ON THE AVERAGE DIFFERENCE BETWEEN THE SOLUTIONS TO LINEAR AND INTEGER KNAPSACK PROBLEMS*

George S. Lueker+

Technical Report #152

September 1980

Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92717

* This work was facilitated by the use of MACSYMA, a large symbolic manipulation program developed at the MIT Laboratory for Computer Science and supported by the National Aeronautics and Space Administration under grant NSG 1323, by the Office of Naval Research under grant N00014-77-C-0641, by the U.S. Department of Energy under grant ET-78-C-02-4687, and by the U.S. Air Force under grant F49620-79-C-020.

+ Supported by the National Science Foundation under grant MCS79-04997.

ABSTRACT

We analyze the expected difference between the solutions to the integer and linear versions of the 0-1 Knapsack Problem. This difference is of interest since it may help understand the efficiency of a fast backtracking algorithm for the integer 0-1 Knapsack Problem. We show that, under a fairly reasonable input distribution, the expected difference is $O(\log^2 n/n)$ and $\Omega(1/n)$.

1. INTRODUCTION

The following optimization problem is known as the 0-1 knapsack problem:

$$\text{maximize } \sum_{i=1}^{N} z_i\, a_i$$

$$\text{subject to } \sum_{i=1}^{N} z_i\, b_i \le B,$$

where a_i, b_i, and B are given and the z_i are to be either 0 or 1. This problem is known to be NP-complete [K72]. Sometimes we will refer to a version in which each z_i may be any real in the interval [0,1]; this will be called the *relaxed* version, as opposed to the *integer* version above, and may easily be solved exactly in O(n log n) time by a greedy algorithm. See, for example, [HS78]. Because of the importance and simple structure of the 0-1 integer knapsack problem, it has been the subject of extensive investigation. For example, it is known [IK75] that it admits a fully polynomial time approximation scheme [GJ79]; that is, we may obtain a worst-case relative error of ε , for any $\varepsilon>0$, by an algorithm whose time is bounded by a polynomial in N and ε^{-1}. See [A78] for an analysis of an algorithm which works well on the average under certain assumptions about the input distribution. The problem also lends itself readily to solution by a backtracking approach; the search tree can be pruned whenever the solution obtained by using the items not yet considered according to the relaxed constraint is not as good as the best integer solution seen previously. See [HS78] for a detailed discussion of this approach. When applied to randomly generated data, this approach, which always yields the exact optimum, seems to run very rapidly even for large values of N; in fact, it seems possible that its expected time is polynomial in N. A proof of this would be very interesting, but probably difficult. A first step towards such a proof might be to obtain a better understanding of the difference between the optimum solutions to the integer and relaxed versions of the problem. (In general, determining the quality of the heuristics that guide a search is useful for understanding the quality of the search algorithm; see, for example, [G77].) This is the goal of this paper.

We will assume that the a_i and b_i are chosen uniformly from the interval [0,1]. Thus the selection of the parameters of the N items can be viewed as the placement of N points at random in the unit square. In order to simplify the analysis, we will assume that N is drawn from a Poisson distribution with parameter n; this will cause the number of points in disjoint parts of the square to be completely independent. (For large n, N will tend to be nearly equal to n.) We will assume that the items are numbered so that the profit density (a_i/b_i) is decreasing. In order to try to cause a constant fraction of the items to be used in the solution as n becomes large, we will assume that for some fixed β, $B=\beta n$. For later convenience, we assume that β lies in the open interval (1/6,1/2); it is not hard to show that this means that the relaxed solution will, almost surely as $n\to\infty$, use more than half but less than all of the items. For a given n, the random problem created this way will be referred to as P_n. The greedy method can be visualized by imagining a ray, which we shall call the profit density ray, which passes through the origin and rotates clockwise; as this ray rotates from pointing up to pointing to the right, it intersects the points in the order in which they are considered. Let $\bar{m}$ be the limit as $n\to\infty$ of the average slope of this ray at the point when the greedy method for the relaxed version fills the knapsack. It is not difficult to show that

$$\beta = \iint_A x\, dx\, dy,$$

where A is the area shown in Figure 1. Then if we let

$$\alpha = \iint_A y\, dx\, dy,$$

it can be shown that the average optimum, to the integer or relaxed version, is asymptotic to αn. By our assumption on β, $\bar{m}$ is in the open interval (0,1); this means that $\bar{m}$ is such that the profit density ray intersects the right edge of the square.

Since the linear and integer solutions are asymptotic to each other, it might not seem interesting to compare them. To obtain an interesting problem, we will look not at the

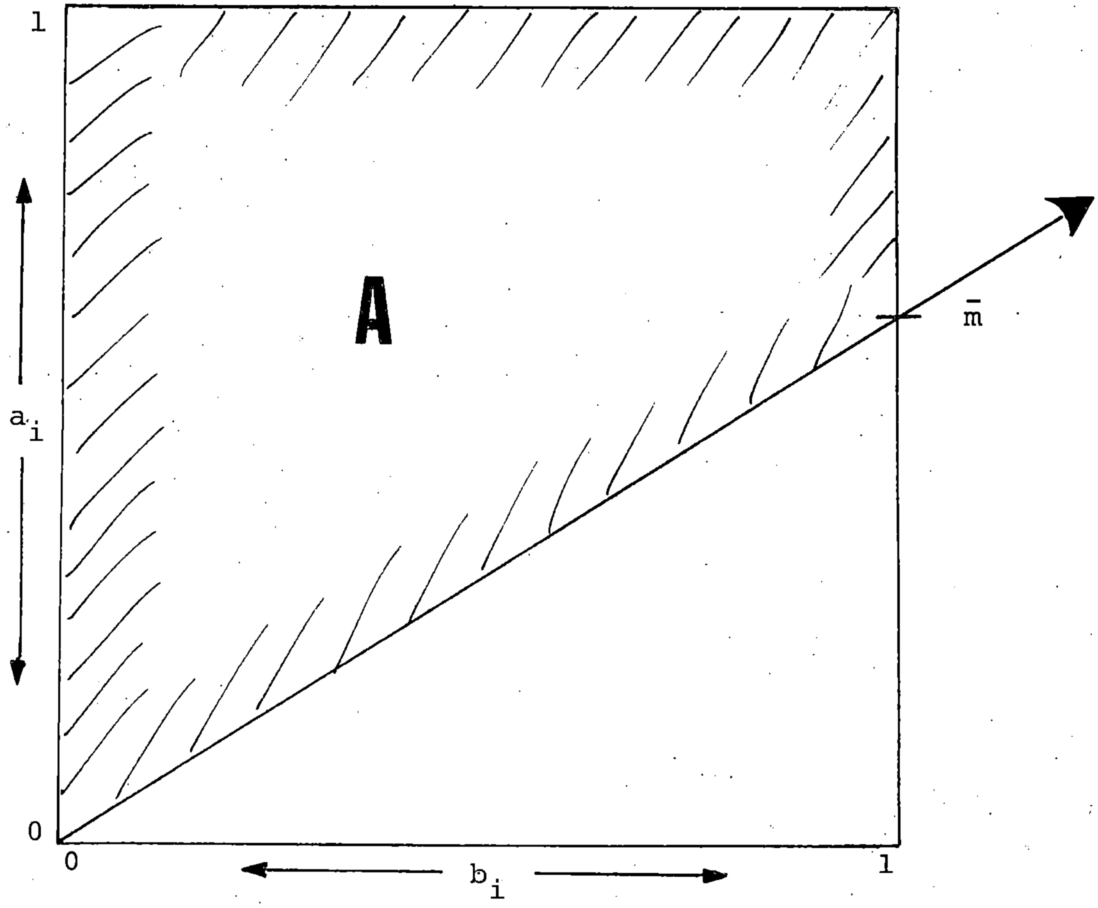

Figure 1. The profit density ray.

ratios of the results, but rather at their differences. In [BZ77] it is observed empirically that for certain data this difference decreases as N increases; this is attributed to the fact that as N decreases, more variables tend to lie in a region of small profit density change, which increases the chances of finding an integer solution with a value close to the relaxed optimum. The results presented in this paper formally establish that the average difference is $O((\log n)^2/n)$, and $\Omega(1/n)$, under our input distribution.

2. A THEOREM ABOUT SUMS OF SUBSETS

Before investigating the knapsack problem further, it is useful to consider the following problem about sums of subsets of random variables. We are given 2k random variables, and we wish to find a subset whose sum is as close as possible to some target x_k. How close can we hope to come? (See [AP80] for an analysis of an algorithm for a related subset sum problem. The method to be used below is nonconstructive, and gives an exponentially tighter bound.)

The following theorem provides a partial answer to this question. Since it appears to be of interest in its own right, we state it in a more general form than is needed for section 3.

Theorem 1. Let g be the probability density function of a variable which assumes values in $[-a,a]$. Suppose g is bounded and has mean 0 and variance 1. Let x_k be a real sequence with $x_k = o(k^{1/2})$. Suppose we draw 2k variables $X_1,\ldots,X_{2k}$ according to g. Then for large enough k, the probability that some subset of k of the 2k variables has a sum in $[x_k-\varepsilon, x_k+\varepsilon]$ is at least 1/2, provided $\varepsilon=7k\ 4^{-k}$.

Proof. A bit of notation is useful. Let G be the cumulative distribution function corresponding to g. Let G_n (resp. g_n) be the cumulative distribution (resp. density) function for the sum of n variables drawn according to g. Let F_n (resp. f_n) be the cumulative distribution (resp. density) function for the sum of n unit normal variables. Hence

$$f_n = C_n e^{-\frac{x^2}{2n}}, \qquad (2.1a)$$

where

$$C_n = \frac{1}{\sqrt{2\pi n}} \, . \qquad (2.1b)$$

Let Y_k be the random variable which tells the number of distinct subsets of size k whose sums lie in $[x_k-\varepsilon, x_k+\varepsilon]$. We seek to show $P\{Y_k > 0\} \geq 1/2$, provided $\varepsilon = 7k4^{-k}$.

First note that the expectation of Y_k is

$$\begin{aligned} E[Y_k] &= \binom{2k}{k} [G_k(x_k+\varepsilon) - G_k(x_k-\varepsilon)] \\ &\sim \binom{2k}{k} f_k(x_k) \; 2\varepsilon \\ &\sim 2\binom{2k}{k} C_k \; \varepsilon \\ &= 2\binom{2k}{k} \varepsilon/\sqrt{2\pi k}, \end{aligned}$$

where we have employed [F66, Theorem 1, page 506] and the fact that $x_k/\sqrt{k} \to 0$. A simple asymptotic analysis of the right hand term shows that it is about 3 for ε as in the lemma. This in itself, however, gives us no proof that the probability that Y is zero is small.

Fortunately, a clever method known as the "second-moment method" (see, for example, [ES74, ER60, BE76, M70]) is useful here; we use the following well-known corollary of Chebyshev's inequality, which holds for arbitrary random variables Y:

$$P\{Y=0\} \le \frac{E[Y^2]}{(E[Y])^2} - 1. \tag{2.2}$$

The computation of $E[Y_k^2]$ is a bit messy, and is deferred to Lemma A1 in the appendix; there it is shown that as $k \to \infty$ and $\varepsilon = o(k^{-1})$,

$$E[Y_k^2] \sim \frac{\binom{2k}{k} 2\varepsilon}{\sqrt{2\pi k}} + \frac{\binom{2k}{k}^2 4\varepsilon^2}{\sqrt{3}\pi k}$$

Hence

$$\frac{E[{Y_k}^2]}{E[Y_k]^2} \sim \frac{4\sqrt{2}\,\binom{2k}{k}\varepsilon + 2\sqrt{3\pi k}}{2\sqrt{6}\,\binom{2k}{k}\varepsilon}$$

Some asymptotic analysis shows that if ε grows as $\alpha k 4^{-k}$, this ratio approaches

$$\frac{2\sqrt{2}\alpha + \sqrt{3}\pi}{\sqrt{6}\alpha} \tag{2.3}$$

Letting $\alpha=7$ causes this expression to achieve a value just under 1.5, which, in view of (2.2), establishes the result. ■

It is interesting to note that letting α become very large does not cause (2.3) to approach 1; rather, it approaches $\sqrt{4/3}$. Thus, to show that a large ε gives a very small probability of failure, some different argument would be needed. Note, on the other hand, that if $\varepsilon = o(k4^{-k})$, one easily shows that $E[Y_k]$ approaches 0, so the probability of finding the desired subset of cardinality k approaches 0.

A similar theorem could be obtained for a more general class of density functions, but we will not pursue that further here.

3. AN UPPER BOUND ON THE AVERAGE DIFFERENCE

Let P_n denoted a random problem generated as explained in the introduction; let INTEGER(P_n) and RELAXED(P_n) denoted the value of the optimum solutions to the integer and relaxed versions of this problem. In order to bound the difference between these solutions, we will employ a procedure, named APPROX, which constructs a feasible solution to the integer version; it appears below. For comparison, we have also presented the greedy procedure which solves the relaxed problem exactly.

Theorem 2. $E[\text{RELAXED}(P_n)-\text{INTEGER}(P_n)]=0(\log^2 n/n)$.

Proof. Since APPROX gives a lower bound on the true optimum, we may bound the difference between INTEGER and RELAXED by that between APPROX and RELAXED. Now the deviation between APPROX and RELAXED is attributable to two causes:

a) we do not completely fill the knapsack during APPROX, and

b) the part we do fill may be filled with items of a lower profit density.

For part (a), note that if the branch to OUT is taken in APPROX, the unused part of the knapsack has size at most 2ε, which is $0(\log n/n)$; the probability that the branch to OUT is never taken can be shown to be $0(n^{-e})$ for any positive integer e. Now consider part (b). By Theorem 1, the probability of success on a single iteration of the second while-loop is at least 1/2. (Note that in this application of that Theorem, the random variables being summed have mean 1/2 and variance 1/12,

```
procedure APPROX;
begin
   BB := B; i := 1; A := 0;
   k := $\lfloor \log_4 n \rfloor$; $\varepsilon$ := $(7/\sqrt{12})k4^{-k}$;
   while BB $\geq$ k/2 and i < N do
      begin
         BB := BB - $b_i$;
         A := A + $b_i$;
         i := i + 1;
      end;
   comment at this point k/2-1 $\leq$ BB $\leq$ k/2;
   while i + 2k $\leq$ N do
      begin
         for all subsets S of {i+1,i+2,...,i+2k} do
            begin
               if the sum of the $b_j$ values over all j in S
                 lies in [BB-2$\varepsilon$,BB] then go to OUT;
            end;
         i := i + 2k;
      end;
   S := the empty set;
   OUT: A := A + the sum of the $a_j$ values over all j in S;
   return A;
end;


procedure RELAXED;
begin
   BB := B; i := 1; A := 0;
   while BB > $b_i$ and i < N do
      begin
         BB := BB - $b_i$;
         A := A + $a_i$;
         i := i + 1;
      end
   A := A + $a_i$ * max(1,BB/$b_i$);
   return A;
end;
```

so some scaling is required.)

Since successive iterations are independent, the expected number of iterations is $0(1)$. Now the extent to which the profit density ray advances at each iteration is independent of the values of the b_i, and can readily be seen to have an expectation of $0(\log n/n)$; this change in density applies to a portion of the knapsack whose capacity is at most $k/2=O(\log n)$ so the expected contribution due to part (b) above is $0(\log^2 n/n)$. ■

4. A LOWER BOUND ON THE AVERAGE DIFFERENCE

An interesting question is whether the bound on the expected difference between the integer and relaxed solutions stated in Theorem 2 is tight. Although we have not been able to answer that question, we have established the following lower bound.

Theorem 3. $E[\text{RELAXED}(P_n) - \text{INTEGER}(P_n)] = \Omega(1/n)$.

Proof Sketch. We will describe a boolean procedure with the following two properties:

a) It returns *true* with probability of at least 1/4 for large n, and

b) if it returns *true*, then for this problem instance the relaxed and integer solutions differ by $\Omega(1/n)$.

From this the Theorem follows readily.

TEST proceeds as follows. First it fills the knapsack as in RELAXED until the remaining capacity BB satisfies

$$1 < BB \le 2. \qquad (4.1)$$

Henceforth in the proof we fix BB at this value; let $\hat{\rho}_0$ be the profit density of the last item used. The procedure rejects (i.e., returns *false*) if the condition (4.1) cannot be met; since all b_i are in [0,1], rejection occurs here only if we run out of items to use in the knapsack, and this occurs with exponentially small probability. TEST also rejects if the profit density ray has not yet advanced past the upper right corner of the square, which again can be seen to have exponentially small probability under our model of input distribution.

Next we look at the next four points as the profit density ray advances. (The probability that fewer than this number remain is again exponentially small.) Call their profits, costs, and densities $\hat{a}_i$, $\hat{b}_i$, and $\hat{p}_i$, respectively, for $i=1,\ldots,4$. Reject unless

$$\hat{p}_0-\hat{p}_1 \geq \frac{1}{20n} \tag{4.2a}$$

$$\hat{p}_3-\hat{p}_4 \geq \frac{1}{20n} \tag{4.2b}$$

$$\hat{p}_4 \geq \bar{m}/2 \tag{4.3}$$

Note that the movement of the profit density ray between these items has an exponential distribution with mean $2/n$; hence the probability of rejection in (4.2a) or (4.2b) is less than (1/20) each, for a total of 1/10. The probability of rejection in (4.3) can be seen to be exponentially small, since it means that we have gone far beyond the $\bar{m}$ of Figure 1.

Next we impose some restrictions on the $\hat{b}_i$ values. (Note that each has, independently, a density function of $2x$ for $x\in[0,1]$.) Reject if <u>any</u> subset of these four values has a sum in the range $BB\pm\frac{1}{600}$. Note that the sum of any fixed nonempty subset of the b_i has a density function uniformly bounded by 2; hence, for any such subset, the probability that its sum lies in the indicated range in at most $\frac{1}{150}$. On the other hand, there are only 15 nonempty subsets, so the probability of rejection here is at most $\frac{1}{10}$. Finally, reject unless

$$\hat{b}_i < BB < \hat{b}_1+\hat{b}_2+\hat{b}_3. \tag{4.4}$$

The left inequality is always true since $1<BB$. The right inequality has probability greater than 1/2; this can be seen by noting that $BB<2$, and performing a tedious but straightforward computation involving convolution of the densities of the $\hat{b}_i$. At this point the description of TEST is complete.

Now since the probability of the union of several events is bounded by the sum of their probabilities, we see that

$$P\{\mathrm{TEST}(P_n) = \underset{\sim\sim\sim}{\textit{false}}\} \le \frac{1}{10} + \frac{1}{10} + \frac{1}{2} + o(1),$$

which is smaller than 0.75 for large enough n, so condition (a) holds.

Next we establish condition (b). Assume that TEST returns *true*. Then we know from (4.4) that the procedure RELAXED fills the knapsack when the profit density ray is lying in the area labeled β_2 in Figure 2. Let B_α, B_β, and B_γ be the total knapsack capacity used in the relaxed solution by items lying in regions α, $\beta_1 \cup \beta_2 \cup \beta_3$, and γ (respectively). Note that $B_\beta = BB$, and $B_\gamma = 0$. Now consider the optimum solution to the integer problem; define $\tilde{B}_\alpha$, $\tilde{B}_\beta$, and $\tilde{B}_\gamma$ for this solution analogously to B_α, B_β, and B_γ. Now by the restriction TEST imposed on sums of subsets of the $\hat{b}_i$, we know that $|B_\beta - \tilde{B}_\beta| \ge \frac{1}{600}$. Hence it can be seen that at least one of the following three conditions must hold:

$$\tilde{B}_\alpha + \tilde{B}_\beta + \tilde{B}_\gamma \le B_\alpha + B_\beta + B_\gamma - \frac{1/3}{600} \tag{4.5}$$

$$\tilde{B}_\alpha \le B_\alpha - \frac{1/3}{600} \tag{4.6}$$

$$\tilde{B}_\gamma \ge B_\gamma + \frac{1/3}{600} \tag{4.7}$$

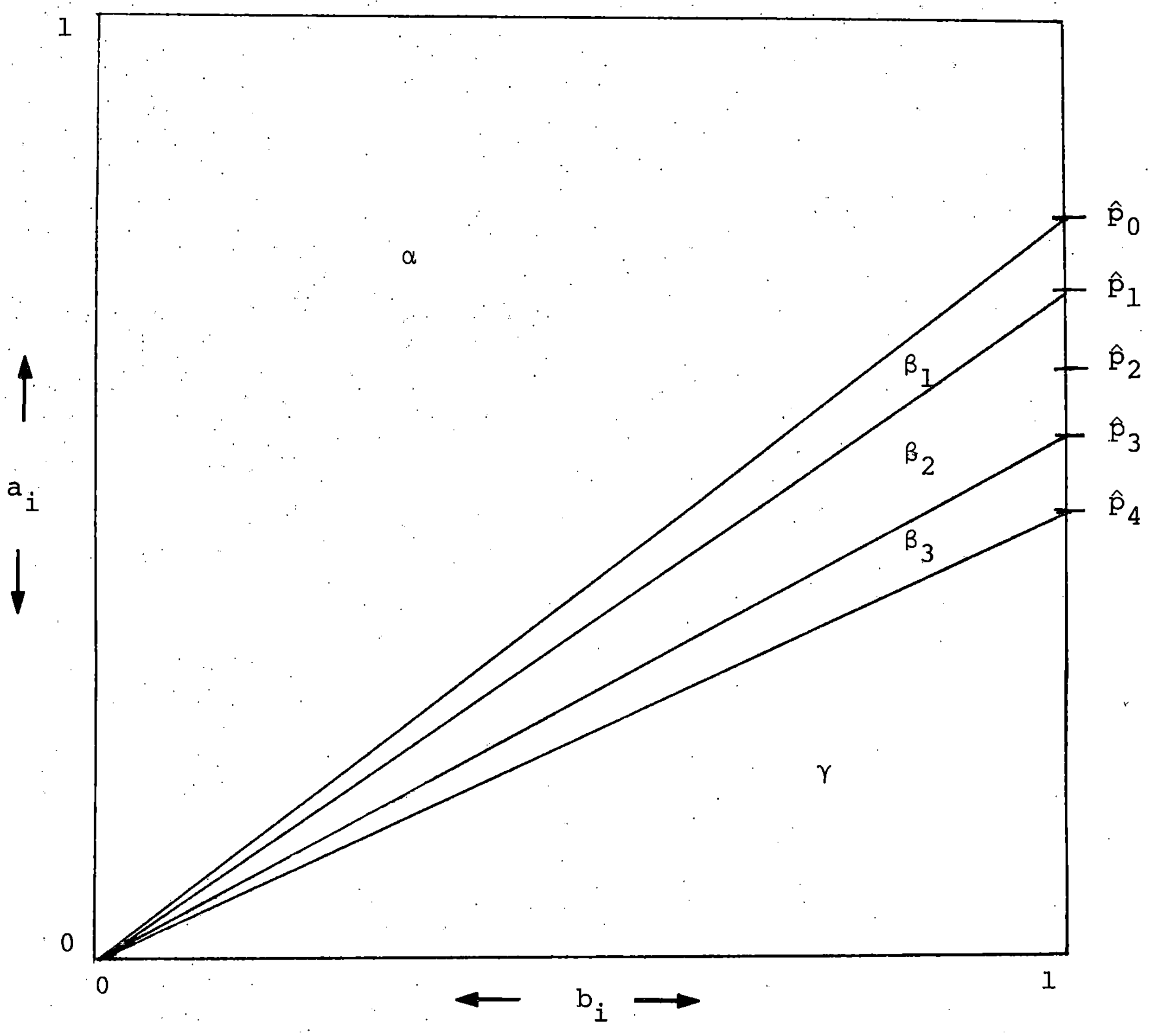

Figure 2. Illustration for the lower bound on the difference. Regions are to include the segment bounding them from below, but not that bounding them above.

If (4.5) holds, at least $\frac{1}{1800}$ units of the knapsack is being wasted, from which it is not hard to see that the difference between the two solutions is $\Omega(1)$. If (4.6) holds, then since in the relaxed solution all items in $\alpha \cup \beta_1$ were used, at least $\frac{1}{1800}$ units of capacity has been shifted by the integer solution from α to $\beta_2 \cup \beta_3 \cup \gamma$, and hence by (4.2a) experienced a decrease of $\frac{1}{20n}$ in profit density; thus the integer solution is worse by $\Omega(1/n)$. A similar argument holds for case (4.7). ■

APPENDIX. Analysis of $E[Y_k^2]$.

Lemma A1. If Y_k is defined as in the proof of Theorem 1, then if $k\to\infty$ and $\varepsilon=o(1/k)$,

$$E[Y_k^2] \sim \frac{\binom{2k}{2}2\varepsilon}{\sqrt{2\pi k}} + \frac{\binom{2k}{k}^2 4\varepsilon^2}{\sqrt{3}\pi k}$$

Proof. Our computation, which is similar to an analogous computation in [BE76], will be facilitated by some further notation. Let $\mathcal{S}$ be the set of all $\binom{2k}{k}$ choices of k elements of $\{1,2,\ldots,2k\}$. If S is a set in $\mathcal{S}$, let Z(S) be the random variable which is one if

$$\sum_{i\in S} X_i \in [x_k-\varepsilon,\ x_k+\varepsilon],$$

and 0 otherwise. Then

$$E[Y_k^2] = \sum_{S_1\in\mathcal{S}} \sum_{S_2\in\mathcal{S}} E[Z(S_1)\ Z(S_2)]. \tag{A.1}$$

It is convenient to reorganize the sum according to the number of elements which S_1 and S_2 share. Let I_{km} be $E[Z(S_1)Z(S_2)]/(2\varepsilon)^2$, assuming that S_1 and S_2 have m elements in common. If we now consider the number of ways in which S_1 and then S_2 may be chosen, we see that (A.1) becomes

$$E[Y^2]=\binom{2k}{k}(2\varepsilon)^2 \sum_{m=0}^{k}\binom{k}{m}\binom{k}{k-m}I_{km} \tag{A.2}$$

We shall break the sum into three parts, as follows.

a) m=k. Then S_1 and S_2 are identical, so

$$I_{kk} = \frac{1}{(2\varepsilon)^2} E\ [Z(S_1)\ Z(S_2)] = \frac{1}{(2\varepsilon)^2} E\ [Z(S_1)]$$

$$= \frac{1}{(2\varepsilon)^2} (G_k(x_k+\varepsilon) - G_k(x_k-\varepsilon))$$

$$\sim \frac{1}{(2\varepsilon)^2} C_k 2\varepsilon = \frac{1}{2\varepsilon\sqrt{2\pi k}}$$

The cases where m≠k will be handled next. Note that if m≠k, we have

$$I_{km} = \frac{1}{(2\varepsilon)^2} \int_{-\infty}^{\infty} g_m(x_k-x)\ [G_{k-m}(x+\varepsilon)-G_{k-m}(x-\varepsilon)]^2 dx.$$

This follows by considering all possible values for the sum of the m variables common to S_1 and S_2, and considering the probability that both of the sets of k-m remaining elements in S_1 and S_2 bring the sum to $x_k \pm \varepsilon$.

b) m<k, and m>3k/4 or m<k/4. Let M be the set of m that satisfy these inequalities. Since we know that g is bounded, say by B, we know that g_n is also bounded by B and hence $G_n(x+\varepsilon)-G_n(x-\varepsilon)$ is bounded by $2\varepsilon B$. Since g_m has unit mass, we see that

$$I_{km} \le \frac{1}{(2\varepsilon)^2} \ (2\varepsilon B)^2 = B^2$$

Thus the contribution to the summation in (A.2) from these terms is bounded by

$$\sum_{m \in M} \binom{k}{m} \binom{k}{k-m} I_{km}$$

$$\leq 2 \sum_{m=0}^{\lceil k/4 \rceil - 1} \binom{k}{m}^2 B^2 = \Theta\left(\binom{k}{\lceil k/4 \rceil}^2\right).$$

c) $k/4 \leq m \leq 3k/4$. (Thus $m = \Theta(k)$.) We begin by noticing that for m in this range,

$$I_{km} = \frac{1}{2\pi\sqrt{k^2 - m^2}} + o(k^{-1}). \tag{A.3}$$

(The proof of this is somewhat messy and is separated out as Lemma A.2.) Now in the sum

$$\sum_{m=\lceil k/4 \rceil}^{\lfloor 3k/4 \rfloor} \binom{k}{m}^2 I_{km}, \tag{A.4}$$

$\binom{k}{m}$ is sharply peaked about $m=k/2$ as k becomes large, while I_{km} is much more slowly varying. Using this observation (as in [P77]), and using (A.3), one may establish that (A.4) is

$$\left(\sum_{m=0}^{k} \binom{k}{m}^2\right) \left(\frac{1}{2\pi\sqrt{k^2 - (k/2)^2}} + o(k^{-1})\right)$$

$$= \binom{2k}{k} \left(\frac{1}{\sqrt{3}\pi k} + o(k^{-1})\right)$$

Adding the contributions from parts (a), (b), and (c), and noting that

$$\binom{k}{\lceil k/4 \rceil}^2 = o\left(\binom{2k}{k}/k\right),$$

we see that (A.2) becomes

$$E[Y_k{}^2] \sim \binom{2k}{k}(2\varepsilon)^2 \left[\frac{1}{2\varepsilon\sqrt{2\pi k}} + \binom{2k}{k}\frac{1}{\sqrt{3}\pi k}\right]$$

$$= \frac{\binom{2k}{k}2\varepsilon}{\sqrt{2\pi k}} + \frac{\binom{2k}{k}^2 4\varepsilon^2}{\sqrt{3}\pi k}\ . \qquad \blacksquare$$

<u>Lemma A.2</u>. Assuming $k/4 \le m \le 3k/4$,

$$I_{km} = \frac{1}{2\pi\sqrt{k^2-m^2}} + o(k^{-1}),$$

where

$$I_{km} = \frac{1}{(2\varepsilon)^2}\int_{-\infty}^{\infty} g_m(x_k - x)\,[G_{k-m}(x+\varepsilon) - G_{k-m}(x-\varepsilon)]^2 dx$$

<u>Proof</u>. We shall use the fact that under our hypotheses (in fact under weaker hypotheses),

$$g_n(x) = (1 + c(\frac{x^3}{n^2} - 3\frac{x}{n}))\ f_n(x) + o(n^{-1}), \qquad \text{(A.5)}$$

for some c which depends only on the distribution g. (This is Theorem 1 on page 506 in [Fe66]. Our notation is different, and we have scaled the axes by factors of $\sqrt{n}$ relative to that Theorem.)

For convenience let

$$P_n(x) = c(\frac{x^3}{n^2} - 3\frac{x}{n}). \qquad \text{(A.6)}$$

Note that $P_n(x)$ of x has the form of $n^{-1/2}$ times a polynomial in $(x/\sqrt{n})$, i.e.,

$$P_n(x) = \frac{(x/\sqrt{n})^3 - 3(x/\sqrt{n})}{\sqrt{n}}; \tag{A.7}$$

this fact will be useful later.

By (A.5) and (A.6), and since $m=\Theta(k)$,

$$g_m(x_k-x) = (1+P_m(x_k-x))\ f_m(x_k-x) + o(k^{-1}). \tag{A.8}$$

It is not hard to see that, since ε is $o(1/k)$ and the right side of (A.5) does not vary too rapidly,

$$g_n(x\pm\varepsilon) = g_n(x) + o(n^{-1}), \quad \text{for } n = \Theta(k),$$

so

$$\begin{aligned} &G_{k-m}(x+\varepsilon) - G_{k-m}(x-\varepsilon) \\ &\quad = 2\varepsilon\ (g_{k-m}(x) + o(k^{-1})) \\ &\quad = 2\varepsilon\ [(1 + P_{k-m}(x))\ f_{k-m}(x) + o(k^{-1})]. \end{aligned} \tag{A.9}$$

Now by (A.8) and (A.9) we obtain

$$I_{km}=\int_{-\infty}^{\infty} (1+P_m(x_k-x))f_m(x_k-x)(1+P_{k-m}(x))^2 f_{k-m}^2(x)dx+o(k^{-1}), \tag{A.10}$$

where the movement of the $o(k^{-1})$ out from the factors is justified using the fact that the area under $g_m(x_k-x)$ and the area under $(1+P_{k-m}(x))^2 f_{k-m}^2(x)$ are both $O(1)$.

Since P_n has the form given in (A.7), and since $m=\Theta(k)$ and $x_k/\sqrt{k}=o(1)$, we may write (A.10) as

$$I_{km}= \int_{-\infty}^{\infty} (1+\frac{1}{\sqrt{k}}\ Q_{km}(\frac{x}{\sqrt{k}}))f_m(x_k-x)f_{k-m}^2(x)dx+o(k^{-1}), \tag{A.11}$$

where Q_{km} is a polynomial of degree 9 whose coefficients are uniformly bounded as m and k vary. The magnitude of the contribution to the integral due to Q_{km} is

$$\left| \int_{-\infty}^{\infty} \frac{1}{\sqrt{k}} Q_{km}\left(\frac{x}{\sqrt{k}}\right) f_m(x_k - x) f_{k-m}^2(x)\,dx \right|$$

$$\leq \int_{-\infty}^{\infty} \frac{1}{\sqrt{k}} \left| Q_{km}\left(\frac{x}{\sqrt{k}}\right) \right| C_m C_{k-m}^2 e^{-\left(\frac{x}{\sqrt{2m}}\right)^2}\,dx, \qquad \text{(A.12)}$$

where we have expanded f_m and f_{k-m} (see (2.1)), ignoring the negative exponential in the latter. Now since for any polynomially bounded function p(x),

$$\int_{-\infty}^{\infty} p(x/c)\, e^{-(x/c)^2}\,dx = O(c),$$

one may establish that (A.12) is

$$\frac{1}{\sqrt{k}} C_m C_{k-m}^2\, O(\sqrt{k}) = O(k^{-3/2}).$$

Thus (A.11) reduces to

$$I_{km} = \int_{-\infty}^{\infty} f_m(x_k - x)\, f_{k-m}^2(x)\,dx + o(k^{-1}).$$

At this point the integral may readily be computed in closed form, and since $x_k/\sqrt{k}$ approaches zero it can be shown to be

$$I_{km} = \frac{e^{-\frac{x_k^2}{k+m}}}{2\pi\sqrt{k^2 - m^2}} + o(k^{-1}) = \frac{1}{2\pi\sqrt{k^2 - m^2}} + o(k^{-1}).$$

References

[A78] G. d'Atri, "Probabilistic Analysis of the Knapsack Problem," Technical Report No. 7, Groupe de Recherche 22, Centre National de la Recherche Scientifique, Paris, October 1978.

[AP80] G. d'Atri and C. Puech, "Probabilistic Analysis of the Subset-Sum Problem," Technical Report No. 1 (1980), Dipartimento di Matematica, Universita' Della Calabria, Italy, March 1980.

[BZ77] Egon Balas and Eitan Zemel, "Solving Large Zero-One Knapsack Problems," Management Sciences Research Report No. 408, Carnegie-Mellon University, July 1977.

[BE76] B. Bollobås and P. Erdös, "Cliques in Random Graphs," Math. Proc. Camb. Phil. Soc. 80 (1976), pp. 419-427.

[ES74] P. Erdös and J. Spencer, Probabilistic Methods in Combinatorics, Academic Press, New York, 1974.

[ER60] P. Erdös and A. Rėnyi, "On the Evolution of Random Graphs," Publ. Math. Inst. Hung. Acad. Sci. 5A (1960), pp. 17-61.

[F66] William Feller, An Introduction to Probability Theory and Its Applications, Volume II, John Wiley and Sons, New York, 1966.

[GJ79] M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman and Company, San Francisco, 1979.

[G77] John Gaschnig, "Exactly How Good are Heuristics?: Toward a Realistic Predictive Theory of Best-First Search," Proc. Intl. Joint Conf. on Artificial Intelligence, Cambridge, Mass., August 1977.

[HS78] E. Horowitz and S. Sahni, Fundamentals of Computer Algorithms, Computer Science Press, Potomac, Maryland, 1978.

[IK75] O. H. Ibarra and C. E. Kim, "Fast Approximation Algorithms for the Knapsack and Sum of Subset Problems," JACM 22:4 (October 1975), pp. 463-468.

[K72] R. M. Karp, "Reducibility among Combinatorial Problems," in R. E. Miller and J. W. Thatcher, eds., Complexity of Computer Computations, Plenum Press, New York, 1972, pp. 85-104.

[M70] L. Moser, "The Second Moment Method in Combinatorial Analysis," in Combinatorial Structures and Their Applications, Gordon and Breach, New York, 1970.

[P77] Nicholas Pippenger, "An Information-Theoretic Method in Combinatorial Theory," J. Comb. Th. 23:1 (July 1977), pp. 99-104.