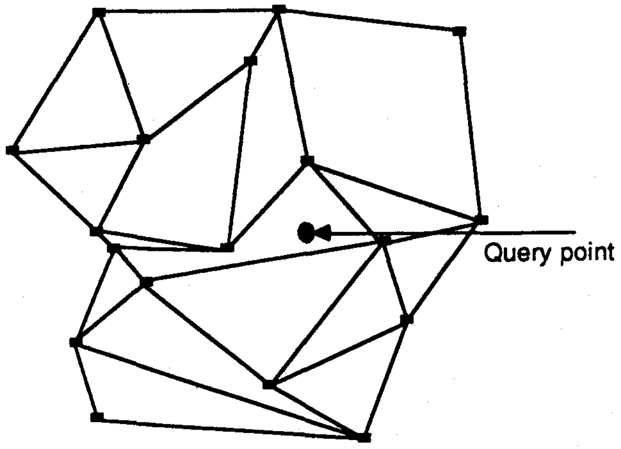


Figure 3.6.
The planar graph point location problem:
return the set in the partition to which the query point belongs.

Remark. [Point location problems.]

In the planar point-location problem, a straight-line planar graph with n vertices is given, and one is asked to find the set in the partition of the plane to which a query point x belongs. In many applications, a large number of queries are raised for one fixed planar partition. We won't be concerned here with worst-case complexities. It suffices to mention that each query can be answered in worst-case time $O(\log(n))$ provided that up to $O(n \log(n))$ time is spent in setting up an appropriate data structure (Lipton and Tarjan, 1977 ; Kirkpatrick, 1983). See also Lee and Preparata (1977) for an algorithm with $O((\log(n))^2)$ worst-case search time, and Shamos and Bentley (1977) for the point-location problem when the space is partitioned into nonoverlapping rectangles. It was pointed out in Asano, Edahiro, Imai, Iri and Murota (1985) that these algorithms can be very slow in practice. In particular, they compare infavorably with a bucket-based algorithm of Edahiro, Kokubo and Asano (1983).

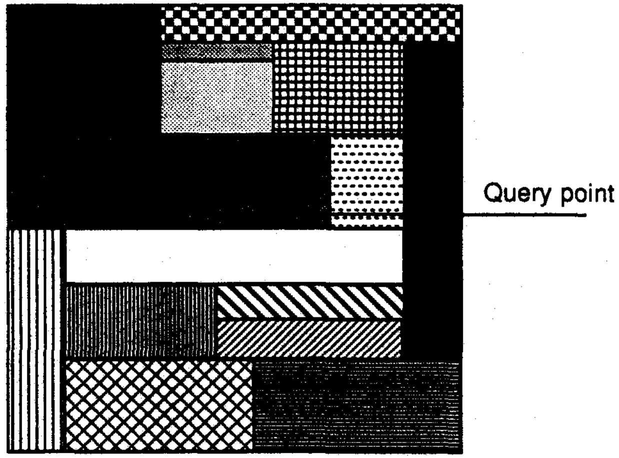


Figure 3.7.
The rectangular point location problem.

Assume for example the following probabilistic model : the n points $X_1, \ldots, X_n$ and the query point are iid random vectors uniformly distributed in the unit square, and the graph is constructed by connecting points in an as yet unspecified manner. In first instance, we will be interested in the expected worst-case time, where "worst-case" is with respect to all possible planar graphs given the data. Let us construct an m-grid where for each bucket the following information is stored : the list of vertices sorted by y-coordinates, the collections of edges intersecting the north, south east and west boundaries (sorted), and the region of the partition containing the north-west corner vertex of the bucket. This assumes that all regions are numbered beforehand, and that we are to return a region number. Partition each bucket in a number of horizontal slabs, where the slab boundaries are defined by the locations of the vertices and the points where the edges cut the east and west boundaries. For each slab, set up a linked list of conditions and region numbers, corresponding to the regions visited when the slab is traversed from left to right. (Note that no two edges cross in our graph.) It is important to recall that the number of edges in a planar graph is $O(n)$, and that the number of regions in the partition is thus also $O(n)$. One can verify that the data structure described above can be set up in worst case time $O(n^{3/2})$ when $m \sim cn$ for some constant c . The expected set-up time is $O(n)$ in many cases. This statement uses techniques similar to those needed to analyze the expected search time. We are of course mainly interested in the expected search time. It should come as no surprise that the expected search time decreases with increasing values of m . If m increases linearly in n , the expected search time is $O(1)$ for many distributions. Those are the cases of interest to us. If m increases faster than n , the expected search time ,while still $O(1)$, has a smaller constant. Unfortunately, the space requirements become inacceptable because $\Omega(\max(m,n))$ space is needed for the given data structure. On the positive side, note that the space requirements are $O(n)$ when m increases at most as $O(n)$.

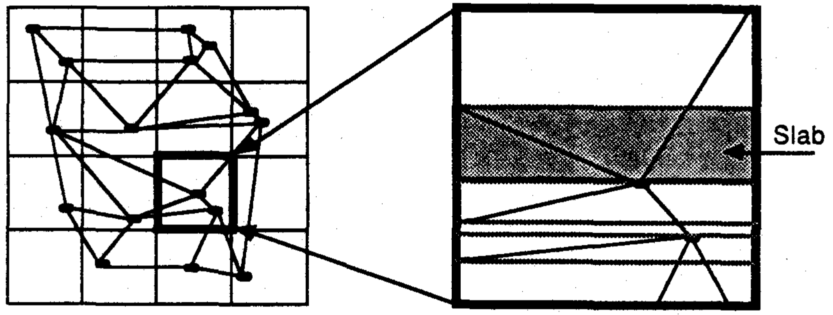


Figure 3.8.

The slab method described above is due to Dobkin and Lipton (1976), and differs slightly from the method described in Edahiro, Kokubo and Asano (1983). The time taken to find the region number for a query point X in a given bucket is bounded by the number of slabs. To see this, note that we need to find the slab first, and then travel through the slab from left to right. Thus, the expected time is bounded by $\sum_{i=1}^{m} p_i S_i$, where S_i denotes the number of slabs in the i-th bucket, p_i is the probability that X belongs to the i-th bucket, and the expected time is with respect to the distribution of X, but is conditional on the data. But $E(S_i) \le np_i + E(C_i)$, where C_i is the number of edges crossing the boundary of the i-th bucket. Without further assumptions about the distribution of the data points and the edges, any further analysis seems difficult, because $E(C_i)$ is not necessarily a quantity with properties determined by the behavior of f in or near the i-th bucket. Assume next that X is uniformly distributed. Then, the expected time is bounded by

$$\sum_{i=1}^{m} \frac{1}{m}(np_i + E(C_i))$$

$$= \frac{n}{m} + \frac{E(C)}{m}$$

where $E(C)$ is the expected value of the overall number of edge-bucket boundary crossings. $E(C)$ can grow much faster than m : just consider a uniform density on $[0,1]^2$. Sort the points from left to right, and connect consecutive points by edges. This yields about n edges of expected length close to $1/3$ each. $E(C)$ should be close to a constant times $n\sqrt{m}$. Also, for any planar graph, $C \le \gamma n \sqrt{m}$ where γ is a universal constant. Thus, it is not hard to check that the conditional expected search time is in the worst-case bounded by

$$\frac{n}{m} + \gamma \frac{n}{\sqrt{m}} .$$

This is $O(1)$ when m increases as $\Omega(n^2)$. Often, we cannot afford this because of space or set-up time limitations. Nevertheless, it is true that even if m increases linearly with n , then the expected search time is $O(1)$ for certain probabilistic models for putting in the edges. Help can be obtained if we observe that an edge of length L cuts at most $2(2+L\sqrt{m})$ buckets, and thus leads to at most twice that number of edge-boundary crossings. Thus, the expected time is bounded by

$$\frac{n}{m} + \frac{1}{m} \sum_{j=1}^{e} 4(2+E(L_j)\sqrt{m})$$

where e is the total number of edges and L_j is the length of the j-th edge. Since $e = O(n)$, and $m \sim cn$ (by assumption), this gives $O(1)$ provided that

$$\sum_{j=1}^{e} E(L_j) = O(\sqrt{m}) .$$

In other words, we have obtained a condition which depends upon the expected lengths of the edges only. For example, the condition is satisfied if the data points have an arbitrary density f on $[0,1]^2$, and each point is connected to its nearest neighbor : this is because the expected lengths of the edges grow roughly as $1/\sqrt{n}$. The condition is also satisfied if the points are all connected to points that are close to it in the ordinary sense, such as for example in a road map.

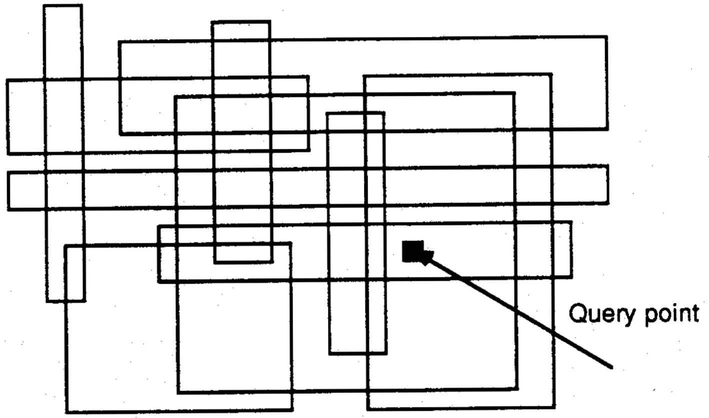


Figure 3.9.
The point enclosure problem:report all rectangles to which query point belongs.

Remark. [Point enclosure problems.]

In point-enclosure problems, one is given n rectangles in R^d. For one query point X, one is then asked to report all the rectangles to which X belongs. Since a rectangle can be considered as a point in R^{2d}, it is clear that this problem is equivalent to an orthogonal range search query in R^{2d}. Thus, orthogonal range search algorithms can be used to solve this problem. There have been several direct attempts at solving the problem too, based mainly on the segment or interval tree (Bentley (1977), Bentley and Wood (1980), Vaishnavi and Wood (1980), Vaishnavi (1982)). For example, on the real line, the algorithm of Bentley and Wood (1980) takes preprocessing time $O(n\log(n))$, space $O(n\log(n))$, and worst-case query time $O(\log(n)+k)$ where k is the number of segments (i.e., one-dimensional rectangles) reported. We will briefly look into the properties of the bucket structure for the one-dimensional point-enclosure problem.

First, we need a good probabilistic model. To this end, assume that (L, R), the endpoints of a segment form a random vector with a density f on the north-west triangle of $[0,1]^2$ (this is because $L \leq R$ in all cases). The n intervals are iid, and the query point has a density g on $[0,1]$. The segment $[0,1]$ is partitioned into m buckets, where typically $m \sim cn$ for some constant c (which we assume from here onwards). For each bucket, keep two linked lists : one linked list of segments completely covering the bucket, and one of intervals only partially covering the bucket. Note that the entire structure can be set up in time proportional to n plus n times the total length of the segments (because a segment of length l can be found in at least 1 and about nl linked lists). The space requirements are formally similar. Under the probabilistic model considered here, it is easy to see that the expected space and expected preprocessing time are both proportional to n times the expected value of the total length. Since the expected value of the total length is n times the expected value of the length of the first segment, and since this is a constant, the expected space and preprocessing requirements increase quadratically in n. The expected search time is small. Indeed, we first report all segments of the first linked list in the bucket of X. Then, we traverse the second linked list, and report those segments that contain X. Thus, the search time is equal to $k+1$ plus the cardinality of the second linked list, i.e. the number of endpoints in the bucket. With the standard notation for buckets and bucket probabilities, we observe that the latter contribution to the expected search time is

$$\sum_{i=1}^{m} P(X \in A_i)\, n\, (P(L \in A_i) + P(R \in A_i)) .$$

In particular, if X is uniformly distributed, then this expression is simply $2n/m$. This can be made as small as desired by the appropriate choice of m. If, however, X is with equal probability distributed as L and R respectively, which seems to be a more realistic model, then the expression is

$$\sum_{i=1}^{m} 2np_i^{\,2} \leq \frac{2n}{m} \int h^2$$

where h is the density of X (i.e. it is the average of the densities of L and R), and $p_i = \int_{A_i} h$. Here we used Lemma 1.1.

There are other probabilistic models with totally different results. For example, in the car parking model, we assume that the midpoints of the segments have density f on $[0,1]$, and that the lengths of the segments are random and independent of the location of the segment : the distribution of the lengths however is allowed to vary with n to allow for the fact that as more segments are available, the segments are more likely to be smaller. For example, if the lengths are all the same and equal to r_n where r_n tends to 0 at the rate $1/n$, the overlap among intervals is quite small. In fact, the preprocessing and set-up times are both $O(n)$ in the worst case. If X has density f as well, then the expected search time is $O(1)$ when $\int f^2 < \infty$.

3.3. THE TRAVELING SALESMAN PROBLEM.

The traveling salesman problem is perhaps the most celebrated of all discrete optimization problems. A traveling salesman tour of $X_1, \ldots, X_n$ is a permutation $\sigma_1, \ldots, \sigma_n$ of $1, \ldots, n$: this permutation formally represents the path formed by the edges $(X_{\sigma_1}, X_{\sigma_2}), (X_{\sigma_2}, X_{\sigma_3}), \ldots, (X_{\sigma_n}, X_{\sigma_1})$. The cost of a traveling salesman tour is the sum of the lengths of the edges. The traveling salesman problem is to find a minimum cost tour. When the lengths of the edges are the Euclidean distances between the endpoints, the problem is also called the Euclidean traveling salesman problem, or ETSP.

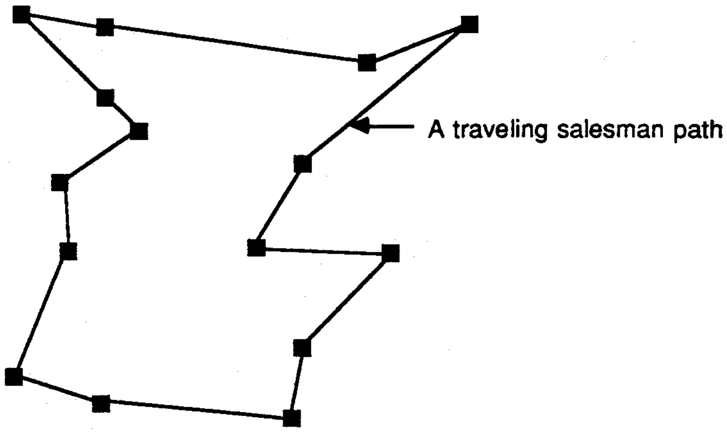


Figure 3.10.
The Euclidean traveling salesman problem:
find the shortest path through all cities.

The ETSP is an NP-hard problem (Papadimitriou (1977), Papadimitriou and Steiglitz (1982)), and there has been considerable interest in developing fast heuristic algorithms (see Papadimitriou and Steiglitz (1982) and Parker and Rardin (1983) for surveys). It should be stressed that these algorithms are nonexact. Nevertheless, they can lead to excellent tours: for example, a heuristic based upon the minimal spanning tree for $X_1, \ldots, X_n$ developed by Christofides (1976) yields a tour which is at worst 3/2 times the length of the optimal tour. Other heuristics can be found in Karp (1977) (with additional analysis in Steele (1981)) and Supowit, Reingold and Plaisted (1983). We are not concerned here with the costs of these heuristic tours as compared, for example, to the cost of the optimal tours, but rather with the time needed to construct the tours. For iid points in $[0,1]^2$, the expected value of the cost of the optimal tour is asymptotic to $\beta\sqrt{n} \int \sqrt{f}$ where $\beta > 0$ is a universal constant (Steele, 1981). For the uniform distribution, this result goes back to Beardwood, Halton and Hammersley (1959), where it is shown that $0.61 \le \beta \le 0.92$.

For the ETSP in $[0,1]^2$, we can capture many bucket-based heuristics in the following general form. Partition $[0,1]^2$ into m equal cubes of side $1/\sqrt{m}$ each. Typically, m increases in proportion to n for simple heuristics, and $m = O(n)$ when the expected cost of the heuristic tour is to be optimal in some sense (see Karp (1977)and Supowit, Reingold and Plaisted (1983)). The bucket data structure is set up (in time $O(n+m)$). The cells are traversed in serpentine fashion, starting with the leftmost column, the second column, etcetera, without ever lifting the pen or skipping cells. The points within the buckets are all connected by a tour which is of one of three possible types:

A. **Random tour.** The points connected as they are stored in the linked lists.

B. **Sorted tour.** All points are sorted according to y coordinates, and then linked up.

C. **Optimal tour.** The optimal Euclidean traveling salesman tour is found.

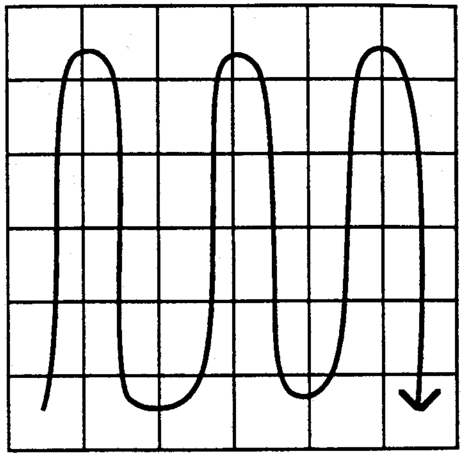


Figure 3.11.
Serpentine cell traversal.

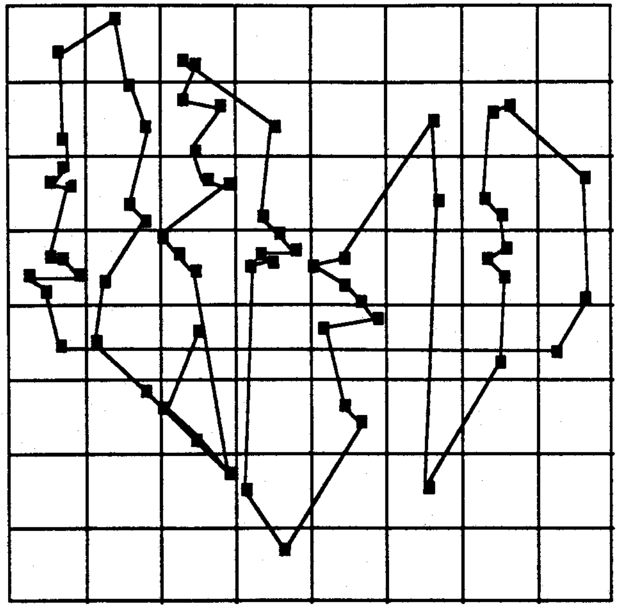


Figure 3.12.
A sorted tour.

The time costs of A, B, C for a bucket with N points are bounded respectivel by

$$CN,$$

$$CN \log(N+1),$$

and

$$CN\ 2^N$$

for constants C. For the optimal tour, a dynamic programming algorithm is use (Bellman, 1962). The m tours are then linked up by traversing the cells in serpentine order. We are not concerned here with just how the individual tours ar linked up. It should for example be obvious that two sorted tours are linked u by connecting the northernmost point of one tour with the southernmost point of the adjacent tour, except when an east-west connection is made at the U-turns in the serpentine. It is easy to see that the total cost of the between-cell connections is $O(\sqrt{m})$, and that the total cost of the tours is $O(n/\sqrt{m})$ for all three schemes. For schemes A and B therefore, it seems important to make m proportional to n so that the total cost is $O(\sqrt{n})$, just as for the optimal tour. In scheme C, as pointed out in Karp (1977) and Supowit, Reingold and Plaisted (1983), if m increases at a rate that is slightly sublinear ($o(n)$), then we can come very close to the globally optimal tour cost because within the buckets small optimal tours are constructed. The expected time taken by the algorithm is bounded by

$$O(n+m) + E(\sum_{i=1}^{m} CN_i),$$

$$O(n+m) + E(\sum_{i=1}^{m} CN_i \log(N_i+1)),$$

and

$$O(n+m) + E(\sum_{i=1}^{m} CN_i 2^{N_i})$$

respectively.

Theorem 3.2.

For the methods A, B, C for constructing traveling salesman tours, the expected time required is bounded by $O(n+m)$ plus, respectively

(A) Cn ;

(B) $Cn \int f \log(2+\frac{n}{m}f) \le Cn \int f \log(2+f) + Cn \log(1+\frac{n}{m})$;

(C) $C2n \int f\ e^{\frac{n}{m}f} \le 2Cn\ \psi(1+\frac{n}{m})$

where $\psi(u)$ is the functional generating function for the density f on $[0,1]^2$.

Remark.

The **functional generating function** for a density f on $[0,1]^2$ is defined by

$$\psi(u) = \int e^{uf(x)}dx \, , \; u \in R \, .$$

By Taylor's series expansion, it is seen that

$$\psi(u) = 1 + u\int f + \frac{u^2}{2!}\int f^2 + \frac{u^3}{3!}\int f^3 + \cdots \, ,$$

which explains the name. Note that the Taylor series is not necessarily convergent, and that ψ is not necessarily finite: it is finite for all bounded densities with compact support, and for a few unbounded densities with compact support. For example, if $f \le f^*$ on $[0,1]^2$, then $\psi(u) \le \frac{1}{f^*}e^{uf^*}$, $u > 0$. Thus, the bound in (C) becomes

$$2Cn\frac{1}{f^*}e^{(1+\frac{n}{m})f^*} \, .$$

(In fact, by a direct argument, we can obtain the better bound $2Cne^{\frac{n}{m}f^*}$.) Note that in the paper of Supowit et al. (1983), m is allowed to be picked arbitrarily close to n (e.g. $m = n/\log\log\log n$). As a result, the algorithm based on (C) has nearly linear expected time. Supowit et al. (1983) provide a further modification of algorithm (C) which guarantees that the algorithm runs in nearly linear time in the worst case.

Proof of Theorem 3.2.

To show (B), we consider

$$E(N_i \log(N_i+1))$$

$$= E(\sum_{j=1}^{n} B_j \log(\sum_{j=1}^{n} B_j + 1))$$

$$= n\, E(B_1 \log(B_1 + \sum_{j=2}^{n} B_j + 1))$$

$$= np_i\, E(\log(2 + \sum_{j=2}^{n} B_j))$$

$$\le np_i \log(2 + (n-1)p_i) \text{ (Jensen's inequality).}$$

where $B_1, \ldots, B_n$ are iid Bernoulli (p_i) random variables. Also, since $p_i \log(2+(n-1)p_i)$ is a convex function of p_i, another application of Jensen's inequality yields the upper bound

$$n \int_{A_i} f \log(2 + \frac{n-1}{m} f) \, ,$$

which is all that is needed to prove the statement for (B). For (C), we argue similarly, and note that

$$E(N_i 2^{N_i})$$

$$= E((\sum_{j=1}^{n} B_j)(\prod_{j=1}^{n} 2^{B_j}))$$

$$= nE(B_1 2^{B_1} \prod_{j=2}^{n} 2^{B_j})$$

$$= 2np_i(2p_i + (1-p_i))^{n-1}$$

$$= 2np_i(1 + p_i)^{n-1}$$

$$\le 2np_i e^{(n-1)p_i}$$

$$\le 2n \int_{A_i} f \, e^{\frac{n-1}{m}f} \quad \text{(Jensen's inequality).}$$

This concludes the proof of Theorem 3.2.

Remark. [ETSP in higher dimensions.]

Halton and Terada (1982) describe a heuristic for the ETSP in d dimensions which is similar to the heuristic given above in which within each cell an optimal tour is found. In particular, for points uniformly distributed on the unit hypercube, it is shown that the tour length divided by the optimal tour length tends with probability one to one as $n \to \infty$. Also, the time taken by the algorithm is in probability equal to $o(n\,\phi(n))$ where ϕ is an arbitrary diverging function picked beforehand and ϕ is used to determine at which rate m/n tends to 0. The divergence of ϕ is again needed to insure asymptotic optimality of the tour's length. The only technical problem in d dimensions is related to the connection of cells to form a traveling salesman tour.

3.4. CLOSEST POINT PROBLEMS.

Local algorithms are algorithms which perform operations on points in given buckets and in neighboring buckets to construct a solution. Among these, we have algorithms for the following problems:

(i) the **close pairs problem:** identify all pairs of points within distance r of each other;

(ii) the **isolated points problem:** identify all points at least distance r away of all other points;

(iii) the **Euclidean minimal spanning tree problem;**

(iv) the **all-nearest-neighbor problem:** for each point, find its nearest neighbor;

(v) the **closest pair problem:** find the minimum distance between any two points.

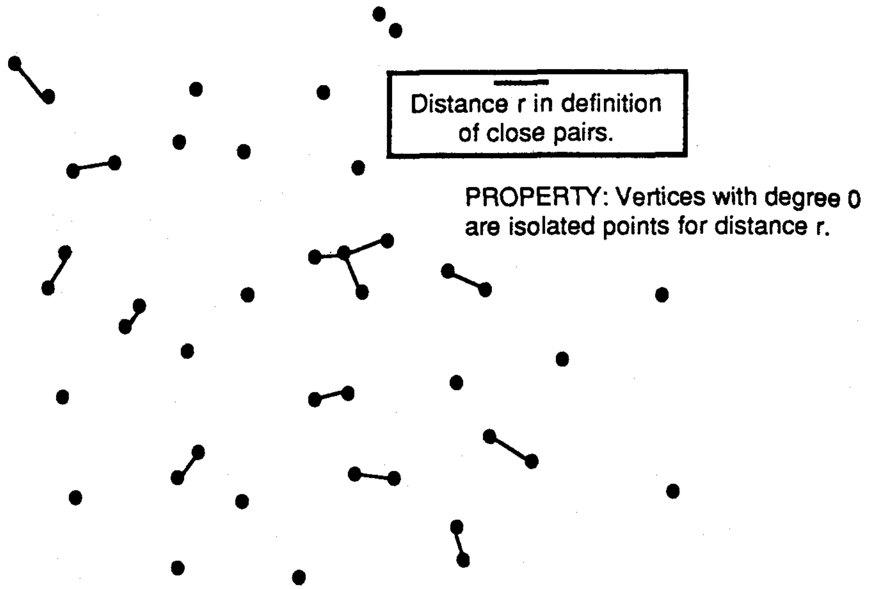

Distance r in definition of close pairs.
PROPERTY: Vertices with degree 0 are isolated points for distance r.

Figure 3.13.
Close pairs graph.

These problems are sometimes called **closest point problems** (Shamos and Hoey, 1975; Bentley, Weide and Yao, 1980). What complicates matters here is the fact that the time needed to find a solution is not merely a function of the form

$$\sum_{i=1}^{m} g(N_i)$$

as in the case of one-dimensional sorting. Usually, the time needed to solve these problems is of the form

$$\sum_{i=1}^{m} g(N_i, N_i^*)$$

where N_i^* is the number of points in the neighboring buckets; the definition of a neighbor bucket depends upon the problem of course. It is quite impossible to give a detailed analysis that would cover most interesting closest point problems. As our prototype problems, we will pick (i) and (ii). Our goal is not just to find upper bounds for the expected time that are of the correct order but possibly of the wrong constant: these can be obtained by first bounding the time by a function of the form

$$\sum_{i=1}^{m} \tilde{g}(N_i + N_i^{*})$$

where $\tilde{g}$ is another function. The overlap between buckets implicit in the terms $N_i + N_i^{*}$ does not matter because the expected value of a sum is the sum of expected values. Our goal here is to obtain the correct asymptotic order **and** constant. Throughout this section too, $X_1, \ldots, X_n$ are independent random vectors with density f on $[0,1]^d$.

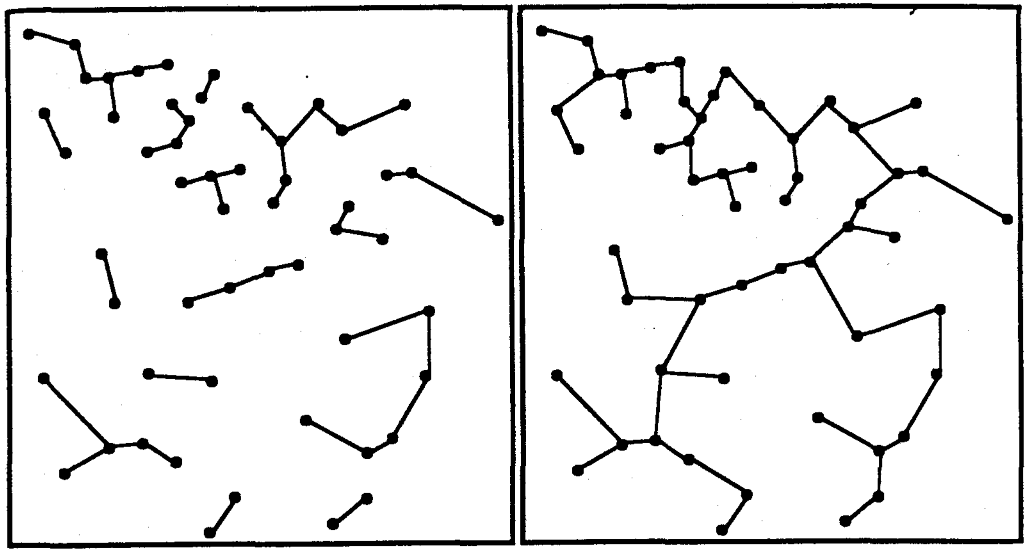


Figure 3.14.
All nearest neighbor graph at left. This graph is a subgraph of the minimal spanning tree, shown at right.

Remark. [Isolated points. Single-linkage clustering.]

If $X_1, \ldots, X_n$ are d-dimensional data points, and $r > 0$ is a number depending upon n only, then X_i is said to be **isolated point** if the closed sphere of radius r around X_i contains no X_j, $j \neq i$.

Isolated points are important in statistics. They can often be considered as "outliers" to be discarded in order not to destabilize certain computations. In the theory of clustering, the following algorithm is well-known: construct a graph in which X_i and X_j are joined when they are within distance r of each other. The connected components in the graph are the clusters. When r grows, there are fewer and fewer connected components of course. Thus, if we can find all pairs (X_i, X_j) within distance r of one another very quickly, then the clustering algorithm will be fast too, since the connected components can be grown by the union-find parentpointer tree algorithm (see e.g. Aho, Hopcroft and Ullman (1983, pp. 184-189)). This clustering method is equivalent to the **single linkage clustering method** (see e.g. Hartigan (1975, chapter 11)). The isolated points algorithms discussed below will all give an exhaustive listing of the pairs (X_i, X_j)

that satisfy $||X_i - X_j|| \leq r$, and can thus be used for clustering too. The problem of the identification of these pairs is called the close pairs problem.

There are two bucket-based solutions to the close-pairs problem. First, we can define a grid of hypercubes (buckets) with sides dependent upon r. The disadvantage of this is that when r changes, the bucket structure needs to be redefined. The advantage is that when n changes, no such adjustment is needed. In the second approach, the bucket size depends upon n only: It is independent of r.

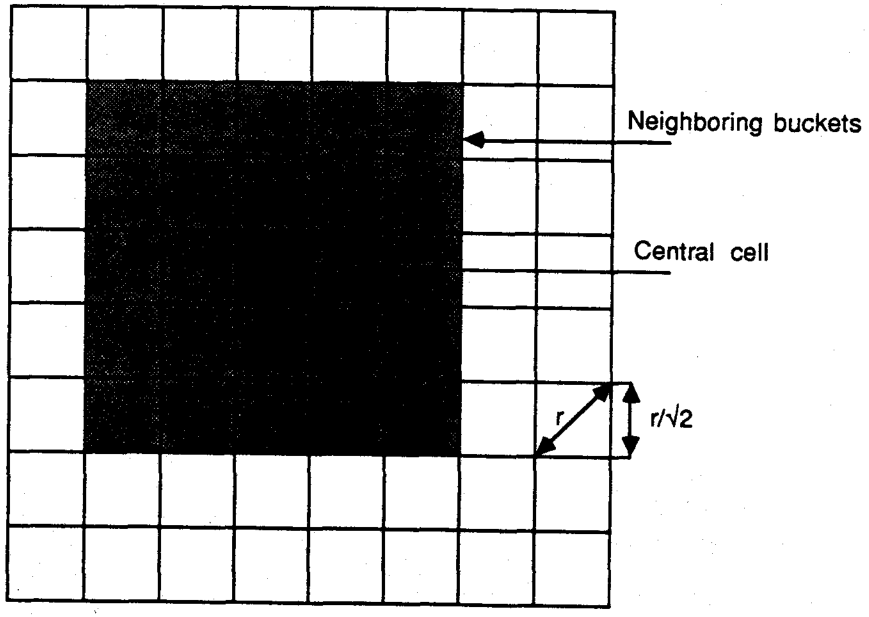


Figure 3.15.

In the r-dependent grid, it is useful to make the sides equal to $r/\sqrt{d}$ because any pair of points within the same bucket is within distance r of each other. Furthermore, points that are not in neighboring buckets cannot be within distance r of each other. By neighboring bucket, we do not mean a touching bucket, but merely one which has a vertex at distance r or less of a vertex of the original bucket. A conservative upper bound for the number of neighboring buckets is $(2\sqrt{d}+3)^d$. In any case, the number depends upon d only, and will be denoted by γ_d. To identify isolated points, we first mark single point buckets, i.e. buckets with $N_i = 1$, and check for each marked point all γ_d neighboring buckets. The sum of distance computations involved is

$$\sum_{i:N_i=1} \quad \sum_{j:A_j \text{ neighbor of } A_i} N_j$$

$$= \sum_j N_j \sum_{i:N_i=1, \text{ and } A_i \text{ neighbor of } A_j} 1$$

$$\leq \gamma_d \sum_j N_j$$

$$= \gamma_d n.$$

The grid initialization takes time $\Omega(r^{-d})$ and $O(\min(r^{-d},1))$. In particular, the entire algorithm is $O(n)$ in the worst-case whenever $rn^{1/d} \geq c > 0$ for some constant c. For r much smaller than $n^{-1/d}$, the algorithm is not recommended because nearly all the points are isolated points - the bucket size should be made dependent upon n instead.

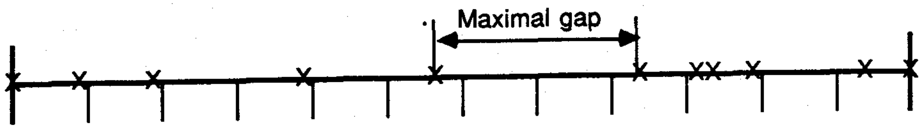


Figure 3.16.
Finding the maximal gap in a sequence of
n points by dividing the range into n+1 intervals.

Remark. [The maximal gap.]

The maximal gap in a sequence of points $x_1, \ldots, x_n$ taking values on [0,1] is the maximal interval induced by these points on [0,1]. As in the case of isolated points, the maximal gap can be found in worst-case time $O(n)$. For example, this can be done by observing that the maximal gap is at least $\frac{1}{n+1}$. Thus, If we organize the data into a bucket structure with $n+1$ intervals, no two points within the same bucket can define the maximal gap. Therefore, it is not necessary to store more than two points for each bucket, namely the maximum and the minimum. To find the maximal gap, we travel from left to right through the buckets, and select the maximum of all differences between the minimum of the current bucket and the last maximum seen until now. This algorithm is due to Gonzalez (1975).

Let us turn now to the close-pairs problem. The time needed for reporting all close pairs is of the order of

$$V = \sum_i N_i^2 + \sum_i N_i \sum_{j : A_j \text{ neighbor of } A_i} N_j$$

where the first term accounts for listing all pairs that share the same bucket, and the second term accounts for all distance computations between points in neighboring buckets.

For this problem, let us consider a grid of m buckets. This at least guarantees that the initialization or set-up-time is $O(n+m)$. The expected value of our performance measure V is

$$E(V) = E(\sum_i N_i^2 + \sum_i N_i \sum_{j : A_j \text{ neighbor of } A_j} N_j)$$

and it is the last term which causes some problems because we do not have a full double sum. Also, when $p_i = \int_{A_i} f$ is large, p_j is likely to be large too since A_i and A_j are neighboring buckets. The asymptotics for $E(V)$ are obtained in the next theorem. There are 3 situations when $m = n$:

A. $nr^d \to \infty$ as $n \to \infty$: the expected number of close pairs increases roughly speaking faster than n.

B. $nr^d \to 0$ as $n \to \infty$: the expected number of close pairs is $O(n)$, and the probability that any given point is an isolated point tends to 1.

C. $nr^d \to \beta \in (0,\infty)$ as $n \to \infty$: the expected number of close pairs increases as a constant times n. This is the critical case.

The upper bound in the theorem is valid in all three cases. In fact, Theorem 3.3 also covers the situation that $m \neq n$: m and / or r are allowed to vay with n in an arbitrary manner.

Theorem 3.3.

Let $\gamma = \gamma(r, d, m)$ be the number of neighboring buckets of a particular bucket in a grid of size m defined on $[0,1]^d$, where r is used in the definition of neighbor. Then

$$E(V) \leq n + \frac{n^2}{m}(\gamma + 1) \int f^2 .$$

If $m \to \infty$, $n \to \infty$, $r \to 0$,

$$E(V) = n + \frac{n^2}{m}(\gamma + 1 + o(1)) \int f^2 .$$

Note that if $mr^d \to \infty$ $r \to 0$, $\gamma(r, d, m) \sim mr^d V_d$ where V_d is the value of the unit sphere in R^d. Thus,

$$E(V) = n + n^2 r^d V_d (1 + o(1)) \int f^2 .$$

If $mr^d \to \beta \in (0,\infty)$, then γ oscillates but remains bounded away from 0 and ∞ in the tail. In that case,

$$E(V) = O(n)$$

when $\int f^2 < \infty$, $m \sim cn$. Note that $E(V) = \Omega(n)$ in all cases.

Finally, if $mr^d \to 0$, such that $r > 0$ for all n, m, then $\gamma \to 3^d - 1$, and

$$E(V) = n + \frac{n^2}{m} 3^d \int f^2 (1 + o(1)) .$$

Proof of Theorem 3.3.

We will use the notation $A(x)$ for the bucket A_i to which x belongs. Furthermore, $B(x)$ is the collection of neighboring buckets of $A(x)$. Define the densities

$$f_n(x) = \frac{1}{|A(x)|} \int_{A(x)} f \ , \ x \in [0,1]^d$$

$$g_n(x) = \frac{1}{|B(x)|} \int_{B(x)} f \ , \ x \in [0,1]^d .$$

Note that by the Lebesgue density theorem, if $m \to \infty$, $r \to 0$ (and thus $|A(x)| \to 0$, $|B(x)| \to 0$), $f_n(x) \to f(x)$ and $g_n(x) \to g(x)$ for almost all x. This result can be obtained without trouble from Lemmas 5.10, 5.11, and the fact that the definition of neighboring bucket is data independent and depends upon r and m only.

The upper bound will be derived first. The sum V is split into $V_1 + V_2$. Only V_2 causes some problems since $E(V_1) \le n^2 \sum_{i=1}^{m} p_i^2 + n \le \frac{n^2}{m} \int f^2 + n$ by Lemma 1.1. Note also for future reference that $E(V_1) \ge n + (1+o(1)) \frac{n^2}{m} \int f^2$ when $m \to \infty$ if we apply the Fatou lower bound argument of the proof of Lemma 1.1. Turning to V_2, we have, by Lemma 5.1,

$$E(V_2) = \sum_{i=1}^{m} \sum_{j:A_j \text{ neighbor of } A_i} p_i p_j \ n(n-1)$$

$$\le n^2 \sum_{i=1}^{m} \lambda(A_i)\lambda(B_i) \ f_n(x_i) g_n(x_i) \ (\text{for any } x_1 \in A_1, \ldots, x_m \in A_m)$$

$$= n^2 \sum_{i=1}^{m} \int_{A_i} \lambda(B_i) \ f_n(x) g_n(x) dx$$

$$= n^2 \gamma \lambda(A_1) \int f_n g_n \ .$$

Since f_n and g_n are probably very close to each other, the integral in the last expression is probably very close to $\int f_n^2$. Therefore, little will be lost if the integral is bounded from above by the Cauchy-Schwartz inequality:

$$\int f_n g_n \le \sqrt{\int f_n^2 \int g_n^2}$$

$$\le \sqrt{\int \frac{1}{\lambda(A(x))} \left(\int_{A(x)} f^2 \right) dx} \ \sqrt{\int \frac{1}{\lambda(B(x))} \left(\int_{B(x)} f^2 \right) dx}$$

(Jensen's inequality)

$$= \sqrt{\sum_{i=1}^{m} \int_{A_i} f^2} \ \sqrt{\sum_{i=1}^{m} \int_{A_i} f^2}$$

$$= \int f^2 .$$

To treat $\int g_n^2$ we have argued as follows:

$$\int_{[0,1]^d} g_n^2 \le \int_{R^d} g_n^2$$

$$= \int \left(\frac{1}{\lambda(B(x))} \int_{B(x)} f(y) dy \right)^2 dx$$

(where $B(x)$ now refers to an infinite grid)

$$\le \int \frac{1}{\lambda(B(x))} \left(\int_{B(x)} f^2 \right) dx \quad \text{(Jensen' s inequality)}$$

$$= \sum_{i=1}^{\infty} \int_{A_i} \frac{1}{\gamma \lambda(A_i)} \sum_{j:A_j \text{ neighbor of } A_i} \int_{A_j} f^2(y) dy \ \ dx$$

$$= \sum_{i=1}^{\infty} \frac{1}{\gamma \lambda(A_1)} \int_{A_j} f^2(y) \left(\sum_{i:A_i \text{ neighbor of } A_j} \int_{A_i} dx \right) dy$$

$$= \sum_{i=1}^{\infty} \int_{A_j} f^2$$

$$= \int f^2 .$$

Note also that

$$\lim\inf \int f_n g_n \ge \int \lim\inf f_n g_n = \int f^2$$

when $m \to \infty$, $r \to 0$. This concludes the proof of the first two statements o the theorem. The remainder of the theorem is concerned with the size of γ as : function of r and m, and follows from elementary geometric principles.

We note for example that when $m \to \infty$, $mr^d \to 0$, the optimal choice fo m would be a constant times $n\sqrt{3^d \int f^2}$ - at least, this would minimiz $Cm + E(V)$ asymptotically, where C is a given constant. The minimizin value is a constant times $n\sqrt{3^d \int f^2}$. The only situation in which $E(V)$ is no $O(n)$ for $m \sim cn$ is when $nr^d \to \infty$, i.e. each bucket has very many dat points. It can be shown that the expected number of close pairs grows as a constant times $n^2 r^d$, and this provides a lower bound for $E(V)$. Thus, th expected time for $E(V)$ obtained in Theorem 3.3 has an optimal asymptotic rate

Remark. [The all-nearest-neighbor problem.]

All nearest neighbor pairs can be found in $O(n \log n)$ worst-case time (Shamos and Hoey, 1975). Weide (1978) proposed a bucketing algorithm in which for a given X_i, a "spiral search" is started in the bucket of X_i, and continues in neighboring cells, in a spiraling fashion, until no data point outside the buckets already checked can be closer to X_i than the closest data point already found. Bentley, Weide and Yao (1980) showed that Weide's algorithm halts in average time $O(n)$ when there exists a bounded open convex region B such that the density f of X_1 is 0 outside B and satisfies $0 < \inf_B f(x) \le \sup_B f(x) < \infty$. (This condition will be called the BWY condition.)

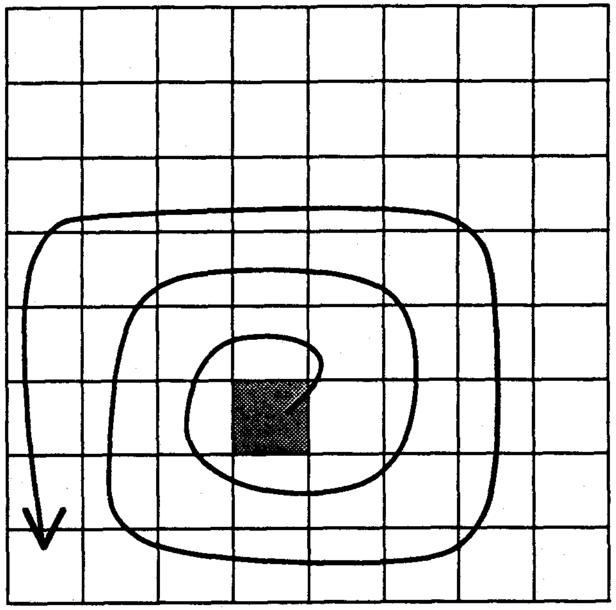


Figure 3.17.
Spiral search for nearest neighbor.

Remark. [The closest pair problem.]

To find the closest pair in $[0,1]^d$, one can argue geometrically and deduce an absolute upper bound of the form C_d/n^d for the smallest distance between any two points among $X_1, \ldots, X_n$ in $[0,1]^d$. Here C_d is a constant depending upon d only. If we construct a grid with buckets having sides C_d/n^d, then we can hope to "catch" the closest pair in the same bucket. Unfortunately, the closest pair can be separated by a bucket boundary. This case can be elegantly covered by shifting the grid appropriately a number of times so that for one of the shifted grids there is a bucket which contains the closest pair (Yuval, 1976). Ignoring the dependence upon d, we see that with this strategy, the time complexity is of the form $c_1 n + c_2 \sum_{i=1}^{n} N_i^2$ where the square accounts for the computations of all pairwise distances within the same bucket, and $c_1, c_2 > 0$ are constants. It is easy to see that if $X_1, \ldots, X_n$ are iid random vectors with density f on $[0,1]^d$, then the shifted grid method takes expected time $O(n)$ if and only if $\int f^2 < \infty$ Rabin (1976) chooses a small subset for which the closest pair is found. Th corresponding minimal distance is then used to obtain the overall closest pair i linear expected time. It is perhaps interesting to note that not much is gaine over worst-case time under our computational model, since there exist algorithm which can find the closest pair in worst case time $O(n \log\log n)$ (Fortune an Hopcroft, 1979).

Remark. [The Euclidean minimal spanning tree.]

For a graph (V,E)) Yao (1975) and Cheriton and Tarjan (1976) give algorithms for finding the minimal spanning tree (MST) in worst-case tim $O(|E|\log\log|V|)$. The Euclidean minimal spanning tree (EMST) of n points i R^d can therefore be obtained in $O(n \log\log n)$ time if we can find a supergraph of the EMST with $O(n)$ edges in $O(n \log\log n)$ time. Yao (1982) suggested to find the nearest neighbor of each point in a critical number of directions; the resulting graph has $O(n)$ edges and contains the MST. This neares neighbor search can be done by a slight modification of spiral search (Weld (1978)). Hence, the EMST can be found in expected time $O(n \log\log n)$ fo any d and for all distributions satisfying the BWY condition. The situation is a bit better in R^2. We can find a planar supergraph of the EMST in expected time $O(n)$ (such as the Delaunay triangulation (the dual of the Voronoi diagram), the Gabriel graph, etc.) and then apply Cheriton and Tarjan's (1976) $O(n)$ algorithm for finding the MST of a planar graph. For a linear expected time Vorono diagram algorithm, see Bentley, Weide and Yao (1980). Thus, in R^2 and for the class of BWY distributions, we can find the EMST in linear expected time.

Chapter 4

THE MAXIMAL CARDINALITY

The expected value of the worst possible search time for an element in a bucket data structure is equal to the expected value of $M_n = \max_{1 \le i \le m} N_i$ times a constant. This quantity differs from the worst-case search time, which is the largest possible value of $\max_{1 \le i \le m} N_i$ over all possible data sets, i.e. n. In a sense, the maximal cardinality has taken over the role of the height in tree structures. Its main importance is with respect to searching. Throughout the chapter, it is crucial to note the dependence of the maximal cardinality upon the density f of the data points $X_1, \ldots, X_n$, which for the sake of simplicity are assumed to take values on $[0,1]^d$. The grid has $m \sim cn$ cells for some constant $c > 0$, unless we specify otherwise.

In section 4.1, we look at the properties of M_n, and in particular of $E(M_n)$ following analysis given in Devroye (1985). This is then generalized to $E(g(M_n))$ where g is a nonlinear work function (see section 4.3). Such nonlinear functions of M_n are important when one particular bucket is selected for further work, as for example in a bucket-based selection algorithm (section 4.2). Occasionally, the maximal cardinality can be useful in the analysis of bucket algorithms in which certain operations are performed on a few buckets, where buckets are selected by the data points themselves. In section 4.4, we will illustrate this on extremal point problems in computational geometry.

4.1. EXPECTED VALUE AND INEQUALITIES.

For the uniform distribution on [0,1], Gonnet (1981) has shown that when $m = n$,

$$E(M_n) \sim \Gamma^{-1}(n)$$

where Γ is the gamma function. For example, when $n = 40320$, $E(M_n)$ is near 7.35 (Gonnet, 1981, table V). In other words, $E(M_n)$ is very small for all practical values of n. Additional information is given in Larson (1982). The situation studied by Gonnet pertains mainly to hashing with separate chaining when a perfect hash function is available. As we know, order-preserving hash functions lead to non-uniform distributions over the locations, and we will see here how $E(M_n)$ depends upon f. This is done in two steps. First we will handle the case of bounded f, and then that of unbounded f.

Theorem 4.1.

Assume that $f^* = \text{ess sup } f < \infty$ (note: $\lambda\{x : f(x) > f^*\} = 0$; $\lambda\{x : f(x) > f^* - \epsilon\} > 0$ for all $\epsilon > 0$). Then, if $m \sim cn$ for some $c > 0$,

$$E(M_n) \sim \frac{\log n}{\log \log n}$$

and, in particular,

$$E(M_n) = \frac{\log n}{\log \log n} + \frac{\log n}{(\log \log n)^2} (\log \log \log n + \log(\frac{f^* e}{c}) + o(1)) .$$

Proof of Theorem 4.1.

We will use a Poissonization device. Assume first that we have shown the statement of the theorem for M_n^* where $M_n^* = \max_i N_i^*$ and N_i^* is the number of X_i' s in $X_1, \ldots, X_N$ belonging to A_i, where N is a Poisson (n) random variable independent of $X_1, X_2, \ldots$. Now, for all $\epsilon > 0$, we have

$$M_n^* \le M_{n(1+\epsilon)} + nI_{N \ge n(1+\epsilon)}$$

and

$$M_n^* \ge M_{n(1-\epsilon)} - nI_{N \le n(1-\epsilon)}$$

where I is the indicator function, and where $n(1+\epsilon)$ and $n(1-\epsilon)$ should be read as "the smallest integer at least equal to ...". By Lemma 5.8,

$$nP(|N-n| \ge n\epsilon) \le \frac{4}{n\epsilon^4} .$$

Define $b(n) = 1 + \log(f^*/c) + \log \log n + \log \log \log n$, $c(n) = \frac{(\log \log n)^2}{\log n}$. Thus, by assumption,

$$o(1) = E(M_n^*)c(n) - b(n) \le E(M_{n(1+\epsilon)})c(n) + \frac{4c(n)}{n\epsilon^4} - b(n)$$

$$\le E(M_{n(1+\epsilon)})c(n(1+\epsilon))\frac{c(n)}{c(n(1+\epsilon))}$$

$$+ o(\frac{1}{n}) - b(n(1+\epsilon)) + (b(n(1+\epsilon)) - b(n)).$$

Now, $b(n(1+\epsilon)) - b(n) = o(1)$, and, for n large enough, $c(n) \ge c(n(1+\epsilon)) \ge c(n)\frac{\log n}{\log(n(1+\epsilon))} \ge c(n)/(1+\epsilon/\log n)$.

Thus,

$$E(M_{n(1+\epsilon)}) \ge \frac{b(n(1+\epsilon)) + o(1)}{c(n(1+\epsilon))(1+\epsilon/\log n)}$$

$$= \frac{b(n(1+\epsilon)) + o(1)}{c(n(1+\epsilon))}$$

Similarly, it can be shown that $E(M_n) \le (b(n) + o(1))/c(n)$, and combining this gives us our theorem.

Lower bounds for M_n^*.

Let $\eta > 0$ be an arbitrary number, and let $\epsilon > 0$ be the solution of $\eta = -2\log(1 - \frac{2}{f^*}\epsilon)$ (this will turn out to be a convenient choice for ϵ). Let A be the set $\{x : f(x) > f^* - \epsilon\}$, and let $\delta = \int_A dx$ (which is positive by definition of f^*). Finally, let $h = h_n$ be the integer part of $\frac{b(n) - \eta}{c(n)}$. We let p_i keep its meaning from the introduction, and note that the function f_n on [0,1] defined by

$$f_n(x) = mp_i \ , x \in A_i \ ,$$

is a density. Because $N_1^*, N_2^*, \ldots, N_m^*$ are independent Poisson random variables with parameters $np_1, np_2, \ldots, np_m$ respectively, we have the following chain of inequalities:

$$P(M_n^* < h) = \prod_{i=1}^{m} P(N_i^* < h)$$

$$\leq \prod_{i=1}^{m} (1 - P(N_i^* = h))$$

$$\leq \exp(-\sum_{i=1}^{m} P(N_i^* = h))$$

$$= \exp\left(-\sum_{i=1}^{m} (np_i)^h \ \frac{e^{-np_i}}{h\,!}\right)$$

$$= \exp\left(-m \int \left(\frac{nf_n}{m}\right)^h \frac{e^{-\frac{n}{m}f_n}}{h\,!}\right).$$

By Lemmas 5.10 and 5.12,

$$\int_A \left(\frac{\frac{n}{m}f_n}{\frac{n}{m}(f^* - 2\epsilon)}\right)^h e^{-\frac{n}{m}f_n} \geq \left(\int_{A\ ,\ f_n > f^* - 2\epsilon} dx\right) e^{-\frac{n}{m}f^*}$$

$$\geq e^{-\frac{n}{m}f^*} \int_{A\ ,\ |f_n - f| \leq \epsilon} dx$$

$$= e^{-\frac{n}{m}f^*} \left(\delta - \int_{A\ ,\ |f_n - f| > \epsilon} dx\right)$$

$$\geq e^{-\frac{n}{m}f^*} \left(\delta - \int_A \frac{|f_n - f|}{\epsilon}\right)$$

$$\geq e^{-\frac{n}{m}f^*} (\delta - o(1)).$$

Thus,

$$P(M_n^* < h) \leq \exp\left(-\frac{m}{h\,!}\left(\frac{n}{m}\right)^h (f^* - 2\epsilon)^h \ e^{-\frac{n}{m}f^*} (\delta - o(1))\right).$$

Using Stirling's approximation for $h\,!$, we see that the exponent is $-e^s$ where

$$s = \log m - h \log h + h - \frac{1}{2}\log(2\pi h)$$

$$+ o(1) + h \log(\frac{f^* - 2\epsilon}{c}) - \frac{n}{m}f^* + \log \delta$$

$$= \log n + \frac{b(n) - \eta}{c(n)} (1 + \log(\frac{f^* - 2\epsilon}{c}))$$

$$- \frac{b(n) - \eta}{c(n)} \log(\frac{b(n) - \eta}{c(n)} - \frac{1}{2}\log(\frac{b(n) - \eta}{c(n)} + t$$

$$(\text{where } t = \log \delta - \frac{1}{2}\log(2\pi) - \frac{f^*}{c} + \log c + o(1))$$

$$= \log n + \frac{b(n) - \eta}{(\log \log n)^2} \log n \ (1 + \log(\frac{f^* - 2\epsilon}{c})$$

$$- \log \log n + \log \log \log n + o(1))$$

$$- \frac{1}{2} \log \log n + \frac{1}{2} \log \log \log n + t + o(1)$$

$$= \frac{\log n}{\log \log n} \left(o(1) + \left(1 + \frac{\log \log \log n}{\log \log n} + \frac{1 + \log(f^*/c) - \eta}{\log \log n}\right)\right.$$

$$\left.\left(1 + \log(\frac{f^* - 2\epsilon}{c}) - \log \log n + \log \log \log n + o(1)\right) + \log \log n\right)$$

$$= \frac{\log n}{\log \log n} \left(\log\left(\frac{f^* - 2\epsilon}{c}\right) - \log\left(\frac{f^*}{c}\right) + \eta + o(1)\right)$$

$$\geq \frac{\eta}{3}\,\frac{\log n}{\log\log n} \text{ (all } n \text{ large enough)}$$

because $\log(f^{\,*}-2\epsilon)-\log(f^{\,*}) = -\frac{\eta}{2}$.

Thus, for all n large enough,

$$E(M_n^{\,*}) \geq h\ P(M_n^{\,*} \geq h) = h(1-P(M_n^{\,*} < h))$$

$$\geq h(1-\exp(-\exp(\frac{\eta}{3}\frac{\log n}{\log\log n}))) \geq h(1-\exp(-\exp(\log\log n)))$$

$$= h(1-\frac{1}{n}) \geq (\frac{b(n)-\eta}{c(n)}-1)(1-\frac{1}{n}) = \frac{b(n)-\eta-o(1)}{c(n)}\ .$$

This concludes the proof of the lower bound, since $\eta > 0$ is arbitrary.

Upper bounds for $M_n^{\,*}$.

Again, we let η be an arbitrary positive number, and choose $h = h_n$ as the integer part of $\frac{b(n)+\eta}{c(n)}$. Let $k \geq h$ be some integer. Then, for $h \geq c$, by Lemma 5.9,

$$P(M_n^{\,*} \geq k) \leq \sum_{i=1}^{n} P(N_i^{\,*} \geq k) \leq n\sum_{j\geq k} c^{\,j}\frac{e^{-c}}{j\,!}$$

$$\leq nc^{\,k}\ \frac{e^{-c}}{k\,!}\ \frac{k+1}{k+1-c}\ .$$

Thus,

$$E(M_n^{\,*}) \leq h + \sum_{k=h}^{\infty} P(M_n^{\,*} \geq k) \leq h + \sum_{k=h}^{\infty} nc^{\,k}\ \frac{e^{-c}}{k\,!}\ \frac{k+1}{k+1-c}$$

$$\leq h + nc^{\,h}\ \frac{e^{-c}}{h\,!}\ (\frac{h+1}{h+1-c})^2\ .$$

By some straightforward analysis, one can show that

$$\log(nc^{\,h}\ \frac{e^{-c}}{h\,!}) \geq -(\eta+o(1))\frac{\log n}{\log\log n}$$

and that

$$(\frac{h+1}{h+1-c})^2 = 1 + \frac{2c}{h+1} + o(\frac{1}{h})\ .$$

Therefore,

$$E(M_n^{\,*} \leq h + (1+\frac{2c}{h}+o(\frac{1}{h}))\ \exp(-(\eta+o(1))\frac{\log n}{\log\log n})$$

$$\leq h + (\frac{1+o(1)}{\log n})(1+\frac{2c}{h}+o(\frac{1}{h})) \leq \frac{b(n)+\eta}{c(n)} + \frac{1+o(1)}{\log n}$$

$$= \frac{b(n)+\eta+o(1)}{c(n)}\ .$$

But η was arbitrary. This concludes the proof of the theorem.

For all bounded f, we have

$$E(M_n) \sim \frac{\log n}{\log\log n}$$

whenever $m \sim cn$. In first approximation, the density does not influence $E(M_n)$. The explanation is due to the fact that the expected value of the maximum of n independent Poisson (λ) random variables is asymptotic to $\log n\ /\ \log\log n$ for any constant λ. The influence of $f^{\,*}$ on $E(M_n)$ is in the third largest asymptotic expansion term only. The proof of Theorem 4.1 is long and tedious because we want to obtain rather refined information. From here onwards, we will content ourselves with main asymptotic terms only.

Theorem 4.1 remains valid when the minimum and the maximum of the $X_i{}'\,s$ are used to determine an initial interval, and the buckets are defined by dividing this interval into n equal sub-intervals. The density f is assumed to have support contained in [0,1] but not in [0, 1−ϵ] or [ϵ,1] for any $\epsilon > 0$.

When f is unbounded, the theorem gives very little information about $E(M_n)$. Actually, the behavior of $E(M_n)$ depends upon a number of quantities that make a general statement all but impossible. In fact, any slow rate of convergence that is $o(n)$ is achievable for $E(M_n)$. Since N_i is binomial (n, p_i) where p_i is the integral of f over the i-th bucket, we have

$$\max_i np_i \leq E(\max_i N_i) = E(M_n) .$$

When f is monotone nonincreasing, the left-hand-side of this inequality is equal to $nF(\frac{1}{n})$ where F is the distribution function corresponding to f . Thus, since any slow rate of decrease to 0 is possible for F, when $n \to \infty$, any slow rate $o(n)$ is achievable for $E(M_n)$. The rate $\log n / \log \log n$, achieved by all bounded densities, is also a lower bound for $E(M_n)$ for all densities.

This note would not be complete if we did not mention how $E(M_n)$ varies when $\max_i np_i$ diverges. Most of this information can be deduced from the inequalities given in Theorem 4.2 below. For example, we will see that $E(M_n) \sim \log n / \log \log n$ (the optimal rate achievable) when q diverges very slowly, and that $E(M_n) \sim \frac{n}{m} q$ when q diverges rapidly.

Theorem 4.2.

Let $q = \max_{1 \leq i \leq m} mp_i$. Then

$$\frac{n}{m} q \leq E(M_n) \leq \frac{n}{m} q + \frac{1}{t} (\log m + \frac{n}{m} q (e^t - t - 1))$$

$$= \frac{\log m}{t} + \frac{n}{m} q (\frac{e^t - 1}{t}), \text{ all } t > 0, m \geq 3.$$

Proof of Theorem 4.2.

The lower bound follows directly from Jensen's inequality. To derive the upper bound, we let $U_i = N_i - np_i$, $U = \max_i U_i$. Note that U is a nonnegative random variable. We have

$$M_n \leq \max_i np_i + \max_i U_i = \frac{n}{m} q + U.$$

For $r \geq 1$, we can apply Jensen's inequality again:

$$E^r(U) \leq E(U^r) = E(\max_i U_i{}^r) \; (u^r \text{ is considered sign-preserving})$$

$$\leq m \max_i E((U_i{}^r)_+) \leq m \max_i E((\frac{r}{e\ t})^r e^{tU_i}) \text{ , all } t > 0.$$

Here we used the inequality $u_+{}^r \leq (\frac{r}{e\ t})^r e^{tu}$, $t > 0$, where $u_+ = \max(u, 0)$. Also,

$$E(e^{tU_i}) = E(e^{-tnp_i} e^{tN_i}) = e^{-tnp_i}(e^t p_i + 1 - p_i)^n \leq e^{np_i(e^t - t - 1)}$$

$$\leq e^{\frac{n}{m} q (e^t - t - 1)} .$$

Thus,

$$E(M_n) \leq \frac{n}{m} q + \frac{r}{et} m^{\frac{1}{r}} \exp(\frac{n}{m} \frac{q}{r} (e^t - t - 1)).$$

This bound is minimal with respect to r when $r = \log m + \frac{n}{m} q (e^t - t - 1)$ (just set the derivation of the logarithm of the second term in the bound equal to 0). Resubstitution give the desired result. The restriction $r \geq 1$ forces us to choose $m \geq 3$.

Theorem 4.2 shows that there are many possible cases to be considered with respect to the rates of increase of q and m . Assume that $m \sim cn$, which is the standard case. Then

$$E(M_n) \sim \frac{n}{m} q \sim \frac{q}{c}$$

when $q / \log n \to \infty$. To see this, observe that

$$e^t - t - 1 \le \frac{t^2}{2} e^t ,$$

so that

$$E(M_n) \le \frac{n}{m} q \frac{1}{t} \log m + \frac{n}{m} q \frac{t}{2} e^t , \; t > 0.$$

Take $t = \sqrt{\frac{2m}{nq} \log m}$ (this minimizes the upper bound when e^t is neglected), and note that

$$E(M_n) \le \frac{n}{m} q + \sqrt{2 \frac{n}{m} q \log m} \; (1+o(1)) \sim \frac{n}{m} q .$$

In this case, the bound of Theorem 4.2 is tight.

Consider a second case at the other end of the spectrum, the very small $q : q = (\log n)^{o(1)}$ (or: $\log q = o(\log \log n)$). Then the upper bound is

$$E(M_n) \le (1+o(1)) \frac{\log m}{\log \left(\frac{\log m}{\frac{n}{m} q} \right)} \sim \frac{\log n}{\log \log n}$$

when we take $t = \log \left(\frac{\log m}{\frac{n}{m} q} \right) - \log \log \left(\frac{\log m}{\frac{n}{m} q} \right)$

(note that this choice of t almost minimizes the upper bound). Thus, Theorem 4.2 provides a considerable short-cut over Theorem 4.1 if one is only interested in first terms.

A third case occurs when $q = o(\log n)$, but q is not necessarily very small. In that case, for the same choice of t suggested above, we have

$$E(M_n) \le (1+o(1)) \frac{\log m}{\log \left(\frac{\log m}{\frac{n}{m} q} \right)} \sim \frac{\log n}{\log \left(\frac{\log n}{q} \right)} .$$

The only case not covered yet is when $q \sim \alpha \log n$ for some constant $\alpha > 0$. It is easy to see that by taking t constant, both the upper and lower bound for $E(M_n)$ vary in proportion to q. Since obviously the bounds implicit in Theorem 4.1 remain valid when $q \to \infty$, we see that the only case in which there might be a discrepancy between the rate of increase of upper and lower bounds is our "third" case.

Remark 4.1. [The behavior of $\max_{1 \le i \le m} mp_i$.]

The behavior of M_n for unbounded densities depends rather heavily on the behavior of $q = \max_{1 \le i \le m} mp_i$. It is useful to relate this maximum to f. In particular, we need to be able to bound the maximum in terms of f. One possible polynomial bound is obtained as follows: for any set A_i, and any $r \ge 1$,

$$\left(\frac{\int_{A_i} f}{\lambda(A_i)} \right)^r \le \frac{1}{\lambda(A_i)} \int_{A_i} f^r \text{ (Jensen's inequality).}$$

Thus,

$$q = \max_{1 \le i \le m} mp_i \le m^{\frac{1}{r}} (\int f^r)^{\frac{1}{r}} .$$

The less outspoken the peakedness of f is (i.e. the smaller $\int f^r$), the smaller the bound. For densities f with extremely small infinite peaks, the functional generating function is finite: $\psi(u) = \int e^{uf} < \infty$, some $u > 0$. For such densities, even better bounds are obtainable as follows:

$$\exp \left(u \frac{\int_{A_i} f}{\lambda(A_i)} \right) \le \frac{1}{\lambda(A_i)} \int_{A_i} \exp(u \; f)$$

$$\le m \; \psi(u).$$

Thus,

$$\max_{1 \le i \le m} mp_i \le \frac{\log m + \log \psi(u)}{u} .$$

The value of u for which the upper bound is minimal is typically unknown. If we keep u fixed, then the upper bound is $O(\log(m))$, and we are almost in the domain in which $E(M_n) \sim \log n / \log\log n$. If $\psi(u) < \infty$ for **all** $u > 0$ then we can find a subsequence $u_m \uparrow \infty$ such that $\psi(u_m) \leq m$ for all m. It is easy to see that the maximum of the $mp_i{}' s$ is $o(\log m)$, so that $E(M_n) \leq \frac{\log n}{\log((\log n)/q)}(1+o(1))$. If $\psi(\log\log m) \leq m^{O(1)}$, then $E(M_n) = O(\frac{\log n}{\log\log\log n})$. Thus, the functional generating function aids in the establishment of simple verifiable conditions for different domains of behavior of $E(M_n)$.

Remark 4.2. [Double bucketing.]

It is a rather straightforward exercise to show that for bounded f on $[0,1]^d$, if all buckets are further subdivided into grids of sizes $N_1, \ldots, N_m$, as is done in section 1.5 for example, then, when $m \sim cn$,

$$E(M_n) \sim \frac{\log\log n}{\log\log\log n} .$$

Here M_n is the maximal cardinality in any of the buckets in the small grids. Intuitively, this can be seen as follows: for the original grid, M_n is very close to $\log n / \log\log n$. For the buckets containing about $\log n / \log\log n$ elements, we obtain an estimate of $E(M_n)$ for the maximal cardinality in its sub-buckets by applying the results of this section after replacement of n by $\log n / \log\log n$. Thus, as a tool for reducing the maximal cardinality in the bucket data structure, double bucketing is quite efficient although not perfect (because $E(M_n) \to \infty$).

Remark 4.3. [Poissonization.]

The proof of Theorem 4.1 is based upon Poissonization of the sample size. The technical advantage is that M_n, a maximum of dependent binomial random variables, is replaced by $M_n{}^*$, a maximum of independent Poisson random variables. In fact, we can do without the Poissonization by using special properties of the multinomial distribution. To illustrate this, we could have used Mallows' inequality:

$$P\Big(\max_{1\leq i\leq m} N_i \leq x\Big] \leq \prod_{i=1}^{m} P(N_i \leq x) \leq \exp(-\sum_{i=1}^{m} P(N_i > x)) , x \geq 0$$

(Mallows, 1968), from which one deduces without work that

$$E(\max_{1\leq i\leq m} N_i) \geq E(\max_{1\leq i\leq m} N_i{}^*)$$

where, $N_1{}^*, \ldots, N_m{}^*$ are independent binomial random variables, distributed individually as $N_1, \ldots, N_m$. This can be used as a starting point for developing a lower bound.

Remark 4.4. [Historical remark.]

Kolchin, Sevast'yanov and Chistyakov (1978, pp. 94-111) have studied in some detail how M_n behaves asymptotically for different rates of increase of m, and for the uniform density on [0,1]. Their results can be summarized quite simply. A critical parameter is $\frac{n}{m}$, the average occupancy of a cell. There are three cases:

Case 1. If $\frac{n}{m \log m} \to 0$ as $n \to \infty$, then

$$\lim_{n\to\infty} P(M_n = r-1) = e^{-\lambda} ,$$

$$\lim_{n\to\infty} P(M_n = r) = 1 - e^{-\lambda} ,$$

where λ is a positive constant, and $r = r_n$ is chosen in such a way that $r > \frac{n}{m}$, $m\frac{(\frac{n}{m})^r e^{-\frac{n}{m}}}{r\,!} \to \lambda$. (Thus, asymptotically, M_n puts all its mass on two points.)

Case 2. $\frac{n}{m \log m} \to x \in (0,\infty)$.

Case 3. If $\frac{n}{m \log m} \to \infty$, then $M_n / (\frac{n}{m}) \to 1$ in probability.

Case 1 is by far the most important case because usually $m \sim cn$. In cases 2 and 3, the asymptotic distribution of M_n is no longer bi-atomic because M_n spreads its mass more out. In fact, in case 3, M_n is with high probability equal to the value of the maximal cardinality if we were to distribute the n points evenly (not randomly!) over the m buckets! The difference $M_n - \frac{n}{m}$ is $\sim \sqrt{2\frac{n}{m} \log m}$ in probability provided that $m > n^{\epsilon}$ for some $\epsilon > 0$.

4.2. AN EXAMPLE : THE SELECTION PROBLEM.

Assume that a bucket structure is used to find the k-th smallest of $X_1, \ldots, X_n$, independent random variables with density f on [0,1]. The m buckets are of size $\frac{1}{m}$ each, but what will be said below remains valid if the m buckets are defined on [min X_i ,max X_i]. In the algorithm, we keep a count for each bucket, so that in one additional pass, it is possible to determine in which bucket the k-th smallest point lies. Within the bucket, this element can be found in several ways, e.g. via a linear worst-case comparison-based algorithm (Schonhage, Paterson and Pippenger, 1976; Blum, Floyd, Pratt, Rivest and Tarjan, 1973), via a linear expected time comparison-based algorithm (Floyd and Rivest, 1975; Hoare, 1961), or via a comparison-based sorting method. In the former two cases, we obtain linear worst-case time and linear expected time respectively, regardless of how large or small m is - we might as well choose $m = 1$. The constant in the time complexity might be smaller though for $m > 1$. If the buckets have cardinalities $N_1, \ldots, N_m$, then the time taken by the linear worst-case algorithm is bounded by

$$V = \alpha n + \beta \max_{1 \le i \le m} N_i + \gamma m$$

where $\alpha, \beta, \gamma > 0$ are constants, and the middle term describes the contribution of the linear worst-case comparison-based selection algorithm. While we can obviously bound all of this by $(\alpha + \beta) n + \gamma m$ (which would lead us to the choice $m = 1$), it is instructive to minimize $E(V)$. As we will see, it will be to our advantage to take m proportional to $\sqrt{n}$, so that $E(V) = \alpha n + 0(\sqrt{n})$ as $n \to \infty$.

The suggestion to take m proportional to $\sqrt{n}$ was also made by Allison and Noga (1980), but their algorithm is different, in that within a selected bucket, the algorithm is applied recursively. Note that the algorithm suggested here is more space efficient (since it is not recursive) but far less elegant (since it is a hybrid of a bucket algorithm and a fairly complicated linear comparison-based selection algorithm).

We note here that max N_i is used in the definition of V because we do not know beforehand which order statistic is needed. For example, the situation would be quite different if we were to ask for an average time, where the average is taken over all n possible values for k - in that case, the middle term would have to be replaced by $\beta \sum N_i^2$, and we can apply some of the analysis of chapter 1.

If sorting is used within a bucket, then the total time for selection is bounded by

$$V = \alpha n + \beta \max_{1 \le i \le m} N_i \log(N_i + 1) + \gamma m ,$$

or

$$V = \alpha n + \beta \max_{1 \le i \le m} N_i^2 + \gamma m ,$$

depending upon whether an $n \log n$ or a quadratic sort is used. To obtain a good estimate for $E(V)$, we need good estimates for $E(M_n \log(M_n + 1))$ and $E(M_n^2)$, i.e. for expected values of nonlinear functions of M_n. This provides some of the motivation for the analysis of section 4.3. In this section, we will merely apply Theorem 4.2 in the design of a fast selection algorithm when a linear worst-case algorithm is used within buckets. The main result is given in Theorem 4.3: this theorem applies to all bounded densities on [0,1] without exception. It is for this reason that we have to appeal, once again, to the Lebesgue density theorem in the proof.

Theorem 4.3.

Define for positive α, β, γ,

$$V = \alpha n + \beta \max_{1 \le i \le m} N_i + \gamma m ,$$

where $X_1, \ldots, X_n$ are iid random variables with bounded density f on [0,1] : $f(x) \le f^* < \infty$ for all x. Then, for any q, m:

$$\alpha n + \gamma m + \beta \frac{n}{m} q \le E(V)$$

$$\le \alpha n + \gamma m + \beta \left(\frac{n}{m} q + \sqrt{2 \frac{n}{m} q \log m} \sqrt{1 + e^s} \right)$$

where $s = \sqrt{2 \frac{m}{nq} \log m}$.

If we choose

$$m = \lfloor \sqrt{\frac{\beta}{\gamma} n f^*} \rfloor$$

then

$$E(V) \leq \alpha n + 2\sqrt{\beta\gamma n f^*} + O(n^{\frac{1}{4}} \log^{\frac{1}{2}} n)$$

and, in fact

$$E(V) = \alpha n + 2\sqrt{\beta\gamma n f^*}\,(1+o(1)).$$

Proof of Theorem 4.3.

The proof of Theorem 4.3 is based upon a crucial lemma.

Lemma 4.1.

For any bounded density f on $[0,1]^d$, and for any sequence $m \to \infty$, $q = \max_{1 \leq i \leq m} mp_i \to f^* = \text{ess sup } f$.

Proof of Lemma 4.1.

We will use the fact that for such f, $\lim_{r \to \infty} (\int |f|^r)^{1/r} = f^*$ (see Wheeden and Zygmund (1977, pp. 125-126)). Defining the density

$$f_m(x) = mp_i \ , \ x \in A_i \ ,$$

on $[0,1]^d$, we note that

$$f^* \geq q = \max_x f_m(x) = \text{ess sup } f_m \geq (\int f_m{}^r)^{1/r} \quad (\text{any } r)\ ,$$

and thus

$$\liminf_{m \to \infty} q^r \geq \int \liminf_{m \to \infty} f_m{}^r \text{ (Fatou's lemma)}$$

$$= \int f^r \text{ (Lemma 5.10)}$$

$$\geq (f^*)^r - \epsilon$$

by choice of $r = r(\epsilon)$, for arbitrary $\epsilon > 0$. This concludes the proof of the Lemma.

We continue now with the proof of Theorem 4.3. The starting point is the bound given immediately following the proof of Theorem 4.2. The choice of t is asymptotically optimal when $nq/m \log m \to \infty$. Since $q \geq 1$ in all cases, this follows if $n/m \log m \to \infty$, which is for example satisfied when $m \sim \sqrt{n}$, a choice that will be convenient in this proof. The upper and lower bounds for $E(V)$, ignoring lower order terms, are thus roughly $\alpha n + \gamma m + \beta \frac{n}{m} q$. Because $q \to f^*$ (Lemma 4.1), the choice $m = \lfloor \sqrt{\frac{\beta}{\gamma} n f^*} \rfloor$ is again asymptotically optimal. Resubtitution of this choice for m gives us our result.

Remark 4.5. [Choice of m.]

With the optimal choice for m, we notice that $E(V) \sim \alpha n$, i.e. the expected value of the time taken by the algorithm has only one main contributor - the set-up of the data structure. The other components, i.e. the traversal of the buckets, and the selection within one particular bucket, take expected time $\sim \sqrt{\beta\gamma n f^*}$ each. Since f^* is unknown, one could use $m \sim \sqrt{n}$ instead, without upsetting the expected time structure: we will still have $E(V) = \alpha n + 0(\sqrt{n}\,)$.

When f is not bounded, and / or m is not of the order of $\sqrt{n}$, the upper bound of Theorem 4.3 should still be useful in the majority of the cases. Recall the inequalities for q obtained in Remark 4.1.

4.3. NONLINEAR FUNCTIONS OF THE MAXIMAL CARDINALITY.

As we have seen in the study of the selection problem, and as we will see in section 4.4 (extremal point problems), it is important to derive the asymptotic behavior of

$$E(g(\alpha_n M_n))$$

where $\alpha_n \uparrow$ is a given sequence of positive integers (most often $\alpha_n \equiv 1$). $M_n = \max_{1 \le i \le m} N_i$, and $g(.)$ is a **work function** satisfying some regularity conditions. The following conditions will be assumed throughout this section:

(i) g is nonnegative and nondecreasing on $[0,\infty)$.

(ii) $g(x) > 0$ for $x > 0$

(iii) $g'(x) \le a + bx^s$ for some $a,b,s > 0$, all $x \ge 0$.

(iv) $\lim_{x \to \infty} g(x) = \infty$

(v) g is convex.

(vi) g is regularly varying at infinity, i.e. there exists a constant $\rho \ge 0$ such that for all $u \in R$,

$$\lim_{x \to \infty} \frac{g(ux)}{g(x)} = u^{\rho} .$$

Examples of such functions include

$$g(x) = x^2;$$

$$g(x) = x^r \ , r \ge 1;$$

$$g(x) = 1 + x \log(1+x).$$

For the properties of regularly varying functions, see Seneta (1976) and Dehaan (1975) for example.

The main result of this section is:

Theorem 4.4.

Let g be a work function satisfying (i-iv, vi), let $X_1, \ldots, X_n$ be iid random vectors with bounded density f on $[0,1]^d$, and let the grid have $m \sim cn$ buckets as $n \to \infty$ for some constant $c > 0$. Then, for α_n as given above,

$$E(g(\alpha_n M_n)) \le (1+o(1))\, g\left(\alpha_n \frac{\log(\alpha_n{}^{s+1}m)}{\log\log(\alpha_n{}^{s+1}m)}\right)$$

$$\sim g\left(\alpha_n \frac{\log(\alpha_n^{s+1}n)}{\log\log(\alpha_n^{s+1}n)}\right).$$

If in addition, $g(u) \ge b^* u^{s+1}$ for some $b^* > 0$, and all $u > 0$, then

$$E(g(\alpha_n M_n)) \le (1+o(1))\, g\left(\alpha_n \frac{\log n}{\log\log n}\right)$$

as $n \to \infty$.

If the work function satifies (i-ii, iv-vi), then

$$E(g(\alpha_n M_n)) \ge g\left(\alpha_n \frac{\log n}{\log\log n}\right)(1+o(1)).$$

If g satisfies (i-vi), $g(u) \ge b^* u^{s+1}$, some $b^* > 0$, all $u > 0$, then

$$E(g(\alpha_n M_n)) \sim g\left(\alpha_n \frac{\log n}{\log\log n}\right).$$

If the work function satisfies (i-vi), then

$$E(g(M_n)) \sim g\left(\frac{\log n}{\log\log n}\right).$$

Proof of Theorem 4.4.

Let us define

$$u = u_n = (1+\epsilon)\alpha_n \frac{\log(\alpha_n^{s+1}m)}{\log\log(\alpha_n^{s+1}m)}$$

where $\epsilon > 0$ is arbitrary. We always have

$$E(g(\alpha_n M_n)) \leq g(u) + \int_{g(u)}^{\infty} P(g(\alpha_n M_n) > t)\, dt$$

$$= g(u) + \int_{u}^{\infty} P(\alpha_n M_n > v)\, g'(v)\, dv$$

$$= g(u) + \int_{u/\alpha_n}^{\infty} P(M_n > v)\, g'(\alpha_n v)\, \alpha_n \, dv$$

$$\leq g(u) + \int_{u/\alpha_n}^{\infty} (a + b\, \alpha_n^s v^s)\, P(M_n > v)\, \alpha_n \, dv$$

$$\leq g(u) + \int_{u/\alpha_n}^{\infty} (a + b\, \alpha_n^s v^s)\, m e^{-\frac{n}{m} q} e^{-v \log(\frac{vm}{enq})} \alpha_n \, dv$$

by Lemma 5.5. If we can show that the integral is $o(1)$, then we have

$$E(g(\alpha_n M_n)) \leq g(u) + o(1)$$

$$\sim (1+\epsilon)^\rho \, g\left(\frac{\log(\alpha_n^{s+1} m)}{\log \log(\alpha_n^{s+1} m)} \right)$$

by conditions (iv) and (vi) on g. Since ϵ was arbitrary, we have shown the upper bound in the theorem. By convexity of g, the lower bound follows easily from theorem 4.1, Jensen's inequality and (vi):

$$E(g(\alpha_n M_n)) \geq g(\alpha_n E(M_n))$$

$$\sim g(\alpha_n \frac{\log n}{\log \log n}).$$

This leaves us with the proof of the statement that the second term is $o(1)$. Note that $q \leq f^*$, and that the bound of Lemma 5.5 remains valid if q is formally replaced by f^*. It suffices to show that

$$\int_{u/\alpha_n}^{\infty} \alpha_n^{s+1} v^s \, m \, e^{-\frac{n}{m} q} e^{-v \log(\frac{um}{e \alpha_n nq})} dv = o(1)$$

because $u/\alpha_n \uparrow \infty$. But the integral can be viewed as a tail-of-the gamma integral with respect to dv. Use $v^s \leq 2^{s-1}((\frac{u}{\alpha_n})^s + (v - (\frac{u}{\alpha_n}))^s)$, and $v = \frac{u}{\alpha_n} + (v - \frac{u}{\alpha_n})$ to obtain an upper bound of the form

$$\alpha_n m u^s 2^{s-1} e^{-\frac{u}{\alpha_n} \log(\frac{um}{e \alpha_n nq})} \cdot \frac{e^{-\frac{n}{m} q}}{\log(\frac{um}{e \alpha_n nq})}$$

$$+ m \alpha_n^{s+1} \frac{s!2^{s-1}}{(\log(\frac{um}{e \alpha_n nq}))^{s+1}} e^{-\frac{u}{\alpha_n} \log(\frac{um}{e \alpha_n nq}) - \frac{n}{m} q}.$$

The first of these two terms is asymptotically dominant. It is easily seen that the first term is

$$o\left(e^{\log(m \alpha_n^{s+1}) + s \log(\frac{u}{\alpha_n}) - \frac{n}{m} q - \frac{u}{\alpha_n} \log(\frac{um}{e \alpha_n nq})} \right).$$

Note that $\frac{m}{nq}$ remains bounded away from 0 and ∞. Trivial calculations show that for our choice of u, the last expression is $o(1)$.

Consider finally all the statements involving the condition $g(u) \geq b^* u^{s+1}$. It is clear that if the upper bounds for the integral are $o(g(u))$ instead of $o(1)$, then we are done. Thus, it suffices that the integrals are $o(u^{s+1})$, or $o(\alpha_n^{\ s+1})$. This follows if

$$\log m + s \log(\frac{u}{\alpha_n}) - \frac{n}{m} q - \frac{u}{\alpha_n} \log(\frac{um}{e \alpha_n nq}) \to -\infty ,$$

which is satisfied for $u = (1+\epsilon) \frac{\log n}{\log \log n}$.

Theorem 4.4 is useful because we can basically take the expected value inside g. Recall that by Jensen's inequality $E(g(M_n)) \geq g(E(M_n))$ whenever g is convex. The opposite inequality is provided in Theorem 4.4, i.e. $E(g(M_n))$ is $1+o(1)$ times larger than $g(E(M_n))$, mainly because M_n concentrates its probability mas near $E(M_n)$ as $n \to \infty$.

The conditions on g may appear to be a bit restrictive. Note however that all conditions are satisfied for most work functions found in practice. Furthermore, if g is sufficiently smooth, then $g'(x) \leq a + bx^s$ and $g(x) \geq b^* x^{s+1}$ can both be satisfied simultaneously.

A last word about Theorem 4.4. We have only treated bounded densities and grids of size $m \sim cn$. The reader should have no difficulty at all to generalize the techniques for use in other cases. For lower bounds, apply Jensen's inequality and lower bounds for $E(M_n)$, and for upper bounds, use the inequalities given in the proof of Theorem 4.4.

4.4. EXTREMAL POINT PROBLEMS.

Extremal point problems are problems that are concerned with the identification of a subset of $X_1, \ldots, X_n$ which in some sense defines the outer boundary of the "cloud" of points. The outer boundary is important in many application, such as:

(i) **pattern recognition:** discrimination rules can be based upon the relative position of a point with respect to the outer boundaries of the different classes (see e.g. Toussaint (1980, 1982)).

(ii) **image processing and computer vision:** objects are often characterized (stored) via the outer boundary.

(iii) **statistics:** points on the outer boundary of a collection of d-dimensional points can be considered as outliers, which need to be discarded before further analysis is carried out on the data.

(iv) **computational geometry:** The convex hull, one particularly simple outer boundary, plays a key role in various contexts in computational geometry. Often, information about the points can be derived from the convex hull (such as the diameter of the collection of points).

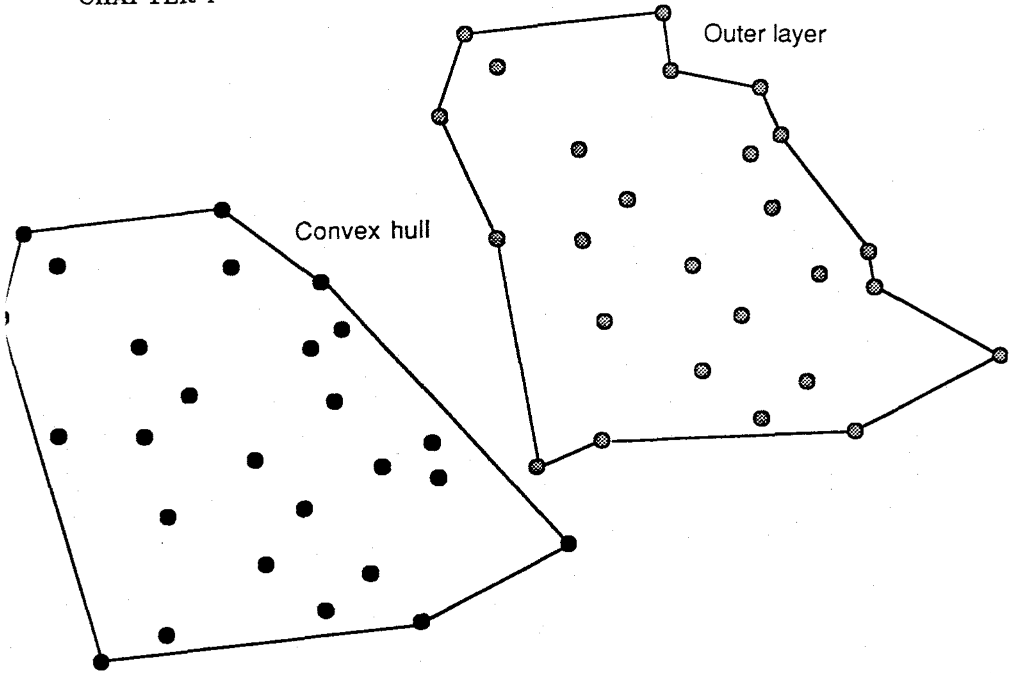


Figure 4.1.
The convex hull and the outer layer of a cloud of points.

We will refer in this short section to only two outer boundaries: the **convex hull** (the collection of all X_i' s having the property that at least one hyperplane through X_i puts all $n-1$ remaining points at the same side of the hyperplane), and the **outer layer**, also called the **set of maximal vectors** (the collection of all X_i' s having the property that at least one quadrant centered at X_i contains no X_j, $j \neq i$). Once again, we will assume that $X_1, \ldots, X_n$ have a common density f on $[0,1]^d$. A grid of size m is constructed in one of two ways, either by partitioning $[0,1]^d$ or by partitioning the smallest closed rectangle covering $X_1, \ldots, X_n$. The second grid is of course a data-dependent grid. We will go through the mechanics of reducing the analysis for the second grid to that of the first grid. The reduction is that given in Devroye (1981). For simplicity, we will consider only $d=2$.

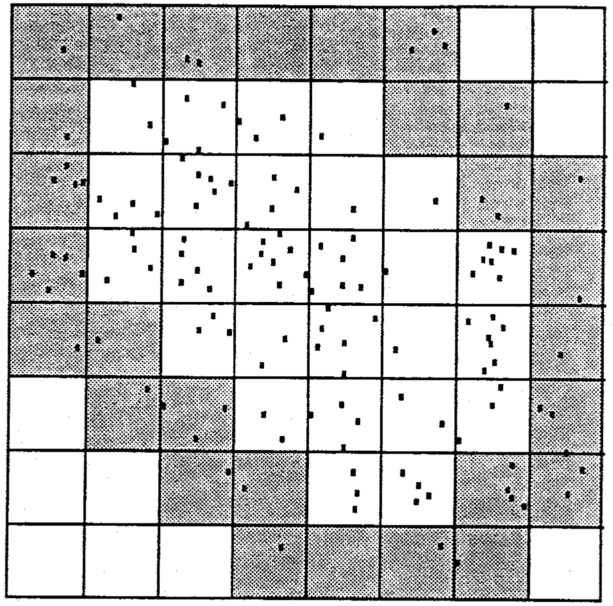


Figure 4.2.
Cell marking procedure.

For the outer layer in R^2, we find the leftmost nonempty column of rectangles, and mark the northernmost occupied rectangle in this column. Let its row number be j (row numbers increase when we go north). Having marked one or more cells in column i, we mark one or more cells in column $i+1$ as follows: (i) mark the cell at row number j, the highest row number marked up to that point; (ii) mark all rectangles between row number j and the northernmost occupied rectangle in column $i+1$ provided that its row number is at least $j+1$. In this manner a "staircase" of at most $2\sqrt{m}$ rectangles is marked. Also, any point that is a maximal vector for the north-west quadrant must be in a marked rectangle. We repeat this procedure for the three other quadrants so that eventually at most $8\sqrt{m}$ cells are marked. Collect all points in the marked cells, and find the outer layer by using standard algorithms. The naive method for example takes quadratic time (compare each point with all other points). One can do better by first sorting according to y-coordinates. In an extra pass through the sorted array, the outer layer is found by keeping only partial extrema in the x-direction. If heapsort or mergesort is used, the time taken to find the outer layer of n elements is $O(n \log n)$ in the worst-case.

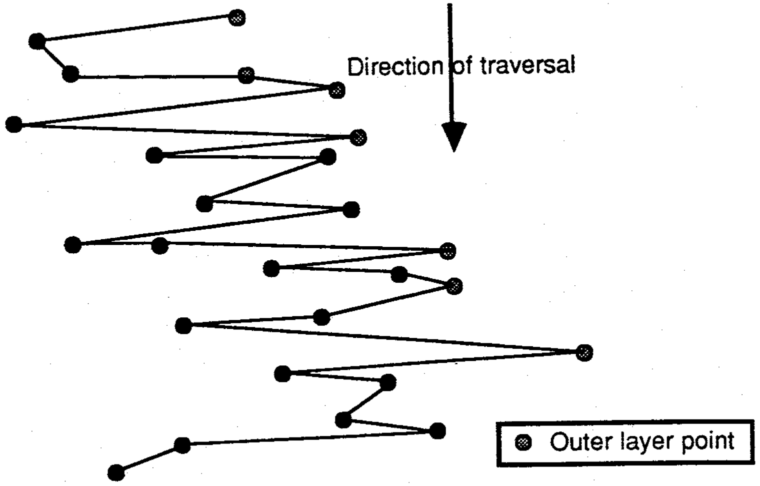


Figure 4.3.
Finding the outer layer points for the north-west quadrant.

Thus, returning to the data-independent grid, we see that the outer layer can be found in time bounded by

$$c_0 m + c_1 n + c_2 \left(\sum_{i \in B} N_i \right)^2$$

$$c_0 m + c_1 n + c_3 \sum_{i \in B} N_i \log(\sum_{i \in B} N_i + 1)$$

where $c_0, c_1, c_2, c_3 > 0$ are constants and B is the collection of indices of marked cells. The random component does not exceed $c_2 (8\sqrt{m} M_n)^2$ and $c_3 8\sqrt{m} M_n \log(1+8\sqrt{m} M_n)$ respectively. Clearly, these bounds are extremely crude. From Theorem 4.4, we recall that when $m \sim cn$, f is bounded, $E(M_n{}^2) \sim (\frac{\log n}{\log \log n})^2$, and $E(M_n \log(1+M_n)) \sim \log n$. Thus, the expected time is $O(n (\frac{\log n}{\log \log n})^2)$ in the former case, and $c_0 m + c_1 n + O(\sqrt{n} \log n)$

in the latter case. In the latter case, we observe that the contribution of the outer layer algorithm is asymptotically negligible compared to the contribution of the bucket data structure set-up. When we try to get rid of the boundedness condition on f, we could argue as follows: first of all, not much is lost by replacing $\log(\sum_{i \in B} N_i + 1)$ by $\log(n+1)$ because $\sum_{i \in B} N_i = \Omega(\sqrt{m})$ and $m \sim cn$. Thus,

$$
\begin{aligned}
&E(\sum_{i \in B} N_i \log(\sum_{i \in B} N_i + 1)) \\
&\leq E(\sum_{i \in B} N_i) \log(n+1) \\
&\leq 8\sqrt{m} \log(n+1) E(M_n) \\
&\leq 8\sqrt{m} \log(n+1) \left(\frac{\log m}{t} + \frac{n}{m} q \left(\frac{e^t - 1}{t}\right) \right) \quad (\text{all } t > 0)
\end{aligned}
$$

where $q = \max(mp_1, \ldots, mp_m)$ (Theorem 4.2). For constant t, we see that the upper bound is $o(n) + 8n \log(n+1) \frac{q}{\sqrt{m}} \frac{e^t - 1}{t}$. This is $O(n)$ for example when $q = O(\frac{\sqrt{n}}{\log n})$, $m \sim cn$. This is the case when

$$\int f^{2+\epsilon} < \infty$$

for some $\epsilon > 0$ (Remark 4.1). See however the important remark below.

Remark 4.6 [Optimization with respect fo m.]

We can once again tailor our grid to the problem by choosing m. Recall that an upper bound for the expected time complexity is $c_1 n + c_2 m + c_3 \sqrt{m} \log(n+1)(\frac{\log m}{t} + \frac{n}{m} q (\frac{e^t - 1}{t}))$ where $c_1, c_2, c_3, t > 0$ are constants. We can first choose t to approximately minimize the bound: for example, minimization of

$$\frac{\log m}{t} + \frac{n}{m} q \frac{t}{2}$$

suggests the value $t = \sqrt{\frac{2m \log m}{nq}}$, and we obtain

$$
\begin{aligned}
&c_1 n + c_2 m + c_3 \sqrt{m} \log(n+1) \left((2+o(1) \sqrt{\frac{nq \log m}{2m}} + \frac{n}{m} q \right) \\
&= c_1 n + c_2 m + c_3 \log(n+1)(\sqrt{2} + o(1)) \sqrt{nq \log m} \\
&+ c_3 \frac{n}{\sqrt{m}} q \log(n+1)
\end{aligned}
$$

if $\frac{m \log m}{nq} \to 0$. If we now minimize $c_2 m + c_3 \frac{n}{\sqrt{m}} q \log(n+1)$, we obtain the recipe

$$m = \left(\frac{c_3}{2c_2} \, . \, nq \log(n+1) \right)^{2/3} .$$

Plugging this back into our condition for the use of the bound, we note that it is satisfied in all cases since $nq \to \infty$. The bound becomes

$$
\begin{aligned}
&c_1 n + c_2^{1/3} c_3^{2/3} \left(\frac{1}{2^{2/3}} + 2^{1/3} \right) (nq \log(n+1))^{2/3} \\
&+ c_3 \left(\sqrt{\frac{4}{3}} + o(1) \right) \log n \sqrt{nq \log(nq)} .
\end{aligned}
$$

Which term is asymptotically dominant depends upon the density f. If f is bounded, then the upper bound is $c_1 n + (K + o(1)) f^{*2/3} (n \log n)^{2/3}$ where K does not depend upon f and f^* is the bound for f. We can also design the grid for a particular class of densities. For bounded densities, we can take

$$m = \left(\frac{c_3}{2c_2} n f^* \log n \right)^{2/3} ,$$

and for densities with $\mu_r = (\int f^r)^{1/r} < \infty$, we can take

$$m = \left(\frac{c_3}{2c_2} n \; m^{\frac{1}{r}} \mu_r \; \log n \right)^{2/3} ,$$

or, solving for m :

$$m = \left(\frac{c_3}{2c_2} n \; \mu_r \; \log n \right)^{\frac{2r}{3r-2}} .$$

This yields useful choices for $r > 2$. Using $q \leq \mu_r m^{1/r}$, we obtain the further bound

$$c_1 n + O((n \log n)^{\frac{2r}{3r-2}}) .$$

The main conclusion is that if m is growing slower than n , then for certain large classes of densities, the asymptotically most important component in the expected time complexity is $c_1 n$. For example, when $\int f^4 < \infty$, we have $c_1 n + 0((n \log n)^{4/5})$.

Of course, the same algorithm and discussion can be used for finding the convex hull of $X_1, \ldots, X_n$ because for arbitrary points there exist simple $O(n \log n)$ and $O(n^2)$ worst-case algorithms (see Graham (1972) , Shamos (1978), Preparata and Hong (1977) and Jarvis (1973)) and all convex hull points are outer layer points. In this form, the algorithm was suggested by Shamos (1979).

Remark 4.7. [Bucket structure in polar coordinates.]

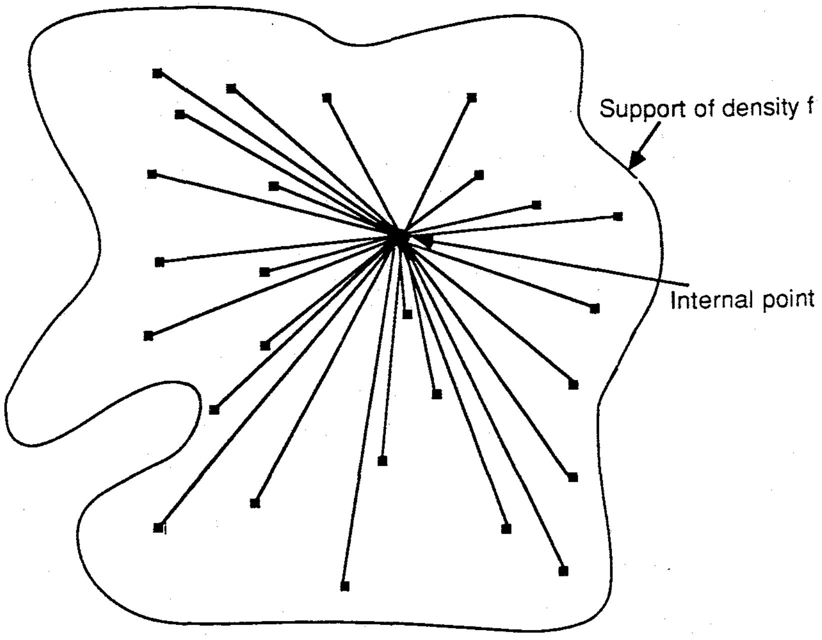


Figure 4.4.
Points are ordered according to angular coordinates
for use in Graham's convex hull algorithm , bucket algorithm.

The bucket data structure can be employed in unexpected ways. For example, to find the convex hulls in R^2, it suffices to transform $X_1 - x, \ldots, X_n - x$ into polar coordinates where x is a point known to belong to the interior of convex hull of $X_1, \ldots, X_n$ (note: we can always take $X = X_1$). The points are sorted according to polar angles by a bucket sort as described in chapter 2. This yields a polygon P. All vertices of P are visited in clockwise fashion and pushed on a stack. The stack is popped when a non-convex-hull point is identified. In this manner, we can construct the convex hull from P in linear time. The stack algorithm is based upon ideas first developed by Graham (1972). It is clear that the expected time of the convex hull algorithm is $O(n)$ if $\int g^2 < \infty$ or $\int g \log_+ g < \infty$ where g is the density of the polar angle of $X_i - x$, $i \geq 1$. For example, when $X_1, \ldots, X_n$ have a radially symmetric density f , and x is taken to be the origin, the g is the uniform density on $[0,2\pi]$, and the algorithm takes $O(n)$ expected time. When x itself is a random vector, one must be careful before concluding anything about the finiteness of $\int g^2$. In any case, g is bounded whenever f is bounded and has compact support.

The results about $E(M_n)$, albeit very helpful, lead sometimes to rather crude upper bounds. Some improvement is possible along the lines of Theorem 4.5 (Devroye, 1985).

Theorem 4.5.

Let $X_1, \ldots, X_n$ be independent random vectors with common density f on $[0,1]^2$, let the grid have m cells, and let $q = \max(mp_1, \ldots, mp_m)$. Then, if B is the collection of indices of marked cell in the extremal cell marking algorithm,

$$E(\sum_{i \in B} N_i) \le 8\sqrt{m} \; \frac{\frac{n}{m} q}{1-e^{-\frac{n}{m}q}} .$$

In particular, if $m \sim cn$ (for some constant $c > 0$),

$$E(\sum_{i \in B} N_i) \le (8+o(1)) \, \frac{\sqrt{\frac{n}{c} q}}{1-e^{-\frac{1}{c}}}$$

and

$$E(\sum_{i \in B} N_i) \le \frac{(8+o(1))}{1-e^{-\frac{1}{c}}} \, n^{\frac{1}{2}+\frac{1}{r}} \, c^{\frac{1}{r}+\frac{1}{2}} (\int f^r)^{\frac{1}{r}}$$

for all $r \ge 1$.

Proof of Theorem 4.5.

We note that each N_i is stochastically smaller than a binomial (n, p_i) random variable conditioned on the variable being at least 1. Thus,

$$E(N_i) \le \frac{np_i}{1-(1-p_i)^n} \le \frac{np_i}{1-e^{-np_i}} \le \frac{\frac{n}{m} q}{1-e^{-\frac{n}{m}q}} .$$

The first inequality follows trivially from this. The second inequality is obvious, and the third inequality is based upon the fact that $q \le m^{1/r}(\int f^r)^{1/r}$.

In the proof of Theorem 4.5, we have not used the obvious inequality $\sum_{i \in B} N_i \le 8\sqrt{m} M_n$. If we find the outer layer or the convex hull by an $O(n \log n)$ worst-case time method, then under the conditions of Theorem 4.5, with $m \sim cn$, the expected time is bounded by

$$O(n) + O(\sqrt{n} \; q) \log n$$

and this does not improve over the bound obtained when the crude inequality was used. For example, we cannot guarantee linear expected time behavior when $\int f^2 < \infty$, but only when a stronger condition such as $\int f^{2+\epsilon} < \infty$ (some $\epsilon > 0$) holds. (We can of course always work on m, see remark 4.6).

There is, however, a further possible improvement along the lines of an outer layer algorithm of Machii and Igarashi (1984). Here we either find the outer layers in all cells A_i, $i \in B$, or sort all points in the individual cells. Then, in another step, the outer layer can be found in time linear in the number of points to be processed. Thus, there are three components in the time complexity: $n+m$ (set-up), $\sum_{i \in B} N_i \log(N_i+1)$ (or $\sum_{i \in B} N_i^{\,2}$) (sorting), and $\sum_{i \in B} N_i$ (final outer layer). It should be clear that a similar strategy works too for the convex hull. The principle is well-known: divide-and-conquer. It is better to delegate the work to the individual buckets, in other words. For example, we always have

$$\begin{aligned} & E(\sum_{i \in B} N_i \log(N_i + 1)) \\ & \le 8\sqrt{m} \; E(M_n \log(M_n + 1)) \\ & \le 8\sqrt{m} \; \log(n+1) \, E(M_n) , \end{aligned}$$

and, if we use a more refined bound from the proof of Theorem 4.5 combined with Lemma 5.6,

$$E(\sum_{i \in B} N_i \log(N_i + 1))$$

$$\le 8\sqrt{m}\,\frac{\frac{n}{m}q\ \log(2+\frac{n}{m}q\,)}{1-e^{-\frac{n}{m}q}}\ .$$

For example, when $m \to \infty$, $n/m \to \infty$, $f \le f^* < \infty$, the bound is

$$\sim \frac{8n}{\sqrt{m}}\ f^*\ \log(\frac{n}{m})\ .$$

The optimal choice for m is proportional to $(f^* n \log n)^{2/3}$, so that the expected time complexity for the algorithm is $c_1 n$ (for the set-up) $+\ O((n \log n)^{2/3})$. In another example, if $m \sim cn$, $q \to \infty$, the upper bound is

$$\sim \sqrt{n}\ \frac{8}{\sqrt{c}}\ q\ \log q\ ,$$

which in turn is $O(n)$ when $q = O(\sqrt{n}/\log n)$.

We turn now to the problem of data-dependent grids, and in particular grids of size m formed by partitioning the smallest closed rectangle covering all the points. For the convex hull and outer layer algorithms considered here, the random terms are either

$$\sum_{i \in B} N_i \ \log(N_i + 1)$$

or

$$\sum_{i \in B} N_i^{\,2}$$

if divide-and-conquer is used, and

$$(\sum_{i \in B} N_i)\ \log(\sum_{i \in B} N_i + 1)\ .$$

or

$$(\sum_{i \in B} N_i)^2$$

otherwise. All these terms are bounded from above by $g(\alpha_n M_n)$ where α_n is an integer, g is a work function and $M_n = \max_{1 \le i \le m} N_i$. Unfortunately, our analysis of M_n and $g(M_n)$ does not apply here because the grid is data-dependent. The dependence is very weak though, and nearly all the results given in this section remain valid if f has rectangular support $[0,1]^2$. (Note: the rectangular support of f is the smallest rectangle R with the property that $\int_R f = 1$.) To keep things simple, we will only be concerned with an upper bound for $E(g(\alpha_n M_n))$ that is of the correct order of increase in n - in other words, we will not be concerned with the asymptotic constant. This case can easily be dealt with via a "shifted grid" argument (Devroye, 1981). Partition $[0,1]^2$ (or $[0,1]^d$ for that matter) into a grid of size $m/2^d$ with member cells B_i. Then consider for each $(j_1, \ldots, j_d) \in \{0,1\}^d$ the shifted grid with member cells $B_i(j_1, \ldots, j_d)$, $1 \le i \le \frac{m}{2d}$, where the shift vector is

$$\left(\frac{j_1}{m^{1/d}}, \frac{j_2}{m^{1/d}}, \ldots, \frac{j_d}{m^{1/d}} \right) .$$

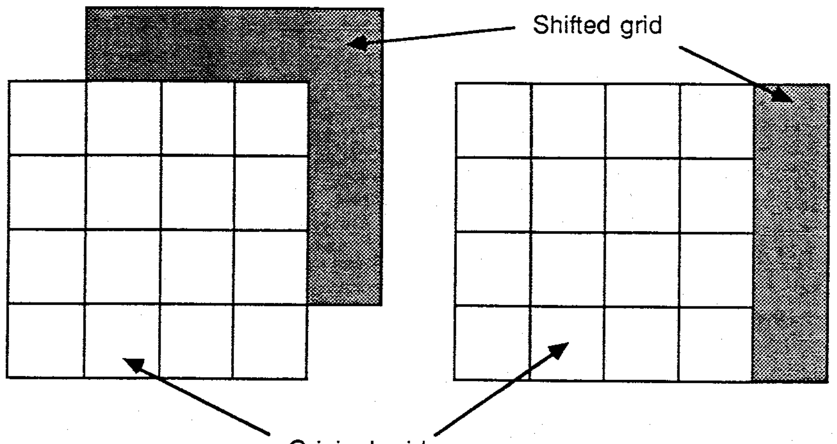


Figure 4.5.
Illustration of the shifted grid argument.

The key observation is that every A_i in the original data-dependent grid is contained in some $B_k(j_1, \ldots, j_d)$. Thus,

$$M_n \le \max_{j_1, \ldots, j_d} M_n^*(j_1, \ldots, j_d)$$

where $M_n^*(j_1, \ldots, j_d)$ is the maximal cardinality for the $(j_1, \ldots, j_d)$ grid. Thus,

$$g(\alpha_n M_n) \le \sum_{j_1, \ldots, j_d} g(\alpha_n M_n^*(j_1, \ldots, j_d))$$

Each individual term on the right hand side is for a data-independent grid, for which we can derive several types of inequalities. Thus, typically, the expected value of the right hand side is about 2^d times the expected value of one term. For example, if f is bounded and $m \sim cn$, then for α_n, g as in Theorem 4.4, the expected value of the right hand side is $\le (1+o(1))2^d \; g(\alpha_n \frac{\log n}{\log \log n})$.

Chapter 5

AUXILIARY RESULTS FROM PROBABILITY THEORY

5.1. PROPERTIES OF THE MULTINOMIAL DISTRIBUTION.

A random vector $(Y_1, \ldots, Y_k)$ is multinomial $(n; p_1, \ldots, p_k)$ when

$$P(Y_1 = i_1, \ldots, Y_k = i_k) = n! \prod_{j=1}^{k} \frac{p_j^{\,i_j}}{i_j!} ,$$
$$i_1 + \cdots + i_k = n, i_j \ge 0, \text{ all } j,$$

where $\sum_{j=1}^{k} p_j = 1$ and all $p_j' s$ are nonnegative. Y_1 is said to be binomial (n, p_1).

Lemma 5.1. [Moments of the multinomial distribution; see e.g. Johnson and Kotz, 1969]

For integer $r, s \ge 1$:

$$E(Y_i(Y_i-1) \cdots (Y_i-r+1)) = p_i^{\,r} \; n(n-1) \cdots (n-r+1) ,$$

$$E(Y_i(Y_i-1) \cdots (Y_i-r+1)Y_j(Y_j-1)...(Y_j-s+1))$$

$$= p_i^{\,r} p_j^{\,s} n(n-1)...(n-r-s+1), \; i \ne j.$$

Thus,

$$E(Y_i) = np_i \ , E(Y_i^2) = np_i + n(n-1)p_i^2 \ ,$$

$$E(Y_i^3) = np_i + 3n(n-1)p_i^2 + n(n-1)(n-2)p_i^3 \ ,$$

$$E(Y_i^4) = np_i + 7n(n-1)p_i^2 + 6n(n-1)(n-2)p_i^2 + n(n-1)(n-2)(n-3)p_i^4 \ ,$$

and for $i \neq j$,

$$E(Y_i Y_j) = n(n-1)p_i p_j \ ,$$

$$E(Y_i(Y_i-1)Y_j(Y_j-1)) = n(n-1)(n-2)(n-3)p_i^2 p_j^2 \ ,$$

and

$$E(Y_i^2 Y_j^2) = n(n-1)(n-2)(n-3)p_i^2 p_j^2$$

$$+ n(n-1)(n-2)(p_i p_j^2 + p_i^2 p_j) + n(n-1)p_i p_j \ .$$

Lemma 5.2. [Moment generating function of the multinomial distribution.]

The random vector $Y_1, \ldots, Y_k$ has moment generating function

$$E(\exp(\sum_{j=1}^{k} t_j Y_j)) = (\sum_{j=1}^{k} p_j \exp(t_j))^n \ .$$

Lemma 5.3. [Uniform bounds for the moments of a binomial random variable.]

If Y is binomial (n,p) and $r > 0$ is a constant, then there exist $a,b > 0$ only depending upon r such that

$$E(Y^r) \leq a(np)^r + b \ .$$

Proof of Lemma 5.3.

When $r \leq 1$, we have $E(Y^r) \leq (np)^r$, by Jensen's inequality. We will thus assume that $r > 1$. Anderson and Samuels (1965) have shown that for all $k \geq np+1$, $P(Y \geq k) \leq P(Z \geq k)$ where Z is a Poisson (np) random variable. Thus,

$$E(Y^r) \leq (np+1)^r + E(Y^r I_{Y \geq np+1}) \leq (np+1)^r + E(Z^r I_{Z \geq np+1})$$

$$\leq (np+1)^r + (np+r)^r + \sum_{k > np+r} k^r \frac{(np)^k}{k!} e^{-np} \ .$$

Because $(u+v)^r \leq 2^{r-1}(u^r+v^r)$, the first two terms in the last sum are not greater than $a(np)^r + b$ for some constants a,b only depending upon r. The last sum can be bounded from above by

$$\sum_{k > np+r} (\frac{k}{k-r})^r \frac{(np)^{k-r}}{(k-r)!} e^{-np} (np)^r \ .$$

Assume that $np \geq 1$. Then this is not greater than

$$(np)^r (1 + \frac{r}{np})^r \leq (1+r)^r (np)^r \ .$$

For $np \leq 1$, we have $E(Y^r) \leq 2^r + E(Z^r)$ where Z is Poisson (1). This concludes the proof of Lemma 5.3.

Lemma 5.4.

Let $g(u)$ be a nonnegative nondecreasing function on $[0,\infty)$, and let Y be binomial (n,p). Then if $g(u) = 0$ on $(-\infty,0)$,

$$E(g(Y)) \geq \frac{1}{2} g(np - \sqrt{np}\ ).$$

If $p \in [0,\frac{1}{4}]$, we have $E(g(Y)) \geq \frac{1}{2} g(\lfloor np \rfloor\ )$.

If also $g(u)/u^k \downarrow$ as $u \to \infty$ for some finite constant k, then

$$E(g(Y)) \leq a \max(g(np), g(1))$$

for some finite constant a depending upon k only.

Proof of Lemma 5.4.

For any $t \geq 0$, we have $E(g(Y)) \geq g(t) P(Y \geq t)$. Now, by the Chebyshev-Cantelli inequality,

$$P(Y \leq np - \sqrt{np(1-p)}) = P\left(\frac{Y-np}{\sqrt{np(1-p)}} \leq -1\right) \leq \frac{1}{1+1} = \frac{1}{2} .$$

Thus,

$$E(g(Y)) \geq \frac{1}{2} g(np - \sqrt{np(1-p)}) \geq \frac{1}{2} g(np - \sqrt{np}).$$

The second inequality follows directly from Theorem 2.1 in Slud (1977). Next,

$$E(g(Y)) = E(g(Y) I_{Y \leq np}) + E(g(Y) I_{Y > np})$$

$$\leq g(np) + E((g(Y)/Y^k)\, Y^k I_{Y > np})$$

$$\leq g(np) + \frac{g(np)}{(np)^k} E(Y^k)$$

$$\leq g(np) + g(np)a + bg(np)/(np)^k$$

where a, b are the constants of Lemma 5.3. If $np \geq 1$, the last sum is not greater than $g(np)(1+a+b)$. If $np \leq 1$, we have $E(g(Y)) \leq E(g(Z)) \leq g(1)(1+a+b)$ where Z is a binomial $(n, \frac{1}{n})$ random variable. This concludes the proof of Lemma 5.4.

Lemma 5.5. [Maximum of a multinomial random vector.]

Let B be a binomial (n,p) random variable. Then, for arbitrary $x > 0$,

$$P(B \geq x) \leq e^{-np} \left(\frac{enp}{x}\right)^x .$$

If $N_1, \ldots, N_m$ is mutinomial $(n; p_1, \ldots, p_m)$, and $x \geq q = \max(mp_1, \ldots, mp_m)$, then

$$P\left(\max_{1 \leq i \leq m} N_i \geq x\right) \leq m e^{-\frac{n}{m} q} \left(\frac{enq}{mx}\right)^x .$$

Proof of Lemma 5.5.

For the first part, we use Chernoff's bounding method (Chernoff, 1952) and note that for any $t > 0$:

$$P(B \geq x) \leq e^{-tx} E(e^{tB}) = (e^t p + 1-p)^n \; e^{-tx}$$

$$\leq e^{-tx + np(e^t - 1)}$$

$$= e^{-\log(\frac{x}{np})x + np(\frac{x}{np} - 1)}$$

where we took $e^t = \frac{x}{np}$, since this choice minimizes the upper bound. Note that the upper bound remains valid when B is binomial (n, p'), $p' \leq p$. For the multinomial distribution, we apply Bonferroni's inequality.

Lemma 5.6. [Logarithmic moment of the binomial distribution.]

Let Y be binomial (n,p). Then

$$E(Y \log(1+Y)) \leq np \log(2+np).$$

Proof of Lemma 5.6.

Let Z be a binomial $(n-1,p)$ random variable. Then, by Jensen's inequality,

$$E(Y\ \log(1+Y)) = \sum_{i=1}^{n} \binom{n}{i}\ i\ \log(i+1)\ p^{i}(1-p)^{n-i}$$

$$= \sum_{i=1}^{n} (np) \binom{n-1}{i-1}\ p^{i-1}(1-p)^{(n-1)-(i-1)} \log(i+1)$$

$$= np\ \ E(\log(Z+2))$$

$$\leq np\ \ \log(E(Z)+2)$$

$$\leq np\ \ \log(np+2).$$

5.2. PROPERTIES OF THE POISSON DISTRIBUTION.

Lemma 5.7. [Exponential inequality for the Poisson tail.]

If Y is Poisson (λ) distributed, then

$$P(|Y-\lambda| \geq \lambda\epsilon) \leq 2\ \exp(-\lambda\epsilon^2/2(1+\epsilon))\ ,\ \text{all}\ \epsilon > 0.$$

Proof of Lemma 5.7.

By Chernoff's bounding technique, we have

$$P(|Y-\lambda| \geq \lambda\epsilon) \leq E(e^{t(Y-\lambda)}+e^{-t(Y-\lambda)})e^{-t\lambda\epsilon}\ ,\ \text{all}\ t > 0,$$

$$= e(e^{\lambda(e^{t}-1-t)}+e^{\lambda(e^{-t}-1+t)})e^{-t\lambda\epsilon}$$

$$= e^{\lambda(e^{t}-1-t)}e^{-t\lambda\epsilon}(1+e^{\lambda(e^{-t}-1+t-e^{t}+1+t)})$$

$$\leq 2\ e^{\lambda(e^{t}-1-t-t\epsilon)}$$

where we used the fact that $e^{-t} \leq e^{t}-2t$. The exponent $e^{t}-1-t(1+\epsilon)$ is

minimal if we take $t = \log(1+\epsilon)$, and this gives the bound

$$2\ \exp(\lambda(\epsilon-(1+\epsilon)\log(1+\epsilon))) \leq 2\ \exp(-\lambda\epsilon^2/2(1+\epsilon))\ .$$

Here we used the Taylor's series with remainder term to obtain the last inequality.

Lemma 5.8. [Fourth moment inequality for the Poisson tail.]

If Y is Poisson (λ) distributed, then

$$P(|Y-\lambda| \geq \lambda\epsilon) \leq \frac{4}{\lambda^2\epsilon^4}\ ,\ \text{all}\ \epsilon > 0\ .$$

Proof of Lemma 5.8.

By Chebyshev's inequality,

$$P(|Y-\lambda| \geq \lambda\epsilon) \leq \frac{E(|Y-\lambda|^4)}{(\lambda\epsilon)^4}$$

$$= \frac{\lambda+3\lambda^2}{\lambda^4\epsilon^4} \leq \frac{4}{\lambda^2\epsilon^4}\ .$$

Lemma 5.9. [Precise estimates of the Poisson tail.]

Let Y be a Poisson (λ) random variable, and let k be a fixed integer. Then, for $k+1 > \lambda$,

$$1 \leq \frac{P(Y \geq k)}{P(Y=k)} \leq \frac{k+1}{k+1-\lambda}$$

Proof of Lemma 5.9.

Observe that

$$\sum_{j \geq k} \lambda^j \frac{e^{-\lambda}}{j\,!} \leq \lambda^k \frac{e^{-\lambda}}{\lambda!} \sum_{j=0}^{\infty} (\frac{\lambda}{k+1})^j .$$

5.3. THE LEBESGUE DENSITY THEOREM.

In this section we give several forms of the Lebesgue density theorem, that will enable us to obtain theorems without continuity conditions on f. For proofs and additional details, we refer to Wheeden and Zygmund (1977) and to de Guzman (1975, 1981).

Lemma 5.10.

Let $\mathbf{A}$ be the class of all rectangles containing the origin of R^d, and with sides $s_1, \ldots, s_d$ satisfying $a_i \leq s_i \leq b_i$ for some fixed positive numbers $a_i \leq b_i$, $1 \leq i \leq d$.

There exists a set $D \subseteq R^d$ such that $\lambda(D^c) = 0$ (D^c is the complement of D) and

$$\sup_{A \in \mathbf{A}} \mid \int_{x+rA} f \,/\lambda(x+rA) - f(x) \mid \to 0 \text{ as } r \to 0, \text{ all } x \in D.$$

Proof of Lemma 5.10.

See Wheeden and Zygmund (1977) or de Guzman (1975, 1981).

Lemma 5.11.

Let C be a fixed rectangle of R^d with sides $c_1, \ldots, c_d$. Let $\{A_n\}$ be a sequence of rectangles tending to C as $n \to \infty$. Let $\mathbf{A}_n$ be the collection of all translates of A_n that cover the origin. Then, for any sequence of positive numbers $r_n \downarrow 0$,

$$\lim_{n\to\infty} \sup_{A \in \mathbf{A}_n} | \int_{x+r_n A} f \,/\lambda(x+r_n A) - f(x)| = 0 \text{ , almost all } x.$$

The set on which the convergence takes place does not depend upon the choice of the sequences A_n and r_n.

Lemma 5.12. [Scheffe's theorem (1947).]

Let f_n be a sequence of densities converging almost everywhere to a density f on R^d. Then

$$\int | f_n - f \mid \to 0$$

as $n \to \infty$.

Proof of Lemma 5.12..

Note that

$$\int | f_n - f \mid = 2 \int (f - f_n)_+ \to 0 \text{ ,}$$

where we used the almost everywhere convergence of f_n to f and the Lebesgue dominated convergence theorem.

REFERENCES

- A.V. Aho, J.E. Hopcroft, and J.D. Ullman, *Data Structures and Algorithms,* Addison-Wesley, Reading, Mass. (1983).
- S.G. Akl and H. Meijer, "Recent advances in hybrid sorting algorithms," *Utilitas Mathematica* **21** pp. 325-343 (1982).
- S.G. Akl and H. Meijer, "On the average-case complexity of bucketing algorithms," *Journal of Algorithms* **3** pp. 9-13 (1982).
- D.C.S. Allison and M.T. Noga, "Selection by distributive partitioning," *Information Processing Letters* **11** pp. 7-8 (1980).
- T.W. Anderson and S.M. Samuels, "Some inequalities among binomial and Poisson probabilities," pp. 1-12 in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press (1965).
- T. Asano, M. Edahiro, H. Imai, M. Iri, and K. Murota, "Practical use of bucketing techniques in computational geometry," pp. 0-0 in *Computational Geometry,* ed. G.T. Toussaint,North-Holland (1985).
- J. Beardwood, J.H. Halton, and J.M. Hammersley, "The shortest path through many points," *Proceedings of the Cambridge Philosophical Society* **55** pp. 299-327 (1959).
- R. Bellman, "Dynamic programming treatment of the travelling salesman problem," *Journal of the ACM* **9** pp. 61-63 (1962).
- J.L. Bentley, "Solutions to Klee's rectangle problems," Technical Report, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. (1977).
- J.L. Bentley, D.F. Stanat, and E.H. Williams, "The complexity of fixed-radius near neighbor searching," *Information Processing Letters* **6** pp. 209-212 (1977).
- J.L. Bentley and J.H. Friedman, "Data structures for range searching," *ACM Computing Surveys* **11** pp. 397-409 (1979).
- J.L. Bentley, B.W. Weide, and A.C. Yao, "Optimal expected-time algorithms for closest point problems," *ACM Transactions on Mathematical Software* **6** pp. 563-580 (1980).

- J.L. Bentley and D. Wood, "An optimal worst-case algorithm for reporting intersections of rectangles," *IEEE Transactions on Computers* **C-29** pp. 571-577 (1980).
- M. Blum, R.W. Floyd, V. Pratt, R.L. Rivest, and R.E. Tarjan, "Time bounds for selection," *Journal of Computers and System Sciences* **7** pp. 448-461 (1973).
- D. Cheriton and R.E. Tarjan, "Finding minimum spanning trees," *SIAM Journal on Computing* **5** pp. 724-742 (1976).
- H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics* **23** pp. 493-507 (1952).
- Y.S. Chow and H. Teicher, *Probability Theory,* Springer-Verlag, New York, N.Y. (1978).
- N. Christofides, "Worst-case analysis of a new heuristic for the traveling salesman problem," Symposium on Algorithms and Complexity, Department of Computer Science, Carnegie-Mellon University (1976).
- L. Dehaan, "On Regular Variation and its Application to the Weak Convergence of Sample Extremes," Mathematical Centre Tracts 32, Mathematisch Centrum, Amsterdam (1975).
- L. Devroye and T. Klincsek, "Average time behavior of distributive sorting algorithms," *Computing* **26** pp. 1-7 (1981).
- L. Devroye, "On the average complexity of some bucketing algorithms," *Computers and Mathematics with Applications* **7** pp. 407-412 (1981).
- L. Devroye, "On the expected time required to construct the outer layer of a set of points," *Information Processing Letters* **0** pp. 0-0 (1985).
- L. Devroye, "Expected time analysis of algorithms in computational geometry : a survey," pp. 0-0 in *Computational Geometry,* ed. G.T. Toussaint,North-Holland (1985).
- L. Devroye, "The expected length of the longest probe sequence when the distribution is not uniform," *Journal of Algorithms* **6** pp. 1-9 (1985).
- L. Devroye and F. Machell, "Data structures in kernel density estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **7** pp. 360-366 (1985).
- D. Dobkin and R.J. Lipton, "Multidimensional searching problems," *SIAM Journal on Computing* **5** pp. 181-186 (1976).
- W. Dobosiewicz, "Sorting by distributive partitioning," *Information Processing Letters* **7** pp. 1-6 (1978).

- M. Edahiro, I. Kokubo, and T. Asano, "A new point-location algorithm and its practical efficiency — comparison with existing algorithms," Research memorandum RMI 83-04, Department of Mathematical Engineering and Instrumentation Physics, University of Tokyo (1983).
- G. Ehrlich, "Searching and sorting real numbers," *Journal of Algorithms* **2** pp. 1-12 (1982).
- J.D. Esary, F. Proschan, and D.W. Walkup, "Association of random variables, with applications," *Annals of Mathematical Statistics* **38** pp. 1466-1474 (1967).
- 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** pp. 315-344 (1979).
- P. Flajolet, "On the performance evaluation of extendible hashing and trie search," *Acta Informatica* **20** pp. 345-369 (1983).
- R.W. Floyd and R.L. Rivest, "Expected time bounds for selection," *Communications of the ACM* **18** pp. 165-172 (1975).
- S. Fortune and J. Hopcroft, "A note on Rabin's nearest-neighbor algorithm," *Information Processing Letters* **8** pp. 20-23 (1979).
- J. Galambos, *The Asymptotic Theory of Extreme Order Statistics,* John Wiley, New York, N.Y. (1978).
- J. Geffroy, "Contributions a la theorie des valeurs extremes," *Publications de l'ISUP* **7** pp. 37-185 (1958).
- B.V. Gnedenko, "Sur la distribution du terme maximum d'une serie aleatoire," *Annals of Mathematics* **44** pp. 423-453 (1943).
- G.H. Gonnet, "Expected length of the longest probe sequence in hash code searching," *Journal of the ACM* **28** pp. 289-304 (1981).
- G.H. Gonnet, *A Handbook of Algorithms and Data Structures,* Addison-Wesley, Reading, Mass. (1984).
- T. Gonzalez, "Algorithms on sets and related problems," Technical Report, Department of Computer Science, University of Oklahoma (1975).
- R. Graham, "An efficient algorithm for determining the convex hull of a finite planar set," *Information Processing Letters* **1** pp. 132-133 (1972).
- J. Gurland, "Inequalities for expectations of random variables derived by monotonicity or convexity," *The American Statistician* **22** pp. 26-27 (1968).
- M. de Guzman, "Differentiation of integrals in R^n," Springer Lecture Notes in Mathematics # 481, Springer-Verlag, Berlin (1975).

- M. de Guzman, "Real Variable Methods in Fourier Analysis," North-Holland Mathematical Studies #46, North-Holland, Amsterdam (1981).
- J.H. Halton and R. Terada, "A fast algorithm for the Euclidean traveling salesman problem, optimal with probability one," *SIAM Journal on Computing* **11** pp. 28-46 (1982).
- J.A. Hartigan, *Clustering Algorithms,* John Wiley, New York, N.Y. (1975).
- C.A.R. Hoare, "Find (algorithm 65)," *Communications of the ACM* **4** pp. 321-322 (1961).
- M. Iri, K. Murota, and S. Matsui, "Linear-time approximation algorithms for finding the minimum-weight perfect matching on a plane," *Information Processing Letters* **12** pp. 206-209 (1981).
- M. Iri, K. Murota, and S. Matsui, "Heuristics for planar minimum-weight perfect matching," *Networks* **13** pp. 67-92 (1983).
- R.A. Jarvis, "On the identification of the convex hull of a finite set of points in the plane," *Information Processing Letters* **2** pp. 18-21 (1973).
- N.L. Johnson and S. Kotz, *Urn Models and Their Application,* John Wiley, New York, N.Y. (1977).
- R.M. Karp, "Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane," *Mathematics of Operations Research* **2** pp. 209-224 (1977).
- D. Kirkpatrick, "Optimal search in planar subdivisions," *SIAM Journal on Computing* **12** pp. 28-35 (1983).
- D.E. Knuth, *The Art of Computer Programming, Vol. 1: Fundamental Algorithms,* Addison-Wesley, Reading, Mass. (1973). 2nd Ed.
- D.E. Knuth, *The Art of Computer Programming, Vol. 3 : Sorting and Searching,* Addison-Wesley, Reading, Mass. (1973).
- V.F. Kolchin, B.A. Sevastyanov, and V.P. Chistyakov, *Random Allocations,* V.H. Winston and Sons, Washington, D.C. (1978).
- P.-A. Larson, "Expected worst-case performance of hash files," *The Computer Journal* **25** pp. 347-352 (1982).
- D.T. Lee and F.P. Preparata, "Location of a point in a planar subdivision and its applications," *SIAM Journal on Computing* **6** pp. 594-606 (1977).
- E.L. Lehmann, "Some concepts of dependence," *Annals of Mathematical Statistics* **37** pp. 137-1153 (1966).
- R.J. Lipton and R.E. Tarjan, "Applications of a planar separator theorem," pp. 162-170 in *Proceedings of the 18th IEEE Symposium on the Foundations of Computer Science,* (1977).

- M. Loeve, *Probability Theory*, Van Nostrand, Princeton, New Jersey (1963).
- M. Machii and Y. Igarashi, "A hashing method of finding the maxima of a set of vectors," Technical Report CS-84-2, Department of Computer Science, Gunma University, Gunma, Japan (1984).
- M.D. Maclaren, "Internal sorting by radix plus sifting," *Journal of the ACM* **13** pp. 404-411 (1966).
- C.L. Mallows, "An inequality involving multinomial probabilities," *Biometrika* **55** pp. 422-424 (1968).
- H. Meijer and S.G. Akl, "The design and analysis of a new hybrid sorting algorithm," *Information Processing Letters* **10** pp. 213-218 (1980).
- G. Nagy and S. Wagle, "Geographic data processing," *Computing Surveys* **11** pp. 139-181 (1979).
- C.H. Papadimitriou and K. Steiglitz, "Some complexity results for the traveling salesman problem," pp. 1-9 in *Eighth Annual ACM SIGACT Symposium*, (1976).
- C.H. Papadimitriou, "The Euclidean traveling salesman problem is NP-complete," *Theoretical Computer Science* **4** pp. 237-244 (1977).
- R.G. Parker and R.L. Rardin, "The traveling salesman problem : an update of research," *Naval Research Logistics Quarterly* **30** pp. 69-96 (1983).
- J. PickandsIII, "Moment convergence of sample extremes," *Annals of Mathematical Statistics* **39** pp. 881-889 (1968).
- F.P. Preparata and S.J. Hong, "Convex hulls of finite sets of points in two and three dimensions," *Communications of the ACM* **20** pp. 87-93 (1977).
- M. Rabin, "Probabilistic algorithms," in *Algorithms and Complexity*, ed. J. Traub,Academic Press, New York, N.Y. (1976).
- H.L. Royden, *Real Analysis*, Macmillan, London (1968).
- H. Samet, "The quadtree and related hierarchical data structures," *Computing Surveys* **16** pp. 187-260 (1984).
- H. Scheffe, "A useful convergence theorem for probability distributions," *Annals of Mathematical Statistics* **18** pp. 434-458 (1947).
- A. Schonhage, M. Paterson, and N. Pippenger, "Finding the median," *Journal of Computers and System Sciences* **13** pp. 184-199 (1976).
- E. Seneta, "Regularly Varying Functions," Springer Lecture Notes in Mathematics #508, Springer-Verlag, Berlin (1976).
- M.I. Shamos and D. Hoey, "Closest-point problems," *Proceedings of the 16th IEEE Symposium on the Foundations of Computer Science*, pp. 151-162 (1975).

- M.I. Shamos and J.L. Bentley, "Optimal algorithms for structuring geographic data," *Proceedings of the Symposium on Topological Data Structures for Geographic Information Systems*, pp. 43-51 (1977).
- M.I. Shamos, "Computational Geometry," Ph.D. Dissertation, Yale University, New Haven, Connecticut (1978).
- M.I. Shamos, 1979.
- G. Simons and N.L. Johnson, "On the convergence of binomial to Poisson distributions," *Annals of Mathematical Statistics* **42** pp. 1735-1736 (1971).
- E.V. Slud, "Distribution inequalities for the binomial law," *Annals of Probability* **5** pp. 404-412 (1977).
- J.M. Steele, "Subadditive Euclidean functionals and nonlinear growth in geometric probability," *Annals of Probability* **9** pp. 365-376 (1981).
- K.J. Supowit, E.M. Reingold, and D.A. Plaisted, "The traveling salesman problem and minimum matching in the unit square," *SIAM Journal on Computing* **12** pp. 144-156 (1983).
- M. Tamminen, "Order preserving extendible hashing and bucket tries," *BIT* **21** pp. 419-435 (1981).
- M. Tamminen, "Analysis of recursive bucket methods," Report HTKK-TKO-B44, Helsinki University of Technology, Laboratory for Information Processing Science, Otakaari 1, SF-02150 Espoo 15, Finland (1982).
- M. Tamminen, "The extendible cell method for closest point problems," *BIT* **22** pp. 27-41 (1982).
- M. Tamminen, "Analysis of N-trees," *Information Processing Letters* **16** pp. 131-137 (1983).
- M. Tamminen, "Two levels are as good as any," *Journal of Algorithms* **6** pp. 138-144 (1985).
- M. Tamminen, "On search by address computation," *BIT* **25** pp. 135-147 (1985).
- G.T. Toussaint, "Pattern recognition and geometrical complexity," pp. 1324-1347 in *Fifth International Conference on Pattern Recognition*, (1980).
- G.T. Toussaint, "Computational geometric problems in pattern recognition," pp. 73-91 in *Pattern Recognition Theory and Applications*, ed. J. Kittler, K.S. Fu and L.F. Pau,D. Reidel (1982).
- V.K. Vaishnavi and D. Wood, "Data structures for the rectangle containment and enclosure problems," *Computer Graphics and Image Processing* **13** pp. 372-384 (1980).

- V.K. Vaishnavi, "Computing point enclosures," *IEEE Transactions on Computers* **C-31** pp. 22-29 (1982).
- W.B. VanDam, J.B.G. Frenk, and A.H.G. Rinnooy Kan, "The asymptotic behaviour of a distributive sorting method," *Computing* **31** pp. 287-303 (1983).
- B.W. Weide, "Statistical Methods in Algorithm Design and Analysis," Ph.D.Dissertation, Carnegie-Mellon University, Pittsburgh, Pennsylvania (1978).
- R.L. Wheeden and A. Zygmund, *Measure and Integral*, Marcel Dekker, New York, N.Y. (1977).
- A.C. Yao, "An $O(|E| \text{loglog} |V|)$ algorithm for finding minimum spanning trees," *Information Processing Letters* **4** pp. 21-23 (1975).
- A.C. Yao, "On constructing minimum spanning trees in k-dimensional space and related problems," *SIAM Journal on Computing* **11** pp. 721-736 (1982).
- G. Yuval, "Finding nearest neighbors," *Information Processing Letters* **5** pp. 63-65 (1976).